BIOTECHNO 2015

The Seventh International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-409-1

May 24 - 29, 2015

Rome, Italy

BIOTECHNO 2015 Editors

Alexey Cheptsov, High Performance Computing Center Stuttgart, Germany

Hesham H. Ali, University of Nebraska at Omaha, USA

BIOTECHNO 2015

Forward

The Seventh International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2015), held between May 24-29, 2015 in Rome, Italy, continued a series of events covering topics related to 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 aspects concerning biocomputation technologies, bioinformatics mechanisms dealing with chemoinformatics, bioimaging, and neuroinformatics.

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.

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, bionanotechnologies, drugs, or microbial propagation and immunity.

The conference had the following tracks:

- Microbiology

- Bioinformatics
- Bio-medical technologies

The conference also featured the following symposium:

- **BIOCOMPUTATION 2015, *The International Symposium on Big Data and BioComputation***

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2015 technical program committee, as well as the numerous reviewers. The creation of such a 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 effort to contribute to BIOTECHNO 2015. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations and sponsors. We also gratefully thank the members of the BIOTECHNO 2015 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope BIOTECHNO 2015 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of bioinformatics, biocomputational systems and biotechnologies. We also hope that Rome, Italy provided a pleasant environment during the conference and everyone saved some time to enjoy the historic beauty of the city.

BIOTECHNO 2015 Chairs

BIOTECHNO Advisory Chairs

Stephen Anthony, The University of New South Wales, Australia
Petre Dini, Concordia University, Canada / China Space Agency Center-Beijing, China
Hesham H. Ali, University of Nebraska at Omaha, USA
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada

BIOTECHNO Industrial/Research Chairs

Yili Chen, Monsanto Company - St. Louis, USA
Attila Kertesz-Farkas, University of Washington, USA
Igor V. Maslov, EvoCo Inc. - Tokyo, Japan
Tom Bersano, Google, USA
Clara Pizzuti, ICAR-CNR - Rende (Cosenza), Italy
John Spounge, National Center for Biotechnology Information /National Library of Medicine - Bethesda, USA

BIOCOMPUTATION 2015 Advisory Committee

Hesham H. Ali, University of Nebraska at Omaha, USA
Bing Wang, Tongji University, China
Alexey Cheptsov, High Performance Computing Center Stuttgart, Germany

BIOTECHNO 2015

Committee

BIOTECHNO 2015 Advisory Chairs

Stephen Anthony, The University of New South Wales, Australia
Petre Dini, Concordia University, Canada / China Space Agency Center-Beijing, China
Hesham H. Ali, University of Nebraska at Omaha, USA
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada

BIOTECHNO 2015 Industrial/Research Chairs

Yili Chen, Monsanto Company - St. Louis, USA
Attila Kertesz-Farkas, University of Washington, USA
Igor V. Maslov, EvoCo Inc. - Tokyo, Japan
Tom Bersano, Google, USA
Clara Pizzuti, ICAR-CNR - Rende (Cosenza), Italy
John Spounge, National Center for Biotechnology Information /National Library of Medicine - Bethesda, USA

BIOTECHNO 2015 Technical Program Committee

Basim Alhadidi, Albalqa' Applied University - Salt, Jordan
Hesham H. Ali, University of Nebraska at Omaha, USA
Jens Allmer, Izmir Institute of Technology, Turkey
Stephen Anthony, The University of New South Wales, Australia
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada
Siegfried Benkner, University of Vienna, Austria
Gilles Bernot, University of Nice Sophia Antipolis, France
Tom Bersano, University of Michigan, USA
Christian Blum, IKERBASQUE, Basque Foundation for Science - Bilbao, Spain
Razvan Bocu, University of Brasov, Romania
Magnus Bordewich, Durham University, UK
Sabin-Corneliu Buraga, "A. I. Cuza" University - Iasi, Romania
Eduardo Campos dos Santos, Universidade Federal de Minas Gerais (UFMG), Brazil
Yang Cao, Virginia Tech – Blacksburg, USA
Cesar German Castellanos Dominguez, Universidad Nacional de Colombia - Manizales,Colombia
Yili Chen, Monsanto Company - St. Louis, USA
Rolf Drechsler, DFKI Bremen || University of Bremen, Germany
Esmaeil Ebrahimie, University of Adelaide, Australia

- CBBC, Colombia
J. Cristian Salgado, University of Chile, Chile
Luciano Sanchez, Universidad de Oviedo, Spain
Steffen Schober, Ulm University, Germany
Sylvain Sené, Aix-Marseille University, France
Avinash Shankaranarayanan, Aries Greenergie Enterprise (P), Ltd., India
Patrick Siarry, Université Paris 12 (LiSSi), France
Anne Siegel, CNRS - Rennes, France
Raj Singh, University of Houston, USA
Zdenek Smékal, Brno University of Technology, Czech Republic
Bin Song, Oracle - Redwood shores, USA
Ondrej Strnad, Masaryk University, Czech Republic
Andrzej Swierniak, Silesian University of Technology, Poland
Sing-Hoi Sze, Texas A&M University, USA
Yoshihiro Taguchi, Chuo University, Japan
Sophia Tsoka, King's College London, UK
Marcel Turcotte, University of Ottawa, Canada
Ugo Vaccaro, Universita` di Salerno, Italy
Chun Wu, Mount Marty College - Yankton, USA
Boting Yang, University of Regina, Canada
Wang Yu-Ping, Tulane University, USA
Alexander Zelikovsky, Georgia State University, USA
Erliang Zeng, University of Notre Dame, USA
Chunchao Zhang, The University of Texas MD Anderson Cancer Center -Houston, USA

BIOCOMPUTATION 2015 Advisory Committee

Hesham H. Ali, University of Nebraska at Omaha, USA
Bing Wang, Tongji University, China
Alexey Cheptsov, High Performance Computing Center Stuttgart, Germany

BIOCOMPUTATION 2015 Program Committee Members

Hesham H. Ali, University of Nebraska at Omaha, USA
Khalid Belhajjame, Paris Dauphine University, France
Zhiwei Cao, Tongji University, China
John Carlis, University of Minnesota, USA
Alexey Cheptsov, High Performance Computing Center Stuttgart, Germany
Sung-Bae Cho, Yonsei University, South Korea
Matthias Chung, Virginia Tech, USA
Raffaele A. Calogero, University of Torino, Italy
Angelo Facchiano, Istituto di Scienze dell'Alimentazione - CNR, Italy
Fabio Fumarola, University of Bari "Aldo Moro", Italy
Saman K. Halgamuge, University of Melbourne, Australia

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

Overcoming Ambiguous Gene Name Synonyms in MEDLINE Searches by Context Mining

Word-vector based text classification of PubMed records

Modest von Korff, Thomas Sander
Research Information Management
Actelion Pharmaceuticals Ltd., Allschwil, Switzerland
Email: modest.korff@actelion.com, thomas.sander@actelion.com

***Abstract*—Classification of ambiguous gene name synonyms is a necessity when mining PubMed Central records with gene-related queries. This work introduces the use of word-vectors for gene name disambiguation. PubMed Central was queried for gene names and their synonyms. The retrieved records were filtered and automatically separated into train- and test-data. A similarity threshold was derived from the similarity matrix of every training word-vector set. The classification performance of the word-vectors was compared to a gene name similarity classification. Both methods showed good results, but the word-vector classification was superior in terms of precision and recall.**

Keywords-Gene name disambiguation; classification; word-vectors; datamining; algorithm.

I. INTRODUCTION

Searching MEDLINE for information about genes is a common task. Retrieving results that are not related to the gene under consideration is a common experience. One reason is the existence of ambiguous gene name synonyms. Many gene name synonyms are shared by two or more genes. This means that a PubMed search for an ambiguous gene name synonym will retrieve the publications for at least two genes. If the search is performed by a scientist, he will be burdened by the additional workload to sort out the unwanted publications. Even worse, a data-mining tool, without the capability of recognizing the ambiguity, will confound the information for the gene under consideration with the information from the other gene. Problems with ambiguous gene names were already reported by Jenssen et al. [1], and were also the topic of the BioCreative 1 and 2 challenges [2] [3]. Hakenberg et al. [4] and Wermter et al. [5] undertook huge efforts to normalize gene names. More recent approaches were published by Neves et al. [6] and by Li et al. [7]. The work presented here demonstrates a solution for the gene name disambiguation problem as it was described by Li et al. and Hakenberg et al. [8]. Our method solves the issue of ambiguous gene name synonyms by context similarity classification. In Section II the applied methods and the datasets are described. Section III gives a summary of the results for the classification of ambiguous PubMed records. The conclusions for the experiments and their results are given in Section IV.

II. METHODS

A. Gene names and synonyms

Gene names and their synonyms were the starting point for our shared synonyms experiments. Two sources were used to retrieve the synonyms. A table with Human Genome Organization (HUGO) ids, gene names, approved symbols and synonyms was retrieved from HUGO Gene Nomenclature Committee (HGNC) [9]. The second source was the MEDLINE database Entrez Gene [10], which also delivered HUGO ids, gene names, and synonyms. Both these databases were used because they do not completely overlap. The combined database, **LG,** is a list of records for each gene. A single record $\mathbf{lg}_{\mathrm{Gene}}$ from this list contains the approved symbol as the approved name and a list with all synonyms from the two data sources. A scheme for the complete algorithm is given in Fig. 1.

B. Ambiguous synonyms detection

The algorithm for detecting ambiguous synonyms consists of two parts. For the detection of ambiguous approved symbols, an approved symbol $as_{\mathrm{Gene,query}}$ from a gene record $\mathbf{lg}_{\mathrm{Gene,query}}$ is taken and compared to the synonyms from all other records in **LG**. This is done for every approved symbol in **LG**. If the approved symbol $as_{\mathrm{Gene,query}}$ matches a synonym, an ambiguous approved symbol is found. Detecting ambiguous synonyms works analogously. From a record $\mathbf{lg}_{\mathrm{Gene,query}}$, a synonym s_{query} is taken and compared to all other synonyms in **LG**. If s_{query} matches any other synonym, an ambiguous synonym has been found. This is done for every synonym in **LG**. If a record $\mathbf{lg}_{\mathrm{Gene,query}}$ contains an ambiguous approved symbol, ambiguous approved name or an ambiguous synonym the gene record receives the label *ambiguous*.

C. Querying PubMed Central with gene name synonyms

For all ambiguous records from **LG**, queries are generated to search the PubMed Central database. One PubMed query is created for every single approved symbol, approved name or synonym. Without any further specification, all fields in the PubMed Central database are searched. Depending on the query, no records at all up to several tens of thousands are retrieved.

 ISBN: 978-1-61208-409-1

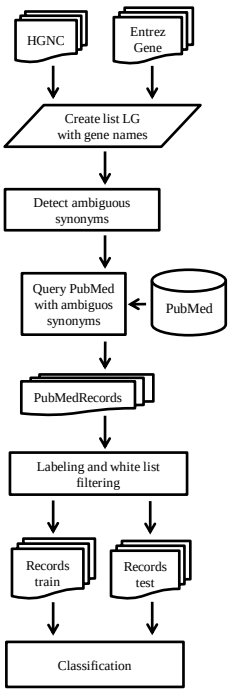

Figure 1: Architecture of the gene name disambiguation.

The result is a dataset, $\mathbf{R}_{\text{Gene}}$, for each gene, containing the retrieved records. If a PubMed record contains an unambiguous approved name or an unambiguous approved symbol it receives the label *train*. If the PubMed record contains only ambiguous gene name information, approved symbol, approved name or synonym, it receives the label *ambiguous*. PubMed queries with the Entrez tool did not distinguish between lower-case and upper-case letters. Unfortunately, many letter combinations exist which differ in capitalization and are shared by different terms. Consequently, up to tens of thousands of false-positive records were retrieved for a single gene.

D. Whitelist filtering of PubMed records

A post-processing step was added to get rid of the false-positive records. If a synonym consisted of less than six characters and did not contain a space, the retrieved PubMed records were filtered for the exact upper- and lower-case pattern of the synonym. However, after this filtering process, many false-positive records still remained. These records contained terms with an identical synonym to the gene under consideration. False-positive records that contain the exact synonym can only be detected by analysing the context of the synonym. The context of the synonyms we were looking for was related to the concept 'gene'. For the record filter in G2DPubMedMiner, a gene context list of 25 terms was defined: activation, activator, allosteric, chromatin, chromosome, codon, exon, expression, gene, genome, genotype, histone, homolog, inhibitor, inhibition, intron, modulator, mutant, nucleosome, peptide, phenotype, phenotypic, polymerase, protein, target, transcript, and transposon. If a PubMed record did not contain any of these words, it was very unlikely that the record was related to a gene. Consequently, a PubMed record was only accepted if it contained at least one of the words from the context list. Furthermore all records were skipped that did not contain a disease MeSH term.

E. Test dataset with ambiguous gene records

From the list with the ambiguous genes a test dataset with eleven pairs of ambiguous gene records was selected at random (Table 1). A gene test set record $\mathbf{gtr}_{\text{Gene1,Gene2,Synonym}}$ contained two approved symbols and the ambiguous synonym they shared. For each gene test set record the corresponding train- and test-sets from PubMed Central were compiled. The training set contained all PubMed records where the approved gene name or the approved symbol was found. The test set contained the records with the ambiguous gene name synonym. All test records were manually classified and received the label *genename1*, *genename2* or *none*. None was given if the text in the PubMed record summary indicated that the gene name synonym referred to neither of the two genes.

TABLE I. TRAIN AND TEST DATA SETS.

Approved symbols		**Shared synonym**	**Number of records in data sets**		
Gene 1	***Gene 2***		***Train 1***	***Train 2***	***Test***
ANPEP	TOR1AIP1	LAP1	24	6	72
APEX1	TEAD1	REF1	84	38	18
APEX1	TFPI2	REF1	84	59	7
CCNL2	FAM58A	cyclin M	11	4	3
CD200R1	HCRTR2	OX2R	31	21	96
CNGB1	LRRC32	GARP	29	20	101
DPYSL2	SDF2L1	dihydropyrimidinase-like 2	32	5	10
ERCC3	GTF2H1	TFIIH	90	9	51
HSD17B7	SKAP2	PRAP	22	5	22
MECOM	RUNX1	AML1-EVI-1	14	90	1
POU2F1	SLC22A1	OCT1	23	90	37
			444	347	418

F. Classification of ambiguous records

Two methods were used for the classification of the test records. A simple gene name similarity search was used as standard method. Word-vectors were used as a second classification method. A word-vector encodes a text as an integer vector. Every field in the vector corresponds to one word, and the field value is equal to the frequency count of the word in the text. Two word-vectors are compared by calculating their similarity coefficient. The method was adapted from Lewis et al. [11]. Because of their results, we decided to use the cosine similarity together with inverse-document-frequency (IDF) weighting. We changed only their formula for the similarity calculation by multiplying x^2

 ISBN: 978-1-61208-409-1

and y^2 with IDF_i (Eq. 1). Consequently, the similarity is scaled between zero and one:

$$\text{Cosine coefficient}(x) = \frac{\sum_{i=1}^{n}(x_i y_i \text{IDF}_i)}{\sum_{i=1}^{n} x^2 \text{IDF}_i \times \sum_{i=1}^{n} y^2 \text{IDF}_i}. \quad (1)$$

For a train data set, the complete similarity matrix was calculated. This means that all pair-wise similarities were calculated between the word-vectors that were compiled from the PubMed records for one gene. The similarity values were sorted and the value at a given percentile of the sorted vector was taken as threshold value. The classification of the test data was done for different percentile values: 0.75, 9.5, 0.25, 0.05, and 0. A percentile of 0 meant that no threshold was used.

III. RESULTS

A total of 35,631 gene names were extracted from HUGO. The ambiguous synonyms detector found 7166 pairs of genes that shared at least one synonym. From this set of gene pairs, eleven were selected for the test dataset. The processing time for a dataset, including querying PubMed and the consequent processing of the results, strongly depended on the number of retrieved PubMed records and took up to 30 minutes.

The results for the classification experiments are given in Table 2 with precision and recall as figures of merit.

TABLE II. RESULTS FOR THE CLASSIFICATION EXPERIMENTS.

Method	Result		
	Precision	Recall	Harmonic mean
GenenameSim	0.83	0.28	0.42
WVSim 0	0.52	1	0.68
WVSim 0.05	0.63	0.87	0.73
WVSim 0.25	0.68	0.57	0.62
WVSim 0.5	0.88	0.29	0.44
WVSim 0.75	0.91	0.19	0.31

In the last column of the table the harmonic mean combines precision and recall. For the simple approach with the gene name similarity classification a precision of 0.83 and a recall of 0.28 was reached. The next five rows show the results for the classification using word-vectors and the five different threshold percentiles. The maximum harmonic mean was reached for a threshold of 0.05 (WVSim 0.05). To compare our results with other approaches like those of Li at al. [7], or Xu at al. [12] et al. is difficult, because gene name normalization and disambiguation are often done together. Or, supervised methods are used, with the disadvantage of being successful only in the training domain.

IV. CONCLUSIONS

Identification of more than 7,000 gene pairs sharing at least one synonym demonstrated that the classification of ambiguous gene names is a worthwhile undertaking. With a test data set, compiled from eleven pairs of ambiguous gene names, it was shown that word-vector classification reduced the ambiguity significantly. A similarity threshold value, which was automatically derived from the similarity matrix of the training data, increased the precision of the classification results. The entire process, starting with querying PubMed Central, followed by filtering and train- and test-set generation, and the classification is unsupervised and can be fully automated. Word-vector classification for gene name disambiguation is a valuable addition to every data-mining tool working on PubMed records with gene-related queries.

REFERENCES

[1] T. K. Jenssen, A. Laegreid, J. Komorowski and E. Hovig, "A literature network of human genes for high-throughput analysis of gene expression," Nature Genetics, vol. 28, 2001, pp. 21-28, doi: 10.1038/ng0501-21.

[2] L. Hirschman, M. Colosimo, A. Morgan and A. Yeh, "Overview of BioCreAtIvE task 1B: normalized gene lists," BMC Bioinformatics, vol. 6 Suppl 1, 2005, pp. S11.11-S11.10, doi: 10.1186/1471-2105-6-S1-S11.

[3] A. A. Morgan et al., "Overview of BioCreative II gene normalization," Genome Biology, vol. 9 Suppl 2, 2008, pp. S3.1-S3.19, doi: 10.1186/gb-2008-9-s2-s3.

[4] J. Hakenberg et al., "Gene mention normalization and interaction extraction with context models and sentence motifs," Genome Biology, vol. 9 (Suppl 2) , S14, 2008, doi: 10.1186/gb-2008-9-S2-S1.

[5] J. Wermter, K. Tomanek and U. Hahn, "High-performance gene name normalization with GeNo," Bioinformatics, vol. 25, 2009, pp. 815-821, doi: 10.1093/bioinformatics/btp071.

[6] M. L. Neves, J. M. Carazo and A. Pascual-Montano, "Moara: a Java library for extracting and normalizing gene and protein mentions," BMC Bioinformatics, vol. 11, 2010, pp. 157, doi: 10.1186/1471-2105-11-157.

[7] L. Li, S. Liu, W. Fan, D. Huang and H. Zhou, "A multistage gene normalization system integrating multiple effective methods," PLoS One, vol. 8, 2013, pp. e81956, doi: 10.1371/journal.pone.0081956.

