BIOTECHNO 2011

The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

May 22-27, 2011

Venice/Mestre, Italy

BIOTECHNO 2011 Editors

Pei-Yuan Qian, KAUST, University of Science and Technology, Hong Kong

Son V. Nghiem, Jet Propulsion Laboratory / California Institute of Technology - Pasadena, USA

BIOTECHNO 2011

Foreword

The Third International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2011), held between May 22 -27, 2011 in Venice, Italy, covered the these three main areas: bioinformatics, biocomputational systems, and biotechnologies.

Bioinformatics deals with the system-level study of complex interactions in biosystems providing a quantitative systemic approach to understand them and appropriate tool support and concepts to model them. Understanding and modeling biosystems requires simulation of biological behaviors and functions. Bioinformatics itself constitutes a vast area of research and specialization, as many classical domains such as databases, modeling, and regular expressions are used to represent, store, retrieve and process a huge volume of knowledge. There are challenging aspects concerning biocomputation technologies, bioinformatics mechanisms dealing with chemoinformatics, bioimaging, and neuroinformatics.

Brain-computing, biocomputing, and computation biology and microbiology represent advanced methodologies and mechanisms in approaching and understanding the challenging behavior of life mechanisms. Using bio-ontologies, biosemantics and special processing concepts, progress was achieved in dealing with genomics, biopharmaceutical and molecular intelligence, in the biology and microbiology domains. The area brings a rich spectrum of informatics paradigms, such as epidemic models, pattern classification, graph theory, or stochastic models, to support special biocomputing applications in biomedical, genetics, molecular and cellular biology and microbiology. While progress is achieved with a high speed, challenges must be overcome for large-scale bio-subsystems, special genomics cases, bio-nanotechnologies, drugs, or microbial propagation and immunity.

Biotechnology is defined as the industrial use of living organisms or biological techniques developed through basic research. Bio-oriented technologies became very popular in various research topics and industrial market segments. Current human mechanisms seem to offer significant ways for improving theories, algorithms, technologies, products and systems. The focus is driven by fundamentals in approaching and applying biotechnologies in terms of engineering methods, special electronics, and special materials and systems. Borrowing simplicity and performance from the real life, biodevices cover a large spectrum of areas, from sensors, chips, and biometry to computing. One of the chief domains is represented by the biomedical biotechnologies, from instrumentation to monitoring, from simple sensors to integrated systems, including image processing and visualization systems. As the state-of-the-art in all the domains enumerated in the conference topics evolve with high velocity, new biotechnologes and biosystems become available. Their rapid integration in the real life becomes a challenge.

We welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard forums or in industry consortia, survey papers addressing the key problems and solutions on any of the above topics short papers on work in progress, and panel proposals.

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2011 technical program committee as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors that dedicated much of their time and efforts to contribute to BIOTECHNO 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We hope that BIOTECHNO 2011 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in biotechnology.

We are certain that the participants found the event useful and communications very open. We also hope the attendees enjoyed the beautiful surroundings of Venice.

BIOTECHNO 2011 Chairs
Olivier Bodenreider, National Institutes of Health - Bethesda, USA
Cristina Seceleanu, Mälardalen University, Sweden
Saman Kumara Halgamuge, University of Melbourne, Australia
Gunnar Klau, Centrum Wiskunde & Informatica, The Netherlands
Ravi Radhakrishnan, University of Pennsylvania - Philadelphia, USA

BIOTECHNO 2011

Committee

BIOTECHNO Advisory Chairs

Olivier Bodenreider, National Institutes of Health - Bethesda, USA
Cristina Seceleanu, Mälardalen University, Sweden
Saman Kumara Halgamuge, University of Melbourne, Australia
Gunnar Klau, Centrum Wiskunde & Informatica, The Netherlands
Ravi Radhakrishnan, University of Pennsylvania - Philadelphia, USA

BIOTECHNO 2011 Technical Program Committee

Luis Alexandre, University of Beira Interior, Portugal
Basim Alhadidi, Albalqa' Applied University - Salt, Jordan
Christine Amaldas, RMIT International University Vietnam - Hanoi, Vietnam
Nizamettin Aydin, Yildiz Technical University - Istanbul, Turkey
Ganesharam Balagopal, Ontario Ministry of the Environment, Canada
Siegfried Benkner, University of Vienna, Austria
Tom Bersano, University of Michigan, USA
Christian Blum, Universitat Politècnica de Catalunya, Spain
Razvan Bocu, University College Cork, Ireland
Olivier Bodenreider, National Institutes of Health - Bethesda, USA
Magnus Bordewich, Durham University, UK
Sabin-Corneliu Buraga, "A. I. Cuza" University - Iasi, Romania
Paul Bustamante, CEIT and TECNUN, Spain
Yang Cao, Virginia Tech - Blacksburg, USA
Keith C.C. Chan, The Hong Kong Polytechnic University - Kowloon, Hong Kong
Yili Chen, Monsanto - St. Louis, USA
David Corne, Heriot-Watt University-Edinburgh, UK
Coral del Val Muñoz, Universidad de Granada, Spain
Benjamin Doerr, Max-Planck-Institut für Informatik-Saarbrücken, Germany
Rolf Drechsler, University of Bremen, Germany
Andreas Dress, Shanghai Institutes for Biological Sciences (SIBS) / Chinese Academy of Sciences (CAS) , China
Victor Felea, "Al.I.Cuza" University - Iasi, Romania
Bin Fu, University of Texas-Pan American - Edinburg, USA
Xin Gao, Carnegie Mellon University, USA
Alejandro Giorgetti, University of Verona, Italy
Paul Gordon, University of Calgary, Canada
Jun-Tao Guo, The University of North Carolina at Charlotte, USA
Steffen Heber, North Carolina State University-Raleigh, USA
Yaochu Jin, University of Surrey, UK
Attila Kertesz-Farkas, International Centre of Genetic Engineering and Biotechnology, Italy
Daisuke Kihara, Purdue University - West Lafayette, USA
DaeEun Kim, Yonsei University - Seoul, South Korea

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Singular Spectral Analysis on Doppler Ultrasound Records Acquired from Healthy Subjects and Intrinsic Sphincter Deficiency Subjects 1
Kadir Tufan, Sadik Kara, Fatma Latifoglu, Sinem Aydin, Adem Kiris, and Unsal Ozkuvanci

Separating Drivers from Passengers in Whole Genome Analysis: Identification of Combinatorial Effects of Genes by Mining Knowledge Sources 6
Stephen Anthony, Vitali Sintchenko, and Enrico Coiera

Bio-UnaGrid: Easing Bioinformatics Workflow Execution Using LONI Pipeline and a Virtual Desktop Grid 12
Mario Villamizar, Harold Castro, David Mendez, Silvia Restrepo, and Luis Rodriguez

Automated Identification of Micro-Embolic Events Using Auditory Perception Features Extracted from Mel Frequency Cepstrum Coefficients 20
Lingke Fan and David Evans

SimGenex: A System for Concisely Specifying Simulation of Biological Processes and Experimentation 26
Anyela Camargo and Jan T. Kim

Automated Image Processing for the Analysis of DNA Repair Dynamics 31
Thorsten Riess, Christian Dietz, Martin Tomas, Elisa Ferrando-May, and Dorit Merhof

Quantifying the Accuracy of C. elegans Image Analysis 37
Jacob Graves and Roger Mailler

Automated Segmentation and Measurement for Cancer Classification of HER2/neu Status in Breast Carcinomas 43
Lee Sing Cheong, Angela Jean, Tsu Soo Tan, Waiming Kong, and Soo Yong Tan

Incorporating Protein Sequence and Evolutionary Information in a Structural Pattern Matching Approach for Contact Maps 49
Hazem R. Ahmed and Janice I. Glasgow

Methodology to explore coexpression in microarray data 56
Bertrand De Meulder, Eric Bareke, Michael Pierre, Sophie Depiereux, and Eric Depiereux

System Biology on Mitochondrion Genomes 61
Michael Sadovsky, Natalia Zaitseva, and Yulia Putintzeva

Computational Modeling of Robust Figure/Ground Separation 67
Marc Ebner and Stuart Hameroff

Anace: Phylogenetic Trees Drawing Web Service 73

Helio Moura and Gerardo Viana

Computation of Dynamic Channels in Proteins 78
Petr Benes, Petr Medek, Ondrej Strnad, and Jiri Sochor

EnzymeTracker: A Web-based System for Sample Tracking with Customizable Reports 84
Thomas Triplet, Justin Powlowski, Adrian Tsang, and Gregory Butler

In silico Identification of Drug Targets in Methicillin/Multidrug-Resistant Staphylococcus aureus 91
Nichole Haag, Kimberly Velk, and Chun Wu

On the Distribution of the Distances Between Pairs of Leaves in Phylogenetic Trees 100
Arnau Mir and Francesc Rossello

De Novo Draft Genome Assembly Using Fuzzy K-mers 104
John Healy and Desmond Chambers

Internal force field in proteins 110
Damian Marchewka and Irena Roterman

Structural Characterization of the Rieske Oxygenase Complex from Burkholderia fungorum DBT1 Strain: Insights from bioinformatics 116
Stefano Piccoli, Silvia Lampis, Giovanni Vallini, and Alejandro Giorgetti

UMPIRE: Ultimate Microarray Prediction, Inference, and Reality Engine 121
Jiexin Zhang and Kevin Coombes

Singular Spectral Analysis on Doppler Ultrasound Records Acquired from Healthy Subjects and Intrinsic Sphincter Deficiency Subjects

Kadir Tufan, Department of Computer Engineering, Fatih University, Turkey ktufan@fatih.edu.tr

Sadık Kara, Institute of Biomedical Engineering, Fatih University, Turkey, skara@fatih.edu.tr

Fatma Latifoğlu, Department of Biomedical Engineering, Erciyes University, Turkey, flatifoglu@erciyes.edu.tr

Sinem Aydın, Radiology Department, Haseki Training and Research Hospital, Turkey, sinem.rad@gmail.com

Adem Kırış, Radiology Department, Haseki Training and Research Hospital, Turkey, ademkiris@hotmail.com

Ünsal Özkuvancı, Urology Department, Haseki Training and Research Hospital, Turkey, unsalozkuvanci@hotmail.com

Abstract **- Stress Urinary Incontinence is a common form of women urinary incontinence and has some surgical therapeutic methods. The quality of surgery merely depends on the quality of diagnosis. Before deciding a surgery type, urodynamic testing is applied. Since urodynamic testing is invasive, and difficult to apply, many subjects do not seek therapy. A non-invasive method can encourage more subjects for searching a remedy. At this point, Doppler ultrasound can be a good choice. In this study, we have demonstrated that the blood flow characteristics of healthy subjects and intrinsic sphincter deficiency type stress urinary incontinence subjects have different characteristics and can be classified by using Doppler ultrasound recording.**

Keywords **-** ***intrinsic sphincter deficiency; Doppler ultrasound; singular spectral analysis,emprical mode decomposition***

I. INTRODUCTION

Urinary **I**ncontinence (UI) is a common disorder [1]. It negatively affects the lifestyle of women, although it is not fatal [2].

UI has four basic types; namely **S**tress **U**rinary **I**ncontinence (SUI), Urge Incontinence, mixed incontinence, and overflow incontinence. Each of them has different grounds and needs a different method of handling. SUI is described as the involuntary leakage of urine under stress conditions like coughing, sneezing, laughing. For only SUI, there are some therapeutic surgeries [3].

The **I**nternational **C**onsultation on **I**ncontinence (ICI) set a series of guidelines for the diagnosis of urinary incontinence [4]. The first step is taking the history of the patient followed by physical examination [5]. For SUI suspected cases, the stress test is applied. In this procedure, the patient who is under the stress condition is examined whether there is any leakage or not.

If surgery is needed, urodynamic methods are involved to decide the appropriate surgical method. Pressure profiles, water-filling cystometry, electrophysiological studies, **P**ostvoid **R**esidual (PVR) measurement, urodynamic testing, cystogram are some examples of diagnostic methods. These methods very helpful but they are invasive.

Invasive methods like urodynamic methods are more successful but less comfortable for patients. Therefore, a new method that is non-invasive will encourage subjects who are suffering from SUI is a necessity. In this point ultrasonography can be an alternative method because of its noninvasive nature. Perineal ultrasound is applied in some studies [6] [7] [8]. In these studies, some parameters of blood flow dynamics and some angles, orientations and distances on ultrasound image are used. Unfortunately, these methods have low accuracy.

In urodynamic testing, **A**bdominal **L**eak **P**oint **P**ressure (ALPP) is measured for deciding a surgery type. If it is less than 60 cmH_2O, then it is called **I**ntrinsic **S**phincter **D**eficiency (ISD). If ALPP is greater than 90 cmH_2O, it is named as **U**rethral **H**yper **M**obility (UHM). The ones who have ALPP value between 60 and 90 mH_2O are called as mid - type. ALPP is also high for continent subjects.

Since the ground of SUI is the physical changes in the anatomy of the pelvic system, it is natural to expect some changes in the blood flow characteristics of SUI subjects. In our previous study, power spectral density of healthy subjects and SUI subjects

 ISBN: 978-1-61208-137-3

are used for classification [9]. In this study, it is aimed to differentiate healthy subjects and ISD type SUI subjects by analyzing Doppler ultrasound records. Nonlinear analysis was successfully applied for Doppler signals in some studies [10].

In this study, **E**mpirical **M**ode **D**ecomposition (EMD) is used to estimate **I**ntrinsic **M**ode **F**unctions (IMF) first, and then **S**ingular **S**pectral **A**nalysis (SSA) of each IMF is calculated. Features extracted from SSA result are used for classification of healthy subjects and ISD subjects.

This article is prepared as follows. In section II, materials used in this study and the methods for classification of healthy subjects and ISD subjects are given. The results of the study are given in section III. In section IV, the discussion and conclusion of the study are given.

II. MATERIALS AND METHODS

A. Subjects

Doppler ultrasound signals were recorded from GE Logiq 9 Doppler Ultrasound Unit in the Radiology Department of Haseki Training and Research Hospital, Istanbul, Turkey. Output of Doppler Ultrasound unit was connected to an Olympus LS-10 Digital Voice Recorder, thatstores data in (.wav) format. Recorded signals then transferred to a Laptop where the signal processing and classification processes were performed. The MATLAB® program is used at the signal processing and the classification steps. The study was approved by the ethic committee of Haseki Training and Research Hospital, and informed consent was given by the participating subjects.

A perineal ultrasound probe (7.5 MHz) was used to transmit pulsed Doppler ultrasound signals to the urethral artery. The reflected signals were recorded by the digital sound recorder. These signals give Doppler shift frequencies for healthy and ISD subjects. During recording, the insonation angle and the presetting of the ultrasound were kept constant. This angle was tuned via electronic steering methods to keep at 45° on a longitudinal view and the sampling volume was placed within the center of the artery. The audio output of the Doppler ultrasound was recorded at 44100 Hz, 16 bits, in a stereo channel format. The graphical illustration of data recording scheme is given in Figure 1.

Urethral arterial signals were recorded from 16 ISD patients and 18 healthy volunteers. Patients were clinically tested and proved to be ISD by urodynamic testing. SUI subjects are between 37 and 56 years old (average is 49). Control group was formed from young volunteers whose age is between 38 and 55, with an average of 46.

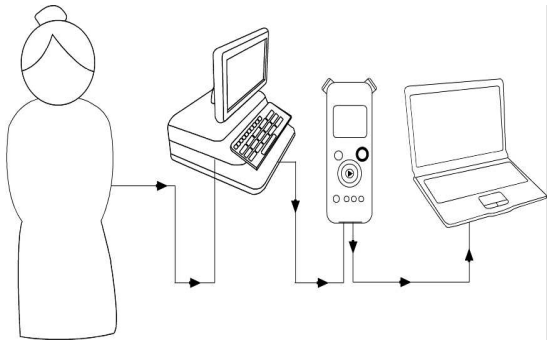


Figure 1. The data recording scheme illustration

B. Data Analysis

B.1. Empirical Mode Decomposition

The empirical mode decomposition (EMD) was proposed by Huang et al., [11]. It is a numerical model to decompose the signal into a finite number of IMF. The IMF function must satisfy two conditions:

(a) The number of the zero crossings and the number of extreme must be either equal or differ by one.

(b) The mean value of the local minima and maxima envelope of the function must be zero.

Once the first IMF is obtained, the restingportion can also include more IMFs. The remaining data can be further decomposed successively to get other IMFs by utilizing EMD.

In the representation of signals in IMF forms, the lower order IMFs hold fast oscillating components. Similarly, higher order IMFs represent slow oscillations.

B.2. Singular Spectrum Analysis

Singular Spectrum Analysis (SSA) is a power time series analysis technique. It is a nonparametric spectral estimation method that is based on embedding a time series $X(t):t=1..N$ in a vector space having M dimension. Here, the time series assumed to be stationary.

SSA is suitable for extracting meaningful information from noisy time series. For the assumption of being stationary, the time series must

 ISBN: 978-1-61208-137-3

have a short duration. It performs an eigen decomposition of the lagged covariance matrix, that is constructed from the time series.

SSA approach is an effective tool for nonlinear analysis and has an important advantage over other time series analysis techniques; it does not need any priory information.

III. RESULTS

The classification features are extracted from SSA calculation of IMFs. First three IMF of Doppler ultrasound signals recorded from healthy subjects and ISD subjects are extracted, firstly. Then SSA is applied for each IMF and covariance matrix (M=10) is calculated. From the covariance matrix, Eigen values are calculated.

In Figure 2, EMD view of one subject is given. The top three are the first, second, and third IMFs of Doppler ultrasound record of the subject. The last plot is the residue of EMD.

In Table 1, the data used for classifications of healthy subjects and ISD subjects. In this study, 16 ISD subjects and 18 healthy (continent) subjects are used. Healthy subjects are called as CONT.

The classification parameters (features) are given at the second and third columns of Table 1. The explanation of the classification parameters extracted from SSA of IMFs.

$$\text{AreaRatio1} = Area1 / Area2 \quad (1)$$

$$\text{AreaRatio2} = Area2 / Area3 \quad (2)$$

Where

AreaRatio1 and AreaRatio2 are classification features. Here Area1 is the sum of Eigen values for IMF1. Similarly, Area2 and Area3 are the sum of Eigen values for IMF2 and IMF3, respectively.

In the Figure 3, scatter plot of classification parameters is given. The difference between two classes is seen clearly in this illustration. When considering the class centers, these classes are separated definitely. There are some members of ISD and CONT classes very close to each other, so the separation is difficult by pattern recognition techniques.

TABLE 1. Features used for Classification

Subject Name	AreaRatio1 (IMF1/IMF2)	AreaRatio2 (IMF2/IMF3)
ISD01	1.61	1.93
ISD02	1.06	1.74
ISD03	1.05	1.67
ISD04	3.32	1.66
ISD05	1.44	1.57
ISD06	1.90	1.55
ISD07	0.94	1.23
ISD08	1.21	1.18
ISD09	0.66	1.13
ISD10	0.70	1.11
ISD11	0.44	1.09
ISD12	0.82	1.07
ISD13	1.00	1.00
ISD14	1.18	0.97
ISD15	0.48	0.85
ISD16	0.27	0.82
CONT01	0.63	0.97
CONT02	0.96	0.92
CONT03	0.60	0.90
CONT04	0.66	0.79
CONT05	0.77	0.76
CONT06	1.14	0.72
CONT07	0.50	0.71
CONT08	0.14	0.60
CONT09	0.24	0.59
CONT10	1.91	0.52
CONT11	0.15	0.45
CONT12	1.10	0.44
CONT13	0.00	0.39
CONT14	1.30	0.35
CONT15	1.86	0.28
CONT16	0.86	0.19
CONT17	0.61	0.19
CONT18	2.39	0.01

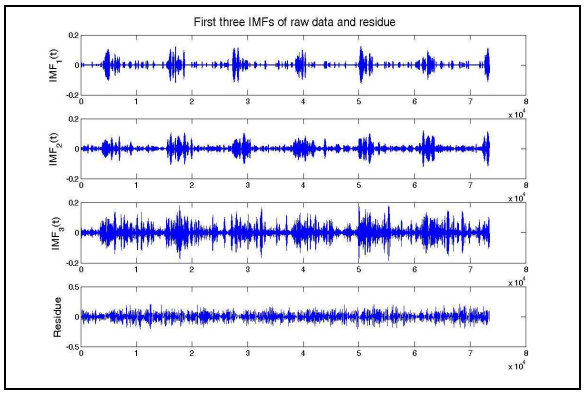


Figure 2. Empirical mode Decomposition (EMD) of one subjects

 ISBN: 978-1-61208-137-3

IV. DISCUSSION AND CONGLUSION

In this study, healthy (continent) and ISD type SUI subjects are classified by using Doppler ultrasound records.

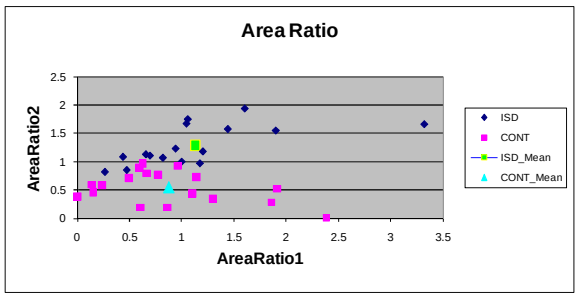


Figure 3. Scatter of features of ISD and healthy (continent) subjects

ISD: Subjects having ISD disorder

CONT: Healthy (continent) subjects

ISD_Mean: Average of ISD subjects

CONT_Mean: Average of healthy subjects

In the first step, signal is divided into IMFs by using empirical mode decomposition. Then Singular Spectrum Analysis (SSA) is applied to each IMF. Then the area ratios of SSA of IMFs are calculated as the feature for classification.

When the scatter plot of classification parameters given in Figure 3 is analyzed, AreaRatio2 values of ISD subjects are generally higher than that of healthy subjects. It means that the ratio between IMF2 and IMF3 of ISD subjects shows the greater difference when compared with healthy subjects. This is actually theoretically expected. In ISD subjects, since the sphincter muscles loose their contractility, the blood flow of the urethral artery should be more laminar than healthy subjects. In this type flow, the Eigen values should be concentrated in a few IMFs. In other words, the area ratios of more laminar flows are higher than more complex flows. This is proven in this study.

SUI is the result of physical deformation on the continence supporting system. It is natural to expect some changes in the blood flow dynamics of urethral artery that feeds the pelvic region. In this study, we have demonstrated that healthy subjects and ISD type SUI subjects can be classified by using Doppler ultrasound records of subjects.

The noninvasive nature of Doppler ultrasound makes it a suitable candidate for diagnosis of SUI. The gold standard method used in diagnosis of the disorder is invasive urodynamic testing that prevents many subjects from seeking a remedy. Method proposed in this study can be a good alternative for urodynamic testing.

In future studies, the discrimination of subtypes of SUI subjects (UHM or ISD types) will be tried by using Doppler ultrasound records of urethral artery.

ACKNOWLEDGEMENT

This work is supported by the Scientific Research Fund of Fatih University under the project number P50060901-2.

REFERENCES

[1] Urinary incontinence - ACOG Technical Bulletin, No. 213, October 1995 (Replaces No. 100, January 1987). Int J Gynecol Obstet. 1996; 52: 7586.

[2] Blok, B.F. M. and Corcos, J. Surgery for stress urinary incontinence in women: A 2006 review. Indian J Urol. Apr 2007; 23(2): 148–152.

[3] Haab, F., Zimmern, P.E., and Leach, G.E. Female Stress Urinary Incontinence Due to Intrinsic Sphincteric Deficiency: Recognition and Management. The Journal of Urology. July 1996; 156(1): 3-17.

[4] Abrams, P., Cardozo, L., Fall, M., Griffiths, D., Rosier, P., Ulmsten, U., et al. The standardisation of terminology in lower urinary tract function: report from the standardisation sub-committee of the International Continence Society. Urology. June 2003; 61: 37-49.

[5] Pantazis, K. and Freeman, R.M. Investigation and treatment of urinary incontinence. Current Obstetrics and Gynaecology. December 2006; 16(6): 344-352.

[6] Aksoy, F. and Kiris, A. Stress üriner İnkontinansın Transperineal Ultrasonografi ile Değerlendirilmesi. Uzmanlık Tezi. 2005.

[7] Korda, A., Krieger, M., Hunter, P., and Parkin, G. The value of clinical symptoms in the diagnosis of urinary incontinence in the female. Aus NZ L Obstet Gynecol. May 1987; 27: 149-151.

[8] Bergman, A., Ballard, C.A., and Platt, L.D. Ultrasonic evaluation of urethrovesical junction in women with stress urinary incontinence. J. Clin. Ultrasound. June 1988; 16(5): 295–300.

 ISBN: 978-1-61208-137-3

[9] Tufan, K., Kara, S., Latifoğlu, F., Aydın, S., Kırış, A., and Özkuvancı, M. Comparison of ISD Type Stress Urinary Incontinence and Healthy Subjects by Analyzing Doppler Ultrasound Data. ICADIWT 2010 The third International Conference on the Applications of Digital Information and Web Technologies. July 2010: 132-135.

[10] Uzunhisarcıklı, S. Nonlinear dynamic analysis of mitral valve doppler signals: surrogate data analysis. Turk J Elec Eng and Comp Sci. 2010; 18(2): 327-337.

[11] Huang, N.E., Shen, Z., Long, S.R., Wu, M.C., Shih, H.H., Zheng, Q., et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical Physical and Engineering Sciences. March 1998; 454(1971): 903–995.

 ISBN: 978-1-61208-137-3

Separating Drivers from Passengers in Whole Genome Analysis: Identification of Combinatorial Effects of Genes by Mining Knowledge Sources

Stephen Anthony
Centre for Health Informatics
University of New South Wales
Sydney, Australia
e-mail: s.anthony@unsw.edu.au

Enrico W. Coiera
Centre for Health Informatics
University of New South Wales
Sydney, Australia
e-mail: e.coiera@unsw.edu.au

Vitali Sintchenko
Centre for Infectious Diseases and Microbiology
Sydney Medical School, University of Sydney
Sydney, Australia
e-mail: vitali.sintchenko@sydney.edu.au

***Abstract*—This study aimed to develop a new informatics platform for the discovery, recovery and multi-level analysis of the effects of individual genes and multiple gene combinations on pathophenotypes of bacteria. Natural language processing algorithms were employed to extract gene-disease associations from PubMed literature and annotated genomes of bacteria with epidemic potential. From these associations gene virulence profiles were generated enabling the comparison of gene signatures within and across genomes. It allowed the identification of virulence genes and construction of their association networks as well as the detection of knowledge gaps. This proof-of-concept study confirmed the feasibility of our original approach for integrating bacterial genome level knowledge with published observations from clinical settings.**

Keywords: structural bioinformatics; whole genome analysis; text mining; infectious diseases; knowledge discovery

I. INTRODUCTION

The exponential growth in whole genome sequencing has created new challenges for data integration and analysis. It has crucially increased the rate of discovery of associations between genes and diseases [1]. The mapping of relationships between genes and disease phenotypes has become possible due to synergistic advances in text mining and the availability of quality data and indexed text in the public domain. For example, online catalogs of human genome-wide association studies exceeded 700 publications linking genetic variations and diseases.

Knowledge about the etiology and pathogenesis of diseases has increasingly been stored in literature and in databases, including sets of fully or partially annotated bacterial genomes [2]. Text mining approaches are gaining importance in the extraction and collation of data and text mining with bioinformatics databases [3-5]. In particular, significant progress has been made in building applications for the knowledge-based profiling of individual genes [6, 7], gene mining and mapping to diseases [8, 9] and mining complex features for predicting drug resistance [10]. These developments have drawn attention to the problem of diminishing returns of existing analytic approaches. The challenge of potential non-linearity in the mapping of genotypes to phenotypes and our ability to address it has been emphasized, with calls for an analytical retooling to address the combinatorial nature of gene-disease effects [11]. Bioinformatics techniques such as collapsing or binning have been borrowed from SNP-based genome-wide studies and applied to study human diseases that can be affected by multiple genes. However, microbial genome-wide association studies have not received due attention so far and bioinformatics applications for microbial genome analyses remain relatively underdeveloped.

The infectious disease (ID) domain presents a new frontier of high-throughput sequence analysis adding another, pathogen-specific, dimension to genome and disease association studies. Early findings indicate that disease-defining properties of pathogens are both multi-factorial and combinatorial [12] but further progress has been limited by the lack of analytical tools. The integrated microbial genome initiatives [13, 14] so far have focused attention on the alignment and comparison of genome sequences without linking sequencing data to ID attributes or clinical outcomes [15].

Whole genome sequence analysis has promised to improve the accuracy of ID risk assessment and the discriminatory power of tracking of outbreaks. Several new tools have been proposed to leverage these resources and to assist researchers with identification of genomic targets for vaccine and drug discovery [16] and the study of pathogen evolution [17]. Until recently, each gene or protein was studied as a single entity. However, new 'omics' technologies have allowed the analysis of large numbers of genes simultaneously and the generation of complex networks [18].

The aim of this study was to develop a new informatics platform for the discovery, recovery and multi-level analysis of the effects of individual genes and multiple gene combinations on properties of pathogenic bacteria.

 ISBN: 978-1-61208-137-3

II. METHODS

A. Approach and Definitions

The assessment of the impact of individual genes was based on searches for literature-based associations of genes with ID syndromes in order to separate key genes responsible for pathogenic phenotype of bacteria ('*drivers*') from non-pathogenic genes of little consequences to bacterial virulence ('*passengers*'). The working hypothesis was that virulence genes can be identified through the combination of virulence profiles, within and across bacterial genomes. Virulence profiles were generated from gene-disease associations.

The concepts of 'core' and 'dispensable' (or unique) genomes were considered [19]. The core genome consisted of the set of common genes conserved across all strains, encoding those functions necessary for the basic biology of the species. The dispensable genes contribute to the diversity within the species, including virulence, transmissibility, antibiotic resistance and niche adaptation [19]. Core genes were identified and the combined virulence profiles across species were compared to the virulence profiles of individual species. Core genes were also contrasted against genes that are unique to individual microorganisms.

B. Data Sources

The 2011 MEDLINE/PubMed Baseline Distribution was employed as the primary source of literature for generating gene virulence profiles. The 2011 baseline comprises 10,891,200 citations that contain text from article abstracts. The citations were loaded into a Postgres Database Management System (DBMS) with associated full text indexes generated for the title and abstract. The text vectors were generated by the DBMS using a template for English that is based on the Porter stemming algorithm [20].

Fully sequenced genomes available through the National Center for Biotechnology Information (NCBI) GenBank [21] were employed for gene symbol and gene sequence location information for a number of bacterial genomes with pathogenic potential representing both Gram positive and Gram negative obligatory and opportunistic pathogens with the genome size of 3-4Mb and a range of core genome sizes (Table 1).

A list of syndromes associated with infectious diseases (108 items) was constructed as reported previously [22]. Names of syndromes included *sepsis*, *pneumonia*, *meningitis*, *encephalitis*, *cellulitis*, *wound infection*, and *urinary tract infection*, among others. However, ID syndromes uniquely associated with specific pathogens such as tuberculosis, malaria, dengue, etc. were excluded from the list to minimize the detection of trivial associations.

C. Gene Symbol Classification

Each PubMed abstract was indexed for mentions of a syndrome or species related gene symbol. The SPECIALIST lexicon [23] supported the identification of variants in nomenclature, spelling, and clinical abbreviations.

TABLE I. BACTERIAL GENOMES UTILISED IN THE STUDY

Bacterial Genome	GenBank Accession	Proteins
Listeria monocytogenes 4b F2365	AE017262	2821
Mycobacterium tuberculosis H37Rv	AL123456	3988
Neisseria meningitidis 053442	CP000381	2020
Pseudomonas aeruginosa PA01	CP000744	5566
Salmonella typhi CT18	AL513382	4391
Staphylococcus aureus MRSA252	BX571856	2650
Streptococcus pyogenes MGAS6180	CP000056	1894

The identification of gene symbols was performed using a combined search and classification strategy. A gene symbol classification model was constructed by acquiring a set of 5,003 unique gene symbols for training sourced from nine fully sequenced NCBI genomes. Inclusion in the training set comprised the following criteria. Gene symbols had to contain at least two out of three characters from the classes: upper case, lower case, and numeric. Gene symbols could not contain underscores or hyphens or were otherwise excluded. A full text search was performed for each of the symbols in the training set. As the full text search configuration collapses case, a further constraint was applied to ensure each gene mention in the abstract appeared with identical orthography.

The rule extraction from each symbol's contextual window utilized morphosyntactic and lexical features. The features were generated from a set of base templates. Base template classes employed to generate features included tokens either side of the symbol, tokens from both sides of the symbol, a function to determine whether the symbol was enclosed in parentheses, and a function to determine whether mixed case tokens were present in the context. Tokens used in feature generation through templates were lemmatized and had their orthographic case folded.

An entropy maximization approach was employed to rank each contextual feature. The measure was calculated by determining the number of times each pattern co-occurred with gene symbols in relation to the total number of occurrences of the contextual pattern in the training corpus. The higher the entropy of the feature, the greater its contribution to the classification of a symbol. An evaluation of the approach is presented in the results section.

D. Representation of Multiple Gene-Disease Associations

A virulence profile was generated for each gene based on the co-occurrence between syndromes and individual pathogen-specific gene symbols. The latter was defined as a gene symbol that co-occurs with a bacterial species name in the same document (e.g., *Staphylococcus aureus* AND *dnaA*). Co-occurrence was calculated using pointwise mutual information (PMI) [24]. The PMI formula (1) provides a measure of association where p represents probability, and both x and y represent terms.

$$\mathrm{PMI}(x,y) = \log_2 (p(x, y) / (p(x) * p(y))) \qquad (1)$$

 ISBN: 978-1-61208-137-3

The association measures for pathogen-specific gene symbols were employed in the generation of vectors intended to express virulence potential of a gene. The summation of individual vector components represented a gene *virulence factor*. Each gene's vector was compared to every other vector in the genome, and across genomes, and similarities between expression vectors were estimated using a Euclidean distance measure [25].

III. RESULTS

A. Gene Classification

Our text mining strategy successfully extracted gene names and syndrome entities. For example, it differentiated the 1,670,132 abstracts that contain the word stem *year* from the three abstracts that contained the gene symbol *yeaR*.

Let us illustrate the performance of our approach using the fully sequenced genome for *Neisseria meningitidis* 053442 as an example. Specifically, it encodes the gene product *porA* (porin, class I outer membrane protein). A PubMed-wide search was initiated for all article titles or abstracts that contained the pathogen-specific symbol *porA* in association with any one of the 108 ID-related syndromes. The symbol *porA* was found to occur in 312 distinct articles, and the pathogen-specific form appeared in 175 distinct articles. A total of 2,079,834 distinct articles were found to contain any one or more of the syndrome terms. The terms co-occured in a total of 41 documents. Unsurprisingly in this case, over half of the associations could be attributed to co-occurrence with a mention of some form of the syndrome *meningitis*. The PMI association scores between the pathogen-specific gene symbol *porA* and individual ID syndromes constituted a virulence profile.

The gene classification approach was evaluated using an independent test set. The test set was sourced from an additional three NCBI genomes. Gene symbols from these three genomes that were present in the training set were excluded. The test set contained 3,649 gene symbols in total, reduced to 1,271 once common genes were excluded. The search with exact match criteria produced 4,128 contexts from documents that contained a test set gene symbol. The approach correctly classified 4,032 of the 4,128 test set instances, achieving and overall accuracy of 97.67%. The largest source of errors originated from instances where the context contained never before seen tokens, a high proportion of punctuation characters, or no mixed case tokens that are often indicative of gene symbols. For example, the gene *lys1* that is found exclusively in the test set and embedded in the following context was misclassified:

(UGA), lys1-1' (UGA) [PMID: 782552]

The contextual tokenization resulted in the following 4-token window either size of the target gene *lys1*:

| (| UGA |) | , | lys1 | -1 | ' | (| UGA |

A number of suggestions for the remediation of these types of errors are presented in the discussion section.

B. Genome-wide Virulence Profiling

The first representation generated from the expression vectors resulted in a genome-wide virulence factor. Table II illustrates overall virulence factors for bacterial genomes that have been calculated by summing the PMI ID-association scores for each pathogen-specific gene within a genome.

TABLE II. TOTAL VIRULENCE FACTOR PER MICROORGANISM

Bacterial Genome	Total No. of Genes	Virulence Factor (bits)
Listeria monocytogenes 4b F2365	756	78.95
Mycobacterium tuberculosis H37Rv	1672	371.04
Neisseria meningitidis 053442	869	183.16
Pseudomonas aeruginosa PA01	1736	810.92
Salmonella typhi CT18	2558	332.07
Staphylococcus aureus MRSA252	782	369.39
Streptococcus pyogenes MGAS6180	776	340.81

Genome-wide virulence profiles for individual genes can be presented in a circularized form (Figure 1) to reflect the exact position of genes in a genome and identify most commonly reported genes of an individual bacteria (*Neisseria meningitidis* in this case) that have been associated with clinical presentations or adverse outcomes of ID. These profiles were generated by plotting each gene's frequency of co-occurrence with ID-syndromes in the literature.

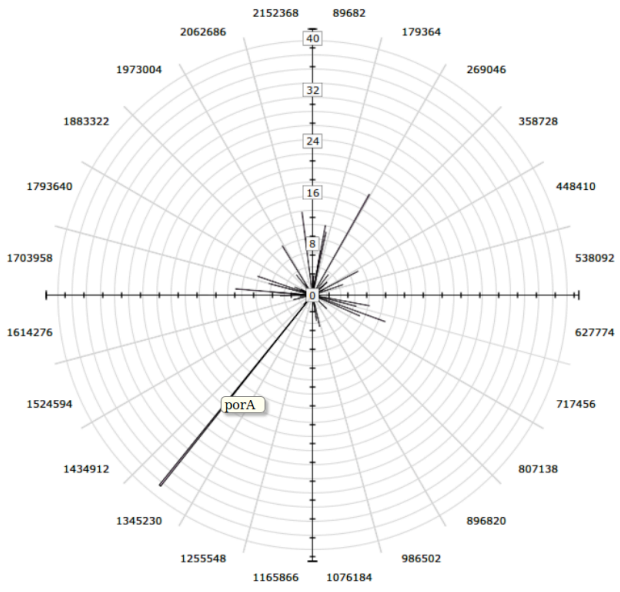


Figure 1. Genome-wide virulence profile for *Neisseria meningitidis*. Individual gene frequency of co-occurence with ID syndromes in the literature are plotted according to each gene's location in the genome.

1) Potential Knowledge Gap Identification

Core genes were determined and visualized by plotting the virulence profile for each of the genes that were common across multiple genomes (Figure 2). Such virulence profiling

contrasts core genes that have been associated with pathogenicity in one (e.g., *leusS, miaA*) or several (e.g., *murE, mutL*) bacterial species.

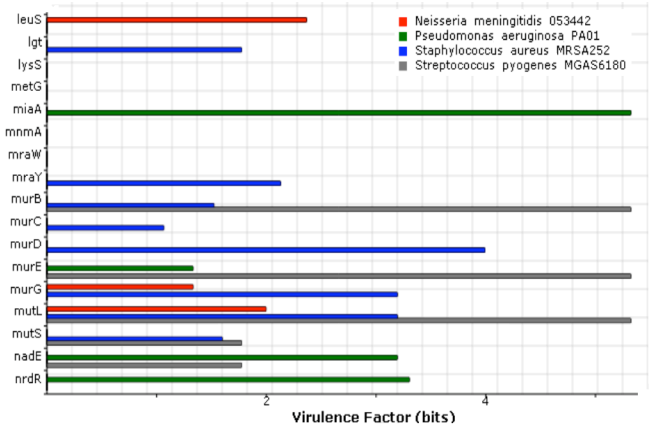


Figure 2. Core Gene Virulence Factor Comparison. A side-by-side comaprison of gene virulence factors for a fragment of alphabetically sorted genes common across selected genomes.

Gene virulence profiles were subsequently employed to identify potential gaps in the knowledge regarding a gene's impact on ID outcomes. The process was initiated by identifying common genes across all selected genomes. This set of genes was then restricted to those that contained at least one positive expression value in any genome. Each of the residual genes was plotted against its respective total (across-genome) virulence profile. For example, a striking discrepancy was identified between a high frequency of associations of many genes of bacteria with ID syndromes and the lack of those in the *Staphylococcus aureus* genome (Figure 3). Specifically, this observation suggested two knowledge gaps may exist for our understanding of *ftsZ* and *glyA* gene functions in the genome of *S. aureaus* (Figure 3). Interestingly, *ftsZ* gene has been associated with virulence properties of other bacteria due to its role in the synthesis of the protein tubulin that participates in replication of toxin-encoding plasmids [26]. Glycinecin gene (*GlyA*) has been also identified as a putative virulence gene in other bacteria because of its involvement in the synthesis of bacteriocins [27].

2) Identification of key virulence genes

The first approach to identifying drivers involved the identification of common genes. The simplest determination of a driver gene could be defined as a gene that is common across genomes and expresses an above-average overall virulence factor. This approach implicates the genes *dnaK*, *eno*, *folD*, *ftsZ*, and *glyA* amongst others as can be seen in Figure 3 as their virulence factors fall above the overall across-genome average indicated by the horizontal line.

The next approach to driver gene identification resulted in a network-based linkage analysis. The analysis was performed to detect combinations of genes that are typically associated with syndromes. The visualization of relationships between individual genes in Figure 4 shows links that are formed between genes within the *Listeria monocytogenes* 4b F2365 genome. Edge weights reflect the strength of association between gene virulence profiles. This network representation (Figure 4) also highlights indirect relationships between some genes (e.g., *plcB* and *prfA* through *actA*) and identifies highly connected virulence genes of particular relevance for the virulence assessment (e.g., *actA* and *prfA*). The *actA* gene, which is involved in the synthesis of bacterial protein actine, presents a compelling example of a 'driver' gene. A literature review, conducted following our experiments, confirmed the *actA* gene of *Listeria monocytogenes* has been a key player in several biological mechanisms relevant to virulence, such as in escaping from vacuoles, undergoing intracellular growth, and spreading to neighboring cells in cell cultures [28].

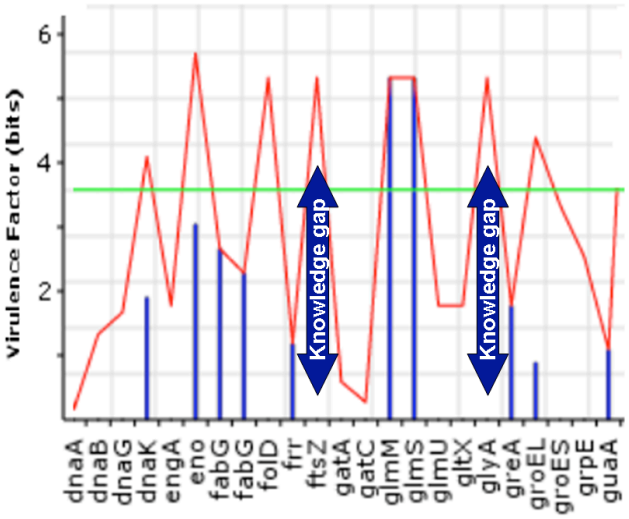


Figure 3. Gene Virulence Knowledge Gap Identification. Represents the virulence profiles for a fragment of core genes for the species *S. aureus* (vertical bars) in relation to the combined virulence profile across all genomes (line plot). The horizontal line represents the average virulence score of the combined virulence profiles.

3) Combinatorial effects of virulence genes

Similarities between virulence profiles were also compared both within and across microbial genomes in order to link individual genes with each other when they were found to be associated with ID syndromes. Figure 5 illustrates our approach to matching virulence profiles of individual genes within and between bacterial genomes. It seems to make more explicit the potential indirect links between genes of different function and origin that could be clustered according to their virulence. The links across the top of Figure 5 represent within-genome gene virulence similarity and the links across the bottom reflect between genomes similarity. Common or core genes were color coded in blue and genes unique to the genome of *S. pyogenes* were coded in red. Gene names in black represented house-keeping genes that could be found in several different bacterial species from our dataset.

 ISBN: 978-1-61208-137-3

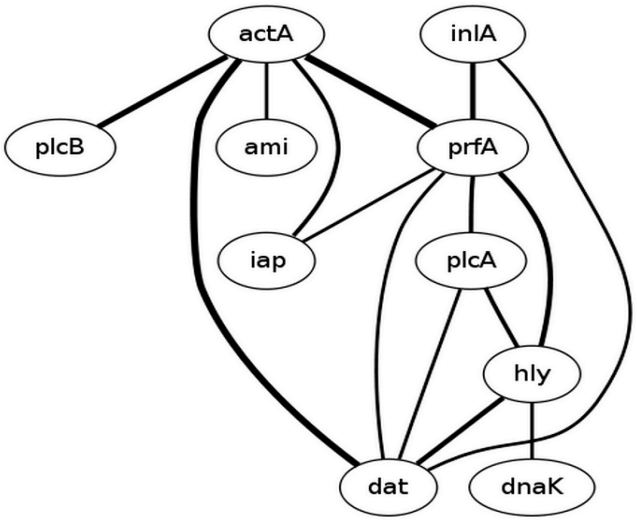


Figure 4. Gene virulence networks for *L. monocytogenes*. All pairs of gene expression vectors within each genome are compared and connected if their virulence profiles are similar.

4) Relative Impact of Unique Genes

The relative impact of unique or disposable versus core genes significantly differed between bacterial genomes in our dataset. This difference could be explained by their different propensity for lateral gene transfer. Table III lists the virulence factors for genes that are unique to each genome when compared against the remainder of genomes listed in the table. The table quantifies the aggregated total virulence factors for unique versus core genes for each species. There are a total of 202 core genes across all species. The number of core and unique gene occurrences per species are listed separately.

TABLE III. COMMON AND UNIQUE GENE VIRULENCE FACTORS PER MICROORGANISM

Bacterial Genome	Genes Virulence Factor (bits)		Number of Genes	
	Core	Unique	Core	Unique
L. monocytogenes 4b F2365	15.97	11.65	210	131
M. tuberculosis H37Rv	53.86	203.55	217	942
N. meningitidis 053442	14.81	54.84	205	153
P. aeruginosa PA01	102.13	367.60	208	719
S. typhi CT18	19.63	173.03	206	1408
S. aureus MRSA252	112.31	153.47	208	232
S. pyogenes MGAS6180	68.82	124.60	210	261

IV. DISCUSSION

Omics-based medicine demands significant re-tooling for continuous re-assessment of evidence for genome-phenome associations. The methods presented in this study can facilitate identification of combinations of genes within and across annotated bacterial genomes, differentiation of key virulence genes from genes of limited clinical relevance, detection of potential knowledge gaps, and measurement of the relative impact of individual genes and gene combinations. The methods are based on a set of tools and resources that comprise a large text corpus with full text search capabilities, a gene symbol classifier, and techniques for gene virulence profiling through literature-based association mining.

A key innovation in this work resulted from the establishment of literature-based pathogen-specific ID association measures spurred by a novel approach to increasing the specificity of gene symbol retrieval. For example, our text mining strategy differentiated the 171,203 PubMed mentions of the different forms of the indexed token *mode* from the 14 mentions of the gene symbol *modE*.

The transformation of our text mining measures into vectors that expressed gene virulence profiles led to a number of applications. The virulence profiles were applied both genome wide and across genomes. This resulted in the ability to identify combinations of genes that share virulence profiles. Comparing genome-wide gene profiles with overall virulence profiles across genomes also identified potential knowledge gaps. The next application identified potentially hidden driver genes indirectly linked by other genes that expressed similar virulence profiles.

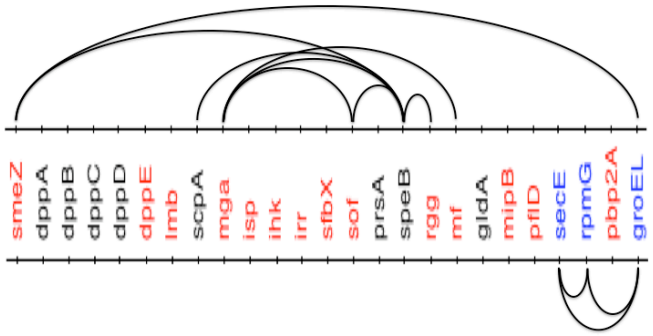


Figure 5. Within and between genome expression profile macthing. Links across the top represent genes that have similar expression profiles within the genome *S. pyogenes*. Links across the bottom represent genes with similar expression profiles across all genomes.

The novel approach taken to gene classification, as opposed to recognition, greatly improves search specificity. A number of further modifications have been discussed previously including a corresponding version of the feature generation templates that preserve case, and experiments with varying context-window sizes. Other potentially useful feature types include unlemmatized tokens, morphological analysis to inspect affixes, and semantic type analysis.

However, some potential limitations of the study have to be acknowledged. One of the major limitations to the gene classification evaluation and approach in general is the absence of negative examples. A possible solution to this problem could come in the form of leveraging examples from a text genre outside of the biomedical domain. Importantly, the aforementioned strategy does not eliminate false negatives that are collected by the search for gene symbols that are presented using a single orthographic case, particularly problematic are those that overlap with common English words such as *era*, *lip*, *map*, and *trap*. In order to combat this issue the abstract text for each article found to contain a gene symbol mention was tokenized and a four-token window either side of the symbol was extracted.

 ISBN: 978-1-61208-137-3

Training set gene symbols were found to occur in their exact form in 102,399 articles. Although not applied to this work the exact matching constraint employed to extract context could be relaxed to allow variants of gene symbols and capture other gene products. A number of adjustments to the classification algorithm could be made to redress potential limitations. For example, a parallel context could be constructed that excludes punctuation characters. Composite approaches are conceivable given the relatively high incidence of parenthesized gene symbols (11,296 out of 102,398 training instances). Another parameter that could be adjusted in future experiments is the context window size.

V. CONCLUDING REMARKS

This proof-of-concept study confirmed the feasibility of our original approach for integrating bacterial genome level knowledge with published observations from clinical settings. It opens a new opportunity for real-time assessment of virulence of bacterial genomes and for identification of high-impact genes and their combinations. Further 'wet lab' experiments are required to validate the utility of the knowledge gap detection function.

This work has culminated in an online toolset that enables researchers to explore, recover, and potentially discover new insights into microbiological mechanisms that contribute to infection. An online implementation of this work can be accessed from http://purl.org/infectious/genome.

REFERENCES

[1] K.E. Ormond, M.T. Wheeler, L. Hudgins, T.E. Klein, A.J. Butte, R.B. Altman, et al., "Challenges in the clinical application of whole-genome sequencing", Lancet, vol. 375, May. 2010, pp. 1749-1751, doi:10.1016/S0140-6736(10)60599-5.

[2] T. Korves, and M.E. Colosimo, "Controlled vocabularies for microbial virulence factors", Trends Microbiol, vol. 17, Jul. 2009, pp. 279-285, doi:10.1016/j.tim.2009.04.002.

[3] S. Ananiadou, D.B. Kell, and J. Tsujii, "Text mining and its potential applications in systems biology", Trends Biotechnol, vol. 24, Dec. 2006, pp. 571-579, doi:10.1016/j.tibtech.2006.10.002.

[4] Y. Kano, P. Dobson, M. Nakanishi, J. Tsujii, and S. Ananiadou, "Text mining meets workflow: linking U-Compare with Taverna", Bioinformatics, vol. 26, Oct. 2010, pp. 2486-2487, doi:10.1093/bioinformatics/btq464.

[5] J.O. Korbel, T. Doerks, L.J. Jensen, C. Perez-Iratxeta, S. Kaczanowski, S.D. Hooper, et al., "Systematic association of genes to phenotypes by genome and literature mining", PLoS Biol, vol. 3, May. 2005, pp. e134, doi:10.1371/journal.pbio.0030134.

[6] H. Xu, J.W. Fan, G. Hripcsak, E.A. Mendonca, M. Markatou, and C. Friedman, "Gene symbol disambiguation using knowledge-based profiles", Bioinformatics, vol. 23, Apr. 2007, pp. 1015-1022, doi:10.1093/bioinformatics/btm056.

[7] W. Xuan, P. Wang, S.J. Watson, and F. Meng, "Medline search engine for finding genetic markers with biological significance", Bioinformatics, vol. 23, Sep. 2007, pp. 2477-2484, doi:10.1093/bioinformatics/btm375.

[8] M. Garcia-Remesal, A. Cuevas, D. Perez-Rey, L. Martin, A. Anguita, D. de la Iglesia, et al., "PubDNA Finder: a web database linking full-text articles to sequences of nucleic acids", Bioinformatics, vol. 26, Nov. 2010, pp. 2801-2802, doi:10.1093/bioinformatics/btq520.

[9] T. Matsunaga, and M. Muramatsu, "Disease-related concept mining by knowledge-based two-dimensional gene mapping", J Bioinform Comput Biol, vol. 5, Oct. 2007, pp. 1047-1067, doi:10.1142/S0219720007003077.

[10] H. Saigo, T. Uno, and K. Tsuda, "Mining complex genotypic features for predicting HIV-1 drug resistance", Bioinformatics, vol. 23, Sep. 2007, pp. 2455-2462, doi:10.1093/bioinformatics/btm353.

[11] J.H. Moore, F.W. Asselbergs, and S.M. Williams, "Bioinformatics challenges for genome-wide association studies", Bioinformatics, vol. 26, Feb. 2010, pp. 445-455, doi:10.1093/bioinformatics/btp713.

[12] D.G. Lee, J.M. Urbach, G. Wu, N.T. Liberati, R.L. Feinbaum, S. Miyata, et al., "Genomic analysis reveals that Pseudomonas aeruginosa virulence is combinatorial", Genome Biol, vol. 7, Oct. 2006, pp. R90, doi:10.1186/gb-2006-7-10-r90.

[13] T. Davidsen, E. Beck, A. Ganapathy, R. Montgomery, N. Zafar, Q. Yang, et al., "The comprehensive microbial resource", Nucleic Acids Res, vol. 38, Jan. 2010, pp. 340-345, doi:10.1093/nar/gkp912.

[14] V.M. Markowitz, I.M. Chen, K. Palaniappan, K. Chu, E. Szeto, Y. Grechkin, et al., "The integrated microbial genomes system: an expanding comparative analysis resource", Nucleic Acids Res, vol. 38, Jan. 2010, pp. 382-390, doi:10.1093/nar/gkp887.

[15] N. Mulder, H. Rabiu, G. Jamieson, and V. Vuppu, "Comparative analysis of microbial genomes to study unique and expanded gene families in Mycobacterium tuberculosis", Infect Genet Evol, vol. 9, May. 2009, pp. 314-321, doi:10.1016/j.meegid.2007.12.006.

[16] A. Muzzi, V. Masignani, and R. Rappuoli, "The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials", Drug Discov Today, vol. 12, Jun. 2007, pp. 429-439, doi:10.1016/j.drudis.2007.04.008.

[17] G.S. Vernikos, and J. Parkhill, "Resolving the structural features of genomic islands: a machine learning approach", Genome Res, vol. 18, Feb. 2008, pp. 331-342, doi:10.1101/gr.7004508.

[18] S.C. De Keersmaecker, I.M. Thijs, J. Vanderleyden, and K. Marchal, "Integration of omics data: how well does it work for bacteria?", Mol Microbiol, vol. 62, Dec. 2006, pp. 1239-1250, doi:10.1111/j.1365-2958.2006.05453.x.

[19] S. Bentley, "Sequencing the species pan-genome", Nat Rev Microbiol, vol. 7, Apr. 2009, pp. 258-9, doi:10.1038/nrmicro2123.

[20] M.F. Porter, "An Algorithm for Suffix Stripping", Program-Automated Library and Information Systems, vol. 14, 1980, pp. 130-137.

[21] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, "GenBank", Nucleic Acids Res, vol. 36, Jan. 2008, pp. D25-30, doi:10.1093/nar/gkm929.

[22] V. Sintchenko, S. Anthony, X.H. Phan, F. Lin, and E.W. Coiera, "A PubMed-wide associational study of infectious diseases", PLoS One, vol. 5, Mar. 2010, pp. e9535, doi:10.1371/journal.pone.0009535.

[23] A.C. Browne, A.T. McCray, and S. Srinivasan, The SPECIALIST Lexcon, Lister Hill National Center for Biomedical Communications, National Library of Medicine, Bethesda, Maryland, 2000.

[24] K.W. Church, and P. Hanks, "Word association norms, mutual information, and lexicography", Computational Linguistics, vol. 16, March. 1990, pp. 22-29, doi:10.3115/981623.981633.

[25] E. Deza, M.M. Deza, and SpringerLink (Online service), Encyclopedia of Distances, Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2009.

[26] Y.J. Pan, T.L. Lin, C.R. Hsu, and J.T. Wang, "Isolation of genetic loci associated with phagocytosis and virulence in Klebsiella pneumoniae using a Dictyostelium model", Infect Immun, vol. Dec. 2010, pp. doi:10.1128/IAI.00906-10.

[27] W. Oswald, D.V. Konine, J. Rohde, and G.F. Gerlach, "First chromosomal restriction map of Actinobacillus pleuropneumoniae and localization of putative virulence-associated genes", J Bacteriol, vol. 181, Jul. 1999, pp. 4161-9.

[28] M. Conter, A. Vergara, P. Di Ciccio, E. Zanardi, S. Ghidini, and A. Ianieri, "Polymorphism of actA gene is not related to in vitro virulence of Listeria monocytogenes", Int J Food Microbiol, vol. 137, Jan. 2010, pp. 100-5, doi:10.1016/j.ijfoodmicro.2009.10.019.

 ISBN: 978-1-61208-137-3

Bio-UnaGrid: Easing Bioinformatics Workflow Execution Using LONI Pipeline and a Virtual Desktop Grid

Mario Villamizar, Harold Castro, David Mendez
Department of Systems and Computing Engineering
Universidad de los Andes
Bogotá D.C., Colombia
{mj.villamizar24, hcastro,
dg.mendez67}@uniandes.edu.co

Silvia Restrepo, Luis Rodriguez
Department of Biological Sciences
Universidad de los Andes
Bogotá D.C., Colombia
{srestrep, luisrodr}@uniandes.edu.co

***Abstract*—Bioinformatics researches use applications that require large computational capabilities regularly provided by cluster and grid computing infrastructures. Researchers must learn tens of commands to execute bioinformatics applications, to coordinate the manual workflow execution and to use complex distributed computing infrastructures, spending much of their time in technical issues of applications and distributed computing infrastructures. We propose the Bio-UnaGrid infrastructure to facilitate the automatic execution of intensive-computing workflows that require the use of existing application suites and distributed computing infrastructures. With Bio-UnaGrid, bioinformatics workflows are easily created and executed, with a simple click and in a transparent manner, on different cluster and grid computing infrastructures (line command is not used). To provide more processing capabilities, at low cost, Bio-UnaGrid use the idle processing capabilities of computer labs with Windows, Linux and Mac desktop computers, using a key virtualization strategy. We implement Bio-UnaGrid in a dedicated cluster and a computer lab. Results of performance tests evidence the gain obtained by our researchers.**

Keywords; Bio-UnaGrid; grid computing; cluster computing; bioinformatics; BLAST; mpiBLAST; desktop grid; UnaGrid; LONI Pipeline

I. INTRODUCTION

Today genomic projects involve the use of two complementary approaches: cluster computing which allows to aggregate homogeneous computational resources for a specific research group or organization, and grid computing which aggregates heterogeneous computational resources of different organizations to support larger computational capabilities in e-Science projects.

High performance computing (HPC) infrastructure, such as cluster or grid computing, provide results in shorter time, however when bioinformatics researchers want to execute applications, they face several consuming time problems: a) there are many application suites to run different analysis, so researchers must learn command line syntax (options, input and output files, distributed execution environment, etc.) for hundreds of applications. b) Researchers must learn commands to manage and use distributed computing infrastructures. c) Applications require specific and complex configurations to operate in distributed environments. d) Researchers frequently execute a set of applications in a sequential and/or coordinated manner, called pipelines or workflows. Workflows are regularly manually executed by researchers, so they must wait an application finishes, before executing the next. e) Very often researchers want to execute applications that require larger processing capabilities than those provided by dedicated computing infrastructures, so they must wait weeks or months before getting their results.

Different approaches have been developed for solving these problems partially on dedicated computing infrastructures. In this work we propose and evaluate an integral infrastructure that allows bioinformatics researchers, requiring large computing capabilities, to focus in bioinformatics analysis and not on technical computing issues of distributed computing infrastructures. The infrastructure, called Bio-UnaGrid, allows researchers to define workflows using graphical user interfaces (GUIs) and drag and drop tools provide by the LONI Pipeline [1] application. Workflows are executed and distributed in a transparent manner on a grid infrastructure. To support larger processing capabilities to those provided by dedicated infrastructures, Bio-UnaGrid also uses the UnaGrid [2] desktop grid infrastructure. UnaGrid permanently takes advantage of the idle processing capabilities available in desktop computers, while students do their daily activities.

Several existing bioinformatics application suites have been ported into Bio-UnaGrid. Performance tests were executed using the BLAST algorithm and its distributed implementations using *query segmentation*, provided by NCBI BLAST [3], and *database segmentation*, provided by mpiBLAST [4]. Test results show that Bio-UnaGrid allows to define workflows and to execute them on dedicated and desktop grid infrastructures, providing speedups 10 times higher than the speed on a desktop computer.

This paper is organized as follows: section 2 presents the related works. Section 3 describes the BLAST application, the UnaGrid infrastructure and the LONI Pipeline application. Section 4 details the Bio-UnaGrid architecture, including its integration with MPI (Message Passing Interface) applications. Section 5 describes the implementation deployed on a university campus. Results and performance tests are described in Section 6. Section 7 concludes and presents future work.

 ISBN: 978-1-61208-137-3

II. RELATED WORKS

Bioinformatics projects require the use of application suites for analysis that regularly require large computing capabilities. An analysis executed like a sequential computational job on a personal computer, may take weeks or months. Cluster computing infrastructures solve this problem aggregating the computational capabilities of several homogeneous computers, a cluster, to parallel execute a set of computational jobs. In a computational cluster there is a cluster master, a computer that submits a subset of jobs to other cluster computers: the slave nodes. Computational clusters are created by individual research groups to support their computational requirements.

Grid computing emerged as a technology that allows different organizations with common goals to create a virtual organization (VO) to share resources such as data, hardware and software. A grid computing infrastructure can group a great number of heterogeneous resources to support the computational requirements of the VO. In a grid infrastructure a set of computational jobs can be distributed among clusters belonging to different administrative domains. A drawback of this approach is that grid infrastructures require complex management processes.

To execute a bioinformatics analysis on a cluster or grid infrastructure, the application must be wrapped or designed to operate on these infrastructures. Wrapped application adapt an existing standalone application so it can be executed in parallel like a set of smaller and independent jobs, this approach is known as bags-of-tasks (BoT). In applications designed to operate on distributed infrastructures, a single job is automatically executed using several processes executed coordinately in a computational cluster or grid, using, regularly a MPI implementation.

Like BLAST, other bioinformatics applications are frequently used by researchers to execute different analysis in a coordinated and dependent manner, called workflows or pipelines, which require HPC infrastructures. The manual and command-based workflow execution requires bioinformatics researchers to spend most of their time in: configuring application parameters, managing HPC infrastructures, managing scientific data, and linking partial results from one application to another. Several projects have been developed to facilitate the automatic workflow execution in other scientific fields. Projects like Khoros [5], 3D Slicer [6], SCIRun/BioPSE [7] and Karma2 [8] for image processing; MAPS [9] for brain images; Trident [10] for oceanography; Kepler [11] and Swift [12] for agnostic area; MediGRID [13] for biomedical; Pegasus [14], OpenDX [15], and Triana [16] for heterogeneous applications; Taverna [17] is a framework to executed bioinformatics applications in distributed environments using MyGrid [18] middleware.

Most of workflow tools have limitations for bioinformatics applications such as: they require applications to be recompiled or modified, they support internal data sources with specific data structures; they operate with specific platforms, and cluster or grid middlewares; they are designed for solving needs of specific scientific fields; and they regularly require researchers to use complex commands involving consuming-time tasks.

Although dedicated cluster and grid infrastructures provides large computational capabilities, in a university or enterprise campus there are tens or hundreds of desktop computers that are under-utilized. Some clusters or grid solutions take advantage of the idle computing capabilities for e-Science projects, these systems are called Desktop Grid and Volunteer and Computing Systems (DGVCSs). Several DGVCS solutions has been developed using different approaches, solutions and architectures, including SETI@home [19], BOINC [20], OurGrid [21], Integrade [22] and UnaGrid [2].

From the conducted survey and our experience working on the field, we concluded that an infrastructure allowing existing bioinformatics applications to be easily incorporated without modifications in workflows created through GUIs and executed on different HPC infrastructures is needed. The infrastructure also must allow the use of different data sources (internal and external), and the incorporation of MPI applications commonly used in bioinformatics projects. For getting results faster the infrastructure should operate on dedicated computing infrastructures and on DGVCS that take advantage of the idle processing capabilities of tens of desktop computers with different operating systems. Bio-UnaGrid integrates the capabilities of LONI Pipeline and UnaGrid to provide these features.

III. LONI PIPELINE, UNAGRID AND BLAST

A. LONI Pipeline

LONI Pipeline [1] is a free framework used to execute neuroscience workflows (see [23]); however its design and implementation allows any command line application (like most bioinformatics applications) to be incorporated. From a computational perspective, LONI differs from other workflow tools in several features: it does not require external application being recompiled, it supports external data storage sources, it is hardware platform independent, and it can be installed on different cluster or grid middlewares using the Pipeline plugin Application Programming Interface (API) [1]. From a research user perspective LONI was designed having in mind features like usability, portability, intuitiveness, transparency and abstraction of cluster or grid infrastructures.

LONI Pipeline uses a client/server model. The server is installed on the master computer of a computational cluster, managed by a distributed resource management (DRM) system such as Oracle Grid Engine (OGE). LONI uses standard execution commands, so any applications executed through command line can be incorporated without requiring recompilations or new developments. An application is defined and loaded in the server through a GUI, defining an XML file, called a LONI Module that contains information about path application executable, and number and types of input and output parameters.

LONI client is a lightweight and standalone Java application that can be executed on Windows, Linux or Mac desktops. A researcher connects through the LONI client

 ISBN: 978-1-61208-137-3

with the LONI server, using a customizable authentication protocol (LDAP, Database, etc.). The LONI modules are downloaded into the client and the researcher can begin to create workflows using simple drag and drop tools provided by an intuitive GUIs. Once a workflow is created, the researcher can begin its validation and execution. The workflow is executed (no commands are required) on the HPC infrastructure available for the master of LONI Pipeline. During its execution the researcher can disconnect of the LONI Server, and reconnect to query the partial results and to monitor the workflow state.

B. UnaGrid solution

UnaGrid is a DGVCS, developed at Universidad de los Andes, which provides the processing capabilities required by applications of different research areas at a university through the use and deployment of Customizable Virtual Clusters (CVCs) composed dynamically on demand, and executed on conventional desktop machines with Linux, Windows or Mac operating systems. The UnaGrid solution does not require applications to be recompiled or rewritten, and it has been used in projects of different research areas for executing BoT applications [2].

UnaGrid is a DGVCS that takes advantage of the idle processing capabilities of desktop computers within computer labs, in a non-intrusive manner, through the execution of Customized Virtual Clusters (CVCs). A CVC is a set of commodity and interconnected physical desktops computers executing virtual machines (VMs). While a student do his/her activities, a VM, playing a slave role of the CVC, is executed as a low-priority and background process on each desktop computer used by a student. A dedicated machine for the CVC plays the role of the cluster master. All of these VMs in execution make up a CVC, which has the operating system (mainly Linux), applications, and middleware required by the research group.

This model allows researcher to continue executing applications within their native environments, guaranteeing high usability of the infrastructure. Users access a CVC through a SSH connection to the CVC master. The use of virtualization tools such as VMware, Oracle Virtual Box, Citrix or Microsoft System Center, allows adding and taking advantage of the capabilities of tens or hundreds of machines in computer labs that have Windows, Linux or Mac operating systems, as well as the faculty to assign and limit the resources consumed by the VMs.

When a research group requires its CVC, they can deploy it on demand using a novel Web application called GUMA (Grid Uniandes Management Application). GUMA allows deploying on demand a previously configured CVC. A researcher securely access GUMA (using a Web browser) and defines the size (number of VMs) of the CVC he/she requires. GUMA automatically deploy the VMs on selected desktops, hiding the complexities associated with the location, distribution and heterogeneity of computing resources, and providing an intuitive graphical interface. GUMA also provides services for selection, shutdown and monitoring of physical computers and VMs. GUMA offers high usability to the UnaGrid solution, using the on-demand approach.

C. BLAST algorithmn

BLAST is a heuristic based application widely used to search for similarities between a set of biological sequences *S* and sequences in a database *D*. Wrapped applications using the NCBI BLAST implementation use the *query segmentation* approach. In this case a sequence query set *S* is divided in *n* subsets of sequences S_i. Each subset S_i is compared with the complete database *D*. A number *n* of NCBI BLAST independent jobs are executed on a computational cluster. Each independent job, executed by a cluster slave, compares a subset S_i with the database *D*. After the *n* jobs have been executed, the files generated by each job are merged in a single file containing all results.

Query segmentation offers great performance when the complete database *D* can be stored on the RAM memory of each cluster node. mpiBLAST adds *database segmentation*. mpiBLAST divides a sequence database *D* in *m* subsets D_j, and the sequence set *S* is compared with each subset D_j in a coordinated manner. A search using mpiBLAST is executed through a single mpiBLAST job executed coordinately on *m* slave nodes (logical MPI cluster) of a physical computational cluster. Each logical slave node compares the sequence set *S* with a subset D_j, and partial results of all logical slave nodes are stored in a unique shared file.

Several BLAST variations modify BLAST algorithm to execute relevant specific analysis and accelerate searches [24] such as PSI-BLAST, PHI-BLAST, Mega-BLAST MPBLAST, WU-BLAST2 and BLASTZ. Other BLAST solutions adapt or implement BLAST to operate efficiently on specific computing infrastructures [24]. BLAST solutions such as HGBS and FPGA-based BLAST require specialized hardware. BLAST implementation for dedicated cluster computing infrastructures are BeoBlast, NBLAST, Soap-HT-BLAST, mpiBLAST, Hyper-BLAST, dBLAST parallelBLAST, BLAST.pm, Parallel BLAST++, ScalaBLAST, pioBLAST and avaBLAST. These solutions use processing scheduling systems like Condor, PBS, OGE or Torque, and distributed storage systems like NFS or PVFS.

Solutions such as TurboBLAST, GBTK, GridBLAST CloudBLAST, G-BLAST, PackageBLAST and mpiBLAST-PIO, operates on dedicated grid computing infrastructures with standard middleware like Globus. W.ND BLAST [25], BOINC BLAST [26], and BLAST on BitDew [27], use the processing capabilities of DGVCSs to execute searches. W.ND BLAST only operates on Windows desktops. BOINC and BitDew require every application being modified.

IV. BIO-UNAGRID INFRASTRUCTURE

We propose the Bio-UnaGrid infrastructure, which was designed to facilitate to bioinformatics researchers the use of bioinformatics applications, the automatic execution of workflows, and the use of HPC infrastructures. With Bio-UnaGrid bioinformatics researchers can use the processing capabilities of dedicated computational clusters, and the idle processing capabilities provided by the UnaGrid DGVCS.

 ISBN: 978-1-61208-137-3

Using the LONI features, Bio-UnaGrid allows that existing and new bioinformatics application can be easily integrated and used by researchers without been recompiled or modified. Current version of LONI Pipeline does not support the execution of MPI applications such as mpiBLAST, however, Bio-UnaGrid support the execution of this type of applications.

A. Bio-Unagrid Architecture

Bio-UnaGrid is based on the integration of two main solutions: LONI Pipeline and UnaGrid. The Bio-UnaGrid architecture is shown in Figure 1.

The *LONI Pipeline Client* is the main entry point of bioinformatics researchers to the Bio-UnaGrid infrastructure. Researchers use the LONI client to connect with the *LONI Pipeline Server*, through an authentication process against a *User Database*. Each researcher receives a username and a password. After authentication the researcher can view the bioinformatics applications installed on the *LONI Pipeline Server* through the *LONI Pipeline Client*, and he/she can proceed to create the workflows, using the bioinformatics *LONI Module Library*, drag and drop, and other tools.

For each *LONI Module* (a bioinformatics application), the researcher must define the input and output parameters (which can take values such as strings, path files, path directories, numbers and related lists), and how each *LONI Module* is connected with other LONI Modules of the workflow. A single *LONI Module* may be used to execute several jobs, for example if the BLASTn *LONI Module* of the NCBI BLAST suite, receives as input parameters ten sequence query files and a database, like *nt*, the *LONI Pipeline Server* will execute ten independent BLASTn jobs, each one comparing a sequence query file with the *nt* database. Once a workflow has been created, researchers can execute it, sending it to the *LONI Pipeline Server*. Researcher can monitor and visualize the status of the workflow, and download partial results, through GUIs of the *LONI Pipeline Client*.

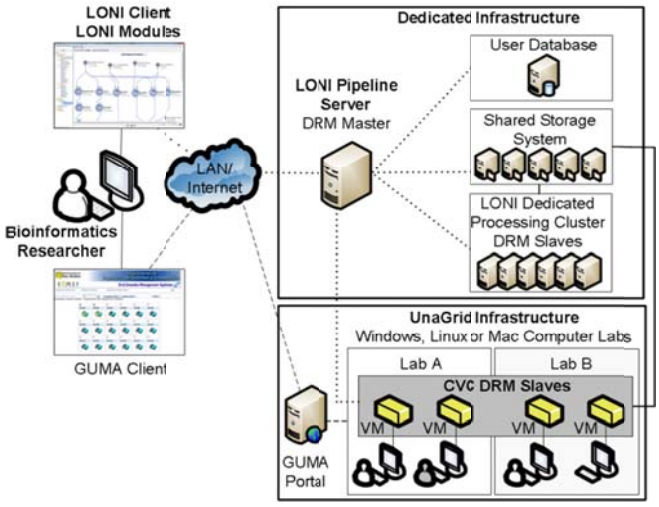


Figure 1. Bio-UnaGrid architecture.