[8] J. Hakenberg, C. Plake, R. Leaman, M. Schroeder and G. Gonzalez, "Inter-species normalization of gene mentions with GNAT," Bioinformatics, vol. 24, 2008, pp. i126-132, doi: 10.1093/bioinformatics/btn299.

[9] HUGO Gene Nomenclature Committee at the European Bioinformatics Institute. [retrieved: Mar. 2015]. Available from: http://www.genenames.org

[10] D. Maglott, J. Ostell, K. D. Pruitt and T. Tatusova, "Entrez Gene: gene-centered information at NCBI," Nucleic Acids Res, vol. 33, 2005, pp. D54-58, doi: 10.1093/nar/gki031.

[11] J. Lewis, S. Ossowski, J. Hicks, M. Errami and H. R. Garner, "Text similarity: an alternative way to search MEDLINE," Bioinformatics, vol. 22, 2006, pp. 2298-2304, doi: 10.1093/bioinformatics/btl388.

[12] H. Xu et al., "Gene symbol disambiguation using knowledge-based profiles," Bioinformatics, vol. 23, 2007, pp. 1015-1022, doi: 10.1093/bioinformatics/btm056.

 ISBN: 978-1-61208-409-1

Tabu Search Algorithm for RNA Degradation Problem

Agnieszka Rybarczyk, Marta Kasprzak, Jacek Blazewicz
Institute of Computing Science
Poznan University of Technology
Piotrowo 2, 60-965 Poznan, Poland
and
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Noskowskiego 12/14, 61-704 Poznan, Poland
Email: arybarczyk@cs.put.poznan.pl
Email: mkasprzak@cs.put.poznan.pl
Email: jblazewicz@cs.put.poznan.pl

***Abstract*—In the last few years, there has been a great interest in the RNA (ribonucleic acid) research due to the discovery of the role that RNA molecules play in biological systems. They do not only serve as a template in protein synthesis or as adaptors in translation process but also influence and are involved in the regulation of gene expression. It was demonstrated that most of them are produced from larger molecules due to enzyme cleavage or spontaneous degradation. In this work, we would like to present our recent results concerning the RNA degradation process. In our studies, we used artificial RNA molecules designed according to the rules of degradation developed by Kierzek and co-workers. On the basis of the results of their degradation, we have proposed the formulation of the RNA Partial Degradation Problem (RNA PDP) and we have shown that the problem is strongly NP-complete. We would like to propose a new efficient heuristic algorithm based on tabu search approach which allows us to reconstruct the cleavage sites of the given RNA molecule.**

Keywords–RNA degradation; tabu search; computational complexity.

I. INTRODUCTION

For the last few years, there has been an increased interest in RNA because of its involvement in controlling gene expression. The sequencing projects and analysis of higher eukaryotic genomes have revealed that in contrast to prior expectations, only a small fraction of the genetic material codes for proteins. It has been demonstrated that the vast majority of genomes of complex organisms are transcribed into an abundance of non-protein coding RNA molecules that perform not only housekeeping but also regulatory functions in the cell. It became evident that not only large RNAs (e.g., messenger (mRNA) or transfer RNA (tRNA)) are responsible for proper functioning of living organisms. There exist plenty of smaller RNAs, called small regulatory RNA, which are involved in the cellular processes. In contrast to housekeeping RNAs (like tRNA, rRNA (ribosomal RNA), tmRNA (transfer-messenger RNA), snRNA (small nuclear RNA)), which are constitutively expressed, regulatory RNAs exhibit the expression pattern dependent on organism's developmental stage, cell differentiation or the interaction with the external environment [1].

Small RNAs are 19-28 nucleotide sequences, which are produced through enzyme cleavage or spontaneous degradation of large RNA molecules. The nature of the latter mechanism is now heavily investigated by the researchers [2].

Numerous studies demonstrated that all components of the degradation machinery affect the RNA molecules in the same manner, but some RNAs apparently are more stable than others e.g., in mammalian cells c-fos mRNA lasts 15 minutes while β–globuline over 24 hours [3]. It has been noted that there exist many sequence elements, which can influence the mRNA stability. Those sequences act through stimulating or inhibiting the degradation of mRNA and, e.g., in mammalian cells were identified as rich in adenosine and uridine elements (AURE). The mRNA stability is strongly dependent on the location of AURE elements within the molecule. It is suggested that the spatial structure combined with the sequence motifs decides that the degradation process performs differently in case of distinct RNAs.

In this paper, we would like to report our recent results concerning the degradation of RNA molecules. In our studies we have used artificial RNA molecules, which were designed in such a way that they should be unstable, according to the rules developed by Kierzek and co-workers [4]. On the basis of the results of their degradation, we have proposed a formulation of a new problem, called RNA PDP and we have shown that the problem is strongly NP-complete [5]. Here, we would like to propose a new efficient heuristic algorithm based on tabu search approach that allows to reconstruct the cleavage sites of the given RNA molecule.

The organization of the paper is as follows. In Section 2, the combinatorial model of the decision version of RNA PDP problem is presented. Section 3 introduces the tabu search algorithm for RNA PDP problem. In Section 4, the results of the computational tests are given, while Section 5 points out the directions for further research.

II. PROBLEM FORMULATION

To analyze the degradation process dependent on RNA structure, we carried out two types of experiments [5]: involving multi-labeled and single-labeled RNA. The aim of the first type of experiment was to visualize all fragments generated during degradation. In this case, the exact location of the fragments within analyzed RNA molecule is missing. The second type of experiment was carried out to visualize only those fraction of degradation products, which contained labeled 5' end of the RNA molecule. In this case, their location within the tested molecule is known. In this way, two collections of fragments are created. Each fragment is represented by its

 ISBN: 978-1-61208-409-1

length.

We will denote as $D = \{d_1, \ldots, d_k\}$ the multiset of fragments (lengths) obtained during the multi-labeled RNA degradation and as $Z = \{z_1, \ldots, z_n\}$ the set of fragments (lengths) coming from the single-labeled RNA degradation. Moreover, we will distinguish between two types of cleavage sites: primary and secondary. Primary cleavage sites occur only within input RNA molecule of the full length while secondary cleavage sites only within lengths obtained as a result of primary cleavages. The computational phase of the reconstruction of the cleavage sites of the given RNA molecule is strongly NP–hard [5]. Hence, there is a need for developing an efficient heuristic.

The mathematical formulation of the RNA PDP is presented below [5]. P_1 stands for the set of primary cleavage sites in the solution and P_2 for the set of secondary ones.

Problem 1: RNA PDP — decision version (Π_{RNAPDP}) .
Instance: Multiset $D = \{d_1, ..., d_k\}$ and set $Z = \{z_1, ..., z_n\}$ of positive integers, positive integer L, constant $C \in \mathbb{Z}^+ \cup \{0\}$.
Question: Do there exist sets P_1 and P_2 such that:

$$P_1 \cup P_2 = P = \{p_1, ..., p_m\}, \ \forall_{p_i \in P} \ 0 < p_i < L, \tag{1}$$

$$D \subseteq D', \ D' \supseteq R = \{p_i - p_j : \ p_i, p_j \in P_1 \cup \{0, L\} \ \wedge \ p_i > p_j\}, \tag{2}$$

$$D' \setminus R = \bigcup_{i=1}^{|T|} D'_i, \tag{3}$$

$$T = \{t_i = (p_a, p_b, p_c) : \ p_a, p_c \in P_1 \cup \{0, L\} \ \wedge \ p_b \in P_2 \ \wedge \ p_a < p_b < p_c \ \wedge \ d'_{i1} = p_b - p_a \ \wedge \ d'_{i2} = p_c - p_b \ \wedge \ \{d'_{i1}, d'_{i2}\} = D'_i\}, \tag{4}$$

$$\forall_{t_i, t_j \in T, \ t_i = (p_{ia}, p_{ib}, p_{ic}), \ t_j = (p_{ja}, p_{jb}, p_{jc})} \quad i \neq j \rightarrow \{p_{ia}, p_{ic}\} \neq \{p_{ja}, p_{jc}\}, \tag{5}$$

$$Z \subseteq Z', \ Z' \subseteq P \cup \{L\}, \ Z' \supseteq P_1 \cup \{L\}, \tag{6}$$

$$Z' \setminus [P_1 \cup \{L\}] = \{p_b : \ (p_a, p_b, p_c) \in T \ \wedge \ p_a = 0\}, \tag{7}$$

$$P_2 = \{p_b : \ (p_a, p_b, p_c) \in T\}, \tag{8}$$

$$|D'| + |Z'| \leq k + n + C \ ? \tag{9}$$

III. Tabu search algorithm

The heuristic algorithm that works for the case of RNAPDP problem with negative experimental errors (i.e., missing fragments in D and Z) presented in this section is based on tabu search metaheuristic, which is a kind of local search procedure. Its aim is to find the coordinates of primary and secondary cleavage sites in a given RNA molecule, taking negative errors into account. The algorithm is implemented in C programming language and runs in a Unix environment. The algorithm takes as an input the data containing fragment lengths obtained via the biochemical experiments, i.e., multiset D of k positive integers and set Z of n positive integers. The main algorithm consists of two parts: inner and outer tabu search method. The outer tabu search method was designed to find the coordinates of primary cleavage sites including negative errors. In this part, four kinds of moves are initially defined: add missing primary fragment to set Z, add missing secondary fragment to set D, prevent from considering element of Z as a primary cleavage site, consider element of Z as a primary cleavage site. The inner tabu search of the algorithm was designed to reconstruct the coordinates of the secondary cleavage sites including negative errors, basing on the results of the inner part. After performing a number of moves not leading to an improvement of the solution quality, the method is restarted, i.e., some randomly generated feasible solution becomes an initial solution. If a specified number of restarts is performed than the algorithm stops. By analyzing the obtained results, we noticed that the algorithm performs very efficiently and fast.

IV. Tabu search algorithm

In this section, results of the tests of the algorithm solving the RNA PDP problem in the case of erroneous data are presented. The algorithm has been tested on PC with Pentium(R) 4, 2.40 GHz processor and 1 GB RAM in Unix environment. As a testing set, a group of randomly generated data was prepared (See [5] for details). Table I summarizes exemplary average running time results for the proposed tabu search algorithm tested on the random instances. In the results, each entry corresponds to 100 instances of the same number of secondary and primary cleavage sites and number of negative errors, i.e., to 100 runs of the algorithm.

TABLE I. Average computational times for randomly generated erroneous instances. Secondary cleavage sites occur in 75% of all primary fragments, the number of the latter being equal to $\binom{r+2}{2}$, where r is the number of primary cleavage sites in the instance. Additionally, the input data set was separately tested with the number of missing fragment lengths ranging from 1 to 5.

No. of reconstructed primary cleavage sites	No. of reconstructed secondary cleavage sites	Average computational time [s] Number of negative errors				
		1	2	3	4	5
4	10	0.05	0.06	0.05	0.05	0.04
5	15	0.40	0.54	0.44	0.41	0.41
6	20	0.30	0.33	0.34	0.33	0.31
7	26	0.82	0.93	0.77	0.85	0.90
8	33	1.18	1.34	1.37	1.35	1.34
10	49	4.71	5.26	5.40	5.48	5.43
12	68	16.44	19.21	19.72	20.12	20.00
14	90	57.13	61.23	63.99	64.95	64.82

Tabu search algorithm was also tested on 3630 randomly generated instances without negative errors and was able to find optimal solution in 3578 cases (98.57% of the whole data).

V. Conclusion

The computational phase of the reconstruction of the cleavage sites in RNA PDP is a computationally hard problem if there are errors in the input data. Hence, the need of developing good polynomial time heuristics arises. In this work, an algorithm based on tabu search approach has been presented and its effectiveness has been tested. The computational tests confirmed high efficiency of the proposed algorithm. This algorithm may be very useful in practice as a tool that facilitates the analysis of the output of the biochemical experiment.

References

[1] M. Szymanski, M. Barciszewska, M. Zywicki, and J. Barciszewski, "Noncoding RNA transcripts," Journal of Applied Genetics, vol. 44, 2003, pp. 1–19.

[2] M. Nowacka and et al., "2D-PAGE as an effective method of RNA degradome analysis," Mol Biol Rep, vol. 39, 2011, pp. 139–146.

[3] M. Deutscher, "Degradation of Stable RNA in Bacteria," Journal Biol. Chem., vol. 278, 2003, pp. 45 041–45 044.

[4] R. Kierzek, "Nonenzymatic Cleavage of Oligoribonucleotides," Methods Enzymol., vol. 341, 2001, pp. 657–675.

[5] J. Blazewicz, M. Figlerowicz, M. Kasprzak, M. Nowacka, and A. Rybarczyk, "RNA Partial Degradation Problem: motivation, complexity, algorithm," J Comput Biol, vol. 18, 2011, pp. 821–834.

 ISBN: 978-1-61208-409-1

RNA_Seq Viewer: A Mobile App for Displaying NGS Gene Expression Data

Chien-Yuan Chiu, Yu-Chieh Lin and Wen-chang Lin
Institute of Biomedical Sciences, Academia Sinica
Taipei, Taiwan, R. O. C.
e-mail: jychiu_tw@yahoo.com.tw, joycelin18@gmail.com, wenlin@ibms.sinica.edu.tw

Abstract— **In recent years, Next Generation Sequencing (NGS) technology has dramatically advanced genome researches and biomedical sciences, which generates and accumulates a large amount of biological data rapidly. Nowadays, there is a rapid expansion of tablets and smart phone devices, replacing the traditional personal computers and notebooks commonly used for information retrieval and display. However, there are little bioinformatic related mobile applications developed specifically for NGS data visualization on mobile devices. To demonstrate the feasibility of displaying large NGS data using the mobile devices, we designed and implemented a new iOS App - the RNA_Seq Viewer, for the visualization of The Cancer Genome Atlas (TCGA) gene expression datasets. This iOS App could efficiently display gene expression information systematically over the human chromosome framework. We have processed over 2,500 human cancer patients in nine cancer types retrieved from the TCGA web site. This mobile App could be potentially utilized in future personalized medicine applications by serving as the essential visualization core component to easily access the personal genomic and medical information through the cloud based data warehouses.**

Keywords-bioinformatics; NGS; App; TCGA; RNA-Seq.

I. INTRODUCTION

In recent years, genome sequence information has been accumulated rapidly thanks to the completion of human genome project and the development of Next Generation Sequencing (NGS) platform technologies. Therefore, many bioinformatic databases and tools are generated for the visualization and analysis of massive genomic sequence information [1]. It is not an easy task to efficiently display vast amount of biological annotation information in addition to the genomic sequences and chromosomal organizations using web browsers. There are several renowned web-based bioinformatic tools and databases to accomplish such challenges, including UCSC Genome Browser [2] and Ensembl [3]. However, over the last few years, it is worth noting that tablets and smart phone devices (such as iPad and iphone) are widely popularized and disseminated on the World, replacing the traditional personal computers. These devices utilize the multi-touch technology as the core user interface, and gradually transform the user experience and machine interface design.

The computer software development sector is also heavily affected by this trend, with the rapid expansion of mobile applications (Apps) for mobile devices. Many developers have been concentrated on the App developments in addition to the traditional desktop software development. Few tablet Apps were created to retrieve and display text based genome annotation information [4]. However, there are no bioinformatic mobile applications developed specifically for the visualization of huge genome information as well as the NGS sequence data. Since the widely adaptation of handheld smart devices, there become the common platform for daily information retrieval and exchange. Therefore, we attempt to develop such novel mobile application software for mobile devices and demonstrate the feasibility of visualization of genome information on the mobile devices.

We first focused on the visualization of The Cancer Genome Atlas (TCGA) data. TCGA is a comprehensive and coordinated effort to accelerate the understanding of the molecular basis of cancer through the utilization of large scale sequencing experiments using NGS platforms [5]. We have retrieved NGS RNA-Seq cancer gene expression information from thousands of cancer patients in various cancer types collected in TCGA project. Our RNA_Seq Viewer mobile application is designed to display the RNA-Seq gene expression information systematically for easy interrogation and visualization.

Herein, in Section II, the development environment and data source were introduced. In Section III, the represented screen displays were demonstrated and basic functionalities were briefly described. Finally, we discussed the future improvements of this new App in Section IV.

II. MATERIALS AND METHODS

The development and implement of RNA_Seq Viewer was done using the Apple development software, Xcode 5.02. The programming language used was Object-C and database used was SQLite. All TCGA cancer level-3 RNA-Seq data were retrieved and gene expression RPKM (reads per kilobase per million) values were processed for SQLite tables. We obtained a total of nine different cancer types with over 2,500 cancer samples. The RNA_Seq Viewer App can be obtained freely through Apple iTune App Store. Additional information can be accessed from [6].

III. RESULTS

RNA_Seq Viewer is a mobile application built for iOS to provide biologists a new user experience in interrogating large-scale NGS data intuitively on the iPad or iPhone. Users can further interrogate differentially expressed genes by interrogating differentially expressed regions from the overall chromosome display view and quickly zooming into regions of particular interest.

The RNA_Seq viewer displays gene expression information (RPKM value) on the chromosome level. Initially, an illustration of the 22 autosomes and two sex chromosomes are displayed (Figure 1). The navigational and

 ISBN: 978-1-61208-409-1

functional buttons are arranged on the bottom row. Users should first select the cancer type of interest by clicking the list button on the left. Detailed information includes the TCGA sample ID and sex/age, as well as cancer stage information on the bottom line (Figure 1).

Once selected, the cancer gene expression information of an individual patient is loaded and displayed. Since the matching normal tissues were not always available in TCGA, we used the average expression values from all available normal tissues for that particular cancer type to represent the background expression levels. The gene expression value of cancer tissue is illustrated by a blue line, and the expression value of normal tissue is illustrated by the overlapping gray line in the background. User then can zoom in and out of the screen to interrogate the regions of their interest (Figure 2).

By adjusting the intensity bar on the right of the function row, the user can dim the blue line to a lighter color, which allows users to compare the expression difference between cancer and normal tissues, especially on the overexpressed genes in cancer. Once a region of interest is identified, users can click on the gene(s) of interest to activate the red-pin icon and show detailed gene information and expression values (Figure 2). The gene name/RefSeq ID and expression values are now displayed before the patient's sample information. The RPKM values of tumor (T) and normal (N) tissues are displayed (Figure 2). Users can then click on the NCBI (National Center for Biotechnology Information) button to retrieve the detailed gene information.

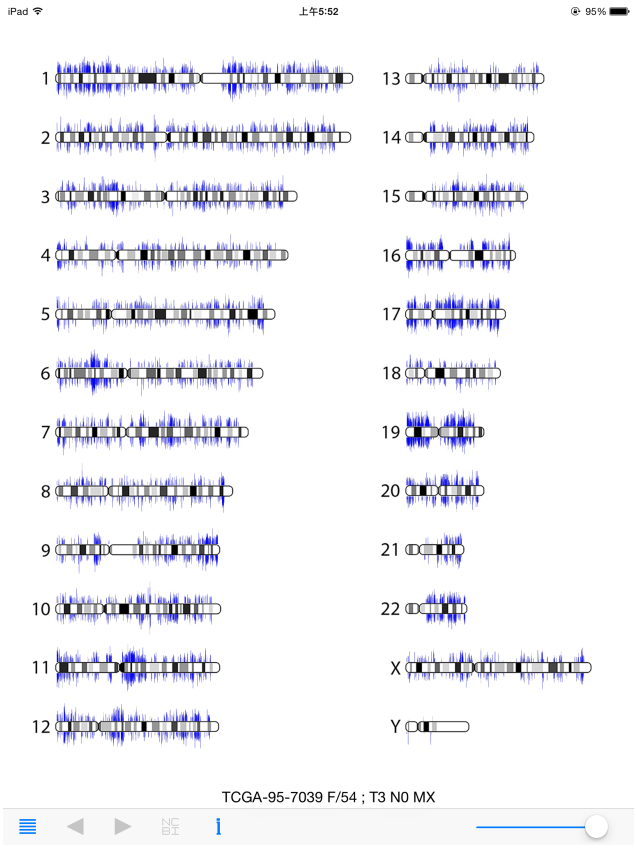

Figure 1. User interface of RNA_Seq Viewer (global view).