The *LONI Pipeline Server* is a dedicated server with the master component of a DRM such as OGE installed locally. The *LONI Pipeline Server* receives the workflows sent by researchers; divide the workflows in individual jobs and send them, in an order manner, to the *DRM Master*. The *DRM Master* distributes the jobs to the *DRM Slaves* which can be running on a *Dedicated Processing Cluster* or on a CVC of UnaGrid, composed of VMs executed on heterogeneous desktop computers, which we will call *CVC DRM Slaves*.

The *Dedicated Processing Cluster* is permanently available for researchers; however its capabilities are limited. To provide more processing capabilities, the *CVC DRM Slaves* can be deployed on demand using the *GUMA Portal* of the UnaGrid infrastructure. Before a researcher execute a workflow or during the workflow execution, he/she can access the *GUMA Portal* using a username and a password. Through *GUMA Portal* researchers define the execution time and number of virtual machines (*VMs*), previously configured by the UnaGrid support team, he/she required for the workflow execution. *GUMA Portal* deploys bioinformatics *VMs* on different desktop computers distributed along computers labs in the university campus. When the *CVC DRM Slaves* are turn on by GUMA, they contact the *DRM Master*. The *DRM Master* begins to submit individual jobs to *CVC DRM Slaves*.

A *Shared Storage System* is used to store the *LONI Modules*, the binaries of bioinformatics applications, and input and output data required for the workflow execution, including genomic databases. The *DRM Master*, the *Dedicated Processing Cluster* and the *CVC DRM Slaves* access the *Shared Storage System* during workflow executions. During workflow creation researchers can select local files of the personal computer where the *LONI Pipeline Client* is executed, and these files are transferred automatically to the *LONI Pipeline Server* and stored in the *Shared Storage System*.

For executing workflows bioinformatics researchers do not have to worry about applications' commands or complex HPC infrastructures. They only need to define workflows using drag and drop tools provided by *LONI Pipeline Client*, and run workflows using a single click. Workflows are executed faster and transparently on dedicated clusters, and on tens or hundreds of commodity desktop computers provided by UnaGrid. When a workflow is finished, researchers can download the results of all *LONI modules* using GUIs, and they are ready to do their bioinformatics analysis.

B. Bioinformatics applications on Bio-UnaGrid

The process to incorporate existing and new bioinformatics application suites, like NCBI BLAST, into Bio-UnaGrid infrastructure are summarized in five steps.

1) Installation and configuration of the application suite on the Shared Storage System: This installation is executed from the *LONI Pipeline Server*.

2) Creation of a LONI Pipeline Module using the LONI Pipeline Client for each executable of the suite: for this it is

necessary to identify the input and output parameters of the executable, specifying its dependences and data types.

3) Individual tests for each LONI Pipeline Module of the suite from the LONI Pipeline Client.

4) Storage of the LONI Pipeline Module in the LONI Pipeline Server.

5) Using the modules from workflows created by bioinformatics researchers in the LONI Pipeline Client.

. This process is executed by the HPC infrastructure's system administrator only once for each bioinformatics application. After a *LONI Module* for a bioinformatics application is created, all researchers can use it in their workflows. This process allows new applications to be agilely incorporated, facilitating the creation of more complex workflows. Our experience in the deployment of several application suites show that any application that can be used from the command line can be integrated into the Bio-UnaGrid solution.

C. MPI applications on Bio-UnaGrid

At the time of this development, LONI did not support MPI applications. To allow the execution of MPI applications, we developed a wrapper, called *mpiJobManager*, which uses the Distributed Resource Management Application API (DRMAA). DRMAA allows sending and monitoring jobs executed on computational cluster using a Distributed Resource Manager (DRM) such as the Oracle Grid Engine (OGE). To execute an MPI Application, such as mpiBLAST, a researcher uses the *MPI LONI Module* of the MPI application to specify the number of MPI processes to be used. *MPI LONI Modules* are received by the *LONI Pipeline Server* and executed on a *DRM Slave* (a dedicated server or desktop computer) using the *mpiJobManager* wrapper.

When an *mpiJobManager* job is executed, it executes a Linux script, called *launcherMPIJob*. The *launcherMPIJob* script sends an MPI job to the *LONI Pipeline Server*, specifying the process number of the job. The *LONI Pipeline Server* receives the MPI job and proceeds to execute it on the *DRM slaves*. During its execution, the MPI job is monitored by the *LONI Pipeline Server* using the *mpiJobManager*. Because of the design of the LONI Pipeline Server, researchers can query the results of a whole MPI job but not the status of its individual processes.

V. IMPLEMENTATION

Bio-UnaGrid was implemented at Universidad de los Andes, involving the current bioinformatics dedicated processing cluster from the Department of Biological Sciences, some dedicated servers and a computer lab with Windows desktop computers from the Department of Systems and Computing Engineering. Bio-UnaGrid was implemented to support several genomic projects related to coffee, potato and cassava, which seek genomic analysis to improve coffee, potato and cassava production affected by different biological organisms that decrease their production [28] [29]. The implementation is illustrated in Figure 2.

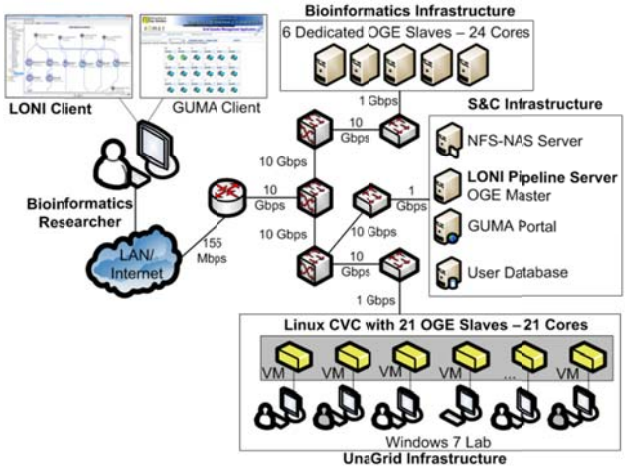


Figure 2. Bio-UnaGrid implementation.

The *LONI Pipeline Server* version 5.0.2 and the *DRM Master* of OGE version 6.2 update 5, were installed on a dedicated server. We use an NFSv3-NAS (Network File System version 3 - Network Attached Storage) solution like the *Shared Storage System*. The *Database User* was installed on a dedicated server using MySQL version 5.1.5. On the dedicated cluster used by the Department of Biological Sciences composed by 6 servers each one with an Intel Xeon X5560 Quad Core processor of 2.8 GHz and 4 GB of RAM, was installed the Slave component of OGE, so these servers now can process jobs sent from *LONI Pipeline Clients*.

To test the performance of bioinformatics applications on the desktop computers, a CVC with 21 VMs was deployed on a computer lab with Windows 7 desktops computers, which have an Intel i5 processor of 3.46 GHz and 8 GB of RAM. On the CVC we installed Debian 4.0 and the Slave component of OGE, so these VMs can also process jobs sent from *LONI Pipeline Clients*. VMs were configured with a CPU core and 4 GB of RAM. All nodes are interconnected through 1 GbE and 10GbE links and fault tolerance mechanisms were configured for BoT applications using the OGE capabilities.

Four bioinformatics application suites have been installed on the Bio-UnaGrid implementation: NCBI BLAST version 2.2.20, HMMER version 2.3.2, InterProScan version 4.6 and mpiBLAST version 1.6.0. Because mpiBLAST requires the use of an MPI implementation, the MPICH-2 implementation version 1.2.7 was installed. 21 *LONI Modules* were created for the application suites: 6 for NCBI BLAST (BLASTn, BLASTp, BLASTx, tBLASTn, tBLASTx and MEGABLAST); 3 for HMMER (Build, Search and Calibrate); 11 for InterProScan (BlastProdom, Coils, Gene3D, PIR, Panther, Pfam, SEG, SMART, SuperFamily, TIGRfam and fPrintScan) and 1 for mpiBLAST.

Researchers now can execute these bioinformatics applications from workflows. An example of a workflow created through the *LONI Pipeline* Client with these applications, and executed on the HPC infrastructure is shown in Figure 3.

 ISBN: 978-1-61208-137-3

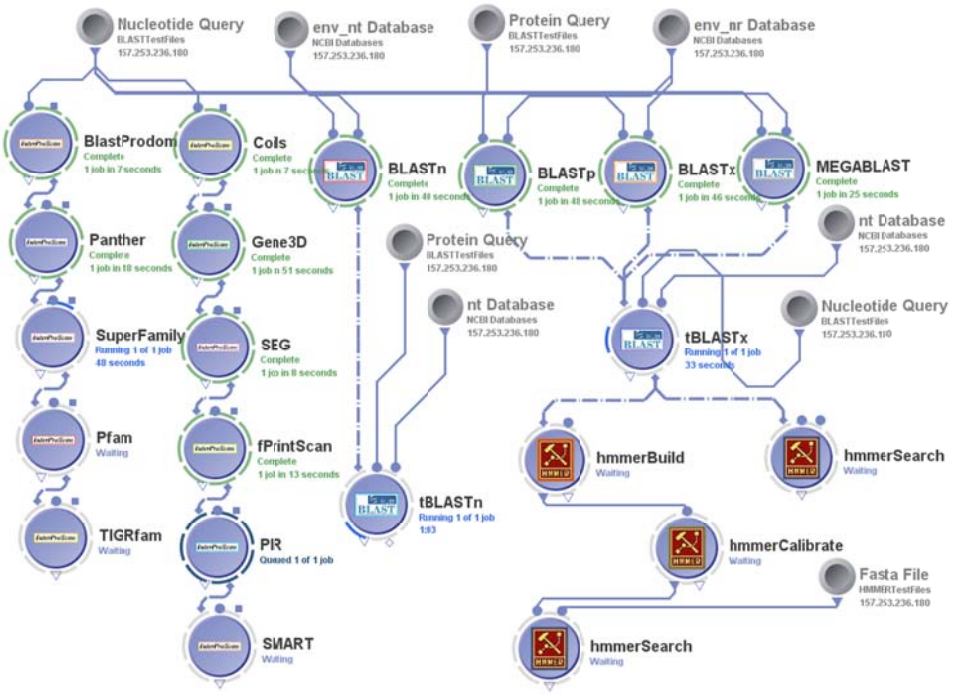


Figure 3. A bioinformtics workflow executed on Bio-UnaGrid.

During the workflow execution bioinformatics researchers can see the status of the complete workflow, and of each individual *LONI Module* (Completed, Running, Queued, and Waiting). In the workflow shown in Figure 3, all files (query, databases and FASTA) are stored in the *Shared Storage System*. This workflow is being executed on the configured distributed infrastructure and researchers do not have to use any command; all workflow operations are executed using GUIs.

VI. PERFORMANCE TESTS AND RESULTS

Bio-UnaGrid support dedicated clusters and desktop machines; we executed performance tests in both environments, in the dedicated cluster and in the UnaGrid CVC. We selected the high popular BLAST algorithm to execute these tests. A detailed time-performance characterization for the execution of BLAST (using *query segmentation*) and mpiBLAST in cluster and grid infrastructures can be found in [30] and [4] respectively.

For testing the performance of Bio-UnaGrid for executing BoT applications we used the NCBI BLASTn application, varying the number of CPU cores and the size of the sequence set. We executed BLASTn searches between nucleotide sets of 480 (487 KB) and 960 (954 KB) sequences, and the *nt* database (28 GB). The searches were executed using 5, 10, 15 and 20, BLASTn independent processes (BoT BLASTn). In both environments each independent process was executed using a CPU core to compare a sequence set with the complete *nt* database. We use *query segmentation*, using a LONI Module, for dividing the sequence sets in the same number of processes; subsets of 96, 72, 48 and 24 sequences were used for the 480 sequence set.

All tests were executed 3 times and we calculated the average time. Two metrics were used in the performance tests, the execution time and the speedup. The speedup is a measure that indicates the improvement in the execution time of a parallel algorithm when it is compared with same algorithm executed in a sequential manner [31]. The sequential searches were executed in a desktop of the UnaGrid CVC and in a server of the dedicated cluster, using a standalone BLASTn search. Sequential searches were executed using a single CPU core. The execution times of the sequential searches with the sets of 480 and 960 sequences on the dedicated server were 6,436 and 13,140 seconds respectively, and on the desktop were 12,288 and 19,488 seconds.

The execution times of the searches on the dedicated cluster and the UnaGrid CVC, using BoT BLASTn, are shown in figure 4.

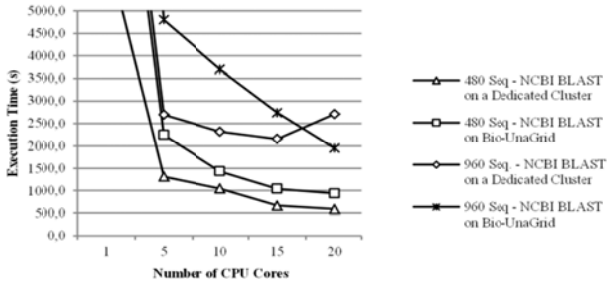


Figure 4. Execution time of BoT BLASTn searches between sets of 480 and 960 sequences, and the *nt* database.

 ISBN: 978-1-61208-137-3

Results in Figure 4 show that the execution time of the searches in both environments, decrease when the process number (CPU cores) is increased (except in the search with 960 sequences on the dedicated cluster using 20 processes). Taking as reference the execution time of the sequential searches on each environment, we calculated the speedups shown in Figure 5. The execution time of the search with 480 sequences on the dedicated cluster was reduced from 6,436 to 589 seconds (using 20 processors), providing results 10.9 times faster (10.9x). On the Bio-UnaGrid CVC the execution time was reduced from 12,288 to 936 seconds, using 20 processors (13.1x faster). For the search with 960 sequences, the execution time on the dedicated cluster was reduced from 13,140 to 2,152 seconds when 15 processors were used (6.1x faster), and on the Bio-UnaGrid CVC from 19,488 to 1,946 second using 20 processors (10x faster).

For testing the execution of MPI applications on both environments, we executed the same tests using mpiBLAST. In these tests we executed MPI processes coordinated on 5, 10, 15 and 20 CPU cores. In each test, the additional process called *mpiJobManager* was executed in another CPU core (this is the reason to show 21 OGE Slaves on Figure 2). We used *database segmentation* to divide the database in 5, 10, 15 and 20 partitions using the *mpiformatdb* installed with mpiBLAST, using a LONI Module. The execution times of these searches using mpiBLAST are shown in Figure 6. The speedups for these searches are shown in Figure 7.

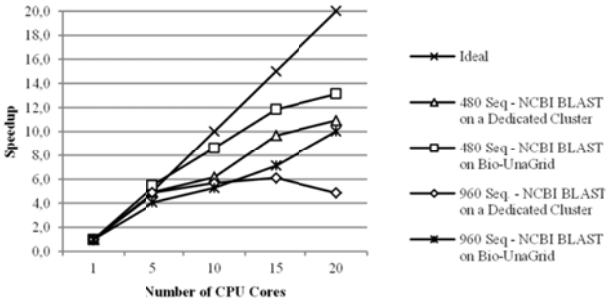


Figure 5. Speedups of BoT BLASTn searches between sets of 480 and 960 sequences, and the *nt* database.

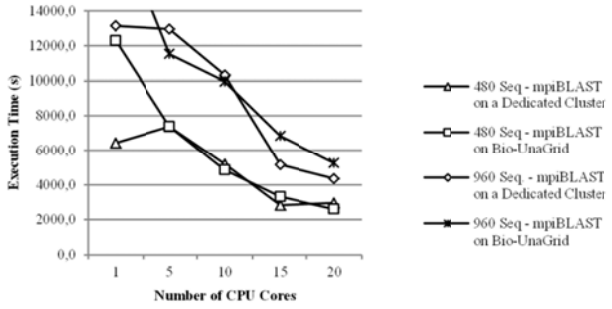


Figure 6. Execution times of mpiBLAST searches between sets of 480 and 960 sequences, and the *nt* database.

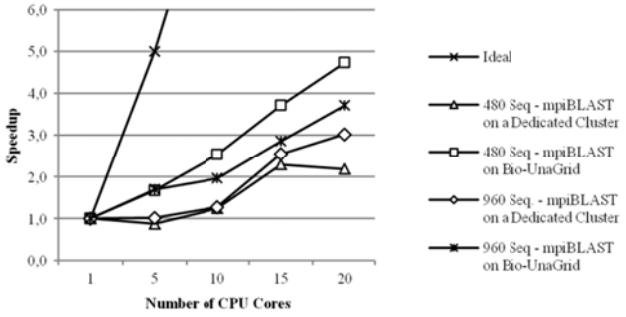


Figure 7. Speedups of mpiBLAST searches between sets of 480 and 960 sequences, and the *nt* database.

Similar to previous tests, using mpiBLAST on both environments, in most tests the execution time is reduced when the number of processor is increased. On the dedicated cluster the execution time for the search with 480 sequences was reduced from 6,436 to 2,815 seconds when 15 processors were used (2.3x faster). On the Bio-UnaGrid CVC the execution time was reduced from 12,288 to 2,597 seconds, using 20 processors (4.7x faster). For the search with 960 sequences, the execution time on the dedicated cluster was reduced from 13,140 to 4,154 seconds when 20 processors were used (3x faster), and on the Bio-UnaGrid CVC from 19,488 to 5,255 second using 20 processors (3.7x faster).

Results from Figures 5 and 7 show that Bio-UnaGrid can provide speedups similar or higher to those provided by a dedicated cluster, decreasing the result generation time when the number of processes is increased. Using the idle processing capabilities, Bio-UnaGrid provides additional processing capabilities to those provide by dedicated computational clusters, decreasing more than 13.1 times (13.1x) the execution time of bioinformatics workflows.

VII. CONCLUSIONS AND FUTURE WORK

We proposed Bio-UnaGrid, an infrastructure that allows bioinformatics researchers to easily define workflows using GUIs and execute them on different cluster and grid computing infrastructures with a simple click. Bio-UnaGrid is highly extensible as existing bioinformatics applications can be incorporated without modifications. With Bio-UnaGrid, researchers focus on analysis of application results, not on technical issues of distributed computing infrastructures. To take advantage of more processing capabilities, besides using dedicated computing infrastructures, Bio-UnaGrid also use the idle processing capabilities of tens or hundreds of desktop computers commonly available in computer labs.

We implemented Bio-UnaGrid in a dedicated cluster composed of 6 servers, and a computer lab with several bioinformatics application suites such as NCBI BLAST, HMMER, InterProScan and mpiBLAST. Performance tests with NCBI BLASTn and mpiBLAST show that Bio-UnaGrid can reduce the sequential execution time up to 13.1 times faster. These results show promising

 ISBN: 978-1-61208-137-3

opportunities for bioinformatics researchers to get results in shorter times using the idle processing capabilities available in computer labs using Windows, Linux or Mac desktops.

As future work, we will incorporate more bioinformatics applications suites in Bio-UnaGrid and we will execute new performance tests with other applications such as HMMER, MPI-HMMER, ClustalW and ClustalW-MPI, in a larger grid deployment involving hundreds of heterogeneous desktop computers in different administrative domains. We will implement a shared storage system more scalable than NFSv3. We also plan to implement a fault tolerance mechanism for MPI applications.

REFERENCES

[1] I. Dinov et al., "Neuroimaging Study Designs, Computational Analyses and Data Provenance Using the LONI Pipeline," PLOS ONE, vol. 5, Sep. 2010, doi:10.1371/journal.pone.0013070.

[2] H. Castro, E. Rosales, M. Villamizar, and A. Miller, "UnaGrid - On Demand Opportunistic Desktop Grid," Proc. 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing, June 2010, pp. 661-666, doi:10.1109/CCGRID.2010.79.

[3] S. Altschul et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Research, vol. 25, July 1997, pp. 3389-3402, doi:10.1093/nar/25.17.3389.

[4] A. Darling, L. Carey, and W. Feng, "The design, implementation, and evaluation of mpiBLAST," Proc. ClusterWorld 2003, June 2003.

[5] S. Kubica, T. Robey, and C. Moorman, "Data parallel programming with the Khoros data services library," Lecture Notes in Computer Science, vol. 1388, 1998, pp. 963-973, doi:10.1007/3-540-64359-1_762.

[6] S. Pieper, M. Halle, and R. Kikinis, "3D SLICER," Proc. IEEE International Symposium on omedical Imaging: Nano to Macro, April 2004, pp. 632-635, doi:10.1109/ISBI.2004.1398617.

[7] R. MacLeod, D. Weinstein, D. Germain, D. Brooks, C. Johnson, and S. Parker, "SCIRun/BioPSE: integrated problem solving environment for bioelectric field problems and visualization," Proc. IEEE International Symposium on Biomedical Imaging: Nano to Macro, April 2004, pp. 640-643, doi:10.1109/ISBI.2004.1398619.

[8] Y. Simmhan, B. Plale, and D. Gannon, “Karma2: Provenance Management for Data Driven Workflows,” International Journal of Web Services Research, vol. 5, April 2008, pp. 1-22, doi:10.4018/jwsr.2008040101.

[9] B. Lucas, B. Landman, J. Prince, and D. L. Pham, “MAPS: A Free Medical Image Processing Pipeline,” Proc. 14th Annual Meeting of the Organization of Human Brain Mapping, June 2008.

[10] Y. Simmhan, R. Barga, C. Ingen, E. Lazowska, and A. Szalay, “Building the Trident Scientific Workflow Workbench for Data Management in the Cloud,” Proc. Third International Conference on Advanced Engineering Computing and Applications in Sciences, Oct. 2009, pp. 41-50, doi:10.1109/ADVCOMP.2009.14.

[11] B. Ludäscher et al., “Scientific workflow management and the Kepler system,” Concurrency and Computation: Practice & Experience, vol. 18, Aug. 2006, pp. 1039-1065, doi:10.1002/cpe.v18:10.

[12] Y. Zhao et al., “Swift: Fast, Reliable, Loosely Coupled Parallel Computation,” Proc. IEEE Congress on Services, IEEE, July 2007, pp. 199-206, doi:10.1109/SERVICES.2007.63.

[13] D. Krefting et al., “MediGRID: Towards a user friendly secured grid infrastructure,” The international journal of grid computing-theory methods and applications, vol. 25, March 2009, pp. 236-336, doi:10.1016/j.future.2008.05.00.

[14] E. Deelman et al., “Pegasus: Mapping Scientific Workflows onto the Grid,” Lecture Notes in Computer Science, vol. 3165, 2004, pp. 131-140, doi:10.1007/978-3-540-28642-4_2.

[15] D. Thompson, J. Braun, and R. Ford, “OpenDX: Paths to Visualization,” Proc. VIS, Inc.

[16] D. Churches et al., “Programming scientific and distributed workflow with Triana services,” Concurrency and Computation: Practice & Experience, vol. 18, Aug. 2006, pp. 1021-1037, doi:10.1002/cpe.v18:10.

[17] T. Oinn et al., “Taverna: lessons in creating a workflow environment for the life sciences,” Concurrency and Computation: Practice and Experience, vol. 18, Aug. 2006, pp. 1067-1100, doi:10.1002/cpe.v18:10.

[18] MN Alpdemir et al., “Contextualised workflow execution in MyGrid,” Lecture Notes In Computer Science, vol. 3470, 2005, pp. 444-453, doi:10.1007/11508380_46.

[19] D. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, “SETI@home An Experiment in Public-Resource Computing,” Communications of the ACM, vol. 45, Nov. 2002, pp. 56-61, doi:10.1145/581571.581573.

[20] D. Anderson, “BOINC: A System for Public-Resource Computing and Storage,” Proc. in 5th IEEE/ACM International Workshop on Grid, IEEE, Nov. 2004, doi:10.1109/GRID.2004.14.

[21] B. Francisco and M. Rodrigo, “The OurGrid Approach for Opportunistic Grid Computing,” Proc. First EELA-2 Conference, Feb. 2009.

[22] A. Goldchleger, F. Kon, A. Goldman, M. Finger, and G. Bezerra, “InteGrade: object-oriented Grid middleware leveraging the idle computing power of desktop machines,” Concurrency and Computation: Practice and Experience, vol. 16, March 2004, pp. 449-459, doi:10.1002/cpe.824.

[23] LONI Pipeline, "Laboratory of Neuro Imaging", [Online], http://pipeline.loni.ucla.edu.

[24] M. Sousa, A. Melo, and A. Boukerche, “An adaptive multi-policy grid service for biological sequence comparison,” Journal of parallel and distributed computing, vol. 70, Feb. 2010, pp. 160-172, doi:10.1016/j.jpdc.2009.02.009.

[25] S. Dowd, J. Zaragoza, J. Rodriguez, M. Oliver, and P. Payton, “Windows.NET network distributed basic local alignment search toolkit (W.ND-BLAST),” BMC Bioinformatics, vol. 6, April 2005, doi:10.1186/1471-2105-6-93.

[26] S. Pellicer, G. Chen, K. Chan, and Y. Pan, “Distributed Sequence Alignment Applications for the Public Computing Architecture,” IEEE Transactions on NanoBioscience, vol. 7, March 2008, pp. 35-43, doi:10.1109/TNB.2008.2000148.

[27] H. He, G. Fedak, B. Tang, and F. Cappello, “BLAST Application with Data-aware Desktop Grid Middleware,” Proc. 9th IEEE/ACM International Symposium on Cluster Computing and the Grid, IEEE, May 2009, pp. 284-291, doi:10.1109/CCGRID.2009.91.

[28] A. Vargas, et al., “Characterization of Phytophthora infestans Populations in Colombia: First Report of the A2 Mating Type,” Phytopathology, Sep. 2009, pp. 82-88, doi:10.1094/PHYTO-99-1-0082.

[29] S. Restrepo et al., “Computational Biology in Colombia,” PLOS Computational Biology, vol. 5, Oct. 2009, doi:10.1371/journal.pcbi.1000535.

[30] E. Afgan, and P. Bangalore, "Performance Characterization of BLAST for the Grid," Cluster Computing, vol. 13, Feb. 2010, pp. 385–395, doi:10.1007/s10586-010-0121-z.

[31] A. Shah, G. Folino, and N. Krasnogor, “Toward High-Throughput, Multicriteria Protein-Structure Comparison and Analysis,” IEEE Transactions on NanoBioscience, vol. 9, June 2010, pp. 144-155, doi:10.1109/TNB.2010.2043851.

 ISBN: 978-1-61208-137-3

Automated Identification of Micro-Embolic Events Using Auditory Perception Features Extracted from Mel Frequency Cepstrum Coefficients

Lingke Fan
Department of Medical Physics
University Hospitals of Leicester NHS Trust
Leicester, UK
lingke.fan@uhl-tr.nhs.uk

David H. Evans
Department of Cardiovascular Sciences
University of Leicester
Leicester, UK
dhe@le.ac.uk

***Abstract*—Directional auditory perception features and energy information were extracted from transcranial Doppler (TCD) ultrasound signals and were used to automatically identify micro-embolic events (MEEs) using a novel embolic signal analysis approach. Three directional analysis methods were evaluated for their MEE identification performance using data recorded during cardiac surgery. The analysis methods were based on Mel frequency cepstrum coefficients (MFCCs), linear frequency cepstrum coefficients (LFCCs) and linear spectral components (LSCs). The results of these preliminary off-line evaluations showed that: a) the auditory perception and energy features of Doppler signals could play an important role in the identification of MMEs; b) MFCC-based analysis seems to be superior to the other two methods, achieving a sensitivity of 95.99%, a specificity of 96.43% and a positive predictive value (PPV) of 95.64%. Future studies using larger data sets and more complicated detection implementation (a rather basic rule-based system was used in the detection stage here) could further confirm or improve the identification performance and robustness of MFCC-based systems.**

Keywords-Doppler ultrasound; automated emoblus identification; TCD; MFCC application; directional analysis, auditory perceptul evaluation; artifact rejection, knowledge-based systems.

I. INTRODUCTION

Identification of cerebral micro-embolic events (MEEs) utilizing transcranial Doppler (TCD) ultrasound is often used to provide valuable information in clinical and research settings [1]-[3]. Therefore considerable effort has been made using a wide range of signal analysis approaches (e.g., [4]-[7]), trying to improve the reliability of automated MEE identification (AMEEI) systems.

A novel AMEEI approach is proposed in this study with the following objectives:

1. To emulate and use human auditory perceptual features in AMEEI. Human auditory perception plays an important role in manual MEE detection procedures that are often regarded as quite reliable in research settings [8], [9]. Hence perceptual features are studied here to explore their potential to improve AMEEI performance.
2. To extract the directional signatures of embolic signals and apply them in the classification of MEEs. It is widely acknowledged that, in general, MEEs are unidirectional events whilst artifacts are bi-directional ones [10], [11]. However, the way to efficiently use this directional information for AMEEI purposes is yet to be addressed in detail and in depth.

Three AMEEI methods were evaluated in this study. Mel frequency cepstrum coefficients (MFCCs), linear frequency cepstrum coefficients (LFCCs) and linear spectral components (LSCs) were used to extract basic perceptual and perception-related energy features from recorded signals. Doppler signals containing both sporadic MEEs and embolic showers recorded from patients during cardiac surgery were used to evaluate these methods. A "gold standard" based on the results from manual MEE detection was applied to all the evaluations of the AMEEI performance.

This paper is divided into 5 sections. The details for the design and implementation of the AMEEI methods are given in the next section. Section III contains the details of the clinical data and evaluation procedures. Experimental results, AMEEI performance comparisons and discussion are provided in Section IV, which is followed by conclusions and future work in Section V.

II. METHODS

A. The MFCC-based Directional Analysis Approach

Imagining that a trained human operator is listening to both the forward and reverse Doppler signals to compare the auditory perceptions, the perceived difference between the two signals can be emulated by calculating the directional perceptual distance (DPD) between the two Doppler signals. The DPD at time index n is defined as (the sampling frequency is 12.5kHz and the data window length is 10.24ms):

$$DPD_n = \left[\sqrt{\sum_{i=1}^{N} [MfccF(i) - MfccR(i)]^2} \right]_{t=n} \quad (1)$$

Here MfccF(i) and MfccR(i) are the ith MFCC elements (from the "classical" MFCC calculation [12]) for the forward and reverse signals at time index n, respectively; and N=7 is the number of the MFCC elements used in the DPD calculation. Based on the DPD, the mean directional perceptual distance (MDPD) and the differential directional

 ISBN: 978-1-61208-137-3

perceptual distance (DDPD) at time index n can be calculated as:

$$MDPD_n = \frac{1}{100}\sum_{i=n-50}^{n+49} DPD_n \quad (2)$$

$$DDPD_n = DPD_n - MPD_n \quad (3)$$

An energy parameter is defined in this study as the estimated signal to background ratio (ESBR):

$$ESBR_i = 10\log_{10}\left(EngS_i/EngB_i\right) \quad (4)$$

$EngS_i$ is the estimated signal energy at time index i and $EngB_i$ is the estimated energy (averaged in a 154-ms window) for the background signals around time index i. These energy estimations are calculated using the spectral magnitude elements obtained using a 128-point FFT and a Hamming window (overlap=50%).

Using above results and an 8 frame (40.96ms) moving window, the relative perceptual and energy correlation (RPEC) at time index n can be defined as:

$$RPEC_n = \frac{8\sum_{i=n-4}^{n+3} DDPD_i\ ESBR\_F_i - \sum_{i=n-4}^{n+3} DDPD_i \sum_{i=n-4}^{n+3} ESBR\_F_i}{\sqrt{8\sum_{i=n-4}^{n+3} DDPD_i^2 - \left[\sum_{i=n-4}^{n+3} DDPD_i\right]^2}\sqrt{8\sum_{i=n-4}^{n+3} ESBR\_F_i^2 - \left[\sum_{i=n-4}^{n+3} ESBR\_F_i\right]^2}} \quad (5)$$

where $ESBR\_F_i$ is the ESBR for the forward signal at time index i. The $RPEC_n$ is actually the correlation coefficient between the differential perceptual change and the relative energy variation for the forward signal within the moving window. It should become high for MEEs (the higher the relative embolic signal level in the forward direction, the larger the relative directional perceptual difference) and low for normal artifacts due to their bi-directional properties.

Another directional parameter derived from the RPEC is the averaged relative perceptual and energy correlation (ARPEC) within a positive ESBR_F peak between the two time indexes n1 and n2 that form the two boundaries of the peak. The ARPEC corresponding to the ESBR_F peak is defined as:

$$ARPEC = \begin{cases} \frac{1}{J2-J1+1}\sum_{i=J1}^{J2} RPEC_i, & SBR\_F_i > 0 \text{ for } i = J1, J1+1, \cdots, J2; \\ 0, & \text{else} \end{cases} \quad (6)$$

An MFCC-based automated identification unit was designed and developed to detect MEEs with a certain signal threshold (i.e., ESBR≥7dB), using the above defined parameters. Fig.1 shows the block-diagram of this rather basic rule-based system and Table I lists the parameters used and their details.

B. The LFCC-based DirectionalAanalysis Approach

An LFCC-based MEE detection unit was designed and developed using a system structure similar to that shown in Fig.1. The main difference, however, is that the MFCCs and all the MFCC-based parameters were replaced by their LFCC counterparts in the directional/perceptual evaluations. The calculation of the LFCCs is same as that of the MFCCs, except that mel scale filters are not used [12].

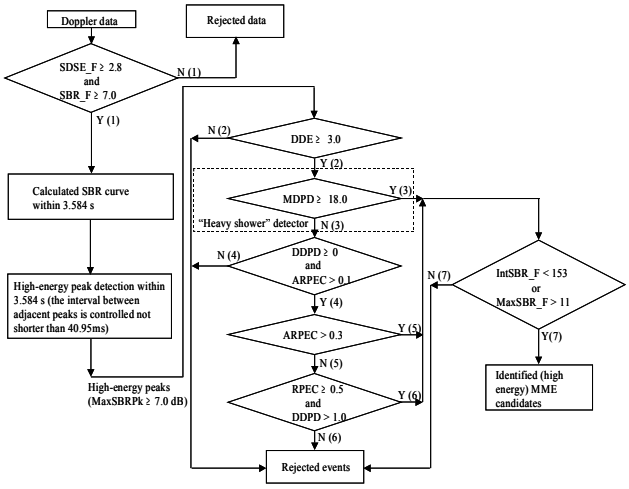


Figure 1. An MFCC-based MEE detection unit. The details of all the parameter are listed in Table I and all the evaluation thresholds were empirically chosen.

TABLE I. THE PARAMETERS USED IN THE MFCC-BASED MEE DETECTION UNIT

Parameter	Details	Units
SDSE_F	The standard deviation of the energy for the forward signal within a 102 ms window.	dB
ESBR_F	The estimated signal to background ratio for the forward signal.	dB
MaxSBRPk	The maximum magnitude within a single positive ESBR_F peak.	dB
DDE	The directional differential energy: the energy of the forward signal minus the energy of the reverse signal.	dB
MDPD	The mean directional perceptual distance within a 512 ms window—see (2).	N/A
DDPD	The differential directional perceptual distance—see (3).	N/A
RPEC	The relative perceptual and energy correlation—see (4).	N/A
ARPEC	The averaged relative perceptual and energy correlation—see (5).	N/A
IntSBR_F	The interval for the longest ESBR_F segment containing all-positive points (within a 3.6 s window). The interval could contain multiple positive ESBR_F peaks.	ms
MaxSBR_F	The maximum ESBR_F value within the IntSBR_F.	dB

The directional linear frequency distance (DLFD) is used as the LFCC counterpart of the (MFCC-based) DPD. The DLFD at the time index n can be defined as:

$$DLFD_n = \left[\sqrt{\sum_{i=1}^{N} [LfccF(i) - LfccR(i)]^2} \right]_{t=n} \quad (7)$$

where LfccF(i) and LfccR(i) are the ith LFCC elements for the forward and reverse signals at time index n, respectively; and N=7 is the number of the LFCC elements used.

Similar to the calculations in (2), (3), (5) and (6), four more LFCC-based parameters were derived from the DLFD, as the LFCC counterparts of MDPD, DDPD, RPEC and ARPEC. These five LFCC-based parameters were then used to form an LFCC-based MEE detection unit according to the same rule-based system structure shown in Fig.1.

C. The LSC-based Directional Analysis Approach

Again, the same system structure shown in Fig.1 was used to design an LSC-based MEE detection unit. This time, the MFCCs and all the MFCC-based parameters were replaced by their LSC counterparts in the directional/perceptual evaluations.

The directional spectral distance (DSD) is used here to replace the MFCC-based DPD in (1). The DSD at the time index n can be defined as:

$$DSD_n = \left[\sqrt{\sum_{i=1}^{N1} [|XF(i)| - |XR(i)|]^2} \right]_{t=n} \quad (8)$$

where $|XF(i)|$ and $|XR(i)|$ are the spectral magnitudes for the forward and reverse signals sampled at the ith frequency index, and N1=46 was chosen to match the corresponding frequency bandwidth used in the MFCC and LFCC methods.

Using formulae similar to (2), (3), (5) and (6), four more LSC-based parameters are deduced from the DSD, as the analogues of MDPD, DDPD, RPEC and ARPEC. These five LSC-based parameters were then used in an LSC-based MEE detection unit with the same rule-based system structure shown in Fig.1.

III. Clinical Data and Evaluation Procedures

A. Data Acquisition

The data used in this study were selected from Doppler recordings on two patients during cardiac surgery (a mitral valve replacement and an aortic root replacement). These recordings were made using a modified version of the in-house multi-gate TCD system previously developed [5]. Doppler signals from only one chosen gate were used in this study. A transmitted frequency of 2 MHz and a pulse repetition frequency of 12.5 kHz were chosen during the signal acquisitions. The receive gate width was set to 10 mm and the sample depth was adjusted to give the optimal signal from the middle cerebral arteries of the patients.

B. MEE Verification Methods

The evaluation started with the identification of significant events (SEs). Here, a SE was defined as an event with the ESBR higher than or equal to 7dB. A 40-ms time-domain resolution was used to find the numbers of SEs in data recordings. A "gold standard" based on the manual "case study decision method" [8] was then used to verify MEEs that were part of the SE family.

C. Data and Procedures for theTraining Phase

Four recordings containing 998 seconds of recorded signals in total were used to train the knowledge-based system shown in Fig.1. About 12350 SEs were found in these training data, which included 1756 verified MEEs.

The training data were purposely selected from clinical recordings, to contain sporadic MEEs, MEE groups and artifacts including those caused by diathermy signals.

The training data were used for setting and tuning the thresholds and parameters in all the three analysis methods (i.e., the MFCC-based, the LFCC-based and the LSC-based approaches), until the "optimized" training performances were reached for these three AMEEI systems.

D. Data and Evaluation Procedures for the Testing Phase

The system was evaluated in two ways using data different from those recordings used in the training phase.

First, an evaluation of sporadic MEE identification was performed using three recordings, which had a total signal duration of 827 seconds. These recordings contained 2,835 SEs, 128 sporadic MEEs and many artifacts including those caused by diathermy signals.

Second, an evaluation of the ability to detect closely packed MEEs was carried out, using one recording that was dominated by heavy embolic showers. The recording contained 4,547 SEs within the recording duration of 310 seconds, amongst which 3,189 were verified as MEEs (closely packed MEEs were separated using 7-dB ESBR peaks during the verification).

The results of the two evaluations were then used to obtain the overall evaluation results for all three AMEEI methods being tested.

IV. Evaluation Results and Discussion

A. Evaluation Results for Sporadic MEE Identification

The sensitivity, the specificity and the positive predictive value (PPV) results for the evaluation of sporadic MEE identification are shown in Table II. It can be seen that the MFCC-based AMEEI method outperforms the other two methods in all the categories, although the performances of the other two methods are also reasonably good.

An example of identification of sporadic MEEs using the MFCC-based method is shown in Fig.2 (the display results for the LFCC-based and the LSC-based methods are not provided since they are identical to Fig.2 in this example).

B. Evaluation Results for the Identification of MEEs in Heavy Embolic Showers

The results for this evaluation are shown in Table III. The performance for the MFCC-based method demonstrates an excellent sensitivity, a moderate PPV and a slightly lower specificity. The number of non-embolic events is relatively

 ISBN: 978-1-61208-137-3

low in this 5 minute recording containing heavy embolic showers, compared with the number of MEEs.

As an example for comparison, Fig.3, Fig.4 and Fig.5 show the MFCC-based, the LFCC-based and the LSC-based methods used to detect a group of MEEs within an embolic shower, and yellow indicator lines are used to mark identified MEEs.

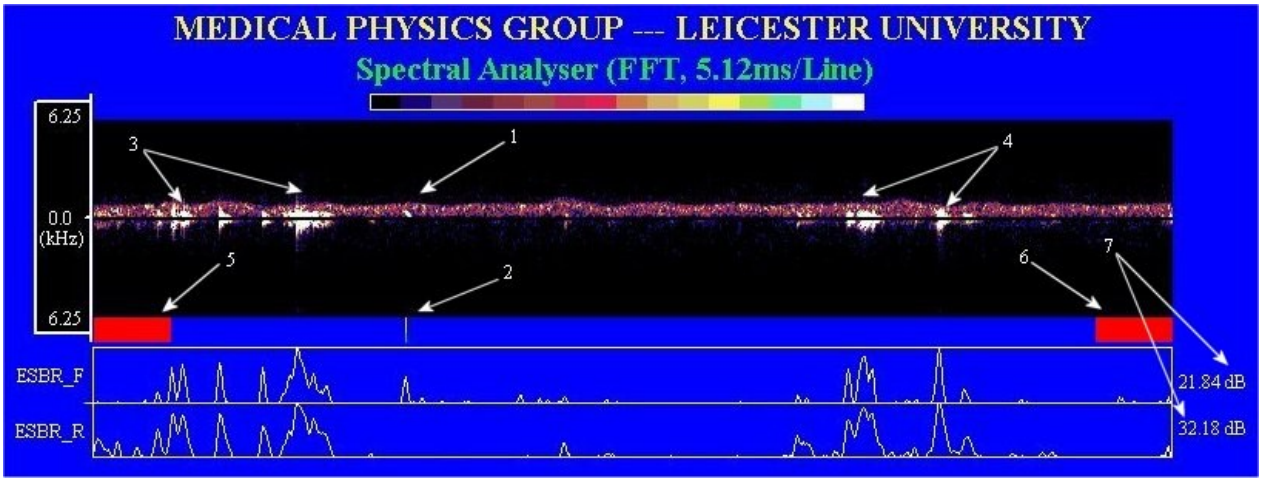


Figure 2. An example of detecting sporadic MEEs using the MFCC-based method. The sonograms for the forward and reverse Doppler signals are displayed on the upper part of the screen, whilst the estimated signal to background ratios (ESBRs) are shown on the lower part of the screen (ESBR_F is for the forward signal and ESBR_R is for the reverse). Arrow 1 shows a sporatic MEE. Arrow 2 demonstrates that the MEE has been detected by the MFCC-based emthod, with a yellow indicator dispayed under the sonogram. Arrows 3 and 4 point to artefacts.Arrows 5 and 6 indicate non-detection regions of the sonograms for the current screen (i.e., any MEE within these regions is to be indentified in the previous or subsequent screen). Arrow 7 points to the display range for the ESBR_F and ESBR_R.

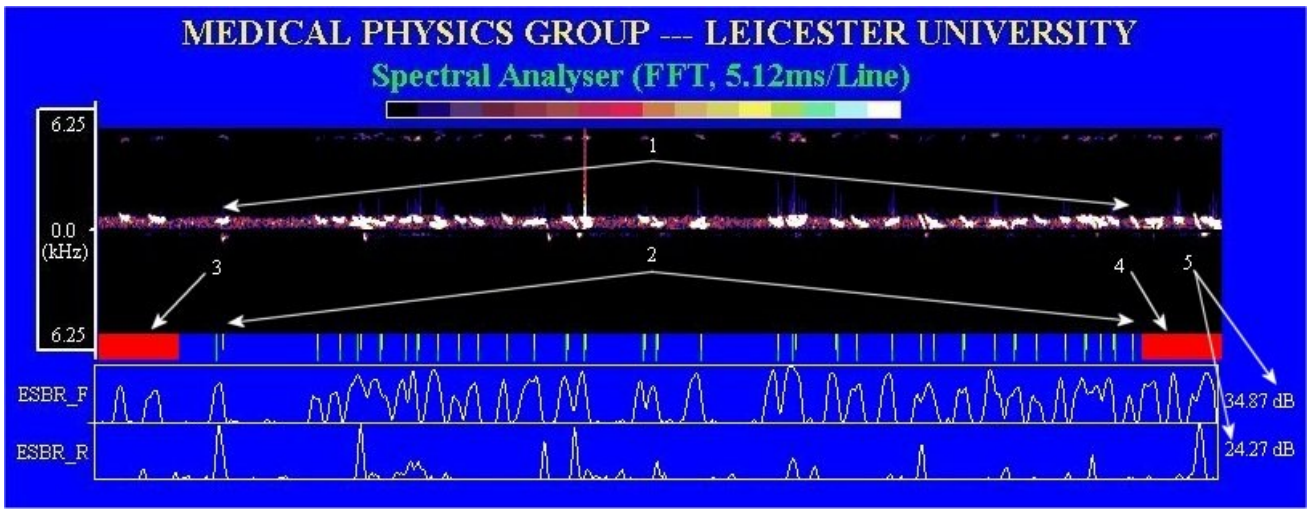


Figure 3. An example of detecting MEEs in the heavy embolic showers using the MFCC-based method. The sonograms for the forward and reverse Doppler signals are displayed on the upper part of the screen, whilst the estimated signal to background ratios (ESBRs) are shown on the lower part of the screen (ESBR_F is for the forward signal and ESBR_R is for the reverse). Arrow 1 shows a group of MEEs within an embolic shower. Arrow 2 demonstrats that the MEEs have been verifiied using the manual "gold standard" (green lines displayed under the sonogram) and also detected by the MFCC-based emthod (with yellow indicator lines dispayed). It can be seen that the automated method detects the EBSR_F peaks more accurately, compared to the manual verifications. Arrows 3 and 4 indicate non-detection regions of the sonograms for the current screen (i.e., any MEE within these regions is to be indentified in the previous or subsequent screen). Arrow 5 points to the display range for the ESBR_F and ESBR_R.

 ISBN: 978-1-61208-137-3

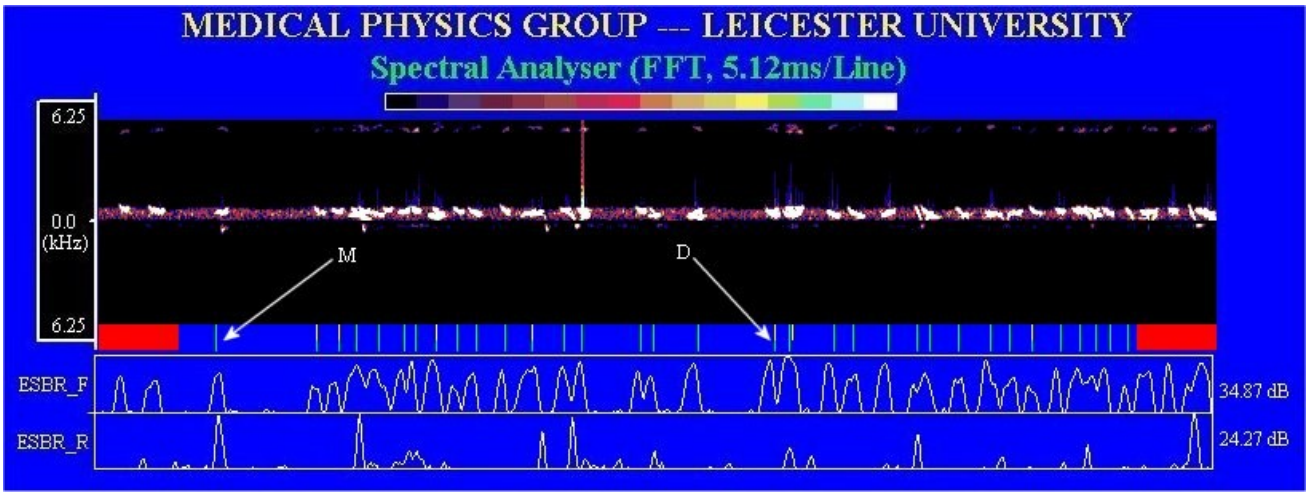


Figure 4. An example of detecting MEEs in the heavy embolic showers using the LFCC-based method. The sonogram and ESBR displays are the same as those in Fig.3, except for the following differences: (a) a green indicator shows that a MEE is verifiied by the manual "gold standard" but missed by the LFCC-based method (as shown in an example indicated by Arrow "M"); (b) a green indicator overlapped by a yellow one demonstrates that a MEE is manually verified and also automatically detected by the LFCC-based method (as shown in an example indicated by Arrow "D").

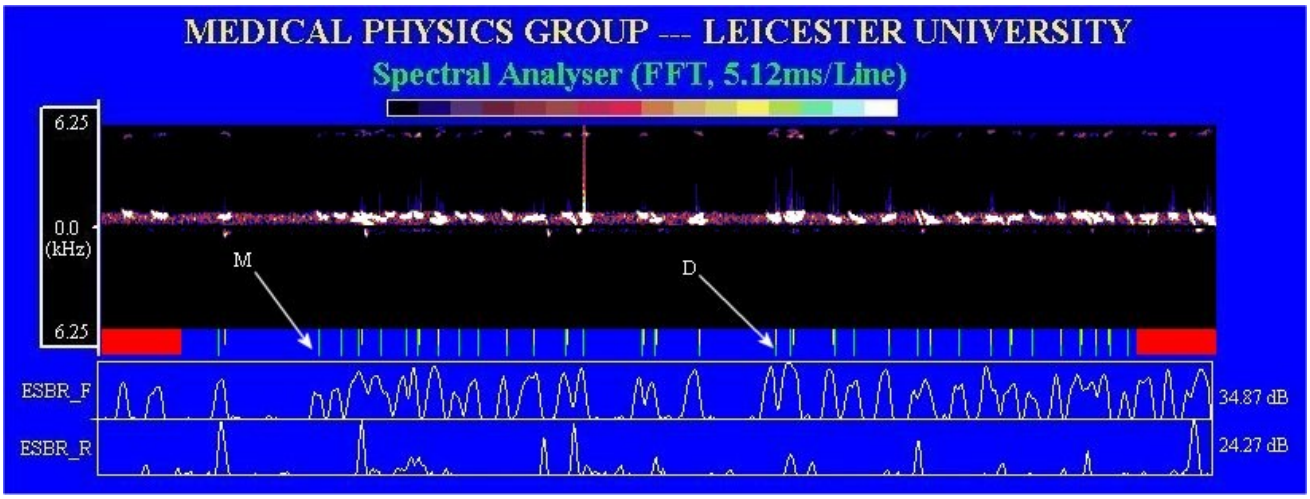


Figure 5. An example of detecting MEEs in the heavy embolic showers using the LSC-based method. The sonogram and ESBR displays are the same as those in Fig.3, except for the following differences: (a) a green indicator shows that a MEE is verifiied by the manual "gold standard" but missed by the LSC-based method (as shown in an example indicated by Arrow "M"); (b) a green indicator overlapped by a yellow one demonstrates that an MEE is manually verified and also automatically detected by the LSC-based method (as shown in an example indicated by Arrow "D").

C. *Overall Evaluation Result*

The overall evaluation results for the automated MEE identification are shown in Table IV. It shows that, with similar specificity and PPV results for all the three methods, the MFCC-based AMEEI system seems to be significantly superior in terms of sensitivity performance.

TABLE II. EVALUATION RESULTS FOR SPORADIC MEE IDENTIFICATION

AMEEI Method	Automated Identification Performances (Data duration: 827s, 2835 SEs & 128 MEEs)		
	Sensitivity	Specificity	PPV
MFCC-based	97.67%	99.93%	98.43%
LFCC-based	88.28%	99.89%	97.41%
LSC-based	96.09%	99.08%	83.11%

 ISBN: 978-1-61208-137-3

TABLE III. EVALUATION RESULTS FOR THE IDENTIDICATION OF MEES IN HEAVY EMBOLIC SHOWERS

AMEEI Method	Automated Identification Performances (Data duration: 310s, 4547 SEs & 3189 MEEs)		
	Sensitivity	Specificity	PPV
MFCC-based	95.92%	89.47%	95.53%
LFCC-based	38.29%	95.58%	95.32%
LSC-based	64.85%	95.14%	96.91%

TABLE IV. OVERALL EVALUATION RESULTS

AMEEI Method	Automated Identification Performances (Data duration: 1137s, 7382 SEs & 3317 MEEs)		
	Sensitivity	Specificity	PPV
MFCC-based	95.99%	96.43%	95.64%
LFCC-based	40.22%	98.45%	95.49%
LSC-based	66.05%	97.76%	96.01%

D. Discussion

It seems from the results that all the three methods performed reasonably well while detecting sporadic MEEs. It should be noticed that the MFCC-based method is superior to other two methods, with a higher sensitivity as well as a higher specificity.

Furthermore, the MFCC-based method coped quite well with heavy embolic showers, whilst the other two methods suffered heavy sensitivity losses. Our in-depth investigations revealed that the LFCC and the LSC parameters were not sensitive enough to follow the rapid directional signal changes due to heavy embolic showers, which could be the cause for the significant numbers of MEEs missed by these two detection approaches.

The results also demonstrated that even the simpler conventional spectral based LSC method outperformed the LFCC-based approach in general.

V. CONCLUSIONS AND FUTURE WORK

A novel MFCC-based embolic signal analysis has been proposed to explore the potentials of auditory perception and energy features of Doppler signals in automated embolus identification. Initial off-line evaluations were carried out using Doppler signals recorded during cardiac surgery and the results show the proposed method could play an important role in a high-performance automated MEE identification system. Compared with the LFCC-based and the traditional LSC-based approaches, the MFCC-based method seems to have a superior performance in both cases of sporadic MEEs and with heavy embolic showers.

Further clinical evaluations with a larger data set will be necessary in future studies.

All the approaches at this stage were applied based on an assumption that an artery had already been located by an operator (i.e., sonograms had been established and displayed). Future studies could be carried out to include the searching phase while the operator is locating the artery. Since unidirectional artifacts could be occasionally generated (e.g., due to a finger touching on the probe) in this rather special detection phase, additional signal analysis measures may be needed to cope with these unwanted events.

ACKNOWLEDGMENT

The authors sincerely thank Emma Chung and Rachel Sumner for their help collecting the clinical data during cardiac surgery.

REFERENCES

[1] E.V. Van Zuilen, J. Van Gijn and R.G.A. Ackerstaff, “The clinical relevance of cerebral microemboli detection by transcranial Doppler ultrasound,” J. Neuroimaging, vol. 8, pp. 32-37, Jan. 1998.

[2] D.H. Evans, “Detection of microemboli,” in Transcraninal Doppler ultrasonography, VL Babikian and LR Wechsler, Eds. Boston: Butterworth Heinemann, 1999, pp. 141-155.

[3] R. Dittrich and E.B. Ringelstein, “Occurrence and clinical impact of microembolic signals during or after cardiosurgical procedures,” Stroke, vol. 39, pp. 503-511, Feb. 2008.

[4] H. Markus, M. Cullinane and G. Reid, “Improved automated detection of embolic signals using a novel frequency filtering approach,” Stroke, vol. 30, pp. 1610-1615, August 1999.

[5] L. Fan, E. Boni, P. Tortoli and D.H. Evans, “Multigate transcranial Doppler ultrasound system with real-time embolic signal identification and archival,” IEEE Trans Ultrason Ferroelectr Freq Control. Vol. 53, pp. 1853-1861, Oct. 2006.

[6] D. Kouamé, M. Biard, J.M. Girault and A. Bleuzen, “Adaptive AR and neurofuzzy approaches: access to cerebral particle signatures,” IEEE Trans Inf Technol Biomed., vol. 10, pp. 559-566, July 2006.

[7] H.S. Ng, Q. Hao, T. Leung, K.S. Lawrence Wong, H. Nygaard, J.M. Hasenkam and P. Johansen, “Embolic Doppler ultrasound signal detection using continuous wavelet transform to detect multiple vascular emboli,” J Neuroimaging, vol. 18, pp. 388-95, Oct. 2008.

[8] L. Fan, D.H. Evans and A.R. Naylor, “Automated embolus identification using a rule-based expert system,” Ultrasound Med Biol. Vol. 27, pp. 1065-1077, August 2001.

[9] Dittrich R, M.A. Ritter, M. Kaps, M. Siebler, K. Lees, V. Larrue, D.G. Nabavi, E.B. Ringelstein, H.S. Markus and D.W. Droste, “The use of embolic signal detection in multicenter trials to evaluate antiplatelet efficacy: signal analysis and quality control mechanisms in the CARESS (Clopidogrel and Aspirin for Reduction of Emboli in Symptomatic carotid Stenosis) trial,” Stroke, vol. 37, pp. 1065-1069, April 2006.

[10] EB Ringelstein, DW Droste, VL Babikian, DH Evans, DG Grosset, M Kaps, HS Markus, D Russell and M Siebler, “Consensus on microembolus detection by TCD. International Consensus Group on Microembolus Detection,” Stroke, vol. 29, pp. 725-729, March 1998.

[11] M Saqqur, N Dean, M Schebel, MD Hill, A Salam, A Shuaib and AM Demchuk, “Improved detection of microbubble signals using power M-mode Doppler,” Stroke, vol. 35, pp. e14-17, Jan. 2004.

[12] S.B. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” IEEE Trans Acoustics, Speech and Signal Processing. Vol. ASSP-28, pp. 357-366, August 1980.

 ISBN: 978-1-61208-137-3

SimGenex: A System for Concisely Specifying Simulation of Biological Processes and Experimentation

Anyela Camargo
School of Computing Sciences
University of East Anglia
Norwich, Norfolk, UK
Email: a.camargo-rodriguez@uea.ac.uk

Jan T. Kim
School of Computing Sciences
University of East Anglia
Norwich, Norfolk, UK
Email: j.kim@uea.ac.uk

***Abstract*—Computational models enable advances in understanding essential features of living systems. Such models can be used to simulate data that can also be measured empirically. Generating such simulated data is frequently a key step in developing and validating models. However, precisely specifying a complex procedure of simulating data is notoriously difficult. The SimGenex language reported here is designed to simplify this task as it is applicable in research scenarios where several candidate models are considered, the mathematical details of regulatory interactions are only known partially or described semiquantitatively, the majority of kinetic parameters are not empirically measured, and a gene expression matrix is available as a basis of identifying the best model. SimGenex enables succinct and flexible descriptions of simulating the biological processes and experimental procedures that are the building blocks of most current wet lab experimental protocols. It enables specification of reproducibly executable workflows for validating computational models of biological systems, it facilitates pre-processing and transformation of data as it is frequently applied in gene expression data analysis and it provides support for comparing and discriminating alternative candidate models based on their ability to approximate the empirical dataset. The result of applying a SimGenex program to a computational model is a simulated dataset that can directly be compared to empirically measured *omic* data through the specification of a distance measure which can be used to discriminate the best model among a number of candidates.**

***Keywords*-Gene Regulatory Networks; Simulation; Systems Biology;**

I. INTRODUCTION

Gene expression measurements determine the amount of product of one or more genes in a biological sample. The amount or concentration of a gene product is called the *expression level* of the gene that encodes the product.

Samples for gene expression measurement are typically cultivated at controlled conditions. While the specific conditions depend on the object of research and the research question, the properties that are subject to control can generally be classified into *genetic properties* and *environmental conditions*. Genetic properties pertain to the genetic makeup of the subjects. Specifically, genes may be *knocked out* (loss of function mutations), or *overexpressed* (gain of function mutations). Typical environmental conditions applied in lab experimentation include treatment with agents such as hormones or drugs, variations in temperature, pH or salinity, and differences in supply of energy or nutrition.

Such signals are effectors impinging on cellular activities and on expression of some genes, or they result in the activation of such effectors. The affected genes frequently encode transcription factors which in turn alter expression of further genes. Perception of environmental signals can thus ripple through a cell's gene regulatory network (GRN), and ultimately change expression levels of many genes.

GRNs enable cells to react to environmental conditions in a genetically determined way. GRNs generate complex dynamics and patterns and attract much scientific interest, particularly since drug development and genetic engineering often involve targeted modification of GRN dynamics. GRN models offer a comprehensive understanding of disease progression, and they can help to predict clinical responses and can be vital to streamline efforts to identify the most promising candidates as early as possible in the drug development pipeline [1]. Therefore, computational GRN models are often used to predict GRN dynamics and to investigate the principles of GRN organisation.

The collection of expression levels of all genes in all samples is called an *expression set*, or, in recognition of the "genes $\times$ conditions" format of the set, an *expression matrix* $X = (x_{gc})$, where g indexes genes and c indexes conditions. The set of expression levels of a given gene g, measured in different samples, is called the *expression profile* (or profile, for short) of that gene, denoted by $\mathbf{x}_g$.

Gene expression measurements are obtained by wet lab methods such as rtPCR or microarrays. The readouts from these techniques are subjected to mathematical operations (e.g. for background correction, normalisation) to obtain estimates of gene expression levels. Gene expression data frequently contain artifacts that require some form of pre-processing (e.g. if a data set contains few negative expression values, this problem can be solved by adding a small offset to all values). After pre-processing, expression data typically is transformed into log-ratios [6], [2] by designat-

 ISBN: 978-1-61208-137-3

ing a reference condition c^* of expression levels (typically corresponding to the unperturbed wild type sample), and calculating $\log(x_{gc}/x_{gc^*})$ for all genes g and all columns c.

Gene expression profiles can be compared by choosing a *profile distance measure* d that quantifies how similar two profiles are. The distance measures currently supported are the Euclidean distance and the correlation distance [8]. The semi-metric correlation distance is defined as $1-r(\mathbf{x}_g, \mathbf{x}_{g'})$, where $r(\mathbf{x}_g, \mathbf{x}_{g'})$ denotes the sample correlation coefficient between the expression profiles of genes g and g', and captures the similarity of "shape" of the profiles being compared [3].

A *simulated expression matrix* can be constructed by applying *in silico* operations to a suitable computational GRN model. Each individual operation reflects biological process or an experimental procedure. By comparing the results of such a simulation to empirical observations, GRN models can be systematically validated. Specifically, a simulated expression matrix Y can be compared to *target matrix* X of empirical gene expression levels by computing $\sum_g d(\mathbf{x}_g, \mathbf{y}_g)$. This *matrix distance* gives an indication of how well the GRN model captures the gene regulatory dynamics of the system from which the empirical measurements were taken, and it can be used to discriminate alternative computational GRN models.

Computational biology often requires reproducible performance of complex workflows to try to simulate the biological processes and experimental procedures that are the building blocks of most current wet lab experimental protocols. Performing *in silico* operations on GRN models typically requires programming in a general purpose language. For data analysis purposes, tools such as Taverna [7], designed to automatise bioinformatics analyses and EXACT [9], designed to represent biological laboratory protocols, have recently been developed. The SimGenex language defines a set of primary operations that are sufficiently general to simulate most standard experimental procedures. This is done within the transsys framework for GRN modelling [5], [4]. SimGenex specifications of operations that model experimentation are declarative and much shorter and simpler than equivalent simulations coded in a computer programming language. SimGenex also provides facilities for specifying mathematical transformations of the primary simulated expression values, and for specifying a distance measure for comparing matrices.

II. SimGenex Feature Overview

The core of a SimGenex program describes how to use a transsys GRN model to produce a simulated gene expression matrix. The `measurementmatrix` block describes how to transform the primary simulated matrix into a *measurement matrix* by e.g. computing log-ratios. Finally, the `discriminationsettings` block configures computation of the distance of the measurement matrix to a target matrix.

A. Simulating Gene Expression

The empirical data in the *target matrix* are normally produced by wet lab means such as rtPCR or microarray. It follows that a number of genotypes are exposed to a number of environmental conditions. In the simulated scenario, transsys GRN models represent genotypes. These are subjected to simulated conditions to produce simulated gene expression values that match the empirical scenario.

The columns of a matrix simulated by SimGenex are generated by creating an initial state and applying a sequence of primary simulation instructions to that state. The primary instructions provided by SimGenex are:

- `runtimesteps` to run a specified number of time steps,
- `knockout` to remove the specified gene from the transsys GRN model,
- `treatment` to set the expression level of a factor to a specified value,
- `overexpress` to insert a new, constitutively expressed gene into the GRN model,
- `setproduct` to alter the product encoded by a gene.

Instruction sequences that are used repeatedly can be declared as a `procedure`. Procedures may in turn invoke other procedures Thus, procedures can straightforwardly be reduced to sequences of primary instructions.

Columns in the simulated matrix are specified by `simexpression` declarations. Like procedures, simexpressions may be composed of primary instructions and procedure invocations. In addition, they also may contain `foreach` instructions. Such simexpressions define multiple columns in the simulated matrix. The `foreach` instruction enables very compact specifications of setups (e.g. when a number of strains are subjected to the same set of experimental conditions). For example, the declaration

```
simexpression s
{
  foreach: wildtype komutant;
  equilibration;
  foreach mock real;
  onehour;
}
```

specifies four columns in which the genotypes `wildtype` and `komutant` are subjected to `mock` and the `real` treatment. The procedures `komutant`, `mock` and `real` have to be defined in order for the above code fragment to work.

B. Computing the Simulated Matrix

In line with the wet lab scenario, the columns of a matrix simulated by SimGenex need to be transformed following the same protocols that were applied to compute the target

 ISBN: 978-1-61208-137-3

matrix of empirical data. SimGenex uses the following blocks within the `measurementmatrix` section to specify such procedures:

- `measurementprocess`: specifies an `offset` parameter to normalise individual gene expression values and a `transformation` equation to indicate how expression values are transformed to simulate a column in a gene expression matrix, e.g. by a log-ratio transform.
- `measurementcolumns`: specifies the columns in the simulated expression matrix. Columns are computed by subjecting the expression levels in one or more simexpressions to mathematical operations, resulting in a column containing one value for each mapped factor of the candidate program. The idea is that the mathematical operations should be the same as those applied to the raw empirical data that have resulted in the empirical expression matrix (e.g. log-ratios where the ratio of a treatment to a control is calculated).

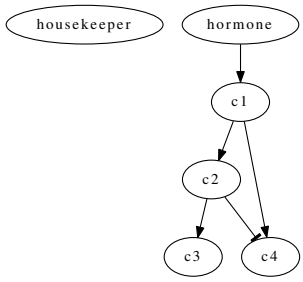
C. Discrimination Settings and Gene mapping

SimGenex allows the specification of a `distance` measure to compare the simulated matrix to a target matrix which can e.g. be used to discriminate the best GRN model from among a number of candidates. In addition SimGenex allows the specification of a mapping scheme, `genemapping`, whereby names of genes in the computational model can be mapped to names in the target matrix. These may e.g. be IDs designated by the microarray provider.

Beyond configuring matrix distance, the `discriminationsettings` section may provide further configuration to be used in the process of discriminating GRN models. Currently, SimGenex supports a `whitelist` of factors or genes which a discriminator may adjust. This feature is useful where parts of the GRN model are unknown, and the discriminator should therefore explore various alternatives for the unknown parts. As an example, where numerical parameters are unknown, these can be set by numerical optimisation.

III. Results

We demonstrate the use of SimGenex on the very simple regulatory network shown in Fig. 1, which is comprised of a constitutively expressed housekeeping gene and a cascade of four genes encoding factors `c1`, `c2`, `c3` and `c4`. The code of the SimGenex protocol to simulate measurements for the wildtype and all single gene knockout mutants is partially shown in Fig. 2. The complete code is posted on the transsys website [4]. Note that specifications for all columns in the `measurementcolumns` block and for all factors in the genemapping were not included. In line with standard practice, we use the expression levels in the wild type as the reference.




```
gene c1gene
{
  promoter
  {
    hormone: activate(0.01, 1.0);
  }
  product
  {
    default: c1;
  }
}

gene c2gene
{
  promoter
  {
    c1: activate(0.01, 1.0);
  }
  product
  {
    default: c2;
  }
}

gene c4gene
{
  promoter
  {
    constitutive: 0.1;
    c1: activate(0.01, 1.0);
    c2: repress(0.01, 1.0);
  }
  product
  {
    default: c4;
  }
}
```

Figure 1. Example network for demonstrating simulation of gene expression measurements with SimGenex. The graph shown at the top shows the overall network topology. The code below shows a part of the transsys model. The `housekeeper` is constitutively expressed and not subject to any regulation. The `hormone` activates `c1`, which is not expressed in the absence of `hormone`. Likewise, `c2` is not expressed in the absence of `c1`. In contrast to this, `c3` is expressed without `c1`, but `c2` increases its rate of expression and `c4` is activated by `c1` and repressed by `c4`.

 ISBN: 978-1-61208-137-3

```
procedure hormtreat
{ treatment: hormone = 1.0; }
procedure equilibration
{ runtimesteps: 100; }
procedure ko_c1
{ knockout: c1gene; }

simexpression all
{
  foreach: wt ko_hk
           ko_c1 ko_c2 ko_c3 ko_c4;
  equilibration;
  foreach: notreat hormtreat;
  treatmenttime;
}

measurementmatrix
{
  measurementprocess
  {
    offset: 0.1;
    transformation:
       log2(offset(x1)) - log2(offset(x2));
  }
}

measurementcolumns
{
  wt_notreat: x1 = all_wt_notreat,
    x2 = all_wt_notreat;
  kohk_notreat: x1 = all_ko_hk_notreat,
    x2 = all_wt_notreat;
  koc1_notreat: x1 = all_ko_c1_notreat,
    x2 = all_wt_notreat;
  wt_hormtreat: x1 = all_wt_hormtreat,
    x2 = all_wt_hormtreat;
  kohk_hormtreat: x1 = all_ko_hk_hormtreat,
    x2 = all_wt_hormtreat;
  koc1_hormtreat: x1 = all_ko_c1_hormtreat,
    x2 = all_wt_hormtreat;
}
}

discriminationsettings
{
  genemapping
  { factor housekeeper = "housekeeper";
     factor c1 = "c1";  }
  distance: correlation;
  whitelistdefs
  { factor: housekeeper c1 c2 c3 c4;
     gene: hkgene c1gene c2gene c3gene
       c4gene;}
}
```

Figure 2. Partial SimGenex code to simulate measurements for the wildtype and all single gene knockout mutants.

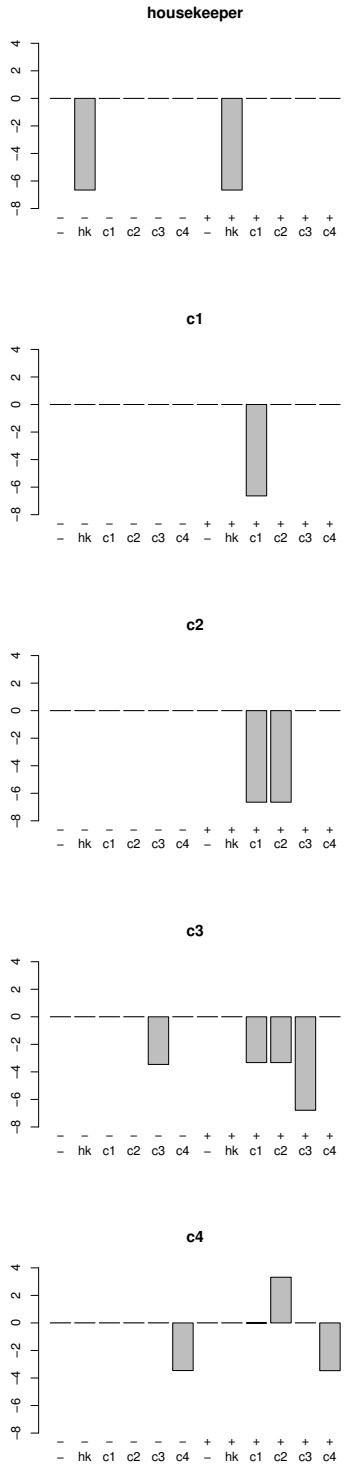


Figure 3. Simulated log-ratio gene expression profiles using the expression levels in the untreated wild type as the reference. The first five bars show expression measurement without, the second five bars show expression measurements with hormone treatment. For non-treated and treated samples, the wild type knockout mutants for all four genes are included.

Fig. 3 shows the simulated gene expression measurements. The two negative bars in the profile of `housekeeper` reflect the fact that this gene product's expression is abolished when the housekeeping gene is knocked out. Otherwise, the housekeeper's expression does not respond to any of the simulated conditions. Without `hormone` treatment, `c1gene` is not expressed, and as a consequence, `c2gene` is not expressed either (see Fig. 1). Therefore, knocking out these genes does not cause any changes in gene expression when no hormone treatment is applied. However, with hormone treatment, the genes in the cascade are expressed and as a consequence, knockouts have detectable effects on the downstream genes in the cascade.

IV. Conclusion

Computational biology often requires reproducible performance of complex workflows. For data analysis purposes, tools such as Taverna [7] and EXACT [9] have recently been developed. SimGenex complements these tools by enabling reproducible specification of simulations of biological processes and experimental procedures. The current main use of SimGenex is generating simulated matrices of expression values. Further, it facilitates specification of a distance measure to compare the simulated matrix to a target matrix comprised of gene expression data externally provided by wet lab means and provides support for discriminating the best gene regulatory network model from among a number of candidates. SimGenex is based on a small and generic set of operations that can be supported by many computational systems biology simulators, and provides new opportunities for unified description and comparison of computational models of living systems.

In our experience, most of the time required to execute a SimGenex program is typically used for simulation of gene expression dynamics. Therefore, significant speed-up can be achieved by re-using intermediate results where the instruction sequences of multiple columns in the simulated matrix share common prefixes, and the underlying GRN model is deterministic. In the near future we plan to optimise the SimGenex implementation accordingly. We also plan to set up a web service for accessing SimGenex and for assessing computational GRN models based on empirical target gene expression data.

Acknowledgement

This work was supported by the Biotechnology and Biological Sciences Research Council, grant number BB/F009437/1.

References

[1] D. K. Arrell and A Terzic. Network systems biology for drug discovery. *Nature*, 88, 2010.

[2] Patrik D'haeseleer. How does gene expression clustering work? *Nature Biotechnology*, 23:1499–1501, 2005.

[3] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.

[4] Jan T. Kim *et.al.* The `transsys` home page, 2001-2010. `http://www.transsys.net/`.

[5] Jan T. Kim. `transsys`: A generic formalism for modelling regulatory networks in morphogenesis. In Jozef Kelemen and Petr Sosík, editors, *Advances in Artificial Life (ECAL 2001)*, volume 2159 of *Lecture Notes in Artificial Intelligence*, pages 242–251, Berlin Heidelberg, 2001. Springer Verlag.

[6] MAQC Consortium. The microarray quality control project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24:1151–1161, 2006.

[7] Tom Oinn, Matthew Addis, Justin Ferris, Darren Marvin, Martin Senger, Mark Greenwood, Tim Carver, Kevin Glover, Matthew R. Pocock, Anil Wipat, and Peter Li. Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17):3045–3054, 2004.

[8] John Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–426, 2001.

[9] Larisa Soldatova, Wayne Aubrey, Ross D. King, and Amanda Clare. The EXACT description of biomedical protocols. *Bioinformatics*, 24:i295–i303, 2008.

 ISBN: 978-1-61208-137-3

Automated Image Processing for the Analysis of DNA Repair Dynamics

Thorsten Rieß*, Christian Dietz*†, Martin Tomas†‡, Elisa Ferrando-May‡ and Dorit Merhof*§
*Interdisciplinary Center for Interactive Data Analysis, Modelling and Visual Exploration (INCIDE)
†Department of Physics, Center for Applied Photonics (CAP)
‡Department of Biology, Bioimaging Center (BIC)
§Visual Computing

University of Konstanz, 78457 Konstanz, Germany

E-Mail: {thorsten.riess, christian.dietz, martin.tomas, elisa.may, dorit.merhof}@uni-konstanz.de

***Abstract*—The efficient repair of cellular DNA is essential for the maintenance and inheritance of genomic information. In order to cope with the high frequency of spontaneous and induced DNA damage, a multitude of repair mechanisms have evolved. These are enabled by a wide range of protein factors specifically recognizing different types of lesions and finally restoring the normal DNA sequence. This work focuses on the repair factor XPC (xeroderma pigmentosum complementation group C), which identifies bulky DNA lesions and initiates their removal via the nucleotide excision repair pathway. The binding of XPC to damaged DNA can be visualized in living cells by following the accumulation of a fluorescent XPC fusion at lesions induced by laser microirradiation in a fluorescence microscope.**

In this work, an automated image processing pipeline is presented which allows to identify and quantify the accumulation reaction without any user interaction. The image processing pipeline comprises a preprocessing stage where the image stack data is filtered and the nucleus of interest is segmented. Afterwards, the images are registered to each other in order to account for movements of the cell, and then a bounding box enclosing the XPC-specific signal is automatically determined. Finally, the time-dependent relocation of XPC is evaluated by analyzing the intensity change within this box.

Comparison of the automated processing results with the manual evaluation yields qualitatively similar results. However, the automated analysis provides more accurate, reproducible results with smaller standard errors.

The image processing pipeline presented in this work allows for an efficient analysis of large amounts of experimental data with no user interaction required.

***Index Terms*—automated intensity measurement; DNA repair; fluorescence microscopy;**

I. Introduction

Damage to the DNA of cells occurs either due to environmental factors *(exogenous damage)* or due to natural metabolic processes *(endogenous damage)*. Exogenous damage may be caused by exposure to UV light or other types of radiation including γ-rays, or toxins and chemicals. Endogenous damage is mostly caused by reactive oxygen species produced from normal metabolic byproducts, and also includes replication errors during mitosis.

DNA damage occurs at a rate of 1,000 to 1,000,000 molecular lesions per cell per day [1], affecting only 0.000165 % of the human genome's approximately 6 billion bases (3 billion base pairs). However, unrepaired damage in critical genes (such as tumor suppressor genes) may interfere with normal cell physiology and significantly increase the likelihood of tumor formation.

Nucleotide excision repair (NER) [2] is a fundamental protective system that promotes genome stability by eliminating a wide range of DNA lesions. Transcription-coupled repair (TCR), which takes place when the transcription machinery is blocked by obstructing lesions in the transcribed strand [3], and global genome repair (GGR) are the two alternative mechanisms of the NER pathway.

The xeroderma pigmentosum group C (XPC) protein is an important protein involved in the GGR pathway. It recognizes DNA damage and initiates the DNA excision repair of helix-distorting base lesions. In healthy cells, the damage is excised by endonucleases, the missing sequence is replaced by DNA polymerase, and a ligase completes the reaction.

An important research question in the field of DNA repair concerns the mechanisms by which the sensor-like protein XPC actually finds base lesions among a large excess of native DNA in a typical mammalian genome [4], [5]. One approach to investigate how XPC searches for aberrant sites within the DNA consists in the visualization of the time-dependent relocation of fluorescently labeled XPC to sites of DNA damage induced at high spatial resolution by irradiation with a femtosecond laser [6]. For this purpose, XPC was marked with green fluorescent protein (GFP), which allows to investigate the damage-dependent recruitment of the fusion product XPC-GFP by confocal microscopy.

In the analysis pursued in [6] the accumulation of XPC-GFP at the induced lesions was quantified by manually defining a bounding box enclosing the lesion and measurement of the intensity change due to accumulation of XPC-GFP in this box over time.

From an image processing point of view, such a manual analysis is unsatisfactory and error-prone for several reasons: First of all, a manual analysis of a significant amount of mammalian cells is a tedious and time consuming task for the investigator. Furthermore, the results obtained from a manual analysis are not rater-independent and lack robustness and

 ISBN: 978-1-61208-137-3

reproducibility.

The challenges for implementing an automated image processing pipeline are the low resolution of the image stacks, movements of the cells over time, other obscuring cells and structures, and the low contrast between the DNA damage and the surrounding nucleus. An overview of current methods for the analysis of fluorescent microscopy images can be found in [7] and references therein.

In this work, these issues are addressed and an automated image processing approach is proposed. This approach allows for automatically detecting the region of XPC accumulation on image stacks showing the cell nuclei and for evaluating the damage-induced changes of XPC dynamics in the nuclear compartment over time. The approach comprises a preprocessing stage, where the image stack data is filtered and the nucleus under consideration is segmented in each image. The images are then registered to each other in order to account for movements of the cell and a bounding box enclosing the DNA damage is automatically determined. Finally, the time-dependent relocation of XPC is evaluated by analyzing the intensity change due to accumulation of XPC-GFP.

II. Material and Methods

In this section, the image acquisition and the image processing methods for evaluating the time-dependent relocation of XPC-GFP are presented. The microscopy image stacks are acquired in an experimental setup explained in Section II-A. The image stacks are then imported into the software framework KNIME, which is described in Section II-B. The image processing pipeline is implemented in KNIME and consists of a segmentation and registration step, as outlined in Section II-C, and the detection of the region of interest (ROI). A scoring algorithm for this detection is presented in Section II-D. Finally, the pixel intensity measurement within the ROI is explained in Section II-E.

A. Biological model system and image data acquisition

Recruitment of the DNA repair factor XPC to sites of DNA damage was monitored in live cells by confocal microscopy. To this end fluorescent fusions of wildtype (WT) XPC and of various XPC mutants (F733A, F756A, F797A, F799A, N754A, W531A, W542A, W690A, W690S) were expressed in either Simian virus 40-transformed human XP-C fibroblasts or in Chinese Hamster Ovary cells (CHO). Cells were then irradiated at the microscope stage (Zeiss LSM 5 Pascal) using an in-house femtosecond Er-fiber laser focused through a 40x oil immersion objective lens. DNA lesions were induced at $775nm$ by multiphoton absorption [8]. A cell nucleus expressing XPC-GFP was placed in the center of the field of view and imaged prior to and at defined time intervals after femtosecond laser irradiation. XPC-GFP fluorescence was detected using a $488nm$ Ar-laser. The acquired image stacks consist of one pre-irradiation frame, one dark frame recorded while scanning with the fiber laser and 60 or 52 post-irradiation frames for experiments with XP-C cells and CHO cells, respectively. The frames were acquired at time steps of 6–7 seconds and the frame size is either 512×512 pixels or 580×580 pixels. The femtosecond laser was scanned along a vertical line of $10\mu m$ in length [6], [9].

B. Software framework

The software platform KNIME (The Konstanz Information Miner [10]) is an open-source tool for data integration, processing, analysis and exploration. Essentially, KNIME is designed to import, transform and visualize large data sets in a convenient and easy to use way. KNIME workflows consist of interacting nodes, which may each represent an algorithm, a single import routine or a visualization tool. The exchange of data between nodes is accomplished via data tables which are passed from one node to another by node connections. The graphical user interface makes it possible to construct workflows consisting of different nodes and their interconnection via a simple drag-and-drop mechanism. The data flow is visually represented by connections between the nodes, typically starting with a node to import the data, followed by one or more processing nodes and finally one or more output nodes. Recently, KNIME has been extended to provide basic image processing nodes such as image input/output and standard thresholding algorithms.

In this work, KNIME is used as a basis to implement a fully automated system that measures fluorescence and quantifies the accumulation of XPC-GFP. The image processing workflow consists of several custom KNIME nodes that are combined with standard image processing nodes. This concept allows to batch process large amounts of image stacks and automatically save the results. Additionally, due to the modular design of KNIME workflows, it is possible to assess intermediate results at every stage of the processing pipeline. In contrast to other software frameworks (e.g. ImageJ), the concept of KNIME workflows allows to apply the same processing pipeline to an arbitrary number of images and to have full control over the output of the analysis.

C. Segmentation and registration

The image processing steps required to quantify the accumulation kinetics of XPC comprise the segmentation of the nucleus, a registration step, and the identification of the irradiated area (the ROI).

In order to identify the nucleus of interest (NOI) in each frame of the image stack, the image is smoothed using a standard Median filter with radius three, and an Otsu thresholding algorithm [11] is applied. This results in a coarse segmentation of each NOI candidate. Since the region scanned during image acquisition is adjusted such that the NOI is located at the center of the image, the NOI can be identified even if multiple cell nuclei are visible in the image by comparing the center of gravity of the NOI candidates. The center of gravity of the NOI is then used to create a polar image [12], [13], which is convolved with a Gaussian Blur filter with $\sigma = 2$ and a standard Median filter of radius three. Then Otsu thresholding is applied to the filtered polar image, which results in a binary image separating the nucleus from the background. In this

 ISBN: 978-1-61208-137-3

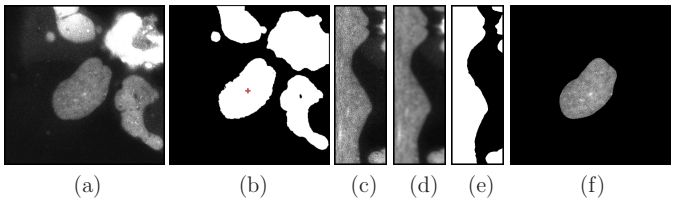


Fig. 1. Segmentation process for an exemplary image. *(a)* Original image with the nucleus of interest (NOI) in the middle and surrounding (undamaged) objects. *(b)* Binary image after a Otsu thresholding, the center of gravity of the NOI is marked. *(c)* Polar image. *(d)* Smoothed polar image. *(e)* Binary image resulting from thresholding the smoothed polar image. *(f)* Masked original image (segmentation result).

binary image, the contour of the nucleus can be detected easily. Finally, the original image is masked using this contour, which concludes the segmentation process. The usage of the polar image improves the segmentation result compared to pure smoothing and thresholding of the cartesian image. To get the same result on the cartesian image only, a much more elaborate pipeline than in the proposed approach would be needed.

Figure 1 illustrates the different steps of the segmentation process and shows the original image including the NOI and surrounding nuclei *(a)*, then the Otsu-thresholded image with the marked center of gravity *(b)* and the polar image *(c)*. Panel *(d)* shows the smoothed polar image, which is again thresholded and the binary image is shown in panel *(e)*. The final result is the masked original image shown in panel *(f)*. The resulting segmentation is very accurate and suitable for further processing steps.

In the next step, registration of the individual frames is required due to potential movement and deformation of the nucleus over time. For this purpose, a rigid body registration algorithm [14] is applied to the cropped images from the previous segmentation step. As a result, the position of the NOI remains the same across all frames, and also the area of induced damage remains stationary. This allows us to determine a single ROI which comprises the area of intensity change due to accumulation of XPC-GFP for all image frames in the stack. In very rare cases the nucleus deforms during image acquisition and this deformation also affects the treated area. The rigid body registration does not compensate for this deformation, hence the ROI that is used in the following step is chosen slightly larger than necessary if the nucleus does not deform, but still gives a sufficient accuracy.

D. Detection of the area of XPC-GFP accumulation

In order to detect the ROI, a time-averaging projection, avg-t-projection for short, of the registered image stack is created in a first step. In this projection, the damaged area is expected to show up as a vertical line of high intensity. In some image stacks, this area can be clearly identified also by a non-expert, in others it is at the limit of visual detection. The ROI is a box of fixed size that exactly covers the irradiated area; for the automatic detection the user can adjust the exact width and height that is used in the algorithm. The box orientation is always such that the long side is vertical, since the irradiated area is a vertical line segment, and the movement/deformation of the nucleus after registration is negligible.

For the automated detection of the correct region, a sum-y-projection is applied to the avg-t-projection image, which sums up the intensity values for each column of the image. This results in an intensity profile of the columns, where the damaged area is expected to show up as a peak. A Median filter with radius three is applied to this intensity profile for smoothing. Due to noise, poor image quality and additional bright spots in the nucleus, this peak is not unique and not straight-forward to identify in some experimental images. For this reason, a combination of three different scoring methods is employed to detect the correct peak. The first scoring method is the height of the peak compared to the neighbouring peaks. The second score is obtained by calculating the response of a "Haar-like" feature [15] centered at the peak and with a fixed width slightly larger than the width of the ROI. This can be interpreted as locally measuring the difference of the area underneath the peak with neighbouring areas. Finally, the third score is the distance of the peak to the center of the nucleus. In an ideal experiment, the irradiation is applied exactly at the center of the NOI, and hence a peak close to the center is more likely to be the correct one. For the automatic evaluation, all three scores are weighted equally and the peak with the highest overall score is chosen.

After the correct peak has been identified, the ROI is chosen such that it is centered at the x-position of the peak and has the predefined width and height. If the height of the nucleus exceeds the height of the ROI, the y-position of the ROI is adjusted such that the intensity in the avg-t-projection is maximized. Examples are shown in Figure 3 below.

E. Intensity measurement

The time-dependent relocation of XPC is evaluated by analyzing the intensity change due to accumulation of XPC-GFP. The measurement is performed on the pre-processed and registered image stack and is accomplished as follows: First, the background is subtracted from the ROI for each image frame in order to improve the results of the intensity measurement. The average pixel intensity of the region of interest I_{ROI}, as well as the average pixel intensity value I_{NOI} of the whole NOI, is then measured at every time-step. The quotient I_{ROI}/I_{NOI} of these values is computed and scaled such that the value of the first image (pre-irradiation) is always equal to one.

In order to validate the presented approach, time-lapse series of a total of 11 different XPC mutants have been analyzed, with each series repeated at least 8 times. The XPC mutants were expressed in two different cell types, human XPC fibroblasts and CHO cells.

The measurement values for each mutant and each cell type were averaged at each time-step and the standard errors were computed, which allows easy comparison to the manual evaluation described in [6], [9].

 ISBN: 978-1-61208-137-3

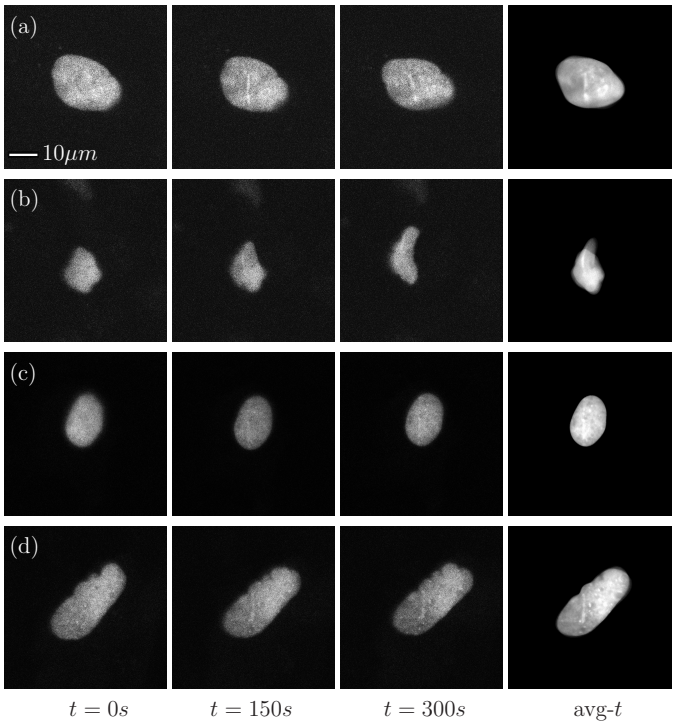


Fig. 2. Microscopy images. Each row shows four images, the first one recorded at the beginning of the experiment (pre-irradiation), the second after 150 seconds and the third after 300 seconds. The fourth image shows the avg-t projection *after* the segmentation and registration algorithm has been applied. The XPC-GFP accumulation within the nucleus in row *(a)* is clearly visible. The nucleus in row *(b)* deforms and moves during image acquisition. Row *(c)* shows a nucleus where the accumulation is hardly detectable. The whole nucleus in row *(d)* is very bright, hence the XPC-GFP accumulation is hardly visible. Note that all microscopy images shown here are contrast enhanced.

III. Results

The results of the automated measurement are presented in the following. In particular, the segmentation performance is evaluated in Section III-A, and the measurement results are compared to the manual evaluation in Section III-B.

A. Segmentation Performance

In order to evaluate the image processing pipeline proposed in this work, approximately 100 image stacks with 52 or 60 frames per stack were processed. The processing pipeline comprises a user interface which is presented to the user at the end of the analysis and allows to review the results. The segmentation results were verified by an expert biologist who rejected image stacks where the first or second image processing step failed.

The first processing step (segmentation of the nuclei) was successful in almost 99 % of all cases. Out of the correctly segmented image stacks, the second processing step (identifying the site of irradiation and defining the ROI) was accurate in 99 % of all cases.

It should be noted that the 100 image stacks included into the analysis also comprise very difficult segmentation scenarios, e.g. cases where the cell moves significantly during image acquisition, or where confounding factors such as air bubbles or other cells (albeit without lesion) are present in the image. A selection of difficult segmentation scenarios where our approach was still successful in most cases is shown in Figure 2. Each row shows sample images that were acquired during the experiment and the avg-t-projection image of the segmented and registered image stack. Row *(a)* shows a very good example where the accumulation is clearly visible and the cell hardly moves. Rows *(b)* and *(c)* are difficult to segment and to measure due to cell movement and deformation of the nucleus *(b)* or hardly detectable accumulation *(c)*. Row *(d)* is another difficult example where the intensity within the nucleus is so high that our algorithm detects a wrong ROI. The ROI detection for all of these image stacks is shown in Figure 3. The top row shows the avg-t-projection and the detected ROI is overlaid. The bottom row shows the corresponding intensity profile, indicating the peak that the scoring algorithm chooses. Note that the ROI detection in image *(d)* failed, the actual accumulation takes place to the left of the detected ROI; this is the only case of all processed image stacks where the ROI detection failed.

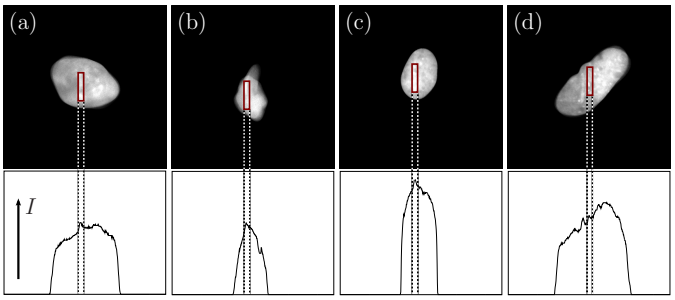


Fig. 3. The detection of the ROI. The top row shows representative images of avg-t-projections, the bottom row shows the associated intensity profile, the intensity is denoted by I. The detected ROI is marked in each avg-t-projection image, as well as the associated peak in the intensity profile. The panels *(a)–(d)* correspond to the images shown in Figure 2 *(a)–(d)*. Note that panel *(d)* shows a negative example: the irradiated area lies to the left of the incorrectly determined ROI.

The processing time per image stack is in the range of 1–2 minutes per stack on a PC with a 2.83 GHz processor and 4GB RAM. In the whole processing pipeline, the registration and preprocessing stages are the most time consuming steps.

B. Manual vs. Automated Analysis

For a more detailed assessment of the results of the automated image processing approach, the automated measurements are compared to the manual evaluation in [6] and [9]. The manual evaluation is achieved similarly to the automated evaluation. First, the image stack is registered, then the ROI is determined manually and the intensity of the ROI in each image of the stack is measured and the quotient I_{ROI}/I_{NOI} is determined. In contrast to the automated evaluation, the ROI size is not fixed, but adjusted to fit the most visible part of the lesion.

Figure 4 *(a)* shows the result of our automatic inten-

 ISBN: 978-1-61208-137-3

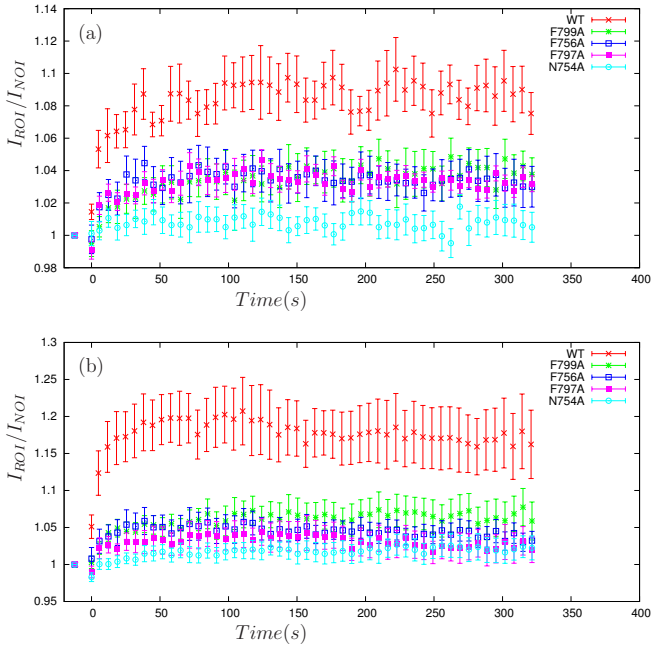


Fig. 4. Results of the automatic measurement *(a)* and the manual measurement *(b)*. Shown are the mean values and their standard errors of the intensity quotient I_{ROI}/I_{NOI} of several image stacks for five different XPC mutants in CHO cells.

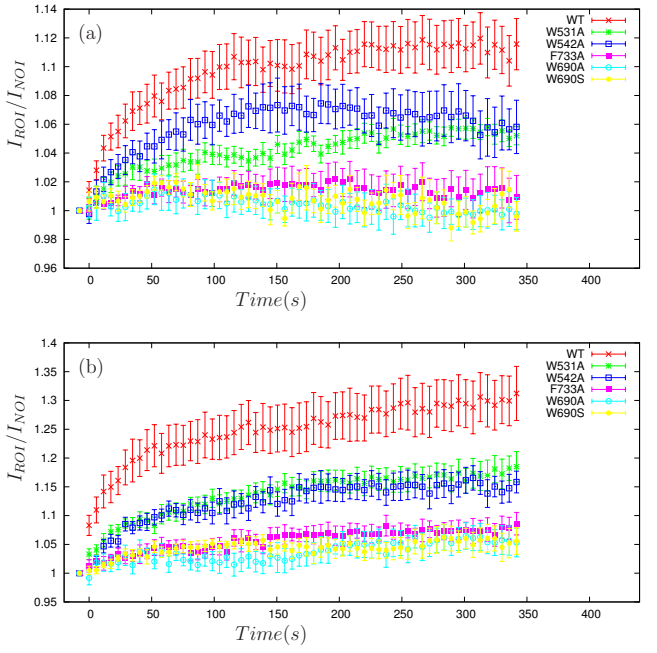


Fig. 5. Results of the automatic measurement *(a)* and the manual measurement *(b)*. Shown are the mean values and their standard errors of the intensity quotient I_{ROI}/I_{NOI} of several image stacks for different XPC mutants expressed in XP-C fibroblasts.

sity measurements for four different XPC mutants (F756A, F797A, F799A, N754A) and wild-type XPC expressed in CHO cells. The quotient I_{ROI}/I_{NOI} is plotted versus the time of acquisition. Shown are the mean values of at least 8 image stacks per mutant, the error bar is given by the standard error of these mean values. The computed results are compared to those obtained by manual evaluation [9] as shown in Figure 4 *(b)*. Note that the graphs show the same qualitative behaviour, but there is a quantitative difference. The intensity increase determined by the automated approach is much lower. However, at the same time there is a reduction of the standard error, hence preserving e.g. the statistically significant difference between the data from XPC-WT and the various XPC mutants (in this case, the statistical significance is simply provided by non-overlapping standard error bars). What is more, the automatically measured intensity changes of the N754A mutant are lower than intensity changes of the F-mutants. Whether the improved ability of the automated analysis to differentiate between the N and the F mutants has a biological correlate is a question beyond the scope of this work and will be investigated in future studies.

Figure 5 *(a)* shows the automatically measured intensity changes for five different XPC mutants (W531A, W542A, F733A, W690A, W690S) and wild-type XPC expressed in XP-C cells. The quotient I_{ROI}/I_{NOI} is plotted versus the time, and again the mean values of several image stacks per XPC mutant are shown. The results of the manual evaluation from [6] are shown in Figure 5 *(b)*; note that the graphs show the same quantitative differences as the graphs in Figure 4. The qualitative features are largely preserved, with the exception that the automatically measured values of W531A and W542A are separated in the time interval starting at $50s$ and ending at $220s$. This separation cannot be seen in the manual evaluation and it is not entirely clear if there are biological reasons for this separation or if it is due to the image quality of the W542A mutants, which is also expressed in the high standard error in the automated analysis result.

Summarizing, the results of the automated measurements preserve the features that were discovered in the manual evaluation, and on top of that have the advantage of reproducibility and unbiasedness, and, last but not least, time saving. Moreover, the measurement seems to be more accurate, the N754A mutant appears clearly separated from the F-mutants in the experiments involving the CHO cells, which is not the case for the manual evaluation.

IV. Discussion

The results of our automated image processing approach are satisfactory for our purposes. Almost 99 % of the cell nuclei were correctly segmented and 99 % of the ROIs were correctly determined in our data set of approximately 100 image stacks. Moreover, the comparison with a manual evaluation of the data set shows that the automated measurement not only supports the qualitative results from the manual evaluation, but also has a lower standard error. In the presented case, the automated evaluation shows that the N754A mutants

 ISBN: 978-1-61208-137-3

behave significantly different from the other evaluated mutants. Whether the improved ability of the automated analysis to differentiate between the N and the F mutants has a biological correlate is a question beyond the scope of this work and will be investigated in future studies.

The quantitative differences of the measured intensity values in the manual and in the automated evaluation are likely due to the choice of the ROI size. In the manual evaluation, the ROI has been chosen smaller for the XPC mutants where the XPC-GFP accumulation is clearly visible, and larger for the XPC mutants where hardly any accumulation is visually detectable. This choice influenced the measured intensity changes. In particular, the values for mutants with high accumulation are much higher than in the automated analysis, where the ROI has a constant size throughout the analysis. Although this automated approach performs very well for our experimental data sets, there is room for improvements regarding the registration algorithm — a nonlinear registration algorithm [16] would be able to compensate deformations of the irradiated area — and computation time.

V. Conclusion

We presented an automated system for measuring the performance of XPC in the DNA repair process based on intensity changes in microscopy images. The image processing pipeline comprises several steps that are based on standard image processing algorithms in combination with a customized segmentation algorithm and a specific scoring method to detect the correct ROI.

Laser microirradiation in combination with fluorescence microscopy has become a popular method for studying the dynamics of DNA repair in live cells. Computational tools that facilitate the extraction of quantitative data from such experiments are therefore of great interest to the biology community. Further work will be directed at recognizing more complex irradiation and damage patterns.

VI. Acknowledgments

The Interdisciplinary Center for Interactive Data Analysis, Modelling and Visual Exploration (INCIDE) is funded via a grant of the German Excellence Initiative by the German Research Foundation (DFG) and the German Council of Science and Humanities awarded to the University of Konstanz. The Center of Applied Photonics (CAP) is supported by the Ministry of Science, Research and the Arts Baden-Württemberg. We thank D. Träutlein and H. Naegeli for providing image data. We gratefully acknowledge A. Leitenstorfer, M. Horn and O. Deussen for fruitful discussions and support.

References

[1] H. Lodish, A. Berk, P. Matsudaira, C. A. Kaiser, M. Krieger, M. P. Scott, L. Zipursky, and J. Darnell, *Molecular Biology of the Cell*. New York: WH Freeman, 2004.

[2] L. C. Gillet and O. D. Schärer, "Molecular mechanisms of mammalian global genome nucleotide excision repair," *Chemical Reviews*, vol. 106, no. 2, pp. 253–276, 2006.

[3] P. C. Hanawalt and G. Spivak, "Transcription-coupled DNA repair: two decades of progress and surprises," *Nature Reviews Molecular Cell Biology*, vol. 9, pp. 958–970, 2008.

[4] O. D. Schärer, "Achieving broad substrate specificity in damage recognition by binding accessible nondamaged DNA," *Molecular Cell*, vol. 28, no. 2, pp. 184–186, 2007.

[5] K. Sugasawa and F. Hanaoka, "Sensing of DNA damage by XPC/Rad4: one protein for many lesions," *Nature Structural & Molecular Biology*, vol. 14, no. 10, pp. 887–888, 2007.

[6] U. Camenisch, D. Träutlein, F. C. Clement, J. Fei, A. Leitenstorfer, E. Ferrando-May, and H. Naegeli, "Two-stage dynamic DNA quality check by xeroderma pigmentosum group C protein," *The EMBO Journal*, vol. 28, pp. 2387–2399, 2009.

[7] Z. Gitai, "New fluorescence microscopy methods for microbiology: sharper, faster, and quantitative," *Current Opinion in Microbiology*, vol. 12, pp. 341–346, 2009.

[8] D. Träutlein, F. Adler, K. Moutzouris, A. Jeromin, A. Leitenstorfer, and E. Ferrando-May, "Highly versatile confocal microscopy system based on a tunable femtosecond Er:fiber source," *Journal of Biophotonics*, vol. 1, no. 1, pp. 53–61, 2008.

[9] F. Clement, N. Kaczmarek, N. Mathieu, M. Tomas, A. Leitenstorfer, E. Ferrando-May, and H. Naegeli, "Dissection of the xeroderma pigmentosum group C protein function by site-directed mutagenesis," *Antioxid Redox Signal*, vol. (to appear), 2010.

[10] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Proc. Data Analysis, Machine Learning and Applications*, pp. 319–326, 2008.

[11] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[12] P. Bamford and B. Lovell, "Unsupervised cell nucleus segmentation with active contours," *Signal Processing*, vol. 71, no. 2, pp. 203–213, 1998.

[13] M. Kvarnström, K. Logg, A. Diez, K. Bodvard, and M. Käll, "Image analysis algorithms for cell contour recognition in budding yeast," *Optics Express*, vol. 16, no. 17, pp. 12943–1257, 2008.

[14] P. Thévenaz, U. Ruttimann, and M. Unser, "A pyramid approach to subpixel registration based on intensity," *IEEE Transactions on Image Processing*, vol. 7, no. 1, pp. 27–41, 1998.

[15] R. Lienhart, E. Kuranov, and V. Pisarevsky, "Empirical analysis of detection cascades of boosted classifiers for rapid object detection," in *In DAGM 25th Pattern Recognition Symposium*, pp. 297–304, 2003.

[16] I.-H. Kim, R. Eils, and K. Rohr, "Non-rigid alignment of multi-channel fluorescence microscopy images of live cells for improved classification of subcellular particle motion," vol. 7262, p. 72620S, SPIE, 2009.

 ISBN: 978-1-61208-137-3

Quantifying the Accuracy of C. elegans Image Analysis

Jacob Graves and Roger Mailler
Department of Computer Science
University of Tulsa
Tulsa, United States
Email: jacob-graves@utulsa.edu, mailler@utulsa.edu

***Abstract*—The nematode *Caenorhabditis elegans* is an important model organism for many areas of biological research including genetics, development, and neurobiology. A common technique used in studying the locomotion of the worm is to take video of the worm in motion and analyze it to extract relevant data. A number of different software solutions exist to analyze these videos, yet there is no technique to determine the accuracy of the statistics being produced. We have developed a method to quantify the accuracy of a given analysis pipeline by using video of a biologically accurate simulation. Using this process we develop a metric to quantify the accuracy of a given pipeline, and we demonstrate this metric by comparing different implementations of a popular pipeline.**

***Keywords*-Image analysis; Feature extraction; Caenorhabditis elegans; Simulation**

I. INTRODUCTION

The nematode *Caenorhabditis elegans* is an important model organism for many areas of biological research including genetics, development, and neurobiology. Extensive research has been performed on video analysis and feature extraction on microscopic video of *C. elegans* locomotion. A number of different software solutions exist to analyze these videos, yet there is no method to determine the accuracy of the statistics being produced.

In this work, a method is developed to validate the accuracy of video analysis pipeline through use of an accurate simulated model of the worm. The ALIVE simulator is a simulation that produces a biologically accurate 3D model of the worm [6]. By using video from this simulator coupled with the data containing the exact underlying position and shape of the worm, we are able to correlate the raw video data with the actual values that describe the worm. We can then take this video and feed it into a given video analysis pipeline and compare its results to the actual values.

To test our method, we took a popular analysis pipeline and varied the thinning algorithm it uses to create skeletons. Then we validated each of these pipelines on three variations of worm movements all produced in the ALIVE simulator. We derived an accepted result for segment angle and segment angle velocities for the video clips, and compared that to the result measured from the pipelines. This exposed a few flaws in the accepted pipelines and allowed us to empirically measure the fitness of an analysis pipeline.

II. BACKGROUND

A. Imaging

A common form of worm analysis involves tracking a single worm at a higher magnification over time to analyze its motion and behavior. One use for this research is in identifying different strains of worm based on their motion patterns over time. Since a complete neurological mapping of the worm exists, studying different mutants of the worm allows for a further understanding of the working of nervous systems. Many mutants cause disruption in the typical sinusoidal motion in *C. elegans*. These range from obvious changes in locomotion that can be spotted through a microscope, to very subtle changes that show themselves through a thorough statistical analysis of the worm's motion [9].

Tracking a single worm is also useful for researchers who are trying to abstract information about a worm's motion in order to artificially reproduce realistic locomotion in a simulator. For systems such as [8], [1], [5], a neurological representation of the locomotion functions of the worm is reproduced that must be verified against the locomotion of a worm *in vitro*. This requires an accurate and thorough set of statistics on worm motion to verify the images against.

Our system focuses on single worm tracking at a relatively high magnification and feature extraction about its forward locomotion. Imaging of *C. elegans* for this purpose is typically performed by a "pipeline" of effects through which a video is processed to produce a "skeleton" of the worm. Data analysis is performed on this skeleton and the prominent features are extracted. A current pipeline, utilized in [4], performs the skeletonization operation by thinning the binarized image of the worm, then discretizing the skeleton into the desired number of points. This skeleton is usually stored in an intermediate data file to be processed for a number of statistics such as amplitude, velocity, segment angle, and segment velocity. These are the four statistics we will compute and compare in this paper. This pipeline will be discussed in detail in Section III.

B. Simulators

Due to the relative simplicity of *C. elegans* and the ready availability of a complete neural mapping, frequent attempts

 ISBN: 978-1-61208-137-3

have been made to develop simulations of the organism. A number of different approaches have been taken in simulating the motion of the worm. Neibur and Erdos produced a simulation in part to show that the muscle activations that cause movement could be reproduced assuming the worm has stretch receptors that allow for the propagation of movement down the worm [7]. Bryden [1] and Karbowski [5] developed simulators that make biologically accurate neural networks and apply them to their simulators to reproduce the range of motion for the organism. These simulations represent the body as a set of uniformly distributed points in two-dimensional space. This prevents them from replicating the proper weight distribution, and more importantly, the non-uniform placement of the muscles that are used to generate locomotive force in the actual worm. They also fail to directly simulate the environment, but instead apply constant frictional forces at these discrete points along the body.

To remedy this, the ALIVE simulator was developed. This simulator uses a biologically accurate three dimensional model of the worm and implements physically accurate interactions of the worm with its environment. The simulator produces photo-realistic images of the worm moving (for the context of image analysis). The simulator will be discussed in more detail in Section III-B.

III. MATERIALS

A. C. elegans *Video Toolkit*

Currently, various software solutions exist to analyze and extract features from videos of the worm *C. elegans*. However, these software packages have not been widely accepted by the community because they are implemented with a single process for converting video into data, documented poorly, tend to be slow, and implemented using MATLAB's outdated, undocumented image processing algorithms.

By developing a standard and easy to use interface for creating and sharing worm analysis pipelines, we hope to foster a convenient and open comparison of existing pipelines. Also, a video analysis pipeline program has the potential to allow researchers without experience in programming to develop customized analysis pipelines. By incorporating multiple different algorithms for performing the same operations, it is easy to compare and contrast the benefits of competing algorithms for image manipulation. A pipeline that provides a way to view the processed image during each effect allows each effect to be more finely tuned, producing better overall results.

The software we have developed rectifies these deficiencies by providing a versatile toolkit of video analysis techniques where the end user can customize the analysis pipeline. Our software, written in Java, not only provides versions of the image analysis algorithms found in current software packages, but also provides alternate and improved techniques to allow the user to mix and match processing steps to accommodate their specific needs. Each effect includes settings and a preview function to allow a researcher to fine tune each effect to their video or microscope, allowing for more accurate and specialized results. Since the software is developed in Java, the program is lightweight and fast, especially compared to other solutions that use MATLAB image processing algorithms.

Our extensive toolkit contains common pre-coded effects used in *C. elegans* feature extraction, such as bend angles, segment length, and segment velocities. The visual interface allows easy connection of effects to create a custom video pipeline, which can then be saved and shared with other CVT users to allow a standard for comparing analysis pipelines. CVT also incorporates modern and legacy image manipulation algorithms. It takes input from video files and folders, allowing for real time or batch processing. The use of color coded "ports" clearly demonstrate the expected inputs and outputs for an effect, allowing easy creation and editing of pipelines.

To facilitate the pipeline creation and management process, our program provides an intuitive drag and drop interface that allows a user to select processing algorithms, set their parameters, and connect them together. Users can then test and refine their pipeline by viewing the video at each stage of processing. The output of the process is a standardized file format that contains the key characteristics of the worm's motion in a format that is compatible with statistical analysis software.

For developer support, we have also incorporated a standard interface for designing and plugging effects into the toolkit. This allows any effect to be inserted into a pipeline, including effects that may output data in a custom format. This flexibility allows any pipeline and data format to be supported with minimal Java coding.

B. *ALIVE Simulator*

Validating a pipeline requires a realistic depiction of the worm in which we know the underlying values of the worms position for each frame. To do this, a biologically accurate model of the worm was required. This was obtained by using the ALIVE simulator developed at the University of Tulsa. This powerful simulator was developed as a biologically accurate simulation of the worm *C. elegans*. This simulator produces realistic locomotion of a three dimensional model of an adult *C. elegans*.

As discussed earlier, other simulations have done a number of significant simplifications in order to model the worm's motion, including depicting the worm as a uniform two-dimensional model. These simplifying assumptions limit the ability of these simulations to accurately depict the non-uniform friction that results from the worm's contact with the world around it and subsequently the complex neural control that is needed to generate the worm's characteristic sinusoidal pattern of locomotion.

 ISBN: 978-1-61208-137-3

Leveraging the tremendous increases in computational power and advances in numeric methods, the ALIVE simulator rectifies these deficiencies by representing a biologically accurate 3D model of the body of *C. elegans* in a virtual environment that mirrors the physical properties of its natural world. This simulator, which has been under development for nearly two years, is built using an open-source 3D game and physics engine. The model accurately depicts the physical properties of the real organism including its nonuniform weight, size, shape, and musculature. In addition, the simulator models the interaction between the worm and its environment to include surface tension, friction, inertia, and gravity. The simulator faithfully reproduces forward and reverse locomotion of *C. elegans* on an agar surface, and the model is cross validated using video recordings of worms that were converted to quantitative data by image analysis software [6].

The worm model itself contains 25 discrete segments of biologically correct length. This is then powered by a sine wave propagated down the worm to simulate forward locomotion. This has been tuned so that the worm moves in a realistically validated way. The frequency and update rate can be easily adjusted to create unusual movement patterns that can make for theoretically realistic movement variants [6].

To further test the various pipelines, we decided to use the power of the ALIVE simulator to create three distinct variants of worm motion to test the versatility of the pipelines. The first worm variant is the typical forward locomotion of the *C. elegans*. The ALIVE simulator by default creates a worm and moves it in the form of a typical adult worm. The simulator does this by calculating muscle activations by propagating a sin wave down the different segments of the body. This creates a typical movement that has been cross validated with video of real worms in forward locomotion. This is the typical configuration of the worm that is analyzed in most video analysis pipelines, and the data that is extracted represents such statistics as amplitude, velocity, bend angles, and bend angle velocities. The second variant moves with very high bend angles and a low bend angle velocity. This makes for a very slow moving worm that appears to make many pirouettes. This covers the pipeline's ability to handle very high bend angle situations. The last variant moves with very low bend angles, but at a very high speed. This worm quickly moves its segment back and forth which allows the worm to quickly move across the plate. This motion is not necessarily represented in the motion of the real worm, but it does put the worm in low angle situations that tests the sensitivity of the pipeline.

These three variants demonstrate a range of the worm's potential motion and test the sensitivity of the pipeline to changes in the worm's style of motion.

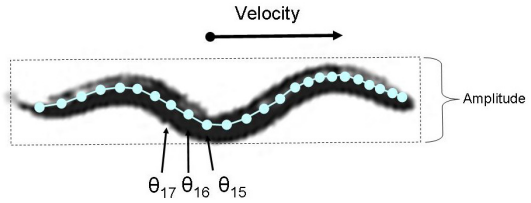


Figure 1. Features of *C. elegans* locomotion

C. WormAnalyzer

To produce the statistical analysis of the skeleton files, we used the WormAnalyzer software that was also developed at The University of Tulsa. This software package takes in skeleton files and batch processes them to produce the relevant statistics. The software extracts 49 features from the worm's motion, four of which are relevant here: amplitude, velocity, segment angle, and segment angle velocity. Amplitude is the measure of the maximum amplitude at any given time. After a file is processed the average amplitude for the worm over the duration of the file is given. Velocity measures the worm's magnitude of velocity (speed) from one frame to the next by tracking the worm's centroid. The average velocity over the file is reported. For the segment angles, one measurement is produced for each joint in the worm's skeleton where two segments meet. This segment angle is a measurement of a single joint's angle at each frame. The average angle over the course of the video is reported. The segment angle velocity is the rate of change of segment angle over time. The average angular velocity over the course of the file is produced. The software outputs these statistics, which can then be further analyzed using statistical packages [6].

IV. METHODS

A. Pipeline Creation

The analysis pipeline we constructed is similar to that described in Geng et. al [4] with some minor differences. To create the pipeline, we used the CVT software described in Section III-A. The first element in the pipeline is an input cell that allows a batch process of videos to be read in and analyzed. Upon each new video a message is sent to all the nodes in the pipeline so that they can initialize in whatever means appropriate. Each frame of the videos are read in and then passed along to the next pipeline processor until the output nodes are reached.

First, each frame is converted from color to grayscale using a simple combination of the 24-bit Red/Green/Blue values for each pixel. The next processor in the pipeline performs a binarization of the video using a local thresholding algorithm. This algorithm uses a sliding 3 X 3 window to determine whether a given pixel represented by a single grayscale value should be converted to black (foreground) or white (background).

 ISBN: 978-1-61208-137-3

Table I
SEGMENT ANGLE ERROR OF DIFFERENT THINNING METHODS

	OPATA-8		SPTA		TPTA	
(in degrees)	PEV	MaxErr	PEV	MaxErr	PEV	MaxErr
Default	1.90	4.36	2.07	7.44	2.03	5.55
Large Ang	1.87	4.62	2.48	9.63	2.14	6.40
Small Ang	1.76	4.01	1.88	6.17	2.07	4.52

Table II
SEGMENT VELOCITY ERROR OF DIFFERENT THINNING METHODS

	OPATA-8		SPTA		TPTA	
(in degrees)	PEV	MaxErr	PEV	MaxErr	PEV	MaxErr
Default	11.28	38.76	15.87	49.99	13.37	46.75
Large Ang	27.82	92.66	32.01	109.48	27.92	95.00
Small Ang	7.79	23.57	13.01	42.18	10.77	33.80

This processor colors the pixel black if the standard deviation of the intensity of the pixel and its surrounding pixels is greater than the mean of the entire image, or if the mean intensity of the pixel and its surrounding pixels was greater than the background pixel intensity.

Next, we remove small objects from the image using a region labeling algorithm that indexes each pixel in the image according to the region to which it belongs. Once all of the black regions are labeled, we remove all of the regions except for the largest. This isolates the worm (which is black) onto a white background. We then used the same method to fill holes in the worm's image by inverting the colors such that only the background region is maintained.

With a binary image of just the background and worm in hand, we apply a thinning algorithm to reduce the worm's image down to a core body that represents the worm's basic shape. A number of options are available to thin the worm. For this paper, we chose to vary the thinning algorithms in order to determine which is the most accurate to use in a *C. elegans* imaging pipeline. This decision allows us to use our newly developed metrics to track a very small difference in a pipeline, demonstrating how this validation method can be used to incrementally choose and validate individual elements in a pipeline in order to create the most robust and accurate pipeline possible. Traditional means of choosing a thinning algorithm might include thinning an image or video of a worm and then judging by hand how accurate it seems to be. There are no real metrics to guarantee that this is actually computing relevant values, so the choice of which thinning algorithm to use would be made almost arbitrarily.

To demonstrate the power of this metric, we chose three published thinning algorithms: a single pass thinning algorithm (SPTA) [12], a triple pass thinning algorithm (TPTA) [11], and a one-pass parallel asymmetric thinning algorithm (OPATA-8) [3]. SPTA is a single pass sequential thinning algorithm that uses both flag map and bitmap simultaneously to decide if a boundary pixel can be deleted. TPTA is an older and simpler thinning algorithm that tends to take longer than SPTA and in theory yield less desirable results. OPATA-8 is a more complicated thinning algorithm based on pattern matching each 3x3 set of pixels in the image to one of 18 pre-established patterns. This is an expensive operation when it must be done for every pixel in every frame of a video, but theoretically it produces very good results.

This shape produced by these thinning algorithms is often not a single line, but has multiple endpoints. We reduce it to a single line by selecting the endpoints that are furthest from one another and removing all others.

The output node produces a set of text files, which we refer to as body files. Each body file contains one line per frame of video with each line giving a timestamp and the pixel locations of the body of the worm. The number of pixels (or length of the body) is dependent on the size of the worm and also exhibits some variability due to the binarization of the image and subsequent thinning.

Because we wanted to gather statistics based on a non-uniform segmentation of the worm's body, it was necessary to identify the location of the worm's head. To do this a simple heuristic was applied to capturing video. We made sure that the first frame of every video captured featured the head of the worm to the right of the tail of the worm. This was possible since rotating and placing the camera in the simulator is quick and easy. The first frame of each file was then marked with a head tag on the end that who had the largest x coordinate. With the head tag in place, subsequent frames of the video were properly rearranged such that the end point closest to the last head location was identified as the head. This turns out to be a very robust and reliable mechanism if the video being processed was taken at high enough frame rates.

These head-tagged body files are then post-processed in batches. The WormAnalyzer software takes a directory of head-tagged body files and produces skeleton files that provide a description of the location and position of each segment of the worm. Like the simulator, the size of these segments are not uniform, but are based on the muscle placement as reported in Varshney et al [10].

For convenience, the simulator also creates skeleton files in the same format as the WormAnalyzer. This allows us to directly compare the output from both processes using the same metrics calculated in exactly the same way. The skeleton files are then processed to extract features of the worm's movement using a technique similar to the one reported in Cronin et al [2]. Currently, this software extracts 49 features from the worm's motion including the velocity of the centroid of the worm, the amplitude of the worm's body, the average angle at each joint location, and the angular velocity of each joint (see Figure 1). The software outputs data files that give these features on a frame by frame basis and a set of summary statistics that can be further analyzed using statistical packages.

 ISBN: 978-1-61208-137-3

Table III
AMPLITUDE AND VELOCITY AVERAGES

	Default		Large Angle		Small Angle	
	Amp	Vel	Amp	Vel	Amp	Vel
Sim	168.37	204.08	180.10	40.79	62.64	374.70
OPATA-8	186.39	259.30	176.11	113.02	59.98	391.56
SPTA	182.34	304.71	170.10	131.56	59.03	404.95
TPTA	177.73	409.15	165.13	201.29	58.21	506.10

B. Testing

To test the pipelines, first video was captured for each of the worm variants. To do this, the simulation was run numerous times for each variant. The zoom level was held constant, and the ”camera” remained in one place while the worm performed forward locomotion. Screen capture software was used to record video of the worm moving. No video was taken while the camera was repositioning. The simulator produces a file which details the underlying values of the worm body in the simulation. Each set of values was timestamped, and the timestamp was also captured in the video. The data file from the simulator was then split into smaller files corresponding with the video clips such that each video file would have a corresponding data file that contained a record of the underlying values during that clip. Over 20,000 frames of worm motion were collected over the three variants. Three distinct sets of video was produced: default, large angle, and small angle movement.

Each set of video was batch processed through the CVT three times. Each time a different thinning algorithm was used. The pipeline was constructed such that data files were outputted that contained the thinned skeleton and a timestamp for each frame of the video. These files were then fed into the WormAnalyzer software for statistical analysis. All pipelines used the same segmentation code to produce the 24 anatomically correct skeleton points and analysis code to produce the statistics. Statistics were recorded for amplitude, velocity, segment angles, and segment angle velocities.

Then, the values produced from the simulator were fed into the WormAnalyzer software to produce the same statistics as those produced from the video analysis. These values produced from the simulator files were used as the ”accepted value” since the underlying value was known. The results from the other pipelines are used as the ”experimental value.” Using these values the root mean square (RMS) error for each pipeline/statistic combination was computed. This value represents how accurate the pipeline was in creating accurate skeletons and therefore accurate statistics. The testing was focused on isolating only the thinning algorithm every other operation in the pipeline is exactly identical, including the statistical analysis. This allowed for the differences in accuracy to be attributed to the accuracy of the thinning algorithm. This demonstrates how a broad measure of accuracy of a pipeline can be focused to help validate each video effect in a pipeline.

V. RESULTS

Approximately 20,000 frames of video were processed to produce their relevant statistics. Each statistic was calculated for each combination of worm variant and thinning algorithm. These statistics were then treated as an experimental value, and the statistics produced for the corresponding videos from the simulator were treated as the accepted value. From this a simple root mean square error term was created measuring the average error over each segment of the worm. This error was calculated for all the worm variant/thinning algorithm pairs. In addition, the maximum error for a segment was recorded. This metric will be referred to here as ”pipeline error value” or PEV. The PEV for a given pipeline is a measurement of its accuracy. A perfect pipeline would produce a PEV of 0 (since it perfectly matched the underlying data values). Table I and II shows the PEV values for each pipeline on the various worms used in the simulation. Figure 2 shows how the pipelines performed over the three different worm variants.

In addition to this, the amplitude and velocity for the worm was measured. Each pipeline produced a single amplitude and velocity value for a given worm variant. This is compared against the given value in Table III.

By all metrics, OPATA-8 is the most accurate thinning algorithm we tested when used on *C. elegans*. All of our metrics point towards its accuracy in analyzing various the movements of all the variants of the worm. Thus, this indicates that any future pipelines that require a thinning algorithm should chose OPATA-8 if their main goal is accuracy of statistics.

In measuring the velocities of the worm as a whole, the two older thinning algorithms (SPTA and TPTA) performed very poorly. After examining the data, we found that these algorithms sometimes greatly alter the length of the worm by making poor choices in the thinning algorithm. The newer and more advanced OPATA-8 goes much further in preserving the length of the worm. Preserving as much of a thinned object as possible was actually a goal in development of OPATA-8 [3]. Velocity of the worm is calculated by first obtaining its centroid, and tracking changes to the centroid over time. When the length of the worm changes wildly from frame to frame, the centroid also changes position. This inaccurate change of position is expressed in the poor performance of these thinning algorithms when calculating velocity of the worm.

It should be noted that our metrics do not take into account the speed of processing of these algorithms. If a pipeline is developed whose main goal is speed and not accuracy, one may chose to go with a different algorithm. This metric still has a place in that it can allow researchers to make an informed decision about the trade off in accuracy and speed, as now both are quantifiable results that are easily obtained.

 ISBN: 978-1-61208-137-3

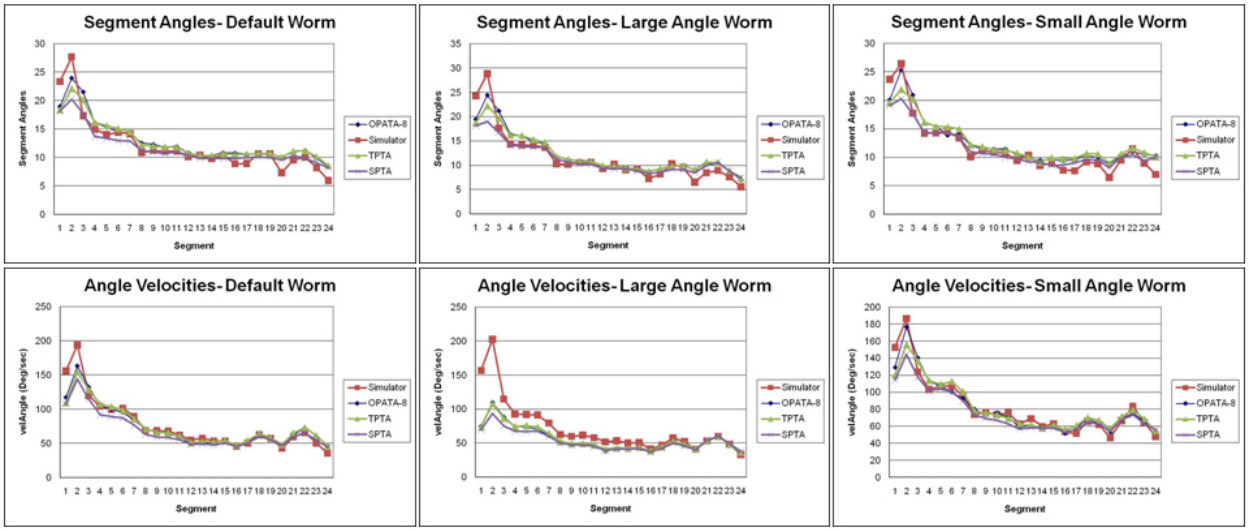


Figure 2. Results of comparison of segment angles and velocities for three locomotion types.

VI. Conclusions

We have demonstrated a new technique for analyzing and quantifying the accuracy of a pipeline, and have shown how this tool can be used to make a simple decision such as deciding which thinning algorithm to choose in the implementation of a pipeline. This is a powerful metric that has not existed before in the *C. elegans* community, and should be used as a tool to measure and compare future pipelines for accuracy. This can also be used in the tuning of pipelines for more reliable results in the future. If research is published with this metric, it would provide further confidence in the results obtained with a pipeline, and allow for easier comparison of data across different institutions by having an independent measurement of the accuracy and confidence in the data obtained on other locations.

References

[1] J. Byden and N. Cohen, "Neural control of caenorhabditis elegans forward locomotion: the role of sensory feedback," *Biological Cybernetics*, vol. 98, pp. 339–351, 2008.

[2] C. J. Cronin, J. E. Mendel, S. Mukhtar, Y.-M. Kim, R. C. Stirbl, J. Bruck, and P. W. Sternberg, "An automated system for measuring parameters of nematode sinusoidal movement," *BMC Genetics*, vol. 6, no. 5, February 2005.

[3] W. Deng, S. S. Iyengar, and N. E. Brener, "A fast parallel thinning algorithm for the binary image skeletonization," *International Journal of High Performance Computing Applications*, vol. 14, no. 1, pp. 65–81, 200.

[4] W. Geng, P. Cosman, C. C. Berry, Z. Feng, and W. R. Schafer, "Automatic tracking, feature extraction and classification of c. elegans phenotypes," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 10, pp. 1811–1820, October 2004.

[5] J. Karbowski, G. Schindelman, C. Cronin, A. Seah, and P. Sternberg, "Systems level circuit model of c. elegans undulatory locomotion: mathematical modeling and molecular genetics," *Journal of Computational Neuroscience*, vol. 24, pp. 253–276, 2008.

[6] R. Mailler, J. Avery, J. Graves, and N. Willy, "A biologically accurate 3d model of the locomotion of caenorhabditis elegans," in *Proceedings of The First International Conference on Computational and Systems Biology and Microbiology (BIOSYSCOM 2010)*, March 2010.

[7] E. Neibur and P. Erdös, "Theory of the locomotion of nematodes: Dynamics of undulatory progression on a surface," *Biophysics Journal*, vol. 60, pp. 1132–1146, November 1991.

[8] ——, "Theory of locomotion of nematodes: Control of the somatic motor neurons by interneurons," *Mathematical Biosciences*, vol. 118, no. 1, pp. 51–82, 1993.

[9] G. D. Tsibidis and N. Tavernarakis, "Nemo: a computational tool for analyzing nematode locomotion," *BMC Neuroscience*, vol. 8, no. 86, 2007.

[10] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii, "Structural properties of the caenorhabditis elegans neuronal network," 2009. [Online]. Available: http://www.citebase.org/abstract?id=oai:arXiv.org:0907.2373

[11] T. Zhang and C. Y. Suen, "A fast parallel algorithm for thinning digital patterns," *Communications of the ACM*, vol. 27, no. 3, pp. 236–239, March 1984.

[12] R. Zhou, C. Quek, and G. Ng, "A novel single-pass thinning algorithm and an effective set of performance criteria," *Pattern Recognition Letters*, vol. 16, pp. 1267–1275, 1995.

 ISBN: 978-1-61208-137-3

Automated Segmentation and Measurement for Cancer Classification of HER2/neu Status in Breast Carcinomas

Lee Sing CHEONG[1] and Angela JEAN[2]
Bioinformatics Group
Nanyang Polytechnic
Singapore
Bioinformatics Research Centre
Nanyang Technological University
Singapore
[1]e-mail: Alice_CHEONG@nyp.gov.sg
[2]e-mail: AngelaJEAN@gmail.com

Tsu Soo TAN[3] and Waiming KONG[4]
Bioinformatics Group
Nanyang Polytechnic
Singapore
[3]e-mail: TAN_Tsu_Soo@nyp.gov.sg
[4]e-mail: KONG_Wai_Ming@nyp.gov.sg

Soo Yong TAN
Department of Pathology
Singapore General Hospital
Singapore
e-mail: tan.soo.yong@sgh.com.sg

***Abstract*—The HER2/neu protein is over-expressed in 20%-30% of breast cancer cases, and is significantly associated with increased breast cancer recurrence and worse prognosis. The assessment of HER2 protein level is visualized using immunohistochemistry (IHC) assays, which is subjected to inter-observer and intra-observer variability. In this paper, we reduce variability by an automated segmentation and measurement system for IHC-stained breast tissue images. From the dataset of breast tissue images, the system is able to obtain the nuclei and the orange stained cell membranes of the cells, quantify the continuous orange hue of the cell membranes, and identify nuclei that are bounded by orange stained cell membrane. This system also suggests a putative assessment classification score for each image based on the same assessment protocol specified for histopathologist. Using the dataset of 42 clinically IHC-scored images, the system correctly suggested the corresponding putative assessment classification score for 39 of the images, achieving an accuracy of 92%.**

Keywords - HER2; neu; breast carcinoma; automated segmentation and measurement; cancer classification.

I. INTRODUCTION

HER2 is the abbreviation for "Human Epidermal growth factor Receptor 2", while oncogene neu was named after the cell line being derived from a rodent glioblastoma cell line, which is a type of neural tumor. It is also named as ErbB-2 for its similarity to ErbB (avian erythroblastosis oncogene B).

The HER2/neu is a cell membrane surface-bound receptor tyrosine kinase protein and is a member of the human epidermal growth factor receptor (HER) family. It interacts with other HER receptors to regulate cell growth, differentiation and survival [1].

HER2/neu protein is over-expressed in 20%-30% of breast cancer cases [2,3], and is significantly associated with increased breast cancer recurrence and worse prognosis [4]. It is the target of the monoclonal antibody trastuzumab (Herceptin US brand name) [5,6], and is FDA approved as part of a treatment plan for the adjuvant treatment of HER2 positive breast cancer patient [6].

In this paper, we will cover the current practice in HER2 assessment, the dataset used, the method employed, the results obtained and discuss the significant observations.

II. HER2 ASSESSMENT

HER2 protein's level of expression is measured through the use of immunohistochemistry (IHC) assays. In an IHC assay, the antibodies bind to the antigen (HER2/neu protein) in the biological tissue.

The assessment protocol assigns a score of 0 to a sample tissue when no observable staining occurs or when less than 10% of the membrane of tissue cells is stained. A score of 1 is assigned when greater than 10% of the membrane of tissue cells is faintly or barely perceptibly stained. A score of 2 is assigned when greater than 10% of the complete membrane of tissue cells is weakly or moderately stained. Lastly, the maximum score of 4 is assigned when greater than 10% of the complete membrane of tissue cells is strongly stained.

While a protocol is in place to assess biological tissue according to membrane staining, inter-observer and intra-observer variability exist. The inter-observer variability reflects the systematic differences among the observers, while the inter-observer variability reflects the discrepancy

 ISBN: 978-1-61208-137-3

resulting from the use of human perception to classify the continuous hue of staining and the completeness of enclosure of the membrane staining.

Thus, an automated segmentation and measurement method that quantify the classification of continuous hue of staining and the completeness of enclosure of the membrane staining, will relieve the histopathologist from performing quantitative classification, and instead concentrates on the assessment of the tissue sample using expert knowledge. This will reduce the amount of assessment variability, and decrease the time used to assess each tissue sample.

III. DATASET

The dataset used in this paper consist of 42 IHC-stained and clinically IHC-scored breast tissue histological images, which were obtained from Dr Tan Soo Yong. Of these 42 images, 3 images were of IHC score 0, 12 images were of IHC score 1, 11 images were of IHC score 2 and 16 images were of IHC score 3.

Each of these images is of dimensions 1300 pixels by 1030 pixels and is in the TIFF image format. They are manually acquired through microscopic imaging of normal breast tissue and varying grades of breast cancer tissue that had been IHC-stained and mounted on histological slides.

For the IHC-stained normal breast tissue cells, the nuclei are stained blue-purple and the cell membranes are hardly stained. For IHC-stained abnormal breast tissue cells, the nuclei are also stained blue-purple, while the cell membranes are stained in varying orange hue.

IV. METHOD

Here, we present an automated segmentation and measurement system that is able to identify nuclei that are bounded by orange stained cell membrane and quantify the continuous hue of orange stained cell membrane. In addition, based on the assessment protocol specified for histopathologist, but without any additional expert knowledge that is accumulated through experience for an expert, a putative assessment classification score is suggested.

A. *Identify orange hue in tissue image*

The breast tissue images are identified as either containing cell membranes that are not stained orange or are stained orange in the following steps:

Step 1: The color space of the image is converted from RGB color space to CIE-Lab color space.

Step 2: The frequency of pixels having each of the possible chromaticity (chromatic component a and chromatic component b) is calculated.

Step 3: The extremal chromaticity maxima values of the frequency of pixels are obtained.

Step 4: The image is classified as breast tissue containing cell membranes that are without orange stain if there are two extremal chromaticity maxima points (It can be seen from Figure 1 that there is two extremal chromaticity maxima points in a plot of maxima chromaticity points for breast tissue image without orange stained cell membranes).

Likewise, it is classified as breast tissue containing cell membranes that are with orange stain if there are three extremal chromaticity maxima points (It can be seen from Figure 2 that there is three extremal chromaticity maxima points in a plot of maxima chromaticity points for breast tissue image with orange stained cell membranes).

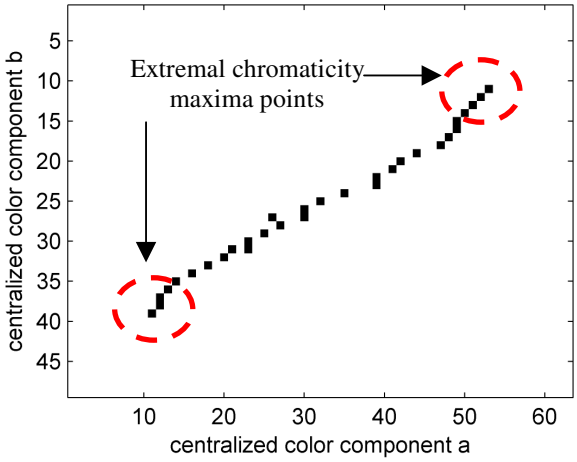


Figure 1. Plot of maxima chromaticity points for breast tissue image without orange stained cell membranes

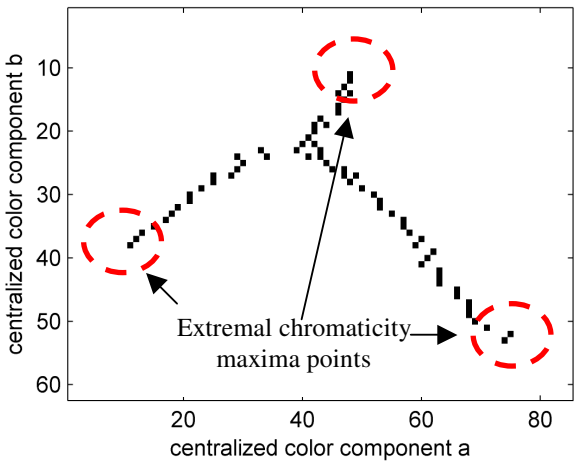


Figure 2. Plot of maxima chromaticity points for breast tissue image with orange stained cell membranes

B. *Segment breast tissue image with orange hue*

The images identified as breast tissue image with orange stained cell membranes is further processed in the following sections. In this section, the breast tissue image with orange stained cell membranes is segmented to obtain the nuclei cluster and the orange stained cell membranes cluster in the following steps:

 ISBN: 978-1-61208-137-3

Step 1: The chromaticity value corresponding to the maximum frequency of chromaticity of the pixels is obtained.

Step 2: The chromaticity value corresponding to the average chromaticity of the maximum frequency's chromaticity obtained in the previous step and the extremal chromaticity maxima value corresponding to the orange hue is obtained.

Step 3: The three extremal chromaticity maxima values obtained in the previous section, the maximum frequency's chromaticity value obtained in Step 1, and the average chromaticity value obtained in Step 2, are used as seed values to perform k-means clustering on the image.

Step 4: The cluster corresponding to the k-means centroid with least color-component b is classified as the nuclei cluster, and is shown in Figure 3.

Step 5: The two clusters corresponding to the k-means centroids with the greatest two values of the sum of the color-components are both classified as the orange stained cell membranes cluster, and is shown in Figure 4.

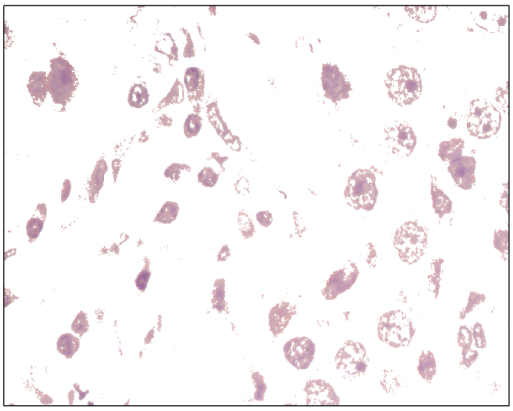


Figure 3. The cluster classified as nuclei cluster

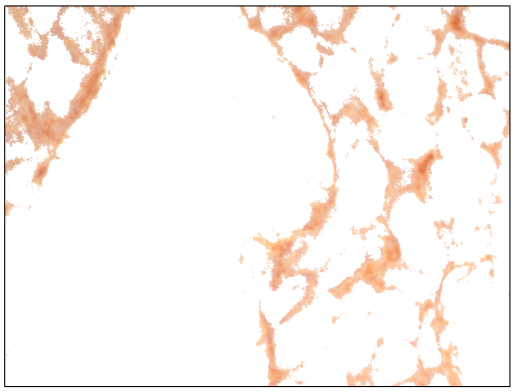


Figure 4. The cluster classified as orange stained cell membranes

C. *Refine the nuclei cluster*

The nuclei cluster, being identified by k-means clustering, contains stray pixels and pixel regions that do not belong to any nucleus. Based on each of the connected region's measurement, the nuclei cluster is further refined by the following steps:

Step 1: The nuclei cluster mask of the nuclei cluster is obtained.

Step 2: Morphological opening is performed on the nuclei cluster mask to remove stray pixels.

Step 3: Each connected region of the nuclei cluster mask, based on neighboring 8-connectivity measurement is individually labeled.

Step 4: For each of the connected region, based on the regional measurement properties of Euler number, major axis, minor axis length and filled area, each connected region is classified as being a nucleus region or not a nucleus region.

D. *Identify orange stained cell membrane bounded nucleus*

The region surrounding each of the connected nucleus region that had been obtained in the previous section of nuclei cluster refinement, is inspected to identify whether it is bounded by orange stained cell membrane in the following steps:

Step 1: For each connected nucleus region obtained in the previous section, its connected nucleus region mask is obtained.

Step 2: Morphological dilation of the connected nucleus region mask is performed to obtain the extended region mask that includes both the connected nucleus region and the surrounding region.

Step 3: The cell membrane region mask is obtained by identifying pixels that are in the orange stained cell membranes cluster obtained by k-means in the earlier section and is within the extended region mask obtained in the previous step.

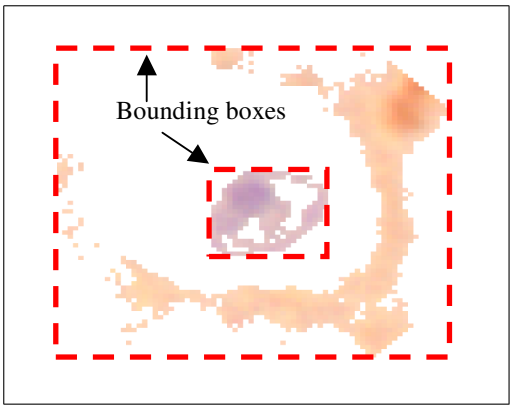


Figure 5. The bounding boxes of the nucleus region mask and the cell membrane region mask for a nucleus that is bounded by orange stained cell membrane

 ISBN: 978-1-61208-137-3

Step 4: The regional measurement property of bounding box for both the nucleus region mask and the cell membrane region mask are obtained.

Step 5: The nucleus is classified as bounded by the orange stained cell membrane if the bounding box of the cell membrane region mask is enclosing the bounding box of the nucleus region mask, and is shown in Figure 5.

E. Quantify continuous hue of orange stain cell membrane

In the following steps, the continuous hue of orange stain for each orange stained cell membrane bounded nucleus is quantify into three ranges; The first range quantifies the description of faintly or barely perceptibly stained cell membrane, the second range quantifies the description of weakly or moderately stained cell membrane, and the third range quantifies the description of strongly stained cell membrane.

Step 1: Using the cell membrane region mask obtained in the previous section, the mean CIE lightness of the orange stained cell membrane region is obtained.

Step 2: Using thresholding, the mean CIE lightness for each orange stained cell membrane bounded nucleus is quantified into one of the possible three ranges corresponding to (1) faintly or barely perceptibly stained membrane, (2) weakly or moderately stained membrane or (3) strongly stained membrane.

F. Provide putative assessment classification score

Based on the percentage of orange stained cell membrane bounded nuclei for each range over the total number of nuclei, a putative assessment classification score is provided for each of the image by the following steps:

Step 1: The number of orange stained cell membrane bounded nuclei for each range is obtained using the quantification of each orange stained cell membrane bounded nucleus identified in the previous section.

Step 2: The total number of connected nucleus regions is obtained based on neighboring 8-connectivity measurement of the refined nuclei cluster obtained in the earlier section.

Step 3: The percentage of the number of orange stained cell membrane bounded nuclei for each range obtained in Step 1 over the total number of nuclei obtained in Step 2 is obtained.

Step 4: The tissue image containing cell membranes that are not stained orange or consists of less than 10% of the orange stained cell membrane bounded nuclei in the range corresponding to faintly or barely perceptibly stained membrane is given a putative assessment classification score of 0.

The tissue image that consists of greater than 10% of the orange stained cell membrane bounded nuclei in the range corresponding to faintly or barely perceptibly stained membrane is given a putative assessment classification score of 1.

The tissue image that consists of greater than 10% of the orange stained cell membrane bounded nuclei in the range corresponding to weakly or moderately stained membrane is given a putative assessment classification score of 2.

The tissue image that consists of greater than 10% of the orange stained cell membrane bounded nuclei in the range corresponding to strongly stained membrane is given a putative assessment classification score of 3.

G. Visualize the orange hue quantification

Visualization of the identified nuclei and the quantification of the continuous hue of orange of the cell membrane bounded nuclei are provided by the use of four different colors. In the following steps, four different colors are used to correspondingly outline nuclei's whose cell membranes are (1) not stained, (2) faintly or barely perceptibly stained, (3) weakly or moderately stained, or (4) strongly stained.

Step 1: The perimeter of each of the nucleus identified in the refined nuclei cluster is obtained.

Step 2: Based on the quantification of each of the orange stain cell membrane bounded nucleus, the perimeter of these nuclei are colored with one of the three colors according to their quantification of orange hue.

Step 3: The remaining nuclei whose perimeter are colored in the Step 2, are those whose cell membranes are not stained. They are colored a different color from the three colors used in Step 2.

V. RESULT

The putative assessment classification score suggested by the automated segmentation and measurement system described in this paper achieved an accuracy of 92% with 39 out of 42 images in the dataset having the same corresponding classification score as the IHC score provided by clinician.