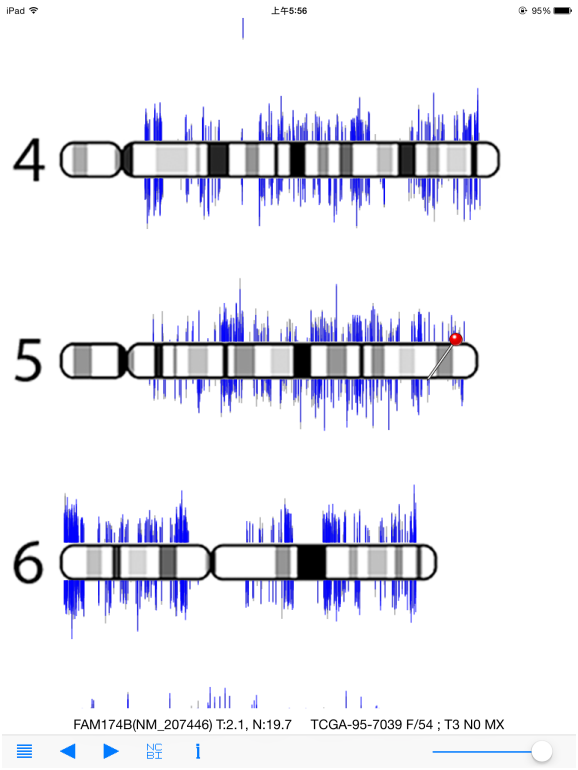

Figure 2. User interface of RNA_Seq Viewer (detailed view).

Users can also submit their very own RNA-Seq data to display in our App here. Users simply obtain the user.agv db file from the web site and fill in the RPKM values of the normal and cancer tissues using any SQLite tools before synchronization to the mobile devices.

IV. DISCUSSION

As proof of concept, we demonstrated that RNA-Seq data could be visualized and interrogated efficiently on the mobile devices. This mobile App is useful for displaying multiple genome data onto global chromosome context. In the future, it can be expanded to integrate additional NGS data and developed for Android platform. This App could be utilized as a fundamental visualization component for personalized genome information display and retrieval.

REFERENCES

[1] S. C. Li, et al., "Identification of homologous microRNAs in 56 animal genomes," Genomics, vol. 96, Jan. 2010, pp. 1-9.

[2] D. Karolchik, et al., "The UCSC Genome Browser database: 2014 update," Nucleic Acids Res., vol. 42, Jan. 2014, pp. D764-D770.

[3] P. Flicek, et al., "Ensembl 2014,", Nucleic Acids Res., vol. 42, Jan. 2013, pp. 749-755.

[4] C. Wu, I. Macleod, and A. I. Su, "BioGPS and MyGene.info: organizing online, gene-centric information," Nucleic Acids Res., vol. 41, Jan. 2013, pp. D561-D565.

[5] Cancer Genome Altas Netwrok, "Comprehensive molecular portraits of human breast tumours," Nature, vol. 490, Oct. 2012, pp. 61-70.

[6] http://tdl.ibms.sinica.edu.tw/NGS_viewer_1/NGS_Viewer/General_Introduction.html.

 ISBN: 978-1-61208-409-1

Cluster Based Analysis Of Petri Net Properties

Marcin Radom
Institute of Computing Science
Poznan University of Technology
Poznań, Poland
mradom@cs.put.poznan.pl

Agnieszka Rybarczyk
Institute of Computing Science
Poznan University of Technology
and
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Poznań, Poland
arybarczyk@cs.put.poznan.pl

Piotr Formanowicz
Institute of Computing Science
Poznan University of Technology
and
Institute of Bioorganic Chemistry
Polish Academy of Sciences
Poznań, Poland
piotr@cs.put.poznan.pl

***Abstract*— Systems biology is an interdisciplinary field of study based mainly on biology, mathematics and computer science. In systems biology, the analysis of complex biological systems of interacting objects is being made. Before the analysis starts, a model is being created, representing the analysed system. The theory of Petri nets offers a necessary tool for such a task. Places in such a net correspond to the static elements of the system like chemical compounds. Transitions correspond to the reactions. The rules of transitions activation and firing allow the modelling of the biological system dynamics. Having a Petri net one can start the analysis which is based on the invariants, maximum common transition sets and t-clusters. Such analysis requires various tools using different file formats which adds to the complexity of such a task. We have developed a Java-based environment to help in the process of net creation, simulation and the analysis in the classical and time Petri nets, while the different formats in which Petri net data is being stored can be used there. Exporting data to other Petri net tools has also been implemented.**

Keywords–Systems biology; Petri nets; model analysis.

I. INTRODUCTION

Systems biology is an interdisciplinary field of study based on biology, mathematics and computer science which focuses on the analysis of complex biological systems. The first step before the analysis is the creation of a model of the system, and its further analysis can lead to the potential discoveries of the behaviors of the system [1] [2]. One of the possible tools for the creation and analysis of such a model is the Petri net theory. A structure of a Petri net is a bipartite directed graph of places and transitions. Places correspond to the static element of the system like chemical compounds, products, substrates, etc. Transitions on the other hand allow modelling of the system dynamics. They represent the reactions taking place within the system. Transitions in the Petri net can be activated and fired depending on the status of the places. The latter depends on the number of tokens in each place, i.e., the number of products of substrates a given place represents. A semi-positive (i.e., non-negative) vector describing the number of tokens in each place is called a marking. Finally, weighted arcs in a Petri net define how many tokens in places a fired transition consumes and produces.

Analysis of such a model is not an easy task and it is divided into different consecutive steps. First, the qualitative and behavioral properties of the net must be determined. Classical Petri net can always be described using a so called incidence matrix. Each entry of this (place $\times$ transition) matrix C gives the token change on a particular place by firing of the respective transition.

On the base of such matrix, the transition invariants can be computed. A t-invariant is a vector $x \in N^m$ where $C \cdot x = 0$ where m is the number of transitions. Every t-invariant corresponds to a set of transitions whose firing a number of times reproduces a given marking of the net. Firing number for each transition is defined within the x vector. Such a chain of reactions that each t-invariant describes, represents some basic behavior of the modelled biological system.

Next, the maximum common transition sets (MCT) can be computed based on the set of t-invariants. The MCT-sets partition the set of transitions into disjoint subsets whose biological meaning can be determined [2]. In other words, the MCT sets consist of the reactions linked with each other.

The most difficult task, yet giving the most valuable knowledge about the modelled system is the cluster analysis. In this phase the set of t-invariants is being divided into disjoint subsets called clusters. To perform this task various tools and methods are necessary. Invariants have to be converted into csv file, which then can be used in, e.g., R language environment. Using cluster algorithms and distance metrics from the R libraries and scripts specially prepared for this task, one can compute different clusterings (i.e., sets of clusters). In order to evaluate clusterings and choose the optimal one, the additional evaluation metrics are necessary. Often Mean Split Silhuette (MSS) and Celiski-Harabasz index [3] are used for such a task. Finally, when the correct clustering has been chosen, the biological meaning of clusters can be determined. This step can lead to the potential discoveries of the new facts concerning the modelled biological system.

All these steps are quite difficult and the number of methods and tools used adds to the overall complexity of Petri net analysis. Drawing a Petri net requires knowledge not only about the biological system, but also about current structural properties of the created Petri net that must be computed. All these problems on a way of successful Petri net model analysis were the reasons to develop a Java-based program, providing aid in all steps of the task: creating, simulating and analysing the given net.

 ISBN: 978-1-61208-409-1

Program interface

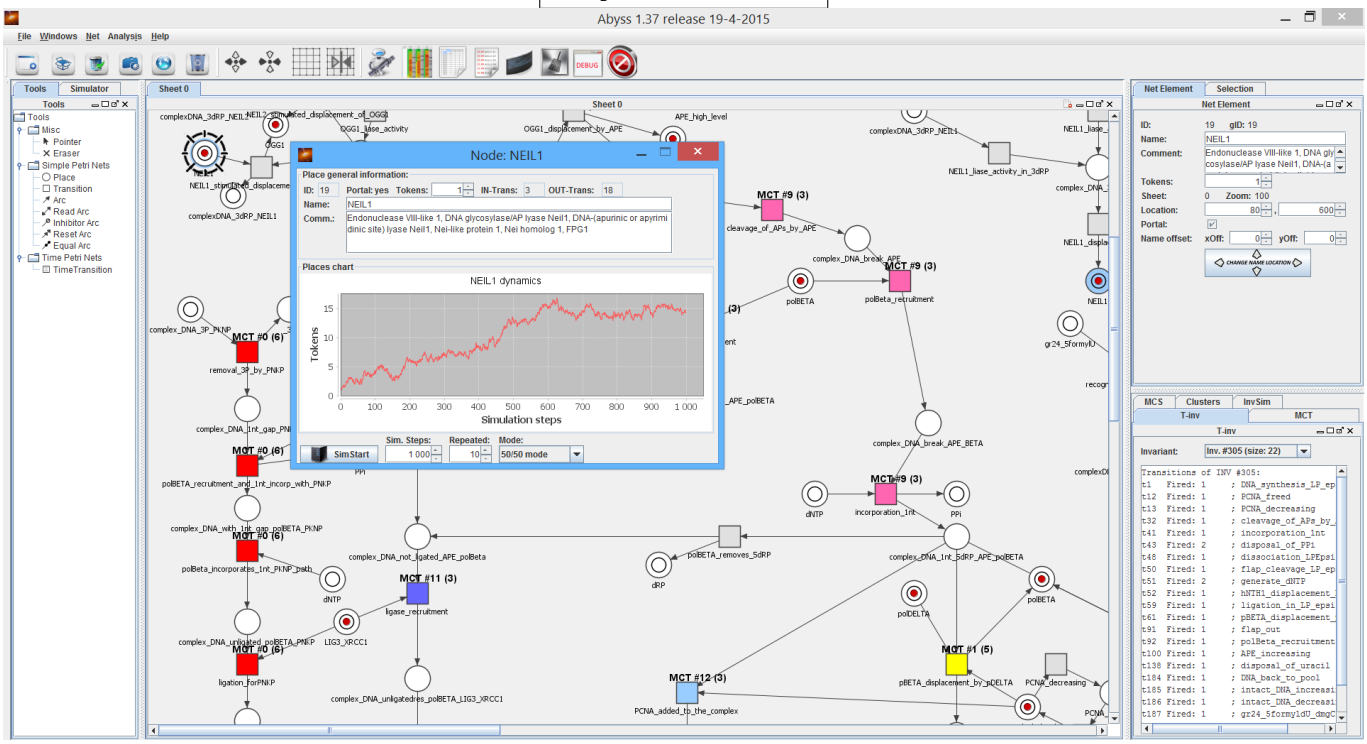

Figure 1. IPNE Java-based environment.

A. *Software*

The proposed tool is a Java-based program consisting of various modules supporting different phases of Petri net-based model analysis [4] [5]. In the current version, the user can draw a classical Petri net, time Petri net and also a net consisting of both classical and time components. The main window of the proposed tool is given in Figure 1. Net qualitative properties can be easily computed and displayed right from the very beginning of drawing the net, which often helps in detection and avoidance of potential problems within the net structure. One of the popular editors for Petri nets is a program called Snoopy [6]. Our application can both load and save nets in the Snoopy format.

When the net is ready, invariants can be obtained using either implemented algorithm within the program or from the external source like Integrated Net Analyzer (INA) [7]. MCT sets are computed as well by our algorithm and both the invariants and the MCT sets can be displayed on the structure of the net within program main window. Additionally, when displaying each given invariant, the number of firings for each transition is also given.

The most complex task lies in the cluster analysis as we have already stated. Our program integrates various R scripts used to compute clusterings using different cluster algorithms and different metrics. Seven cluster algorithms and eight distance metrics can be used, for each of them results for a range of numbers of clusters are given. This allows the user to choose the optimal clustering based on MSS and Celiski-Harabasz evaluation metrics which are computed as well. All these computations are performed in the background by our scripts within the R language, and when the computations are finished our program integrates the data into a table of clusterings. Every clustering (i.e., set of clusters for a given clustering algorithm, distance metric and the number of clusters) can be chosen for more detailed computations. Given the information about the appearance of each invariant within a specific cluster one can obtain important knowledge about the clusters which helps in finding their biological meaning. The number of MCT sets within the cluster and the number of firings of transition are given, along with the MSS measures for each cluster within the clustering. The user can send this data into the Petri net displayed in the main window of the program and analyse the structure of the clusters seeing each transition and its role within them.

II. Conclusion

The described tool is still being developed, but even now we obtained a quite powerful tool aiding the tasks of Petri net drawing, simulation and analysis. The most time consuming task, i.e., the clusters analysis have been made much simpler and faster. Therefore, more time can be devoted to detailed analysis, while all the computations and potential problems with different file formats describing Petri nets are solved in the background by the program. Each clustering and every cluster within it can be analysed thoroughly, while the data concerning them are given in separate windows.

Acknowledgment

This research has been partially supported by the Polish National Science Centre grant No. 2012/07/B/ST6/01537.

 ISBN: 978-1-61208-409-1

REFERENCES

[1] D. Formanowicz, A. Kozak, T. Glowacki, M. Radom, and P. Formanowicz, "Hemojuvelin-hepcidin axis modeled and analyzed using Petri nets," Journal of Biomedical Informatics, vol. 46, 2013, pp. 1030–1043.

[2] D. Formanowicz, A. Sackmann, A. Kozak, J. Blazewicz, and P. Formanowicz, "Some aspects of the anemia of chronic disorders modeled and analyzed by petri net based approach," Bioprocess and Biosystems Engineering, vol. 34, 2011, pp. 581–595.

[3] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," Communications in statistics, vol. 3, 1974, pp. 1–27.

[4] H. Andrzejewski, P. Chabelski, and B. Szawulak, "Integrated system for the creation, simulation and analysis of the Petri nets," Poznan University of Technlogy, B.Sc. (eng.) thesis, 2013.

[5] B. Szawulak, "Integrated system for creation, simulation and analysis of timed Petri nets," Poznan University of Technlogy, M.Sc. thesis, 2014.

[6] M. Heiner, M. Herajy, F. Liu, C. Rohr, and M. Schwarick, "Snoopy A Unifying Petri Net Tool," Lecture Notes in Computer Science, vol. 7347, 2012, pp. 398–407.

[7] "Integrated Net Analyzer," 2001, URL: http://www2.informatik.huberlin.de/ starke/ina.html [retrieved: 2015-03-02].

 ISBN: 978-1-61208-409-1

Investigating Reactions of Laccase and Small Molecule Medium as a Substrate

Zhiyu Liao
Shanghai High School
Shanghai China
e-mail: liaozhiyujonathan@outlook.com

Abstract—**Laccase is a ceruloplasmin widely used in industrial production. It is mainly used in catalytic degradation of lignin and its catalytic process is safe and environmentally friendly. Small molecule medium is not only commonly used as a catalyst in the laccase catalytic reactions, but also can participate in the reaction as a substrate. In this study, the reactions of small molecules as substrates and laccase are investigated by the changes in the absorption wavelength of the reaction system recorded by ultraviolet spectrometer and analysis of the reaction of laccase and small molecules using MATLAB R2013a. It can be seen by the final results of hierarchical clustering analysis and calculation of the Euclidean distance that the entire medium system can be divided into three categories, namely, (i) Cumaric Acid, Sinapic Acid and HBT, (ii) Cumaric Acid and ABTS, and (iii) Ferulic Acid and Syringic Acid. The results of this study provide new ideas for finding and verifying potential medium systems.**

Keywords-laccase; small molecule medium ; principal component analysis; hierarchical clustering analysis.

I. INTRODUCTION

With the progress of science and technology, biological knowledge and technology are being gradually blended into traditional chemistry to form the more and more popular biochemistry. Currently, solving the pollution problem by a variety of enzymes is a new direction; this approach has also been shown to be an effective method. Through a variety of enzymes, the final products from decomposition of pollutants are often simple and harmless inorganic compounds. Many biological toxic, low concentration and difficultly degradable substances can often be decomposed by means of biochemical processes. Among many enzymes, laccase has the advantages of a wide range of substrates, high catalytic activity and no pollution, and thus, not only has great industry usefulness, but also has a very important position in dealing with pollution problems. Through continuous trials and studies, types and applications of the small molecule catalysts that react with laccase have begun to take shape. These small molecule mediators often play a very important role in different fields, and greatly accelerate the reaction efficiency. However, since laccase is an enzyme with a very wide range of substrates, many small molecule mediators are also within the scope of its substrates. Originally as a catalyst, small molecule mediators may cause some effect on the reaction when they act as substrates and are involved in the reaction. Further, small molecule mediators also show other characteristics in addition to catalytic properties in the reaction.

Currently, there is not a clear classification for small molecule mediators of laccase. By analyzing the nature of small molecule mediators when reacting with laccase as a substrate, small molecule mediators are classified by biological information and mathematical methods, and the role of small mediator molecules play in the reaction of laccase can be more accurately predicted. Further, the difficulty of predicting the nature of new small molecule mediators can be reduced, and testing complexity can be simplified.

The rest of the paper is structured as follows. Section II presents the experiments we performed. Section III explains principal component analysis, as well as hierarchical cluster analysis. Section IV presents the results and discussion, and we conclude in Section V.

II. EXPERIMENTS

A. Materials

Small molecules commonly used as medium in laccase reaction are selected, including ABTS, Ferulic Acid, HBT, Syringic Acid, Coumaric Acid, Caffeic Acid, and Sinapic Acid.

B. Ultraviolet spectroscopy

Absorption wavelength variation of the reaction system is recorded by Ultraviolet spectrometer and formed matrices to determine the efficiency and extent of the reaction. Data of wavelength of 240-900nm interception are the basis of analysis data set and used for reaction analysis and classification of mediators.

C. Data Analysis

Matlab R2013b [1] is used for principal component analysis of data set, and for comparison, projection and establishing response structure model diagram. The properties of small molecules in the reaction system are initially manifested. The main components of reaction are re-extracted by initial data matrix transpose, and of small molecules are classified by calculation of the Euclidean distance and hierarchical cluster analysis.

 ISBN: 978-1-61208-409-1

III. PRINCIPLE OF PRINCIPAL COMPONENT ANALYSIS AND HIERARCHICAL CLUSTER ANALYSIS

A. Principal component analysis

In order to fully describe the system in analyzing the current problems, analysts tend to select the relevant indicators as thoughtfully as possible. In fact, many of the social, economic, and technical indicators have a synchronized growth trend. When an analyst intentionally or unintentionally describes the feature of a system by namely different indicators but actual relevant indicators, usually, he faces the issue of multiple correlated variables. Multiple correlations of variables imply artificially exaggerate certain features' position in the system analysis, which affect the objectivity of the analysis and impede decision makers' right judgment.