The corresponding putative assessment classification score for 3 out of 3 images of IHC score 0, 10 out of 12 images of IHC score 1, 10 out of 11 images of IHC score 2 and 16 out of 16 images of IHC score were correctly suggested.

For the 2 images of IHC score 1 with incorrectly suggested classification score, they were both scored as one category more serious with classification score 2 that corresponds to IHC score 2. Likewise, for the single image of IHC score 2 with incorrectly suggested classification score, it is scored as one category more serious with classification score 2 that corresponds to IHC score 3.

Thus, the automated segmentation and measurement system provides visualization of the identified nuclei and the quantification of the continuous hue of orange of the cell membrane bounded nuclei, along with a suggestion of putative assessment classification score that is highly accurate.

VI. DISCUSSION

In this section, five observations are being discussed.

 ISBN: 978-1-61208-137-3

A. *Challenge of variable stains' hue*

There is great variance in the stains' hue for the same components of cells in the tissue images. As seen in Figure 6, the same description of blue-purple stain has different chromaticity in different image.

These variances in hue can be due to many reasons, such as the difference in the amount of stains used or amount of washing performed during the process of preparing of the IHC-stained histological slides, the different lighting condition of the microscopy and the difference in automatic white balancing of the image-capturing device.

The stain for the same cell component in each image having different chromaticity pose the difficulty of requiring the identification of the number of hues in the image, and the identification of the nuclei cluster and the cell membranes cluster, to be non-dependent on exact chromaticity and has to be adaptive to each of the image.

In this paper, the identification of the number of hues in the image circumvent this difficulty through the use of the chromaticity frequency matrix of the individual image to obtain the number of extremal chromaticity maxima points, which corresponds to the same number of hues in the image.

Likewise, k-means clustering method used to identify the nuclei cluster and the cell membranes cluster is able to circumvent this difficulty, as the method is not dependent exact chromaticity.

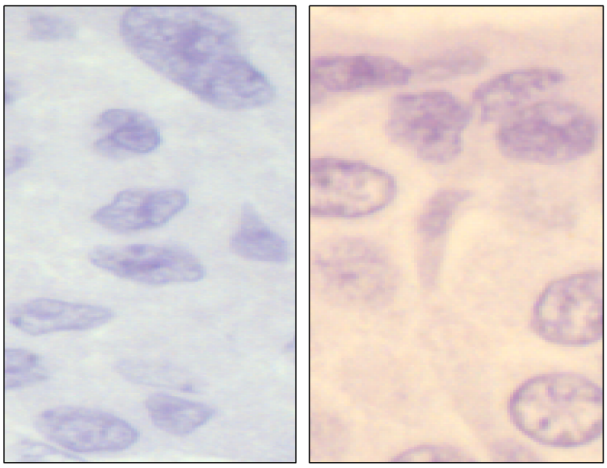


Figure 6. Variance in stains in different image

B. *Challenges in using k-means clustering*

The challenges in using k-means cluster are that the number of seed values has to be known prior and the provision of initial seed values greatly influence the clustering result.

The first challenge of knowing the number of seed values in advance had been solved in the previous section through the use of the chromaticity frequency matrix of the individual image to obtain the number of extremal chromaticity maxima points, which corresponds to the same number of hues in the image.

The second challenge of providing seed values that converge to the required solution is solved through the use of seed values adapted to the individual image, which includes the three extremal chromaticity maxima points and two additional intermediate maxima points.

C. *Reliance on cell characteristics domain knowledge in the refinement of nuclei cluster*

The refinement of nuclei cluster relied on the cell characteristics domain knowledge, in which each nucleus is required to have a minimum size and has the spherical-like shape. This cell characteristics information varies depending on the type of cells that are contained within the breast tissue image.

D. *Challenge of identifying incompletely stained cell membrane as enclosing nucleus*

The incompletely stained cell membrane of nucleus in any of the image can consist of multiple disconnected stained regions of varying distance from the closest point of the nucleus. This poses a challenge in identifying whether each nucleus is being bounded by orange stained cell membrane.

In this paper, in order to overcome the challenge, the bounding box of possibly disconnected regions of the cell membrane pixels that are within an extended distance from the nucleus is evaluated on the criteria of whether it encloses the bounding box of the nucleus under inspection.

The use of bounding box relax the criteria from identifying the entire surrounding cell membrane, which might be incompletely stained, to identifying sufficient number of pixels of the cell membrane to form the bounding box that will enable correct solution to be obtained.

E. *Reliance on domain knowledge of cell membranes staining in quantifying cell membranes' stain*

The quantifying of cell membranes' stain into one of the possible three ranges corresponding to (1) faintly or barely perceptibly stained membrane, (2) weakly or moderately stained membrane or (3) strongly stained membrane rely on domain knowledge of the threshold values for the mean CIE lightness.

F. *Reliance on the assessment protocol specified for histopathologist to suggest putative assessment classification score*

The suggested putative assessment classification score rely on the assessment protocol by the scoring of each image based on the percentage of orange stained cell membrane bounded nuclei for each of the three ranges over the total number of nuclei.

VII. CONCLUSION AND FUTURE WORK

We have presented an automated segmentation and measurement system that is able to identify nuclei that are bounded by orange stained cell membranes, quantify the

 ISBN: 978-1-61208-137-3

continuous hue of orange stained cell membrane, and suggest a putative assessment classification score for each image.

Using the dataset of 42 clinically IHC-scored images, the system correctly suggested the corresponding putative assessment classification score for 39 of the images, achieving an accuracy of 92%.

Of the 2 IHC-scored 1 images and the single IHC-scored 2 image, whose score are incorrectly suggested, the putative assessment classification score suggested were one category more serious.

In addition, we discussed the challenges posed by the nature of the images, the challenges of the method being used, and the reliance of the method used on domain knowledge and assessment protocol.

As the system was tested on 42 clinical IHC-scored images, we propose that future work includes testing on a larger dataset.

ACKNOWLEDGEMENT

This study was funded by the Ministry of Education, Singapore (Innovation Fund grant no. MOE2008-IF-1-018).

REFERENCES

[1] Yarden Y., Biology of HER2 and Its Importance in Breast Cancer, Oncology, Vol. 61, Suppl. 2, Pg. 1-13, 2001.

[2] McCann A. H., Dervan P. A., O'Regan M., Codd M. B., Gullick W. J., Tobin B. M., and Carney D. N., Prognostic Significance of c-erbB-2 and Estrogen Receptor Status in Human Breast Cancer, Cancer Research, Vol. 51, Pg. 3296-3303, 15 June 1991.

[3] Slamon D. J., Godolphin W., Jones L. A., Holt J. A., Wong S. G., Keith D. E., Levin W. J., Stuart S. G., Udove J., Ullrich A., and al. et., Studies of the HER-2/neu Proto-oncogene in Human Breast and Ovarian Cancer, Science, Vol. 244, Issue 4905, Pg. 707-712, 12 May 1989.

[4] Salmon D. J., Clark G. M., Wong S. G., Levin W. J., Ullrich A., and McGuire W. L., Human Breast Cancer: Correlation of Relapse and Survival with Amplification of the HER-2/neu Oncogene, Science, Vol. 235, No. 4785, Pg. 177-182, 9 January 1987.

[5] Albanel J., Codony J., Rovira A.., Mellado B., and Gascon P., Mechanism of Action of Anti-HER2 Monoclonal Antibodies, Advances in Experimental Medicine and Biology, Vol. 532, Pg. 253-268, 2003.

[6] Goldenberg M. M., Trastuzumab, A Recombinant DNA-derived Humanized Monoclonal Antibody, A Novel Agent for the Treatment of Metastatic Breast Cancer, Clinical Therapeutics, Vol 21, Issue. 2, Pg.309-318, February 1999.

 ISBN: 978-1-61208-137-3

Incorporating Protein Sequence and Evolutionary Information in a Structural Pattern Matching Approach for Contact Maps

Hazem Radwan A. Ahmed
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: hazem@cs.queenu.ca

Janice I. Glasgow
School of Computing and Information Science
Queen's University
Kingston, Ontario, Canada. K7L 3N6
Email: janice@cs.queensu.ca

***Abstract*— Protein structure prediction from the primary sequence remains a major challenging problem in bioinformatics. The main issue here is that it is computationally complex to reliably predict the full three-dimensional structure of a protein from its one-dimensional sequence. A two-dimensional contact map has, therefore, been used as an intermediate step in this problem. A contact map is a simpler, yet representative, alternative for the three-dimensional protein structure. In this paper, we propose a pattern matching approach to locate similar substructural patterns between protein contact map pairs using protein sequence information. These substructural patterns are of particular interest to our research, because they could ultimately be used as building blocks for a bottom-up approach to protein structure prediction from contact maps. We further demonstrate how to improve the performance of identifying these patterns by incorporating both protein sequence and evolutionary information. The results are benchmarked using a large standard protein dataset. We performed statistical analyses (e.g., Harrell-Davis Quantiles and Bagplots) that show sequence information is more helpful in locating short-range contacts than long-range contacts. Moreover, incorporating evolutionary information has remarkably improved the performance of locating similar short-range contacts between contact map pairs.**

Keywords-protein structure prediction; protien contact maps; structural pattern matching; evolutionary information; harrell-davis quantiles.

I. INTRODUCTION

Since the human genome sequence was revealed in April 2003, the need to predict protein structures from protein sequences has dramatically increased [1]. Proteins are macromolecules with a wide range of biological functions that are vital for any living cell. They transport oxygen, ions, and hormones; they protect the body from foreign invaders; and they catalyze almost all chemical reactions in the cell. Proteins are made of long sequences of amino acids that fold into three-dimensional structures. Because protein folding is not easily observable experimentally [2], protein structure prediction has been an active research field in bioinformatics as it can ultimately broaden our understanding of the structural and functional properties of proteins. Moreover, predicted structures can be used in structure-based drug design, which attempts to use the structure of proteins as a basis for designing new ligands by applying principles of molecular recognition [3].

In recent decades, many approaches have been proposed for understanding the structural and functional properties of proteins. These approaches vary from time-consuming and relatively expensive experimental determination methods (e.g., X-ray crystallography [4] and NMR spectroscopy [5]) to less-expensive computational protein modeling methods for protein structure prediction (e.g., ab-initio protein modeling [6], comparative protein modeling [7], and side-chain geometry prediction [8]). While the computational methods attempt to circumvent the complexity of the experimental methods with an approximation to the solution (predicted protein structures versus experimentally-determined structures), analyzing the three-dimensional structure of proteins computationally is not a straightforward task. Hence, two-dimensional representations of protein structures, such as distance and contact maps, have been widely used as a promising alternative that offers a good way to analyze the 3D structure using a 2D feature map [9]. This is because they are readily amenable to machine learning algorithms and can potentially be used to predict the three-dimensional structure, achieving a good compromise between simplicity and competency [26].

The paper is organized as follows: Section II provides the reader with the background material required to understand the concepts used in this study. It describes distance and contact maps, gives examples of structural patterns of contact maps, and discusses protein similarity relationships at different representational levels of detail, as well as the structural classification of protein domains. Section III presents the experimental setup and the details of the multi-regional analysis of the contact map method used in our experiments. Section IV discusses the experimental benchmark dataset used in this study and shows the performance of the proposed method using statistical analyses, including a quantile-based analysis as well as a correlation analysis. The final section highlights the contributions and summarizes the main results of the study. It also presents a set of potential directions for future research.

 ISBN: 978-1-61208-137-3

II. BACKGROUND MATERIAL

Contact and distance maps provide a compact 2D representation of the 3D conformation of a protein, and capture useful interaction information about the native structure of proteins. Contact maps can ideally be calculated from a given structure, or predicted from protein sequence. The predicted contact maps have received special attention in the problem of protein structure prediction, because they are rotation and translation invariant (unlike 3D structures). While it is not simple to transfer contact maps back to the 3D structure (unlike distance maps), it has shown some potential to reconstruct the 3D conformation of a protein from accurate and even predicted (noisy) contact maps [10].

A. Distance and Contact Maps

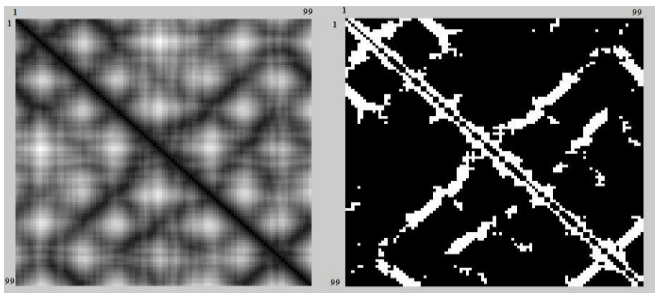
A distance map, *D*, for a protein of *n* amino acids is a two-dimensional *n* x *n* matrix that represents the distance between each pair of alpha-carbon atoms of the protein, as shown in Figure 1(a). The darker the region is, the closer the distance of its corresponding atom pairs is. The distance information can be used to infer the interactions among residues of proteins by constructing another same-sized matrix called a contact map.

A contact map, *C*, is a two-dimensional binary symmetric matrix that represents pairs of amino acids that are in contact, i.e., their positions in the three-dimensional structure of the protein are within a given distance threshold (usually measured in Ångstroms), as shown in Figure 1(b). According to extensive experimental results presented in [11], contact map thresholds, ranging from 10 to 18 Å allow the reconstruction of 3D models from contact maps to be similar to the protein's native structure.

An element of the i^{th} and j^{th} residues of a contact map, C*(i,j)*, can be defined as follows:

$$\mathbf{C}(i,j) = \begin{cases} \mathbf{1}; & \text{if } D(i,j) \leq \text{Threshold} \\ \mathbf{0}; & \text{otherwise} \end{cases}$$

Where *D(i,j)* is the distance between amino acids i and j, ***1*** denotes *contacts* (white), and **0** denotes *no contacts* (black).



(a) (b)

Figure 1. (a) Distance map for a protein of 99 amino acid residues. (b) contact map for the same protein of 99 amino acids after applying a distance threshold of 10 Ångstrom (1 nm) on its distance map. (local contacts < 3.8 Å are ignored – refer to Section III-C for details.)

B. Structural patterns of Contact Maps

Different secondary structures of proteins have distinctive structural patterns in contact maps. In particular, an α-helix appears as an unbroken row of contacts between i, i ± 4 pairs along the main diagonal, while beta-sheets appear as an unbroken row of contacts in the off-diagonal areas. A row of contacts that is parallel to the main diagonal represents a pair of parallel β-sheets, while a row of contacts that is perpendicular to the main diagonal represents a pair of anti-parallel β-sheets [12].

C. The Classification of Protein Domains

The Structural Classification of Proteins (SCOP) database was designed by G. Murzin et al. [15] to provide an easy way to access and understand the information available for protein structures. The database contains a detailed and comprehensive description of the structural and evolutionary relationships of the proteins of known structure. Structurally and evolutionarily related proteins are classified into similar levels in the database hierarchy. Evolutionarily-related proteins are those that have similar functions and structures because of a common descent or ancestor. The main levels in the classification hierarchy of the SCOP database are as follows: 1) *Family* level that implies clear evolutionary relationship, 2) *Superfamily* level that implies probable common evolutionary origin, and 3) *Fold* level that implies major structural similarity.

D. Protein Similarity Relationships

Understanding protein similarity relationships is vital for the further understanding of protein functional similarity and evolutionary relationships. Although a protein with a given sequence may exist in different conformations, the chances that two highly-similar sequences will fold into distinctly-different structures are so small that they are often neglected in research practice [13]. This suggests that sequence similarity could generally indicate structure similarity. Furthermore, a pair of proteins with similar structure has similar contact maps [14]. Therefore, as shown in Figure 2, by the transitivity relationship, a logical inference could be drawn regarding the association between sequence similarity and contact map similarity. The premise of the method of multi-regional analysis of contact maps in this paper is based on this transitive similarity relationship between contact map and protein sequence (via structure).

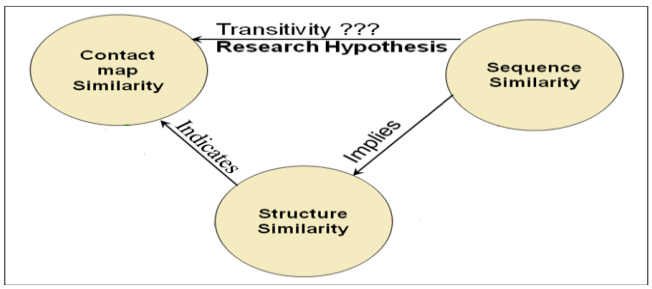


Figure 2. Protein similarity relationships at different levels of detail.

III. METHOD AND EXPERIMENTAL SETUP

This section describes the multi-regional analysis of the contact maps method used in the experiments. The method examines whether sequence similarity information helps in a pattern matching approach to locate regions of similarity in contact maps (the target substructural patterns) that correspond to local similarities in protein structures. The first stage of this method aims to align pairs of protein sequences for each combination pair of contact maps to find the most local similar subsequences. The next stage aims to quantify the similarity of contact maps regions that correspond to these similar subsequences found in the first stage. Finally, different statistical analyses were considered to evaluate the performance of the method, and to determine how well local protein sequence similarity leads to corresponding local contact map similarity.

A. Experimental Dataset

The benchmark Skolnick dataset is adopted for our experiments. The Skolnick dataset is a standard benchmark dataset of 40 large protein domains, divided into four categories as shown in Table I. It was originally suggested by J. Skolnick and described in [16]. The dataset has been used in several recent studies related to structural comparison of proteins [16][17][18].

TABLE I. PROTEIN DOMAINS IN SKOLNICK DATASET

Categories	Global sequence similarity	Sequence length (residues)	Domain indices
1	15-30% (low)	124	1-14
2	7-70% (Med)	170	35-40
3	35-90% (High)	99 (Short)	15-23
4	30-90% (High)	250 (Long)	24-34

B. Sequence Analysis

For the sequence analysis stage, we align every combination pair of sequences. The SIM algorithm [19] is used for this purpose. This algorithm employs a dynamic programming technique to find user-defined, non-intersecting alignments that are the best (i.e., with the highest similarity score) between pairs of sequences. The results from the alignments are sorted descendingly according to their similarity score [20].

In this method, we are only interested in alignments of subsequence of at least 10 residues, and at most 20 residues. We are not interested in alignments of length less than 10 residues because these alignments would not form a complete substructural pattern (for example, the lengths of alpha helices vary from 4 or 5 residues to over 40 residues, with an average length of about 10 residues [21]). We are also not interested in long alignments because most methods for contact maps analysis are known to be far more accurate on local contacts (those contacts that are clustered around the main diagonal), than nonlocal (long-range) contacts [22]. Thus, to eliminate one source of uncertainty of the long-range contacts, alignments of a length greater than 20 residues are not considered.

In this experiment, BLOSUM62 [23] is used as the similarity metric to score sequence alignment. As for gaps, the open and extended gap penalties are set to 10 and 1 respectively. This is because a large penalty for opening a gap and a much smaller one for extending it have generally proven to be effective [24]. An open gap penalty is a penalty for the first residue in a gap, and an extended gap penalty is a penalty for every additional residue in it. To analyze pairs of sequences, the best 100 local subsequence alignments are generated from every pair of sequences. Then, a selection strategy is used to select the two alignments of 10-20 residues with the most and least similarity score (to check the performance in case of low and high similarity).

C. Contact Map Analysis

The second stage of the method is to locate contact map regions that correspond to the most and least similar protein subsequences. In order to unbiasedly analyze the diagonal contact map regions, we ignored local contacts between each residue and itself on the main diagonal. Comparing the main diagonal of contact maps (protein backbone) will neither add meaningful information for their similarity nor dissimilarity, (for example, even too distant contact maps will share a similar main diagonal). Based on the fact that the minimum distance between any pair of different residues cannot be less than 3.8 Å [22], every local contact of each residue and itself that is less than this threshold is ignored.

Jaccard's Coefficient (J) [25] is used as a similarity metric to score contact map regions. J is suitable for measuring contact map similarity, because it does not consider counting zero elements in the matrix (no contacts) of both contact maps, removing the effect of the "double absence" that has neither meaningful contribution to the similarity, nor the dissimilarity, of contact maps.

$$J = \frac{C_{11}}{S - C_{00}} \quad (1)$$

Where C_{11} is the count of nonzero elements (contacts) of both contact maps, C_{00} is the count of zero elements (no-contacts) of both contact maps, and S is the contact map size (i.e., the square of the sequence length for the contact map).

D. Sequence Gap and Region Displacement Problem

The displacement problem happens when a pair of aligned subsequences is very similar (greater than 70%), but their corresponding diagonal contact map regions are not as similar (less than 50-60%). This is noticed to occur as a result of a slight shift in the aligned subsequence pair either because of a gap in the alignment, or because of a slightly shifted alignment. In this case, if the right displacement is considered for one of the aligned subsequence in the correct direction with the correct number of residues, their corresponding diagonal contact map regions will perfectly overlay one another and their similarity can go up to 90%, as shown in Figure 3. The current experimental setup, however, (e.g., open gap penalty, extended gap penalty, etc.) are optimized to minimize the displacement problem. As shown

 ISBN: 978-1-61208-137-3

in Figure 4, the proposed method was successful in locating the exact correct boundaries of contact map regions that perfectly overlay one another, in an effort to maximize their similarity. That is, if any boundary is shifted only by one or two residues, the local contact map similarity will be significantly dropped, as shown in Figure 3 and Figure 5.

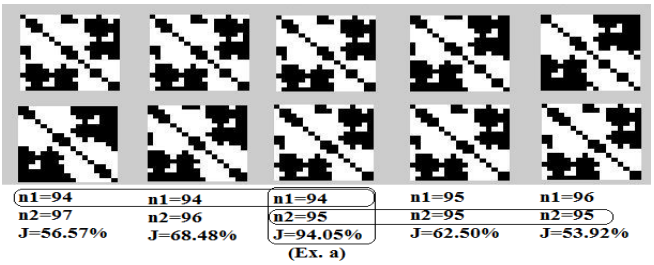


Figure 3. One example of the calculated region boundaries (n1 = 94 & n2 = 95) shows that the selected boundaries have the maximum Jaccard's coefficient (J = 94%) as opposed of 68% and 56% if the lower boundary is shifted by only one residue at a time, or 62% and 53% if the upper boundary is shifted by one residue at a time, instead.

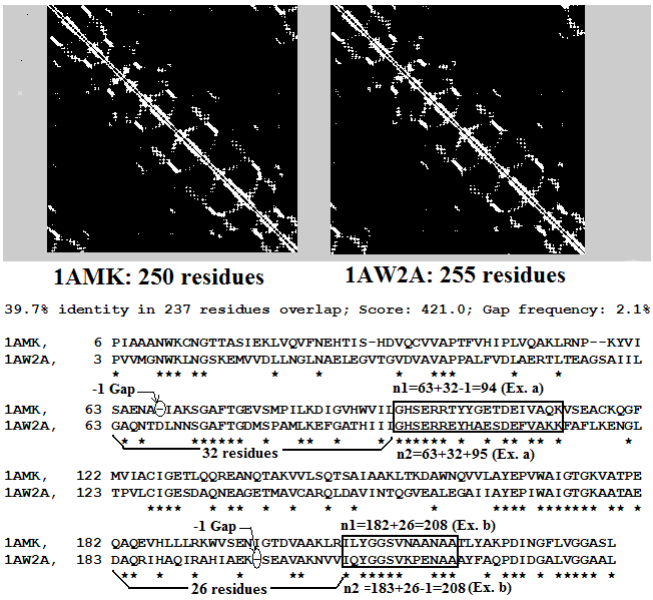


Figure 4. An illustration of the displacement problem between two highly-similar proteins (1AMK & 1AW2A). The gap length is subtracted from the start position of the upper boundary (n1 of Ex. a) and the lower boundary (n2 of Ex. b), since contact maps have no representation of gaps.

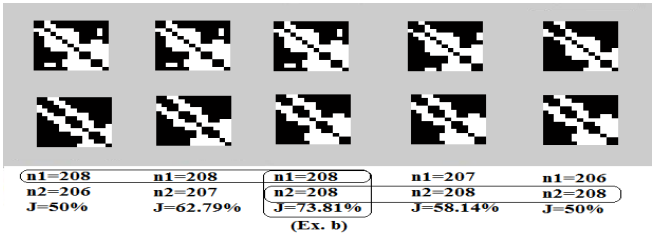


Figure 5. Another example of the calculated region boundaries of (Ex. b) also shows that the selected boundaries have the maximum Jaccard's coefficient (J = 73%) as opposed of 62% and 50% if the lower boundary is shifted by only one residue at a time, or 58% and 50% if the upper boundary is shifted by one residue at a time, instead.

IV. RESULTS AND DISCUSSION

A. *The Big Picture*

To see the big picture of the problem, an all-against-all pair-wise analysis is performed on the benchmark Skolnick dataset, yielding several hundreds of pairwise alignment instances. The entire results of sequence and contact map similarity of each pairwise instance are presented as a 2D scatter plot to study the correlation between them, as shown in Figure 6. This figure draws a clear distinction between the correlation between sequence similarity and their contact map similarity in the diagonal area (short-range contacts), and the correlation between sequence similarity and their contact map similarity in the off-diagonal areas (long-range contacts).

Firstly, for long-range contacts, no matter how high the sequence similarity is the majority of the corresponding contact map similarity is very low (less than 20%). Thus, even high sequence similarity cannot help to suggest corresponding similarity for the long-range contacts. Secondly, for the short-range contacts, the plot reveals two different trends: 1) when sequence similarity is low (less than 60%), contact map similarity is indiscriminately dispersed between a very low similarity level (35%) and a very high one (90%), making it hard to reliably associate low sequence similarity to short-range contact map similarity. 2) When sequence similarity is high (greater than 60%), contact map similarity is apparently clustered in the upper-right corner of the plot (around 80%), suggesting a high correlation between local sequence similarity and short-range contact map similarity.

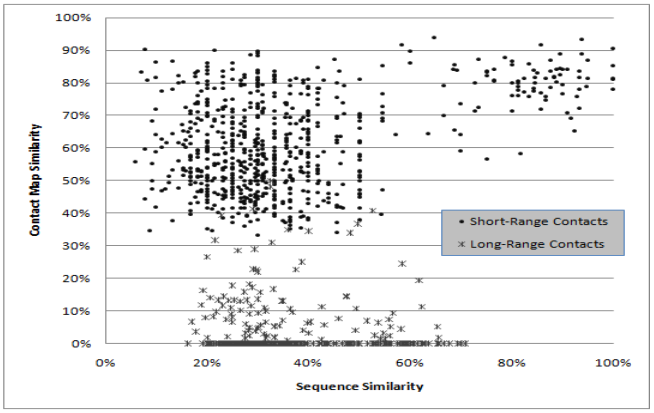


Figure 6. A 2D scatter plot showing the correlation between sequence similarities and their contact map similarities in the diagonal area (short-range contacts) and the off-diagonal areas (long-range contacts).

B. *Harrell-Davis Quantiles*

In an effort to improve performance in locating similar patterns in the diagonal regions of contact map pairs, evolutionary information (represented in SCOP family information) is proposed to be incorporated with the sequence information. As described in [18], the 40 protein

 ISBN: 978-1-61208-137-3

domains of the Skolnick dataset are classified into five SCOP families. Based on SCOP family information, the results are distributed into four different groups: 1) the first group includes the results of pairs of protein subsequences that are most similar and of the same SCOP family. 2) The second group includes the results of pairs of protein subsequences that are most similar and of a different SCOP family. 3) The third group includes the results of pairs of protein subsequences that are least similar and of the same SCOP family. 4) The last group includes the results of pairs of protein subsequences that are least similar and of a different SCOP family.

Quantile-based analysis is performed to compare the different groups. The q^{th} quantile of a dataset is defined as the value where the *q*-fraction of the data is below q and the (1- *q*) fraction of the data is above q. Some *q*-quantiles have special names: the 2-quantile (or the 0.5 quantile) is called the median (or the 50^{th} percentile), the 4-quantiles are called quartiles, the 10-quantiles are called deciles, and the 100-quantiles are called percentiles. For example, the 0.01 quantile = the 1^{st} percentile = the bottom 1% of the dataset, and the 0.99 quantile = the 99^{th} percentile = the top 1% of the dataset.

Using the online R statistics software in [27], the Harrell-Davis method for 100-quantile estimation is computed for this study. The Harrell-Davis method [29] is based on using a weighted linear combination of order statistics to estimate quantiles. The standard error associated with each estimated value of a quantile is also computed and plotted as error bars, as shown in Figure 7. Error bars are commonly used on graphs to indicate the uncertainty, or the confidence interval in a reported measurement. Figure 7(a) clearly shows that the results of contact map similarity of the same family are much better (higher) than those of a different family as in Figure 7(b). This supports the previous hypothesis that incorporating evolutionary information with sequence information improves the performance of locating remarkably better (highly-similar) diagonal contact map region. Comparing Figure 7(a) and Figure 7(c) reveals that low sequence information considerably deteriorates the method performance, even for the results of the same SCOP family. Whereas, comparing Figure 7(c) and Figure 7(d) demonstrates that with low sequence information, the performance is almost the same (poor), no matter if the protein pairs are of the same or of a different SCOP family.

C. Bagplots

A bagplot, initially proposed by Rousseeuwet et al. [30], is a bivariate generalization of the well known boxplot [31]. In the bivariate case, the "box" of the boxplot changes to a convex polygon forming the "bag" of the bagplot. The bag includes 50% of all data points. The fence is the external boundary that separates points within the fence from points outside the fence (outliers), and is simply computed by increasing the bag by a given factor. Data points between the bag and fence are marked by a light-colored loop. The loop is defined as the convex hull containing all points inside the fence. The hull center is the centre of gravity of the bag. It is either one center point (the median of the data) or a region of more than one center points, usually highlighted with a different color. Therefore, the classical boxplot can be considered as a special case of the bagplot, particularly when all points happen to be on a straight line. The bagplot provides a visualization of several characteristics of the data: its location (the median), spread (the size of the bag), correlation (the orientation of the bag), and skewness (the shape of the bag) [30].

In this statistical analysis, we study the effect of the global sequence similarity on the method performance. Thus, the factor that varies in this analysis is the global similarity information, while other factors will be fixed at their best settings obtained from Figure 7(a). In particular, 1) for the local similarity information, the subsequence pairs of the most local similarity will be used. 2) For the region of similarity, short-range contacts in the diagonal area will be considered. 3) For the evolutionary information, protein pairs will be of the same protein SCOP family. According to the global similarity information of the four categories of the Skolnick dataset (shown in Table I), the pair-wise results are further grouped into four clusters. Namely, 1) Low *vs*. Low, 2) Med *vs*. Med, 3) High *vs*. High (Short), and 4) High *vs*. High (Long). Using the online R statistics software in [28], the bagplots are computed for each cluster, in an effort to perform an in-depth correlation study of the experimental results between short-range contacts and most similar local subsequences at different ranges of global similarity. Although the available samples at the best settings are found to be considerably few, the global sequence information does appear to affect the method performance, as shown in Figure 8. For example, in Figure 8(a), even at the best settings, the centre of gravity of the bag is fairly low (around ~62% for contact map similarity) in the case of low global similarity (15-30%). As for the rest of plots, the center of gravity is higher and remains almost the same (around 80% for contact map similarity), when global sequence similarity is medium and high.

V. CONCLUSION AND FUTURE WORK

The paper proposes a pattern matching approach that incorporates both protein sequence and evolutionary information, with the goal of locating similar substructural patterns between contact map pairs. These patterns could ultimately be used as building blocks for a computational bottom-up approach to protein structure prediction from contact maps [9]. A standard benchmark dataset of carefully-selected 40 large protein domains (Skolnick dataset) is adopted for this study as the experimental dataset.

To the best of our knowledge, this is the first-of-its-kind study to utilize sequence and evolutionary information in locating similar contact map patterns, with no comparable state-of-the-art results. The paper provides an extensive analysis for the three different factors believed to affect the performance of short-range pattern matching in the diagonal area, in particular, 1) local sequence information, 2) evolutionary information, and 3) global sequence

 ISBN: 978-1-61208-137-3

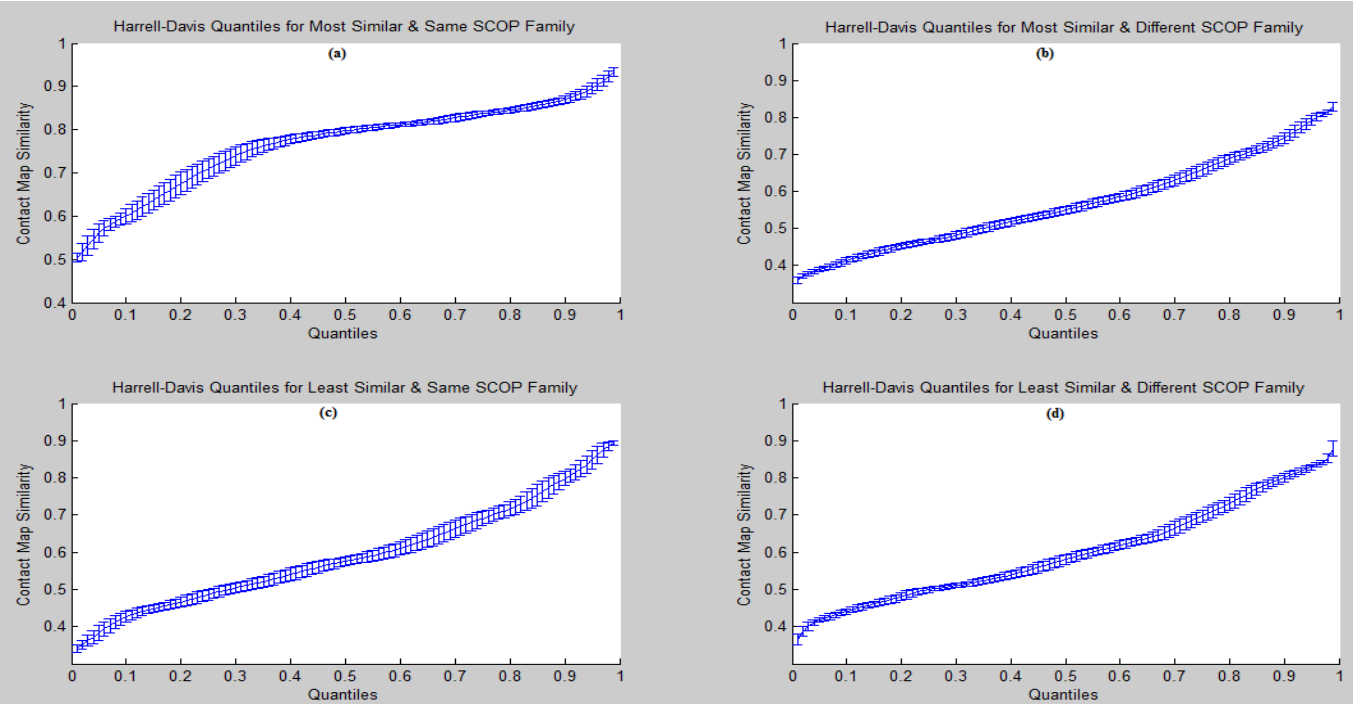


Figure 7. Harrell-Davis quantiles for different categories of the results, along with the error bars of the associated standard error for each reported quantile. (a) Shows the first category of the results of pairs of protein subsequences that are most similar and of the same protein class. (b) Shows category 2 of pairs of protein subsequences that are most similar and of the different protein class. (c) Shows category 3 for pairs of protein subsequences that are least similar and of the same protein class. (d) Shows the last category of pairs of protein subsequences that are least similar and of the different protein class.

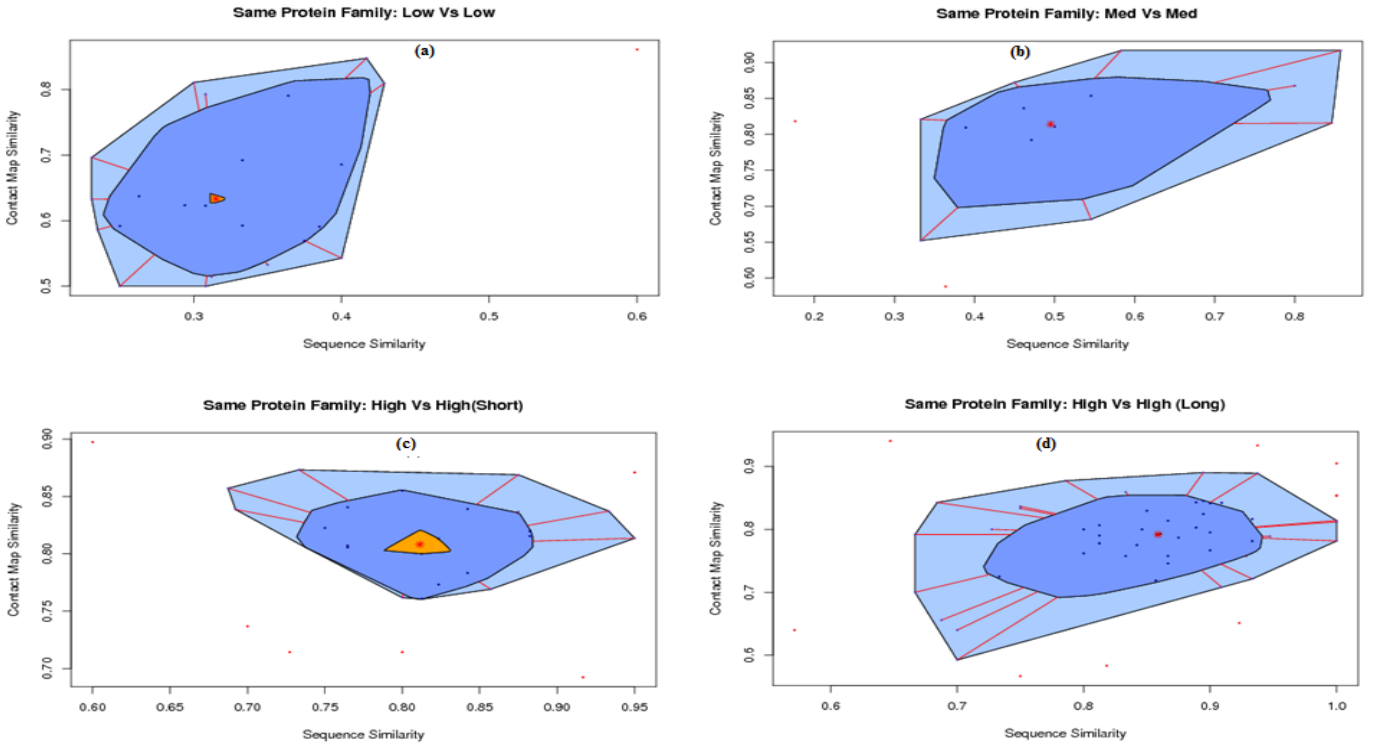


Figure 8. Bagplots for different clusters of the pair-wise results of most similar local subsequences and short-range contacts. (a) Shows the results of first cluster of pairs of protein sequences that are of low global sequence similarity (15-30%). (b) Shows the results of pairs of protein sequences that are of medium global sequence similarity (7 – 70%). (c) Shows the results of pairs of protein sequences that are of high global sequence similarity (35 – 90%) and short length (99 residues). (d) Shows the results of pairs of sequences that are of high global sequence similarity (30-90%) and long length (250 residues).

 ISBN: 978-1-61208-137-3

information. Firstly, for local sequence information, high sequence similarity (above 60%) has demonstrated (using a scatter-plot analysis) to be a good indicator of a corresponding high diagonal contact map similarity (around 70-90%). This correlation, however, does not appear to be suitable when contacts are long-range (i.e., in the off-diagonal areas of contact maps), or when local sequence similarity is low (less than 60%). Secondly, for evolutionary information, the results proved (using a quantile-based analysis) to be considerably higher when protein pairs have a clear evolutionary relationship, i.e. when they are of the same SCOP family. Lastly, for global sequence information, the results are observed (using a bagplot analysis) to be superior when the global sequence similarity is not low (more than 30%).

Possible future work to improve pattern matching in the diagonal area would be to perform a dynamic expandable multi-regional analysis of contact maps to reduce any possibility of region displacement. That is, we may consider looking further in the neighborhood of the corresponding regions of similar local subsequences. As for the off-diagonal areas, alternative approaches could be employed instead of sequence and evolutionary information that both did not appear helpful in these areas. We are currently looking into exploring *Swarm Intelligence* techniques [32] as a promising way to tackle the problem in the off-diagonal areas of contact maps, where the most uncertain, yet important, long-range contacts exist.

REFERENCES

[1] F. Collins, M. Morgan, and A. Patrinos, "The human genome project: lessons from large-scale biology," *Science*, vol. 300, 2003, pp. 286–290.

[2] R. D. Schaeffer and V. Daggett, "Protein folds and protein folding," *Protein Engineering, Design and Selection*, Vol. 24, no. 1-2, 2010, pp. 11-19.

[3] A. C. Anderson, "The process of structure-based drug design," *Chemistry and Biology*, vol. 10, 2003, pp. 787–797.

[4] J. Drenth, "Principles of protein X-ray crystallography," *Springer-Verlag*, New York, 1999, ISBN 0-387-98587-5.

[5] M. Schneider, X. R. Fu, and A. E. Keating, "X-ray versus NMR structures as templates for computational protein design," *Proteins*, vol. 77, no. 1, 2009, pp. 97–110.

[6] A. Kolinski (Ed.), "Multiscale approaches to protein modeling," 1st Edition, Chapter 10, *Springer*, 2011, ISBN 978-1-4419-6888-3.

[7] P. R. Daga, R. Y. Patel, and R. J. Doerksen, "Template-based protein modeling: recent methodological advances," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, 2010 , pp. 84-94.

[8] C. Yang et al., "Improved side-chain modeling by coupling clash-detection guided iterative search with rotamer relaxation," *Bioinformatics*, 2011, doi:10.1093/bioinformatics/btr00.

[9] J. Glasgow, T. Kuo, and J. Davies, "Protein structure from contact maps: a case-based reasoning approach," *Information Systems Frontiers*, Special Issue on Knowledge Discovery in High-Throughput Biological Domains, Springer, vol. 8, no. 1, 2006, pp. 29-36.

[10] I. Walsh, A. Vullo, and G. Pollastri, "XXStout: improving the prediction of long range residue contacts," *ISMB 2006*, Fortaleza, Brazil.

[11] M. Vassura et al., "Reconstruction of 3D structures from protein contact maps," Proceedings of 3rd International Symposium on Bioinformatics Research and Applications, Berlin, Springer, vol. 4463, 2007, pp. 578–589.

[12] X. Yuan and C. Bystroff, "Protein contact map prediction," *in Computational Methods for Protein Structure Prediction and Modeling*, Springer, 2007, pp. 255-277, doi:10.1007/978-0-387-68372-0_8.

[13] E. Krissinel, "On the relationship between sequence and structure similarities in proteomics," *Bioinformatics*, vol. 23, 2007, pp. 717–723.

[14] Dictionary of secondary structure of proteins: available at http://swift.cmbi.ru.nl/gv/dssp/, 14.03.2011.

[15] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: a structural classification of proteins database for the investigation of sequences and structures," *J. Mol. Biol.*, vol. 247, 1995, pp. 536–540.

[16] G. Lancia, R. Carr, B. Walenz, and S. Istrail, "101 Optimal PDB structure alignments: A branch-and-cut algorithm for the maximum contact map overlap problem," *Proceedings of Annual International Conference on Computational Biology (RECOMB)*, 2001, pp. 193-202.

[17] W. Xie and N. V. Sahinidis, "A branch-and-reduce algorithm for the contact map overlap problem," *Proceedings of RECOMB of Lecture Notes in Bioinformatics*, Springer, vol. 3909, 2006, pp. 516-529.

[18] P. Lena, P. Fariselli, L. Margara, M. Vassura, and R. Casadio, "Fast overlapping of protein contact maps by alignment of eigenvectors," *Bioinformatics*, vol. 26, no. 18, 2010, pp. 2250-2258. doi: 10.1093

[19] H. Xiaoquin and W. Miller, "A time-efficient, linear-space local similarity algorithm," *Advances in Applied Mathematics*, vol. 12, 1991, pp. 337-357.

[20] SIM: Alignment Tool for Protein Sequences, available at http://ca.expasy.org/tools/sim-prot.html, 14.03.2011.

[21] V. Arjunan, S. Nanda, S. Deris, and M. Illias, "Literature survey of protein secondary structure prediction," *Journal Teknologi*, vol. 34, 2001, pp. 63-72.

[22] Y. Xu, D. Xu, and J. Liang (Eds.), "Computational methods for protein structure and modeling," *Springer*, Berlin, 2007, ISBN: 978-1-4419-2206-9

[23] Henikoff and Henikoff, "Amino acid substitution matrices from protein blocks," *Proceedings of the national academy of sciences*, USA, vol. 89, 1992, pp. 10915-10919.

[24] S. F. Altschul and B. W. Erickson, "Optimal sequence alignment using affine gap costs," *Bull. Math. Biol*., vol. 48, 1986, pp. 603-616.

[25] L. Lee, "Measures of distributional similarity," *Proceedings of the 37th annual meeting of ACL*, 1999, pp. 25–32.

[26] H. R. Ahmed and J. I. Glasgow, "Multi-regional analysis of contact maps towards locating common substructural patterns of proteins," *J Communications of SIWN*, vol 6, 2009, pp.90-98.

[27] P. Wessa, "Harrell-Davis quantile estimator", *in Free Statistics Software*, Office for Research Development and Education, 2007, URL: http://www.wessa.net/rwasp_harrell_davies.wasp/, 14.03.2011.

[28] P. Wessa, "Bagplot," *in Free Statistics Software*, Office for Research Development and Education, 2009, URL: http://www.wessa.net/rwasp_bagplot.wasp/, 14.03.2011.

[29] F. E. Harrell and C. E. Davis, "A new distribution-free quantile estimator," *Biometrika*, vol. 69, 1982, pp. 635-640.

[30] P. J. Rousseeuw, I. Ruts, and J. W. Tukey, "The bagplot: A bivariate boxplot," *The American Statistician*, vol. 53, 1999, pp. 382–387.

[31] D. F. Williamson, R. A. Parker, and J. S. Kendrick, "The box plot: a simple visual method to interpret data," *Ann Intern Med*, vol. 110, 1989, pp. 916-921.

[32] S. Das, A. Abraham and A. Konar, "Swarm Intelligence Algorithms in Bioinformatics," *Studies in Computational Intelligence*. vol. 94, 2008, pp. 113–147.

 ISBN: 978-1-61208-137-3

Methodology to Explore Co-expression in Microarray Data

Bertrand De Meulder, Eric Bareke, Michael Pierre, Sophie Depiereux, Eric Depiereux
Bioinformatics & Biostatistics lab, URBM
University of Namur
Namur, Belgium
bertrand.demeulder@fundp.ac.be

Abstract - In the past several years, the amount of microarray data accessible on the Internet has grown dramatically, representing millions of Euros worth of underused information. We propose a method to use this data in a coexpression study. The method is simple in principle: the aim is to detect which genes react in the same way in certain circumstances (such as a disease, stress, medication,), potentially highlighting new interaction partners or even new pathways. We propose to study coexpression using a large amount of data, process it through an adequate algorithm and visualize the results with a dynamic graphical representation. We gather the microarray data using the PathEx database developed in our lab, which allows searching through more than 120,000 microarrays experiments on *Homo sapiens* using specific criteria such as the tissue sample, the biological background or any information contained in the metadata describing the experiment. Then, we process the data using the Minet R package, which allows for coexpression analysis using cutting-edge algorithms such as ARACNE or MRNET methods. This step computes the weighted relations between all the probesets in the microarrays and provides a GraphML representation of the relations. In order to explore the relations optimally, we channel the GraphML into a dynamic graphical program we developed called gViz. This program allows for data visualization but also for exploration and post-analysis. We can extract meaningful information from the network computed, compare this information with curated databases such as KEGG (Kyoto Encyclopedia of Genes and Genomes), highlight the discrepancies and hopefully discover new interactions or add new steps in canonic pathways. We present here a fast, free and user-friendly working methodology to analyze co-expression in microarray data

Keywords: Co-expression; microarray; methodology

I. INTRODUCTION

Co-expression analysis in gene networks in an infant domain in the teenager field of network biology. It is only a few years ago that the technology, knowledge, and data mass mandatory for these analysis has been made available to researchers. Co-expression analysis is the study of the similarities in changes of genes expressions in various circumstances (such as diseases). Using this technique, researchers aim to discover new relationships between known genes, new partners in known pathways or even entirely new pathways. This holds promises for new insights into complex biological states, such as cancer or degenerative diseases, as well as further comprehension of the cell fine machinery. Ultimately, this approach could provide a compendium of gene interactions maps in an ever-growing array of cell states.

Currently, there lies in databases millions of Euros worth of underused microarray data, since researchers conducting those experiments usually focus only on a small number of genes among the thousands available on the chip. We propose a way to make use of this data mass, by studying co-expression relations between all genes represented on the chip, using state of the art processing algorithms and an adapted graphical interface for exploration. With this we hope to predict new interactions, which we will then try to confirm using wet-lab analysis.

Our approach implies the possession of several elements: a large deposit of microarray data, if possible in database form; an algorithm to process this data efficiently; a graphical interface to browse the map of interactions and an external verification database for the validations.

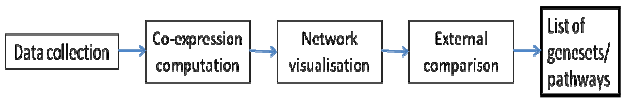


Figure 1. General layout of the co-expression analysis

Each step of this general layout will be developed hereunder. We first discuss about data collection then present the different steps of the co-expression computation process. Section IV is about the visualization solutions, while Sections V and VI present the validations and conclusions respectively.

II. DATA COLLECTION

The main microarray deposits are the well known Gene Expression Omnibus (GEO) [1] and ArrayExpress [2] databases. The data collection, although laborious and time-consuming, could be done directly from these websites. However the construction of the websites does not allow for advanced querying on the biological parameters of the experiment. This aspect is crucial: indeed, the relations we potentially will highlight in the end of the analysis are related to the biological state of the cells on which the microarray experiment was done. In other words, all the information we will extract from the analysis has to be already buried in the microarray experiment we select at this step.

 ISBN: 978-1-61208-137-3

To help the process of selection, we developed a database with the data included in both GEO and ArrayExpress and we included all the information in the description field manually into a database, thus making high-level queries possible. This database – PathEx [3] – allows us to query all the microarray experiments on criterions such as disease, cell lines, age of the patient, organ studied, etc., thus allowing for a much more precise choosing of data.

III. CO-EXPRESSION COMPUTATION

During this part of the methodology, gene expression data from the microarrays selected in the previous step are processed to generate a representation of the co-expressions between those genes. This processing can be done by several algorithms such as the R packages nem, qpgraph or GeneNet [4 , 5 , 6]. We chose to use the R package MINET to process our data based on its speed, ease of use, large choice of methods at each steps and possibility of parameterization. This package was developed by Patrick Meyer in the Machine Learning Group at the University of Bruxelles (ULB) [7]. MINET computes the weighted relations between every probeset in the microarray; it takes as input the preprocessed microarray data and outputs a GraphML representation of the interactions.

The Mutual Information networks are a subcategory of inference methods. In those networks, a link is set if it exhibits a high score based on pairwise mutual information. One of the advantages of these methods is the low computational complexity. This is due to the fact that $n(n-1)/2$ calls of mutual information, based on bivariate probability distributions, are required to compute the mutual information matrix. Since each estimation of a bivariate distribution can be done quickly and does not require a large number of samples, this method is ideal to analyze microarray data [8].

The MINET package consists of three successive steps: Discretization, Mutual Information computation and Network Inference.

A. Discretization

This step is mandatory for the next computing step. However, it is known that there is an inevitable loss of information when discretizing continuous data. To minimize this loss, two discretization algorithms are implemented: *equalWidth* and *equalFreq*. The principle of the first is to divide the interval [a, b] into [X_i] intervals of same size, while the principle of the latter is to divide [a, b] into [X_i] intervals, each having the same number of data points. The number of bins is also important, as it controls the ratio between the bias and the variance. Practically, if m is the number samples, it is considered that a number of bins equal to the square root of m is a fair trade-off between bias and variance [9].

B. Mutual Information Matrix (MIM) computation

This step consists in the computing, between all pairs of genes present in the dataset, of the mutual information, i.e. the similarity in the behaviors of the genes. This step is very tricky, often biased but is crucial for the good results of the whole procedure. It is not surprising that there exist many alternative algorithms for this step. Without going into too many details, here is an overview of the MIM computation algorithms available in the MINET package:

1) General formulation

The MIM computation requires the computation of a square matrix whose m_{ij} element is given by

$$mim_{ij} = I(X_i; X_j) \quad (1)$$

where $I(X_i; X_j)$ is the Mutual Information between variable X_i and X_j.

The difference between the methods lies in the computing of this term $I(X_i; X_j)$ Mutual Information computation requires the determination of three entropy terms:

$$I(X_i; X_j) = H(X_i) + H(X_j) - H(X_i; X_j) \quad (2)$$

where $H(X)$ is the entropy of the variable X.

Entropy has to be estimated and an effective and fast entropy estimator is essential. The reduction of the bias inherent to the entropy estimation has gained much interest over the last years and most approaches have focus on minimizing this bias. However, in the case of microarray analysis, the reduction of the bias should not be the only criterion, as computational complexity/speed should also be minimized. To save space, we only discuss the *Shrink* and the *Schurmann-Grassberger* estimators.

2) Shrink Estimator

The rationale behind this algorithm is to combine two different estimators: one with low bias and one with low variance, by use of a weighting factor λ *[0, 1]*. The general formulation is the following [7]:

$$\hat{p}_\lambda(x) = \lambda \frac{1}{|\chi|} + (1-\lambda)\frac{\#(x)}{m} \quad (3)$$

where λ is the weighting factor [0, 1], $|\chi|$ is the number of non null bins, #(x) is the number of data points having the value x and m is the number of samples.

The entropy can then be estimated with:

 ISBN: 978-1-61208-137-3

$$\hat{H}^{shrink}(X) = -\sum_{x \in \chi} \hat{p}_{\lambda^*}(x) \log \hat{p}_{\lambda^*}(x) \quad (4)$$

where H is the entropy and λ^* is the value of λ minimizing the mean square error [10] [8].

3) The Schurmann-Grassberger estimator

It is a Bayesian estimator which assumes the sample distribution follow a Dirichlet distribution. A Dirichlet distribution is the generalization of the Beta distribution [11]. The density of this distribution is described by [8]

$$p(X;\Theta) = \frac{\prod_{i \in \{1,2,\ldots|\chi|\}} \Gamma(\Theta_i)}{\Gamma(\sum_{i \in \{1,2,\ldots|\chi|\}} \Theta_i)} \prod_{i \in \{1,2,\ldots|\chi|\}} x_i^{\Theta_i - 1} \quad (5)$$

where θ_i is the prior probability of an event x_i, x_i being the ith element of the set χ and Γ (.) is the gamma function.

The entropy can then be estimated by

$$\hat{H}^{dir}(X) = \frac{1}{m + |\chi| N} \sum_{x \in \chi} (\#(x) + N) (\psi(m + |\chi| N + 1) - \psi(\#(x) + N + 1)) \quad (6)$$

where $\psi(z) = \frac{d \ln \Gamma(z)}{dz}$ is the digamma function, N is a weighting factor. Various choices of parameters for this factor N have been proposed [12 , 13].

C. Network inference

Once the MIM computation is done, the network inference step can take place. This step is essentially a translation of the Mutual Information links computed at the previous step into a graph of the probable relations between the variables. In the case of microarray data the nodes in the graph represent probesets and the arcs represent the regulator/regulated relations between them. Various network inference methods are available in the MINET package: *Relevance network, CLR, ARACNE* and *MRNET* algorithms [14-16]. We will only discuss the *MRNET* method, to save space.

MRNET is based on the Maximum Relevance / Minimum Redundancy (MRMR) rationale [7]. Simply put, if we consider a set of genes (*X*) and a target gene *Y*, the algorithm will first select the gene X_i with the highest mutual score to variable *Y*. Then, the next selected, X_j, will be the one with a high $I(X_j; Y)$ score (maximum relevance), and at the same time a low $I(X_i; X_j)$ (minimum redundancy). At each step, the algorithm is thus expected to select the variables with and efficient trade-off between relevance and redundancy, for every gene X_i in the set of X genes. A selection based on a score above a certain threshold $I(X_i; X_j) < \theta$ is performed in both directions: for two genes X_i and X_j, there will be an edge if X_i is a well predictor of X_j ($s_i > \theta$) or if X_j is a well predictor of X_i ($s_j > \theta$). The complexity of this methods lies between $O(n^2)$ and $O(n^3)$.

D. Methods selection

Following MINET author's recommendation based on validations on external datasets, we use the following methods to analyze our data: *equalFreq* discretization, the *Shrink* estimator if the number of replicates is low and the *Schurmann-Grassberger* estimator otherwise and the *MRNET* algorithm [7]. Our figure 1 then becomes:

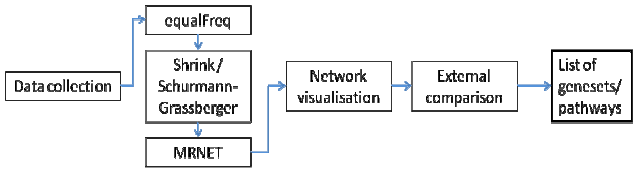


Figure 2. Layout of the analysis with choice of methods in MINET

IV. NETWORK VISUALIZATION

The easiest way to explore the results found at the previous step is to use a graphical representation of the network. Although there are several softwares available for the task [17-19], we were not satisfied either with the functionalities proposed or with the interactions possibilities of said softwares. We decided to develop our visualization software, which we called gViz. An application note describing it has been submitted to Bioinformatics on January 22nd [20].

gViz takes as input the network computed by MINET in GraphML format. It can then translate the probeset IDs into a wide range of mainstream identifiers: Entrez gene, Ensembl, UniGene or KEGG IDs. The advantage of gViz over the other network softwares is that it can be used to visualize specific parts of the network. The user can select one of several identifiers in the left panel (see fig 3) and display the network containing only the relations it wants to focus on. The network displayed in gViz is dynamic and interactive. The user can choose to display the whole network (although it can be resources consuming) or specific parts of it. When clicking on a node, the user can highlight as well the neighbors of the node, with an adjustable deepness. gViz also has a feature capable of filtering the entire network based on the MRMR score given by MINET or by nodes degree (i.e. the number of neighbors) or also based on annotation criteria (involved in same biological process). The user can at any time adjust the value of the exclusion threshold.

Several layout algorithms are available. Other 'visual' features allow to display the thickness of the nodes (representing the degree of said node) and of the edges (representing the MINET score for said edge).

 ISBN: 978-1-61208-137-3

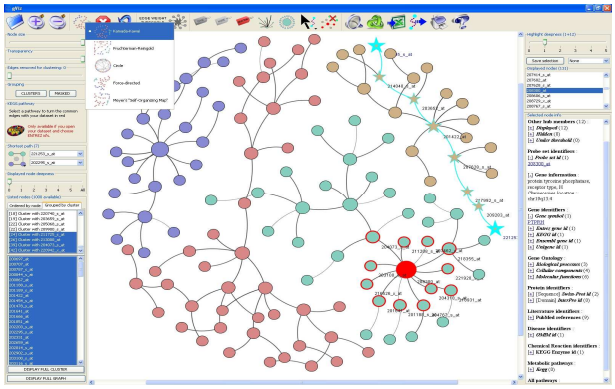


Figure 3. General layout of the gViz interface

V. VALIDATIONS

This part is still in progress. However, we have a clear picture of the three validations that should be done. First we will perform a consistency validation, to ensure that our method can retrieve known interactions, and a coverage validation to estimates the method's ability to retrieve all data present in the original dataset. Once those steps are done, we will be able to make predictions by comparing our results with an external database, such as KEGG [21], highlight discrepancies and test the corresponding genes in wet-lab analysis, therefore biologically validating our approach. These experiments will be performed in the course of 2011.

VI. CONCLUSION

We presented a method to analyze co-expression in microarray data, with the help of a large data mass, cutting-edge algorithms and suitable visualization solution. In a short time we will finish the first validations. We hope to spot new interactions which we will explore further in wet-lab analysis. We will then have produced a reliable, fast, cheap and user-friendly way for researchers to analyze their microarray data prior to the wet-lab. In the near future, we will apply this methodology to try and discover new genes or interactions involved in the metastatic transformation of primary cancer cell and eventually provide new targets for cancer treatments.

ACKNOWLEDGEMENT

B. D.M. thanks Patrick Meyer and Raphaël Helaers for help and technical ideas. B. D.M. benefits from a FRS-FNRS Télévie Grant n° FC 81726.

REFERENCES

1. T. Barrett et al., *NCBI GEO: mining tens of millions of expression profiles--database and tools update.* Nucleic Acids Res, 2007. **35**(Database issue): pp. D760-765.
2. H. Parkinson et al., *ArrayExpress--a public database of microarray experiments and gene expression profiles.* Nucleic Acids Res, 2007. **35**(Database issue): pp. D747-750.
3. E. Bareke et al., *PathEx: a novel multi factors based datasets selector web tool.* BMC Bioinformatics, 2010. **11**: pp. 528-537.
4. H. Frohlich, M. Fellmann, H. Sultmann, A. Poustka, and T. Beissbarth, *Estimating large-scale signaling networks through nested effect models with intervention effects from microarray data.* Bioinformatics, 2008. **24**(22): pp. 2650-2656.
5. R. Castelo and A. Roverato, *Reverse engineering molecular regulatory networks from microarray data with qp-graphs.* J Comput Biol, 2009. **16**(2): pp. 213-227.
6. R. Opgen-Rhein and K. Strimmer, *From correlation to causation networks: a simple approximate learning algorithm and its application to high-dimensional plant gene expression data.* BMC Syst Biol, 2007. **1**: p. 37. Last access date: 22 March 2011.
7. P.E. Meyer, F. Lafitte, and G. Bontempi, *minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information.* BMC Bioinformatics, 2008. **9**: p. 461. Last access date: 22 March 2011.
8. P.E. Meyer, *PhD Thesis.* 2008. Available from http://www.ulb.ac.be/di/map/pmeyer. Last access date: 22 March 2011.
9. Y. Yang and G. Webb, *Discretization for naive-bayes learning: managing discretization bias and variance*, in *Technical report*, S.o.C.S.a.S. Engineering, Editor. 2003, Monash University.
10. J. Hausser, *Improving entropy estimation and inferring genetic regulatory networks.* , in *National Institute of Applied Sciences.* 2006: Lyon. Available from http://jean.hausser.org/site/64. Last access date: 22 March 2011.
11. T. Schurmann and P. Grassberger, *Entropy estimation of symbol sequences.* Chaos, 1996. **6**(3): pp. 414-427.
12. N. Beerenwinkel et al., *Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype.* Proc Natl Acad Sci U S A, 2002. **99**(12): pp. 8271-8276.
13. L. Wu, P. Neskovic, E. Reyes, E. Festa, and W. Heindel, *Classifying nback eeg data using entropy and mutual information features.* in *European symposium on Artificial Neural Networks*. 2007.
14. A.J. Butte and I.S. Kohane, *Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements.* Pac Symp Biocomput, 2000: pp. 418-429.
15. J.J. Faith et al., *Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles.* PLoS Biol, 2007. **5**(1): p. e8. Last access date: 22 March 2011.
16. A.A Margolin et al., *ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context.* BMC Bioinformatics, 2006. **7 Suppl 1**: p. S7. Last access date: 22 March 2011.

 ISBN: 978-1-61208-137-3

17. *yEd- Graph Editor*. 2010 Available from: http://www.yworks.com/en/products_yed_about.html. Last access date: 22 March 2011.
18. N. Salomonis et al., *GenMAPP 2: new features and resources for pathway analysis.* BMC Bioinformatics, 2007. **8**: p. 217. Last access date: 22 March 2011.
19. P. Shannon et al., *Cytoscape: a software environment for integrated models of biomolecular interaction networks.* Genome Res, 2003. **13**(11): pp. 2498-2504.
20. R. Helaers et al., *gViz - A novel co-expression network visualization tool*., unpublished.
21. H. Ogata, S. Goto, K. Sato, W. Fujibichi, H. Bono, and M. Kanehisa, *KEGG: Kyoto Encyclopedia of Genes and Genomes.* Nucleic Acids Res, 1999. **27**(1): pp. 29-34.

 ISBN: 978-1-61208-137-3

System Biology on Mitochondrion Genomes

Michael G.Sadovsky
Institute of computational modelling SB RAS
660036 Krasnoyarsk, Russia
msad@icm.krasn.ru

Natalia A. Zaitseva
Siberian Federal univeristy
Svobodny prosp., 79
660041 Krasnoyarsk, Russia
zaiceva-5@g-service.ru

Yulia A. Putintseva
Siberian Federal univeristy
Svobodny prosp., 79
660041 Krasnoyarsk, Russia
kinommanka5@mail.ru

***Abstract*—Relations between the triplet composition of mitochondria genomes, and the phylogeny of their bearers is considered. It is shown that the genomes are split into several classes in the space of information values of the triplets. The classification exhibits a feasible correlation to the phylogeny attribution of the genomes. The stability of the classification, as well as the impact of various techniques of a data pretreatment is analyzed. A strong and fruitful correlation between the structure of trinucleotide composition of mitochondrion genomes, and the taxonomy of the bearers of these genomes is proven.**

***Keywords*-frequency, entropy, mutual entropy, order, phylogeny, elastic map, knowledge retrieval**

I. INTRODUCTION

A study of statistical properties of nucleotide sequences may bring a lot towards the relation between structure and function encoded in these former. A consistent and comprehensive investigation of the features and peculiarities is based on the study of frequency dictionary of a nucleotide sequence [1], [2], [3]. Such approach answers the questions concerning the statistical and information properties of DNA sequences. A frequency dictionary, whatever one understands for it, is rather multidimensional entity.

In particular, a relation between a structure (i. e., oligonucleotide composition and their frequency), and the taxonomy of the bearers of DNA sequences is of great importance. Here we studied this relation for the set of mitochondrion genomes. They exhibit a significant violation of the second Chargaff's rule. Such violation may provide another opportunity for knowledge retrieval from the statistical properties of them [7], [8].

Consider a continuous symbol sequence from four-letter alphabet $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{T}\}$ of the length N. No other symbols or gaps in a sequence are supposed to take place. Any coherent string $\omega = \nu_1\nu_2 \dots \nu_q$ of the length q is a word. A set of all the words occurred within a sequence makes the support of that latter. Counting the numbers of copies n_ω of the words, one gets a finite dictionary; changing the numbers for the frequency

$$f_\omega = \frac{n_\omega}{N}$$

one gets the frequency dictionary W_q of the thickness q. This is the main object of our study.

Further, we shall study the triplet composition only, i. e., consider the frequency dictionaries W_3. Thus, any genome is represented as a point in 63-dimensional space. What is the pattern of the distribution of genomes in that space, and whether the distribution exhibits a correlation to a phylogeny of the genome bearers are two key questions of our study. Reciprocally, all genomes are known for a symmetry: frequencies of a couple of string composing a complimentary palindrome[1] are pretty close. This proximity is not an absolute equivalence, and the deviation varies for various genomes; mitochondria are well known to be the most variable, from that point of view [1], [4], [5].

To address these questions, we have implemented an unsupervised classification of the mitochondrion genomes, in various spaces of frequencies (or information values) of triplets. Then, the taxa composition of the classes developed due to the classification has been studied; a considerable correlation between taxa composition, and the class occupation was found. Some results of the study of the correlation of the distribution of bacterial taxa in the information value space, developed over 16S RNA are presented in [4], [5].

To study the effects of a structure peculiarities of a genome expressed in the violation of the symmetry mentioned above (for W_3, we developed similar classifications in 32-dimensional symmetrized spaces; the former is the space of differences of the frequencies of two triplets composing a complementary palindrome, and the latter is the space of differences of information values of two triplets composing a complementary palindrome.

The paper presents the evidences of the strong relation between the structure of mitochondrion genomes, and the taxonomy of their bearers. Section II describes the source of the genetic data, Section III provides a short description of the techniques of the classification and knowledge retrieval. The results of the study are present at Section IV, where the subsection IV-A provides the results of the study of the impact of a symmetry violation towards the relation between structure and taxonomy, and the subsection IV-B shows the results proving the high level of interrelation between the

[1]For example, the couples ATC $\leftrightarrow$ GAT and GGCAATC $\leftrightarrow$ GATTGCC are the complementary palindromes. One must bear in mind, that such entities are determined over a single strand!

 ISBN: 978-1-61208-007-9

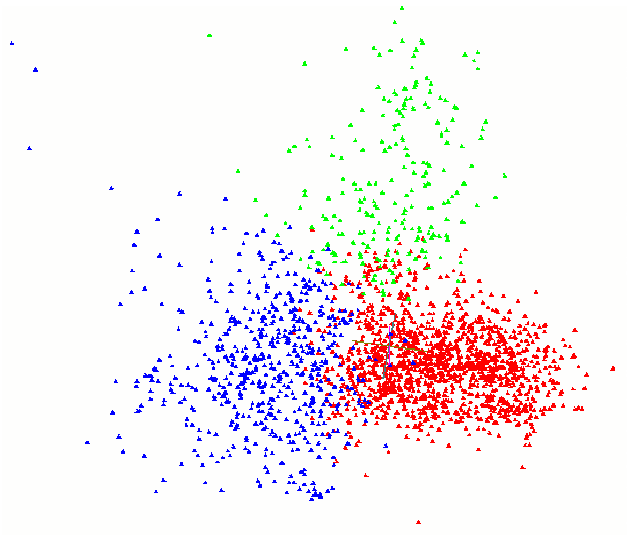


Figure 1. Unsupervised classification developed in 63-dimensional space of frequencies separates mitochondrion genomes into 3 classes; shown in principal components, raw database.

structure, and taxonomy. Finally, the biological issues, as well as some more mathematically oriented questions are discussed in Section V.

II. Databases and methods

Mitochondrion genomes were retrieved from EMBL–bank. The the list of genomes available at EMBL–bank is inhomogeneous, from the point of view of the equity of the number of species of various genders enlisted into the database. The excessive number of closely related genomes (not speaking about the strains and variants) may yield a cluster of increased density[2] that overweights other points at the space, thus resulting in a distortion of the real pattern of a genome distribution at the space of information value of words.

To eliminate the effect of the possible bias described above, we hashed the databases: a single genome from a gender was selected randomly, while the other ones were eliminated from the database. It resulted in a decrease of the number of entries in the database up to 1651 ones.

Another problem in database structure results form an abundant set of entries representing taxonomically rather high clade solely: a single genome is deciphered in a clade. The scattered nature of the database conspires the effects of the correlation between taxonomy and statistical features of the genomes. Thus, we have also hashed the database, excluding the entries which are less than 50 taxons within a class, or an order; finally, 1132 entries was gathered into the database.

[2]And they do, in reality.

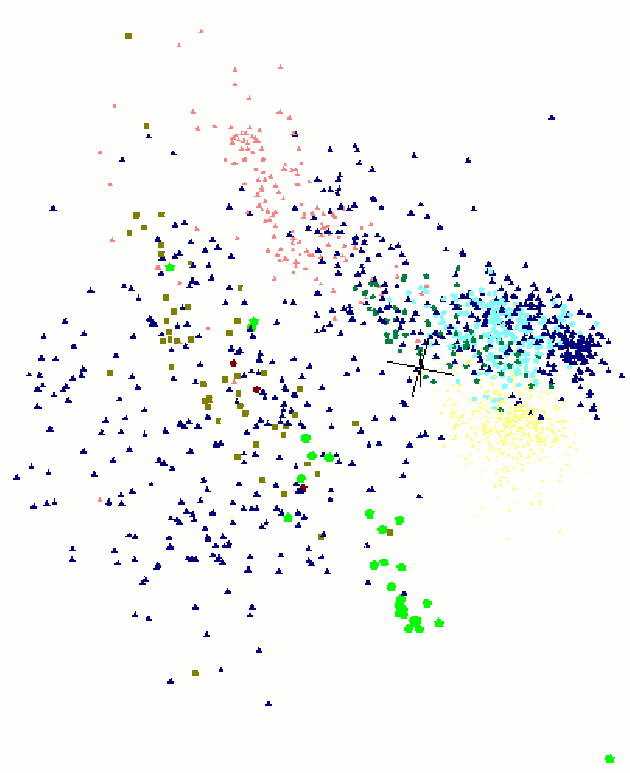


Figure 2. A distribution of seven clades in the space determined by principal components. See text for details.

III. Classification

A standard unsupervised classification technique was implemented to develop a classification of the genomes in the information value space. We used *ViDaExpert* software [9] to do that; the algorithm of the classification development see in [4], [5].

An unsupervised classification does not increases a number of classes: if no separability condition is verified, then the number remains the same. Separability of classes means that two classes are discrete with respect to a relation of a distance between their centroids, and their radii. An excess of a distance between two centroids over the sum of the radii of the relevant classes is the strongest separability condition. On the contrary, one faces the weakest separability condition, if the greater radius (among two classes to be checked for a separability) is not longer than the distance between two centroids of the classes.

We did not check formally the separability conditions, for the classification developed over the genomes. Meanwhile, a stability of the abundance of each class, and the maintenance of the class occupation were checked.

IV. Results and Discussion

Fig. 1 shows an example of an unsupervised classification of the mitochondrion genomes. This is a typical pattern of

 ISBN: 978-1-61208-007-9

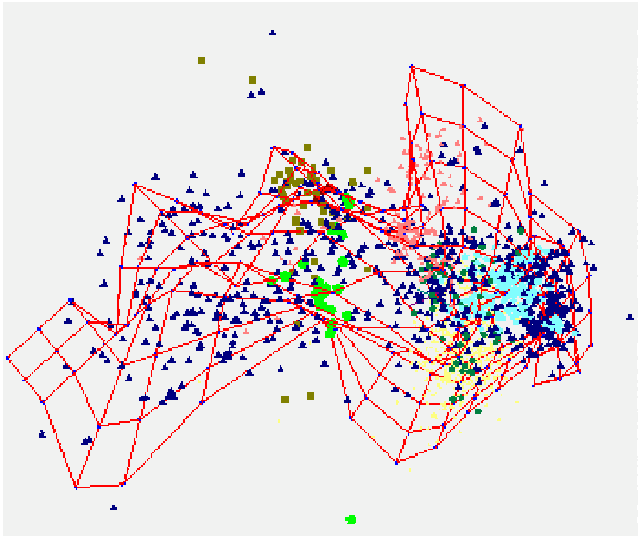


Figure 3. The data distribution around the elastic map (with high rigidity). Seven clades (see text for details) are shown in color.

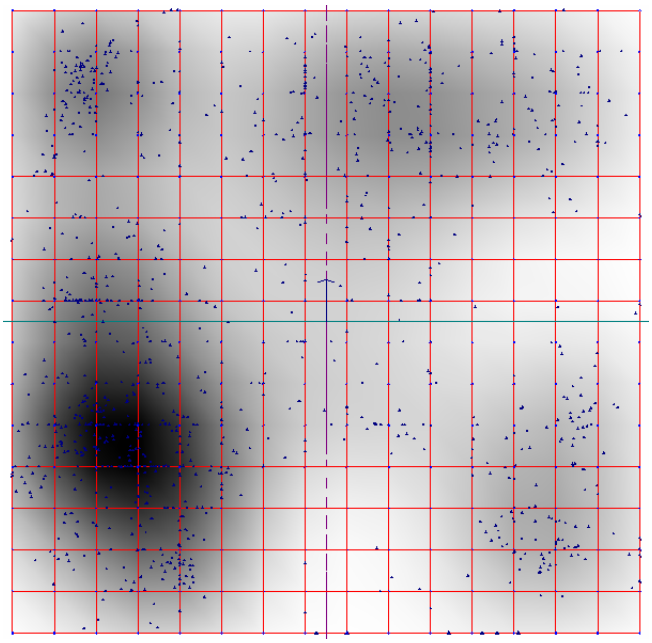


Figure 4. A distribution of 1132 genomes in 63-dimensional space of triplet frequencies. An averaged local density is shown in grey tone.

the separation of the genomes into three classes; namely, in a series of independent experiments with classification development, when a random initial distribution of the genomes takes place, the classification exhibits three outcomes: the genomes could gather into three classes, two classes, and all of them may occupy a single class (with very few exceptions).

Since the classification exhibits not so high instability in the separation of genomes into classes (the separation into

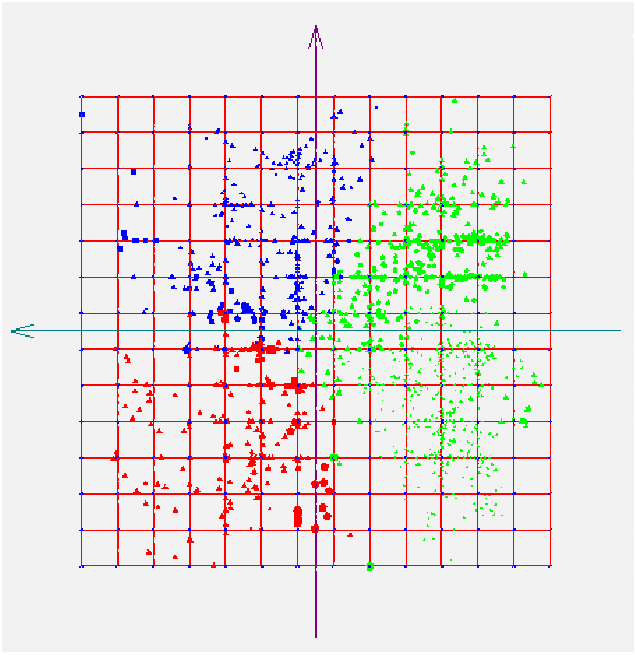


Figure 5. Unsupervised classification of genomes shown on the elastic map (inner coordinates).

three classes seems to be the most stable one), we checked it through a series of experiments. A series of independent classifications were developed, with independent and random redistribution of the genomes into the initial classes. We traced the preference in the class occupation by the same genome; indeed, the greatest majority of the genomes always occupy the same group, if any. Few genomes are quite volatile. They change the class occupation very often, with no regular pattern of the behaviour. Such genomes have been excluded from the database, total number of excluded entities was 41.

A distribution of the clades of various taxonomy level is of great interest; Fig. 2 shows the distribution of seven clades in the space determined by the principal components. This figure shows a distribution of seven (rather different in abundance) clades; moreover, these latter belong to different taxonomy levels. There are three clades of very high taxonomy level: *Fungi* (khaki, 50 entities), *Viridoplantae* (bright green, 31 entities), and *Rhodophyta* (reddish-brown, 3 entities). Four other clades are of lower taxa: fishes (*Actinopterygii*) are shown in yellow (507 entities), *Amphibia* are shown in green-blue (66 entities), and incests (*Neoptera*) are pink (143 entities). Finally, *Mammalia* are shown in light blue (212 entities). Small very dark-blue triangles show all other genomes.

To analyze the patterns of the interdependence of taxonomy and distribution in the space determined by the triplet composition of the genomes, we have implemented elastic map technique; Figure 3 shows such (rigid) map, and the

 ISBN: 978-1-61208-007-9

distribution of the genomes around it. Fig. 4 shows similar elastic map in the inner coordinates; it makes obvious a non-random pattern of the distribution of clades in the space. Moreover, the nonlinear statistical approach (i. e., elastic map implementation) identifies more than three classes: from seven to ten clusters are identified, depending on a rigidity of the map.

An unsupervised classification similar to that one shown in Fig. 1 developed for the distribution of clades projected to the elastic map presented in the inner coordinates is shown in Figure 4. A direct comparison of Figs. 4 and 5 clearly exhibits a character of the distribution of clades over these three classes obtained due to the unsupervised classification implementation. One should avoid a misconduct: the colors in these two figures have nothing to do each other; colors in Fig. 3 represent the clades, while the colors in Fig. 5 represent the classes obtained from the unsupervised classification implementation.

A. Symmetry and asymmetry impact on classification

Complementary palindromes exhibit close figure for the frequencies of the strings composing them. Similar (while less) proximity is observed for information values of the strings. This symmetry has been observed at the very beginning of the decipher of genetic sequences [7]. A violation of the second Chargaff's rule is the example of the asymmetry observed for W_1. Yet, very few attention is paid to the phenomenon of the symmetry [8].

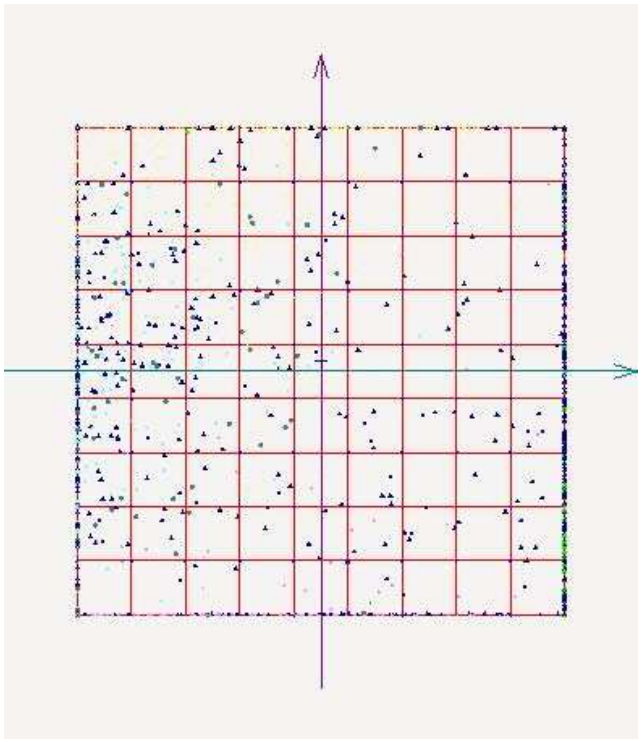


Figure 7. A distribution of 7 clades in 32-dimensional symmetrized space of information values. The distribution of shown on elastic map in internal coordinates.

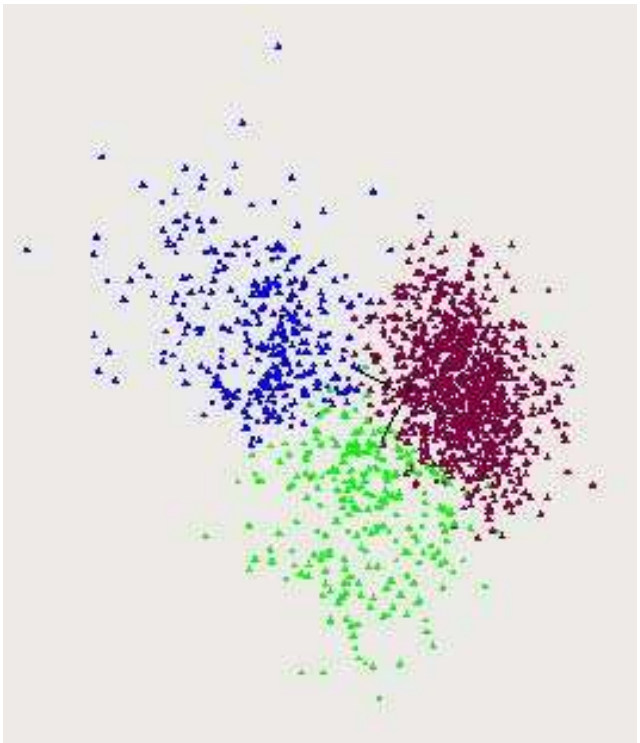


Figure 6. Unsupervised classification of genomes shown in 32-dimensional symmetrized space of information values.

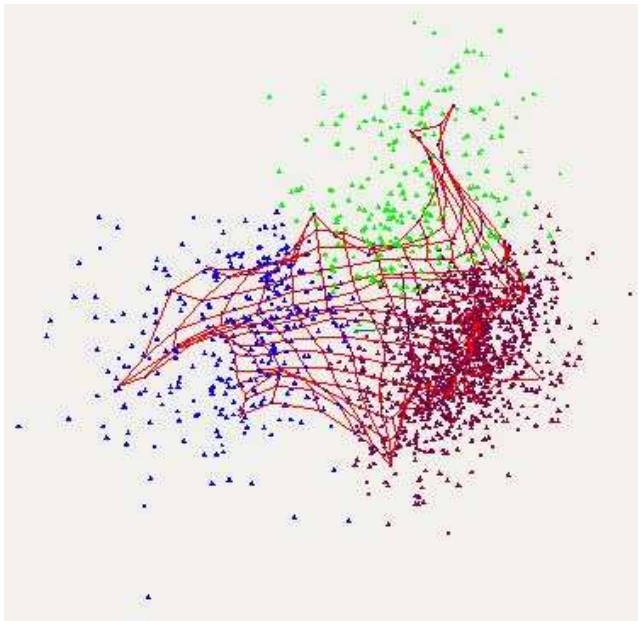


Figure 8. An unsupervised classification shown on elastic map in principal components.

 ISBN: 978-1-61208-007-9

Different genomes exhibit different level of the symmetry violation [8]; mitochondria are known for very high level of the violation. Such effect may bring additional knowledge towards the relation between structure (frequency dictionary) and taxonomy, or a function encoded at genetic entity. To test this idea, we have implemented a series of classifications both in direct, and symmetrized spaces determined by information values of triplets.

Figures 6, 8 and 7 show the distributions similar to those described above, while these latter were developed in the symmetrized 32-dimensional space of information values of triplets. It is evident, that the details of the shape of the distribution, especially the distribution of clades looks rather different. This difference means that the asymmetry in complementary triplets information values play some role in the statistical properties of mitochondrion genomes, while the details of that role are still conspired from a researcher, and await for further investigations.

B. Relation of classification and taxonomy

The relation between the classes obtained due to the separation of the genomes by various statistical techniques, and the taxonomic composition of those classes is the key issue of the paper. In general, the answer on this question is positive.

Fig. 2 shows that taxonomically close entities occupy an obviously isolated subspace, in total genomes distribution. Meanwhile, this figure does not prove an inverse statement. A direct comparison of the composition of the classes shows that there exists a strong correlation in the composition of a class determined statistically, and the taxonomy of its members.

The first class[3] contains genomes of *Actinopterygii* taxon: it contains 457 entries of that taxon, from 506 ones, totally. The second class contains *Neoptera* genomes (137 entries), 8 genomes of *Amphibia* and a genome of *Mammalia*. The third class exhibits the most complicated pattern. It bears 49 *Actinopterygii* genomes, 39 *Amphibia* genomes, the genomes of *Archosauria* and *Lepidosauria*, 97 and 87 entries, correspondingly. Also, this class incorporates almost all mammalian genomes (210 entries) and *Testudines* genomes: 24 entries (all of them are in the class). Besides, the class contains also 4 genomes of insects.

It should be said that the genomes of *Amphibia* are separated between two classes in rather non-random way: the occupation of a class is predominated by the genomes of the same taxon, of lower level. We checked, whether the classification developed over a single clade[4] yields similar pattern, and found that it works. By isolating the genomes of the same class (or other taxonomic level), and developing an unsupervised classification within that latter, one faces the similar behaviour of genetic entities: the statistically determined classes are predominantly occupied by the entities of the same taxon.

[3]Here we understand class as statistically determined cluster.

[4]That must be sufficiently abundant, of course.

V. Conclusion

We addressed the problem of the relation between a structure of DNA sequence, and the sense encoded in that latter. A triplet composition with information values of these latter was considered to be a structure. Genetically, the mitochondrion genomes are rather conserve, thus providing a good raw for knowledge extraction. The distribution of mitochondrion genes in 63-dimensional space of triplets frequency, 32-dimensional space of symmetrized frequencies, and in the relevant spaces determined by information values of triplets is far from a random one.

Statistically (i. e., with no external or additional knowledge, or assumptions), the considered genomes are gathered into several clusters. Linear analysis (unsupervised classification implementation) fails to discrete the genomes properly, providing three clusters, at best. An implementation of some techniques of the nonlinear statistical analysis improves the situation. Self-organized elastic maps clearly identify at least four clusters in the spaces mentioned above.

Reciprocally, the genomes from the came clade always occupy a single cluster; moreover, the clades yield obvious discrete groups within a cluster. It is very important, that the genomes from the same clade never share themselves among two (or more) clusters.

The patterns and relations standing behind the observed distribution of the clades over the clusters is the key issue; a the first glance, there is no simple order in the distribution that could be easily interpreted in the terms of classic systematics or traditional biology. Here are the following options in the studies of the relation between structure and taxonomy (or structure and function encoded in some genetic entities).

Comparative study of the distribution of clades (or functional groups) in frequencies vs. that one in information values: this study aimed to compare and figure out the differences in the composition of the clusters observed in two different spaces may reveal the role and significance of a symmetry observed in triplet composition;

Detailed study of the clade composition of various clusters: this study is the most essential. It is evident, that clades occupy some very peculiar areas in the spaces used to figure out a distribution of genomes. Moreover, there are no clade shared among two (or several) clusters.

Thus, the composition of those clusters is not random, or accidental. The point is that the very different clades occupy the same cluster. What makes them be close each other? Classical biologists consider them to be very far from; meanwhile, the statistics of their genomes seems to be pretty similar. A comprehensive analysis of the triplets that provide the proximity of very far species may reveal

 ISBN: 978-1-61208-007-9

the inner relations in the genetical entities observed in various taxonomy levels. Such relation does not disprove the classical taxonomy; it just provides another dimension for a study of genetic entities and their bearers;

Studying of distributions of the genomes in the space of longer oligonucleotides: this is an obvious expansion of the approach presented above. Here we presented some preliminary results of the comparison of statistical properties of trinucleotide composition of genomes, and the taxonomy of their bearers. Meanwhile, one make pose a question whether the distributions described above are stable against the growth of the strings taken into consideration. Indeed, suppose we changed triplets for octanucleotides; then, what happens with the structure of the clusters described above, and would happen with the distributions?

Exponential growth of the dimension of the relevant space is the main problem here. Indeed, even an abundant database is small enough to exceed the dimension of the space determined by the frequency dictionary W_8. Probably, the most fruitful approach here is the efficient reduction of the space of oligonucleotides. If a greater number of strings have the same frequency, then they make no contribution into the discretion of the genomes, and they might be eliminated from the analysis. In reality, such strings have very proximal, but different frequency. Thus, one needs to identify the strings providing the most significant distinctions of the genetic entities.

A dual problem reveals such strings: one should study the distribution of strings in the "space" of genetic entities, instead of the studying of the distribution of the entities in the space of frequencies. The solution of this dual problem would identify the strings that prove the greatest difference among the genetic entities under consideration. The set of such strings may be rather meager, thus providing a researcher with a good space for direct problem solution. More detailed discussion of all these approaches and options falls beyond this paper.

Another surprising point is that we have studied two, formally speaking, incompatible issues: the former is a proximity between the structures of mitochondrion genomes, and the latter is a taxonomy of the bearers determined morphologically. The point is that the evolution of the bearers, strictly speaking, is quite diverse from the evolution of mitochondrion genomes. The results shown above actually prove that the (co)evolution of mitochondrion genomes and the genetic entities of the bearers runs extremely tightly. This proof could be verified through the study of the relation between taxonomy and statistically defined structuredness of the bearer genomes.

ACKNOWLEDGMENT

The authors are thankful to Andrew Zinovyev for the promoting interest, help in *ViDaExpert* software adoption and valuable discussion.

REFERENCES

[1] M.G. Sadovsky, A.S. Shchepanovsky, and J.A. Putintzeva, *Genes, Information and Sense: Complexity and Knowledge Retrieval* // Theory in Biosciences (2008). V.**127**, pp. 69 – 78.

[2] M.G. Sadovsky, *Comparison of real frequencies of strings vs. the expected ones reveals the information capacity of macromoleculae* // J.of Biol.Physics (2003). V.**29**, pp. 23 – 38.

[3] M.G. Sadovsky, *Information capacity of nucleotide sequences and its applications* // Bulletin of Math.Biology (2006). V.**68**, pp. 156 – 178.

[4] A.N. Gorban, T.G. Popova, M.G. Sadovsky, and D.C. Wunsch, *Information content of the frequency dictionaries, reconstruction, transformation and classification of dictionaries and genetic texts*. *Intelligent Engineering Systems through Artificial Neural Networks*, 11 — *Smart Engineering System Design*, N.-Y.: ASME Press, (2001). pp. 657 – 663.

[5] A.N. Gorban, T.G. Popova, and M.G. Sadovsky, *Classification of symbol sequences over thier frequency dictionaries: towards the connection between structure and natural taxonomy* // Open Systems & Information Dyn. (2000). V.**7**(1), pp. 1 – 17.

[6] N.N. Bugaenko, A.N. Gorban, and M.G. Sadovsky, *Maximum entropy method in analysis of genetic text and measurement of its information content* // Open Systems & Information Dyn. (1998). V.**5**(3), pp. 265 – 278.

[7] G.R. Day, and R.D. Blake *Statistical significance of symmetrical and repetitive segments in DNA* // Nucl. Acids Res. (1982) V.**10**(24), pp. 8323-- 8339.

[8] D. Mitchell, and R. Bridge *A test of Chargaff's second rule* // Biochem. and Biophys. Res. Commun. (2006) V.**340**, pp. 90 – 94.

[9] *http://bioinfo-out.curie.fr/projects/vidaexpert/* March, 2011.

 ISBN: 978-1-61208-007-9

Computational Modeling of Robust Figure/Ground Separation

Marc Ebner
Eberhard Karls Universität Tübingen
Wilhelm-Schickard-Institut für Informatik
Cognitive Systems, Sand 1
72076 Tübingen, Germany
marc.ebner@wsii.uni-tuebingen.de

Stuart Hameroff
Departments of Anesthesiology and Psychology
and Center for Consciousness Studies
The University of Arizona
Tucson, Arizona 85724, USA
hameroff@u.arizona.edu

***Abstract*—It is unknown which computational method the brain uses to perceive a visual scene. Given current advancements, it is now possible to model perceptual processes of the brain using spiking neural network models. We have developed a computational model for robust figure/ground separation. The model is based on a laterally connected sheet of spiking neurons. The sheet of neurons receives its visual input from a virtual retina. It is assumed to be located inside V1 or a higher visual area. The neurons are assumed to be laterally connected to their nearest neighbors through gap-junctions. These lateral connections allow the neurons to exchange information and therefore allow for robust figure/ground separation. Even though we only show results for visual signals, the method is quite general and may be used in various areas of the brain. A result of the lateral coupling is that the neurons synchronize their firing behavior resulting in the so called gamma-synchrony which is also a result of our computational model.**

***Index Terms*—visual perception; spiking neurons; lateral-coupling; gap-junctions; gamma-oscillations**

I. Introduction

In computational neuroscience, one tries to understand how the brain actually processes information at the neural level. The goal is to seek an algorithmic description. Once this description is obtained it may be used to simulate the same behavior in another medium, i.e. on a computer. We are still a long way from being able to fully understand how human visual processing works. However, we have been able to show how the brain can process visual information using a sheet of spiking neurons. Our sheet of neurons is laterally connected to neighboring neurons. The connections (assumed to be due to gap junctions) allow the neurons to exchange data with their neighbors and therefore tune their firing behavior such that the relevant neurons collectively respond to a certain stimulus. Our contribution is to extend the spiking neuron model to include lateral connections. We provide a complete algorithmic description of our theoretical model which can be used for comparison with real data or for predictions. We show how the sheet of neurons automatically adapts its behavior so as to robustly extract a figure from ground.

In our simulations, we model a single sheet of neurons. The input to this sheet of neurons is assumed to come from a virtual retina, i.e. from neural cells responding to visual stimuli. Hence, the sheet of neurons perceives and represents a visual scene. Even though we only show results for visual stimuli, the method is quite general and may be used to process arbitrary signals. It could also be used to process haptic or auditory information. Our model assumes that cells performing a related function are connected through gap junctions while no lateral connection exists between cells tuned to process different kinds of information. Since a gap-junction can be modeled as a resistive connection, the entire set of interconnected neurons form a resistive grid. This resistive grid causes the neurons to laterally exchange part of their activation level with nearby neurons provided that the connected gap junction is in an open state. The resistive grid is also used to temporally and spatially average the incoming spikes. This enables the network to tune their behavior and to perform robust figure/ground separation. The temporally and spatially averaged signal is used as an adjustive signal for the neuron. Depending on this signal, the gap junctions open or close. When the temporal average of the neuron's dendritic input is above the spatial average of the neuron's dendritic input, then the gap junction opens it's connection. If the temporal average is below the spatial average, then the gap junction closes. Once, the gap junction between two neurons is open, then these two neurons exchange part of their activation, thereby synchronizing their firing behavior. Eventually, other nearby neurons will also open their gap junctions, thereby forming an extended zone of laterally connected neurons with synchronized firing behavior. All of the neurons whose receptive field shows part of the figure will fire in synchrony. Neurons for which the figure is outside the receptive field will fire out of sync and at a much lower rate.

II. Spiking Neural Networks

Sensory perception, motor control and learning are due to the neural processing which occurs inside the brain. The brain itself is usually modeled as a set of spiking neurons [2]. In this standard model, each neuron independently integrates the electrical inputs which it receives from other neurons. This happens until the activation of the neuron rises above a certain level or threshold. Once this happens, the neuron is said to fire. The neuron then sends an electrical impulse or signal along the axon. This signal may then be integrated by other neurons which eventually will also fire.

 ISBN: 978-1-61208-137-3

It is standard practice to only model the spiking behavior of neurons as this is thought to be the most relevant aspect of neural information processing. It is assumed that the entire function of the neuron can be replicated by only modeling the spiking behavior of the neuron. Low level interactions between neurons, i.e. at the level of neuro transmitters and ion channels, are thought not to be relevant to replicate neural processing. Hence, these aspects are usually omitted in computational modeling. By abstracting and given powerful computational resources, it is possible to even model thalamocortical systems [1].

In the standard so called integrate and fire model, each neuron is viewed as a functional unit. The neuron integrates the input received through the dendrites. Once a given threshold is reached, then the neuron is said to fire. The input of a neuron is due to electrical signals received via axons from other neurons. Whenever a neuron fires, then a voltage spike is sent along its axon. This electrical signal is received by other neurons through their dendrites (and also via their cell bodies). Each neuron integrates its input over time resulting in a buildup of the activation potential. If the activation potential of a cell is high enough, then the neuron will again send a spike down its axon. This signal will be integrated by other neurons and the process continues.

Let V_i be the activation potential of neuron i of a larger network. The change of the activation potential V_i can be modeled by the following equation (modified from [3]):

$$C\frac{dV_i}{dt} = g_i(E_i - V_i) + I_{\text{tonic}} + I_i + \sum_{j=1}^{N} w_{ij}K_j \qquad (1)$$

Here, C is the capacitance of the neuron. The factor g_i is the leakage conductance. This factor will determine the speed with which the cell will eventually reach the resting potential E_i if no input is received. A tonic current can be modeled through the term I_{tonic}. An input current to neuron i from an external source can be provided through the term I_i. Let K_j be the input received from neuron j. Each input will be weighted with factors w_{ij} describing the connection strength between neurons i and j. The connection strengths can be tuned through neural learning. The input of a neuron is the weighted sum over all its inputs received from other neurons.

In this standard neural model, an important ingredient is missing. Lateral connections between neurons are not considered. We find such lateral connections between neurons to be highly useful for signal processing. The lateral connections allow the neurons to exchange data with their immediate neighbors and thereby to collectively tune the response to a given stimulus.

III. Lateral Connections

Our model neuron extends the standard model by also including lateral connections between neurons. Similar to the standard model, the neuron temporally integrates the incoming spikes. This leads to a rise of the activation voltage until a particular threshold is reached. Once this happens, the neuron sends a spike along its axon, i.e. it fires. In contrast to the integrate and fire model, our neuron includes lateral connections which as assumed to be due to gap junctions between neurons. Only neurons which perform a similar function are assumed to be laterally connected. During development, lateral connections may just occur completely at random. In the course of time, some neighboring neurons will fire together by chance. This may lead to gap junctions between these neurons. The laterally connected neurons will form a sub-network. A gap junction can be modeled as a resistive connection between neurons [4], [5]. Hence, the connected neurons form a resistive grid. Since the gap-junctions are always there, the gap-junction connections form an unconditional resistive grid. This resistive grid is used to adaptively tune the neuron to a given stimulus.

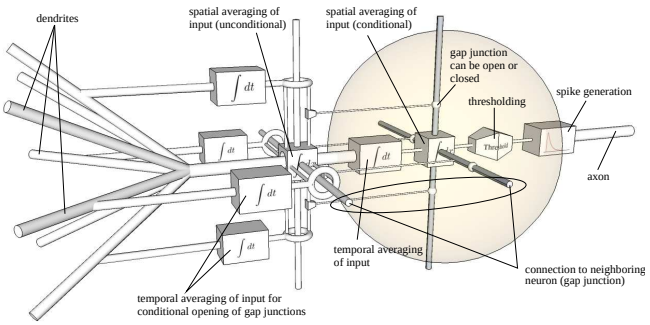


Fig. 1. Artificial neuron. Each neuron is laterally connected via gap junctions to several other neurons (only 4 gap junctions are shown).

A gap junction may be in one of two states. It can be open or closed. The state that is chosen is voltage dependent. A voltage dependent conductance of gap junctions was also used by Traub et al. [6]. In our model, a channel is opened for each open gap junction allowing the connected set of neurons to exchange part of their activation. This leakage current causes the conditionally connected neurons to synchronize their firing behavior. In computational modeling, an open gap junction is modeled as a resistor. The synchronization of laterally connected neurons occurs in the same way that chaotic or non-linear electrical circuits synchronize their behavior if they are resistively connected, i.e. a signal is exchanged between the two circuits [7]–[9].

The input spikes passing next to each gap-junction are temporally integrated and, through the resistive grid, also spatially averaged. The spatially averaged input results in an adjustive signal for the neuron. Gap junctions open and close depending on this signal. In our model, we call this signal the sync-threshold. Gap junctions open if the temporal average of a neuron's input is above the spatial average. Otherwise, the gap junctions close.

An illustration of our neuron including lateral connections is shown in Figure 1. The lateral connections are shown extruding from the body of the neuron in order to make clear that this

 ISBN: 978-1-61208-137-3

is a connection to other neurons on the same level. In reality, the gap-junctions are located on the dendrites which are shown on the left part of Figure 1 leading up to the neuron body. The neuron receives its input through the dendrites. This input is temporally integrated as is illustrated by the center box labeled "$\int dt$". The same input is also temporally integrated (although with a different factor) and also spatially averaged at each gap-junction as illustrated by the boxes "$\int dt$ and "$\int dx$". Figure 1 shows 8 dendritic connections but only 4 gap-junctions in order not to overload the figure. In an actual neuron, the connections are not necessarily uniformly distributed. For each gap-junction, two connections are shown. One dark connection and one light connection. The lighter connection illustrates the resistive grid that is formed because the gap-junction exists. The darker connection illustrates the conditional connection between neighboring neurons as the gap junctions open or close (sphere on the dark lateral connection). If the temporal average of the incoming signal is clearly above the spatial average, then the gap junctions open. If the temporal average is below the spatial average, then the gap junctions close. The dendritic input to the neuron is integrated by the second box labeled "$\int dx$". Note that this input passes through the first box labeled "$\int dx$" which is the integration due to the unconditional resistive grid. If the gap junction is open, part of the activation will be exchanged between the connected neurons. The current will flow from the neuron having a higher activation to the neuron having a lower activation. This causes the connected neurons to synchronize their firing behavior. If the activation of the neuron rises above a threshold (illustrated by the "Threshold"-box), then the neuron will fire. In this case, an electrical impulse is sent along the axon. This is illustrated by the box with the spike.

A connected network of such neurons is able to extract an arbitrary signal which is above the average. The same function could also be achieved with multiple interconnected neurons. It could be that the above behavior illustrated within a single neuron is actually spread over multiple neurons inside a cortical column. See Mountcastle [10] for a review of columnar organization of the neocortex.

IV. Robust Figure/Ground Separation

In order to evaluate our model, we first start off using virtual stimuli. A sheet of 1000 laterally connected neurons is simulated. This sheet of neurons processes input from a virtual retina. The 1000 neurons are randomly placed inside a $100\times100\times2$ area. It would suffice to model a two-dimensional sheet of neurons. However, we have used a three-dimensional sheet in order to include the fact that actual neurons are not perfectly positioned inside a two-dimensional plane. Let (x_i, y_i, z_i) be the position of the i-th neuron inside the three-dimensional area. Each neuron is laterally connected to its 6 nearest neighbors the sheet. Input to neuron i is provided by a virtual retina. The receptive field of neuron i is mapped topographically from its position inside the sheet to the retinal neurons. Let x_i, y_i, and z_I be the normalized coordinates with range $[0, 1]$, then neuron i receives its input from position $(wx_i + x_r, hy_i + y_r)$ where w is the width of the retina and h is the height of the retina and (x_r, y_r) is a random offset selected from $-1, 0, 1$.

Our sheet of neurons could theoretically be located inside V1, however, it is more likely to be located in some higher visual area. It could be used wherever a signal has to be separated from ground. Below, we will show how the network can be used to separate a lighter signal from a darker background. The same network, however, can also be used to separate more complicated signals which depend on motion or texture. Neurons processing these features would be located in V3 or V5 or inside higher areas [11], [12].

The human visual system uses two different types of receptors: rods and cones. The cones are used for color vision. Three different cones can be distinguished. Their peak response lies either in the red, green or blue parts of the spectrum [13]. The retinal receptors measure the light falling onto the retina. The information is then passed on to the lateral geniculate nucleus and finally reaches V1. By the time, the visual information has reached the visual cortex, it has been transformed from a red-green-blue coordinate system to a rotated coordinate system. This rotation is caused by color opponent and double-opponent cells. The axes of the rotated coordinate system are: bright-dark, red-green and yellow-blue [14]. For our experiments, we will be using only the bright-dark channel (also called lightness). We process data which is stored as computer images. The transformation from red, green, and blue non-linear pixel intensities (R, G, B) is given by $L = 0.299R + 0.587G + 0.114B$ [15]. Each neuron i of our sheet receives lightness L from 3 different positions of the virtual retina. The mapping from neurons to their input is defined as described above. Thus, we have for the output o_i of the retinal neuron i: $o_i = L(x'_i, y'_i)$ with $(x'_i, y'_i) = (wx_i + x_r, hy_i + y_r)$.

Each neuron is fully described by the following state variables: a_i activation, t_i fire-threshold, o_i output voltage, $\tilde{a}_i$, temporal average of incoming spikes, $\bar{a}_i$ spatial average of temporal average. The variable $\tilde{a}_i$ is actually associated with every gap-junction. However, we have used one variable per neuron to speed up the simulation. The algorithm which is run by each neuron i is shown in Figure 2. Due to the leakage factors, the state variables can be initialized with random values at the start of the algorithm. For our experiments, we have used the following parameters: $\alpha_a = 0.9995$ decay of activation potential, $\alpha_o = 0.5$ decay of output voltage, $\alpha_t = 0.001$ temporal averaging factor of gap-junction, $\alpha_s = 0.0001$ spatial averaging factor of gap-junction input, $\epsilon = 0.0001$ leakage to adjacent neurons upon firing, $\gamma = 0.0005$ reduction of fire-threshold, $\omega = 1.999$ factor for over-relaxation, $\Delta t_r = 10$ refractory period of neuron, $w_{ij} = 1$ weight between neurons i and j. We have used only positive unit weights because the input image is directly processed by the neural sheet. In the brain, the weights can be found using neural learning, e.g., Hebbian learning [16]. Of couse, it is also possible to include negative weights. Negative weights would represent inhibitory signals. The type of weights that have to be used, are of course

 ISBN: 978-1-61208-137-3

(01) $o_i = (1 - \alpha_o)o_i$ // decay of output
(02) $a_i = (1 - \alpha_a)a_i$ // decay of activation
(03) $a_i = a_i + \alpha_a \sum_j w_{ij} o_j$ // integrate input
(04) $\tilde{a}_i = (1 - \alpha_t)\tilde{a}_i + \alpha_t \sum_j w_{ij} o_j$ // temporal average
(05) $\bar{a}'' = \bar{a}_i$ // save previous result
(06) $\bar{a}' = \frac{1}{1+|N|} \sum_{j \in N} \bar{a}_i$ // compute spatial average
(07) $\bar{a}_i = (1 - \alpha_s)\bar{a}' + \alpha_s \tilde{a}_i$ // add temp. average
(08) $\bar{a}_i = (1 - \omega)\bar{a}'' + \omega \bar{a}_i$ // use over-relaxation
(09) **if** $(\tilde{a}_i > \bar{a}_i)$ open gap junctions
(10) **else** close gap junctions
(11) **if** Neuron i fired within Δt_r **return**
(12) $N = \{j |$Neuron j is laterally connected to
(13) neuron i via open gap junction $\}$
(14) $a' = a_i$; $n = 1$ // initialize spatial averaging
(15) **for all** $j \in N$ **do** : **if** Neuron j did not fire within Δt_r
(16) $\{ a' = a' + a_j; n = n + 1 \}$
(17) $a_i = a'/n$ // spatial averaging completed
(18) // distribute sp. avg to neighboring neurons
(19) **for all** $j \in N$ **do** : **if** Neuron j did not fire within Δt_r
(20) $\{ a_j = a_i; \}$
(21) $t_i = \max[0, 1 - \gamma \cdot N_s]$ // comp. fire-threshold
(22) **if** $(a_i > t_i)$ {// does the neuron fire?
(23) $a_i = 0$ // reset activation
(24) $o_i = 1 - \epsilon|N|$ // output rises to 1
(25) **for all** $j \in N$ **do** : $a_j = a_j + \epsilon$ // distribute leakage
(26) }

Fig. 2. Algorithm of neuron i

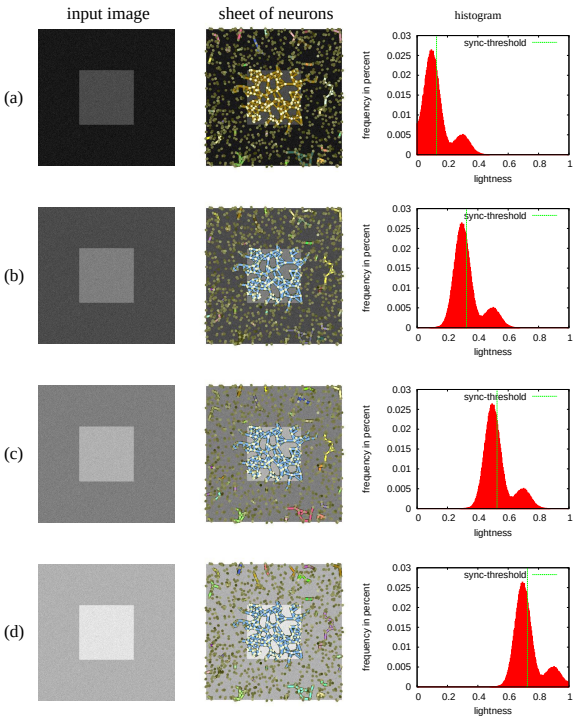


Fig. 3. Experimental results for different noisy input images (zero mean, standard deviation 0.05). The relationship between background lightness L_b and figure lightness L_f is (a) L_b/L_f=0.1/0.3 (b) L_b/L_f=0.3/0.5 (c) L_b/L_f=0.5/0.7 (d) L_b/L_f=0.7/0.9

dependent on the problem that has to be solved. For our task, unit weights suffice. The parameter N_s denotes the number of neurons in the sub-network.

In the following, we will refer to the line numbers of Figure 2 in order to explain what the neuron does. First, the output voltage (01) as well as the activation (02) decays. Each neuron integrates the input (03). The gap junctions are controlled depending on whether the temporal average of the input is above the spatial average (09-10). The temporal average of the input is computed in (04). The spatial average of the temporal average is computed using over-relaxation in (05-08). This spatial average is basically an adaptive threshold which allows for adaptive figure/ground separation.

Note that in an earlier model [17], we have used the firing signal of the neuron as a feedback signal to control all of the gap-junctions at the same time. It is probably more accurate, that each gap-junction is controlled independently by the temporal average of the signal passing through the dendrite where the gap-junction is located. Thus, according to our theory, each gap-junction opens or closed independently of the other gap-junctions depending on the signal that passes through its dendrite. The algorithm that we use for our simulation, nevertheless takes the signal running through all of the dendrites as a single input and controls all gap-junctions of a neuron at the same time. This allows for faster simulation of the entire sheet of neurons.

Condition (09) ensures that the brightest stimulus is extracted. Parts of the image with high lightness correspond to the figure whereas other parts with low lightness correspond to the background. Processing continues if the neuron is no longer in its refractory period (11). Lines (12-20) distribute part of the activation across open gap junctions. The activation flows from the neuron having a higher activation to neighboring neurons having a lower activation. This causes adjacent neurons with open gap-junctions to synchronize their firing behavior. The fire-threshold is set depending on the size of the connected sub-network (21). If the connected sub-network is large, then the threshold is lowered, whereas if the connected sub-network is small, then the threshold is higher. This causes neurons belonging to a larger object to fire with a higher frequency. Once the neuron fires (22-26), most of the activation is sent along the axon. However, part of the activation is also distributed to neighboring neurons.

A single neuron could also perform a bright/dark classification with a proper choice of parameters. However, such a neuron will not be adaptive to the image content. Figure 3 shows the results for different input images with static noise. The input received by the retinal neurons is shown on the left hand side. The sheet of neurons is shown in the middle. Each neuron is marked by a dot. Open gap junctions between neurons are drawn with colored lines. The right hand side shows the distribution of the lightness of the input image. In Figure 3(a) both background and the foreground square (figure) are quite dark. Subsequently, in cases (b-d), the lightness is increased. We can see that for input image (a), a lightness of 0.3 is classified as figure because the background has a lower lightness, e.g., 0.1. However, for case (d), a lightness of as high as 0.7 is classified as background because

 ISBN: 978-1-61208-137-3

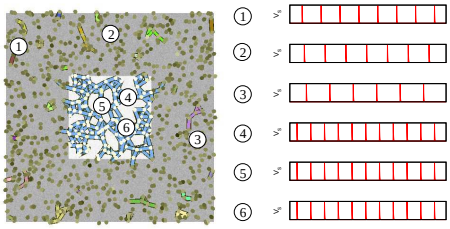


Fig. 4. (a) input stimulus (b) behavior of six different neurons (marked). Neurons 1-3 are located on the figure and show synchronous firing behavior whereas neurons 4-6 are located on the background and fire out of sync.

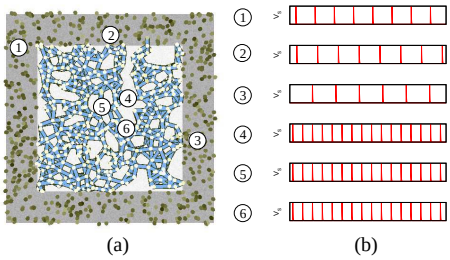


Fig. 5. (a) input stimulus (b) behavior of six different neurons (marked). Neurons 1-3 are located on the figure and show synchronous firing behavior whereas neurons 4-6 are located on the background and fire out of sync.

the figure has a higher lightness of 0.9. Thus, we see that our network is able to adapt to the image content and extract the correct figure. It is also robust in that it is able to cope with noisy input stimuli.

By using other types of input with appropriate weights in (03) and (04), arbitrary stimuli can be extracted. For instance, one could envisage a sheet of neurons processing input from V4 (color) or V5 (motion). Such a sheet could be tuned to extract a moving color stimulus.

Figure 4 shows that the neurons that have their receptive field on the figure fire in synchrony while other neurons fire out of sync. Figure 4(a) shows the neural sheet overlaid on the input image. The output of six different neurons is shown in Figure 4(b). Figure 5 shows what happens for a stimulus of larger size. In this case, the neurons increase their firing rate. This effect is due to the adaptive threshold that is computed in Figure 2(21). Higher visual areas can discern objects of different sizes based on their firing rate.

Figure 6 shows how the network behaves for real input images moving across the virtual retina. As the object or figure moves across the retina, different neurons are activated in the course of time. Neurons of a connected sub-network synchronize their firing rates. Different objects will have different firing rates. This allows for visual servoing techniques [18], [19] which can be used by higher visual areas to track an object.

V. Discussion and Basis of our Model

The sheet of neurons segments the scene into figure and ground. Related work for scene segmentation includes the work of Zhao and Breve [20]. They have used Wilson-Cowan

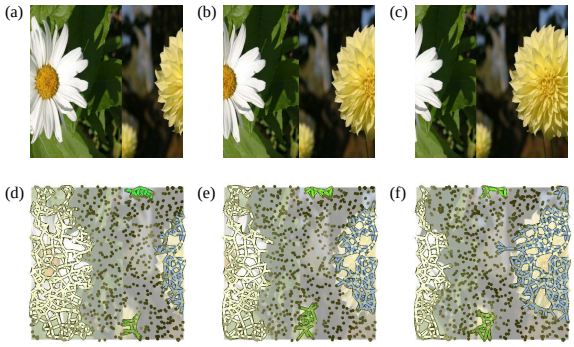


Fig. 6. (a-c) Moving stimulus. (d-f) A connected sub-network tracks the figure.

neural oscillators [21] and segmented static input. Quiles et al. [22] have developed a visual selection mechanism and show how their integrate and fire network responds to different static images. Their model includes short range excitatory connections and long-range inhibitory connections. Eckhorn et al. [23] simulated results from the visual cortex of the cat. They simulated two one-dimensional layers of neurons and used a moving stimulus as input. In contrast to our model, they have used long range feeding connections connecting neurons of the same layer. Our computational model is quite simple, yet it shows how synchronized zones of activity can arise and move around in the brain. These zones of activity are assumed to correlate with conscious perception and control.

Our model of laterally connected neurons show a synchronous firing behavior of neurons responding to the main stimulus (figure) whereas the remaining neurons fire out of sync. Indeed, the electroencephalogram (EEG) shows the synchronized firing behavior of neurons inside the frequency band from 40 to 80Hz [24], [25] This is called gamma synchrony EEG. A review on how gamma synchrony correlates with perception and motor control is given by Singer [26]. The gamma synchrony is due to inter-dendritic gap junctions [27], [28]. Hameroff [29] has put forward the “conscious pilot” model. According to this model, gap junctions open and close, thereby creating synchronized zones of activity. These zones move through the brain and convert non-conscious cognition, i.e. cognition on auto-pilot, to conscious cognition. A review of several different theories of conciousness is given by Kouider [30]. Several theories of consciousness assume re-entrant, i.e. recurrent, processing of information, e.g., the re-entrant dynamic core hypothesis by Tononi and Edelman [31], or the local recurrence theory by Lamme [32]. Crick and Koch [33] have noted that humans appear not to be aware of processing which occurs inside V1. Thus, conscious processing probably starts somewhere above V1. According to Zeki [34], multiple consciousnesses are distributed across different processing sites giving rise to microconsciousness. Attributes such as color, form and motion are bound which then gives rise to macroconsciousness. And finally, there is a global form of conciousness or unified consciousness which involves linguistic and communication skills. Our model is

 ISBN: 978-1-61208-137-3

based on recurrent information processing. Hence, it is in line with theories of Tononi and Edelman as well as the theory of Lamme. In Zeki's terms, our model would be a case of microconsciousness.

Synchronized firing can be achieved through either local or global connections. Our model only uses local connections between neurons. No global connections are required. Nevertheless, global connections could be used to pass information on to higher areas or to provide feedback to lower areas. Wang [35] as well as König and Schillen [36] have used global connections to establish synchronous firing. They use long range excitatory delay connections to achieve desynchronization across different regions. Terman and Wang [37] use a global inhibitor to achieve desynchronization. In our model, neurons responding to the same object will synchronize their firing behavior because they are laterally connected through gap-junctions. Two neurons, each responding to a different object will not be synchronized because of the dependence of the firing threshold on the size of the connected zone of neurons.

VI. Conclusion

The standard integrate and fire model does not take lateral connections between neurons into account. The lateral connections are assumed to occur through gap junctions which behave like resistors. A gap junction may be either in an open state or in a closed state. The gap-junctions form two resistive networks. An unconditional network and a conditional network. The unconditional network is used by our model to tune the network to the correct input level. It computes a spatial average of the temporally smoothed input. This spatial average is used to set the sync-threshold by comparing it to the temporal average of the overall input to the neuron. If the overall input is above the spatial average, then the gap junctions open. This causes the neuron to synchronize its firing behavior such that neurons which have their receptive field above the stimulus fire in synchrony. We have shown that our model allows for robust figure/ground separation both on artificial stimuli as well as with real stimuli.

References

[1] E. M. Izhikevich and G. M. Edelman, "Large-scale model of mammalian thalamocortical systems," *Proceedings of the National Academy of Sciences USA*, vol. 105, no. 9, pp. 3593–3598, 2008.

[2] W. Gerstner and W. Kistler, *Spiking Neuron Models*. Cambridge, UK: Cambridge University Press, 2002.

[3] J.-P. Thivierge and P. Cisek, "Nonperiodic synchronization in heterogeneous networks of spiking neurons," *The Journal of Neuroscience*, vol. 28, no. 32, pp. 7968–7978, Aug. 2008.

[4] J. Herault, "A model of colour processing in the retina of vertebrates: From photoreceptors to colour opposition and colour constancy phenomena," *Neurocomputing*, vol. 12, pp. 113–129, 1996.

[5] M. L. Veruki and E. Hartveit, "All (rod) amacrine cells form a network of electrically coupled interneurons in the mammalian retina," *Neuron*, vol. 33, pp. 935–946, Mar. 2002.

[6] R. D. Traub, N. Kopell, A. Bibbig, E. H. Buhl, F. E. N. LeBeau, and M. A. Whittington, "Gap junctions between interneuron dendrites can enhance synchrony of gamma oscillations in distributed networks," *The Journal of Neuroscience*, vol. 21, no. 23, pp. 9478–9486, Mar. 2001.

[7] T. L. Carroll and L. M. Pecora, "Synchronizing chaotic circuits," *IEEE Trans. on Circuits and Systems*, vol. 38, no. 4, pp. 453–456, Apr. 1991.

[8] L. M. Pecora and T. L. Carroll, "Synchronization in chaotic systems," *Physical Review Letters*, vol. 64, no. 8, pp. 821–824, Feb. 1990.

[9] C. K. Volos, I. M. Kyprianidis, and I. N. Stouboulos, "Experimental synchronization of two resistively coupled Duffing-type circuits," *Nonlinear Phenomena in Complex Systems*, vol. 11, no. 2, pp. 187–192, 2008.

[10] V. B. Mountcastle, "The columnar organization of the neocortex," *Brain*, vol. 120, pp. 701–722, 1997.

[11] S. M. Zeki, "Review article: Functional specialisation in the visual cortex of the rhesus monkey," *Nature*, vol. 274, pp. 423–428, Aug. 1978.

[12] S. Zeki, *A Vision of the Brain*. Oxford: Blackwell Science, 1993.

[13] H. J. A. Dartnall, J. K. Bowmaker, and J. D. Mollon, "Human visual pigments: microspectrophotometric results from the eyes of seven persons," *Proc. R. Soc. Lond. B*, vol. 220, pp. 115–130, 1983.

[14] M. J. Tovée, *An introduction to the visual system*. Cambridge: Cambridge University Press, 1996.

[15] C. Poynton, *Digital Video and HDTV. Algorithms and Interfaces*. San Francisco, CA: Morgan Kaufmann Publishers, 2003.

[16] D. O. Hebb, *The Organization of Behavior*. New York: Wiley, 1949.

[17] M. Ebner and S. Hameroff, "A computational model for conscious visual perception and figure/ground separation," in *Proc. Int. Conf. on Bio-Inspired Systems and Signal Processing, Rome, Italy*, Portugal: Science and Technology Publications, 2011, pp. 112–118.

[18] F. Chaumette and S. Hutchinson, "Visual servo control part I: Basic approaches," *IEEE Robotics & Automation Magazine*, vol. 13, no. 4, pp. 82–90, Dec. 2006.

[19] ——, "Visual servo control part II: Advanced approaches," *IEEE Robotics & Automation Magazine*, vol. 14, no. 1, pp. 109–118, 2007.

[20] L. Zhao and F. A. Breve, "Chaotic synchronization in 2D lattice for scene segmentation," *Neurocomputing*, vol. 71, pp. 2761–2771, 2008.

[21] H. R. Wilson and J. D. Cowan, "Excitatory and inhibitory interactions in localized populations of model neurons," *Biophysical Journal*, vol. 12, pp. 1–24, 1972.

[22] M. G. Quiles, L. Zhao, F. A. Breve, and R. A. F. Romero, "A network of integrate and fire neurons for visual selection," *Neurocomputing*, vol. 72, pp. 2198–2208, 2009.

[23] R. Eckhorn, H. J. Reitboeck, M. Arndt, and P. Dicke, "Feature linking via synchronization among distributed assemblies: Simulations of results from cat visual cortex," *Neural Computation*, vol. 2, pp. 293–307, 1990.

[24] C. M. Gray and W. Singer, "Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex," *Proceedings of the National Academy of Sciences USA*, vol. 86, pp. 1698–1702, Mar. 1989.

[25] U. Ribary, A. A. Ioannides, K. D. Singh, R. Hasson, J. P. R. Bolton, F. Lado, A. Mogilner, and R. Llinás, "Magnetic field tomography of coherent thalamocortical 40-hz oscillations in humans," *Proc. of the National Acad. of Sciences USA*, vol. 88, pp. 11 037–11 041, Dec. 1991.

[26] W. Singer, "Neuronal synchrony: A versatile code for the definition of relations?" *Neuron*, vol. 24, pp. 49–65, 1999.

[27] R. Dermietzel, "Gap junction wiring: a 'new' principle in cell-to-cell communication in the nervous system?" *Brain Research Reviews*, vol. 26, pp. 176–183, 1998.

[28] A. Draguhn, R. D. Traub, D. Schmitz, and J. G. R. Jefferys, "Electrical coupling underlies high-frequency oscillations in the hippocampus in vitro," *Nature*, vol. 394, pp. 198–192, Jul. 1998.

[29] S. Hameroff, "The "conscious pilot" – dendritic synchrony moves through the brain to mediate consciousness," *Journal of Biological Physics*, vol. 36, pp. 71–93, 2010.

[30] S. Kouider, "Neurobiological theories of consciousness," in *Encyclopedia of Consciousness*, W. P. Banks, Ed. Elsevier, 2009, pp. 87–100.

[31] G. Tononi and G. M. Edelman, "Consciousness and complexity," *Science*, vol. 282, pp. 1846–1851, Dec. 1998.

[32] V. A. F. Lamme, "Towards a true neural stance on consciousness," *Trends in Cognitive Sciences*, vol. 10, no. 11, pp. 494–501, 2006.

[33] F. Crick and C. Koch, "Are we aware of neural activity in primary visual cortex?" *Nature*, vol. 375, pp. 121–123, May 1995.

[34] S. Zeki, "A theory of micro-consciousness," in *The Blackwell companion to consciousness*, M. Velmans and S. Schneider, Eds. Malden, MA: Blackwell Publishing, 2007, pp. 580–588.

[35] D. Wang, "Emergent synchrony in locally coupled neural oscillators," *IEEE Trans. on Neural Networks*, vol. 6, no. 4, pp. 941–948, Jul. 1995.

[36] P. König and T. B. Schillen, "Stimulus-dependent assembly formation of oscillatory responses: I. synchronization," *Neural Computation*, vol. 3, pp. 155–166, 1991.

[37] D. Terman and D. Wang, "Global competition and local cooperation in a network of neural oscillators," *Physica D*, vol. 81, pp. 148–176, 1995.

 ISBN: 978-1-61208-137-3

Anacê: Phylogenetic Trees Drawing Web Service

Hélio Augusto Sabóia Moura
MPComp - Integrated Master on Applied Computation
State University of Ceará
Fortaleza, CE, Brazil
helio.moura@uece.br

Gerardo Valdísio Rodrigues Viana
Dept. of Computer Science
State University of Ceará
Fortaleza, CE, Brazil
valdisio@uece.br

***Abstract*—In this paper, we describe a tool, called Anacê[1], to draw phylogenetic trees in diverse topologies. Available via Web service, and developed in Scala, this tool can be used in any computational platform in the interactive form or as subprograms in any application or programming language. The generated trees is exported in SVG (Scalable Vector Graphics) image formats that are independent of computational platforms. The tool easily draws trees from the distinct forms of tree's representation, generated by other software. It is meant to be used by researchers and as a learning tool.**

Keywords-computational biology; phylogenetic trees; Web service.

I. INTRODUCTION

A phylogenetic tree can be depicted in several topologies. For example, a given phylogenetic tree can be represented as a rectangular cladogram, an inclined cladogram, a phylogram, a radial tree, a free tree with or without root or still in the textual form using the Phylip standard [1]. In this paper, we propose a tool, called Anacê, to draw trees in any[2] of these formats. It is available via Web service and is implemented in Scala [2]. We created a site with a Anacê's tutorial [3].

Since a user developing a Web service does not need to know on which programming language a client will implement his/her applications, Anacê uses RESTful [4] Web services, and so, the only resource needed is a library to use HTTP protocol, wich is available in the most of the programming languages.

The main objective with this tool is to have a unique point of execution, preventing each different computational environment to have a copy of the library installed in the client application. Another advantage is that the improvements in this service and new methodologies developed become automatically available on the Anacê web site, thus assuring that the latest version of the tool is used.

Several tools have been developed to draw phylogenetic trees. For example, DrawTree [1], TreeView [5], PhyloDraw [6] and Spectrum [7]. In order to be used all these tools need to be installed in the user machine, and so they request specific computational resources. There exists also the Web server called Phyml [8] that uses the maximum-likelihood method to infer phylogenetic trees.

In this work our intention is not to infer [8] neither to reconstruct [9] phylogenies. Actually, we aim to draw trees from its distinct forms of representation generated from other software.

In Section II, we review concepts about Phylogenetic Systematics with emphasis on phylogeny, cladograms and phylogenetic trees. In Section III, we describe methods for phylogeny construction from matrices of distances and sequences of genes and proteins [8]. In Section IV, we describe the functionalities of our tool. Finally, in Section V, we present the conclusions of the work.

II. PHYLOGENETIC SYSTEMATICS

The fundamental concept of the evolution is that for any two species there was at least one common ancestral species; for three species, the hypothesis is that two of them have an ancestral that is not common to the third one [10]. Following this reasoning for all species, we get a sequence of fragmented divisions from the first ancestral species. The diagram that represents this evolutive history of the species is called, generically, of phylogeny or phylogenetic tree [11].

In modern biology there exists a research area called Phylogeny Systematics that has as objective to understand the relationships between all living beings and then to infer the history of their lives and origins [12]. The term phylogeny is used to designate any diagram that presents the phylogenetic relations between the species in study. A cladogram corresponds to the relationships of a group of species (taxons) with common ancestor, whereas a phylogenetic tree, moreover, expresses the relations of the type ancestral-descendants. There exists still the term phylogram that is a special rectangular cladogram, in which the size of its branches is proportional to the evolutive distances between taxons.

To illustrate these differences, we present in Figure 1.a an inclined cladogram with four recent species A, B, D, E and a fossil species $C^{\star}$. One of the possible phylogenetic trees for this cladogram is shown in Figure 1.b, where F is

[1]The name Anacê comes from Tupi, a brazilian native idiom, and means parenthood.

[2]In time of this article not all formats are available yet.

 ISBN: 978-1-61208-137-3

a common ancestral to A and B and G is a species common to all other species.

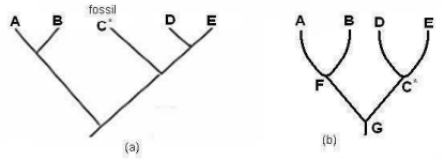


Figure 1. (a) Inclined cladogram with four recent species and one fossil. (b) A possible phylogenetic tree [10].

Phylogenetic trees can be rooted or not. In a rooted tree it is possible to introduce the notion of ancestral traces (plesiomorphics) and derivatives (apomorphics). It is observed in Figure 2.a that the evolutive sequence of some tetrapodies (vertebrate terrestrial that possess four members) is clear and the control group *fishes* is identified, what does not occur in the non-rooted tree shown in Figure 2.b.

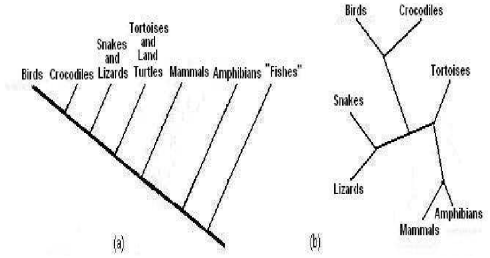


Figure 2. (a) A rooted tree. (b) A non-rooted free tree [13].

In general, in a phylogeny it is not necessary to identify ancestors nor the dates in the ramifications. However, the sequential order of the evolution must always be shown. Therefore, if the external group is biased to the left, recent will be biased to the right, or vice versa, in such a way that it is possible to distinguish the ancestral characteristics from the derivatives [14].

A phylogenetic tree can also express the evolutive distance between the species, in a such way that the length of its branches or edges is proportional to that distance. Thus, two species that have higher similarity will be near, otherwise they will be distant from each other.

III. Methods for Construction of Phylogenetic Trees

Anacê has as input, in the basic format, a text file that contains the distance matrix or a representation enclosed in parentheses in the Phylip standard format [1]. Both generated from programs that analyze and make alignments of sequences of nucleotides or amino acids [15].

To illustrate some of these forms, we use an example of DNA test to identify which of the two suspects A or B had transmitted the HIV/Aids virus to a victim V of rape. In Figure 3 are shown sequences with 30 nucleotides of the involved ones in the test, where sequence X belongs to a person carrying the virus, however, not related with the crime. In this case, he/she corresponds to the called control group, or external group.

X-control group:	AAGCTTCATAGGAGCAACCATTCTAATAAT
A-suspect1:	AAGCTTCACCGGCGCAGTTATCCTCATAAT
B-suspect2:	GTGCTTCACCGACGCAGTTGTCCTTATAAT
V-victim:	GTGCTTCACCGACGCAGTTGCCCTCATGAT

Figure 3. DNA Sequences [13].

Comparing every pair of sequences in Figure 3 we have the following percentages of distinct characters: $XA = 8/30$, $XB = 12/30$, $XV = 13/30$, $AB = 5/30$, $AV = 6/30$ and $BV = 3/30$. From these values we obtain the results in Figure 4.a, represented by the matrix of distances in relative values. Figure 4.b shows the corresponding rectangular cladogram. Similar matrix, with proportional values, can be obtained by running the programs Clustal and ProtDist in Phylip package [1].

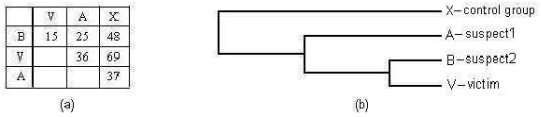


Figure 4. (a) Matrix of distances of the elements in the test. (b) Corresponding phylogeny represented by a rectangular cladogram [13].

From Figure 4.b we conclude that suspect2 (B) is the culprit for the crime, since that one has more common characteristics with the victim (V).

The method used to generate the phylogenetic tree shown in Figure 4.b was the Neighbor-Joining [12]. This method uses the concept of distance in a metric space [11] given by the function $d : E \times E \rightarrow R$, such that are valid the following properties for any distinct elements x, y and z of E:

- $d(x, x) = 0$ and $d(x, y) > 0$ (i.e., d is a non-negative function)
- $d(x, y) = d(y, x)$ (i.e., d is symmetric)
- $d(x, y) \leq d(x, z) + d(y, z)$ (i.e., d satisfies the triangle inequality)

We say that a metric space is additive if, and only if, given four elements i, j, k and l, represented, for example, in Figure 5.a, the relations shown in Figure 5.b are valid.

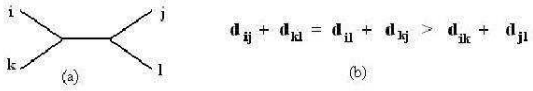


Figure 5. Valid relations in an additive metric space.

In the matrix in Figure 4.a, B and V are the nearest neighbors, and $AV > AB$, and so in an additive metric space we would have the situation presented in Figure 6.

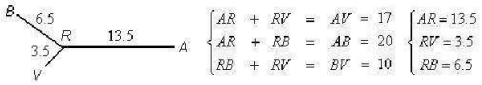


Figure 6. (a) Determining the ramification point (R) of the line segment BV in an additive metric space.

We observe that the system in Figure 6 is always feasible, with unique solution given by $AR = (AV + AB - BV)/2$, $RV = (AV+BV-AB)/2$ and $RB = (AB+BV-AV)/2$. In this case, the ramification RB starting at BV always exists if the solution of the system is positive.

In case this does not occur, the space is not metric and point R is not defined. To fix this problem, in order to show the relationship between the species, we use the methods Unweighted Pair-Group Method using Arithmetic average (UPGMA) when the evolution taxes are approximately constant between different species and Weighted Pair-group Method using Arithmetic average (WPGMA) corresponding to a weighed mean, in order to better reflect the neighborhoods. For example, for $AB = 19$, $AV = 36$ and $BV = 15$, the solution of the system would be $AR = 20$, $RV = 16$ and $RB = -1$. Using the methods mentioned above, we obtain the solutions shown in Figure 7 (method UPGMA) and in Figure 8 (method WPGMA).

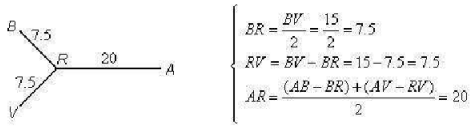


Figure 7. Solution obtained by the method UPGMA.

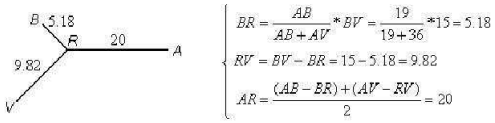


Figure 8. Solution obtained by the method WPGMA.

We observe that UPGMA halves a segment, while WPGMA makes it proportionally, better reflecting the similarities. For this reason, we chose this method in our implementation.

The algorithm finds, in the distance matrix, the nearest neighbors, and then the ramification point is computed, initially in accordance with the criterion presented in Figure 6 (Neighbor-Joining), otherwise, with that described in Figure 8 (WPGMA). We observe that segment BV was divided proportionally and accurately, while segment RA has reduced size to make the considered points pertaining to a *new* metric space. These procedures are successively repeated until all elements in the matrix have been considered.

Applying this method for the matrix in Figure 4.a, we obtain the phylogenetic tree indicated in two distant forms in Figure 9, whose representation in the text format using the Phylip standard is given by $(X : 25, A : 2, (B : 3.5, V : 6.5) : 11.5)$.

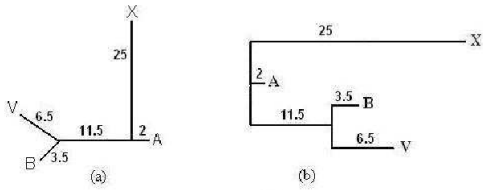


Figure 9. (a) Free and non-rooted phylogenetic tree. (b) Corresponding phylogram.

IV. Functionalities of the Anacê

Anacê is a set of functions written in the Scala [2] programming language available as a set of REST [16] services. Using the RESTful [4] pattern, any program in any programming language can access this services. The only resource needed is a library to use HTTP protocol, wich is available in the most of the programming languages.

The interactive use of this service via Web is by selecting options in forms. The input file is a symmetric matrix of distances in text format that must be pasted in a specific area of work with the structure indicated in Figure 10. Note that de distance's matrix starts with the heading between bracktes, following the heading are the lines of de superior-triangular matrix of distances, stating from zero, that corresponds to the element in the main diagonal in each line.

An alternative input form for Anacê is the Phylip standard format that represents a tree via a text. Figure 12 shows the phylogenetic tree generated by Anacê for the following data that use the Phylip standard format: ((((Y arrowia : 0.57, (Orthopsilosis : 0.03, Candida : 0.01) : 0.43) : 0.20, M archantia : 0.52) : 0.35, Caenorhabditis : 1.71) : 0.23, (Drosophila : 0.30, M elipona : 0.61) : 0.54, ((Rattus : 0.13, (P an : 0.03, Homo : 0.03) : 0.10) : 0.15, (Cobitis : 0.46, Oreochromis : 0.27) : 0.33) : 0.30).

A. *Using* curl *command to access Anacê*

The command *curl* [17] is a tool to transfer data from or to a server, using one of the supported protocols (DICT, FILE, FTP, FTPS, GOPHER, HTTP, HTTPS, IMAP, IMAPS, LDAP, LDAPS, POP3, POP3S, RTMP, RTSP, SCP, SFTP, SMTP, SMTPS, TELNET and TFTP). The command is designed to work without user interaction.

In the next examples the word *prefix* must be replaced by *http://anace.uece.br:9080/anace/rs/*.

```
[Seven Nodes Test]
A 0  63  94 111  67  23 107
B     0  79  96  16  58  92
C        0  47  83  89  43
D           0 100 106  20
E              0  62  96
F                 0 102
G                    0
```

Figure 10. Input example in the basic format.

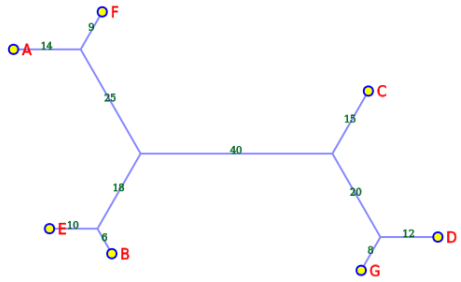


Figure 11. Phylogenetic tree for the input in Figure 10.

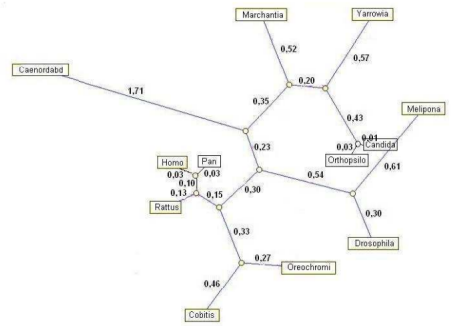


Figure 12. Phylogenetic tree for 12 species generated by Anacê.

The example that follow informs a distance's matrix and get a phylogenetic tree in Phylip's format.

```
curl -X POST prefix/toTree \
-d matrix="[7 Nodes] A 0 63 94 111 67 \
23 107 B 0 79 96 16 58 92 C 0 47 83 89 \
43 D 0 100 106 20 E 0 62 96 \
F 0 102 G 0"
```

Note that all the commands might be written in a single line, here we use de \ character to continue the command in the next line only for visualization purpose. The result is:

```
[&&HM:title=7 Nodes]((F:9.00000,A:14.0000)
:25.0000,(E:10.0000,B:6.00000):18.0000,((
G:8.00000,D:12.0000):20.0000,C:15.0000)
:40.0000)
```

The example that follow informs a phylogenetic tree in Phylip's format and get a distance's matrix.

```
curl -X POST prefix/toMatrix \
 -d tree="((F:9,A:14):25,(E:10,B:6):18, \
          ((G:8,D:12):20,C:15):40)"
```

The result is:

```
[no title]
E 0  62  67  96  16  83 100
F 0  23 102  58  89 106
A 0 107  63  94 111
G 0  92  43  20
B 0  79  96
C 0  47
D 0
```

Note that this distance's matrix can be rewroted like:

```
[no title]
E 0 62  67  96  16 83 100
F    0  23 102  58 89 106
A       0 107  63 94 111
G          0  92 43  20
B             0 79  96
C                0  47
D                   0
```

The example that follow informs a phylogenetic tree in Phylip's format and get a SVG image for the tree like a cladogram.

```
curl -X POST prefix/toCladogram \
  /400/400/1.0/0.2 \
  -d tree="((F:9,A:14):25,(E:10, \
     B:6):18,((G:8,D:12):20, \
     C:15):40)"
```

The result is:

```
<svg:svg width="400" height="400" ...
 ... here comes all the SVG's commands
     to trace the phylogenetic tree ...
</svg:svg>
```

The Anacê's site, with its tutorial, may be visited at:

http://anace.uece.br:9080/anace/home

V. Conclusion

In this paper, we have described a tool, called Anacê, that makes it simple the task of drawing phylogenetic trees. This tool is especially useful to researchers in computational biology that work with phylogeny. The Anacê is also useful for people working in graph theory, since this tool makes more enjoyable the process of checking visually if a given graph satisfies a specific property.

References

[1] J. P. Felsenstein, "Phylogeny inference package computer programs for inferring phylogenies," URL:http://evolution.genetics.washington.edu/phylip.html, Seattle - WA, EUA, 1993, last time accessed: January 2011.

[2] Odersky, Martin, Spoon, Lex, and Venners, Bill, *Programming in Scala*. EUA: Artima, 2008.

[3] G. V. R. Viana and H. A. S. Moura, "Anacê," URL:http://anace.uece.br/anace/home, Fortaleza, CE, Brazil, 2011, last time accessed: January 2011.

[4] Richardson, Leonard and Ruby, Sam, *RESTful Web Services*. USA: OReilly, 2007.

[5] R. D. M. Page, "Treeview for win32," URL:http://taxonomy.zoology.gla.ac.uk/rod/rod.html, last time accessed: January 2011.

 ISBN: 978-1-61208-137-3

[6] Choi, J., Jung, H., KIM, H., and Cho, H., "Phylodraw: A phylogenetic tree drawing system," *Bioinformatics*, vol. 16, pp. 1056–1058, 2000.

[7] M. A. Charleston, "Spectrum: Spectral analysis of phylogenetic data," *Bioinformatics*, vol. 11, p. 9899, 1998.

[8] O. Gascuel, "A web server for fast maximum likelihood-based phylogenetic inference," 2004.

[9] Swofiord, D. L. and Olsen, G. L., "Phylogeny reconstruction molecular systematics," Massachusetts - EUA, 1990.

[10] D. S. Amorim, *Elementos Básicos da Sistemática Filogenética*. Holos, 1997.

[11] J. C. Setubal and J. Meidanis, *Introduction to Computational Molecular Biology*. Brooks/Cole Publishing Co., 1997.

[12] N. Saitou and N. Nei, "The neighbor-joining method: A new method for reconstructiong phylogenetic trees," *Molecular Biology and Evolution*, vol. 4, p. 4.

[13] S. C. Stearns, *Evolution: an Introduction*. Oxford University Press, 2000.

[14] Purves, W.K., Sadawa, D., Orians, G. H., and Heller, C., *Life: The Science of Biology*. W.H. Freeman Co., 2003.

[15] Kumar, S., Tamura, K., and Nei, M., "Mega3: Integrated software for molecular evolutionary. genetic analysis and sequence alignment." *Briefing in Bioinformatics*, vol. 5(2), pp. 150–163, 2004.

[16] T. Fielding, "Architectural styles and the design of network-based software architectures," Doctor of Philosophy, University of California, Irvine - CA, USA, 2000.

[17] "curl," URL:http://curl.haxx.se, last time accessed: January 2011.

 ISBN: 978-1-61208-137-3

Computation of Dynamic Channels in Proteins

Petr Beneš, Petr Medek, Ondřej Strnad, Jiří Sochor
Faculty of Informatics
Masaryk University
Brno, Czech republic
xbenes2@fi.muni.cz, peme@peme.cz, xstrnad2@fi.muni.cz, sochor@fi.muni.cz

***Abstract*—In this paper, we propose a new method which considers the movement of a protein molecule as a whole for the computation of so called dynamic channels in a molecular dynamics trajectory. The method is based on maximizing the information about the empty space over time and is built on basic computational geometry principles. The dynamic channels highlight pulsing and flexible parts of the molecule. It is believed that such parts allow a ligand to pass into or out from the active site. The method was tested on real protein data and the results indicate that it presents new information about the molecule.**

***Keywords*-protein; dynamic channel; molecule; trajectory; computational geometry**

I. Introduction

The shape of a protein molecule is complicated and contains many cavities and pockets. In our research, we are primarily interested in specific cavities connecting a part of a protein molecule (active site) with the surface of the protein. Such cavities are denoted as channels. Channels are used by a substrate molecule to pass into the active site where it can react and may also be used by the products to leave the active site.

The structure of a protein molecule does not remain static over time. Atoms are continuously moving. With this movement, cavities and channels are also changing. The movement of atoms (protein dynamics) is represented as a set of states of the protein molecule which we call snapshots. The whole set is called a trajectory and may contain thousands of snapshots.