Principal Component Analysis (PCA) [2] is a basic method for overcoming multiple correlations of variables. PCA is method to establish as few as possible new variables for all variables originally proposed, so that these new variables are independent and uncorrelated, and these new variables should maintain information of the original proposed variables as much as possible. By means of an orthogonal transformation, PCA converts original correlated component random vectors into relevant new random vectors which are corrected to each other. This is manifested on algebra as transforming a covariance matrix of original random vectors into a diagonal matrix, and on geometry as transforming an original coordinate system into a new orthogonal coordinate system, in which sample points spread in the most open p orthogonal directions. Then, multidimensional variables are treated by reducing the dimension of the system so that variables can be converted into a low-dimensional variable system in high accuracy. The low-dimensional variable system is then further transformed into a one-dimensional system through constructing an appropriate value function.

B. Hierarchical cluster analysis

Hierarchical Cluster Analysis (HCA) [3] is one of the basic methods of exploratory research work. The so-called clustering is clustering subjects of study into classes based on the degree of closeness between the study subjects, and in accordance with the principle of "Similar Together, Different Apart".

The idea of hierarchical clustering analysis is that data with higher similarity are closer to each other, have higher tendency with similar properties, and therefore are grouped into the same category. Data can be hierarchical clustered according to the Euclidean distance of matrix, and preliminarily analyzed and classified.

IV. RESULTS AND DISCUSSION

A. Experiment results

The data collected by an ultraviolet spectrometer are expressed in a matrix and formed an original 598 × 660 matrix, wherein each row represents the instrument data measured in every 0.5s in the 190nm to 1850nm wavelength range. Finally, the original matrix is cut, and data within the wavelength range of 240-900nm are selected as the initial data set.

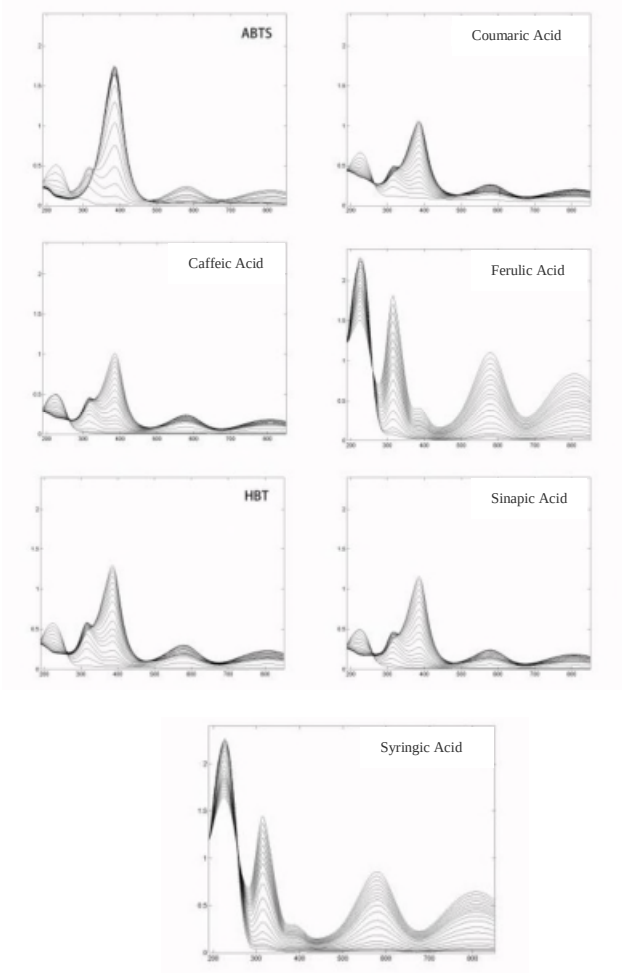

Figure 1. Figures of the raw data collected by the ultraviolet.

Figure 1. shows wavelength changes for small mediator molecules in the reaction system. Among them, there is no significant similarity between each two data sets.

B. Principal component analysis[4]

- Data with time information
  Table 1 summarizes the first and second principal component contribution rates and their sum of each mediator by principal component analysis of the experimental data containing time information. It can be seen from the table that the sums of all of the first and second main component contribution rates are in excess of 95%. The results analysis illustrates that the first and second principal components can be a good representative of the relevant data sets.

 ISBN: 978-1-61208-409-1

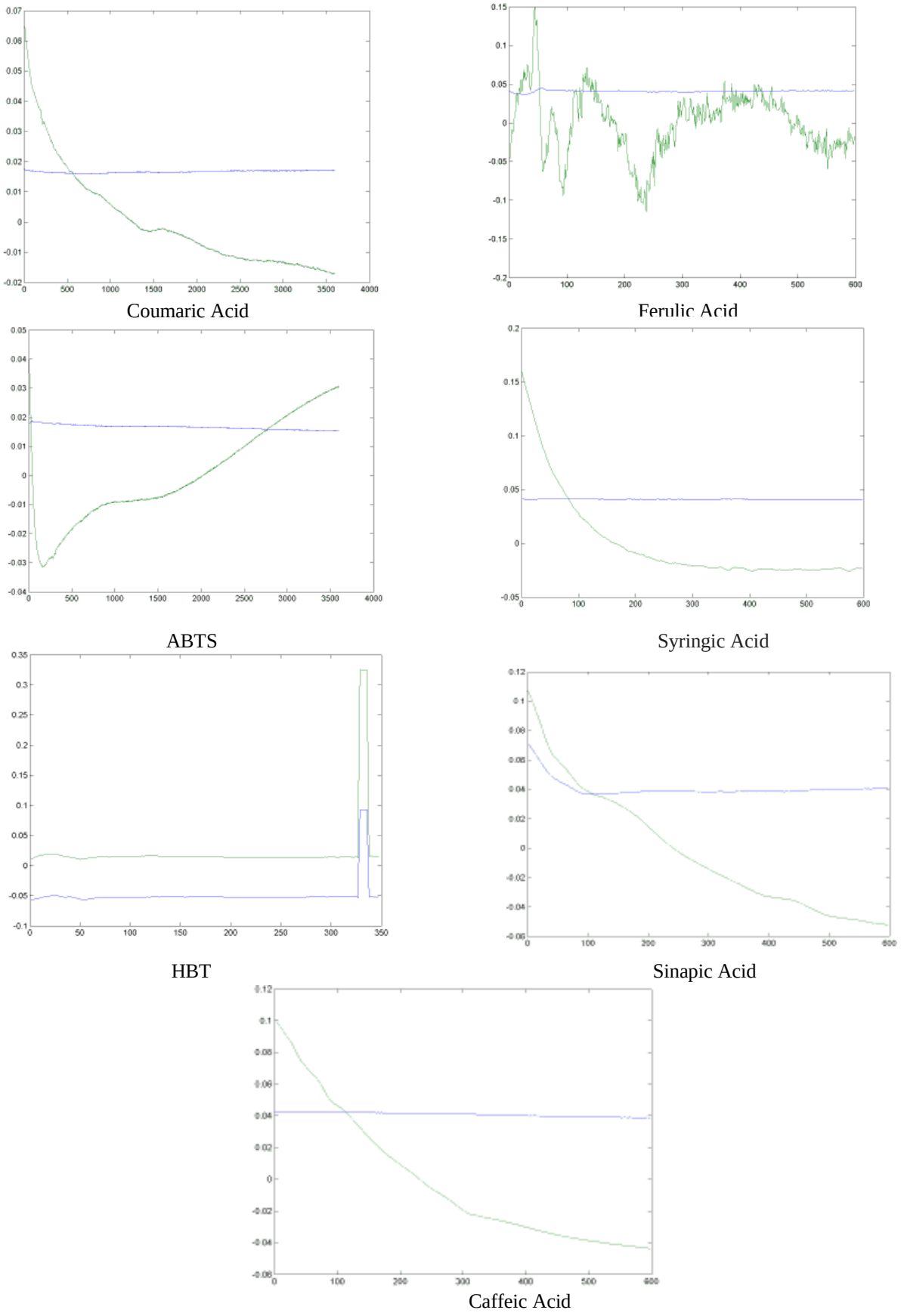

Figure 2. Curves of the first two principal components of the changes of reactions between the small molecule mediums and laccase.

 ISBN: 978-1-61208-409-1

After initial overlap comparison, it can be observed that there is no significant similarity among the above images. The corresponding one-dimensional matrixes of the first and second main components are combined to form a two-dimensional vector. Figure 3 shows the two-dimensional vectors of the tested small molecule mediators.

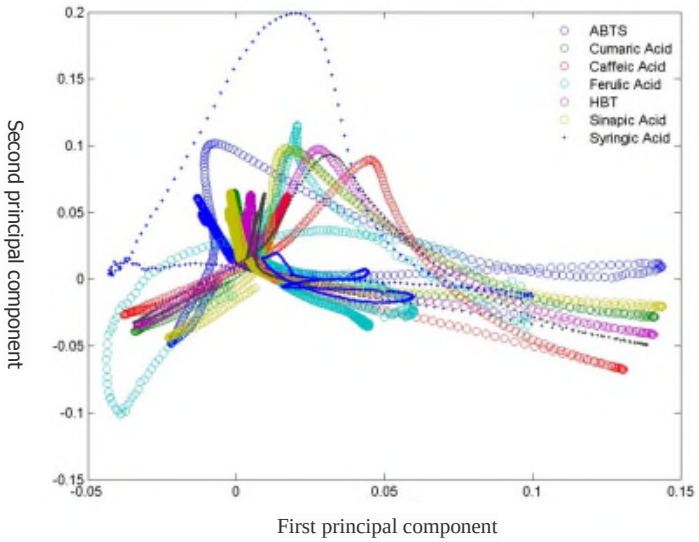

Figure 3. Put all of the seven figures of Figure 2 together.

It can clearly be seen through the above figure that in all the data, some parts are intensively overlapped, while the rest is very discrete. There is no significant similarity and obvious trend among the data.

- Data with wavelength information
  Table 2 summarizes the first and second principal component contribution rates and their sum of each mediator by principal component analysis of the experimental data containing wavelength information. It can be seen from the table that the sums of all of the first and second main component contribution rates are in excess of 95%. The results analysis illustrates that the first and second principal components can be a good representative of the relevant data sets.

  Wavelength data of the first and second principal components are also plotted to obtain seven different curves, respectively representing the properties of the selected seven small molecule mediators in the reaction of laccase. These seven curves are drawn together to get Figure 4:

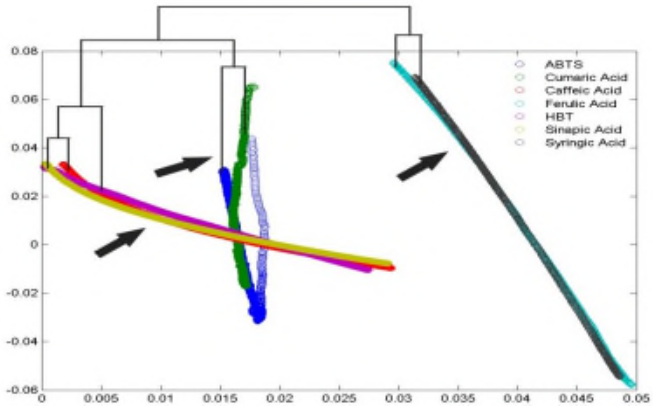

Figure 4. Combination of the seven curves.

C. *Hierarchical cluster analysis*

From Figure 4, the tendency of each image and the similarity between each other can be seen more clearly. Among them, caffeic acid, Sinapic Acid and HBT as a group, p-coumaric acid and ABTS as a group, and ferulic acid and syringic acid as a group， have high degree of overlap of curves within each group, indicating that the nature of the reactions of the small molecule mediators within each group has high similarity. On the other side, there is significantly different tendency between each groups, indicating that the grouping of mediators is qualitative and efficient.

The wavelength changes during the reaction of small molecules with laccase are shown in the reaction structure diagram, wherein horizontal coordinate is the first principal component and vertical coordinates is the second principal component. Reaction structure diagram represents the full nature of small molecules react with laccase, but there is a large number of irregular curves overlap, and there is no clear tendency. Therefore, principal components are re-analyzed by transpose of the original data matrix, and the first and second principal components are combined into a vector. Then, hierarchical clustering analysis is done by calculating the Euclidean distance. The results are shown in the Figure 4, wherein smaller Euclidean distance indicates the nature of the reaction chemistry and the catalytic mechanism are similar. The slopes of curves in the figure are similar, and the degree of correlation is high.

V. CONCLUSION

When small molecule mediators as a substrate are involved in the reaction with laccase, Coumaric Acid, Sinapic Acid, HBT; Coumaric Acid, ABTS; Ferulic Acid and Syringic Acid respectively have substantially the same catalytic mechanism, and can be divided into three groups. The selected molecules form different catalytic systems with laccase and are divided to different categories of medium, although many medium molecules may contain similar functional group (such as benzene ring) or specific chemical

 ISBN: 978-1-61208-409-1

elements (such as nitrogen). Medium molecules may not have the same catalytic reactions with laccase, although they may have the same or similar groups. Medium molecules may impact the catalytic reaction of laccase very differently, there are belong to different groups. These three types of catalytic medium in the catalytic system with laccase exhibits different catalytic mechanism, changes in specific wavelengths are also different. This new discovered classification system can help study the reaction mechanism of laccase and small molecule medium. The screening method used in the study is based on principal component analysis and cluster analysis, and can be used for classification and clustering of unknown reaction medium system of laccase. New small molecule medium systems can also be incorporated into the above classification, and their properties can be predicted by the categories they are belonging to. Further, the classification can also be used as a screening tool for small molecules that are suitable in certain reactions. This rapid screening method can also be extended to other fields, such as study of the reactions of other enzymes and prediction of the properties of other reactants.

REFERENCES

[1] Yilei Zhao, "Basic tutorial of Matlab", unpublished.

[2] Yanke Bao, "Data Analysis", Tsinghua University Press, Sep. 2011, pp. 68-75, ISBN: 978-7-302-26596-2.

[3] Yanke Bao, "Data Analysis", Tsinghua University Press, Sep. 2011, pp. 202-216, ISBN: 978-7-302-26596-2.

[4] 梁逸曾，许青松,《复杂体系仪器分析:白、灰、黑分析体系及其多变量解析方法》, 化学工业出版社, 第 1 版, 2012, Part 1-Chapter 3-Section 2.

TABLE I. THE FIRST AND SECOND PRINCIPAL COMPONENT CONTRIBUTION RATES

	ABTS	Ferulic Acid	HBT	Syringic Acid	Coumaric Acid	Caffeic Acid	Sinapic Acid
First principal component	0.9633	0.9993	0.8944	0.9992	0.8647	0.9035	0.9021
Second principal component	0.0361	5.585×10^{-4}	0.1055	5.380×10^{-4}	0.0922	0.0964	0.0978
Total of First and Second principal component	0.9994	0.9999	0.9999	0.9997	0.9569	0.9999	0.9999

TABLE II. THE FIRST AND SECOND PRINCIPAL COMPONENT CONTRIBUTION RATES OF DATA WITH WAVELENGTH INFORMATION

	ABTS	Ferulic Acid	HBT	Syringic Acid	Coumaric Acid	Caffeic Acid	Sinapic Acid
First principal component	0.9423	0.9099	0.9206	0.9434	0.9956	0.7904	0.9174
Second principal component	0.0543	0.0900	0.0793	0.0565	0.0042	0.1791	0.0825
Total of First and Second principal component	0.9967	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

 ISBN: 978-1-61208-409-1

Monitoring of Biotechnological Strains of the *Bacillus subtilis* / *Bacillus amyloliquefaciens* Group in Natural Habitats by using Multilocus Genetic Barcodes

Oleg N. Reva
Bioinformatics and Computational Biology Unit, Dep. Biochemistry,
University of Pretoria
Pretoria, South Africa
e-mail: oleg.reva@up.ac.za

***Abstract*—*Bacillus subtilis*, *B. amyloliquefaciens* and other related strains isolated from rhizosphere are widely used in biotechnology as Plant Growth Promoting Rhizobacteria (PGPR) showing variable activity in different soils and on different plants. Despite diverse bioactivity, they are almost indistinguishable by phenotype and by 16S rRNA that makes it problematic to trace them down in nature. Spores of these microorganisms are abundant in many habitats, but it remains unclear whether these organisms thrive in these habitats or simply contaminated the samples? In this work, several new computational approaches were proposed for design, evaluation and application of multilocus genetic barcodes for identification and monitoring of biotechnological strains. The barcodes containing 150 marker genes were designed for 35 strains of the *B. subtilis* group and tested on publically available metagenomic datasets. Clear habitat preferences were observed for different subspecies and even individual strains. Also, an approach of evaluation and improving of sensibility and sensitivity of barcodes was proposed.**

Keywords-genetic barcoding; plant growth promoting; Bacillus; metagenomics.

I. INTRODUCTION

Bacillus subtilis was one of the first known bacterial species described in 1835 by Ehrenberg as *Vibrio subtilis* and then renamed to *Bacillus subtilis* in 1872 by Cohn [1]. This spore-forming bacterium caught eye because of its abundance and ease of isolation. From then on *B. subtilis* has become one of the most popular model organism used in bacteriological and genetic laboratories. At dawn of the genomic era it became obvious that *B. subtilis* in fact is a conglomerate of several closely related taxa vernacularly termed the *B. subtilis / B. amyloliquefaciens* group. Many of these organisms found a wide application in biotechnology as biopesticides and plant growth promoters [2], lytic enzyme producers [3] and in decontamination tests [4]. Activities of a biotechnological interest often are associated with genetically isolated groups of these organisms [5]. For example, many commercial plant growth promoters belong to *B. amyloliquefaciens* ssp. *plantarum* [2]; however, several strains of *B. subtilis* and *B. atrophaeus* also were introduced as successful biopesticides [6].

Surprisingly, despite this long history of study of this group of microorganisms, ecology and preferable habitats of species and subspecies of this group remain unclear. This is a back side of the abundance of these microorganisms. Their spores may tolerate even several minutes of boiling and they easily can be spread by water and wind to any environment. This is why these organisms are the most frequent contaminants in microbiological laboratories and an isolation of *B. subtilis* related organisms from any eco-niche cannot a proof of preferable inhabiting this biotope. Our ignorance of habitat preferences of these biotechnologically important microorganisms limits their application and may be an explanation why *Bacillus* based plant growth promoters and biopesticides sometimes show an unstable efficacy in field conditions and on different plants [8]. Strains for new biopesticides often are selected based on their antagonistic activity against pathogens, but it was shown that the bacteria with extraordinary antagonistic activities not always were able to survive in rhizosphere and to colonize plants [10].

Next generation sequencing and metagenomics allowed studying complex microbial populations without isolation of individual strains. The most common approach is amplification of fragments of 16S rRNA from the whole DNA sample using the universal primers followed by massive parallel sequencing of the amplified fragments. Application of this approach is limited in our case as these bacteria are almost indistinguishable by 16S rRNA [5]. Another metagenomic approach consists in sequencing of randomly generated genomic fragments. As there are plenty of bacterial species in complex habitats, any single marker gene for a species of interest may be absent in the generated metagenomic set of reads. To avoid overlooking of species of interest, the genetic barcode sequence should contain multiple marker genes. The aim of this study was to develop computational approaches for a standardized selecting of marker genes and evaluation of the barcode efficacy. An approach of identification of suitable marker genes by whole genome comparison was proposed in our previous publication [11]. Here, the prepared barcodes were used for identification of closely related *Bacillus* species in publically available metagenomic datasets. Then, an in-house Python script was used for evaluation of performance of different loci in the barcode sequences for further improvement of the barcodes. In Section III.A, the results of identification of barcoded strains of *Bacillus* in plant and rhizosphere associated habitats were presented that followed by the analysis of several gut microbiomes (Section III.B) and

 ISBN: 978-1-61208-409-1

estuarine habitats (Section III.C). The approaches of improving of multi-locus barcode sequences to achieve a higher sensitivity and specificity were discussed in the Section III.D. The paper ends up with conclusive remarks.

II. MATERIALS AND METHODS

A. *Complete genome sequences and metagenomic datasets used in this study.*

Genetic barcodes were developed for 34 strains of the *B. subtilis / B. amyloliquefaciens* group representing different species and subspecies including commercial and biotechnologically potential strains (Table 1).

TABLE I. BACILLUS GENOME SEQUENCES USED IN THIS STUDY

	Species and strain	NCBI ID or BioProject
B. subtilis ssp. *subtilis*		
1	168W	NC_000964
2	6051	NC_020507
3	QB928	NC_018520
4	UCMB5121	Newly sequenced
5	UCMB5021	Newly sequenced
6	BSP1	NC_019896
7	BSn5	NC_014976
8	BAB1	NC_020832
9	XF1	NC_020244
10	RO_NN_1	NC_017195
11	*B. subtilis* ssp. *natto* BEST195	NC_017194
12	*Bacillus* sp JS	PRJNA79217
B. subtilis ssp. *spizizenii*		
13	At3	Newly sequenced
14	UCMB5014	PRJNA176696
15	At2	PRJNA176701
16	TU-B-10	NC_016047
17	W23	NC_014479
18	*B. mojavensis* UCMB5075	Newly sequenced
B. atrophaeus		
19	1942	NC_014639
20	UCMB5137	PRJNA176685
B. amyloliquefaciens ssp. *amyloliquefaciens*		
21	TA208	NC_017188
22	XH7	NC_017191
23	LL3	NC_017189
24	DSM7	NC_014551
25	IT-45	NC_020272
B. amyloliquefaciens ssp. plantarum		
26	CAU_B946	NC_016784
27	YAU_B9601_Y2	NC_017061
28	UCMB5007	PRJNA176687
29	UCMB5044	Newly sequenced
30	UCMB5036	Newly sequenced
31	UCMB5140	PRJNA176688
32	At1	PRJNA176703
33	FZB42	NC_009725
34	AS43.3	NC_019842

Barcodes were designed as it was explained in our previous publication [11]. Shortly: complete genome comparison revealed a set of 150 rather conserved core genes, which have accumulated amino acid substitution with a relatively higher rate than other genes. It implied an evolutionary positive selection of amino acid substitutions leading to adaptation of organisms to specific habitat conditions.

Metagenome datasets representing different eco-niches were obtained from NCBI and MG-RAST databases [12] (Table 2).

TABLE II. METAGENOMIC SUBSETS

Biotope	# reads	Types of reads
Rice phyllosphere	1,026,982	Roche 454, ~500 bp.
Meadow grassland , USA	976,268	Roche 454, ~300 bp.
Rain forest soil	782,404	Roche 454, ~300 bp.
Tropical soil	5,235,352	Illumina, 100 bp.
Soybean rhizosphere from Amazon soils	578,060	Roche 454, ~500 bp
Mediterranean oak forest rhizosphere	561,526	Roche 454, ~150 bp
Mangrove estuarine mud	481,226	Roche 454, ~300 bp.
Anthropogenic estuarine mud	526,919	Roche 454, ~300 bp.
Termite gut	99,776	Roche 454, ~600 bp.
Cow gut	264,849	Roche 454, ~100 bp.
Human gut	1,000,000	Sanger, ~1300 bp.
Canine gut	583,523	Roche 454, ~500 bp.

B. *Application of the barcodes.*

DNA reads of metagenomic datasets were aligned against the barcode sequences by BLASTN. Hits with e-

 ISBN: 978-1-61208-409-1

value smaller than 0.00001 and the scores above 75 for short Illumina and Roche 454 reads; above 100 for Roche 454 reads and short contigs around 500 bp; and above 150 for Sanger reads were selected for barcode scoring according to the (1):

$$score = \frac{\sum \frac{b\_score}{r\_length} \times \frac{G - N_g}{G - 1}}{barcode\_length} \quad (1)$$

where *b_score* – blastn score of a read; *r_length* – length of the read; *G* – number of barcodes in the set, in our case – 34; N_g – number of barcodes with which the read gives reliable blast hits. The logic of this score is that the reads aligned through its whole length contribute more to the score than those aligned partially; and the reads with specific hits against one single barcode sequence contribute more than those with multiple hits against many barcodes. An assumption was that the strain specific barcodes with higher scores would indicate presence of similar organisms in the habitat specific metagenome.

Local blast search and statistics were implemented in Python 2.5 scripts.

C. Phylogenomic tree inferring.

Phylogenomic tree was inferred by the maximum likelihood algorithm based on a super-alignment of amino acid sequences of 2,318 identified core genes common for all tested *Bacillus* genomes.

III. Results and Discussion

A. Analysis of the rhizosphere and phyllosphere associated barcodes

Results of the blast search of reads of 5 metagenomic datasets associated with rhizosphere, soil and plants are shown in Figure 1.

In Figure 1 and in the following figures of mapping of the metagenomic reads, the barcodes are numbered in the same order as in Table 1.

The profiles of strains were quite similar in the rhizosphere associated biotopes with a general predominance of *B. amyloliquefaciens* over *B. subtilis* and other species except for the strain *B. subtilis* ssp. *spizizenii* At3. The latter one was frequent in all metagenomic datasets and in two of them it was the most abundant strain. The bacillary microflora of rice phyllosphere was quite similar to those of rhizosphere samples, but an interesting observation was that *B. amyloliquefaciens* ssp. *amyloliquefaciens* were more frequent in the phyllosphere sample, while their closest relatives of the subspecies *plantarum* were more common in the rhizosphere samples.

Bacillus species profile in the oak forest rhizosphere was quite different with *B. subtilis* ssp. *subtilis* and *B. mojavensis* to be the dominant organisms.

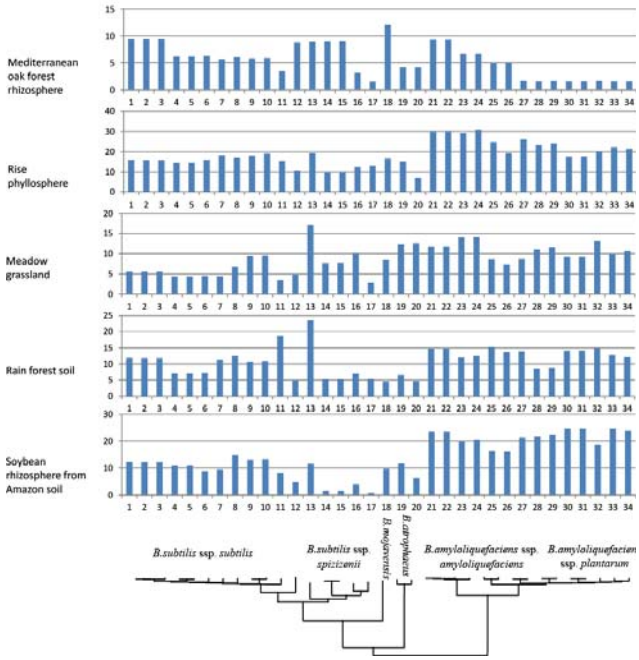

Figure 1. Scores calculated for the barcodes by BLASTN mapping of DNA reads are shown in histograms. Phylogenomic tree of tested strains is shown below.

B. amyloliquefaciens ssp. *plantarum* are widely used in many commercial biopesticides and as plant growth promoting agents. The observation of distribution of *Bacillus* species shown in Figure 1 suggests that strains of *B. amyloliquefaciens* ssp. *amyloliquefaciens* may be a better choice for applying biocontrol agents on leaves; and *B. subtilis* based biopesticides may be more effective in forestry.

B. Analysis of metagenomes of gut microflora

It was interesting to investigate whether the microflora of herbivorous organisms would resemble that one of plants and would there be differences in microbiota of herbivorous and carnivorous organisms. In this study, the metagenomes of gut microflora of cow, human, dog and termite were analysed (Figure 2).

In termite gut, *B. amyloliquefaciens* and *B. atrophaeus* were absolutely dominants. These species were abundant also in guts of cow, but *B. subtilis* ssp. *spizizenii* was also frequent in this microbiota, especially the linage At3 that was although frequent in the rhizosphere (Figure 1). *Bacillus* species were more or less equally distributed in human intestines with minority in *B. subtilis* ssp. *spizizenii* and *B. mojavensis*. Contrary, in canine guts *B. mojavensis* and organisms similar to *B. subtilis* ssp. *spizizenii* W23 were the only present *Bacillus*.

 ISBN: 978-1-61208-409-1

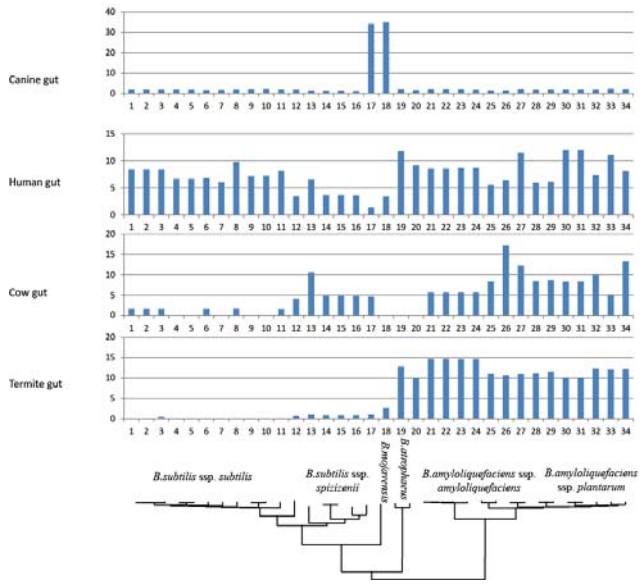

Figure 2. *Bacillus* species profiles in metagenomes of cow, human, canine and termite gut microflora.

Strains of the *B. subtilis* group are used in several medicinal probiotics to prevent gastro-intestinal diseases and dysbacteriosis; however, it has been never reported was there any difference between the strains used for plant protection and those used in probiotics, as they all belonged to the same species [13]. The hypothesis was that although species of *Bacillus* did not belong to human and animals' resident microflora, they constantly had been arriving with the food and might interfere with pathogenic bacteria and the immune system of the higher organism. This research confirmed that *Bacillus* are common in the gut microflora of herbivorous and omnivorous organisms including the human gut microflora. However, a significant difference in species profiles was observed suggesting that human probiotics may not necessary be effective for domestic and farm animals.

C. *Analysis of other metagenomes*

To check how specific the profiles of species of the *B. subtilis* group are in the habitats observed above, several other metagenomic datasets were analysed. In Figure 3, the results of analysis of metagenomes of natural and anthropogenic estuarine mud are shown.

In the natural estuarine mud the species of *B. subtilis* / *B. amyloliquefaciens* were poorer represented than it was in the rhizosphere associated habitats. There were more representatives of *Bacillus* in the estuarine affected by industrial pollution with *B. atrophaeus* became the most abundant species comparing to other species of this group.

Several more metagenomic datasets not mentioned in Table 2 were tried in this study: cliff soil, acid mine drainage, rock biofilm, activated sludge and hydrothermal vent. There were no hits against *Bacillus* specific barcodes indicating that these habitats were not occupied by representatives of the *B. subtilis* and *B. amyloliquefaciens* lineages.

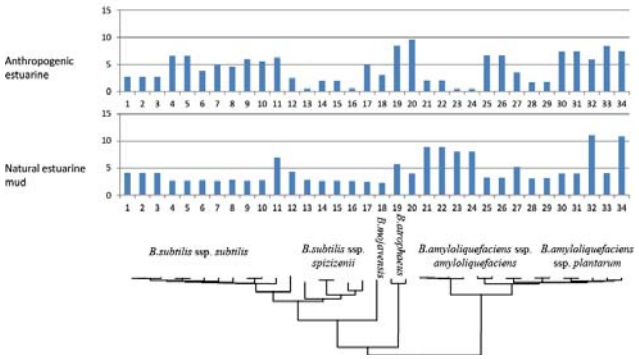

Figure 3. *Bacillus* species profiles in estuarine mud metagenomes.

D. *Further improvement of barcode specificity*

In addition to getting information regarding distribution of *B. subtilis* related species in different habitats, the conducted study aided in identification of loci in the barcode sequences, which contributed to species distinguishing and those which created noise of were silent. A developed Python script returned a graphical representation of efficacy of different barcode loci as shown in Figure 4. In this figure red to brown areas were the most effective in separating even closely related strains. Green areas were species and subspecies specific. Blue areas were not specific and created informative noise, and the white areas never have been hit by any read in this study.

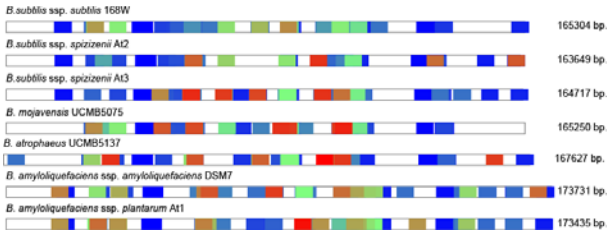

Figure 4. Per locus analysis of discriminative power of several selected barcode sequences.

The barcodes may be improved by removal non-specific loci (in Figure 4 shown in blue) and replacing them with other marker sequences. Than the blastn mapping may be repeated to check whether the specificity of barcodes had been improved.

IV. CONCLUSION

The secret of success of a biotechnological culture is not in its species name but in the strain-specific biological activities. Recent advance in sequencing technologies made it affordable to sequence and compare whole genomes of closely related organism in infancy of clonal segregation and speciation. The most popular next generation sequencing techniques are Roche 454, Illumina and Ion Torrent. They

 ISBN: 978-1-61208-409-1

produce short DNA reads of several hundred nucleotides depending on the used technology, which are randomly generate from genome sequences in a sample. There is an urgent need in new computational techniques for mining of huge amounts of sequence data generated by the next generation sequencers to identify and highlight marker sequences suitable for barcoding of strains of interests and their biological activities. Multilocus barcoding seems to be a promising approach for a reliable identification of strains of closely related bacteria in environmental samples. Prior to this study the genomes of 34 organisms of the *B. subtilis* / *B. amyloliquefaciens* group were compared and 150 core genes with traces of positive selection were chosen for genetic barcode design [11]. The aim of the current research was to evaluate the prepared barcodes on publically available metagenomic datasets and to develop an approach of estimation of efficacy of barcode sequences per individual loci for further improvement of the selectivity and specificity of the method.

Finding of this research was that the species and subspecies of the *B. subtilis* / *B. amyloliquefaciens* group, and even several individual strains of this group, have had preferences in distribution among different biotopes. Rhizosphere biotopes were populated mostly with *B. amyloliquefaciens* ssp. *plantarum* (Figure 1). Representatives of this taxonomic unit are promising biotechnological strains used as components of plant growth promoting preparations and biopesticides. Interestingly, that the strains of *B. amyloliquefaciens* ssp. *amyloliquefaciens*, which are characterized by a stronger enzymatic activity, were more dominant in rice phyllosphere and in the gut microbiota of termites (Figure 2), where they might contribute to enzymatic digestion of complex hydrocarbons [14]. Unexpectedly, the microflora of the oak forest rhizosphere was quite different from that of the grassland and tropical soils with dominance of *B. subtilis* ssp. *subtilis* strains in the former one. This observation may explain why several biotechnological formulations for plant growth promoting and protection not always were equally efficient in laboratory applications and in different field conditions.

One serious problem with multilocus barcode application is that the barcodes comprising different sets of marker sequences may return incomparable results. A Python script was introduced in this work that may be used for evaluation of efficacy of different barcodes and even individual loci in their sequences. This algorithm allows a critical consideration of results of barcode based identification and makes it possible to determine the loci causing noise or false signals. Sensitivity and specificity of barcodes may be improved by removal of these fragments. An ultimate goal of further research on this project will consist in development of a standardized computer based system for a recurrent design, evaluation and improving of barcode sequences for identification and monitoring of biotechnological and pathogenic microorganisms in nature on the level of subspecies or individual strains.

ACKNOWLEDGMENT

This project is supported by the South African National Research Foundation (NRF) grant #93134.

REFERENCES

[1] R. A. Slepecky and H. E. Hemphill, “The genus Bacillus – nonmedical” in The Prokaryotes, vol. 4, M. Dworkin, S. Falkow, E. Rosenberg, K.-H. Schleifer, and E. Stackebrandt, Eds. Singapore: Springer, pp. 530-562, 2006.

[2] X. H. Chen, et al., “Comparative analysis of the complete genome sequence of the plant growth–promoting bacterium Bacillus amyloliquefaciens FZB42,” Nat. Biotechnol., vol. 25, pp. 1007-1014, Aug. 2007, doi:10.1038/nbt1325.

[3] G. R. Castro, B. S. Méndez, and F. Siñeriz, “Amylolytic enzymes produced by Bacillus amyloliquefaciens MIR‐41 in batch and continuous culture,” J. Chem. Tech. Biotechnol., vol. 56, pp. 289-294, Apr. 1993, doi: 10.1002/jctb.280560312.

[4] H. S. Luftman and M. A. Regits, “B. atrophaeus and G. stearothermophilus biological indicators for chlorine dioxide gas decontamination,” Applied Biosafety: Journal of the American Biological Safety Association, Vol. 13, pp. 143-157, Mar. 2008.

[5] L. A. Safronova, L. B. Zelena, V. V. Klochko, and O. N. Reva, “Does the applicability of Bacillus strains in probiotics rely upon their taxonomy?” Can J Microbiol., Vol. 58, pp. 212-219, Feb. 2012, doi: 10.1139/w11-113.

[6] J. W. Kloepper, R. Lifshitz, and R. M. Zablotowicz, “Free-living bacterial inocula for enhancing crop productivity,” Trends in Biotechnology, Vol. 7, pp. 39-44, Feb. 1989, doi:10.1016/0167-7799(89)90057-7.

[7] W. Y. Chan, K. Dietel, S. V. Lapa, L. V. Avdeeva, R. Borriss, and O. N. Reva, “Draft genome sequence of Bacillus atrophaeus UCMB-5137, a plant growth-promoting rhizobacterium,” Genome Announc., Vol. 1, pp. e00233-13, Jun. 2013, doi: 10.1128/genomeA.00233-13.

[8] T. Vasiliki, J. O'Sullivan, A. C. Cassells, D. Voyiatzis, and G. Paroussi, "Comparison of AMF and PGPR inoculants for the suppression of Verticillium wilt of strawberry (Fragaria ananassa cv. Selva)," Applied Soil Ecology, Vol. 32, pp. 316-324, Jul. 2006, doi:10.1016/j.apsoil.2005.07.008.