Recent methods for computation of channels typically detect channels separately in each snapshot. The channels computed in one snapshot are optimized for bottleneck radius in this particular snapshot only, but not in the whole trajectory. Over time, as atoms are moving, the channel may pulse and thus its parts may alternate from really narrow to wide. In each snapshot then, only a part of a channel may be wide while other parts of the same channel are narrow. Evaluating snapshots separately implies that such a channel would not be identified by existing methods. During a given time interval the channel may be wide in different parts. If these parts are considered altogether, we can find that the channel was wide along its whole length and is maximized in width for the whole trajectory (see Fig. 1).

Such a channel which is detectable in multiple snapshots is called dynamic channel. The method proposed in this paper is designed to detect dynamic channels. This approach considers snapshots together which ensures that dynamic channels are optimized for the whole trajectory. In other words, a dynamic channel may be composed of parts from different snapshots.

The dynamic channel is an approximation because it does not take the order of snapshots into account. In spite of this, such information is valuable since the trajectory is the approximation of the reality as well and since it covers only a short interval of protein life.

Each part of the dynamic channel is wide in a certain snapshot and thus it is expected that the protein molecule is flexible in that parts. This means that if the substrate molecule would pass through a dynamic channel, the atoms in the protein molecule may easily move and create the necessary empty space.

Our method assumes that the whole protein molecule does not change its position significantly. The data obtained from molecular dynamics simulations usually satisfy this condition. If not, there are various alignment techniques which are able to omit the global movement of the molecule.

Preliminary testing on haloalkane dehalogenase DhaA indicates that the method provides reasonable results. However, this paper does not address the issue of biochemical relevance of computed channels – it presents the method and its capabilities.

II. Related work

A channel in a protein molecule is defined [1] as a centerline and a volume. The centerline is a three-dimensional continuous curve and the volume is formed by the union of spheres with centers on the centerline and with an appropriate radius so that they do not intersect any atom in the molecule. The example of a channel is demonstrated in Fig. 1.

There are many methods which deal with the issue of detecting cavities in protein molecules. For instance, the method introduced in [2] is based on the alpha shape theory. The latest approaches can be found for instance in [3], [4]. The information about cavities is important, but these methods do not consider the cavities as channels.

 ISBN: 978-1-61208-137-3

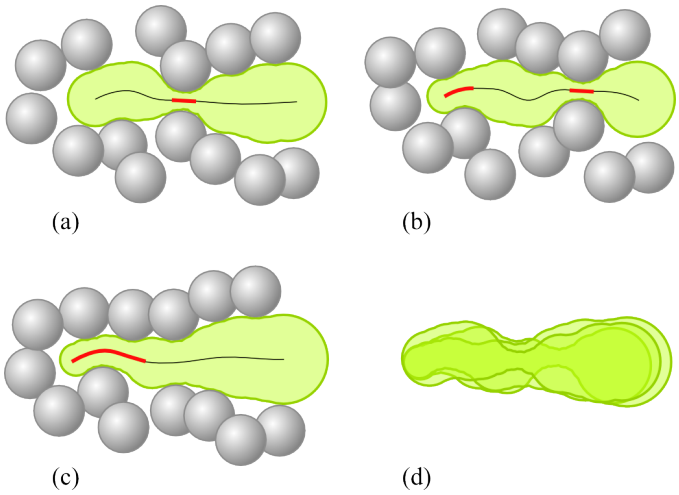


Figure 1. Dynamic channel. (a), (b), (c) A part of a channel is very narrow in each snapshot (emphasized). (d) All channels considered at once form a dynamic channel which is maximized in width across all three snapshots.

There are various approaches to channel computation. Most of them deal with the static case only. The first method which was capable of detecting channels was introduced in [5]. The method was based on sampling the molecule using a three-dimensional grid and suffered from several disadvantages. The methods that followed were based on computational geometry – the first of them was proposed in [1]. Other methods [6], [7] released later are similar. In all these methods, the space partitioning is represented by the Voronoi diagram (VD) or its dual Delaunay triangulation (DT). These structures are processed by an algorithm for finding the shortest path in a graph (typically Dijkstra's algorithm) and a channel which is optimal for given criteria is found. In principle, the centerlines of channels computed using these methods lead along Voronoi edges. A grid-based method introduced in [8] is able to find static molecular channels and voids.

The movement of atoms in a protein molecule is usually obtained by complex physical simulations called molecular dynamics [9]. In reality, the movement is continuous, but the simulation provides discrete results in the form of a set of snapshots as referred above. A variant of molecular dynamics, Random accelerated molecular dynamics (RAMD, [10]) could be also used to detect channels. In RAMD, a small molecule is placed into the active site and a simulation is started. During the simulation, additional random forces are applied to the small molecule and it is probable that it reaches the surface and leaves the protein molecule. If the small molecule leaves, an escape path exists and is detected. The disadvantage of this approach is that it is time consuming and it does not ensure that an escape path will be found.

Other method introduced in [11] computes channels in an existing trajectory. A given number of widest channels is computed in each snapshot and they are partitioned into clusters according to their similarity after all snapshots are processed. Each cluster contains channels from different snapshots which are similar and thus these channels can be considered as states of one channel over time. The similarity of channels is determined by a user-defined distance function. This approach has an obvious disadvantage. The method is not able to detect dynamic channels since it is focused on the computation of channels with the biggest bottleneck radius in each snapshot. The problem is that if we compute a limited number of channels in each snapshot, the dynamic channel may not be present among these channels, because in each snapshot it may not be wide along its whole centerline (see Fig. 1).

Recent methods for computation of channels are based on computational geometry principles. The fundamental geometric structures Voronoi diagram and Delaunay triangulation are of key importance for the proposed method. The duality which exists between these structures allows the easy converting between them. In three dimensions, the important correspondences which are used for our purposes are such that a tetrahedron in the Delaunay triangulation corresponds to Voronoi vertex, a triangle face shared by two neighbouring tetrahedra in DT corresponds to Voronoi edge. Certainly, there are other correspondences, but they are not necessary for the purposes of this paper and are omitted. For more details we refer to [12]. The illustration of the duality in three dimensions can be seen in Fig. 2.

In the dynamic case, when processing the whole trajectory, it is necessary to compute the Voronoi diagram in each snapshot. The issue of computation of the Voronoi diagram in the dynamic environment is described in [13]. Instead of recomputing the Voronoi diagram, the algorithm only performs necessary updates. The complexity of this algorithm is dependent on the number of changes in the Voronoi diagram. This approach assumes that input trajectories of moving spheres are continuous and thus its use in the method presented in this paper is limited. Therefore we recompute the Voronoi diagram for each snapshot using the QuickHull algorithm [14].

III. Proposed method

A dynamic channel can be as well as static channel ([1]) defined as a centerline and a volume which is formed by the union of empty spheres inserted in each point on the centerline. In addition to previous definition, the dynamic channel contains the information about snapshot number for each point on the centerline – a sphere inserted at that point does not intersect nor contain any of the atoms in that particular snapshot.

Definition: Let $M = \{m_1, ..., m_k\}$ be a set of snapshots in the trajectory. A dynamic channel T is defined as $T = \bigcup_{x \in a_T} s(x, r, i)$ where a_T is a three-dimensional curve (centerline of T) and $s(x, r, i)$ is a sphere with center x

and radius r which does not intersect any atom in snapshot $m_i \in M$.

The method we propose is based on the two algorithms described in [1] and [15]. Therefore we first briefly describe the main idea of these algorithms. Then, we describe a novel method which is the main contribution of this paper.

The space partitioning of the protein molecule is stored in the Delaunay triangulation (DT) computed for its atoms. In [1], the DT computed for the set of atom centers is converted to edge-weighted graph G. The nodes of G are formed by Voronoi vertices dual to tetrahedra in the DT. In G, an edge between two nodes exists, if the corresponding tetrahedra in DT share a face (Fig. 2). The value of the edge is equal to the distance from the edge to the surface of the nearest atom. Graph G is then processed using the Dijkstra's algorithm. The cost function maximizes the bottleneck radius of the channel.

The algorithm introduced in [15] is designed to track the centerline of an existing channel in any snapshot in a trajectory. In each snapshot, the algorithm locates the set of tetrahedra intersected by the centerline of the channel and determines Voronoi edges dual to faces shared by neighbouring tetrahedra in this set. As a result of the algorithm, a new centerline composed of Voronoi edges is returned. The centerline is optimized in width while preserving its location. The method uses so-called walks in the Delaunay triangulation to speed up the tracking progress.

We propose a new method for the detection of dynamic channels defined above. The method can be divided into three main parts. Firstly, a graph G_{ini} which represents the molecule appropriately is created. The topology of the graph G_{ini} remains constant during the computation. Secondly, all snapshots in the trajectory are processed. In each snapshot, edges in G_{ini} are updated so that after the processing of the whole trajectory, the value of each edge is maximized. The third step is to compute paths in G_{ini} from the starting node to any of the boundary nodes. The thorough description of these steps follows.

A. G_{ini} creation

There are many possibilities while creating G_{ini}, but not all of them, however, represent the molecule conveniently. In this paper, we propose following two G_{ini} variants which we expect to provide accurate results. We either use the Voronoi diagram (VD) from a selected snapshot of the trajectory or create G_{ini} based on a 3D grid. In case of grid, the bounding volume of the molecule is sampled uniformly using given sampling density. Samples in the grid are used as nodes in G_{ini}. Two nodes are connected by an edge if the corresponding samples are neighbours in the 3D grid. Both G_{ini} variants are discussed in Results section.

The node in G_{ini} that is the nearest to the active site is marked as starting node. The active site is specified by a user as three-dimensional coordinates or by surrounding

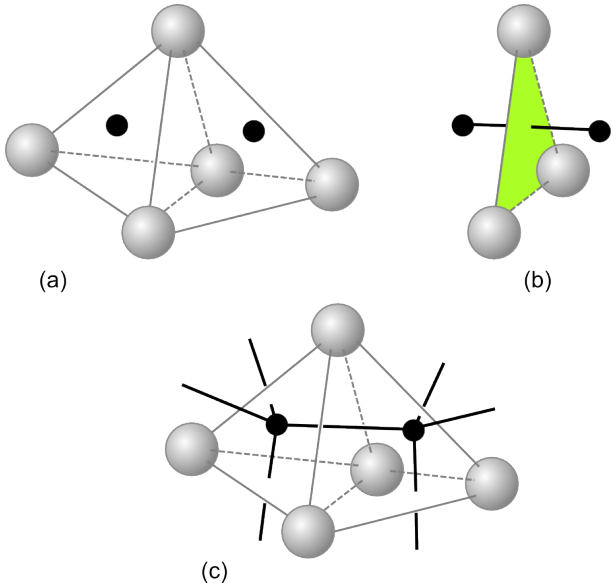


Figure 2. Voronoi – Delaunay duality in three dimensions. (a) Tetrahedra dual to Voronoi vertices. (b) A triangle face shared by two neighbouring tetrahedra is dual to a Voronoi edge. (c) Complete Voronoi – Delaunay duality.

atoms in the molecule. If atoms are used, the coordinates are typically computed by averaging the atom positions in the first snapshot in the trajectory.

For further computation of channels it is necessary to mark certain nodes in G_{ini} as boundary. These nodes lay near the surface of the molecule. In both G_{ini} variants, nodes of edges which are outside or intersect the convex boundary of the molecule are marked as boundary.

B. G_{ini} maximization

The maximization process of values of edges works as follows. All edges in G_{ini} are processed in each snapshot. For each snapshot m_i, each particular edge e in G_{ini} is tracked using the previously mentioned procedure for tracking a channel [15]. Recall that the procedure returns the set of Voronoi edges $e_{track} = \{e_{t_1}, ..., e_{t_n}\}$ which are dual to the tetrahedra intersected by e. The bottleneck of edges in e_{track} is determined and compared against the value of e. If the bottleneck is larger, then the value of e is updated. Additionally, the actual snapshot m_i for this edge as well as its optimized geometry e_{track} is stored for further reconstruction. The process of updating one edge is depicted in Fig. 3.

After processing of all snapshots, the value of each edge in G_{ini} is maximal in the whole trajectory.

It is clear that it would be time consuming to apply the tracking to each single edge e in G_{ini} in each snapshot. Since the edges e_{track} returned as a result of tracking procedure for e are dual to tetrahedra in DT, the last tetrahedron intersected (i.e., the tetrahedron containing end node of e)

 ISBN: 978-1-61208-137-3

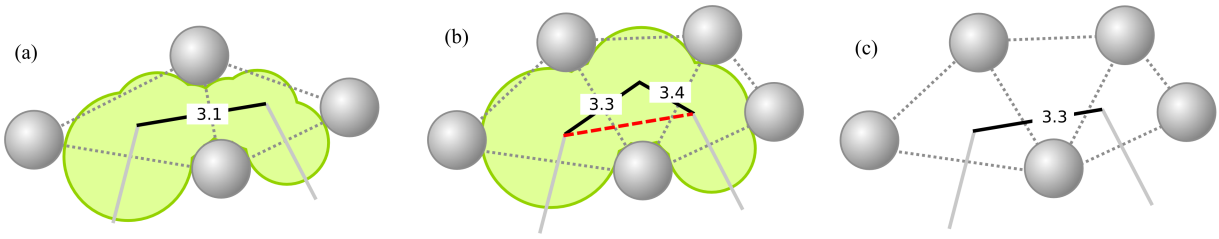


Figure 3. Edge update. (a) Edge in G_{ini} with so far maximal value. (b) Tracked edges e_{track}; bottleneck value: 3.3Å. (c) Edge value is updated.

can be stored with the node in G_{ini} and reused when edges emanating from this node are processed. Together with the breadth-first search (BFS) processing of edges from the starting node it is ensured that when an edge is processed, the tetrahedron for the starting node of the edge is known and the expensive tetrahedron location test has to be performed for the starting node only.

C. G_{ini} Dijkstra's algorithm processing

Finally, updated G_{ini} is processed by the Dijkstra's algorithm in the same way as in the static case [1]. Typically, the path in G_{ini} with the biggest bottleneck which connects the starting node (located in the active site) with arbitrary boundary node is computed. Such path defines the centerline of a dynamic channel. Naturally, for each edge in the centerline of the resulting channel, the information about its geometry as well as the number of snapshot in which the edge is valid is known.

The Dijkstra's algorithm cost function is modified in the way such that channels with the biggest possible bottleneck are computed. In addition, it is ensured that if a boundary node is selected during the algorithm progress, the path with the biggest bottleneck connecting the starting node with the boundary is found. At this point, the computation is terminated and the resulting channel is reported. Alternatively, certain edges in G_{ini} can be disabled and the Dijkstra's algorithm may be run again to find another different dynamic channel. Such approach is widely used when computing channels in static molecules, see [16].

D. Complexity

The number of nodes in G_{ini} is equal to the number of Voronoi edges from any snapshot in case the VD is used. If the grid solution is used, the number of edges is appropriate to the sampling density. Let i be the number of edges in G_{ini}. The theoretical maximum number of tetrahedra processed to track the i edges in G_{ini} in one snapshot is $\mathcal{O}(i \cdot n^2)$. However, for the analysed real data, the expected complexity is significantly better. Each edge in G_{ini} intersects only a limited number of tetrahedra in the DT computed for each snapshot. With the BFS applied, the expected time to track all i edges in G_{ini} is linear with respect to the number of tetrahedra in the DT in each snapshot. After processing all snapshots, the Dijkstra's algorithm is run. Its complexity is $\mathcal{O}(i^2)$ in the worst case. Again, the expected time is smaller since the algorithm can terminate before processing all edges in G_{ini} if a path which ends in a boundary node is computed.

E. Dynamic channels and channel states

Once a dynamic channel is computed, its actual geometry can be visualized in the corresponding valid snapshots. This means that a user would view the snapshots in a trajectory and the corresponding parts of a channel would be visualized in their respective snapshots.

In addition, the behaviour of the empty space near the whole centerline of the dynamic channel can be computed using the method in [15]. In each snapshot, the centerline is tracked and the resulting geometry can be visualized. In this manner, the user can get a state of the dynamic channel in each snapshot and can get a complex view on the behaviour of the molecule near the dynamic channel. Alternatively, the method proposed in [18] could be used. The method can effectively update the triangulation near the centerline of the dynamic channel and use it for the determination of states of dynamic channel in each snapshot.

IV. Results

To show that spatial changes appear in the molecule, we analyzed the width of all edges in G_{ini} for each snapshot. As the input data, trajectories of haloalkane dehalogenase DhaA were used (wild-type wt, mutated with codenames 04, 14, 15; more details on the data can be found in [17]). The number of updates and the average edge value were determined. As mentioned in the previous section, the update happens only when the value of an edge in the current snapshot is larger than the value of the corresponding edge in G_{ini}. As shown in Fig. 4 (a) the average edge value has increased after processing each snapshot. Moreover, the increase is relatively high which indicates that there are significant changes in the behaviour of empty space inside the protein. In addition, the results indicate that the value of edges is increasing despite the fact that the number of edge updates is relatively low and decreases (see Fig. 4 (a, b)). After processing a certain number of snapshots we can compute the channel in G_{ini} with the largest bottleneck – the

 ISBN: 978-1-61208-137-3

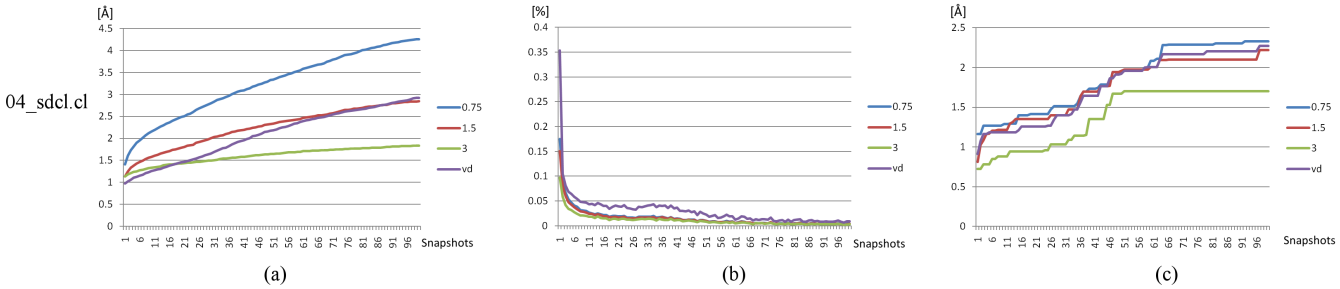


Figure 4. The analysis of 04_rdcl.cl and 04_sdcl.cl (details in [17]). (a) The average width of edge in G_{ini} after processing certain number of snapshots. (b) The percentage of edges updated after adding each snapshot. (c) The bottleneck radius of the widest dynamic channel computed after adding i-th snapshot. Similar trends were observed for other analysed trajectories.

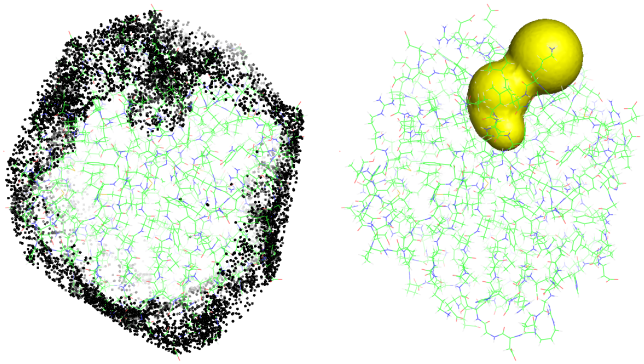


Figure 5. Example on 04_sdcl.cl (a) Edges with value above 3.0Å emphasized – the midpoint of each such edge visualized as a small sphere. The visualization is clipped with a clipping plane perpendicular to the view in order to reveal the inner parts of the molecule. (b) Computed dynamic channel. The molecule is visualized as a line model, first snapshot of the trajectory. Visualized in PyMOL [19].

Table I
COMPARISON OF BOTTLENECK RADII. THE BOTTLENECK OF A DYNAMIC CHANNEL IS COMPARED WITH THE CHANNEL WHICH HAS THE MAXIMUM BOTTLENECK IN THE WHOLE SET OF CHANNELS COMPUTED SEPARATELY IN EACH SNAPSHOT.

Trajectory	**Bottleneck radius**	
	Dynamic channel (VD G_{ini} variant)	**Set of channels** (computed separately, maximum)
wt_sdcl.cl	2.798 Å	2.021 Å
04_sdcl.cl	3.051 Å	1.968 Å
14_sdcl.cl	2.568 Å	2.079 Å
15_sdcl.cl	2.601 Å	1.725 Å

dynamic channel. The results show that there is an increase in the bottleneck radius of the dynamic channel. Such an increase with respect to the number of snapshots processed is shown in Fig. 4 (c). In practical case, the computation of a dynamic channel would be performed only once after processing the whole sequence.

In the experiments, the computed dynamic channels for both proposed G_{ini} variants led in general through the same parts of the molecule. Therefore, for the following tests, we chose the VD variant that we consider to be more prospective for its better computational time.

Fig. 5 shows the visualization of a computed dynamic channel. Edges in G_{ini} with value above certain threshold are visualized as small spheres located at midpoints of each edge (a). The surface of the channel is shown in (b).

The computed dynamic channel was compared against the channel (from the set of channels computed separately in each snapshot) with the biggest bottleneck radius. The comparison shows that the radii of dynamic channels computed for selected trajectories are significantly larger implying that the molecule is flexible at that regions and potentially a small ligand molecule may pass through. Table I illustrates the results of the comparison for protein trajectories of wild-type (wt) and mutated (04,14,15) haloalkane dehalogenase DhaA molecules.

For example, the time requirement to update G_{ini} (maximization) in one snapshot (approx. 4500 atoms), in the case VD is used, is below 8 seconds on the common desktop computer (single-threaded, 2.0GHz, 2GB RAM). We have to point out, that the time for processing the trajectory is linearly dependent on the number of snapshots in the trajectory.

After snapshots are processed, the Dijkstra's algorithm is to be run. The complexity, again, is not dependent on the number of snapshots previously processed since the topology of G_{ini} remains constant over time. The Dijkstra's algorithm running time on the previously mentioned sample G_{ini} and computer is below 5 seconds. Unlike the update of G_{ini} in each snapshot, the Dijkstra's algorithm is run only once after all snapshots are processed.

V. CONCLUSION

We have presented a method capable of computing dynamic channels. Dynamic channels are optimized for the whole trajectory and not for one particular snapshot only. Our method is able to consider the whole trajectory at once – computed dynamic channels are composed of parts from different snapshots.

 ISBN: 978-1-61208-137-3

Dynamic channels are able to highlight pulsing local neighbourhood and parts of the molecule with high flexibility. A ligand may pass through such pulsing areas.

In combination with previous methods, the proposed solution may help chemists to find possible paths which could provide an access to the active site. The residues surrounding the selected channel in the molecule could be replaced so that the active site becomes either more easily accessible or, on the contrary, inaccessible through the particular part of the molecule.

We expect that, by using this method, new information about the behaviour of various protein molecules will be revealed. After the development of sophisticated visualization methods the algorithm will be integrated into protein visualization software Caver Viewer (http://www.caver.cz).

The testing version of the algorithm was implemented in Java programming language and is available for download from http://decibel.fi.muni.cz/~xbenes2/dynChannels.

ACKNOWLEDGMENT

This work was supported by The Ministry of Education of The Czech Republic, Contract No. LC06008 and by The Grant Agency of The Czech Republic, Contract No. P202/10/1435 and GA201/09/0097. We would like to acknowledge Loschmidt Laboratories, Masaryk University for provided protein data.

REFERENCES

[1] P. Medek, P. Beneš, and J. Sochor, "Computation of tunnels in protein molecules using delaunay triangulation," *Journal of WSCG*, vol. 15(1-3), pp. 107–114, 2007.

[2] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design." *Protein Sci*, vol. 7, no. 9, pp. 1884–1897, September 1998. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/9761470

[3] K. Rother, P. W. Hildebrand, A. Goede, B. Gruening, and R. Preissner, "Voronoia: analyzing packing in protein structures." *Nucleic Acids Research*, vol. 37, no. Database-Issue, pp. 393–395, 2009. [Online]. Available: http://dblp.uni-trier.de/db/journals/nar/nar37.html

[4] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: An open source platform for ligand pocket detection," *BMC Bioinformatics*, vol. 10, pp. 168+, June 2009. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-10-168

[5] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, and J. Damborský, "Caver: a new tool to explore routes from protein clefts, pockets and cavities," *BMC Bioinformatics*, vol. 7, no. 1, pp. 316+, June 2006. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-7-316

[6] M. Petřek, P. Košinová, J. Koča, and M. Otyepka, "Mole: A voronoi diagram-based explorer of molecular channels, pores, and tunnels." *Structure*, vol. 15, no. 11, pp. 1357–1363, November 2007. [Online]. Available: http://dx.doi.org/10.1016/j.str.2007.10.007

[7] E. Yaffe, D. Fishelovitch, H. J. Wolfson, D. Halperin, and R. Nussinov, "Molaxis: Efficient and accurate identification of channels in macromolecules," *Proteins: Structure, Function, and Bioinformatics*, vol. 73, no. 1, pp. 72–86, 2008. [Online]. Available: http://dx.doi.org/10.1002/prot.22052

[8] B. Ho and F. Gruswitz, "Hollow: Generating accurate representations of channel and interior surfaces in molecular structures," *BMC Structural Biology*, vol. 8, no. 1, p. 49, 2008. [Online]. Available: http://www.biomedcentral.com/1472-6807/8/49

[9] B. J. Alder and T. E. Wainwright, "Studies in Molecular Dynamics. I. General Method," *jcp*, vol. 31, pp. 459–466, Aug. 1959.

[10] S. K. Ldemann, V. Lounnas, and R. C. Wade, "How do substrates enter and products exit the buried active site of cytochrome p450cam? 1. random expulsion molecular dynamics investigation of ligand access channels and mechanisms," *Journal of Molecular Biology*, vol. 303, no. 5, pp. 797 – 811, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/B6WK7-45F50WB-39/2/503cb97cac5e6a3a94899d6bc61ed2a8

[11] P. Beneš, P. Medek, and J. Sochor, "Computation of channels in protein dynamics," *In Proceedings of the IADIS International Conference Applied Computing 2009*, vol. 2, pp. 251–258, 2009.

[12] F. P. Preparata and M. I. Shamos, *Computational geometry: an introduction*. New York, NY, USA: Springer-Verlag New York, Inc., 1985.

[13] M. L. Gavrilova and J. Rokne, "Updating the topology of the dynamic voronoi diagram for spheres in euclidean d-dimensional space," *Comput. Aided Geom. Des.*, vol. 20, no. 4, pp. 231–242, 2003.

[14] D. D. Barber, C.B. and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software*, vol. 22(4), pp. 469–483, 1996.

[15] P. Beneš, P. Medek, and J. Sochor, "Tracking single channel in protein dynamics," *In WSCG Communication Papers proceedings*, vol. 2, pp. 109–114, 2010.

[16] P. Benes, P. Medek, and J. Sochor, "Computation of more channels in protein molecules," *In Proceedings of Visual Computing for Biomedicine*, vol. 7, pp. 45–51, 2008.

[17] M. Klvaňa, "Structure-dynamics-function relationships of haloalkane dehalogenases," Ph.D. dissertation, Faculty of Science, Masaryk University, Brno, Czech Republic, 2010.

[18] M. Zemek and J. Skala, "Fast method for computation of channels in dynamic proteins," in *Vision, Modeling and Visualization 2008*. Heidelberg: Akademische Verlagsgesellschaft Aka, 2008, pp. 333–343.

[19] W. L. Delano, "The pymol molecular graphics system," Palo Alto, CA, USA., 2002.

 ISBN: 978-1-61208-137-3

EnzymeTracker: A Web-based System for Sample Tracking with Customizable Reports

Thomas Triplet*§, Justin Powlowski†§, Adrian Tsang‡§ and Gregory Butler*§

*Department of Computer Science and Software Engineering, Concordia University
1455 De Maisonneuve Blvd. West, Montreal, Quebec, H3G 1M8, Canada
Email: {thomastriplet@gmail.com} {gregb@encs.concordia.ca}

†Department of Chemistry and Biochemistry, Concordia University
7141 Sherbrooke Street West, Montreal, Quebec, H4B 1R6, Canada
Email: powlow@alcor.concordia.ca

‡Department of Biology, Concordia University
7141 Sherbrooke Street West, Montreal, Quebec, H4B 1R6, Canada
Email: tsang@gene.concordia.ca

§ Centre for Structural and Functional Genomics
7141 Sherbrooke Street West, Montral, Quebec, H4B 1R6, Canada

***Abstract* — In many laboratories, researchers store experimental data on their own workstation using spreadsheets. However, this approach poses a number of problems, ranging from versioning or sharing issues to inefficient data-mining. Standard spreadsheets are also error-prone as data do not undergo any validation process. In this paper, we propose the EnzymeTracker, a web-based laboratory information management system for sample tracking, as a robust and flexible alternative that aims at facilitating entry, mining and sharing of experimental biological data. The EnzymeTracker features online spreadsheets and tools for monitoring numerous experiments conducted by several collaborators to identify and characterize samples, from their basic functional annotations to their complete enzymatic activity. It also provides libraries of shared data such as protocols, and administration tools for data access control using OpenID and user/team management. Our system relies on a database management system for efficient data indexing and management and a user-friendly AJAX interface that can be accessed over the Internet. The EnzymeTracker facilitates data entry by dynamically suggesting entries and providing smart data-mining tools to effectively retrieve data. It also features a number of tools to visualize and annotate experimental data, and export customizable reports. The EnzymeTracker is available online at http://cubique.concordia.ca/enzymedb/index.html under the GNU GPLv3 licence.**

Keywords* — *laboratory information management; enzyme; data warehousing; data integration

I. INTRODUCTION

Spreadsheets (like Excel) are broadly used by the scientific community. Their intuitive and easily understandable user interface is a significant advantage. They are also visually appealing and feature a number of tools to visualize data using charts. Hence, spreadsheets are currently the primary means to store both experimental and manually curated genomics/proteomics data in most laboratories.

A. Spreadsheet/database paradigm

Spreadsheets might be sufficient when one needs to organize simple data. However, this approach raises a number of problems as spreadsheets present numerous well-known deficiencies compared to databases when dealing with involved data. As reported in previous studies [1], [2], [3], [4], spreadsheets do not scale up well and, as the spreadsheet will expand to accommodate a growing number of records of increasing complexity, data handling — from data entry to data mining and analysis — will become increasingly cumbersome, hence reducing the utility of potentially valuable information.

Besides the scalability issue, spreadsheets are subject to data redundancy and consequently data integrity loss. For example, if protein annotations should be displayed in different spreadsheets, they will most likely be duplicated in each document. When an annotation is updated in one place, all occurrences elsewhere may not be updated, which will result in multiple inconsistent versions of the same data. In this case, one does not know which versions are obsolete and which version is correct. Moreover, unlike databases, spreadsheets do not enforce referential integrity: they do not check that resources referenced somewhere in the spreadsheet are still valid, which may be critical, in

 ISBN: 978-1-61208-137-3

particular when those resources are frequently updated or deleted.

Spreadsheets are also error-prone and do not facilitate data entry. Typically, any cell can contain any type of data and validation is optional at best. Spreadsheets may even incorrectly infer a data type based on the data, in particular numbers and dates in Excel. Spreadsheets are also inefficient to handle sparse data, both in terms of storage and performance. Storage is less of a concern nowadays as costs have dramatically decreased in the past few years. However, it should still be taken into consideration when handling millions of records, as is often the case in bioinformatics and large-scale studies in general. In contrast, optimized databases lead to speed improvements.

Furthermore, sharing data using spreadsheets proved to be difficult, when possible. For example, a shared Excel spreadsheet can be checked-out and edited by only one user at a time. Other collaborators can only display a read-only copy of the document until changes are committed by the first user. Neither waiting for a user to complete his work or duplicating resources is a practical satisfactory solution in larger work groups.

Finally, spreadsheets provide little — if any — security or access control mechanisms. Spreadsheets can be password-protected. However, the password of the spreadsheet is unique and known by many users, and they do not offer the possibility to select what users or groups of users can see/edit in the document: once opened, any record can be displayed by the user. The password is also embedded within the document and it is therefore not possible to revoke access remotely. Databases on the other hand provide advanced access control mechanisms, and enable system administrators to precisely grant or revoke permissions to users or groups of users to create, view, update or delete resources as needed.

B. Technology acceptance issue

Despite their deficiencies, spreadsheets have been heavily used by biologists because they offer an intuitive and generic user interface that is applicable to most of their projects. Upgrading from spreadsheets to a more sophisticated laboratory information management systems (LIMS) is not trivial. To be broadly accepted by the scientific community as a valuable replacement for spreadsheets, LIMS need to present the five acceptance characteristics defined by Rogers [5]:

- *relative advantage*: the extent to which the LIMS offers improvements over spreadsheets,
- *compatibility*: its consistency with social practices and norms among its users,
- *complexity*: its ease of use or learning,
- *trialability*: the opportunity to try an innovation before committing to use it,
- *observability*: the extent to which the technology's gains are clear to see.

In this paper, we propose the EnzymeTracker, a generic web-based laboratory information management system for sample tracking, as an efficient and user-friendly alternative that aims at facilitating entry, mining and sharing of samples and experimental biological data. Our system was designed to present the above acceptance characteristics to maximize its utility and features advanced yet intuitive annotation and visualization tools as well as a flexible and customizable report designer.

Sections III and IV give an overview of the unique features of the EnzymeTracker and its web-based graphical user interface respectively. Section V briefly presents a number of visualization tools embedded within our system. Section VI describes data-mining and the generation of reports based on templates. Finally, Sections VII and VIII gives some implementation details and future directions respectively.

II. Related work

To overcome spreadsheets limited capabilities, a number of proprietary LIMS have been developed. However, their expensive license reduces their audience to bigger laboratories or to the industry and very few systems are freely available to the general scientific community.

A. iLAP

Stocker et al. [6] recently developed iLAP, a workflow-driven software for experimental protocol development, data acquisition and analysis. iLAP relies on a relational database and a web-based interface to effectively manage complex work flows derived from biological experimental protocols. Integration of external programs using Java Applets is also possible, in particular the popular image processing library ImageJ [7]. However, iLAP does not manage biological data directly, as data remain in files that should be uploaded and associated with a specific experiment or protocol. It is therefore not possible to search for a particular piece of biological data. iLAP does not provide tools for annotating pictures from experimental results such as SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis) gels or microplates, nor does it provides facilities to generate reports.

B. SLIMS

Daley et al. [8] developed SLIMS, a Sample-based Laboratory Information Management System. SLIMS is a web application that provides members of a laboratory with an interface to view, edit, and create sample information. Unlike iLAP, SLIMS leverages the relational database to store and manage biological data. However, its web-interface does not utilize recent advances in web technologies. For example, most data are displayed to the user as static HTML tables, which cannot be dynamically mined nor customized. SLIMS also features a microplate annotation tool. Microplate pictures, though, cannot be uploaded nor

 ISBN: 978-1-61208-137-3

visualized along with their annotations. Similarly, SLIMS supports SDS-PAGE gels, which can be downloaded as plain text files, but may not be properly visualized using the picture of the gel. Reports can be generated and exported, but cannot be customized.

III. Overview of the EnzymeTracker

Despite their numerous benefits over spreadsheets, database management systems still lack satisfactory user interfaces for data analysis [9] whereas Excel spreadsheets do provide intuitive graphical interfaces for data analysis and consolidation, provided the issues mentioned above are addressed.

Web-based applications are dynamic and interactive websites that offer a rich user interface comparable to standard desktop programs [10], [11]. They can be executed on any connected workstation, without software installation nor specific requirements besides a recent web-browser and an active Internet connection to remotely access data. Web applications have the major advantage of being always up-to-date wherever they are being accessed, thereby eluding the need for multiples copies of the same document on different workstations, effectively solving synchronization issues between local copies.

The EnzymeTracker was thus designed as an integrated collection of online spreadsheets accessible over the Internet and backed-up by a relational database for efficient data management. It features a number of novel online tools to facilitate data entry and visualization. The EnzymeTracker also provides a library of shared records such as experimental protocols for sample assays and a comprehensive set of reporting and system administration tools.

Figure 1 gives an overview of the graphical user interface (GUI). Most pages are composed of three panels: the main menu (A) on the left, a spreadsheet (B), which is the primary means to enter to enter data, and a panel at the bottom (C), whose content depends on the data to display. Others data entry means are presented later in sections V and IV-B. Panels A and C can be dynamically collapsed and resized to customize the workspace as needed. Spreadsheets may also be customized by displaying, hiding, reordering and resizing columns as needed so that only the most relevant data are displayed.

The content of the lower panel (C) varies with the data being shown. On most pages, the panel displays the record selected in the spreadsheet in a more readable format. Depending on the spreadsheet, it can provide links to cross-referenced databases such as the Gene Ontology [12] or the Clusters of Orthologous Groups of proteins (COG) database [13]. It also automatically fetches complete references from the literature using PubMed's public API (http://eutils.ncbi.nlm.nih.gov/) given the PMID of an article and jobs for nucleotide or protein sequence alignment can be submitted to NCBI's BLAST server in one click (F).

IV. Data Entry

The spreadsheet (Figure 1B) is the primary means of entering data in the EnzymeTracker. Each cell is associated with an editor whose format depends on the data within the cell. Most cell editors are simple text fields. More advanced editors are provided where needed. In particular, cross-references to other tables are typically associated with a combo box, whose content is dynamically generated after the content of the referenced table. Figure 1D illustrates the utilization of a combo box to select a clone in the page for *Annotations*. Combo boxes facilitate data entry by suggesting entries as the user types. They also have the added benefit of limiting data entry mistakes, in particular when users enter data that do not exist in the referenced table. Specific editors are also provided for Boolean flags and dates. The EnzymeTracker also supports rich text editors with text formatting capabilities, which are mainly used for comments and free-text cells.

A. Data integrity and validation

To further reduce entry errors, each cell editor can be associated with a *validator*. Validators, which are usually based on regular expressions or more advanced customized functions, check the correctness of data types and send immediate feedback to the user in case of an error. Validators are also useful to enforce data entry conventions and consistency within a group of users.

In addition, to minimize data entry, cells are automatically computed whenever possible. For example, the length of a protein sequence and its molecular weight (E) are automatically calculated when one enters a protein sequence. Calculated fields are also used to reduce data redundancy compared to standard spreadsheets. For instance, the name of a protein should appear on several related spreadsheets. Using standard spreadsheets, the user will copy/paste the name of the protein wherever needed. This will lead to inconsistencies between spreadsheets during their update. In the EnzymeTracker, the underlying relational database is leveraged to display the name of the protein in all tables where it is needed. The first benefit is that the protein name is automatically displayed whenever there exists a relation between proteins and the current spreadsheet. Second, changes to the protein data are automatically reflected in all tables. Data in the various online table are therefore always consistent and up-to-date.

B. Data importation/exportation

In some cases, the different enzyme assays and characterization of samples were already being recorded using Excel spreadsheets. We therefore implemented importation routines to facilitate the migration process to the EnzymeTracker. From experience, basic data importation by uploading and parsing files is error-prone as files formats and layouts tend to vary between files. For example, one column

 ISBN: 978-1-61208-137-3

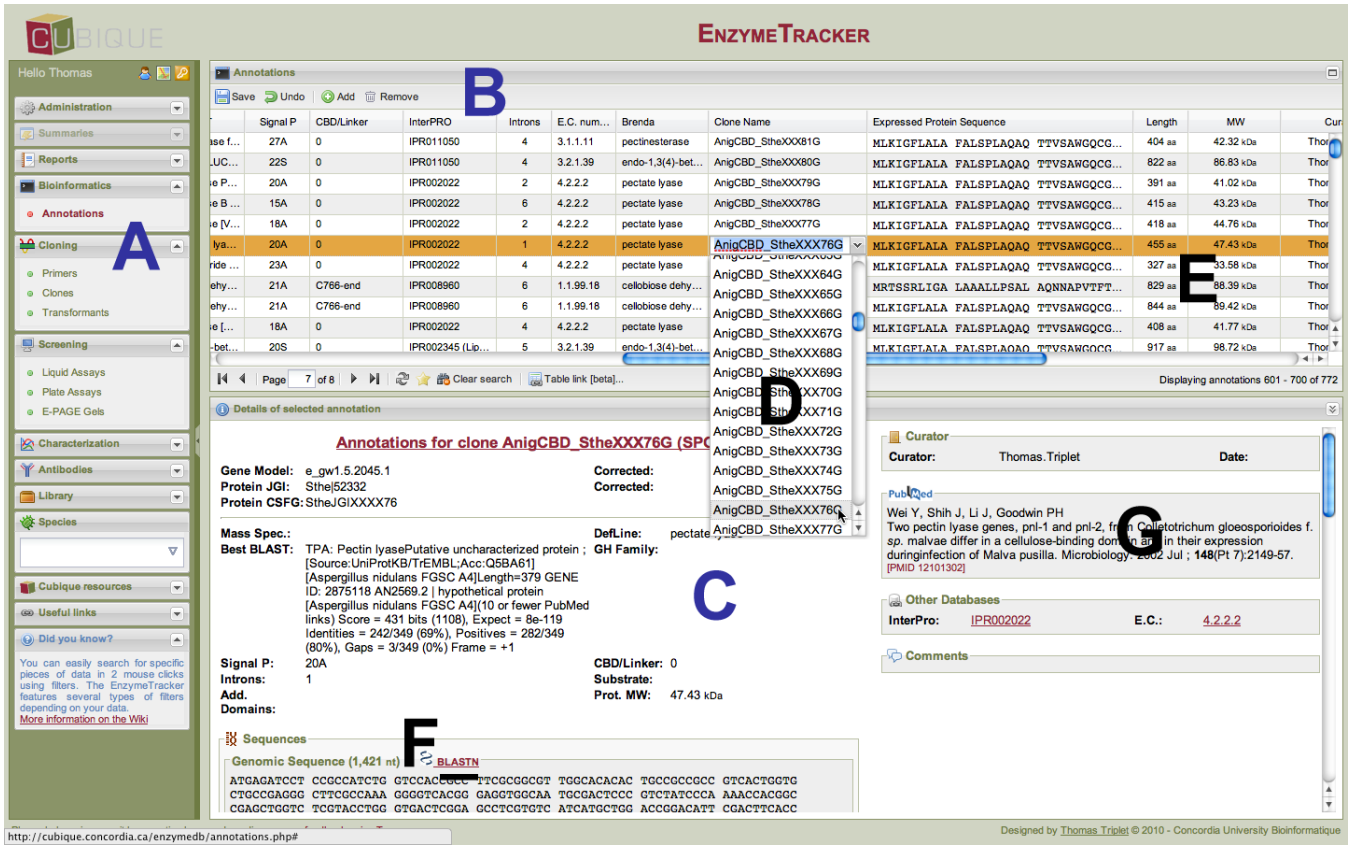


Figure 1. Screenshot of the web-based user interface of the *Annotation* page of the EnzymeTracker. The main menu (A) is on the left. The main panel is usually composed of a spreadsheet (B) and a panel at the bottom to display the entry selected in the spreadsheet using a more readable layout (C). See details in the text.

may be missing in one file, which will shift other columns and lead the parser to import the wrong data.

Instead, we implemented a drag-and-drop importation mechanism where appropriate. The user selects the data to import in the Excel file and drags and drops the selection into the browser's window. The major benefit of this semi-automatic approach is that it makes it easier for the users to review the data before importation, hence reducing the number of errors made. It also gives more flexibility as only specific records can be selected and imported. Finally, users have the possibility to export EnzymeTracker spreadsheets to Excel documents in one click. Data may also be imported programmatically, using JavaScript and RESTful requests.

C. Versioning and backups

Our system is supported by a relational database which efficiently handles versioning and backups. Unlike standard spreadsheets, when a record is updated or deleted in the EnzymeTracker, the current version of the record is flagged as obsolete, backed-up and logged for future reference. As a consequence, while updating a spreadsheet is always possible, no data are ever deleted and restoring a record to a previous state or accessing the complete data modification log in case an error is made while updating a spreadsheet is always possible.

V. VISUALIZATION TOOLS

Most data in the EnzymeTracker can be viewed using tables. In a number of cases however, tables may be improved to give the user a more visual perspective of the data. To enhance the utility of experimental screening data, the EnzymeTracker integrates a number of annotation and visualization tools. Sections V-A and V-B describe in detail how the bottom panel of a spreadsheet (Fig. 1C) can be customized to accommodate plate assays and E-PAGE™48 gels from Invitrogen respectively.

A. Microplate assays

Microplate assays are widely used in research and drug discovery to detect biological or chemical events of samples. Those events are typically detected by measuring the fluorescence intensity of samples from each of the ninety-six wells (labeled A1 to H12) that compose a plate. The plate assay is usually repeated twice, at two different sample dilution factors.

 ISBN: 978-1-61208-137-3

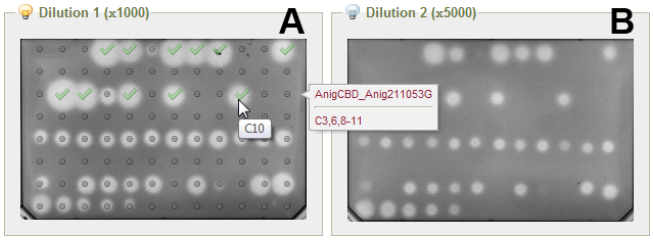


Figure 2. Graphical user interface for the annotations of plate assays. Pictures of the microplates for the two dilutions can be uploaded and automatically annotated based on the content from the tables describing *clones* and *transformants*. High-activity wells can be selected within the web interface by clicking on picture. Annotations can be laid over the picture (A) or hidden (B) as needed.

The EnzymeTracker enables users to upload the two microplate pictures for the two dilutions of each experiment (Figure 2). The tables describing *clones* and *transformants* are leveraged to automatically annotate the plate. A "virtual plate" representing the 96 wells can also be layered over the original picture (A) or hidden (B) as needed. The virtual plate is also convenient to quickly visualize and identify most active wells by simply clicking on the desired wells directly on the picture.

B. SDS-PAGE gels

E-PAGE™48 gels are improved SDS-PAGE (Sodium Dodecyl Sulfate PolyAcrylamide Gel Electrophoresis) gels broadly used for high-throughput protein separation and analysis. Each gel comprises 48 lanes for samples and 4 marker lanes, which define the ladder of the molecular weights of the proteins on the gel. Similarly to plate assays, the picture of the gel can be uploaded and annotated within the user interface (Figure 3). Each sample lane in the gel (A) can be annotated using a form (B) that is displayed upon click. A tooltip summarizing annotations of a lane is displayed when hovered by the cursor (orange). The drop-down menus in (B) to select the clone and the transformation

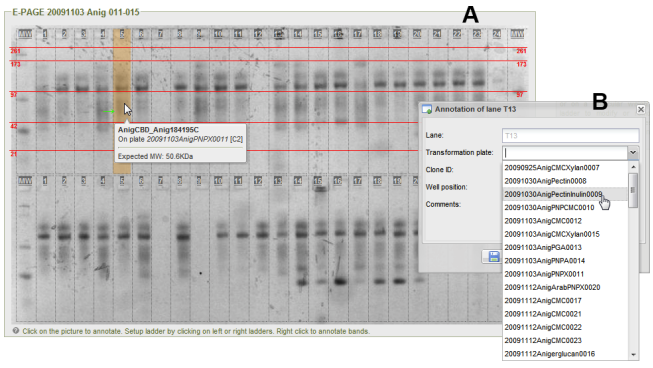


Figure 3. Graphical user interface for the annotations of E-PAGE™48 gels from Invitrogen. Pictures of the gels can be uploaded and annotated within the web interface of the EnzymeTracker.

plate loaded in each lane are dynamically built based on their respective tables. In addition, specific bands can be highlighted (green arrows) and annotated. Finally, the ladder (red) can be easily setup by clicking on one of the four outer marker lanes.

The EnzymeTracker eludes the need for external tools and leverages data from other spreadsheet to facilitate the annotations of hundreds of experimental data and to reduce data entry errors.

C. Chart visualization

In many cases, the experiments aim at characterizing the evolution of a variable given a set of parameters. Representing the data using charts is then is suitable alternative to tables for data presentation.

The EnzymeTracker fully supports charts, in particular to represent the enzymatic activity characterization of a sample. Graphs are usually represented using curves although histograms and pie charts are also supported. The graph is dynamically updated when the underlying data is edited within the interface or imported from Excel as described in Section IV-B.

Graphs are also used in the administration console, in particular to display connection and data logs.

VI. Data-Mining and Reporting

As of January 2011, over 55,000 entries have been saved within the EnzymeTracker and a growing number are being recorded on a daily basis. Despite these large amounts of heterogeneous data, scientists routinely need to search for specific pieces of information. For example, a principal investigator may look for "all enzymatic activities detected during liquid assays performed by his assistants in the past two months on clones from *S. thermophile*".

A. Context-dependent filtering

Each table in the EnzymeTracker is fully searchable and each column is associated with a flexible filter that depends on the type of data the column represents. Five different types of filters can be configured: textual, multi-selection, numerical, calendar and Boolean. Numerical filters let the user query for values above, below or equal to a given threshold. They are most useful to query biochemical properties of enzymes and samples, for example protein sequence length or molecular weight, or the temperature stability of a molecule. Boolean filters are typically used to retrieve records when given a flag. For instance, this filter is convenient to list all assays where a strong activity has been reported. Calendar filters are helpful to search for records given a time frame. The multi-selection filter is most effective for searching for one or more items in a given list. The list may be static or may be dynamically generated by the server based on data from other tables. For example, it is possible to search for samples from a given organism,

 ISBN: 978-1-61208-137-3

the list of organisms being automatically generated by the database server.

B. Reporting

In order to facilitate data sharing among collaborators, the EnzymeTracker provides a flexible and easy-to-use tool for designing report templates. A report template is similar to other tables within the EnzymeTracker, except that the user can dynamically select the pieces of information he/she wants to share. It is also valuable to aggregate data from various tables and display consolidated statistical data. For instance, one can easily create a report template to display the percentage of transformants which were successfully assayed.

The EnzymeTracker allows users to quickly design a report template and assemble relevant pieces of information together. A preview of the report can be automatically displayed when the configuration of the report changes or when filters are set. The report can also be refined using a number of flags, for example to decide whether to display only current values of a record or its modification log also.

More technically, when a report is designed, the corresponding SQL query is automatically generated based on the report configuration. In other words, the designer effectively enables users to design SQL queries, without writing any code. This is particularly useful for more complex queries, such as non-trivial joins: when a user selects two items from two different tables, it may happen that the two tables are not *directly* related. In that case, a number of intermediate tables must be used in order to join the two tables.

For example, consider the case when the user needs to list the plate assays performed on clones related to cellulase.

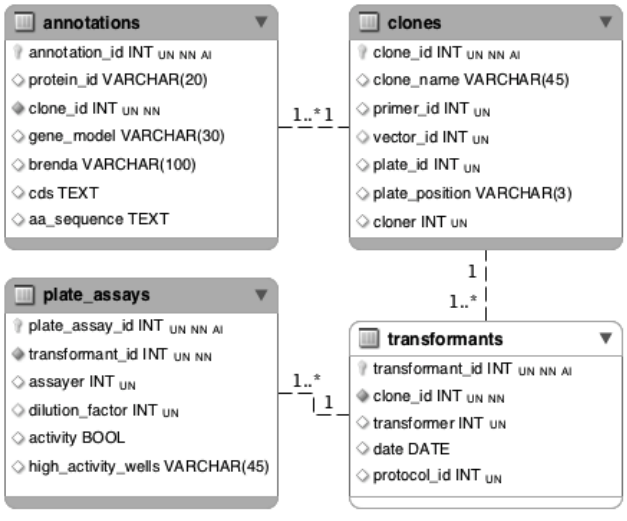


Figure 4. Simplified EER to illustrate reporting mechanisms. Foreign keys are marked with solid bullets. Shaded tables are explicitly listed in the configuration of the report. The *transformants* table (white) is not listed but is implicitly required to perform the join query.

Figure 4, which illustrates a simplified Entity-Relationship diagram of tables relevant to generate this report, shows that plate assays are performed on transformants, not on clones directly. Fortunately, transformants are related to clones, hence it is possible to define an *implicit* relation between clones and plates assays using transformants.

We defined the *cost* of a join between two tables as the length of the shortest path between the two tables in the undirected weighted graph implied by the database structure, where the nodes represent tables and edges, foreign keys. Using the above example, the cost of the join between *clones* and *plate assays* is 2.

We designed an algorithm to determine the optimal implicit joins to relate two tables, which is derived from Dijkstra's shortest path algorithm. The optimal join is defined as the path between the two tables with the lowest cost. Edges were weighted based on the *biological* significance of the foreign keys. For instance, because of the normalization of the database, a number of intermediate joining tables are created to define the relationships between *real* biological entities — in particular in $m : n$ relations — which incorrectly increases the cost of the relationship as the path between the two biologically meaningful tables is longer. The cost of edges in $m : n$ relations was therefore reduced to avoid the bias induced by the normalization process during the database design.

When the configuration of the report is updated, the optimal join is computed and executed by the SQL engine. The results of the query are finally used to build the configuration of the ExtJS spreadsheet used to display the report. Advanced users can also create a report by typing the SQL query directly.

Once a template is created, it can be shared and displayed like other tables. In particular, the report can be further refined using filters as described in section VI-A. In addition, reports are automatically updated as more data is added to the EnzymeTracker: there is therefore no need to re-design a report to display up-to-date data. Finally, reports can be easily shared with collaborators or saved as standard Excel files for further analysis.

VII. Implementation and Availability

The EnzymeTracker aims at providing an interactive web-based user interface. To achieve this goal, the EnzymeTracker of composed of a set of highly dynamic web pages implemented using AJAX (Asynchronous JavaScript and XML) web technologies [11], which enable a web application to communicate with a server in the background using JavaScript and *XMLHttpRequest* objects, without interfering with the current state of the page. The web user interface of the EnzymeTracker was implemented using ExtJS, the general Asynchronous JavaScript and XML (AJAX) framework from Sencha. It is backed-up by the freely available

 ISBN: 978-1-61208-137-3

MySQL relational database management system. The server-side code was implemented using PHP 5.

The EnzymeTracker and its documentation are available at http://cubique.concordia.ca/enzymedb/index.html under the GNU General Public License version 3.

VIII. Usage and future directions

The EnzymeTracker was designed to be flexible, easy to use and offers many benefits over spreadsheets, thus presenting the characteristics required to facilitate acceptance by the scientific community. The EnzymeTracker has been successfully used for 15 months on a daily basis by over 50 scientists to monitor protocols and experiments conducted to identify, annotate and fully characterize thousands of samples from multiple fungal species.

The initial implementation of the EnzymeTracker has focused on facilitating sample tracking and experimental data annotation and visualization. The future development of the EnzymeTracker will focus on the implementation of widgets based on the online spreadsheets, which will facilitate data sharing as widgets can be embedded in virtually any web page. We will also enhance reporting by allowing chart generation in addition to tabular data. Finally, the EnzymeTracker will be expanded to enable bar-coding of samples using QR codes (two-dimensional matrix codes), which will facilitate the identification, physical tracking and long-term storage of samples.

Acknowledgement

We thank Annie Bellemare, Noutcheka St-Felix, Marek Krajewski and their teams for providing data and their valuable feedback.

Funding

This work was supported by the Cellulosic Biofuel Network, funded by Agriculture and Agri-Food Canada.

References

[1] K. J. Gordon, "Spreadsheet or database: Which makes more sense?" *Journal of Computing in Higher Education*, vol. 10, no. 2, pp. 111–116, Mar. 1999.

[2] J. Pemberton and A. Robson, "Spreadsheets in business," *Industrial Management & Data Systems*, vol. 100, no. 8, pp. 379–388, 2000.

[3] E. V. Denardo, "The Science of Decision Making: A Problem-Based Approach Using Excel," *OR/MS Today*, vol. 28, no. 4, 2001.

[4] B. B. Gansel, "About the Limitations of Spreadsheet Applications in Business Venturing," in *Operations Research Proceedings*, ser. Operations Research Proceedings, J. Kalcsics and S. Nickel, Eds., vol. 2007. Berlin, Heidelberg: Springer, 2008, pp. 219–223.

[5] E. Rogers, *Diffusion of Innovations, 5th Edition*. New York, NY, USA: Free Press, 2003.

[6] G. Stocker, M. Fischer, D. Rieder, G. Bindea, S. Kainz, M. Oberstolz, J. G. McNally, and Z. Trajanoski, "iLAP: a workflow-driven software for experimental protocol development, data acquisition and analysis." *BMC bioinformatics*, vol. 10, p. 390, Jan. 2009.

[7] W. Rasband, "ImageJ," Nov. 2010. [Online]. Available: http://imagej.nih.gov/ij/

[8] D. Daley, M. Lemire, L. Akhabir, M. Chan-Yeung, J. Q. He, T. McDonald, A. Sandford, D. Stefanowicz, B. Tripp, D. Zamar, Y. Bosse, V. Ferretti, A. Montpetit, M.-C. Tessier, A. Becker, A. L. Kozyrskyj, J. Beilby, P. A. McCaskie, B. Musk, N. Warrington, A. James, C. Laprise, L. J. Palmer, P. D. Paré, and T. J. Hudson, "Analyses of associations with asthma in four asthma population samples from Canada and Australia." *Human genetics*, vol. 125, no. 4, pp. 445–59, May 2009.

[9] E. Codd, S. Codd, and C. Salley, *Providing OLAP to User-Analysts: An IT Mandate*. San Jose, CA, USA: Codd & Date, Inc, 1993.

[10] J. J. Garrett, *The Elements of User Experience: User-Centered Design for the Web*. Berkeley, CA: Peachpit Press, 2002.

[11] ——, "Ajax: A New Approach to Web Applications," Aug. 2005. [Online]. Available: http://www.adaptivepath.com/ideas/essays/archives/000385.php

[12] The Gene Ontology Consortium, "The Gene Ontology (GO) project in 2006," *Nucleic Acids Research*, vol. 34, pp. D322–D326, 2006.

[13] R. L. Tatusov, N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin, and D. A. Natale, "The COG database: an updated version includes eukaryotes," *BMC Bioinformatics*, vol. 4, p. 41, 2003.

 ISBN: 978-1-61208-137-3

In silico Identification of Drug Targets in Methicillin/Multidrug-Resistant *Staphylococcus aureus*

Nichole Louise Haag, Kimberly Kay Velk, Chun Wu
Division of Natural Sciences
Mount Marty College
1105 W 8th St. Yankton, the United States
nichole.haag@mtmc.edu
kimberly.velk@mtmc.edu
cwu@mtmc.edu

***Abstract*— This paper reports an extension of an established bioinformatics approach to a new organism involving more than one strains for comparison. Methicillin/multidrug-resistant *Staphylococcus aureus* causes serious infections in humans and becomes resistant to increasing numbers of antibiotics. Our approach utilizing CD-HIT and BLASP *in silico* tools identified 133 and 134 genes in MRSA 252 strain and MRSA Mu50 strain respectively that are essential to pathogen survival with E-score < 10^{-4} and absent in the human genome with E-score < 10^{-3}. The genes were further classified according to their known or hypothetical or putative functions annotated by NCBI RefSeq and/or Integr8-Inquisitor. A list of central energy metabolic enzymes, which either do not have human homologues or functionally differentiate themselves from their human counterparts through alternative catalytic mechanisms, were considered as promising antibiotic drug targets. We proposed that the development of central energy metabolic inhibitors is a novel approach to avoid antibiotic resistance.**

Keywords

Methicillin/multidrug-Resistant Staphylococcus aureus (MRSA), essential genes, Database of Essential Genes (DEG), drug targets, central metabolism

I. INTRODUCTION

Methicillin/multidrug resistant *Staphylococcus aureus* (MRSA) infections are caused by antibiotic resistant strains of the common bacterium *Staphylococcus aureus* [1]. The beginning signs of MRSA infections are skin infections that resemble pimples, boils or spider bites. In immune-deficient patients, localized skin infections quickly spread through the bloodstream causing vital organ infections and possible death [2]. In a 2007 Centers for Disease Control and Prevention press release, there were about 94,000 cases of MRSA infections, contributing to around 19,000 deaths in the United States in 2005, which implies a mortality rate higher than that caused by HIV [3, 4].

The first MRSA case presented in the United Kingdom in 1961[5]. Shortly after, more variations were identified to be immune to β-lactam antibiotics (including penicillin, methicillin, oxacillin, and cephalosporins [6, 7]). Newly discovered MRSA strains have evolved to survive sulfa drugs, such as tetracyclines, and clindamycin [8]. Glycopeptide antibiotics, such as vancomycin and teicoplanin, considered drugs of "last resort", were used for the treatment of MRSA infections [9, 10]. However, recently discovered MRSA strains showed resistance even to vancomycin and teicoplanin [11, 12]. As of 2007, one variant found was resistant to six major kinds of antibiotics [13].

The current treatment for MRSA infections is still traditional broad-spectrum antibiotics such as lincosamides, sulfa drugs, glycopeptides [14-16], among which linezolids [17] daptomycin [18], Trimethoprim-sulfamethoxazole and MoxifloxacinHCl were considered relatively more effective [19, 20] though MRSA infections have become increasingly difficult to treat [15-17]. Thus, alternative treatments precisely targeting the root cause of MRSA infections needs to be established.

Novel antibiotic development focuses on the following: target screening *vs.* whole organism screening, microarray and/or proteomics [21]; target identification; rational and computer-assisted drug design [22, 23] and combinatorial chemistry *etc.*. The task falls on the shoulder of academia since the pharmaceutical industry has ceased investing in antibiotic discovery owing to high cost, lengthening developing cycles, complexities and low profits along with failure of several recent investments into target-based approaches [24]. In this paper, we report the initial results of anti-MRSA drug development, *i.e.,* a systematic *in silico* approach for the identification of drug targets in two MRSA strains, MRSA 252 and MRSA Mu50 based on the following two criteria: essentiality to pathogen survival and absence from the human genome [25, 26]. The novelty lays in that a special list of enzymes targeting bacterial metabolism was identified, shedding light on a potentially new approach for antibiotic development.

II. METHOD

The objective of this study was to determine potential drug targets for alternative treatment of MRSA infections, to explore hypothetically the functions of the identified targets

 ISBN: 978-1-61208-137-3

and to shorten the list. We employed a reported *in silico* approach through a systematic and juestified method [27, 28] for the identification of drug targets in two MRSA strains, MRSA 252 and MRSA Mu50. The proteomes of MRSA 252 and Mu50 were retrieved from NCBI gene bank [29]. MRSA genes were purged at 90 % and 60% using CD-HIT [30] to remove paralogues. The resulting sequences were run through the database of essential genes (DEG) [31, 32] at an expectation (E-value) cutoff of 10^{-4}. The database of essential genes includes genes required for basic survival of *Staphylococcus aureus* and other microorganisms according to experimental evidence. The essential genes were subjected to BLASTP against the human genome to exclude any genes that have a significant match (E-value cutoff of 10^{-3} and lower) with human homologs. Genes having BLAST E-scores less than 10^{-3} were considered as having no close relatives in human. Information about the putative gene function was derived from the annotated genome sequence through NCBI RefSeq and Integr8-Inquisitor [33].

III. RESULTS AND DISCUSSION

The goal of this investigation was to determine potential drug targets for alternative treatment of MRSA infections and to classify and to analyze the identified targets. Out of the complete genomes of 13 MRSA strains that were sequenced and deposited in the NCBI gene bank, MRSA 252 and MRSA Mu50 were selected due to the fact that the former is a common strain in USA [34] and UK [35] and the latter, a methicillin and vancomycin resistant strain isolated in Japan [36] is commercially available for future molecular biological study (ATCC). The common method of drug target identification encompasses two steps: the identification of essential genes for bacterial viability [25] and the identification of genes absent in the human genome [26]. The former was performed by adopting the DEG database in our approach because this database compiles a list of all currently available essential genes in more than 10 prokaryotes including *Staphylococcus aureus* [29] and was proved to be more accessible than conventional tools [27, 28]. On the other hand, the availability of the human genome sequence [37, 38] renders the latter step feasible. Following two newly published genomic analysis methods [27, 28], 2656 MRSA 250 and 2697 Mu50 genes were purged at 90 % and 60% using CD-HIT to remove paralogues, respectively. The resulting 2568 MRSA 250 and 2592 Mu50 sequences were run through the database of essential genes (DEG) at an expectation cut-off of 10^{-4}, yielding 499 and 496 essential genes respectively. Those 499 and 496 essential genes identified were subjected to BLASTP against the human genome [37, 38] to exclude any genes that have a significant match (*E*-value cutoff of 10^{-3} and lower) with human homologs. Consensually, 133 MRSA 252 and 134 Mu50 genes respectively were considered as having no close relatives in human. The results are summarized in table 1. Their known or hypothetic or putative functions annotated by NCBI RefSeq Integr8-Inquisitor are listed in table 2.

TABLE 1: GENOMICS ANALYSES OF MRSA 252 AND MRSA MU50 STRAINS.

Genes	MRSA 252	MRSA Mu50
Total number	2656	2697
Duplicates (>60% identical)	88	105
Non-paralogs	2568	2592
Essential genes [cut-off *E*-value < 10^{-4}]	499	496
Essential genes w/o human homologs[cut-off *E*-value < 10^{-3}]	133	134

Among the 133 and 134 essential non-human homologous genes in MRSA 252 and Mu50 strains, respectively, 133 encode proteins that are well conserved between the two strains. Out of this conserved set, 63 are involved in metabolism, 24 participate in the transmission of genetic information, 29 represent transmembrane proteins, 9 are with other functions such as regulation cell division and carrier proteins, *etc.,* and 8 have unknown functions.

Our approach identified 14 genes in cell wall biosynthesis. Other research groups have validated most of these targets [39-41]. Among them, 6 are involved in the elongation of peptidoglycan, in agreement with previous studies [39, 40]. FemA family proteins are currently considered novel anti-staphylococcal targets due to the fact that they are involved in cell wall biosynthesis and expression of a methicillin resistance gene [41]. They are found to be essential in both MRSA 252 (NCBI Gene Accession#: 49484627 and 49483567) and Mu50 (NCBI Gene Accession#: 15925401 and 15924364) strains by our approach. Gene GI#49484133 in MRSA 252 and GI#15924882 in Mu50 respectively represents *Staphylococcus aureus* murE gene encoding UDP-N-acetylmuramyl tripeptide synthetase, which was demonstrated to be essential in *Staphylococcus aureus* through a method incorporating an IPTG controllable promoter [42].

Although the cell wall has long been considered an attractive target for antibiotic development because of its absence in humans, what should not be overlooked is that one of the most common antibiotic resistance mechanisms is the metamorphosis of cell-wall proteins, resulting in inhibiting antibiotic activity. For example, β-lactam

 ISBN: 978-1-61208-137-3

TABLE 2. 133 ESSENTIAL NON-HUMAN HOMOLOGOUS GENES IN BOTH MRSA 252 AND MU50 ENCODING DIFFERENT CLASSES OF PROTEINS AND THEIR PUTATIVE OR HYPOTHETIC FUNCTIONS

Categories	Classes	General Functions	MRSA 252 *NCBI Gene Accession#*	MRSAMu50 *NCBI Gene Accession #*	Specific putative or hypothetic functions
Metabolism	Cellular respiration	Carbohydrate Catabolism	49482458	15923216	Formate acetyltransferase
			49482459	15923217	Formate acetyltransferase activating enzyme
			49482486	15923242	Xylitol dehydrogenase
			49483017	15923750	HPr kinase/phosphorylase
			49483247	15924074	Phosphoenolpyruvate-protein phosphatase ptsI
			49483033	15923765	Phosphoglyceromutase
			49483952	15924701	Acetate kinase
			49484267	15925031	Sucrose-6-phosphate hydrolase
			49484349	15925115	Fructose-bisphosphate aldolase
			49484367	15925133	Mannose-6-phosphate isomerase
			49484381	15925149	Mannitol-1-phosphate 5-dehydrogenase
			49484415	15925185	Galactose-6-phosphate isomerase subunit LacA
		Lipid Catabolism	49483384	15924216	Phosphatase/ dihydroxyacetone kinase
			49483425	15924288	Glycerol uptake operon antiterminator regulatory protein
		Amino acid catabolism	49482426	15923174	N-acetyl-γ-glutamyl-phosphate reductase
			49482779	15923539	N-acyl-L-amino acid amidohydrolase
			49483163	15923990	Thimet oligopeptidase homolog
			49483313	15924141	Glutamate racemase
			49483846	15924589	5'-methylthioadenosine nucleosidase/S-adenosylhomocysteine nucleosidase
			49484504	15925279	Urease subunit β
			49484120	15924869	Aminopeptidase ampS
			49484649	15925422	Glycerate kinase
			49484868	15925663	HisF cyclase-like protein
				15923177	Cystein Hydrolase
			49483520	15924318	Homoserine dehydrogenase
			49483584	15924384	Aspartate semialdehyde dehydrogenase
				15925319	Amino acid amidohydrolase
		Common metabolic pathway	49482818	15923578	Phosphotransacetylase
			49484161	15924909	Putative manganese-dependent inorganic pyrophosphatase
			49484002	57634637	Probable NAD(FAD)-utilizing dehydrogenase
	Bio-synthesis	Amino acid biosynthesis	49484873	15925668	Histidinol dehydrogenase
			49482425	15923173	Ornithine acetyltransferase
			49482586	15923346	5-methyltetrahydropter-oyltriglutamatehomo- cysteine methyltransferase
			49482696	15923462	Glutamate synthase, large subunit
			49483565	15924362	Tryptophan synthase β subunit
			49483583	15924383	Aspartokinase II
			49483655	15924456	Chorismate synthase
			49484279	15925043	dihydroxy acid dehydratase
			49484281	15925046	Ketol-acid reductoisomerase
			4948429	15925060	Alanine racease
			49484794	15925588	Pantoate--β-alanine ligase
		Fatty acid biosynthesis	49483392	15924219	Fatty acid/phospholipid synthesis protein
		Nucleotide biosynthesis	49482382	15923129	Phosphopentomutase
			49483421	15924248	Uridylate kinase
			49483664	15924468	Cytidylate kinase
		Cell wall biosynthesis	49484627	15925401	FemAB family protein
			49483567	15924364	FemA protein
			49482490	15923244	Teichoic acid biosynthesis protein (truncated TagF)
			49482939	15923673	Undecaprenyl Pyrophosphate Phosphatase
			49482995	15923728	UDP-N acetylenolpyruvoyl-glucosamine reductase
			49483182	15924008	UDP-N-acetylmuramoylalanyl-D-glutamate-2, 6-diaminopimelate ligase
			49484307	15925072	UDP-N-acetylmuramoylalanyl-D-glutamyl-2, 6-diaminopimelate-D-alanyl-D-alanyl ligase
			49484133	15924882	UDP-N-acetylmuramyl tripeptide synthetase

 ISBN: 978-1-61208-137-3

			49483346	15924173	UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase
			49484348	15925114	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
			49484309	15925074	Rod shape determining protein RodA
			49483587	15924387	Tetrahydrodipicolinate acetyltransferase
			49483980	15924730	UDP-N-acetyl-muramoyl-L-alanine synthetase
				57634647	UDP-N-acetylglucosamine 1-carboxyvinyltransferase
		Other biosynthesis	49482716	15923479	tetrapyrrole(corrin/porphy-rin) methylase
			49482722	15923485	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase
			49484013	15924759	Riboflavin biosynthesis
			49484795	15925589	3-methyl-2-oxobutanoate hydroxymethyltransferase
Transmission of genetic information	DNA replication, recombination and repair		49482254	15922991	Chromosomal replication initiation protein
			49482255	15922992	DNA polymerase III β subunit
			49482269	15923006	Replicative DNA helicase (DnaB-like)
			49483309	15924136	Excinuclease ABC subunit C
			49483633	15924434	Methyltransferase
			49483747	15924487	Integrase/recombinase
			49483811	15924552	DNA primase
			49483834	15924577	DNA polymerase III subunit delta
			49483926	15924674	Primosomal protein DnaI
			49483944	15924693	DNA polymerase III, β chain
			49484385	15925153	DisA bacterial checkpoint controller nucleotide binding
	Transcription and RNA processing		49483418	15924245	Transcriptional repressor CodY
			49483550	15924347	Transcription antiterminator
			49484097	15924845	SpoU rRNA methylase family protein
			49484908	15925703	Ribonuclease P
			49483433	15924260	Ribosome-binding factor A
			49483855	15924600	Transcription elongation factor
			49482590	15923350	Transcription terminator
			49483976	15924726	Catabolite control protein A
	Translation and posttranslational modifications		49483000	15923733	peptidase T
			49483039	15923772	SsrA-binding protein
			49483384	15924211	Hypothetical translation and posttranslational modifications
			49483609	15924409	Gcn5-related acetyltransferases
			49483778	15924518	Elongation factor P
Trans-membrane Proteins	Antibiotic Resistance		49482275	15923012	Metallo- lactamase
			49483344	15924171	Penicillin-binding protein
	Regulation		49483168	15923996	GTP pyrophosphokinase
			49483425	15924252	Zinc metalloprotease yluc
	Transport		49482431	15923179	Glucose-specific PTS, IIABC component
			49482476	15923232	PTS, IIBC component
			49482956	15923690	Gructose-specific PTS, IIABC component
			49483966	15924716	N-acetylglucosamine specific PTS, IIABC component
			49484378	15925146	Mannitol-specific PTS, IIBC component
			49484380	15925148	Mannitol specific PTS, IIA component
			49484538	15925313	PTS, arbutin-like, IIBC component
			49484739	15925528	Glucose-specific PTS, II ABC component
			49484838	15925631	PTS, IIABC component
			49483148	15923977	Oligopeptide transport system permease protein
			49484706	15925495	Gluconate permease
			49482866	15923628	Teichoic acid ABC transporter permease
			49484434	15925210	Cobalt transport protein
			49484516	15925291	Na^+/H^+ antiporter
			49484891	15925688	Nickel transport protein
			49484846	15925639	Bifunctional Preprotein translocase subunit SecA
			49483881	15924627	Bifunctional preprotein translocase subunit SecD/SecF
			49483265	15924092	Spermidine/putrescine-binding protein precursor homolog
			49482314	15923062	Potassium-transporting ATPase subunit A
			49482353	15923100	L-lactate permease homolog
			49484303	15925067	potassium-transporting ATPase subunit A
			49484446	15925220	Preprotein translocase subunit SecY
			49483071	15923829	ABC transporter substrate-binding protein
			49483075	15923833	ABC transporter-associated protein
			49483078	15923836	ABC transporter-associated protein

 ISBN: 978-1-61208-137-3

Other Proteins	Carrier proteins	49483175	15924003	Sodium/proton-dependent alanine carrier protein
		49482688	15923454	Lipoprotein
	Regulation	49482271	15923008	Response regulator protein
	Cell division	49482736	15923499	C ell division
		49483349	15924176	C ell division protein FtsZ
		49484905	15925700	Glucose-inhibited division protein B
	Other	49484374	15925142	Haloacid dehalogenase-like hydrolase
		49484612	15925386	Nitrate reductase β chain
		49484613	15925387	Respiratory nitrate reductase alpha chain
	Unknown function	49482472	15923228	Unknown
		49483005	15923738	Unknown
		49483022	15923755	Unknown
		49483024	15923757	Unknown
		49483035	15923767	Unknown
		49483546	15924343	Unknown
		49483928	15924676	Unknown
		49484792	15925584	Unknown

resistance was attributed to the expression of a group of cell wall penicillin-binding proteins (PBP-2') encoded by the mecA gene [43, 44]. Glycopeptide resistance is also considered to be caused by cell wall thickening resulting in binding vancomysin extracellularly [45,46] and/or alteration of the drug-acting site in the cell wall from D-alanine-D-alanine to D-alanine-D-lactate owing to the expression of *vanA* resistance gene [47]. Hence, for novel antibiotic development, substances that anchor in sites other than the bacterial cell wall may have more potential because resistance usually arises as the result of gene mutation on the target proteins that are subject to direct antibiotic attack [48]. A 2006 review on mechanisms of bacterial antibiotic resistance suggested the exploration of novel antibiotics with alternative mechanisms of action [49].

Genes involved in transmission of genetic information including DNA replication, recombination and repair, transcription and RNA processing, translation, post-translational modification remain viable targets for antibacterial agent development [33]. Our approach identified 24 of these candidate genes.

Our approach identified 29 membrane bound proteins. A recent review on anti-MRSA drug development indicated that agents anchoring in the bacterial membrane (*e.g.,* ceragenins and lipopeptides) showed great bactericidal effect and may be less prone to drug resistance due to the incapability of bacteria to modify their targets in a way that is compatible with their survival [50]. Among this pool of proteins, 19 are involved in membrane transport, which represent valid drug targets because pathogens usually lose their biosynthetic capabilities and rely on their hosts for the supply of essential nutrients [51, 52].

Our approach identified 30 energy metabolic (*i.e.* cellular respiration) genes in both MRSA 252 and MRSA Mu50, which are essential to staphylococcal survival with *E*-score < 10^{-4} but absent in human genome with *E*-score <10^{-3}. Currently there are limited numbers of commercially available antibiotics targeting energy metabolism. Those existing are mainly biological reagents such as oligomycin [53] and pesticides or piscicides such as antimycin A [54], not commonly used for humans in that they affect both bacterial and human cells. Surprisingly, nature has provided us with a group of energy metabolic enzymes which are essential to pathogen survival while absent in humans. The differentiation lies in that those enzymes function through alternative mechanisms other than their counterpart enzymes in humans. For example, fructose-1, 6-diphosphate aldolase (FBPA) is one of the key enzymes in the glycolytic pathway that involves the breakdown of glucose [55]. FBPA is divided into two classes based on structural properties and catalytic mechanisms [56]. Class I FBPA is mainly found in higher order organisms (*e.g.,* human and animals). Catalysis in Class I FBPA proceeds via a Schiff base intermediate formed by an active site lysine residue [55]. Class II FBPA is usually found in yeasts, bacteria, fungi, and parasites [56]. Catalysis in Class II FBPA centers on the participation of a Zn (II) cofactor that coordinates to an enolate anion intermediate [54]. Multiple alignment of the sequence of *MRSA* FBPA with class II *giardia* FBPA and class I *human* FBPA was shown in Fig. 1. *MRSA* FBPA (NCBI Gene Accession[#]: 49484349 and 15925115 respectively) exhibits 40.8% sequence identity to Class II *giardia* FBPA while it exhibits only 18.8 % sequence identity to class I *human* FBPA [57,58]. Thus, *MRSA* FBPA should be hypothetically classified into class II FBPA, not class I FBPA. Validation of the essential nature of class II *MRSA* FBPA through allelic replacement and inducible expression is underway in our research group. Based on major differences in active site structure and catalytic mechanism, an inhibitor of class II FBPA can be designed which will not inhibit class I FBPA.

Accumulating *in vitro* [59] and *in vivo* [60] evidence suggests that enzymes catalyzing bacterial cellular respiration with differentiated mechanisms of action are promising targets for novel antibiotic development. The

 ISBN: 978-1-61208-137-3

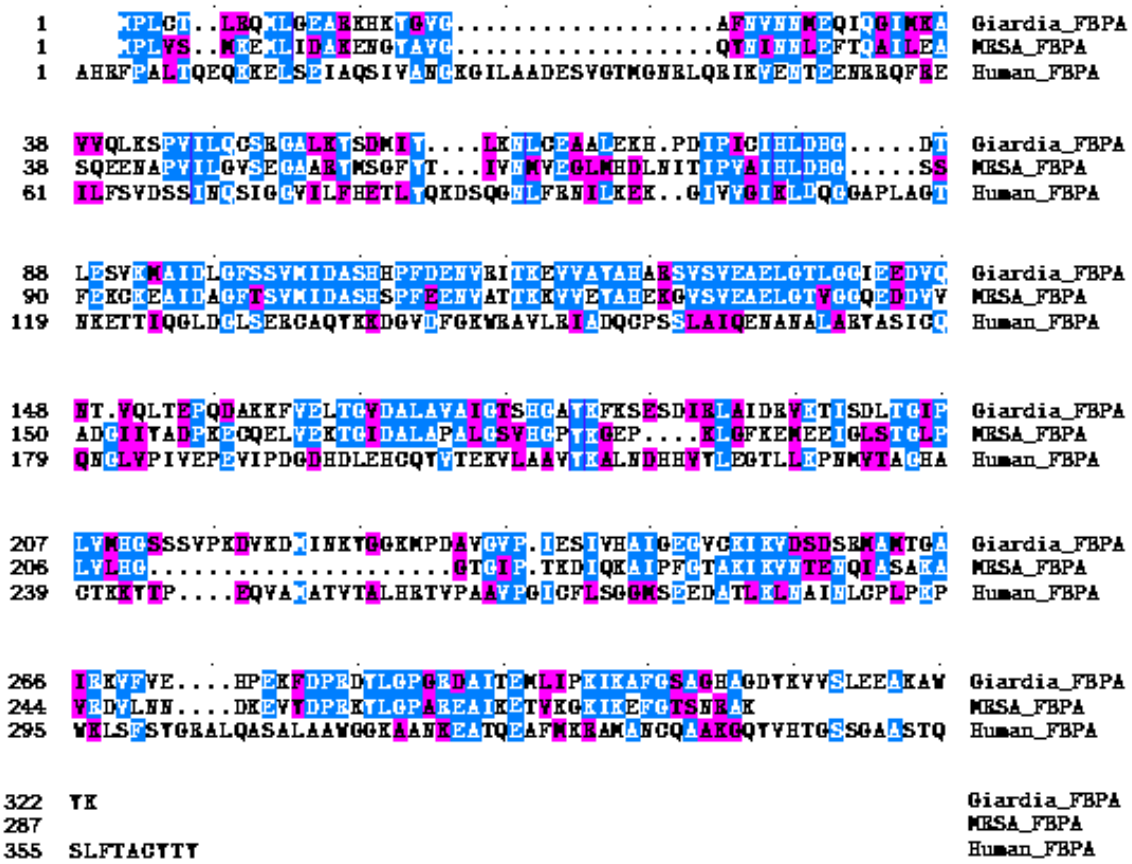


Figure.1 Alignment of the amino acid sequences of *MRSA* FBPA (NCBI GENE ACCESSION#:49484349 and 15925115 respectively) with class II *giardia* FBPA (2ISV) and class I *human* FBPA(1QO5). Numbering of the amino acids is indicated on the left. Identical amino acid residues in the alignment are indicated in light-blue shading and similar amino acid residues are indicated in purple shading. Gaps introduced during the alignment process are indicated as dotes.

inhibitors designed are able to hinder bacterial growth by inhibition of those enzymes without interfering with their human cousins. Most importantly, attacking bacterial energy-making machinery bypasses the usual bacterial mutation sites for drug resistance [61-62]. The rationale lies in that almost all existing antibiotics target only 4 cellular functions: cell wall synthesis, protein synthesis, nucleic acid synthesis and foliate synthesis, though there are hundreds of antibiotics on the market [63]. Repeated exposure of bacteria to antibacterial reagents targeting similar sites increases the chance of bacterial gene mutation, which remains to be the primary cause of the prevalence of antibiotic-resistant bacteria, such as MRSA, NDM-1 induced antibiotic -resistant *Escherichia coli* [62], and *etc*.. Exploration of antibiotics targeting alternative cellular functions such as central metabolic pathways may be a promising direction, and selective inhibition of targets specific to bacterial energy metabolism may be a potentially efficacious alternative in the treatment of MRSA infections. The enzymes on the higher priority list include *MRSA* FBPA, *MRSA* dihydroxyacetone kinase (DAK) 2 Phosphatase, *MRSA* acetate kinase, *MRSA* histidinol dehydrogenase, *MRSA* Phosphotransacetylase, *MRSA* Sucrose-6-phosphate hydrolase and *MRSA* glycerate kinase, which either do not have human homologues or adopt dramatically different catalytic mechanisms comparing to their human cousins.

 ISBN: 978-1-61208-137-3

CONCLUSION AND FUTURE WORK

One of the crucial steps in narrow-spectrum antibiotics development is target identification. In this study, a putative set of candidate drug targets were elucidated by an in silico approach. The candidate genes are hypothetically required for survival of the candidate microorganism and have no close human analogue. Many identified targets have been experimentally validated [41-44, 65-68]. By shortening the list of potential drug targets to a small pool of genes, the data presented in this paper facilitated our group and, may also aid other researchers in pursuing target validation and target characterization for alternative treatment of MRSA infections. Future directions include developing inhibitors. for the candidate proteins. In principle, the premise is that the inhibitors of these targets should only be toxic to pathogens but safe for use by humans.

More importantly, we propose that a class of central metabolic enzymes, such as MRSA FBPA, MRSA dihydroxyacetone kinase (DAK) 2 Phosphatase, MRSA acetate kinase, MRSA histidinol dehydrogenase, MRSA Phosphotransacetylase, MRSA Sucrose-6-phosphate hydrolase and MRSA glycerate kinase (table 2), are promising antibiotic drug targets due to the fact that they either do not have their human counterparts or if they do, different catalytic mechanisms are employed (*e.g.,* class I and class II FBPA). Based on major differences in active site structure and catalytic mechanism, an inhibitor of such a bacterial enzyme can be designed which will not inhibit its human cousin. Also, the risk of bacterial drug resistance against inhibitors of those enzymes may be low because antibiotics targeting bacterial central metabolism are not commonly used. Those cellular sites are not repeatedly exposed to antibacterial agents thus less prone to drug resistance. Proposed long-term work includes utilizing MRSA as a model bacterial system to develop methods combating antibiotic resistance. It is even more crucial that this type of investigation is undertaken in academia than it would be if industry were still heavily investing in it [24, 63].

ACKNOWLEDGMENT

We thank Dr Adhar Manna (University of South Dakota) for the ongoing collaboration on target essentiality validation. This publication was made possible by NIH Grant Number 2 P20 RR016479 from the INBRE Program of the National Center for Research Resources. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

REFERENCES

[1] M. E. Mulligan, K. A. Murray-Leisure, B. S. Ribner, H. C. Standiford, J. F. John, J. A. Korvick, C. A. Kauffman and V. L. Yu, "Methicillin-resistant *Staphylococcus aureus*: a consensus review of the microbiology, pathogenesis, and epidemiology with implications for prevention and management," Am. J. Med. vol.94, Mar. 1993, pp. 313-328.

[2] J. M. Voyich, M. Otto, B. Mathema, K. R. Braughton, A. R. Whitney, D. Welty, R. D. Long, D. W. Dorward, D. J. Gardner, G. Lina, B. N. Kreiswirth and F. R. DeLeo, "Is panton-valentine leukocidin the major virulence determinant in community-associated methicillin-resistant *Staphylococcus aureus* disease?" J. Infect. Dis. vol.194, Dec. 2006, pp. 1761-1770.

[3] Centers for Disease Control and Prevention, (2007). MRSA: Methicillin-resistant *Staphylococcus aureus* in Healthcare Settings.

[4] R. M. Klevens, M. A. Morrison, J. Nadle, S. Petit, K. Gershman, S. Ray, L. H. Harrison, R. Lynfield, G. Dumyati, J. M. Townes, A. S. Craig, E. R. Zell, G. E. Fosheim, L. K. McDougal, R. B. Carey and S. K. Fridkin, "Invasive methicillin-resistant *Staphylococcus aureus* infections in the United States," JAMA, vol. 298, Oct. 2007, pp. 1763-1771.

[5] M. Jevons, Celbenin"-resistant staphylococci, Br. Med. J., vol. 1, 1961, pp.124-125.

[6] T. Foster, (1996). *Staphylococcus*. In: Barron's Medical Microbiology (Barron S *et al*, eds.), 4th ed. Galveston, TX.

[7] K. Okuma, K. Iwakawa, J. D. Turnidge, W. B. Grubb, J. M. Bell, F. G. O'Brien, G. W. Coombs, J. W. Pearman, F. C. Tenover, M. Kapi, C. Tiensasitorn, T. Ito and K. Hiramatsu, "Dissemination of new methicillin-resistant *Staphylococcus aureus* clones in the community", J. Clin. Microbiol. vol. 40, Nov. 2002, pp. 4289–4294, doi: 10.1128/JCM.40.11.4289-4294.2002

[8] H. Huang, N. M. Flynn, J. H. King, C. Monchaud, M. Morita and S. H. Cohen, "Comparisons of Community-Associated Methicillin-Resistant *Staphylococcus aureus* (MRSA) and Hospital-Associated MSRA Infections in Sacramento, California," J. Clin. Microbiol. vol. 44, July 2006, pp. 2423-2427, doi:10.1128/JCM.00254-06

[9] R. C. Moellering, "Vancomycin: A 50-Year Reassessment," Clin. Infect. Dis. vol. 42, Suppl 1, Jan. 2006, pp. S3-4.

[10] P. L. Donald, "Vancomycin: A History," Clin. Infect. Dis. vol., 42, Suppl 1, Jan. 2006, pp. S5-S12.

[11] G. C. Schito, "The importance of the development of antibiotic resistance in *Staphylococcus aureus*," Clin. Microbiol. Infect., Suppl 1, Mar. 2006, pp. 16445718.

[12] K. Sieradzki and A. Tomasz, "Inhibition of cell wall turnover and autolysis by vancomycin in a highly vancomycin-resistant mutant of *Staphylococcus aureus*," J. Bacteriol., vol. 179, April, 1997, pp. 2557-2566.

[13] C. Burlak, C. H. Hammer, M. Robinson, A. R. Whitney, M. J. McGavin, B. N. Kreiswirth and F. R. DeLeo, "Global analysis of community-associated methicillin-resistant *Staphylococcus aureus* exoproteins reveals molecules produced *in vitro* and during infection Cell" Microbiol., vol. 9, Jan. 2007, pp.1172–1190, doi:10.1111/j.1462-5822.2006.00858.x

[14] J. D. Siegel, E. Rhinehart, M. Jackson and L. Chiarello, (2008). Management of multi-drug resistant organisms in healthcare settings, 2006. US Centers for Disease Control and Prevention. Healthcare Infection Control Practices Advisory Committee. Accessed January 25.

[15] L. Nicolle, "Community-acquired MRSA: a practitioner's guide," CMAJ., vol. 175, June 2006, pp. 145, doi:10.1503/cmaj.060457

 ISBN: 978-1-61208-137-3

[16] Centers for Disease Control and Prevention. Epidemiology and management of MRSA in the Community. October 26, 2007. Accessed January 25, 2008.
[17] J. A. Gorchynski and J. K. Rose, "Complications of MRSA Treatment: Linezolid-induced Myelosuppression Presenting with Pancytopenia," Western Journal of Emergency Medicine, vol. 9, August 2008, pp. 177-178
[18] Cubist Pharmaceuticals, Inc. (2003). Daptomycin (Cubicin) package literature. Cubist Pharmaceuticals, Inc., Lexington, Mass.
[19] R. Moellering, Trends in New Drug Development; from Broad- to Narrow-Spectrum Antibiotics Program and abstracts from the 39th ICAAC Symposium 138, F 1363 .
[20] F. Tally, Trends in New Drug Development; from Broad- to Narrow-Spectrum Antibiotics Program and abstracts from the 39th ICAAC Symposium 138, F 1364 .
[21] Y. Zhang, "The Magic Bullets and Tuberculosis Drug Targets. Ann Rev Pharmacol Toxicol, " vol. 45, 2005, pp. 529-564, doi: 10.1146/annurev.pharmtox.45.120403.100120
[22] J. Smith and V. Stein, "SPORCalc: A development of a database analysis that provides putative metabolic enzyme 2008. reactions for ligand-based drug design," Comput. Biol. Chem., vol. 33, Apr.2009, pp.149–159, doi:10.1016/j.compbiolchem.2008.11.002
[23] W. F. de Azevedo and R. Dias, "Computational methods for calculation of ligand-binding affinity". Curr Drug Targets vol.9, Dec. 2008, pp. 1031–1039.
[24] S. Projan, "Why is big Pharma getting out of antibacterial drug discovery?" Curr. Opin. Microbiol. vol. 6, Oct. 2003, pp. 427–430, doi:10.1016/j.mib.2003.08.003
[25] P. F. Boreham, R. E. Phillips and R. W. Shepherd, "Altered uptake of metronidazole in vitro by stocks of Giardia intestinalis with different drug sensitivities," Trans. R. Soc. Trop. Med. Hyg. vol. 82, May 1988, pp. 104-106.
[26] S. D. Mills, "The role of genomics in antimicrobial discovery," J. Antimicrob. Chemother., vol. 51, Mar. 2003, pp.749-752, doi: 10.1093/jac/dkg178
[27] K. R. Sakharkar, M. K. Sakharkar and V. T. Chow, "A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*," *In Silico* Biology, vol. 4, 2004, pp.355-360.
[28] V. Sharma, P. Gupta, and A. Dixit, "Identification of putative drug targets from different metabolic pathways of *Aeromonas hydrophila*," *In Silico.* Biology," vol 4, 2008, pp. 331-338.
[29] http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi(03/15/2011)
[30] W. Li, L. Jaroszewski and A. Godzik, "Clustering of highly homologous sequences to reduce the size of large protein databases," Bioinformatics, vol. 17, Mar. 2001, pp. 282-283, doi: 10.1093/bioinformatics/17.3.282
[31] R. Zhang, H. Ou and C. Zhang, "DEG, a Database of Essential Genes," Nucleic Acids Res. vol. 32, (suppl 1), Jan. 2004, pp. D271-D272, doi: 10.1093/nar/gkh024
[32] R. Zhang and Y. Lin, "DEG 5.0, a database of essential genes in both prokaryotes and eukaryotes," Nucleic Acids Res., vol. 37, (suppl 1): Jan. 2009, pp. D455-D458, doi: 10.1093/nar/gkn858
[33]http://www.ebi.ac.uk/integr8/InquisitorPage.do (03/15/2011)
[34] B. A. Diep, H. A. Carleton, R. F. Chang, G. F. Sensabaugh and F. Perdreau-Remington, "Roles of 34 virulence genes in the evolution of hospital- and community-associated strains of methicillin-resistant *Staphylococcus aureus*," J. Infect. Dis. vol. 193, Apr.2006, pp.1495–1503, doi: 10.1086/503777
[35] A. P. Johnsona, H.M. Auckenb, S. Cavendishc, M. Gannerb, M. C. J. Walec, M. Warnera, D. M. Livermorea and B. D. Cooksonb "Dominance of EMRSA-15 and -16 among MRSA causing nosocomial bacteraemia in the UK: analysis of isolates from the European Antimicrobial Resistance Surveillance System" (EARSS)". J. Antimicrob. Chemother., vol. 4, 2001, pp. 143–144, doi:10.1093/jac/48.1.143.
[36] K. Hiramatsu, N. Aritaka and H. Hanaki, "Dissemination in Japanese hospitals of strains of *Staphylococcus aureus* heterogeneously resistant to vancomycin," Lancet, vol. 350, Dec. 1997, pp. 1670-1673, doi:10.1016/S0140-6736(97)07324-8
[37] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, *et al.*, "Initial sequencing and analysis of the human genome," Nature vol. 409, Feb. 2001, pp. 860-921, doi:10.1038/35057062
[38] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, *et al.*, "The sequence of the human genome" Science, vol. 291, Feb. 2001, pp.1304-1351,doi: 10.1126/science.1058040.
[39] K. L. Longenecker, G. F. Stamper, P. J. Hajduk, E. H. Fry, C. G. Jakob, J. E. Harlan, R. Edalji, D. M. Bartley, K. A. Walter, L. R. Solomon, T. F. Holzman, Y. G. Gu, C. G. Lerner, B. A. Beutel and V. S. Stoll, "Structure of MurF from *Streptococcus pneumoniae* co-crystallized with a small molecule inhibitor exhibits interdomain closure," Protein Sci., vol. 14, Dec. 2005, pp3039-3047, doi: 10.1110/ps.051604805
[40] A. Perdih, M. Kotnik, M. Hodoscek and T. Solmajer, "Targeted molecular dynamics simulation studies of binding and conformational changes in *E. coli* MurD," Proteins, vol.68, Apr. 2007,pp. 243-254. doi: 10.1002/prot.21374
[41] A. Strandén, K. Ehlert, H. Labischinski and B. Berger-Bächi, "Cell wall monoglycine cross-bridges and methicillin hypersusceptibility in a femAB null mutant of methicillin-resistant *Staphylococcus aureus*," J. Bacteriol. vol. 179, Jan. 1997, pp. 9–16.
[42] J. Malabendu, T. Luong, H. Komatsuzawa, M. Shigeta and C. Y. Lee, "A method for demonstrating gene essentiality in *Staphylococcus aureus*," Plasmid, vol. 44, Mar. 2000. pp. 100-104, doi:10.1006/plas.2000.1473
[43] N. H. Georgopapadakou and F. Y. Liu, "Binding of β-lactam antibiotics to penicillin-binding proteins of *Staphylococcus aureus* and *Streptococcus faecalis*: relation to antibacterial activity," Antimicrob. Agents Chemother. vol. 18, Nov. 1980, pp.834-836.
[44] K. Ubukata, N. Yamashita and M. d Konno, "Occurrence of a β-lactam-inducible penicillin-binding protein in methicillin resistant *staphylococci*," Antimicrob. Agents Chemother. vol. 27, May 1985, pp. 851-857.
[45] M. Kuroda, H. Kuroda, T. Oshima, F. Takeuchi, H. Mori and K. Hiramatsu, "Two-component system VraSR positively modulates the regulation of cell-wall biosynthesis pathway in *Staphylococcus aureus*," Mol. Microbiol. vol.49, Aug. 2003, pp807–821, doi:10.1046/j.1365-2958.2003.03599.x
[46] L. Cui, H. Murakami, K. Kuwahara-Arai, H. Hanaki and K. Hiramatsu, "Contribution of a thickened cell wall and its glutamine nonamidated component to the vancomycin resistance expressed by *Staphylococcus aureus* Mu50," Antimicrob. Agents Chemother. vol. 44, Sep. 2000, pp. 2276–2285.
[47] A. Severin, K. Tabei, F. Tenover, M. Chung, N. Clarke and A. Tomasz, "High level oxacillin and vancomycin resistance and altered cell wall composition in *Staphylococcus aureus* carrying the staphylococcal mecA and the enterococcal vanA gene complex," J. Biol. Chem. vol. 279, Jan. 2004, pp. 3398-3407.
[48] A. A. Salyers, and D. D. Whitt, (2005). Revenge Of The Microbes: How Bacterial Resistance Is Undermining The Antibiotic Miracle, American Society for Microbiology Press, Washington, DC.
[49] F. C. Tenover, "Mechanisms of antimicrobial resistance in bacteria," Am J Med., vol. 119 (6 Suppl 1), June 2006, pp. S3–S10, doi:10.1016/j.ajic.2006.05.219

 ISBN: 978-1-61208-137-3

[50] F. Van Bambeke, M. P. Mingeot-Leclercq, M. J. Struelens and P. M. Tulkens, "The bacterial envelope as a target for novel anti-MRSA antibiotics," Trends Pharmacol Sci. vol. 29, Mar. 2008, pp. 124-134.

[51] K. Lewis, "Multidrug resistance: versatile drug sensors of bacterial cells," Curr. Biol., vol. 9, Jun. 1999, pp. R403-R407, doi:10.1016/S0960-9822(99)80254-1

[52] C. M. Fraser, J. D. Gocayne, O. White, M. D. Adams, R. A. Clayton, R. D. Fleischmann, C. J. Bult, A. R. Kerlavage, G. Sutton, J. M. Kelley, R. D. Fritchman, J. F. Weidman, K. V. Small, M. Sandusky, J. Fuhrmann, D. Nguyen, T. R. Utterback, D. M. Saudek, C. A. Phillips, J. M. Merrick, J. F. Tomb, B. A. Dougherty, K. F. Bott, P. C. Hu, T. S. Lucier, S. N. Peterson, H. O. Smith, C. A. Hutchison 3rd and J. C. Venter, "The minimal gene complement of *Mycoplasma genitalium*," Science. vol. 270, Oct. 1995, pp. 397-403, DOI: 10.1126/science.270.5235.397

[53] Y. Kagawa and E. Racker, "Partial resolution of the enzymes catalyzing oxidative phosphorylation. 8. Properties of a factor conferring oligomycin sensitivity on mitochondrial adenosine triphosphatase," J. Biol. Chem., vol.241, May 1966, pp. 2461-2466.

[54] D. Xia, C. H. Yu, H. Kim, J. Xia, A. M. Kachurin, L. Zhang, L. Yu and J. Deisenhofer, "Structure of Antimycin A1, a Specific Electron Transfer Inhibitor of Ubiquinol-Cytochrome c Oxidoreductase" J. Am. Chem. Soc., vol. 121, Aug. 1999, pp.4902–4903, DOI: 10.1002/chin.199933269

[55] T. Gefflaut, C. Blonski, J. Perie and M. Willson, "Class I aldolases: substrate specificity, mechanism, inhibitors and structural aspects," Prog. Biophys. Mol. Biol. vol.63, 1995, pp. 301-340.

[56] A. Galkin, L. Kulakova, E. Melamud, L. Li, C. Wu, P. Mariano, D. Dunaway-Mariano, T. E. Nash and O. Herzberg, "Characterization, Kinetics, and Crystal Structures of Fructose-1,6-bisphosphate Aldolase from the Human Parasite *Giardia lamblia,"* J. Biol. Chem., vol. 282, Feb. 2007, pp. 4859–4867.

[57] S. F. Altschul, T. L. Madden, A. A. Schäffe, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," Nucleic Acids Res. vol. 25, Sep. 1997, pp. 3389-3402.

[58] J.D. Thompson, D.G. Higgins and T.J. Gibson, "CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," Nucleic Acids Res. vol. 22, Nov. 1994, pp. 4673-4680, doi: 10.1093/nar/22.22.4673.

[59] A. Fredenhagen, S. Y. Tamura, P. T. M. Kenny, H. Komura, Y. Naya, K. Nakanishi, K. Nishiyama, M. Sugiura and H. Kita, "Andrimid, a new peptide antibiotic produced by an intracellular bacterial symbiont isolated from a brown planthopper" J. Am. Chem. Soc. vol. 109, Jul. 1987, pp. 4409–4411, doi: 10.1021/ja00248a055

[60] C. Freiberg, J. Pohlmann, P. G. Nell, R. Endermann, J. Schuhmacher, B. Newton, M. Otteneder, T. Lampe, D. Häbich and K. Ziegelbauer, "Novel bacterial acetyl coenzyme A carboxylase inhibitors with antibiotic efficacy *in vivo,"* Antimicrob Agents Chemother, vol. 50, Aug. 2006, pp. 2707-2712.

[61] F. R. Stermitz, P. Lorenz, J. N. Tawara, L. A. Zenewicz and K. Lewis, "Synergy in a medicinal plant: Antimicrobial action of berberine potentiated by 5'-methoxyhydnocarpin, a multidrug pump inhibitor," *Proc. Natl. Acad. Sci. U.S. A.* vol. 97, Feb. 2000, pp. 1433-1437.

[62] N. R. Guz and F. R. Stermitz, "Synthesis and structures of regioisomeric hydnocarpin-type flavonolignans," *J. Nat. Prod.* Vol. 63, Aug. 2000, pp. 1140-1145, DOI: 10.1021/np000166d.

[63] C. Nathan, "Antibiotics at the crossroads," Nature, vol. 431, Oct. 2004, pp. 899–902, doi :10.1038/431899a

[64] K. K. Kumarasamy, M. A. Toleman, T. R. Walsh, J. Bagaria, F. Butt, R. Balakrishnan, and U. Chaudhary, *et al.* "Emergence of a new antibiotic resistance mechanism in India, Pakistan, and the UK: a molecular, biological, and epidemiological study". Lancet Infect Dis vol.10, August 2010, pp. 597–602, doi:10.1016/S1473-3099(10)70143-2.

[65] A. Boniface, A. Bouhss, D. Mengin-Lecreulx and D. Blanot, "The MurE synthetase from Thermotoga maritima is endowed with an unusual D-lysine adding activity," J. Biol. Chem. vol. 281, Jun. 2006, pp. 15680-15686.

[66] T. Deva, E. N. Baker, C. J. Squire and C. A. Smith, "Structure of *Escherichia coli* UDP-N-acetylmuramoyl:L-alanine ligase (MurC)," Acta. Crystallogr. D. Biol. Crystallogr. vol.62, Dec. 2006, pp. 1466-1474, doi:10.1107/S0907444906038376

[67] C. A. Smith, "Structure, function and dynamics in the mur family of bacterial cell wall ligases," J. Mol. Biol. vol. 362, Sep.2006, pp. 640-655, doi:10.1016/j.jmb.2006.07.066.

[68] K. Ehlert, "Methicillin-resistance in *Staphylococcus aureus* - molecular basis, novel targets and antibiotic therapy," Curr. Pharm. Des. vol.5, Feb.1999, pp. 45-55.

 ISBN: 978-1-61208-137-3

On the Distribution of the Distances Between Pairs of Leaves in Phylogenetic Trees

Arnau Mir
Department of Math and Computer Science
University of Balearic Islands
Palma de Mallorca, Spain
arnau.mir@uib.es

Francesc Rosselló
Department of Math and Computer Science
University of Balearic Islands
Palma de Mallorca, Spain
cesc.rossello@uib.es

***Abstract*—The distance, or path length, between two nodes in a phylogenetic tree (rooted or unrooted) is defined as the length of the unique undirected path connecting these nodes. In this paper we study the distribution of the distances between pairs of leaves in fully resolved phylogenetic trees with a fixed number of leaves. More precisely, we prove both in the unrooted and the rooted cases that, when the trees are equiprobably chosen, this distribution approximates a gamma distribution.**

***Keywords*-phylogenetic trees; statistical distribution;**

I. INTRODUCTION

Over the last years there has been an increasing interest in the study of the statistical behaviour of topological features in phylogenetic trees under different evolution models [2], [3], [9], [10]. The motivations for such studies are the assessment of the validity of an evolutionary model for a given set of phylogenetic trees, and the objective evaluation of how atypical a given phylogenetic tree is.

One feature whose behaviour has been studied is the topological distance, or path length, between pairs of leaves. Steel and Penny [11] computed the mean value and the variance of this distance d between two leaves in a fully resolved unrooted phylogenetic tree with n leaves. The statistical analysis of this random variable was continued in [6], where Steel and Penny's results were generalized to rooted phylogenetic trees, and in [7], [8], where the median and the mode of d were computed, both in the rooted and the unrooted cases. Let us mention that the study of the distance between pairs of leaves has a further motivation, as it may be used in the study of the statistical properties of the nodal distance between phylogenetic trees, an interesting and mostly open problem in phylogenetics [6], [11].

In this paper, instead of focusing on the exact computation of the statistical measures for d, we focus on its distribution, and we prove that, in the fully resolved case, it is approximately a gamma distribution, in the sense that the mean quadratic error between the distribution of d and a gamma distribution of the same mean and mode has limit 0 as $n \to \infty$.

The rest of this paper is organized as follows. In Section II, we prove our main result for unrooted fully resolved trees. Then, in Section III we briefly describe how this result translates to the rooted case, and in Section IV we report on some experimental results showing the fast convergence between d and the corresponding gamma distribution. The paper ends with a Conclusions section.

II. THE UNROOTED CASE

Throughout this paper, by a *phylogenetic tree* on a set S we mean a *fully resolved* (that is, with all its internal nodes of degree 3) unrooted tree with its leaves bijectively labelled in the set S. Although in practice S may be any set of taxa, to fix ideas we shall always take $S = \{1, \ldots, n\}$, where n is the number of tree leaves. For simplicity, we shall always identify a leaf of a phylogenetic tree with its label.

Let $\mathscr{T}_n^u$ be the set of (isomorphism classes of) phylogenetic trees with n leaves. It is well known [4] that $|\mathscr{T}_1^u| = |\mathscr{T}_2^u| = 1$ and $|\mathscr{T}_n^u| = (2n-5)!! = (2n-5)(2n-7)\cdots 3 \cdot 1$, for every $n \geqslant 3$.

Let $k, l \in S = \{1, \ldots, n\}$ be any two different labels of trees in $\mathscr{T}_n^u$. The *distance*, or *path length*, $d_T^u(k,l)$ between the leaves k and l in a phylogenetic tree $T \in \mathscr{T}_n^u$ is the length of the unique path between them. Let's consider the random variable

$$d_{kl}^u = \text{distance between } k \text{ and } l \text{ in one tree in } \mathscr{T}_n^u.$$

The possible values of d_{kl}^u are $\Omega^u = \{1, 2, \ldots, n-1\}$.

Our goal is to approximate the distribution of the variable d_{kl}^u on $\mathscr{T}_n^u$ when the tree and their leaves are chosen equiprobably. In this case, $d_{kl}^u = d_{12}^u$, and thus we can reduce our problem to study the distribution of the variable $d_n^u := d_{12}^u$.

For every $i \in \Omega^u$, let

$$c_{i,n}^u = \frac{|\{T \in \mathscr{T}_n^u \mid d_T^u(1,2) = i\}|}{(2n-5)!!}$$

denote the fraction of trees in $\mathscr{T}_n^u$ where the leaves 1 and 2 are at distance i. The sequence $(c_{i,n}^u)_{i=1,\ldots,n-1}$ is the distribution of the variable d_n^u. From [11, p. 140] and [8], we have the following result.

Lemma 1: (a) $c_{n-1}^u = \frac{(n-2)!}{(2n-5)!!}$ and, for every $i = 1, \ldots, n-2$,

$$c_{i,n}^u = \frac{(i-1)(2n-i-4)!}{(2(n-i-1))!! \cdot (2n-5)!!}.$$

 ISBN: 978-1-61208-137-3

(b) The mean of d_n^u is

$$\mu_n = \sum_{i=2}^{n-1} i c_{i,n}^u = \frac{2^{n-2}(n-2)!}{(2n-5)!!}$$

(c) The mode of d_n^u is

$$m_n = \left\lceil \frac{1+\sqrt{8n-15}}{2} \right\rceil.$$

Let $\gamma(k,\theta)$ denote a gamma distribution with parameters k (shape) and θ (scale), and let $f_{\gamma(k,\theta)}$ be its density function. Recall that the mean of $\gamma(k,\theta)$ is $k\cdot\theta$ and its mode is $(k-1)\cdot\theta$. Our goal in this section is to prove the following result:

Theorem 1: Consider the gamma distribution $\gamma(k_n,\theta_n)$, with parameters k_n and θ_n given by

$$k_n \cdot \theta_n = \mu_n, \quad (k_n-1)\cdot\theta_n = m_n,$$

where

$$\begin{aligned} k_n &= \frac{2^{n-2}\cdot(n-2)!}{2^{n-2}\cdot(n-2)! - (2n-5)!!\lceil(\sqrt{8n-15}+1)/2\rceil}, \\ \theta_n &= \frac{2^{n-2}(n-2)!}{(2n-5)!!} - \left\lceil \frac{1+\sqrt{8n-15}}{2} \right\rceil. \end{aligned}$$

In other words, $\gamma(k_n,\theta_n)$ is the gamma distribution with the same mean and mode as d_n^u on $\mathscr{T}_n^u$. Let MQE_n^u be the mean quadratic error between the random variable d_n^u and this gamma distribution:

$$MQE_n^u = \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_n,\theta_n)}(i))^2.$$

Then, $\lim_{n\to\infty} MQE_n^u = 0$.

Proof: Let $g_n(x)$ be the following function:

$$g_n(x) = \frac{(x-1)\cdot 2^{x-1}\cdot\Gamma(2n-x-3)\cdot\Gamma(n-1)}{\Gamma(n-x)\cdot\Gamma(2n-3)}.$$

This function satisfies that $g_n(i) = c_{i,n}^u$ for every $i = 1,\ldots,n-1$, and therefore it can be seen as the extension to $\mathbb{R}^+$ of the discrete distribution of d_n^u.

The sequence $(g_n(i))_{n=1,\ldots,n-1}$ reaches its maximum at m_n of d_n^u. We want to approximate $g_n(m_n)$. To do that, we shall use the following expansion of the logarithm of the Gamma function:

$$\ln\Gamma(x) \approx \frac{\ln(2\pi)}{2} + \left(x-\frac{1}{2}\right)\ln\left(x-\frac{1}{2}\right) - \left(x-\frac{1}{2}\right), \quad (1)$$

for large values of x. Using this expansion and using that

$$m_n = \sqrt{2n} + \frac{1}{2} + O((1/n)^{1/2}),$$

the expansion of the value of $\ln g_n(m_n) = \ln g_n\left(\sqrt{2n}+\frac{1}{2}+O\left(\left(\frac{1}{n}\right)^{1/2}\right)\right)$ is

$$\ln g_n(m_n) = \frac{1}{2}\left(-1+\ln\left(\frac{1}{2n}\right)\right) + O\left(\left(\frac{1}{n}\right)^{1/2}\right).$$

So, we can approximate $g_n(m_n)$ by

$$g_n(m_n) = \frac{\mathrm{e}^{-1/2}}{\sqrt{2n}} + O(1/n).$$

Next, we study the value of $f_{\gamma(k_n,\theta_n)}(m_n)$.

Lemma 2: The expansions of the parameters μ_n, k_n and θ_n are the following:

$$\begin{aligned} \mu_n &= \sqrt{\pi}\sqrt{n} + O(1/n), \\ k_n &= \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}} + O(1/n), \\ \theta_n &= (\sqrt{\pi}-\sqrt{2})\sqrt{n} - \frac{1}{2} + O(1/n). \end{aligned}$$

Proof: The parameter θ_n can be written as:

$$\theta_n = \frac{1}{4}\left(\frac{2^n(n-2)!}{(2n-5)!!} - 2\left(\sqrt{8n-15}+1\right)\right).$$

If we expand the previous expression, we obtain:

$$\theta_n = \frac{1}{4}\left(\frac{2^{2n-2}((n-2)!)^2}{(2n-4)!} - 4\sqrt{2n} - 2\right) + O\left(\frac{1}{\sqrt{n}}\right).$$

Using that $(n-2)! = \Gamma(n-1)$ and the expansion (1), we have:

$$\begin{aligned} \theta_n &= \frac{1}{4}\left(4\sqrt{\pi}\sqrt{n}\mathrm{e}^{O\left(\frac{1}{n}\right)} - 4\sqrt{2n} - 2\right) + O\left(\frac{1}{\sqrt{n}}\right), \\ &= \sqrt{\pi n} - \sqrt{2n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right), \\ &= \left(\sqrt{\pi}-\sqrt{2}\right)\sqrt{n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

Thus, the expression for the parameter θ_n is obtained.

Next, we will proceed similarly with the parameter μ_n. This parameter can be written as:

$$\mu_n = \frac{2^{n-2}(n-2)!}{(2n-5)!!} = \frac{2^{2n-4}(n-2)!^2}{(2n-4)!}.$$

For the second time, using that $(n-2)! = \Gamma(n-1)$ and the expansion (1), we have:

$$\mu_n = \sqrt{\pi}\sqrt{n} + O\left(\frac{1}{n}\right).$$

Finally, using that $k_n = \frac{\mu_n}{\theta_n}$ and the previous expansions for the parameters μ_n and θ_n, we can obtain:

$$\begin{aligned} k_n &= \frac{\sqrt{\pi}\sqrt{n} + O\left(\frac{1}{n}\right)}{\left(\sqrt{\pi}-\sqrt{2}\right)\sqrt{n} - \frac{1}{2} + O\left(\frac{1}{\sqrt{n}}\right)} = \frac{\sqrt{\pi} + O\left(\frac{1}{n\sqrt{n}}\right)}{\sqrt{\pi}-\sqrt{2}+O\left(\frac{1}{n}\right)}, \\ &= \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}} + O\left(\frac{1}{n}\right), \end{aligned}$$

as we claimed.

 ISBN: 978-1-61208-137-3

Using these expansions, we obtain the following expression for the value of $f_{\gamma(k_n,\theta_n)}(m_n) = \frac{m_n^{k_n-1}\cdot e^{-\frac{m_n}{\theta_n}}}{\Gamma(k_n)\cdot\theta_n^{k_n}}$:

$$f_{\gamma(k_n,\theta_n)}(m_n) = \frac{\alpha}{\beta(n)},$$

where:

$$\begin{aligned}\alpha &= 2^{(1/2)\cdot(\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})-1)},\\ \beta(n) &= e^{\sqrt{2}/(\sqrt{\pi}-\sqrt{2})}\cdot\Gamma\left(\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})\right)\\ &\quad\cdot(\sqrt{\pi}-\sqrt{2})^{\sqrt{\pi}/(\sqrt{\pi}-\sqrt{2})}\cdot n^{-1/2}+O\left(\frac{1}{n}\right).\end{aligned}$$

We conclude that:

$$(g_n(m_n)+f_{\gamma(k_n,\theta_n)}(m_n))^2 = \frac{C}{n}+O\left(\frac{1}{n\sqrt{n}}\right),$$

where the constant C could be found using the expansions of $g_n(m_n)$ and $f_{\gamma(k_n,\theta_n)}(m_n)$.

Finally, an upper bound for the mean quadratic error MQE_n^u is found:

$$\begin{aligned}MQE_n^u &= \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_n,\theta_n)}(i))^2\\ &\leqslant \frac{n}{n-1}\cdot((g_n(m_n)+f_{\gamma(k_n,\theta_n)}(m_n))^2\\ &= \frac{C}{n-1}+O\left(\frac{1}{n\sqrt{n}}\right),\end{aligned}$$

and the right hand side term in this inequality tends to zero as n goes to infinity, as we claimed. This finishes the proof of Theorem 1.

III. The rooted case

By a *rooted phylogenetic tree* on S we mean a *fully resolved* (which in this case means with all its internal nodes of out-degree 2) rooted tree with its leaves bijectively labelled in the set S. As in the previous section, for simplicity we consider only the sets of labels $S_n = \{1,\ldots,n\}$, with n the number of leaves of the tree. Let $\mathscr{T}_n^r$ be the set of (isomorphism classes of) rooted phylogenetic trees on S_n. It is well known [4, Ch. 3] that $|\mathscr{T}_n^r| = |\mathscr{T}_{n+1}^u|$ for every $n \geqslant 1$.

Let $k,l \in S_n$ be any two different labels and let $T \in \mathscr{T}_n^r$. The *distance*, or *path length*, $d_T^r(k,l)$ between the leaves k and l in T is the length of the unique *undirected* path between them. We consider now the random variable

$$d_{kl}^r = \text{distance between } k \text{ and } l \text{ in one tree in } \mathscr{T}_n^r,$$

which takes values in $\Omega^r = \{2,3,\ldots,n\}$. Arguing as in Section II, when the trees and the leaves are chosen equiprobably, we are reduced to study the variable $d_n^r := d_{12}^r$.

For every $i \in \Omega^r$, let

$$c_{i,n}^r = |\{T \in \mathscr{T}_n^r \mid d_T^r(1,2) = i\}|.$$

The sequence $(c_{i,n}^r)_{i=2,\ldots,n}$ is the distribution of the variable d_n^r.

We have the following result connecting $c_{i,\cdot}^r$ with $c_{i,\cdot}^u$. For the sake of completeness, we sketch a direct proof, although it could be deduced from the explicit computations provided in [6], [11].

Lemma 3: $c_{i,n}^r = c_{i,n+1}^u$, for every $n \geqslant 2$ and $i = 2,\ldots,n$,

Proof: Consider the usual bijection $\Phi : \mathscr{T}_n^r \to \mathscr{T}_{n+1}^u$ that sends a rooted tree $T \in \mathscr{T}_n^r$ to the unrooted tree $\Phi(T) \in \mathscr{T}_{n+1}^u$ obtained by adding a new leaf labeled $n+1$ and a new edge connecting the root of T with this leaf (cf. [4, Ch. 3]). Then, $d_T^r(1,2) = d_{\Phi(T)}^u(1,2)$, and therefore Φ induces a bijection

$$\{T \in \mathscr{T}_n^r \mid d_T^r(1,2) = i\} \to \{T \in \mathscr{T}_{n+1}^u \mid d_T^u(1,2) = i\}.$$

This lemma allows one to translate Theorem 1 into the rooted case as follows:

Theorem 2: Let $\gamma(k_{n+1},\theta_{n+1})$ be the gamma distribution with parameters k_{n+1} and θ_{n+1} given by

$$\begin{aligned}k_n &= \frac{2^{n-2}\cdot(n-2)!}{2^{n-2}\cdot(n-2)!-(2n-5)!!\lceil(\sqrt{8n-15}+1)/2\rceil},\\ \theta_n &= \frac{2^{n-2}(n-2)!}{(2n-5)!!}-\left\lceil\frac{1+\sqrt{8n-15}}{2}\right\rceil.\end{aligned}$$

Let MQE_n^r be the mean quadratic error between the random variable d_n^r and this gamma distribution:

$$MQE_n^r = \frac{1}{n-1}\sum_{i=1}^{n-1}(c_{i,n}^u - f_{\gamma(k_{n+1},\theta_{n+1})}(i))^2.$$

Then, $\lim_{n\to\infty} MQE_n^r = 0$.

IV. Experimental results

Figure 1 shows the data plot of $(c_{i,n}^u)_{i=1,\ldots,n-1}$ and the gamma density function with parameters $k = \frac{\sqrt{\pi}}{\sqrt{\pi}-\sqrt{2}}$ and $\theta_n = (\sqrt{\pi}-\sqrt{2})\sqrt{n}$ as functions of i, for $n = 5000$ leaves. The figure confirms that the distribution of d_n^u approximates well this gamma density function.

Figure 2 shows the data plot of minus the logarithm of the mean quadratic error function $(-\ln(MQE_n))$ as a function of the number n of leaves. The curve hints at the existence of parameters α and β such that $-\ln(MQE_n) \approx \alpha+\beta\ln(n)$, that is, $MQE_n \approx e^{-\alpha}\cdot n^{-\beta}$. If we adjust the values of the α and β using the least squares method, we obtain $\alpha \approx 4.659$ and $\beta \approx 1.44$. This confirms the theoretical result in Section II.

V. Conclusion

In this paper, we have proven that the distribution of the distance between a fixed pair of leaves in an equiprobably, randomly chosen, fully resolved phylogenetic tree with n leaves approximates a gamma distribution as n goes to ∞. This result holds in the rooted and the unrooted case.

Our result is purely numerical, and it remains to be seen whether there is some deep meaning in the relationship between the distribution of the distances in phylogenetic

 ISBN: 978-1-61208-137-3

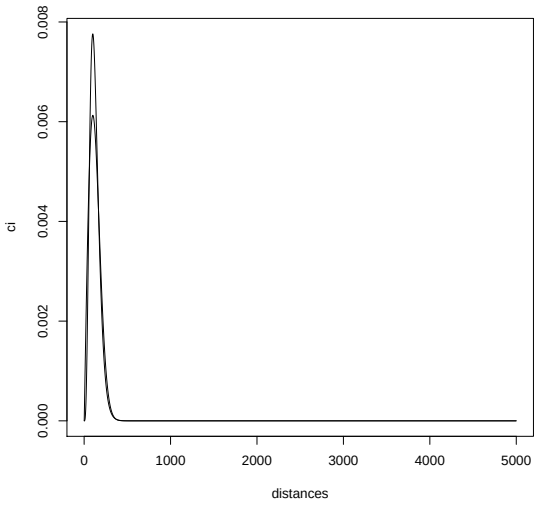


Figure 1. Data plot of $(c^u_{i,n})_{i=1,\ldots,n-1}$ and the corresponding gamma density function for $n = 5000$ leaves. The higher curve corresponds to the gamma density function, the lower one to d^u_n.

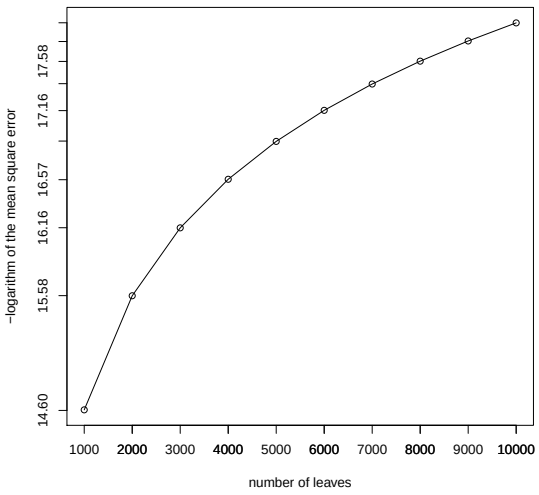


Figure 2. Data plot of $-\ln(MQE_n)$ as a function of the number n of leaves.

trees and a gamma distribution. Another unanswered question is whether the distances between pairs of leaves in the phylogenetic trees contained in some phylogenetic database, see TreeBASE ([1]) or PhylomeDB ([5]) are well approximated by using a gamma distribution. A negative answer would give information on the random model for real-life phylogenetic trees. We are working currently in this topic.

ACKNOWLEDGMENT

This work has been partially supported by the Spanish Government, through projects MTM2009-07165 and TIN2008-04487-E/TIN.

REFERENCES

[1] M. J. Sanderson, M. J. Donoghue, W. Piel, and T. Eriksson, TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. American Journal of Botany 81 (1994), pp. 183-189. http://www.treebase.org/ (last visited: March 11, 2011)

[2] M. Blum, N. Bortolussi, E. Durand, and O. François, AP-Treeshape: statistical analysis of phylo- genetic tree shape. Bioinformatics 22 (2006), pp. 363-364.

[3] H. Chang and M. Fuchs, Limit theorems for patterns in phylogenetic trees. Journal of Mathematical Biology 60 (2010), pp. 481-512.

[4] J. Felsenstein: Inferring Phylogenies. Sinauer Associates Inc. (2004)

[5] J. Huerta-Cepas, A. Bueno, J. Dopazo, and T. Gabaldón, PhylomeDB: a database for genome-wide collections of gene phylogenies. Nucleic Acids Research 36 (2008), D491-6. http://phylomedb.org/ (last visited: March 11, 2011)

[6] A. Mir and F. Rosselló, The mean value of the squared path-difference distance for rooted phylogenetic trees. Journal of Mathematical Analysis and Applications 371 (2010), pp. 168-176

[7] A. Mir and F. Rosselló, The median of the distance between two leaves in a phylogenetic tree. Advances in Bioinformatics, Proc. IWPACBB 2010 (M.P. Rocha *et al*, eds.), Advances in Intelligent and Soft Computing, vol. 74 (Springer, 2010), pp. 131-135.

[8] A. Mir and F. Rosselló, The mode of the distance between two leaves in a phylogenetic tree. X Jornadas de Bioinformática (Málaga, Spain, october 2010).

[9] N. Rosenberg, The mean and variance of the numbers of r-pronged nodes and r-caterpillars in Yule-generated genealogical trees. Annals of Combinatorics 10 (2006), pp. 129-146.

[10] M. Steel and A. Mooers, The expected length of pendant and interior edges of a Yule tree. Applied Mathematics Letters 23 (2010), pp. 1315-1319

[11] M.A. Steel and D. Penny, Distributions of tree comparison metrics—some new results. Systematic Biology 41 (1993), pp. 126-141

 ISBN: 978-1-61208-137-3

De Novo Draft Genome Assembly Using Fuzzy K-mers

John Healy
Department of Computing & Mathematics
Galway-Mayo Institute of Technology
Ireland
e-mail: john.healy@gmit.ie

Desmond Chambers
Department of Information Technology
National University of Ireland Galway
Ireland
e-mail: des.chambers@nuigalway.ie

Abstract- Although second generation sequencing technology can be used to rapidly sequence an entire genome, assembly algorithms require a high level of coverage to produce a complete genomic sequence. We describe a fuzzy *k*-mer approach that is capable of rapidly ordering and orientating low coverage sequence reads with a high level of accuracy. Using this approach, a draft genome of *Mycoplasma genitalium*, sampled at varying low levels of coverage, was accurately anchored against the genome of *Mycoplasma pneumoniae*. The anchored reads were assembled into scaffolds with a vastly increased N50 length and an error rate of <1.5%.

Keywords- genome assembly, anchoring, fuzzy k-mers, fuzzy hash maps

I. INTRODUCTION

The rapid evolution of genome sequencing technologies in recent years has lead to a reappraisal of sequencing alignment and assembly strategies [1-3]. Using massive parallelism, second generation sequencing (SGS) technologies are capable of rapidly producing a very large number of short reads [3-8]. These twin characteristics of read number and read length have resulted in a move away from assembly strategies based on the traditional overlap graph [9] to more *k*-mer centric approaches, such as sequence graphs and de Bruijn graphs [10-12].

Regardless of the sequencing technology employed, the assembly of a set of sequence reads into a complete or draft genome is predicated on a sufficient number of overlapping reads being made available to an assembler. The level of overlaps, or coverage, is a function of the amount of oversampling employed during sequencing. To create a set of read fragments that represents 99.9% of a genome, an eight-fold (8X) level of coverage is required [13, 14]. Notwithstanding this and the recent rapid advances in DNA sequencing technology, many of the published genomes available in public repositories are of draft quality, at coverage levels as low as 2X [15]. Despite an acceleration in the number of completed genomes, the sheer number of potential candidate species available implies that most will either never be sequenced, or will be sequenced to draft quality only [16]. There is thus an evident niche for applications that are capable of generating sizable assemblies from sets of low-coverage sequence reads.

A. Comparative Assembly

As the number of sequenced organisms increases, alternative approaches to genome assembly based on orthologous relationships become ever more viable. Comparative *de novo* assembly algorithms map sequence reads to a high-quality reference genome and use the resultant anchoring information to direct the assembly process. Originally proposed by Pop [17], the AMOS comparative assembler employs an *alignment-layout-consensus* approach to genome assembly. AMOS uses a complete, high-quality sequence of a closely related organism to determine the placement of reads in a layout graph.

More recently, the related, but distinct concept of assisted assembly was proposed by Gnerre [18]. Designed for use with low-coverage sequences, assisted assembly reinforces information already present in reads to detect erroneous or missed overlaps during the initial phase of genome assembly. Simultaneously constructing both a *de novo* and a comparative assembly, proximity relationships between reads are used to guide the assembly process.

B. Hash Tables and Variability

Given the recent trend towards *k*-mer centric genome assembly, the application of a similar approach to comparative assembly is worthy of consideration. Although the use of *k*-mers has a long history in both sequence alignment [19, 20] and sequence assembly [10-12], the underlying implementation typically manifests itself in the form of hash tables or hash maps. Hash tables and maps are dictionary data structures that use a key and hashing function to provide rapid, **O**(1), access to a set of mapped values [21]. As hash maps are capable of quickly detecting exact matches between keys, they are an ideal data structure for use in *k*-mer centric alignment and assembly applications. In a hash data structure, the hash key is used to functionally determine a mapped value. The implication of this property is that, although redundancy is permitted among the values in a hash map, the hash keys must be unique. While this uniqueness requirement provides hash structures with the underlying property to facilitate speed, it constrains access to exact matches of keys. This renders hash data structures intolerant of variations in sequence composition, such as sequence errors, polymorphisms, insertions and deletions, common in biological sequences.

 ISBN: 978-1-61208-137-3

To circumvent the constraints imposed by the uniqueness requirement of hash keys, alignment applications have employed a number of different strategies. Chief among these is the "seed and extend" strategy used by BLAST [19], which applies a hash table to seed exact *k*-mer matches, before attempting to join high-scoring alignments using dynamic programming. An alternative approach is the use of spaced seeds [22], which permit a degree of mismatch at pre-determined positions in a sequence. However, both approaches facilitate sensitivity by sacrificing the speed inherent in the hash structure.

Richer, object-oriented, programming languages permit the extension of hash maps to provide built-in support for key variability. Originally proposed by Topac [23], a Fuzzy Hash Map (FHM) applies fuzzy capabilities to traditional hash structures, with a minimal reduction in access speed. FHMs permit a degree of variation in hash keys and can be used for inexact sequence comparison.

To illustrate the applicability of FHM to sequence alignment and assembly, a draft genome of *M.genitalium*, sampled at very low levels of coverage, was anchored against the complete genome of *M.pneumoniae*, which was then used to guide the assembly process. The approach is highly effective for both ordering and orientating low coverage sets of reads into assembly contigs and scaffolds. The remainder of this discussion includes a description of how fuzzy *k*-mers can be implemented using a FHM. This is followed by a description of the anchoring and assembly process and the presentation of results. Finally, the mechanism used to test the validity of the approach is described and conclusions presented.

II. FUZZY HASH MAPS AND FUZZY *K*-MERS

Unlike procedural programming languages, object-oriented languages allow arbitrary objects to act as keys and values in a hash map [21]. The rapid access time of hash maps is accomplished by transforming a key value to an integer value that corresponds to a table index. When a collision between a search term and a hash key is detected, this transformation is applied to provide access to the mapped value. In the Java language, the semantics of object equality is determined by the implementation of the *hashCode()* and *equals()* methods [24]. When searching a hash map for a given key, if two *hashCode()* methods return the same integer value, an initial collision is detected. The *equals()* method is then executed to resolve any ambiguity and determine if a full collision has occurred.

FHMs manipulate the relationship between both of these methods by encouraging initial collisions based on part of the hash key and using the *equals()* method to permit a degree of variability in the remainder of the key. The degree of similarity is determined by the implementation of the *equals()* method, which can employ any sequence similarity algorithm that is capable of returning a fuzzy value between 0 and 1.

As depicted in Figure 1, fuzzy *k*-mers can be accommodated in a FHM by specifying the part of the *k*-mer to be used when computing the hash code. The remainder of the *k*-mer is evaluated by encapsulating a sequence similarity algorithm inside the *equals()* method. Using standard object-oriented techniques such as composition and inheritance, any edit distance algorithm such as Levenshtein Distance [25], Hamming Distance [26] and the Smith-Waterman algorithm [27] can be used. Consistent with traditional "seed and extend" strategies, the design of a fuzzy hash key is a trade-off between speed and sensitivity. Computing a hash code on too small a part of a hash key has the effect of flattening the FHM into a list, reducing the speed in proportion to the time complexity of the sequence similarity algorithm. In practice, specifying a minimum word size of 11 bases in the *hashCode()* implementation allows variability in the remainder of a *k*-mer, with little or no impact on running time.

III. ANCHORING AND ASSEMBLING FUZZY K-MERS

To illustrate the relevance and utility of FHMs to comparative genome assembly, the approach was used to anchor and assemble a draft genome of *Mycoplasma genitalium* using the complete genome of *Mycoplasma pneumoniae* to direct part the assembly. A *k*-mer centric anchoring and assembly strategy was applied, which consists of four main phases: anchor detection and extraction, anchor alignment, contig assembly and contig scaffolding.

A. Anchor Detection and Extraction

Given a complete, high-quality reference genome, a de

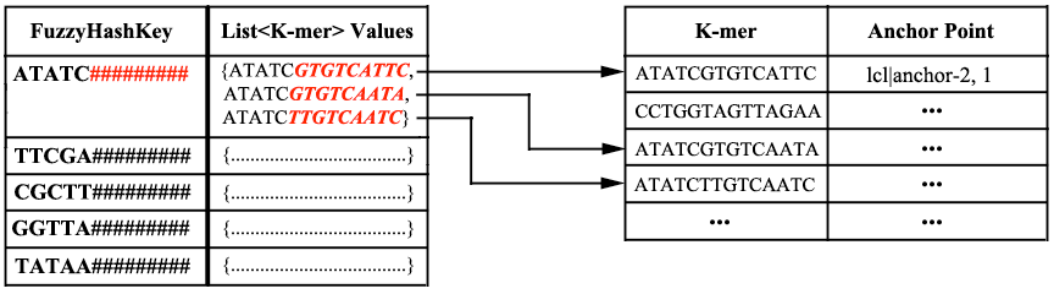


Figure 1. A Fuzzy Hash Map with a hash key initialised to cause collisions if the first five bases in a sequence are the same.

 ISBN: 978-1-61208-137-3

Bruijn graph can be created that represents a perfect tiling path through the genome. For a genome of size n, the de Bruijn graph will have $\mathbf{O}(n)$ nodes and $\mathbf{O}(n)$ edges, regardless of the number of reads in an assembly [28]. Each node in the graph represents a k-mer and can be weighted to reflect the multiplicity of matches to the sequence it contains. In the FHM approach, the multiplicity is not denoted by an integer value, but by labelling each node with read edges. A read edge represents the alignment of part of a genome with the k-mer contained by a node. Thus, nodes composed with more than one read edge represent repetitive sequences and are easily identified and, if necessary, avoided. The anchor detection and extraction process requires that the full reference genome be parsed and transformed into a de Bruijn graph (Figure 2). Although the memory requirements of a de Bruijn graph are huge, the memory consumption can be greatly reduced by merging all nodes with an in-degree and out-degree of 1. In the case of anchor detection, an additional constraint of merging only nodes with a multiplicity of 1 yields a sequence graph, where each merged edge represents a unique anchoring region. These anchoring regions are easily detected using a Depth-First Search [29] and are extracted and written to a FASTA file and to a serialized map. It is noteworthy that this process is executed once for each reference genome and is not undertaken as part of the assembly.

B. *Anchor Alignment*

Anchoring the reads from a draft genome requires that all reads, in both forward and reverse orientations, be compared against each anchor sequence. Before the alignment and assembly phases commence, the anchoring sequences are first parsed and read into a FHM. The FHM must be configured with a *FuzzyHashKey* that specifies the parts of each k-mer to be used to compute similarity. In addition, the *FuzzyHashKey* must also be configured with the sequence similarity algorithm to use and a fuzzy threshold value. Only alignment matches above the fuzzy threshold will cause a full collision in the FHM and indicate a match. Thus, a fuzzy threshold of 0.65 will only result in a match if a hash code causes an initial collision in the FHM and 65% of the remainder of the k-mer matches a hash key. Read alignment is accomplished by decomposing each read into a set of overlapping k-mers and attempting to add each k-mer to the FHM. If a match is found in the FHM, the name and index of the anchor is recorded, along with the orientation of the read. After the read has been aligned in both orientations, a majority count is use to determine the correct orientation of the read and its order with respect to the anchor. Each anchor maintains a list of the name, orientation and starting

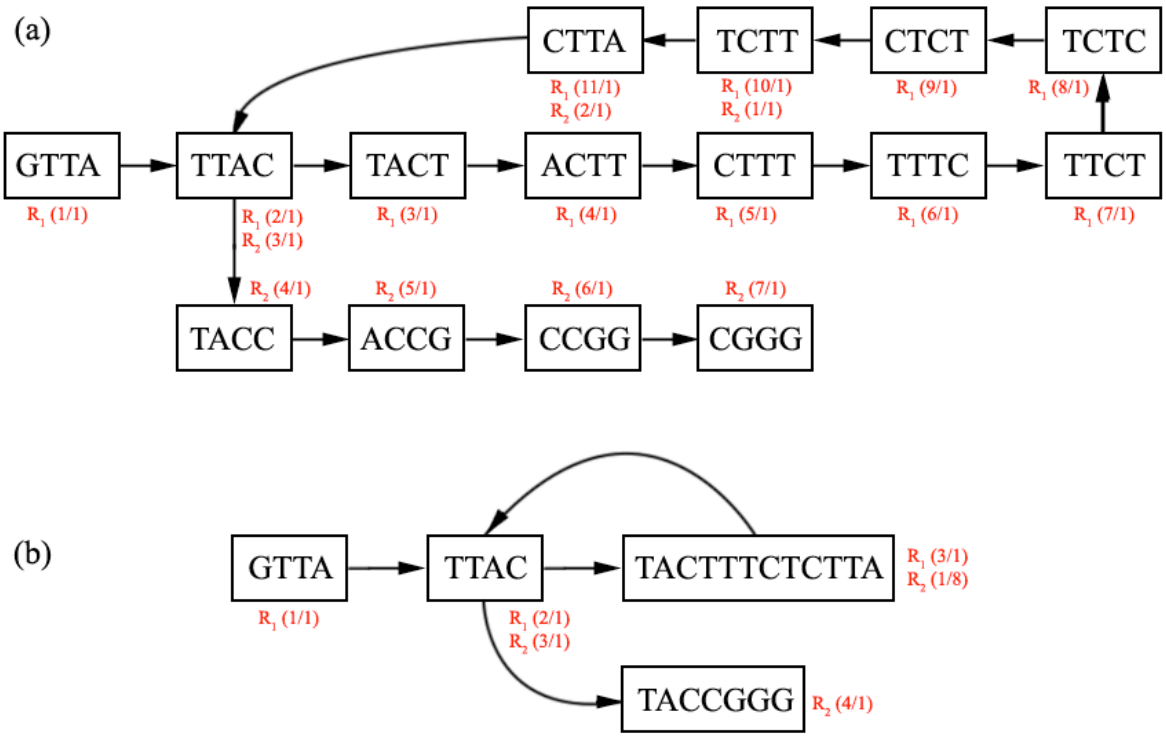


Figure 2. (a) A 4-mer de Bruijn graph for the overlapping sequences GTTACTTTCTCTTA and TCTTACCGGG. In practice k-mer sizes of at least 24 are used. As each read is added to the graph, the index position of the read (in red) relative to the sequence of the graph node is recorded. This information enables reliable transversal through highly repetitive nodes, by following the indices of the current read in increasing order. (b) Transformation to a sequence graph can be achieved by merging together nodes with an in-degree and out-degree of 1. The starting index of each read with respect to the merged node is altered to reflect the length of the newly merged sequence. The transformation to a sequence graph has the effect of significantly reducing the number of nodes and edges in the graph.

 ISBN: 978-1-61208-137-3

position of each read and automatically sorts the set of reads using a priority queue. As each anchor knows its own starting index with respect to the reference genome, the alignment process not only orientates, but also orders, each anchored read.

C. Contig Assembly

The assembly of overlapping reads into contigs is based on the application of a de Bruijn graph, in a manner similar to that used for anchor detection and extraction. As each draft read is parsed, an attempt is made to anchor the read using the procedure described above. If a read has been anchored, it is added to the de Bruijn graph only in the orientation given by the anchor. Otherwise, a read is added to the graph in both orientations. After parsing the full set of draft reads, the de Bruijn graph is transformed into a sequence graph by merging together all nodes with an in-degree and out-degree of 1. Assembly commences by generating a stack representing the set of source nodes in the graph. At low levels of coverage, there is an insufficient number of overlapping reads to provide a single path through the graph. Thus, the total number of source nodes in the graph indicates the minimum number of contigs available to the assembler.

Contig assembly is facilitated by the labelling of graph nodes with read edges. Starting at a source node, the list of read edges is processed to identify the starting sequence of an unprocessed read. The best candidate read edge can be identified by its index with respect to its own read and its index relative to the graph node. Using the current read as a heuristic value, neighbouring edges are evaluated and offered to a priority queue. The priority queue applies a distance constraint to determine if a read edge is permissible and, after examining all neighbouring edges, returns the next best edge to the assembler. By applying a distance constraint in this manner, the assembler will select the correct path when it encounters a branch in the graph. In addition to selecting the next edge, the priority queue is also responsible for determining the next read to process.

The labelling of nodes with read edges also facilitates the transversal of loops in the graph, which represent repetitive sequences. Using the current read as a heuristic, if an adjacent node has more than one read edge for the current read, the priority queue will only select a read edge that meets the distance constraints.

D. Contig Scaffolding

Given a set of contigs from the initial assembly phase, the anchoring information can now be applied to order and group the contigs. Although the absolute index of each anchor sequence with respect to the reference genome is known, this information cannot be used to determine the distance between contigs, as there may be large insertions or deletions in the reference sequence. Thus, the scaffolding of contigs involves the full ordering of contigs and the grouping of contigs linked by anchors into sub-contigs.

This final phase of assembly involves polling each anchor from a sorted queue of anchors. If an anchor spans more than one contig, at least two reads must be present in the adjacent contig to establish a join. As each anchor is processed, its aligned reads are removed in ascending order of alignment index. When all the reads have been polled from an anchor, the next anchor is removed from the anchor queue and the process iterates. The final assembly output is a FASTA file containing the assembled reads and an XML document. The XML document contains information about the set of contigs, including the constituent reads, read index and orientation.

IV. Results and Discussion

The 0.58Mb genome of *M.genitalium* was randomly sampled at coverage levels from 0.1X-2.0X and assembled using the 0.81Mb genome of *M.pneumoniae* as a reference sequence. Using a *k*-mer size of 24 and, with fuzzy index and fuzzy threshold values of 11 and 0.8 respectively, the FHM approach anchored 65.56% of the *M.genitalium* reads. Empirical evidence demonstrates that, for genomic sequences, as the fuzzy index decreases below 11, the number of collisions in the FHM increase exponentially until the access time reaches $\mathbf{O}(n)$, at which point the running time of the FHM is no better that that of an indexed list. The fuzzy threshold of 0.8 reflects the close genetic relationship between *M.genitalium* and *M.pneumoniae*. For more divergent species, this parameter should be relaxed to permit a greater tolerance of sequence variability during the anchoring phase.

The results of the assembly at various levels of coverage are shown in Table 1. Among the more salient features of the fuzzy assembly approach, is the low percentage of orientation errors. This illustrates the utility of genome anchoring in general and the FHM in particular, for determining the correct orientation of a sequence read, even at very low levels of coverage.

TABLE I. Summary of assembly results at varying levels of coverage.

Coverage	N50 Contig	N50 Scaffold	% Ordering Errors	% Orientation Errors	Time (s)
2.0	2141	51215	1.21	0.12	19.2
1.8	1918	14787	4.60	1.00	17.3
1.6	1798	14853	1.12	0.17	16.2
1.4	1739	9734	0.39	0.59	14.6
1.2	1456	10322	2.18	0.46	12.8
1.0	1269	6001	1.24	0.97	11.1
0.8	1228	8240	1.55	0.69	7.5
0.6	992	4743	0.92	0.46	7.8
0.4	-	2450	2.07	0.00	6.0
0.2	-	2539	1.38	2.07	4.1

The N50 metric indicates that 50% of bases are in contigs of size *n* or greater. Again, the effectiveness of the approach can be seen by comparing the N50 size for the contigs generated by the initial assembly with the N50 size of scaffolded contigs. Even at ultra-low levels of coverage,

 ISBN: 978-1-61208-137-3

the assembler is capable of generating sizable contigs with low ordering and orientation errors. The execution time exhibits a logarithmic growth rate in the order **O**(log *n*). Allowing for the parsing of an ever-increasing number of reads, this slow growth rate illustrates that using fuzzy *k*-mers in this manner does not compromise running time.

V. VALIDATION OF APPROACH

To establish the validity of the approach, an automated testing framework was developed that is capable of examining and scoring the order and orientation of each read in the assembly. The set of reads for each draft genome was randomly sampled from a complete genome, by providing information such as desired coverage level, average read length and clone insert length to a validation framework. The output of the read generation process is a set of randomly sampled and oriented reads in FASTA format, representing the draft genome, and a specialised data structure containing validation data for each read. The validation data includes the correct order and orientation of reads, the length of each read and the distance between adjacent reads.

After anchoring and assembling the draft genome, the set of assembled reads was validated, by computing a local alignment of each contiguous set of reads against the full ordered set of randomly sampled sequences. This was accomplished by creating a dynamic programming matrix and scoring the list of reads in each contig against the full list of reads generated by the framework. It should be noted that the dynamic programming matrix requires only the names of reads and their relative distances to compute an alignment score. Furthermore, a positive score in the programming matrix requires a read to be both in order and at the correct distance relative to its adjacent reads.

In addition to ascertaining the correct order of reads, the test framework also computes, from a suffix of the FASTA sequence name, the correct orientation of each read. The local alignment was implemented using a modification of the Smith-Waterman [27] algorithm. Ancillary information, such as N50 size and Lander-Waterman [14] statistics, is also generated by the validation framework.

VI. CONCLUSION

The anchoring and assembly process described is capable of rapidly ordering and orientating reads from a draft genome with a low level of errors. In particular, the anchoring mechanism is highly effective in orientating reads, thereby reducing the size of the graph created by the assembler and simplifying the assembly of contigs. Directing assembly using sets of anchored reads enables the construction of large contig scaffolds, even at low levels of coverage. Furthermore, the application of a FHM data structure permits a degree of variability between sequences, without sacrificing execution speed. Finally, the fuzzy k-mer approach allows a high degree of error tolerance that is invaluable when processing biological sequences that contain sequence errors, insertions and deletions.

VII. REFERENCES

[1] S. Batzoglou, "The many faces of sequence alignment," *Briefings in bioinformatics,* vol. 6, p. 6, 2005.

[2] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," *Brief Bioinform,* p. bbq015, 2010.