[9] B. Fan, R. Borriss, W. Bleiss, and X. Wu, “Gram-positive rhizobacterium Bacillus amyloliquefaciens FZB42 colonizes three types of plants in different patterns,” J. Microbiol., Vol. 50, pp. 38-44, Feb. 2012, doi: 10.1007/s12275-012-1439-4.

[10] O. N. Reva, C. Dixelius, J. Meijer, and F. G. Priest, “Taxonomic characterization and plant colonizing abilities of some bacteria related to Bacillus amyloliquefaciens and Bacillus subtilis,” FEMS Microbiol. Ecol., Vol. 48, pp. 249-259, May 2004, doi: 10.1016/j.femsec.2004.02.003.

[11] O. N. Reva, et al, “Genetic barcoding of bacteria and its microbiology and biotechnology applications,” In Bioinformatics and Data Analysis in Microbiology, O. Tastan Bishop Ed., Caister Academic Press, pp. 229-244, Mar. 2014, ISBN 978 190823 0737.

[12] F. Meyer, et al., “The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes,” BMC Bioinformatics, Vol. 9, p. 386, Sep. 2008, doi: 10.1186/1471-2105-9-386.

[13] I. B. Sorokulova, et al, “The safety of two Bacillus probiotic strains for human use,” Dig. Dis. Sci.,Vol. 53, pp. 954-963, Apr. 2008, doi: 10.1007/s10620-007-9959-1.

[14] S. Um, A. Fraimout, P. Sapountzis, D. C. Oh, and M. Poulsen, “The fungus-growing termite Macrotermes natalensis harbors bacillaene-producing Bacillus sp. that inhibit potentially antagonistic fungi,” Sci. Rep., Vol. 3, p. 3250, Nov. 2013, doi: 10.1038/srep03250.

 ISBN: 978-1-61208-409-1

Prokaryotes, Metagenomics, and GC-Content

Erin R. Reichenberger*, Gail L. Rosen†, Uri Hershberg*‡, and Ruth Hershberg§
*Department of Biomedical Engineering, Science & Health Systems, Drexel University, Philadlephia, PA 19104 USA
†Department of Computer and Electrical Engineering, Drexel University, Philadelphia, Pennsylvania 19104 USA
‡Department of Microbiology and Immunology, Drexel University, Philadelphia, Pennsylvania 19104 USA
§Rachel and Menachem Mendelovitch Evolutionary Processes of Mutation and Natural Selection Research Laboratory
Department of Genetics, the Ruth and Bruce Rappaport Faculty of Medicine, Technion University, Haifa, 31096 Israel

***Abstract*—The degree of variation in nucleotide content across all prokaryotic genomes is expansive and ranges from ∼15% to ∼75% guanine and cytosine (GC). There is an ongoing debate as to the causes of this extensive variation, however, since variation in nucleotide content is a genome-wide trait that affects the genome as a whole, it is highly interesting to understand what drives such variation. Employing 183 metagenomic datasets (959G) from numerous types of environments, a Unix environment pipeline of command-line bioinformatics tools, scripting languages, and statistical programs was employed to investigate the influence of environment on GC-content. Using several statistical approaches, we show that each type of environment has a distinct GC-signature that cannot be entirely explained by disparities in phylogenetic composition. Further, our results indicate that environment *and* phylogeny impact nucleotide composition.**

Keywords–GC-Content; Prokaryotes; Metagenomics; Mutation; Genomic Variation; Big Data, Environmental Influence.

I. INTRODUCTION

The causes of the great variation in nucleotide composition of prokaryotic genomes have long been disputed [1]–[3]. In our previous work, we used extensive metagenomic and whole-genome data containing over 31 million sequences to demonstrate that both phylogeny and the environment shape prokaryotic nucleotide content [4]. The GC-content – which is the percentage of guanine and cytosine in a genome or fragment of DNA is important as it can describe the makeup of an organism, provide insight into an organism's evolution, and expand our understanding of gene expression.

II. METHODOLOGY

Shotgun-sequenced fasta files (183 datasets) from 14 environments were obtained from MG-Rast [5]. The details of each project's methodology, metadata and geographic location can be found utilizing a mapping API we created (http://simlab.biomed.drexel.edu/maps/map.php) [6]–[17].

After screening each dataset (e.g., ambiguous/short reads), the remaining sequences were extracted and classified according to phylogeny [18]. The GC-content was calculated for each classified read, followed by a mean GC-content calculation for each phylum, each sample (there were multiple samples in an environmental category), and each environmental category.

III. RESULTS

After calculating the mean GC for all environments, we found that each environment carried a distinct GC-content signature. We found a similarly distinct GC-level trend in 111 samples that comprised a single type of environment. To rule out the possibility that variation in GC-composition between environments could be explained by differences in phylogenetic composition, each environment's prokaryotic community was investigated from two standpoints; the microbial composition and the phylum pair-wise correlation level in GC-content in each environmental category.

A. Microbial Composition

The relative abundance of each phylum in an environment was calculated. Additionally, to assess whether phyla differed at the genus-level, a taxonomic list of the genus names present in each environment was compiled. Using the intersection and union of the lists, the level of similarity (Jaccard similarity coefficient) in the genera contained within two environments was calculated.

B. Phyla and GC-Content

In the process of looking at phylogenic distribution, we found that different phyla were characterized by different mean GC-contents. Additionally, some phyla were characterized by a much broader GC-content range than others. These averages and possible ranges of nucleotide compositions for each taxonomic classification (phylum-level) were, to a large extent, maintained across different environments and were in accord with the GC-levels of fully-sequenced prokaryotic genomes. Phylogeny therefore seems to impose a clear limit on the range of nucleotide content a prokaryote can adopt.

C. Hypergeometric Distribution, Phyla, and GC-Content

The GC-content variation seen in prokaryotes provided an opportunity to observe the behavior of a phylum. Using our largest environmental dataset (111 samples), we found that the GC-content of a phylum with a high range of variability would be at its upper bounds in a high GC sample and the lower bounds in a low GC sample.

D. Correlations, Phyla, and GC-Content

The correlative relationship between the GC-content of each phylum was assessed using the Spearman correlation coefficients. Our analysis showed a number of statistically significant correlations which appeared at a frequency much greater than expected by chance. A significant correlation would indicate that whatever force influenced the nucleotide content in one phylum, had a similar effect on the nucleotide content of the remaining phyla.

E. Assessing Correlations: Phyla, GC-Content, and the 3rd Codon Position of 4-fold Redundant Amino Acids

We confirmed our results and ensured that our findings were not related to artifacts due to amino acid usage by annotating the classified sequences and re-running the correlative analysis on them [19]. The annotated sequences were

 ISBN: 978-1-61208-409-1

examined for the location of those amino acid with four-fold redundancies (Alanine, Arginine, Glycine, Leucine, Proline, Serine, Threonine, Valine) and the 3rd codon positions of these codons were extracted for GC-content calculations. As the third codon positions of fourfold degenerate codons do not affect the amino acid sequence of a protein, their nucleotide content should not be affected by selection at the level of amino acid usage. We found that the GC-content of the 3rd codon position of fourfold degenerate codons within protein-coding genes was correlated between phyla across environments far more frequently than expected by chance.

IV. Conclusion

Employing numerous shotgun-sequenced datasets as well as data from all currently available fully-sequenced genomes, we show that both phylogeny and environment influence prokaryotic nucleotide composition. We demonstrate that, across environments, different phyla have distinct nucleotide compositions. We then show that GC-levels vary by environment in a manner that can not be explained solely by differences in phylogenetic composition. Combined, our results demonstrate that both phylogeny and the environment significantly affect nucleotide composition and that the environmental differences affecting nucleotide composition are far subtler than previously appreciated.

Acknowledgment

The authors would like to thank Calvin Morrison for his assistance, and Yemin Lan for gathering the AAI genomic data.

This short paper is an advance of the publication *Prokaryotic nucleotide composition is shaped by both phylogeny and the environment*. This collaborative project is supported by the Louis and Bessie Stein Foundation. ERR is supported by a Ford Foundation; RH is supported by ERC FP7 CIG grant [321780], a Yigal Allon Fellowship, and by the Robert J. Shillman Career Advancement Chair. Research reported in this publication is supported by the NIH [P01AI106697], NSF [0845827, 1120622] and the Department of Energy [DE-SC0004335].

References

[1] K. Foerstner, C. von Mering, S. Hooper, and P. Bork, "Environments shape the nucleotide composition of genomes," EMBO Reports, vol. 6, no. 12, Dec 2005, pp. 1208–1213.

[2] F. Hildebrand, A. Meyer, and A. Eyre-Walker, "Evidence of Selection upon Genomic GC-Content in Bacteria," PLOS Genetics, vol. 6, no. 9, Sep 2010.

[3] R. Raghavan, Y. D. Kelkar, and H. Ochman, "A selective force favoring increased G plus C content in bacterial genes," Proceedings of the National Academy of Sciences of the United States of America, vol. 109, no. 36, Sep 4 2012, pp. 14 504–14 507.

[4] E. R. Reichenberger, G. Rosen, U. Hershberg, and R. Hershberg, "Prokaryotic nucleotide composition is shaped by both phylogeny and the environment," Genome Biology and Evolution, 2015 (Accepted).

[5] F. Meyer, D. Paarmann, M. D'Souza, R. Olson, E. M. Glass, M. Kubal, T. Paczian, A. Rodriguez, R. Stevens, A. Wilke, J. Wilkening, and R. A. Edwards, "The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes," BMC Bioinformatics, vol. 9, Sep 19 2008.

[6] F. E. Angly, D. Willner, A. Prieto-Davo, R. A. Edwards, R. Schmieder, R. Vega-Thurber, D. A. Antonopoulos, K. Barott, M. T. Cottrell, C. Desnues, E. A. Dinsdale, M. Furlan, M. Haynes, M. R. Henn, Y. Hu, D. L. Kirchman, T. McDole, J. D. McPherson, F. Meyer, R. M. Miller, E. Mundt, R. K. Naviaux, B. Rodriguez-Mueller, R. Stevens, L. Wegley, L. Zhang, B. Zhu, and F. Rohwer, "The GAAS Metagenomic Tool and Its Estimations of Viral and Microbial Average Genome Size in Four Major Biomes," PLOS Computational Biology, vol. 5, no. 12, Dec 2009.

[7] P. Belda-Ferre, L. Alcaraz, R. Cabrera-Rubio, H. Romero, A. Simon-Soro, M. Pignatelli, and A. Mira, "The oral metagenome in health and disease," ISME Journal, vol. 6, no. 1, Jan 2012, pp. 46–56.

[8] C. Desnues, B. Rodriguez-Brito, S. Rayhawk, S. Kelley, T. Tran, M. Haynes, H. Liu, M. Furlan, L. Wegley, B. Chau, Y. Ruan, D. Hall, F. E. Angly, R. A. Edwards, L. Li, R. V. Thurber, R. P. Reid, J. Siefert, V. Souza, D. L. Valentine, B. K. Swan, M. Breitbart, and F. Rohwer, "Biodiversity and biogeography of phages in modern stromatolites and thrombolites," Nature, vol. 452, no. 7185, Mar 20 2008, pp. 340–U5.

[9] E. A. Dinsdale, R. A. Edwards, D. Hall, F. Angly, M. Breitbart, J. M. Brulc, M. Furlan, C. Desnues, M. Haynes, L. Li, L. McDaniel, M. A. Moran, K. E. Nelson, C. Nilsson, R. Olson, J. Paul, B. R. Brito, Y. Ruan, B. K. Swan, R. Stevens, D. L. Valentine, R. V. Thurber, L. Wegley, B. A. White, and F. Rohwer, "Functional metagenomic profiling of nine biomes," Nature, vol. 452, no. 7187, APR 3 2008, pp. 629–U8.

[10] R. A. Edwards, B. Rodriguez-Brito, L. Wegley, M. Haynes, M. Breitbart, D. M. Peterson, M. O. Saar, S. Alexander, E. C. Alexander, Jr., and F. Rohwer, "Using pyrosequencing to shed light on deep mine microbial ecology," BMC Genomics, vol. 7, Mar 20 2006.

[11] V. Kunin, J. Raes, J. K. Harris, J. R. Spear, J. J. Walker, N. Ivanova, C. von Mering, B. M. Bebout, P. N. R., P. Bork, and P. Hugenholtz, "Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat," Molecular Systems Biology, vol. 4, 2008, p. 198.

[12] X. Mou, S. Sun, R. A. Edwards, R. E. Hodson, and M. A. Moran, "Bacterial carbon processing by generalist species in the coastal ocean," Nature, vol. 451, no. 7179, Feb 7 2008, pp. 708–U4.

[13] D. T. Pride, J. Salzman, M. Haynes, F. Rohwer, C. Davis-Long, R. A. White, III, P. Loomer, G. C. Armitage, and D. A. Relman, "Evidence of a robust resident bacteriophage population revealed through analysis of the human salivary virome," ISME Journal, vol. 6, no. 5, MAY 2012, pp. 915–926.

[14] B. Rodriguez-Brito, L. Li, L. Wegley, M. Furlan, F. Angly, M. Breitbart, J. Buchanan, C. Desnues, E. Dinsdale, R. Edwards, B. Felts, M. Haynes, H. Liu, D. Lipson, J. Mahaffy, A. Belen Martin-Cuadrado, A. Mira, J. Nulton, L. Pasic, S. Rayhawk, J. Rodriguez-Mueller, F. Rodriguez-Valera, P. Salamon, S. Srinagesh, T. F. Thingstad, T. Tran, R. V. Thurber, D. Willner, M. Youle, and F. Rohwer, "Viral and microbial community dynamics in four aquatic environments," ISME Journal, vol. 4, no. 6, Jun 2010, pp. 739–751.

[15] B. K. Swan, C. J. Ehrhardt, K. M. Reifel, L. I. Moreno, and D. L. Valentine, "Archaeal and Bacterial Communities Respond Differently to Environmental Gradients in Anoxic Sediments of a California Hypersaline Lake, the Salton Sea," Applied and Environmental Microbiology, vol. 76, no. 3, Feb 2010, pp. 757–768.

[16] L. Wegley, R. Edwards, B. Rodriguez-Brito, H. Liu, and F. Rohwer, "Metagenomic analysis of the microbial community associated with the coral Porites astreoides," Environmental Microbiology, vol. 9, no. 11, Nov 2007, pp. 2707–2719.

[17] T. Yatsunenko, F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. N. Baldassano, A. P. Anokhin, A. C. Heath, B. Warner, J. Reeder, J. Kuczynski, J. G. Caporaso, C. A. Lozupone, C. Lauber, J. C. Clemente, D. Knights, R. Knight, and J. I. Gordon, "Human gut microbiome viewed across age and geography," Nature, vol. 486, no. 7402, Jun 14 2012, pp. 222+.

[18] A. Brady and S. Salzberg, "PhymmBL expanded: confidence scores, custom databases, parallelization and more," Nature Methods, vol. 8, no. 5, May 2011, pp. 1208–1213.

[19] M. Rho, H. Tang, and Y. Ye, "FragGeneScan: predicting genes in short and error-prone reads," Nucleic Acids Research, vol. 38, no. 20, Nov 2010.

 ISBN: 978-1-61208-409-1

Gene Expression Profile of a Plant Growth Promoting Rhizobacterium *Bacillus atrophaeus* UCMB-5137 in Response to Maize Root Exudate Stimulation

Liberata Mwita, Wai Yin Chan, Oleg N. Reva
Department of Biochemistry
University of Pretoria
Pretoria, South Africa
E-mails: libemwi@yahoo.co.uk,
annie.chan@fabi.up.ac.za,oleg.reva@up.ac.za

Sylvester Lyantagaye
Department of Molecular Biology and Biotechnology
University of Dar-es-salaam
Dar-es-salaam, Tanzania
E-mail: lyantagaye@gmail.com

Svitlana V. Lapa, Lilija V. Avdeevad
Department of Antibiotics
D.K Zabolotny Institute of Microbiology and Virology
Kiev, Ukraine
E-mails: slapa@ukr.net, avdeeva@imv.kiev.ua

Abstract— **Plant Growth Promoting Rhizobacteria (PGPR) are widely used in agriculture as an ecologically safe replacement of chemical pesticides and fertilizers. In this work, gene expression regulation under stimulation by maize root exudate was studied in a PGPR strain *Bacillus atrophaeus* UCMB-5137. A strong up-regulation of synthesis of (p)ppGpp (guanosine pentaphosphate) alarmone was observed. This alarmone is associated with the stringent response in bacteria, but its involvement in regulation of plant colonization has never been reported. Comparison of the profiles of gene expression at the classical stringent response and in the current experiment showed only a partial overlap that allowed us to make a conclusion that the tested bacteria had not been starved but responded specifically to the root exudate stimulation and (p)ppGpp alarmone is involved in this complex response. Comparison of the gene expression regulation in *B. atrophaeus* UCMB-5137 with a report obtained on a similar experiment with another PGPR strain *B. amyloliquefaciens* FZB42 showed that these two closely related organisms had used different strategies of plant colonization.**

Keywords—Bacillus atrophaeus; plant growth promoting; gene expression; transcriptional regulation.

I. INTRODUCTION

Application of beneficial microorganisms residing in the rhizosphere to promote plant growth is a promising and efficient method of biocontrol of plant pathogens; it is safe for human health, other soil microorganisms and the environment. Plant growth promoting rhizobacteria (PGPR) is a diverse group of microorganisms living in the plant rhizosphere. They exert beneficial direct or indirect effects on plant growth [1]. Till now, the gene regulation in PGPR *Bacillus* during root colonization was studied only in a few bacteria [13], and it remains unclear to which extend other PGPR bacteria follow the same pattern of gene regulation. Introduction of new generation sequencing techniques (NGS) made it possible to investigate gene expression profiles under similar experimental conditions in multiple taxonomically related PGPR bacteria by using the RNA-Seq approach. Here, we want to present our very first result in a planned series of experiment. *Bacillus atrophaeus* is one of PGPR able to colonize plants by forming thick colonies on the root surface. This study aimed at investigating the gene expression profile in UCMB-5137 stimulated by maize root exudate and compare the results to those obtained for PGPR *B. amyloliquefaciens* [13]. Understanding gene regulation underlying interaction between this bacteria and plants will broaden our understanding on how the PGPR bacteria belonging to *Bacillus subtilis* taxonomic clade can be effectively used to promote plant growth taking into consideration the commonalities and differences in gene regulation in these bacteria. The next section will provide information on the previous related researches. Then, the section on the methods is followed by a discussion of results obtained in this study and a conclusion section.

II. STATE OF THE ART

Use of Bacilli as biocontrol agents is an advantage over other bacteria because they form thermostable and chemically resistant endospores [12]. Complete genome sequence of *Bacillus atrophaeus* UCMB-5137 showed

 ISBN: 978-1-61208-409-1

multiple horizontally acquired unique genes, which were hypothesized as possible source of an extraordinary activity in plant root colonization [2]. This strain has shown a significant potential of promoting plant growth in greenhouse trials (ongoing research). It was therefore interesting to study gene regulation mechanisms taking place as it interacts with plant. The interaction between *Bacillus amyloliquefaciens* FZB42 under maize root exudate stimulation was studied previously using a method developed by Fan et al. [13]. It is demonstrated in this publication that the maize root exudate added to liquid medium may simulate plant derived signals stimulating root colonization behavior in PGPR bacteria. We used their approach except for transcriptomic profiling, where we used RNA-Seq. RNA-Seq is a sequence based approach, which has several advantages over microarray techniques including a broader dynamic range of expression levels [14]. Our expectation was that the gene expression profiles obtained by both methods will be comparable and it will be possible to identify similarities and differences in plant colonization strategies of these two PGPR strains. *B. atrophaeus* UCMB-5137 was selected for this study because being as active PGPR bacterium as *Bacillus amyloliquefaciens* FZB42, it showed some alterations in root colonization behaviour, which will be discussed in the conclusion.