[3] M. Schatz, A. Delcher, and S. Salzberg, "Assembly of large genomes using second-generation sequencing," *Genome Research,* vol. 20, p. 1165, 2010.

[4] P. Flicek and E. Birney, "Sense from sequence reads: methods for alignment and assembly," *Nature Methods,* vol. 6, pp. S6-S12, 2009.

[5] C. Fuller, L. Middendorf, S. Benner, G. Church, T. Harris, X. Huang, S. Jovanovich, J. Nelson, J. Schloss, and D. Schwartz, "The challenges of sequencing by synthesis," *nature biotechnology,* vol. 27, pp. 1013-1023, 2009.

[6] C. Hutchison III, "DNA sequencing: bench to bedside and beyond," *Nucleic Acids Research,* 2007.

[7] J. Shendure and H. Ji, "Next-generation DNA sequencing," *nature biotechnology,* vol. 26, pp. 1135-1145, 2008.

[8] A. Sundquist, M. Ronaghi, H. Tang, P. Pevzner, and S. Batzoglou, "Whole-genome sequencing and assembly with high-throughput, short-read technologies," *PLoS One,* vol. 2, 2007.

[9] J. Kececioglu and E. Myers, "Combinatorial algorithms for DNA sequence assembly," *Algorithmica,* vol. 13, pp. 7-51, 1995.

[10] J. Butler, I. MacCallum, M. Kleber, I. Shlyakhter, M. Belmonte, E. Lander, C. Nusbaum, and D. Jaffe, "ALLPATHS: De novo assembly of whole-genome shotgun microreads," *Genome Research,* vol. 18, p. 810, 2008.

[11] J. Simpson, K. Wong, S. Jackman, J. Schein, S. Jones, and Birol, "ABySS: A parallel assembler for short read sequence data," *Genome Research,* vol. 19, p. 1117, 2009.

[12] D. Zerbino and E. Birney, "Velvet: Algorithms for de novo short read assembly using de Bruijn graphs," *Genome Research,* vol. 18, p. 821, 2008.

[13] R. Fleischmann, M. Adams, O. White, R. Clayton, E. Kirkness, A. Kerlavage, C. Bult, J. Tomb, B. Dougherty, and J. Merrick, "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science,* vol. 269, p. 496, 1995.

[14] E. Lander and M. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics,* vol. 2, pp. 231-239, 1988.

[15] I.-M. A. C. Konstantinos Liolios, Konstantinos Mavromatis, Nektarios Tavernarakis, Philip Hugenholtz, Victor M. Markowitz and Nikos C. Kyrpides, "The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata," *Nucleic Acids Research,* vol. 38, 2010.

[16] N. Hall, "Advanced sequencing technologies and their wider impact in microbiology," *Journal of Experimental Biology,* vol. 210, p. 1518, 2007.

[17] M. Pop, A. Phillippy, A. Delcher, and S. Salzberg, "Comparative genome assembly," *Briefings in bioinformatics,* vol. 5, p. 237, 2004.

[18] S. Gnerre, E. Lander, K. Lindblad-Toh, and D. Jaffe, "Assisted assembly: how to improve a de novo genome assembly by using related species," *Genome biology,* vol. 10, p. R88, 2009.

[19] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology,* vol. 215, pp. 403-410, 1990.

[20] W. Pearson and D. Lipman, "Improved tools for biological sequence comparison," *Proceedings of the National Academy of Sciences,* vol. 85, p. 2444, 1988.

[21] M. Goodrich and R. Tamassia, "Data Structures and Algorithms in Java," John Wiley & Sons, 2001.

[22] B. Ma, J. Tromp, and M. Li, "PatternHunter: faster and more sensitive homology search," *Bioinformatics,* vol. 18, p. 440, 2002.

 ISBN: 978-1-61208-137-3

[23] V. Topac, "Efficient fuzzy search enabled hash map," 2010, pp. 39-44.

[24] J. Gosling, B. Joy, G. Steele, and G. Bracha, *Java (TM) Language Specification, The (Java (Addison-Wesley))*: Addison-Wesley Professional, 2005.

[25] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," 1966.

[26] R. Hamming, "Error detecting and error correcting codes," *Bell System Technical Journal,* vol. 29, pp. 147-160, 1950.

[27] T. Smith and M. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology,* vol. 147, pp. 195-197, 1981.

[28] M. Chaisson and P. Pevzner, "Short read fragment assembly of bacterial genomes," *Genome Research,* vol. 18, p. 324, 2008.

[29] R. Tarjan, "Depth-first search and linear graph algorithms," 1971, pp. 114-121.

 ISBN: 978-1-61208-137-3

Internal Force Filed in Proteins

Marchewka Damian
Department of Bioinformatics and Telemedicine
Medical College Jagiellonian University
Lazarza 16, 31-530 Krakow, Poland
damian.a.marchewka@gmail.com

Roterman Irena*
Department of Bioinformatics and Telemedicine
Medical College Jagiellonian University
Lazarza 16, 31-530 Krakow, Poland
corresponding Author: myroterm@cyf-kr.edu.pl

***Abstract*— The proteins representing the structures of ordered form in respect to their tertiary structure are discussed. The hydrophobic core of 3-D Gauss function ("fuzzy oil drop" model) appeared to be present in some proteins. The proteins of the structure with ordered form of vdW and/or electrostatic internal interaction in protein body are discussed in the paper. The vdW interaction distribution in proteins appeared to be accordant with assumed 3-D Gauss function while the electrostatic represents the distribution of random form. This characteristics allows interpretation of the tertiary structure as the effect of external/internal force field influence expressed by 3-D Gauss function in respect to hydrophobic and/or vdW interaction. It is postulated that the additional introduction of 3-D Gauss function representing the influence of external environment (in contrast to internal force field of protein body) in the simulation of protein folding process in silico may simplify the optimization procedure leading to the appropriate order on the level of tertiary structure of protein molecule and directing the hydrophobic residues toward the center of the protein body with the exposure of hydrophilic residues on the surface. The procedure minimize the differences between internal interactions and the idealized one expressed by external force field.**

Internal force field; External force field; Interactions in protein; 3-D Gauss function; Information theory

I. INTRODUCTION

The procedure of protein structure prediction is based on the search for polypeptide structure of low internal energy expressed mostly by side chain - side chain interaction of electrostatic, vdW forms. Thus the optimization procedure (minimization of the energy) is the most important element of the procedure oriented on the protein structure prediction especially in ab initio (new fold – according to CASP nomenclature – Critical Assessment of Protein Structure Prediction) as well as homology search based (comparative modeling - according to CASP nomenclature) computational techniques. The CASP initiative organized every second year is the event assumed to monitor the progress in protein structure prediction [1].

The groups of proteins representing highly ordered structure in respect to the hydrophobic density distribution have been found: downhill proteins [2], antifreeze proteins [3] and some proteins acting in form of homodimers [4]. The high accordance with the 3-D Gauss function with the empirically observed hydrophobicity density distribution [5] in protein body suggested the search for possible ordered force field of other character like electrostatic, dipole-interaction oriented etc.

This is why the analysis of the distribution of the energy components like: electrostatic interaction and vdW interaction additionally to hydrophobic interaction in protein body was undertaken. The question was, whether other than hydrophobic interaction represents the ordered character on the tertiary structure level.

The search for the order system on the level of tertiary structure of proteins is presented in this work based on earlier observed organization of hydrophobic core organization accordant with 3-D Gauss function.

II. MATERIAL AND METHODS

A. Data

The proteins representing different status of hydrophobic core organization were selected for analysis (Tab. I). The selection of proteins was done according to the analysis of down-hill proteins presented in [2].

TABLE I. THE LIST OF PROTEINS UNDER CONSIDERATION. THE PDB ID, LENGTH OF POLYPEPTIDE CHAIN, SOURCE ORGANISM, SECONDARY STRUCTURE DESCRIPTION, BIOLOGICAL FUNCTION AND REFERENCES ARE GIVEN

Protein	N	Source	Structure description	Biological function	Reference
1HZC	66	bacteria	β-barrel	Cold shock protein	[6]
1BDC	60	bacteria	Mainly helical	Immunoglobulin binding domain	[7]
1VII	36	chicken	Mainly helical	Villin subdomain Actin binding	[8]
2I5M	66	bacteria	Mainly β-structural	Cold shock protein	[9]
1CSP	67	bacteria	Mainly β-structural	Cold shock protein	[10]
1RIJ	23	De novo design	Mainly helical	De novo design	[11]

B. Energy optimization using Gromacs

The Gromacs program was applied to run the energy optimization procedure to relax the crystal structure.
All EMs (energy minimization) have been performed with Gromacs software package v4.0.3 and Gromos96 43a1 force field [12-16]. The coordinates for starting structures have been taken from the Protein Data Bank. In first step, all EMs have been compared an in vacuo model to a solvated model.

 ISBN: 978-1-61208-137-3

Default protonation states and hydrogen positions were generated by pdb2gmx utility of the Gromacs package. The energy optimization procedure was performed in water solvent. SPC water model was used [16]. The total charge of the molecule was null.

The parameters for energy minimization procedure were as follows:

Maximum number of iterations - 100 steps; The minimization was converged when the max force was smaller than 1000.0 kJ mol-1 nm-1; Initial step size - 0,01 nm; Method to determine neighbour list – Grid; Treatment of long range electrostatic interactions – cut-off; Long range electrostatic cut-off - 1.0 nm; Long range van der Waals cut-off – 1.0nm; Cut-off distance for short-range neighbour list - 1.0nm; Constraint algorithm used to restrain bond lengths – none;

Frequency to update the neighbour list -10 steps;

The individual interactions of particular residues with the rest of the protein molecule was performed using the make_ndx procedure defining the "group" under consideration and g_energy program in order to extract data from output energy files.

Each residue was taken as one group while the rest of protein molecule was defined as second group. The energy calculation was performed for each amino acid in this system. The set of individual interactions was standardised to the unit making the interaction distribution unified allowing the comparison with other distributions (theoretical and random one).

C. *"Fuzzy oil drop" model*

The assumption of this model is the accordance of hydrophobic (and possible other) interactions in protein with the idealized one expressed by 3-D Gauss function [5]. The procedure allowing generation of this type of force field is shown below.

1) The theoretical hydrophobicity density distribution: The geometric center of the protein molecule is localized in the origin of coordinate system. The longest distance between two effective atoms (averaged position of atoms belonging to side chain) determines the orientation of the X-axis. The longest distance between two projections (on the YZ-plane) of effective atoms determines the orientation of the Y-axis. The longest distance between elements along each axis in coordinate system is expressed by 3σ. The hydrophobicity density in each position of effective atom can be calculated as follows:

$$\widetilde{Ht}_j = \frac{1}{\widetilde{Ht}_{sum}} \exp\left(\frac{-(x_j-\bar{x})^2}{2\sigma_x^2}\right)\exp\left(\frac{-(y_j-\bar{y})^2}{2\sigma_y^2}\right)\exp\left(\frac{-(z_j-\bar{z})^2}{2\sigma_z^2}\right)$$

where $\bar{x},\bar{y},\bar{z}$ are the coordinates of the geometric center of the molecule (usually located in the origin of the coordinate system). This is why these values can be considered equal to zero. The size of the molecule is expressed by the triple σ_x, σ_y, σ_z, which is calculated for each molecule individually provided that the orientation of the molecule with the longest possible inter-effective atoms distance is determined according to the appropriate coordinate system axis. The σ values are calculated as the 1/3 of the longest distance between two effective atoms calculated along each axis. The value of the Gauss function at any point of protein body is treated as the idealized hydrophobic density defining the hydrophobic core.

2) Observed distribution: On the other hand, the empirical hydrophobicity distribution is calculated according to the function presented by Levitt [17]:

$$\widetilde{Ho}_j = \frac{1}{\widetilde{Ho}_{sum}} \sum_{i=1}^{N} (H_i^r + H_j^r) \begin{cases} \left[1 - \frac{1}{2}\left(7\left(\frac{r_{ij}}{c}\right)^2 - 9\left(\frac{r_{ij}}{c}\right)^4 + 5\left(\frac{r_{ij}}{c}\right)^6 - \left(\frac{r_{ij}}{c}\right)^8\right)\right] & \text{for } r_{ij} \le c \\ 0 \text{ for } r_{ij} > c \end{cases}$$

where N expresses the number of amino acids in the protein (number of grid points), $\widetilde{H}_i^r$ expresses the hydrophobicity of the i-th residue according to the accepted hydrophobicity scale (the scale presented in [18]) was applied in this work, r_{ij} expresses the distance between the i-th and j-th interacting residues, and c expresses the cutoff distance, which according to the original paper [17] is assumed to be 9 Å. The values of $\widetilde{Ho}_j$ are standardized by dividing them by the coefficient $\widetilde{Ho}_{sum}$, which is the sum of all hydrophobicities attributed to grid points.

3) Electrostatic and vdW interactions: The individual interactions of particular residues with the rest of the protein molecule was performed using the make_ndx procedure defining the "group" under consideration. Each residue was taken as one group while the rest of protein molecule was defined as second group. The energy calculation was performed for each amino acid in this system. The set of individual interactions was standardised to the unit making the interaction distribution unified allowing the comparison with other distributions (theoretical and random one).

4) The analysis of distributions: To evaluate quantitatively the accordance between the idealized and empirically observed distribution of the density of selected parameter (interaction), divergence entropy (also known as Kullback-Leibler entropy [19] was calculated:

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

where $D_{KL}(P \parallel Q)$ denotes the distance entropy (also called deficiency/divergence entropy), which is a measure of the distance between P(i) and Q(i) distributions (probabilities), where Q(i) plays the role of target distribution.

The values of Q(i) were taken according to the 3-G values for the ellipsoid of particular protein. This target (reference)

 ISBN: 978-1-61208-137-3

function was commonly used for all types of interaction (electrostatic and vdW) under consideration. The values of P(i) expressed the density of particular type of interaction calculated in relation to the sum of interaction of each residue with the entire protein molecule. The values expressing particular type of interaction were standardized to make the sum of all values equal to 1 after the unified rescaling of negative and positive values.

Since the entropy values can be interpreted only in the relative scale the comparison of observed distribution with the random one was performed. The protein of the distance between observed distribution (O) and theoretical one (T) expressed as O/T lower in relation to the distance between O and random distribution (R) (O/R) was treated as the protein of distribution accordant with expected one.

III. RESULTS

Hydrophobicity distribution: the 3-D Gauss function was taken as the target distribution for hydrophobic force field. The Kulback Leibler entropy values are given in Tab. II. All the proteins classified as downhill proteins appeared to represent the structure accordant with the idealized hydrophobic core.

A. Density distribution in proteins under consideration

The distribution profile of each component of the force field for selected proteins: 1HZC and1BDC are shown in Fig. 1 and Fig.2 respectively.

The profiles visualize the range of similarity/discrepancy between expected and observed distribution. The high accordance between idealized hydrophobic distribution and observed one can be seen except 1HZC. The high accordance between random and observed electrostatic density can be seen in all cases.

B. Summary of the internal interaction in proteins

The summary characterizing the structure of internal force field is given in Tab.II.

TABLE II. THE O/T AND O/R ENTROPY VALUES CALCULATED FOR INDIVIDUAL TYPES OF INTERACTIONS (HYDROPHOBIC, ELECTROSTATIC AND VDW) TAKING THE IDEALIZED 3-D GAUSS FUNCTION (T) AS THE TARGET AND THE RANDOM DISTRIBUTION (R) TO MAKE POSSIBLE INTERPRETATION OF THE ENTROPY VALUES. THE VALUES FOR STRUCTURES ACCORDANT WITH ASSUMED MODEL ARE GIVEN IN BOLD.

Protein	Hydrophobicity		Electrostatic		vdW	
	O/T	O/R	O/T	O/R	O/T	O/R
1HZC	0.213	0.207	0.323	0.057	0.307	0.252
1BDC	**0.121**	**0.141**	0.348	0.097	**0.149**	**0.186**
1VII	**0.223**	**0.568**	0.336	0.115	0.226	0.101
2I5M	**0.188**	**0.559**	0.444	0.215	**0.116**	**0.188**
1CSP	**0.134**	**0.466**	0.266	0.035	**0.103**	**0.190**
1RIJ	**0.171**	**0.583**	0.244	0.110	**0.079**	**0.087**

The proteins characterized in Tab. II. were selected to represent different status in respect to the order of energy components distribution.

The protein 1HZC (cold shock protein) of the form of compact β-barrel proteins without disulfide bonds and cis-proline residues has been recognized as the molecule of low stability [6]. The absence of ordered form of hydrophobic core as well as absence of any ordered form of internal force field seems to explain the low stability of this molecule (Fig.1.).

The protein 1BDC (immunoglobulin binding domain) represents the mainly helical structural form [7]. According to the analysis presented in this paper its stability may the result of the ordered structure of hydrophobic core as well as ordered vdW internal force field (Fig. 2.).

IV. CONCLUSIONS

The results presented in this paper suggest that the tertiary level organization is expressed by the ordered form of hydrophobic as well as vdW interaction force field. No regularity (random distribution) was found for electrostatic interaction. The local, biological function related charge presence (enzymatic active site) was not taken into account. It was aimed to analyze the non-specific distribution of charges (the electrostatic interaction).

The regularity of the hydrophobicity distribution identified in downhill proteins (as well as in antifreeze proteins [3] and some homodimers [4]) suggests that the folding process is directed by the hydrophobic interaction in the form accordant with the 3-D Gauss function. The introduction of the external force field of 3-D Gauss function during the folding process simulation may facilitate the structure optimization process in silico. The presence of external force field may direct the hydrophobic residues toward the center of the protein body with the exposure of hydropholic residues on the surface [18]. The folding process accordant with high density of vdW interactions in the center of the protein molecule may additionally introduce the expected order of residues in the space. The “fuzzy oil drop” model was proved performing the molecular dynamics simulation of trans-membrane protein. The simulation performed in the presence of external force field in form of 3-D Gauss function for hydrophobic interaction revealed high accordance of results with those received using the traditional simulation performed in the presence of membrane and water molecules [20]. The regression function comparing the results received using the explicite water molecules and “fuzzy oil drop” model was of the form y=1*x. The time consumption for “fuzzy oil drop” model was significantly lower in comparison with traditional molecular dynamics simulation in all-atoms form [20].

The proteins representing different secondary structures and different biological function appeared to represent also different accordance with the assumed model expecting the density distribution of particular type of interaction accordant with 3-D Gauss function. The ordered distribution of particular type of interaction seems to generate the ordered internal force field probably responsible for tertiary stabilization. The differences between proteins of different

 ISBN: 978-1-61208-137-3

structural (secondary structure) characteristics of the internal force field suggest different mechanism of the structure generation. The absence of the accordance of the assumed order in respect to electrostatic interactions suggests low influence of external force field of electrostatic character. The discordance between the expected (3-D Gauss distribution) and the observed one was recognized to appear due to the presence of ligand (including also the protein-protein complexation interaction area) [5]. It may suggests the aim-oriented local disorder related to specific biological function [18].

ACKNOWLEDGMENT

The work was financially supported by Jagiellonian University – Medical College grant K/ZDS/001531.

REFERENCES

[1] CASP http://predictioncenter.org/

[2] Roterman I., Konieczny L., Jurkowski W., Prymula K. and Banach M. (2011) Two-intermediate model to characterise the structure of fast-folding proteins – submitted

[3] Banach M., Prymula K., Jurkowski W., Konieczny L. and Roterman I. (2011) Fuzzy oil drop model to interpret the structure of antifreeze proteins and their mutants. J. Mol. Mod. – in press

[4] Banach M., Konieczny L. and Roterman I. (2011) "Fuzzy oil drop" model to identify the complexation area in protein homodimers. – submitted

[5] Konieczny L., Brylinski M. and Roterman I. (2006) Gauss-function-based model of hydrophobicity density in proteins In Silico Biol 6, 0002.

[6] Delbruck H., Mueller U., Perl D., Schmid F.X. and Heinemann U. (2001) Crystal structures of mutant forms of the Bacillus caldolyticus cold shock protein differing in thermal stability. J.Mol.Biol. 313: 359-369

[7] Gouda H., Torigoe H., Saito A., Sato M., Arata Y. and Shimada I. (1992) Three-dimensional solution structure of the B domain of staphylococcal protein A: comparisons of the solution and crystal structures. Biochemistry. 31(40): 9665-72

[8] McKnight C.J., Matsudaira P.T. and Kim P.S. (1997) NMR structure of the 35-residue villin headpiece subdomain. Nat.Struct.Biol. 4: 180-184

[9] Max K.E., Wunderlich M., Roske Y., Schmid F.X. and Heinemann U. (2007) Optimized variants of the cold shock protein from in vitro selection: structural basis of their high thermostability. J.Mol.Biol. 369: 1087-1097

[10] Schindelin H., Marahiel M.A. and Heinemann U. (1993) Universal nucleic acid-binding domain revealed by crystal structure of the B. subtilis major cold-shock protein. Nature 364: 164-168

[11] Liu Y., Liu Z., Androphy E., Chen J. and Baleja J.D. (2004) Design and characterization of helical peptides that inhibit the E6 protein of papillomavirus. Biochemistry 43: 7421-7431

[12] Berendsen H.J., van der Spoel D. and van Drunen R. (1995) GROMACS: a message-passing parallel molecular dynamics implementation. Comput Phys Commun 91: 43-56

[13] Berendsen H.J.C., Postma J.P.M., van Gunsteren W.F. and Hermans J. (1981) Interaction models for water in relation to protein hydration. In: Intermolecular Forces. Pullman, B. ed. . D. Reidel Publishing Company Dordrecht pp 331–342

[14] van der Spoel D., Lindahl E., Hess B., Groenhof G., Mark A.E. and Berendsen H.J. (2005) GROMACS: fast, flexible, and free. J Comput Chem 26: 1701-1718

[15] Lindahl E., Hess B. and van der Spoel D. (2001) GROMACS 3.0: a package for molecular simulation and trajectory analysis. J Mol Model 7: 306-317

[16] van der Spoel D., Lindahl E., Hess B., van Buuren A.R., Apol E., Meulenhoff P.J., Tieleman D.P., Sijbers A.L., Feenstra K.A., van Drunen R. and Berendsen H.J. Gromacs User Manual version 3.3. www.gromacs.org; 2005b

[17] Levitt M. (1976) A simplified representation of protein conformations for rapid simulation of protein folding. J Mol Biol 104: 59-107

[18] Brylinski M., Konieczny L. and Roterman I. (2007) Is the protein folding an aim-oriented process? Human haemoglobin as example. Int J Bioinform Res Appl. 3: 234-260

[19] Nalewajski R. F. (2006) Information theory of molecular systems. Amsterdam: Elsevier

[20] Zobnina V. and Roterman I. (2009) Application of the fuzzy-oil-drop model to membrane protein simulation Proteins 77: 378-394

 ISBN: 978-1-61208-137-3

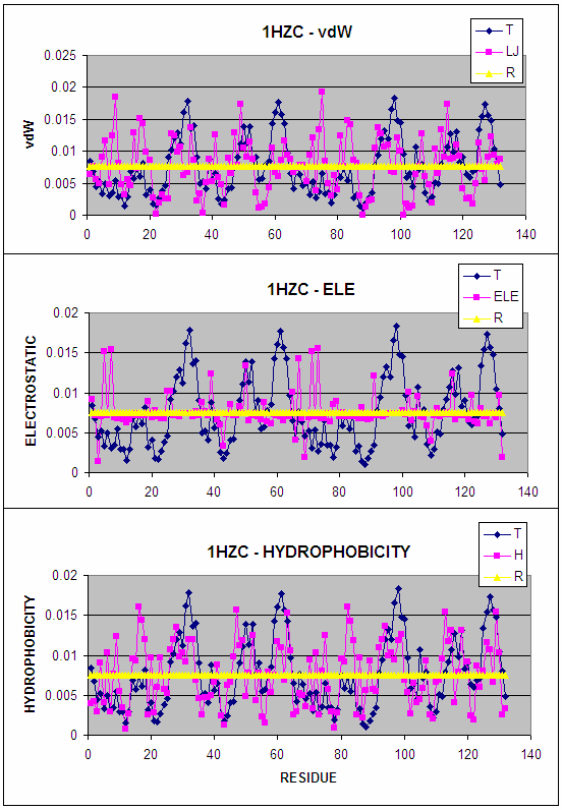


Figure 1. The profiles representing the density distribution of electrostatic, vdW and hydrophobic interaction in protein body in 1HZC. The lack of accordance can be seen in all profiles.

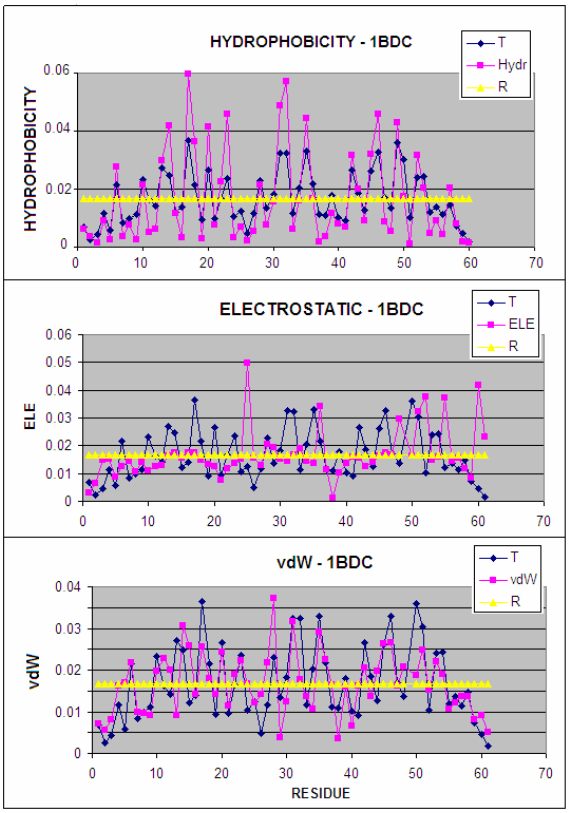


Figure 2. The profiles representing the density distribution of electrostatic, vdW and hydrophobic interaction in protein body in 1BDC. The lack of accordance can be seen in profile of electrostatic interaction.

 ISBN: 978-1-61208-137-3

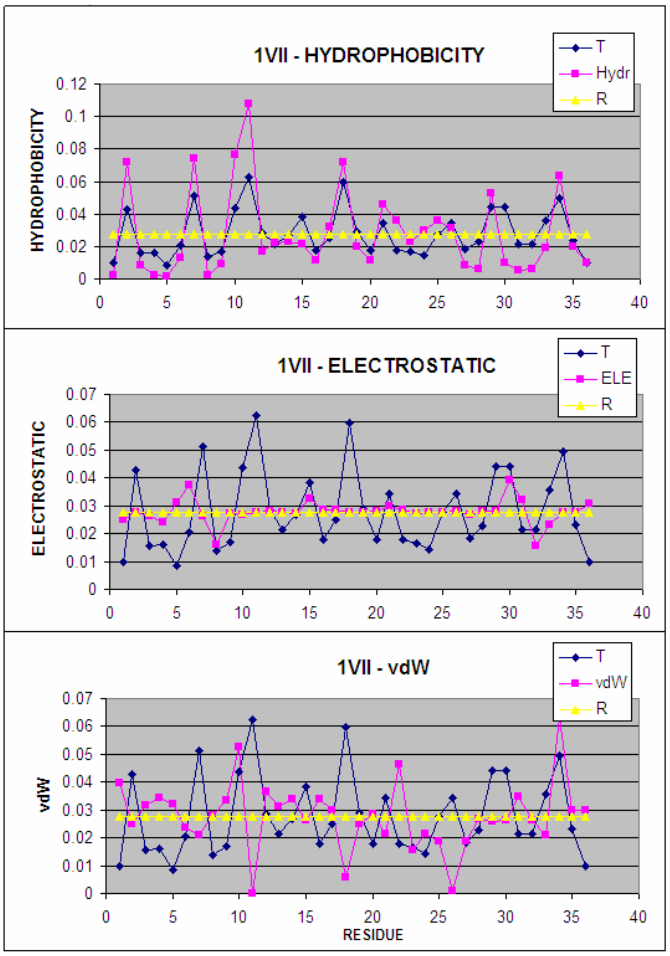


Figure 3. The profiles representing the density distribution of electrostatic, vdW and hydrophobic interaction in protein body in 1RIJ. The high accordance between idealized and observed distribution can be seen for hydrophobic and vdW interactions.

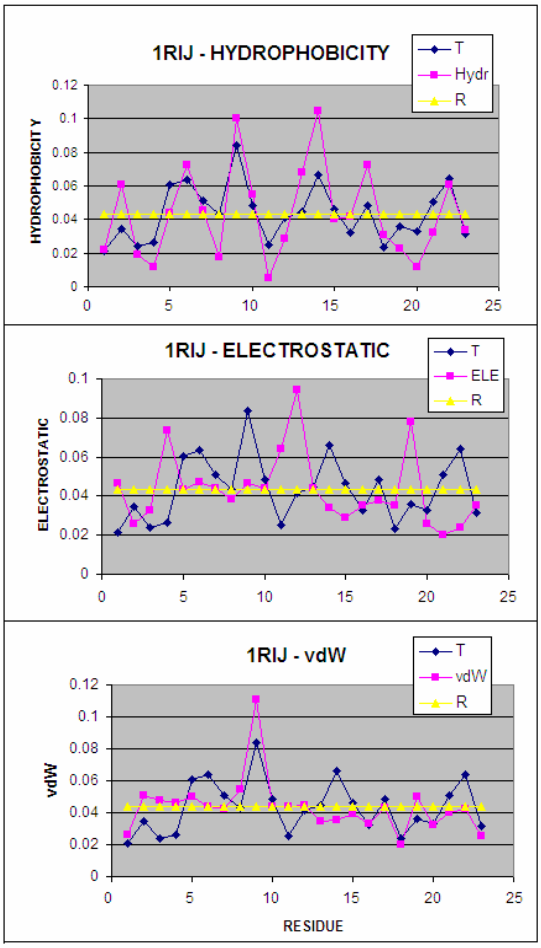


Figure 4. The profiles representing the density distribution of electrostatic, vdW and hydrophobic interaction in protein body in 1VII. The lack of accordance can be seen in profile of electrostatic interaction.

 ISBN: 978-1-61208-137-3

Structural Characterization of the Rieske Oxygenase Complex from *Burkholderia fungorum* DBT1 strain: Insights from bioinformatics

Stefano Piccoli[1], Silvia Lampis[2], Giovanni Vallini[3], Alejandro Giorgetti[4]

Department of Biotechnology, University of Verona, Ca' Vignal 1 Strada Le Grazie 15, I-37134 Verona, Italy
e-mail: 1 stefano.piccoli@univr.it; 2 silvia.lampis@univr.it; 3 giovanni.vallini@univr.it; 4 alejandro.giorgetti@univr.it

***Abstract*—Polycyclic aromatic hydrocarbons (PAHs) represent a class of organic compounds that negatively affect human health. These compounds are of toxicological concern because some of them have been identified as carcinogenic, mutagenic, and teratogenic. *Burkholderia fungorum* DBT1 is a bacterial strain, first isolated from an oil refinery discharge, which can utilize dibenzothiophene (DBT), phenanthtrene and napthalene as substrates for growth. This strain is capable of degrading DBT nearly completely through the "Kodama pathway" more efficiently than others. The work presented here is aimed at a structural characterization at the molecular level of the proteins involved in the first step of the PAH degradation pathway, i.e., the Rieske Oxygense (RO) complex. Thus, using state-of-the-art structural bioinformatics tools we have built the structural models of each of the members of the RO complex encoded in the *Burkholderia fungorum* DBT1 strain. The structural characterization combined with future molecular biology experiments may give important insights into the functioning of this particular strain.**

Keywords - biocomputing; *Burkholderia fungorum DBT1*; dibenzothiophene; modeling; PAHs.

I. INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) are an ubiquitous class of hydrophobic organic compounds consisting of two or more fused aromatic rings. PAHs are widespread in the environment and persist over long periods of time: many polycyclic aromatic hydrocarbons (PAHs) are largely suspected to be mutagenic or carcinogenic [1], and their contamination in soil and aquifer is of great environmental concern.

Of the PAHs occurring in soils and groundwaters, about 0.04 - 5% (wt/wt) are sulfur heterocycles [2] among which dibenzothiophene (DBT) represents the prevailing compound. This is therefore taken into account as model chemical structure in studies dealing with either biodegradation of organo-sulfur contaminants by petroleum biodesulfurisation through the "4-S pathway"[3] or through the "Kodama pathway" [4] [5] [6]. The latter transforms the molecule to the final product 3-hydroxy-2-formylbenzothiophene (HFBT). Denome *et al.* [7] cloned the genes responsible for the Kodama pathway from *Pseudomonas fungorum*. These genes, that are organized in a single operon, encode enzymes of the upper naphthalene catabolic pathway, and belong to a group of genes showing a high degree of sequence identity with the *nah* genes from *Pseudomonas putida* G7 [8]. The *nah*-like class of genes, cloned from different microorganisms, are highly conserved and are involved in the transformation of molecules that constitute the low molecular weight fraction of PAHs, including DBT [9] [10] [11] [12] [13]. They are normally clustered in a single operon under the control of a single promoter.

A novel genotype for the initial steps of the oxidative degradation of dibenzothiophene was recently found in the bacterial strain *Burkholderia fungorum* DBT1 isolated from a drain receiving oil refinery wastewater [14]. *Burkholderia fungorum* DBT1 is a NON-PATHOGENIC strain capable of transforming DBT completely through the "Kodama pathway" with higher efficiency than other microorganisms. This strain shows a particular genomic organization for the initial steps of the oxidative degradation of PAHs when compared to previously described genes capable of PAHs catabolism. In fact the genes are organized in two operons instead of one: they are called pH1A-p46 and p51 respectively [14] (GenBank accession numbers AF380367 and AF404408 respectively). In DBT1genes involved in DBT transformation show only low similarity with the corresponding conserved isofunctional oxidative genes. The unusual gene organisation suggests the possibility of novel features of DBT transformation in natural context.

Rieske Oxygenase (RO) systems have been shown to catalyze the first step in the Kodama pathway [15]. ROs produce *cis*-dihydrodiols from a large variety of substrates and molecular oxygen (dioxygen).

RO systems use electrons from NAD(P)H to activate molecular oxygen, which is then used to oxidize the substrate. RO systems are composed of two or three components, including a reductase, a ferredoxin (not found in all systems), and an oxygenase (Fig. 1). The reductase component liberates electrons from NAD(P)H and transfers the electrons to the ferredoxin. The ferredoxin shuttles the electrons to the oxygenase, where they are used in catalysis. In systems where the ferredoxin is absent, the reductase transfers electrons directly to the oxygenase. The oxygenase component of these systems is responsible for catalysis. This component consists of an alpha subunit, which contains both a Rieske binding domain and a catalytic domain. In some cases, a beta subunit is present, which is believed to

 ISBN: 978-1-61208-137-3

primarily function as a stabilizer for the alpha subunits. Rieske Oxygenase (RO) systems of *Burkholderia fungorum* DBT1 is called dibenzothiophene dioxygenase [14]. While structural studies have been performed on a number of ROs, no crystal structure exists for the DBT1 enzymes. Dibenzothiophene dioxygenase is able to degrade a wide spectra of molecules, including naphthalene, phenanthrene and DBT [16]. For its particular characteristics, *Burkholderia fungorum* DBT1 might be interestingly exploited in bioremediation protocols of PHA-contaminated sites. Ultimately, therefore, understanding the molecular basis of ligand-target interactions in this system may be fundamental for a complete characterization of the mechanism of action of the present strain and for future applications in bioremediation protocols of PHA-contaminated sites. Computational molecular biology (CMB) and protein structural bioinformatics approaches are keys to face these challenges, especially when, as for this particular case, no crystal structures exist for the different molecular components.

The work presented here is aimed at a structural characterization at the molecular level of the proteins involved in the first step of the PHA degradation pathway, i.e., the RO system. We have built the structural models of each of the components encoded in the *Burkholderia fungorum* DBT1 strain. Although the work is in a preliminary phase, still several conclusions can already been drawn specially in the formation of protein complexes involved in the Kodama pathway.

II. MATERIALS AND METHODS

The sequence alignment of the different targets and their corresponding structural templates were extracted from the multiple sequence alignment considering the entire families of interest.

We then constructed the models for each of the members of the RO complex of *Burkholderia fungorum* DBT1 strain. This was calculated as follows: all sequences and those of their families were retrieved from the Uniprot [17] database using *ssearch* [18]. They were aligned with PROMALS [19]. This multiple sequence alignment was then used for the definition of the Hidden Markov profile (HMM) of each of the target sequences. The profiles were then funneled through the HHsearch [20] program to identify the most plausible homologous structural templates. Such procedure is currently one of the best ones as evaluated from CASP7 experiment [21]. The multiple sequence alignments obtained in this way were used as the reference for the structural prediction of the different targets by homology modeling. The models were then built up by the use of the program Modeller9v4 [22]. Superposition of the structures, protein visualization and figures were carried out using the program VMD [23].

III. RESULTS AND DISCUSSIONS

Ferredoxin reductase. The nucleotide sequence of ferredoxin reductase component of dibenzothiophene dioxygenase complex was taken from some recent studies (data not published) that indicates the presence of a putative ORF that probably encodes for this subunit. This component of the dibenzothiophene dioxygenase complex was modeled by standard homology modeling [24] [25] procedures by the use of the program HHpred [26] in a semi-automatic fashion, as described in the methods section. The subunit was modeled based on the corresponding subunit of Benzoate 1,2-Dioxygenase Reductase from *Acinetobacter fungorum* Strain ADP1 [27] (PDB code 1KRH). The sequence identity shared between the ferredoxin reductase subunit with its template is 24%. The sequence alignments are available as supplementary material (Fig. S1 [28]). The selected template was not co-crystallized with NADP ligand. So the functional structure was built by optimal superposition of the main chain of ferredoxin reductase with the main chain of the X-ray solved structure. To improve the quality of the model we modeled the enzyme in a putative catalytic conformation, that is, with NADP ligand in the putative binding cavity. The ligand was manually transferred to the model from the structure of Pea FNR Y308S mutant from *Pisum sativum* (PDB code 1QFY) following the procedure used by Deng *et al.* [28]. That is, using the program VMD, we superposed the main chain coordinates of our model to the main chain of 1QFY protein transferring NADP coordinates into our model.

Ferredoxin (Ac. N. AAK96190). This subunit of the dibenzothiophene dioxygenase was modeled by standard homology modeling using the same protocol and criteria of ferredoxin reductase (Fig. S2 [28]). The subunit model was built using as template the corresponding subunit from *Sphingomonas yanoikuyae* B1 ferredoxin [29] (PDB code 2I7F), which shears 50% the sequence identity. This operation was performed as described in the Methods section. Studies of the interaction of ferredoxin with the reductase and with the oxygenase will be carried in a near future by the use of protein-protein docking techniques and validated by experiments.

Oxygenase (Ac. N. AAK62353 and AAK62354). The oxygenase component (α and β subunits) of dibenzothiophene dioxygenase was modeled by standard homology modeling using the same protocol seen for the others components. The alpha/beta complexes were modeled based on the alpha/beta subunits of the Bifenyl 2,3-dioxygenase from *Sphingobium yanoikuyae* B1. The sequence identity shared between the alpha and beta subunits with their templates is 49 and 37 % respectively. The sequence alignments are available as supplementary material (Fig. S3, S4 [28]). The selected template was not

 ISBN: 978-1-61208-137-3

co-crystallized with ligands. The functional hexamer (three alpha and three beta subunits) was built by optimal superposition of the main chain of the alpha/beta complexes with the main chain of the X-ray solved structure. To test the validity of our models we modeled the hexamer in a putative catalytic conformation, that is, with the ligands bound in the putative binding cavity. The moiety of dioxygenase with their cognate ligands were solved for a variety of members of the family, nevertheless, two structures were co-crystallized with ligands that also bind to dibenzothiofene, i.e., naphthalene and phenanthrene. These structures are the Naphthalene 1,2-Dioxygenase from *Pseudomonas* sp. strain NCIB 9816-4 [30] (PDB code 2HMK) and Naphthalene 1,2-Dioxygenase [31] (PDB code 1O7G) from *Pseudomonas putida*. Therefore, using the program VMD, we superposed the main chain coordinates of our model to the main chain of 2HMK and of 1O7G protein. We then transferred naphthalene and phenanthrene coordinates respectively, into our hexamer model. In the supplementary material (Fig. S5 [28]), the high structural conservation of the binding sites can be appreciated, showing the full conservation of the latter, although the ligands were transferred from independent crystal structures belonging to different species. The conservation of all the interacting residues, albeit the low sequence identity between templates and target, provide an initial, although non definitive, validation of the reliability of our models. Indeed, future work will include the virtual docking of dibenzothiofene into the binding cavity combined with experimental validation. The final modeled RO complex members can be appreciated in Figure 1.

IV. CONCLUSION

In the present work, we aimed at the modeling of the initial step in the Kodama pathway for the degradation of PAHs by the *Burkholderia fungorum* DBT1 strain. The availability of several templates covering the entire RO complex gave us the possibility of building not only the structural models of the isolated components but also to gain insight into the big protein complexes involved in the process. Although, the obtained results are preliminary and correspond to the first step of a lengthy iterative process of experimental and computational work, the possibility of modeling one of the most important proteins complexes and its validation with experiments extracted from literature, prompted us to hypothesize that more refined models will offer more a important overview of the system under study and may allow the full characterization of the entire pathway. Moreover, at the present stage site directed mutagenesis experiments can be already designed and proposed from the models. The extremely high three-dimensional conservation observed in the binding cavities will allow the production of targeted mutants that may permit a deeper characterization of the enzymatic mechanisms. Our analysis and modelling procedures allowed us to find, in the alpha subunit of RO complex of *Burkholderia fungoru*m DBT1 strain, an insertion of three amino acids very close to the active site that seems to be a duplication. This insertion is present only in *Burkholderia fungoru*m DBT1 strain and could represent a peculiar characteristic of DBT1 strain for substrate degradation efficiency. We are also involved in modelling of the complex between the alpha subunit of RO and the ferredoxin component and the transient complex between the ferredoxin and ferredoxin reductase subunit in different activation states. This work, will be extended to the entire Kodama pathway with the aim of characterizing such an interesting and efficient PAH-contaminated degradation organism.

V. REFERENCES

[1] Fujikawa, K., F.L. Fort, K. Samejima, and Y. Sakamoto. 1993. "Genotoxic potency in *Drosophila melanogaster* of selected aromatic amines and polycyclic aromatic hydrocarbons as assayed in the DNA repair test", *Mutat. Res.*, 290, 175-182.

[2] Thompson, C.J. 1981. "Identification of sulfur compound in petroleum and alternative fossil fuel", In *Organic sulfur chemistry* R.K. Freidlina and A.e. Skorova (eds)., 189-208. Pergamon Press, Oxford.

[3] Gallagher, J.R., E.S. Olson, and D.C. Stanley. 1993. "Microbial desulfurization of dibenzothiophene: a sulfur specific pathway", *Fems Microbiol. Lett.*, 107, 31-36.

[4] Kodama K., Umehara K., Shimizu K., Nakatani S., Minoda Y., and Yamada K. 1973. Identification of microbial products from dibenzothiophene and its proposed oxidation pathway. *Agric. Biol. Chem.* 37: 45-50.

[5] Kodama K., Nakatani S., Umehara K., Shimizu K., Minoda Y., and Yamada K. 1970. Microbial conversion of petro-sulfur compounds. Part III. Isolation and identification of products from dibenzothiophene. *Agric. Biol. Chem.* 34: 1320-1324.

[6] Kropp K.G. and Fedorak P.M. 1998. A review of occurrence, toxicity, and biodegradation of condensed thiophenes found in petroleum. *Can. J. Microbiol.* 44: 605-622.

[7] Denome S.A., Stanley D.C., Olson E.S., and Young K.D. 1993*b*. Metabolism of dibenzothiophene and naphthalene in *Pseudomonas* strains: complete DNA sequence of an upper naphthalene catabolic pathway. *J. Bacteriol.* 175: 6890-6901.

[8] Simon MJ., Osslund TD., Saunders R., Ensley BD., Suggs S., Harcourt A., Suen WC., Cruden DL., Gibson DT., and Zylstra GJ. 1993. Sequences of genes encoding naphthalene dioxygenase in *Pseudomonas putida* strains G7 and NCIB 9816-4. Gene 127: 31–37.

[9] Denome S.A., Olson E.S., and Young K.D. 1993. Identification and cloning of genes involved in specific desulfurization of dibenzothiophene by *Rhodococcus fungorum* Strain IGTS8. *Appl. Environ. Microbiol.* 59: 2837-2843.

[10] Menn FM., Applegate BM., and Sayler GS. 1993. NAH plasmidmediated catabolism of anthracene and phenanthrene to naphthoic acids. *Appl. Environ. Microbiol.* 59: 1938–1942.

[11] Sanseverino J., Applegate BM., King JM., and Sayler GS. 1993. Plasmid-mediated mineralization of naphthalene,

 ISBN: 978-1-61208-137-3

phenanthrene, and anthracene. *Appl. Environ. Microbiol.* 59: 1931–1937.

[12] Kiyohara H., Torigoe S., Kaida N., Asaki T., Iida T., Hayashi H., and Takizawa N . 1994. Cloning and characterization of a chromosomal gene cluster, *pah*, that encodes the upper pathway for phenanthrene and naphthalene utilization by *Pseudomonas putida* OUS82. *J. Bacteriol. 176*: 2439–2443.

[13] Geiselbrecht AD., Hedlund BP., Tichi MA., and Staley JT. 1998. Isolation of marine polycyclic aromatic hydrocarbon (PAH)-degrading *Cycloclasticus* strains from the Gulf of Mexico and comparison of their PAH degradation ability with that of Puget Sound *Cycloclasticus* strains. *Appl. Environ. Microbiol.* 64:4703–4710.

[14] Di Gregorio S., C. Zocca, S. Sidler, A. Toffanin, D. Lizzari, and G. Vallini. 2004. Identification of two new sets of genes for dibenzothiophene transformation in *Burkholderia* sp. DBT1. Biodegradation 15:111–123.

[15] Habe H. and Omori T. 2003. Genetic of polycyclic aromatic hydrocarbon metabolism in diverse aerobic bacteria. *Biosci Biotechnol Biochem*. 67: 225–243.

[16] Andreolli M., Lampis S., Zocca C., and Vallini G. 2008. Biodegradative potential of *Burkholderia fungorum* DBT1 in the abatement of polycyclic aromatic hydrocarbons. Proceedings (124) of the *4th European Bioremediation Conference*, Chania, Crete, Greece.

[17] Ballesteros JA., and Weinstein H. 1992. Analysis and refinement of criteria for predicting the structure and relative orientations of transmembranal helical domains. *Biophys J.* 62: 107–109.

[18] Scheerer P., Park JH., Hildebrand PW., Kim YJ., Krauss N., Choe HW., Hofmann KP., and Ernst OP. 2008. Crystal structure of opsin in its G-protein-interacting conformation. *Nature* 455:497–502.

[19] Altenbach C., Kusnetzow AK., Ernst OP., Hofmann KP., and Hubbell WL. 2008. High-resolution distance mapping in rhodopsin reveals the pattern of helix movement due to activation. *Proc Natl Acad Sci* U.S.A 105: 7439–7444.

[20] Wu CH., Apweiler R., Bairoch A., Natale DA., Barker WC., Boeckmann B., Ferro S., Gasteiger E., Huang H., Lopez R., Magrane M., Martin MJ., Mazumder R., O'Donovan C., Redaschi N., and Suzek B. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* 34: D187–D191.

[21] Ropelewski AJ., Nicholas HB., and Deerfield DW. 2004. Mathematically complete nucleotide and protein sequence searching using Ssearch. *Curr Protoc Bioinformatics*, Chapter 3, Unit3.

[22] Sali A. and Blundell T.L.. 1993. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* 234, 779-815.

[23] Humphrey W., Dalke A., and Schulten K. 1996. VMD: visual molecular dynamics. *J Mol Graph*. Feb;14(1):33-8, 27-8.

[24] Chothia C., and Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins."*EMBO J.* 5, 823–826.

[25] Tramontano, A. 2006. Protein Structure Prediction: Concepts and Applications, JohnWiley & Sons, Ltd.,Weinheim, Germany.

[26] Söding J. 2005. Protein homology detection by HMM-HMM comparison. *Bioinformatics* 21, 951-960.

[27] Karlsson A., Beharry ZM., Matthew D., Coulter ED., Neidle EL., Kurtz DM. Jr, Eklund H., and Ramaswamy S. 2002. X-ray crystal structure of benzoate 1,2-dioxygenase reductase from *Acinetobacter fungorum* strain ADP1. *Mol Biol. Apr.* 26;318(2):261-72.

[28] Supplementary Material are available from the web page: http://molsim.sci.univr.it/RO_complex.

[29] Deng Z., Aliverti A., Zanetti G., Arakaki AK., Ottado J., Orellano EG., Calcaterra NB., Ceccarelli EA., Carrillo N., and Karplus PA. 1999. A productive NADP+ binding mode of ferredoxin-NADP+ reductase revealed by protein engineering and crystallographic studies. *Nat Struct Biol.* Sep;6(9):847-53.

[30] Ferraro DJ., Brown EN., Yu CL., Parales RE., Gibson DT., and Ramaswamy S. 2007. Structural investigations of the ferredoxin and terminal oxygenase components of the biphenyl 2,3-dioxygenase from *Sphingobium yanoikuyae* B1. *MC Struct Biol*. 9;7:10.

[31] Ferraro DJ., Okerlund AL., Mowers JC., and Ramaswamy S. 2006. Structural basis for regioselectivity and stereoselectivity of product formation by naphthalene 1,2-dioxygenase. *J Bacteriol.* 188(19):6986-94.

[32] Karlsson A., Parales JV., Parales RE., Gibson DT., Eklund H., and Ramaswamy S. 2003. Crystal structure of naphthalene dioxygenase: side-on binding of dioxygen to iron. *Science* 14;299(5609):1039-42.

 ISBN: 978-1-61208-137-3

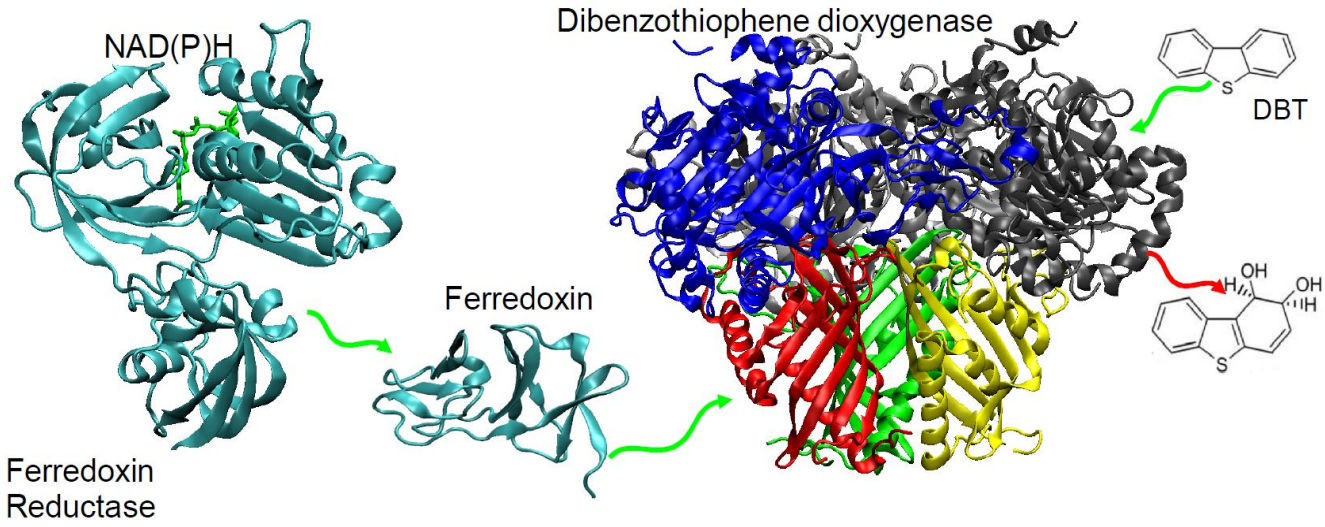


Figure 1. Schematic description of the oxidation process of polycyclic aromatic hydrocarbons (PAHs) catalyzed by Dibenzothiophene dioygenase complex (Rieske oxygenase system). The enzymes depicted were modeled using the protocol described in this article.

 ISBN: 978-1-61208-137-3

UMPIRE: Ultimate Microarray Prediction, Inference, and Reality Engine

Jiexin Zhang and Kevin R. Coombes
Department of Bioinformatics and Computational Biology
University of Texas M.D. Anderson Cancer Center
Houston, TX 77005, USA
kcoombes@mdanderson.org

***Abstract*—High-throughput measurements of gene expression pose a challenge to analysts attempting to learn models that predict treatment response or survival. One possible explanation for the lack of significant progress in this area is the limited sample size of most experiments. Realistic simulations could help with the development and assessment of analytical methods; however, existing simulation tools have focused more on the technology and less on the biological complexity. In this paper, we introduce a package of simulation tools to address this problem. Our model incorporates additive and multiplicative noise, transcriptional activity or inactivity, and block correlation structures. More importantly, it models the multi-hit theory of cancer via latent variables that link gene expression, binary outcome, and survival data. We illustrate the use of the simulation package by showing that standard analysis methods (i.e., univariate Cox models) are only likely to recover the true structure with more samples than are included in most current studies of survival.**

***Keywords*-gene expression; microarray; simulation; class prediction; multi-hit theory of cancer**

I. Introduction

The introduction of gene expression microarrays in the 1990's ushered in an era of high-throughput biology that has required the development of novel methods for the statistical and computational analysis of large biological datasets. Richard Simon and colleagues [1] identified three kinds of problems addressed by these technologies: class comparison, class discovery, and class prediction. The current state-of-the-art has evolved reasonable methods for class comparison (e.g., gene-by-gene t-tests or ANOVA coupled with estimates of the false discovery rate) and class discovery (e.g., hierarchical clustering coupled with resampling techinques to assess robustness) [2]. However, there is less agreement on the best (or even consistently good) methods for discovering complex models that can accurately predict biologically relevent outcomes such as treatment response or survival.

Part of the difficulty is that prediction is inherently harder than class comparison or class discovery. It is conceivable that the the number of samples (typically between 100 and 300) included in most of the current studies is simply inadequate to learn effective predictive models. It is, however, extremely difficult to assess this possibility. Although some progress has been made for binary classifiers [3]–[5], we do not have general theoretical ways to justify formal sample size computations that address the combination of feature selection and model building that goes into the discovery of predictive models from high-throughput biological datasets. Nor is it possible to collect gene expression data on $10,000$ patients in order to test empirically how many samples are really needed to learn good predictive models.

The obvious solution is to use simulation. If we can simulate many datasets, of different sizes, with realistic biological properties, then we can use those datasets to evaluate proposed methods for class prediction. The simulation of microarray gene expression datasets has a long history. However, none of the existing simulation tools was designed to focus on the biological diversity related to such important outcomes as treatment response or survival. Many of the earliest simulation tools focused on the simulation of microarray images, and were useful for developing better image processing algorithms [6]–[8]. Other simulation tools have attempted to explicitly model the steps in a microarray experiment, including printing, hybridization, dye effects, and scanning [9], [10]. As with many of the early statistical simulations [11]–[14], however, most tools use a model that simply compares two homogeneous populations of samples. Even more recent and more detailed simulations still assume that the data come from two homogenous popluations [5], [15]–[17].

To address this gap, we have developed a simulation package that incorporates a heterogeneous model that is consistent with the multiple hit theory of carcinogenesis [18], [19]. Moreover, our package uses latent variables to simulate the connections between gene expression and either binary or time-to-event outcomes.

II. Homogeneous gene expression model

Version 1.0 of the Ultimate Microarray Prediction, Inference, and Reality Engine (Umpire) is an R package that allows researchers to simulate complex, realistic microarray data that is linked to binary or time-to-event outcomes. The package is available from the R reposi-

 ISBN: 978-1-61208-137-3

tory at http://bioinformatics.mdanderson.org/OOMPA; detailed instructions on how to install the package can be found at http://bioinformatics.mdanderson.org/Software/OOMPA.

The fundamental object in `Umpire` is a "random-vector generator" (RVG), which is represented by the `Engine` class. Equivalently, each `Engine` object represents a specific multivariate distribution, from which random vectors can be generated using the generic `rand` method. In Version 1.0 of `Umpire`, we include three basic components for these kinds of distributions: independent normal, independent log normal, and multivariate normal. A general `Engine` is simply a list of RVG components. Because `Umpire` is implemented using S4 classes in R, adding additional components to implement alternative models of gene expression generation is a straightforward application of object-oriented programming.

A. Additive and Multiplicative Noise

The observed signal, Y_{gi}, for gene g in sample i is:

$$Y_{gi} = S_{gi} * \exp(H_{gi}) + E_{gi}$$

where

$$S_{gi} = \text{true biological signal}$$

$$H_{gi} = \text{multiplicative noise}$$

$$E_{gi} = \text{additive noise.}$$

The noise model represents technical noise that is layered on top of any biological variability when measuring gene expression in a set of samples. For example, background noise is usually additive, while the variation between the signal pixels is multiplicative noise. We modeled additive and multiplicative noise as normal distributions:

$$E_{gi} \sim \text{Normal}(\nu, \tau)$$

$$H_{gi} \sim \text{Normal}(0, \phi)$$

Note that we allow the additive noise to include a bias term (ν) that may represent, for example, a low level of cross-hybridization providing some level of signal at all genes. The noise model is represented in the `Umpire` package by the `NoiseModel` class. Again, the object-oriented and modular design make it possible to add more elaborate noise models in the future, such as those described by Nykter and colleagues [9].

B. Active and Inactive Genes

We model the true biological signal S_{gi} as a mixture:

$$S_{gi} \sim (1 - z_g) * \delta_0 + z_g * T_{gi}$$

In this model, δ_0 is a point mass at zero, z_g defines the activity state (1 = active, 0 = inactive), and T_{gi} is the expression of a transcriptionally active gene. By allowing for some genes to be transcriptionally inactive, this design takes into account that the transcriptional activity of most genes is conditional on the biological context. Activity is modeled in `Umpire` using a binomial distribution, $z_g \sim Binom(p_0)$.

C. Expression Distributions

For most purposes, we assume that the expression, T_{gi}, of a transcriptionally active gene follows a log-normal distribution, $\log(T_g) \sim Normal(\mu_g, \sigma_g)$. In a class of samples, the mean expression of gene g on the log scale is denoted by μ_g and the standard deviation on the log scale is σ_g. Both μ_g and σ_g are properties of the gene itself and the sample class. Within a given simulation, we typically place hyperdistributions on the log-normal parameters μ_g and σ_g. We take $\mu_g \sim Normal(\mu_0, \sigma_0)$ to have a normal distribution with mean μ_0 and standard deviation σ_0. We take σ_g to have an inverse gamma distribution with *rate* and *shape* parameters. Reasonable values for the hyperparameters can be estimated from real data. For instance, $\mu_0 = 6$ and $\sigma_0 = 1.5$ are typical values on the log scale of a microarray experiment using Affymetrix arrays. The parameters for the inverse gamma distribution are determined by the method of moments from the desired mean and standard deviation; we have found that a mean of 0.65 and a standard deviation of 0.01 (for which $rate = 28.11$ and $shape = 44.25$) produce reasonable data.

D. Correlated blocks of genes

Biologically, genes are usually interconnected in networks and pathways. In fact, clustering methods are often used to group genes into correlated blocks. Thus, it is natural to simulate microarray experiments from this perspective. In our simulations, we usually allow the mean block size, bs, to range from 1 to 1000, and the sizes of gene blocks to vary around the pre-defined mean block size. To be more specific, the block size follows a normal distribution with mean bs and standard deviation $0.3{*}bs$. The case $bs = 1$ is special, since we take the standard deviation of the block size to be zero so all genes are independent. The correlation matrix for a block b, has 1's on the diagonal and ρ_b in the off-diagonal entries. We usually allow $\rho \sim Beta(pw, (1-p) * w)$ to follow a beta distribution with parameters $p = 0.6$ and $w = 5$.

We mentioned above that some genes would be transcriptionally inactive under certain biological conditions. Instead of simulating this active status for genes individually, we simulate the whole block of genes being transcriptionally active or inactive. This models the idea that the entire pathway or network could be turned on or off under certain biological conditions.

III. The Multi-hit Model of Cancer

The multiple hit theory of cancer was first proposed by Carl Nordling in 1953 [18] and extended by Alfred

 ISBN: 978-1-61208-137-3

Knudson in 1971 [19]. The basic idea is that cancer can only result after multiple insults (mutations; hits) to the DNA of a cell. We use the combinatorics of multiple hits to simulate heterogeneity in the population. Let H be the number of possible hits (typically on the order of 10 to 20). We define a cancer subtype as a collection of hits (usually 5 or 6 out of those possible). Each subtype has a prevalence; by default, each subtype is equally likely to occur in the population. To simulate a set of patients, we start by assigning them to one of the cancer subtypes (with probabilities equal to the prevalences). We then use the individual hits as (unobserved) latent variables that influence gene expression, survival, and binary outcomes. Specifically, let Z_h be a binary variable that indicates the presence ($Z_h = 1$) or absence ($Z_h = 0$) of a hit h. Then the probability p of an unfavorable (binary) outcome is simulated from a logistic model

$$log\left(\frac{p}{1-p}\right) = \sum_{h=1}^{H} \beta_i Z_i,$$

where the parameters $\beta_i \sim N(0, \sigma_B)$ are simulated from a normal distribution. We simulate survival times from a Cox proportional hazards model, with

$$h(t) = h_0(t) \sum_{h=1}^{H} \alpha_i Z_i,$$

where $h_0(t)$ can be taken to be any desired survival model (usually exponential) and the coefficients $\alpha_i \sim N(0, \sigma_A)$ can be taken to be either independent of or related to the β_i depending on the goal of the simulation. Finally, each hit is assumed to affect the expression of one correlated block of genes (representing the effect on a single biologically pathway) by altering the mean expression of the genes in that block. More elaborate models can also be generated, by altering the variances or the correlation structure within the block.

IV. Simulation Results

To illustate the `Umpire` simulation package, we have simulated a microarray data set with associated survival data. We assumed that there are 20 possible hits, and that 5 hits at a time defined a cancer subtype. For this simulation, we assumed that there were 6 distinct, equally likely, cancer subtypes. As above, each of the 20 hits corresponds to a correlated block of gene expression and also affects survival. We also assumed that there were 100 correlated blocks of genes that were unrelated to cancer or to survival. Blocks were simulated to contain a mean of 100 genes with a standard deviation of 30. Gene means, standard deviations, and correlation structures were simulated using the distributions and hyperparameters described above. We simulated survival by assuming an exponential baseline hazard function.

Table I
Number of significant genes, by sample size and FDR.

	N = 100	N = 300	N = 500
FDR = 0.01	12	86	144
FDR = 0.05	22	135	209
FDR = 0.1	37	169	253
FDR = 0.2	74	249	354
FDR = 0.3	127	346	446

We analyzed the simulated data using an approach that is common in the field. Specifically, we fit gene-by-gene univariate Cox proportional hazards models. We recorded the p values for a log-rank test of the significance of each gene. We then fit a beta-uniform mixture (BUM) model to the set of p-values, and used the BUM model to estimate the false discovery rate (FDR). Table I shows the number of genes called significant as a function of the FDR and the sample size. For an FDR of 20%, Table II separates these results into groups depending on the membership of genes in different correlated blocks. Recall that 20 correlated blocks of genes were associated with cancer-related hits; the blocks of "irrelevant" genes are collected in the row of the table labeled "FP" to denote obvious false positive findings. The first column of Table II shows the number of cancer subtypes (patterns) that included each hit; the second column shows the coefficient of that (latent) hit in the simulated survival model. Note that even though there were 20 possible hits, four of them (G4, G7, G10, and G14) were not actually included in the patterns of 5 hits that defined the 6 cancer subtypes in this simulation. Using 100 samples, we only discovered multiple genes that represented 5 of the cancer-related gene blocks. Using 500 samples, we discovered multiple genes representing all 16 "active" cancer-related gene blocks.

Figure 1 displays heatmaps of the genes selected as significant at the 20% FDR level using either 100 or 500 samples. The color bar along the top reflects the true cancer subtype for each patient. The color bar along the side displays the cancer-related gene block, with false positive genes colored white. When using 100 samples, only two or three of the six cancer subtypes can be seen in the heatmap, and only four of the cancer-related gene blocks. With 500 samples, all six cancer subtypes are visible in the heatmap, along with almost all of the cancer-related gene blocks. In both heatmaps, the false positive genes are recognizable by their lack of correlation with other selected genes.

V. Conclusion

We have described the Umpire simulation package and shown that it can be used to simulate microarray data that is related to survival outcomes in complex ways. An initial simulation using this package suggests, using

 ISBN: 978-1-61208-137-3

Table II
NUMBER OF SIGNIFICANT GENES AS A FUNCTION OF THE SAMPLE SIZE AND THE TRUE HIT STATUS.

	Patterns	Alpha	N = 100	N = 300	N = 500
G1	4	0.291	0	8	10
G2	2	0.366	0	5	11
G3	1	0.090	0	3	11
G4	0	0.278	0	1	0
G5	1	1.428	0	2	2
G6	3	0.313	0	1	2
G7	0	0.496	0	0	0
G8	1	-0.428	1	5	13
G9	3	-2.135	6	34	40
G10	0	0.631	2	1	0
G11	1	0.047	17	38	44
G12	2	0.422	0	13	27
G13	2	1.062	1	7	12
G14	0	1.433	0	2	0
G15	2	2.514	0	6	15
G16	1	-0.384	0	3	3
G17	1	-0.841	1	10	14
G18	2	0.299	0	13	16
G19	2	1.358	10	25	32
G20	2	-1.674	6	35	41
FP	0	0.000	30	37	61

a plausible set of biologically meaningful parameters, that studies to discover signatures that predict time-to-event outcomes may need more than the 100 samples that have frequently been used in practice. More detailed simulation studies will be required to test this idea further.

The results of the simulation also suggest that we may need better methods for combining gene expression values into predictive signatures. First, the common statistical approach that tries to optimize the coefficients of all 354 selected genes using 500 samples is unlikely to succeed. Moreover, since we know "ground truth" for this particular simulation, we know that there are 16 independent factors that influence survival. From the heatmap on the bottom of Figure 1, we would also estimate that there are many distinct expression patterns that contribute to survival. This observation suggests two possible approaches. On the one hand, we could group correlated genes together into simpler factors that can be included in predictive models. For example, we could perform a principal components analysis and use the first few principal components (PCs) as predictors. For our simulated data, a scree plot of the variance explained by each PC suggests that there are approximately five non-random PCs (data not shown). A Cox proportional hazards models identifies all five of those PCs as significant predictors of survival (data not shown). On the other hand, the same heatmap indicates the presence of six subtypes of cancer. An alternative approach would be to use those six subtypes as a categorical predictor; a Cox model successfully identifies these categories as significant predictors (data not shown). In this case, the obvious next step would be to develop a robust multi-category classifier.

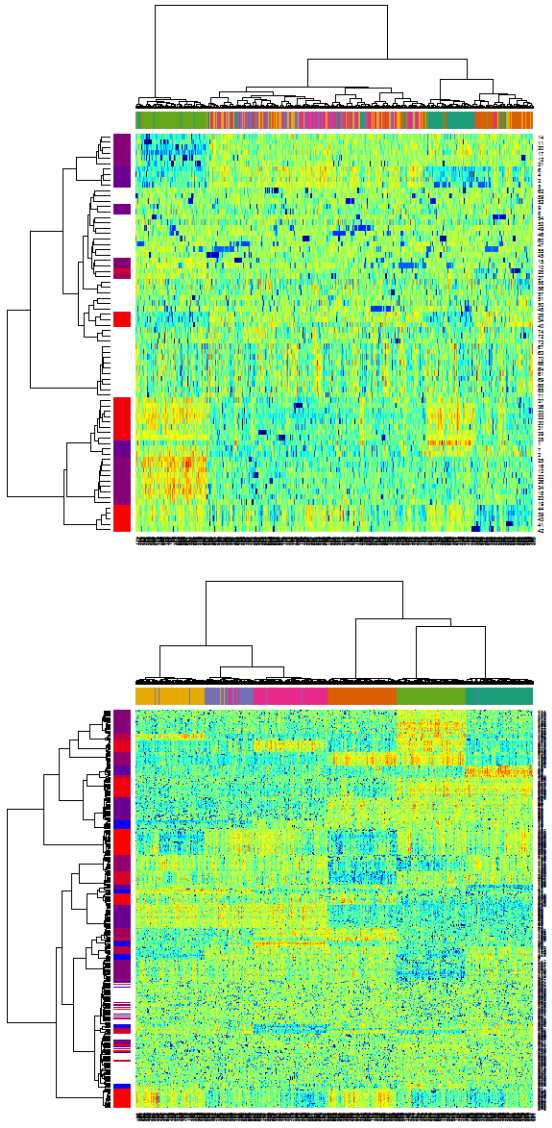


Figure 1. Heatmaps of the significant genes at FDR = 20% using 100 (top) or 500 (bottom) samples.

We do not pursue these approaches in the current paper. However, the Umpire simulation package provides the tools that are necessary to evaluate a range of analytical methods on data sets with different sizes and properties. The availability of this tool should contribute to the development of better methods to learn useful predictors of biologically relevant outcomes.

ACKNOWLEDGMENT

This research was supported by grants P30 CA016672, R01 CA123252, P50 CA070907, and P50 CA140388

 ISBN: 978-1-61208-137-3

from the National Cancer Institute of the United States National Institutes of Health.

This document was prepared using Sweave, a literate programming tool for the R statistical software environment. Complete source code, including all code necessary to run the simulations and generate the figures and tables, is available upon request.

References

[1] R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao, *Design and Analysis of DNA Microarray Investigations*, ser. Statistics for Biology and Health. New York, NY: Springer-Verlag, 2003.

[2] D. B. Allison, X. Cui, G. P. Page, and M. Sabripour, "Microarray data analysis: from disarray to consolidation and consensus," *Nat Rev Genet*, vol. 7, no. 1, pp. 55–65, 2006.

[3] K. K. Dobbin and R. M. Simon, "Sample size planning for developing classifiers using high-dimensional DNA microarray data," *Biostatistics*, vol. 8, no. 1, pp. 101–17, 2007.

[4] K. K. Dobbin, Y. Zhao, and R. M. Simon, "How large a training set is needed to develop a classifier for microarray data?" *Clin Cancer Res*, vol. 14, no. 1, pp. 108–14, 2008.

[5] C. F. Aliferis, A. Statnikov, I. Tsamardinos, J. S. Schildcrout, B. E. Shepherd, and F. E. Harrell Jr., "Factors influencing the statistical power of complex data analysis protocols for molecular signature development from microarray data," *PLoS One*, vol. 4, no. 3, p. e4922, 2009.

[6] C. K. Wierling, M. Steinfath, T. Elge, S. Schulze-Kremer, P. Aanstad, M. Clark, H. Lehrach, and R. Herwig, "Simulation of DNA array hybridization experiments and evaluation of critical parameters during subsequent image and data analysis," *BMC Bioinformatics*, vol. 3, p. 29, 2002.

[7] Y. Balagurunathan, E. R. Dougherty, Y. Chen, M. L. Bittner, and J. M. Trent, "Simulation of cDNA microarrays via a parameterized random signal model," *J Biomed Opt*, vol. 7, no. 3, pp. 507–23, 2002.

[8] D. S. Lalush, "Characterization, modeling, and simulation of mouse microarray data," in *Methods of Microarray Data Analysis III*, S. M. Lin and K. F. Johnson, Eds. Boston: Kluwer Academic Publishers, 2003, pp. 75–92.

[9] M. Nykter, T. Aho, M. Ahdesmaki, P. Ruusuvuori, A. Lehmussola, and O. Yli-Harja, "Simulation of microarray data with realistic characteristics," *BMC Bioinformatics*, vol. 7, p. 349, 2006.

[10] C. J. Albers, R. C. Jansen, J. Kok, O. P. Kuipers, and S. A. van Hijum, "Simage: simulation of DNA-microarray gene expression data," *BMC Bioinformatics*, vol. 7, p. 205, 2006.

[11] K. Dobbin and R. Simon, "Comparison of microarray designs for class comparison and class discovery," *Bioinformatics*, vol. 18, no. 11, pp. 1438–45, 2002.

[12] A. Szabo, K. Boucher, W. L. Carroll, L. B. Klebanov, A. D. Tsodikov, and A. Y. Yakovlev, "Variable selection and pattern recognition with gene expression data generated by the microarray technology," *Math Biosci*, vol. 176, no. 1, pp. 71–98, 2002.

[13] I. Lonnstedt and T. Speed, "Replicated microarray data," *Statistica Sinica*, vol. 12, pp. 31–46, 2002.

[14] M. S. Pepe, G. Longton, G. L. Anderson, and M. Schummer, "Selecting differentially expressed genes from microarray experiments," *Biometrics*, vol. 59, no. 1, pp. 133–42, 2003.

[15] P. de Valpine, H. M. Bitter, M. P. Brown, and J. Heller, "A simulation-approximation approach to sample size planning for high-dimensional classification studies," *Biostatistics*, vol. 10, no. 3, pp. 424–35, 2009.

[16] R. S. Parrish, H. J. Spencer III, and P. Xu, "Distribution modeling and simulation of gene expression data," *Computational Statistics and Data Analysis*, vol. 53, pp. 1650–1660, 2009.

[17] Y. Guo, A. Graber, R. N. McBurney, and R. Balasubramanian, "Sample size and statistical power considerations in high-dimensionality data settings: a comparative study of classification algorithms," *BMC Bioinformatics*, vol. 11, p. 447, 2010.

[18] C. O. Nordling, "A new theory on cancer-inducing mechanism," *Br J Cancer*, vol. 7, no. 1, pp. 68–72, 1953.

[19] J. Knudson, A. G., "Mutation and cancer: statistical study of retinoblastoma," *Proc Natl Acad Sci U S A*, vol. 68, no. 4, pp. 820–3, 1971.

 ISBN: 978-1-61208-137-3