III. MATERIALS AND METHODS

Root exudate was extracted as it was described previously [13]. *B. atrophaeus* UCMB-5137 was grown up on the liquid medium (1 C medium containing 0.1% glucose) with (experiment) and without (control) 0.25 mg/ml maize root exudate. RNA from two independent experiment samples and three control samples were extracted using ZR Fungal/ Bacteria RNA Mini Prep kit and sequenced by MiSeq 500 Illumina in Inqaba (Pretoria, South Africa). They were quality controlled and trimmed, and then mapped against the available complete genome sequence of UCMB-5137 [2] using CLC Genomics Workbench 7. EDGE statistics approach was used to identify significantly up- and down- regulated genes (at least 3 folds difference) with a p-value ≤ 0.01.

IV. RESULTS AND DISCUSSION

Tiny amounts of root exudate caused a significant up and down regulation of many genes. A remarkable up-regulation was observed for genes responsible for the following mechanisms: stress response, biofilm formation, transcription and translation regulation, heme synthesis, ribosome proteins, cell division, vitamin synthesis and quorum sensing. Meanwhile, down-regulation of genes belonging to the following functional categories: transport proteins, DNA replication, purines and pyrimidines synthesis, sporulation and secondary metabolite biosynthesis were observed. Remarkable, it was the complete silencing of multiple phage associated genes. Approximately, 70% of the genes regulated by root exudates were annotated with known functions whilst 30% were annotated as hypothetical, putative or unknown proteins. An overview of regulated genes is shown in Table 1.

TABLE I. GENES REGULATED BY ROOT EXUDATE

Pathway	# up-regulated	# down-regulated
Aerobic respiration	4	1
Amine and polyamine synthesis	2	1
Amino acid synthesis	1	18
Anaerobic respiration	1	3
Antibiotic and bacteriocin synthesis	1	7
Aromatic compound synthesis	1	1
Biofilm formation and regulation	1	2
Carbohydrate synthesis and degradation	3	19
Cell division	4	0
Cell wall protein synthesis	12	5
Chemotaxis and motility	1	5
Co-factor and vitamin synthesis and utilization	8	12
DNA replication	2	2
Fatty acid synthesis and degradation	0	6
Formaldegyde assimilation	1	1
Nucleotide synthesis	0	9
Other pathways of degradation of complex compounds	2	4
Protein maturation and activation	14	3
Ribosomal proteins	10	4
Sporulation regulation	3	3
Stress response and detoxication	28	5
Transcriptional regulation	31	2
Transport and uptake	5	21
Urea degradation	0	2

One of the important observation was a strong up-regulation of synthesis of the alarmone (p)ppGpp (guanosine pentaphosphate). This alarmone is known to be responsible for the stringent response in bacteria, usually associated with starvation [3]. It controls many metabolic reactions including inhibition of protein synthesis, DNA replication and transcription when bacteria experience a shortage of nutrients. Surprisingly, this alarmone activation was in response to the root exudate stimulation while the control bacteria did not experience any shortage of nutrients. It was hypothesized that the stringent response in this PGPR bacterium was needed to prepare the organism for root colonization. It was interesting to compare genes expression profiles under root exudate stimulation against the classical stringent response described by Eymann et al. [3]. One similarity was that the general stress response pathways were

 ISBN: 978-1-61208-409-1

up regulated in both cases. Another similarity was down-regulation of the amino acid biosynthesis, DNA replication, thiamine biosynthesis, fatty acid biosynthesis and nucleotide biosynthesis metabolic pathways. Silencing of phage-related genes was also reported in both cases. However, a number of discordances were also observed. Pathways of cell wall biosynthesis, ribosome proteins, protein translation, maturation, activation and utilization were negatively controlled by stringent response, but they were up-regulated in this study. Contrary, cross-membrane transportation, urea degradation and preparation for sporulation were up-regulated by the stringent response but down-regulated by the root exudate.

Stress related proteins were up-regulated by maize root exudates. A stress state in bacteria was defined by Lengeler et al. [4] as any change in the environment that provokes a significant change in cell physiology. Up-regulated stress related proteins included heat stress induced protein, manganese superoxide dismutase, organic hydroperoxide resistance reductase B, alkyl hydroperoxide reductases, cold shock protein CspA and integral membrane protein YggT responding to heat, oxidative, cold and osmotic stresses respectively. Bacteria may produce multiple stress resistance proteins to anticipate for future stress [5]. Some of these stress resistance proteins are detoxifying enzymes, e.g., superoxide dismutase and thiol peroxidase, which are used to neutralize oxidative burst from plants [6].

Transcriptional regulators mediate bacteria adaptive responses to continuous changes in their environment [7]. In this study, many transcriptional regulators responsible for variety of responses were up regulated. Among them there were SinR regulator and transition state regulatory protein AbrB, which regulate post exponential phase responses (competence, sporulation and biofilm formation); transport and uptake regulators Fur (iron uptake) and Zur (zinc uptake); peroxide stress regulator PerR; thiol specific oxidative response regulator Spx; and RsbR positive regulator of sigma-B. Sigma-B is a general stress transcription top level regulator, which is activated when bacteria is exposed to environmental stresses.

Remarkably, most of carbohydrate, amino and fatty acid biosynthesis and degradation metabolic pathways were down regulated by the root exudate. This might be connected to the observed up-regulation of the GntR family transcriptional regulator, which represses the general metabolism.

Communication within and between bacterial occur through autoinducers, among which the most studied are quorum sensing autoinducers [8]. Bacteria use quorum sensing to control their cell population density and sense metabolic potential of the environment [4]. Up regulation of *luxS* gene, which is involved in synthesis of autoinducer 2 (AI-2), was observed. AI-2 is used for interspecies cell-to-cell communication. Up-regulation of *luxS* may be important to enable communication between bacteria cells and plant cells during root colonization and biofilm formation. Biofilm is an extracellular matrix produced by complex aggregate of cells that adhere to each other [9]. Bacteria produce biofilm to attach and maintain contact with the host plant during bacteria-plant interaction [10]. It was observed in previous research that *B. atrophaeus* UCMB-5137 colonize plants by forming biofilm on the root surface. Up-regulation of Veg protein, which stimulates biofilm formation [9] and YMCA protein, – a master regulator for biofilm formation, – is in agreement to this observation.

V. CONCLUSION AND FUTURE WORK

Even though (p)ppGpp was up regulated and some pathways might have been controlled by the stringent response, we concluded that the bacteria have not been starved in our experiment and the gene transcription was regulated by several other signalling pathways. Anyway, the involvement of (p)ppGpp alarmone regulation in plant root colonization have never been reported before including the recently published paper on gene expression regulation by root exudate in a PGPR strain *B. amyloliquefaciens* FZB42 [11]. In general, the gene expression profile in *B. amyloliquefaciens* FZB42 was quite different to that in *B. atrophaeus* UCMB-5137 with strong up-regulation of motility proteins and antibiotic biosynthesis. Also, dissimilar was the pattern of root colonization observed by luminescent microscopy (Figure 1).

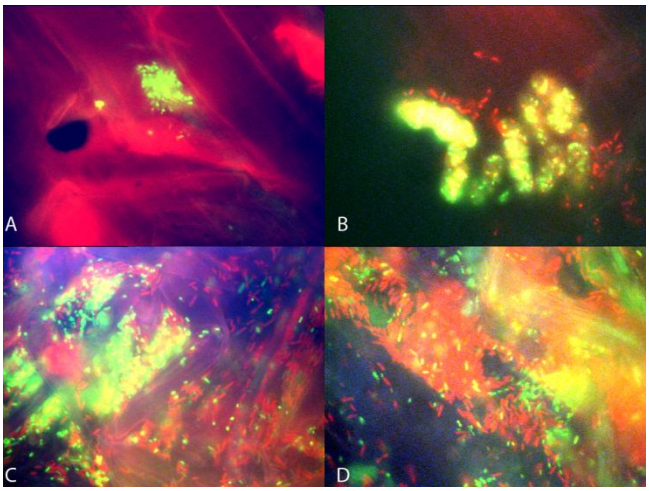

Figure 1. *Colonization of barley roots by B. atrophaeus* UCMB-5137 *(A and B) and by B. amyloliquefaciens* UCMB-5113 *(C and D).*

B. atrophaeus had formed thick colonies on the root surface, while *B. amyloliquefaciens* had been characterized by active penetration into plant tissues and an endophytic growth [12]. This difference in behavioural strategies may explain the observed striking differences in the root exudate stimulated gene expression profiles in these two relative PGPR bacteria. In the future, we want to repeat this experiment with a number of selected PGPR *Bacillus*, belonging to closely relates species *B. amyloliquefaciens*, *B. subtilis* and *B. mojavensis*, which have been sequenced recently (see NCBI bio-projects with the reference numbers: 176685, 176687, 176696, 176688, 176703 and 176701).

 ISBN: 978-1-61208-409-1

ACKNOWLEDGMENT

This work was supported by the funding from the Southern African Biochemistry and Informatics for Natural Product [http://www.sabina-africa.org/], and by a grant provided by the Genomic Research Institute (GRI) at the University of Pretoria for genome sequencing.

REFERENCES

[1] B. Lugtenberg and F. Kamilova, "Plant Growth Promoting Rhizobacteria," Annu. Rev. Microbiol., vol. 63, 2009, pp. 541-56, doi:10.1146/annurev.micro.62.081307.162918.

[2] W. Y. Chan et al. "Draft genome sequence of Bacillus atrophaeus UCMB-5137, a plant growth-promoting rhizobacterium," Genome Announc., vol. 1, 2013, pp. e00233-13, doi: 10.1128/genomeA.00233-13.

[3] C. Eymann, G. Homuth, C. Scharf, and M. Hecker, "Bacillus subtilis functional genomics : global characterization of the stringent response by proteome and transcriptome analysis," J. Bacteriol. vol. 184, 2002, pp. 2500-2520.

[4] J. W. Lengeler, G. Drews, and H. G. Schlegel, "Biology of Prokaryotes," Blackwell Science Ltd, Germany, 1999, pp. 652.

[5] C. W. Price, "Protective function and regulation of the general stress response in Bacillus subtilis and related gram positive bacteria," *In* G. Storz and R. Hengge-Aronis (ed), Bacteria stress responses. ASM press, Washington, D.C, 2000, pp. 179-197.

[6] R. B. Abramovitch, J. C. Anderson, and G. B. Martub, "Bacteria elicitation and evasion of plant innate immunity," Nat Rev Mol Cell Biol., vol. 7, 2006, pp. 601-611.

[7] J. L. Ramos et al. "The TetR family of transcriptional repressors," Microbiology and Molecular Biology Reviews, 2005, pp. 326–356, doi:10.1128/MMBR.69.2.326–356.2005.

[8] M. B. Miller and B. L. Bassler, "Quorum sensing in bacteria," Annu. Rev. Microbiol., vol. 55, 2001, pp. 165-99.

[9] Y. Lei, T. Oshima, N. Ogasawara, and S. Ishikawa, "Functional Analysis of the Protein Veg, which stimulates biofilm formation in Bacillus subtilis,"Journal of Bacteriology, vol. 195, 2013, pp. 1697-1705.

[10] B. E. Ramey, M. Koutsoudis, S. B. von Bodman, and C. Fuqua, "Biofilm formation in plant microbe associations," Current Opinion in Microbiology, vol. 7, 2004, pp. 602–609.

[11] K. Kierul et al. "Influence of root exudates on the extracellular proteome of the plant growth-promoting bacterium Bacillus amyloliquefaciens FZB42," Microbiology, vol. 161, 2015, pp. 131-147, doi: 10.1099/mic.0.083576-0.

[12] B. Fan, R. Borriss, W. Bleiss, and X. Wu, "Gram-positive rhizobacterium Bacillus amyloliquefaciens FZB42 colonizes three types of plants in different patterns," J. Microbiol., vol. 51, 2013, pp. 544, doi: 10.1007/s12275-012-1439-4.

[13] B. Fan et al. "Transcriptomic profiling of Bacillus amyloliquefaciens FZB42 in response to maize root exudates," BMC. Microbiol., vol. 12, 2012, pp. 116.

[14] Z. Wang , M. Gerstein, and M. Synder, "RNA-Seq : a revolutionary tool for transcriptomics," Nat Rev Genetics, vol. 10, 2009, pp. 57-63.

 ISBN: 978-1-61208-409-1

Automated Quantification of the Capacitance of Epithelial Cell Layers from an Impedance Spectrum

Thomas Schmid
Department of Computer Engineering
Universität Leipzig
Leipzig, Germany
schmid@informatik.uni-leipzig.de

Dorothee Günzel
Institute of Clinical Physiology
Charité
Berlin, Germany
dorothee.guenzel@charite.de

Martin Bogdan
Department of Computer Engineering
Universität Leipzig
Leipzig, Germany
bogdan@informatik.uni-leipzig.de

***Abstract*—Quantifying the intestinal surface area of epithelia is crucial to assess changes in protein expression during disease. A convenient alternative to microscopic evaluation of serial sections is capacitance measurement by impedance spectroscopy. While the underlying theoretical relations are well-known, in practice data scatter considerably decreases precision of estimations. Estimations are even less precise if obtained impedance spectra cannot be approximated by a semicircle. Here, we demonstrate that using machine learning techniques together with detailed modeling of cell layers allows reliable predictions of epithelial capacitance. Our results show that estimates for modeled impedance spectra can be obtained with less than 20 percent relative deviation from the target value. In particular, this is shown for spectra that deviate from a semicircular shape.**

Keywords–Physiology, Epithelia, Impedance Spectroscopy, Artificial Neural Networks, Clustering.

I. Introduction

The intestinal epithelium is the inner-most cell layer lining the gut wall and forms the primary barrier between the gut contents and the body. To maintain a tight barrier against toxins and pathogens, neighbouring epithelial cells are connected by tight junctions, arrays of transmembrane proteins that seal the space between two neighbouring cells. Acute (e.g., norovirus infection, giardiasis) and chronic intestinal diseases (e.g., Crohns disease, celiac disease) cause a restructuring of the gut mucosa due to loss of damaged surface cells and compensatory cell division within the crypts. Depending on the different rates, mucosal area may be enlarged [1] or reduced [2][3]. When investigating molecular processes underlying these diseases, an exact knowledge of changes in mucosal area is indispensable.

Commonly, mucosal area is determined morphometrically, i.e., by microscopic evaluation of serial sections, a process that is both tedious and time consuming. A much more elegant way is to determine the epithelial capacitance as a surrogate marker for the epithelial surface area. Capacitative properties of a cell are due to the lipid bilayer of the cell membrane. Capacitance of the unit cell membrane is in the order of 1 μF/cm^2 [4] and considered to be widely constant. In epithelial cells, the tight junction divides the plasma membrane into two compartments and as a consequence the total epithelial capacitance (C^{epi}) is composed of two capacitances in series. This subdivision is asymmetrical, as the tight junction is located close the apical side facing the outer environment. Intestinal epithelial cells are columnar (height $\gg$ diameter), therefore the apical membrane area is considerably smaller than the opposing basolateral membrane and C^{epi} is, as a first approximation, proportional to the apical membrane area.

A fast, convenient and noninvasive method to determine electrical properties of tissues is impedance spectroscopy. By measuring current-voltage relationships under alternate current (AC) at frequencies between 1 Hz and 100 kHz, typically 40 to 50 complex impedance values Z are obtained [5]. These spectra are often displayed in so-called Nyquist diagrams (Figure 1a), where the real part $\Re$ of each impedance value is plotted against its imaginary part $\Im$. To explain properties of the measured samples, it is common to derive an equivalent electric circuit. To describe epithelial cell layers, circuits of different degrees of complexity are used [6]. The simplest circuit that incorporates C^{epi} is a resistor-capacitor (RC) circuit (Figure 1b). To reflect physiological polarity of epithelial cells explicitly, two RC subcircuits in series and a resistor in parallel are used (Figure 1c). Electric behavior of the subepithelium may be considered by a further resistor in series.

In previous work, we have demonstrated that conventional analysis of impedance spectra like visual extrapolation of plots can lead to non-neglectable errors in parameter estimations. At the same time, we have shown that the precision in estimating epithelial and subepithelial resistance can be improved substantially for the epithelial cell lines HT-29/B6 and IPEC-J2 by using machine learning techniques [7][8]. Rationale behind this approach is that for a given electric circuit, the theoretical impedance at a given frequency can be calculated if the values of all circuit components are known. As exact target values are known for such synthetic data, too, this data can be used to train artificial neural networks or other machine learning techniques. In order to draw conclusions for data from laboratory measurements, however, optimal modeling of such training data according to the respective cell line and cell conditions is required.

Here, we adapt this approach to efficiently predict C^{epi} from an impedance spectrum. The tasks to be carried out include cell line modeling, feature selection, training and evaluation of the predictions. As model cell line, we investigate Madin-Darby canine kidney cells type I (MDCK-I). These cells have been studied since the 1960s [9] and are typically described as possessing a high transepithelial resistance [10].

 ISBN: 978-1-61208-409-1

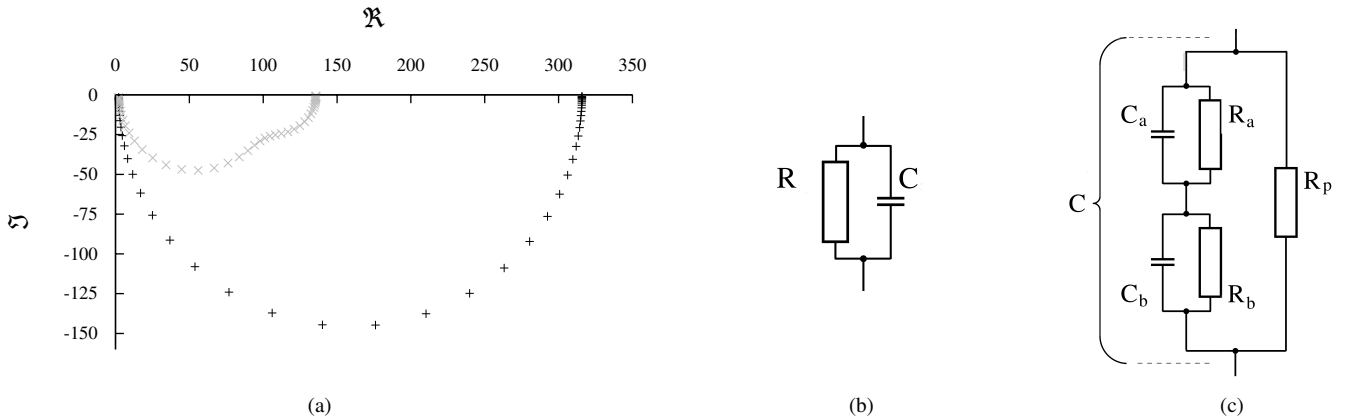

Figure 1. (a) Overlay of a semi- and a nonsemicircular Nyquist plot where real ($\Re$) and imaginary ($\Im$) part of each complex-valued impedance Z are plotted against each other. The displayed impedance spectra reflect AC application at n=42 frequencies between 1.3 and 16,000 Hz on an epithelial cell layer with capacitance $1/C = 1/C_a + 1/C_b$. (b) A simple resistor-capacitor (RC) circuit that can be used as equivalent circuit. This circuit yields semicircular impedance spectra. (c) Equivalent electric circuit discriminating between apical and basolateral properties of an epithelial cell layer. This circuit can yield semicircular or nonsemicircular impedance spectra.

II. Methods

A. Modeling Impedance Spectra

To model realistic impedance spectra for a given cell line, three prerequisites are required: a) a matching equivalent electric circuit, b) appropriate ranges for the parameters of the circuit, and c) an error model that reflects the data scatter intrinsic to the electrophysiological measurement set-up.

The equivalent circuit considered here (Figure 1c) consists of two RC subcircuits a (R_a, C_a) and b (R_b, C_b) located in series and a resistor in parallel (R_p). Using Kirchhoff's laws, the impedance Z of an electric circuit at an angular frequency ω can be derived from the impedances of its components:

$$Z(\omega) = \frac{R_p(R_a + R_b) + i\omega[R_p(R_a\tau_b + R_b\tau_a)]}{R_a + R_b + R_p(1 - \omega^2\tau_a\tau_b) + i\omega[R_p(\tau_a + \tau_b) + R_a\tau_b + R_b\tau_a]} \quad (1)$$

where $i = \sqrt{-1}$, and $\tau_a = R_aC_a$ and $\tau_b = R_bC_b$.

In practice, however, the electrophysiological set-up used for measurements induces data scatter and thus systematic deviation from the theoretical impedance value. In order to mimic realistic data, such systematic deviations can be modeled as function of the transepithelial resistance R^T [7] and added to $Z(\omega)$. For simplicity, such data scatter was not considered here.

Using $n = 42$ frequencies (1.3 to 16,350 Hz), n tupels of real and imaginary parts $((\Re(\omega_0), \Im(\omega_0)), \ldots, (\Re(\omega_{n-1}), \Im(\omega_{n-1})))$, are obtained from measurements or calculations, respectively. Alternatively, complex impedance values can be transformed into polar coordinates, i.e., into phase angle ϕ and magnitude r $(((\phi(\omega_0), r(\omega_0)), \ldots, (\phi(\omega_{n-1}), r(\omega_{n-1}))))$.

Real and imaginary parts of a spectrum can be regarded as separate feature sets $S_\Re$ and $S_\Im$:

$$S_\Re = \{\Re(\omega_0), \ldots, \Re(\omega_{n-1})\} \quad (2)$$

$$S_\Im = \{\Im(\omega_0), \ldots, \Im(\omega_{n-1})\} \quad (3)$$

Analogously for phase angles and magnitudes:

$$S_\phi = \{\phi(\omega_0), ..., \phi(\omega_{n-1})\} \quad (4)$$

$$S_r = \{r(\omega_0), ..., r(\omega_{n-1})\} \quad (5)$$

B. Sampling the MDCK-I Cell Line

For MDCK-I cells published values for the transepithelial resistance R^T range from 1500 to 14000 Ωcm^2 [11]. For the parameters R_p, R_a, R_b, C_a and C_b to the best of our knowledge published estimates exist neither for physiological conditions nor for drug applications. Therefore these parameter ranges were initially estimated from laboratory experiences and evaluated and optimized analogously to [7]. As two distinct cell conditions, physiological conditions ("Control") and conditions after the application of EGTA ("EGTA") were modeled. A table of the final parameters can be found in the appendix.

Estimating C^{epi} for semi-circular spectra is often considered a simple task with little error potential. At the same time, reliable estimation of C^{epi} for spectra deviating from this shape is thought to be considerably more difficult. Reassessing this assumption, we considered both cases individually and separated spectra reflecting control and EGTA conditions accordingly. As separation criterion, we assumed that spectra possessing greatly asymmetrical time constants express a nonsemicircular shape. To this end, we defined the τ ratio of the used electric circuit (Figure 1c) as the larger time constant divided by the smaller time constant, and a nonsemicircular shape was assumed for data with a τ ratio greater than five. This parameter can not only be calculated directly for modeled impedance spectra, but also be predicted with good precision for measured spectra [12].

C. Reference Methods to Determine Epithelial Capacity

Analogously to our previous work [7], we used two different conventional approaches as reference methods for estimating the parameter C^{epi}. Additionally, we employed a theoretical relation of the underlying circuit.

1) Nearest Data Point (Method M1): Assuming that a semicircular shape results from a single RC circuit (Figure 1b), $C^{epi} = 1/(\omega_c R)$ holds true. ω_c is the characteristic frequency, at which the spectrum reaches its minimal turning point. The frequency related the minimum of $S_\Im$ was used to approximate ω_c and the maximum value of $S_\Re$ to approximate $R = R^{epi}$.

 ISBN: 978-1-61208-409-1

2) Frequency-blind Circle Fit (Method M2): Analogously to M1, C^{epi} was calculated from the substitute parameters ω_c and R^{epi}. A Cole-Cole fit [13] was carried out on a Nyquist diagram, i.e., a circle was fitted as described by Kasa [14]. The frequency of the data point nearest to the circle center was used to approximate ω_c; the intercept with the x-axis at the low frequency end was taken as R^{epi}.

3) High-Frequency Limit Approximation (Method M3): Given the electric circuit in Figure 1b or 1c, respectively, the theoretical high-frequency limit for the imaginary part of the impedance is the reciprocal of the overall capacitance [15]:

$$-\lim_{\omega\to\infty}\omega\Im(\omega)=\frac{1}{C} \tag{6}$$

Of all 42 impedances obtained here, the data point with the highest frequency was $Z_{42}=\Re(\omega_{42})-i\Im(\omega_{42})$. Thus, we used the value given by $-1/(\omega_{42}\Im(\omega_{42}))$ to approximate C^{epi}.

D. Machine Learning Approach

The given reference methods represent two distinct estimation approaches: solving either a primarily geometric fitting problem (M2) or an idealized physics formula with error-prone data (M3). Therefore, we investigated these two representations of the same problem by two individual machine learning approaches. Also, we considered semicircular and nonsemicircular spectra as separate problem domains. For each representation and domain, we assessed the prediction quality of the respective problem representation by decision trees (using *R* and the package *rpart*), artificial neural networks with backpropagation (using the FORWISS Artificial Neural Network Toolbox [16]) and random forests (using *R* and the package *randomForest* [17]).

1) Training data: For the semicircular, as well as for the nonsemicircular domain, a sample of 30,000 random spectra was selected each. Data for each domain was split into a training dataset of 20,000 spectra (circa 66 percent) and a test dataset of 10,000 spectra (circa 33 percent), respectively.

As geometric data representations, either cartesian ($\perp$) or polar ($\angle$) coordinates were used where

$$S_{\perp}=\{\Re_0,...,\Re_{n-1},\Im_0,...,\Im_{n-1}\} \tag{7}$$

$$S_{\angle}=\{r_0,...,r_{n-1},\phi_0,...,\phi_{n-1}\} \tag{8}$$

As representation related to the high-frequency limit of the imaginary part, we transformed the spectra into the preciprocal products of frequencies and imaginary parts (cf. (6)):

$$S_{\omega\Im}=\left\{-\frac{1}{\omega_0\Im(\omega_0)},...,-\frac{1}{\omega_{n-1}\Im(\omega_{n-1})}\right\} \tag{9}$$

Note that while polar and cartesian representations possessed a total of $2\cdot n$ features, $S_{\omega\Im}$ possesses only n features.

2) Algorithm settings: Given a multivariate regression task, decision trees were created using analysis of variance ("anova" method). As ANNs, multilayer perceptrons (MLP) with one hidden layer were used. Depending on the number of input features, a $2n$-20-1 or n-10-1 architecture was used where hidden units employed sigmoid and input and output units linear activation functions; as learning algorithm Quickprop [18] was used. For random forests, 50 trees were used and variable importance was assessed by 25 consecutive runs of the *Boruta* algorithm (using the *R* package *Boruta* which searches all relevant variables by iterative removal of features that are statistically less relevant than random probes [19]). For evaluation, only test data was used. As exact target values were known, predictions were evaluated by relative deviation from the target, i.e., in percent.

3) Clustering: Using the best performing data representation, the sample was clustered by *k-means* where $k=\{1,..,10\}$; semicircular und nonsemicircular spectra were clustered separately. For each clustering, individual clusters were evaluated by decision trees. The best clustering for semi- and nonsemicircular data was determined by the least predictive cluster of each clustering and re-evaluated with ANNs.

III. Results

A. Evaluation of Reference Methods

Estimations of C^{epi} for nonsemicircular spectra showed in general greater deviations from the target value than those for semicircular spectra. For semicircular spectra, M2 showed least maximum deviations, while M3 showed least interquartile distance (Figure 2). For nonsemicircular spectra, M3 showed both least maximum deviations and least interquartile distance (Figure 3). Numerically, the maximum deviations of M3 was 90% for semicircular and 222% for nonsemicircular spectra.

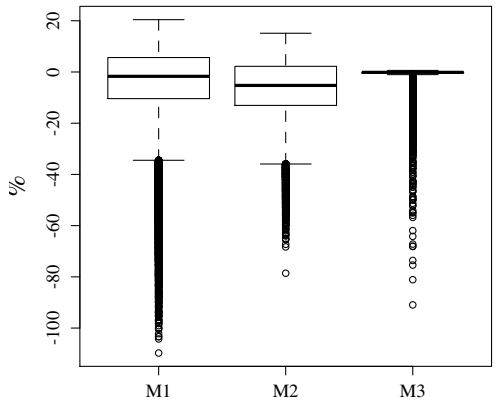

Figure 2. Relative deviation from the target value using reference methods M1, M2 and M3 for semicircular test data (n=30,000).

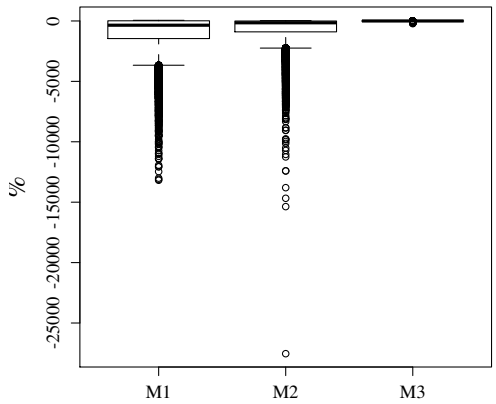

Figure 3. Relative deviation from the target using reference methods M1, M2 and M3 value for nonsemicircular test data (n=30,000).

B. Predictions of C^{epi}

Application of decision trees yielded maximum relative deviations from the target between 42 and 80 percent for semicircular and between 80 and 229 percent for nonsemicircular spectra. Application of ANNs yielded maximum relative deviations between 23 and 40 for semicircular and between 37 and 167 percent for nonsemicircular spectra. Application of random forests yielded constantly maximum relative deviations larger than the largest maximum relative deviation observed for decision trees; these results are therefore omitted here.

TABLE I. RELATIVE DEVIATION OF PREDICTIONS FROM THE TARGET VALUE C^{epi} FOR SEMICIRCULAR SPECTRA (IN PERCENT).

	cartesian		polar		high-frequency limit	
	tree	ANN	tree	ANN	tree	ANN
Minimum	-15.4	-39.1	-14.7	-40.3	-15.1	-23.5
1. Quartile	-4.6	-1.0	-4.1	-0.8	-4.2	-0.2
Median	0.1	0.1	0.4	0.1	0.3	-0.1
Mean	0.5	0.0	0.6	0.0	0.7	-0.1
3. Quartile	4.4	1.0	4.3	0.8	4.4	0.1
Maximum	80.7	29.9	42.1	30.7	80.6	11.4

TABLE II. RELATIVE DEVIATION OF PREDICTIONS FROM THE TARGET VALUE C^{epi} FOR NONSEMICIRCULAR SPECTRA (IN PERCENT).

	cartesian		polar		high-frequency limit	
	tree	ANN	tree	ANN	tree	ANN
Minimum	-22.5	-68.6	-20.0	-70.5	-22.2	-37.3
1. Quartile	-5.3	-1.7	-4.1	-1.6	-5.4	-0.1
Median	0.1	-0.2	0.1	-0.2	0.0	0.0
Mean	1.2	-0.1	0.6	-0.1	1.4	0.0
3. Quartile	4.8	1.7	4.4	1.4	5.0	- 0.2
Maximum	229.4	28.1	80.8	36.6	167.1	21.6

C. Variable Importance

Assessing the three data representations $S_{\perp}$, $S_{\angle}$ and $S_{\omega\Im}$ using the *Boruta* algorithm, neither for the semicircular nor for nonsemicircular domain a relevant number of features was removed. For the high-frequency limit representation $S_{\omega\Im}$ of the semicircular domain, e.g., only one feature was removed while 41 features were kept as relevant. Consequently, these findings were not used for explicit feature selection.

In all three data representations, however, analysis of variable importance showed that features reflecting the highest five frequencies yielded higher variable importance than the remaining features. This was observed for the semicircular, as well as for the nonsemicircular domain and was exploited in the next step of the analysis.

D. Cluster evaluation

K-means clustering was applied either to the full (84 or 42 features, respectively) or partial data representation (five features). When evaluating with decision trees, most clusterings did not yield less maximum deviations from the target value C^{epi} than seen in previous evaluations (Table I and II). An exception was the $S_{\omega\Im}$ representation clustered by the five features related to the five highest frequencies. For a number of five clusters, the highest maximum deviation observed for all clusters was 30.1 percent for semicircular spectra (Figure 4) and 58.8 percent for nonsemicircular spectra (Figure 5).

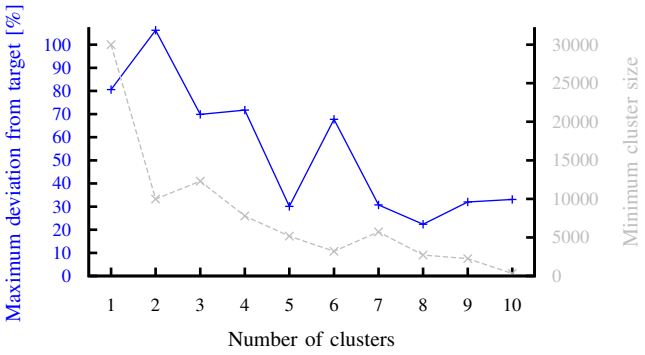

Figure 4. Cluster analysis for semicircular test data split into a variable number of clusters by k-means.

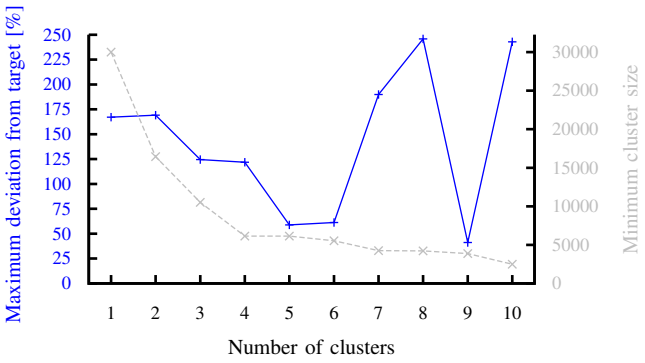

Figure 5. Cluster analysis for nonsemicircular test data split into a variable number of clusters by k-means.

E. Cluster-based estimations

Employing the findings from the cluster analysis, we split the $S_{\omega\Im}$ representation of the sample data into five clusters (cf. section III.D). For each cluster, an individual ANN was trained (analogously to section II.D). Relative deviations of the predictions did not exceed ±15.9 percent for any cluster of semicircular spectra (Table III), and did not exceed ±18.1 percent for nonsemicircular spectra (Table IV).

TABLE III. RELATIVE DEVIATION OF PREDICTIONS FROM C^{epi} FOR CLUSTERED SEMICIRCULAR SPECTRA (IN PERCENT).

Cluster	1	2	3	4	5
Minimum	-5.1	-15.9	-16.4	-3.3	-2.7
1. Quartile	-0.1	-0.1	-0.2	-0.1	-0.2
Median	0.0	0.0	0.0	0.0	0.0
Mean	0.0	0.0	0.0	0.0	0.0
3. Quartile	0.1	0.1	0.2	0.1	0.2
Maximum	8.9	6.3	6.2	8.3	2.3
$n_{cluster}$	7413	4526	7659	5240	5162

TABLE IV. RELATIVE DEVIATION OF PREDICTIONS FROM C^{epi} FOR CLUSTERED NONSEMICIRCULAR SPECTRA (IN PERCENT).

Cluster	1	2	3	4	5
Minimum	-9.3	-18.1	-7.4	-8.2	-11.3
1. Quartile	-0.1	-0.1	-0.2	-0.2	-0.1
Median	0.0	0.0	0.0	0.0	0.0
Mean	0.0	0.0	0.0	0.0	0.0
3. Quartile	0.1	0.2	0.2	0.2	0.1
Maximum	9.6	15.0	7.9	10.8	5.3
$n_{cluster}$	7191	7190	2969	6518	6132

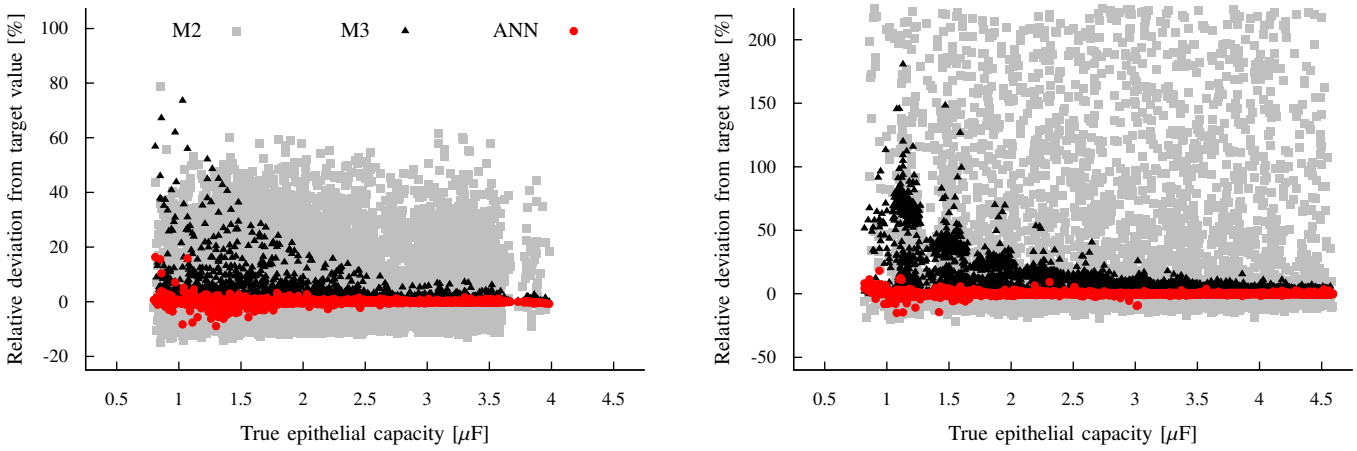

Figure 6. Relative deviations of estimated from true values for C^{epi} using reference methods M2 (grey ■) and M3 (black ▲) and a cluster-based ANN approch (red ●) are plotted against the respective true target value; method M1 is omitted as estimations are considerably less precise than those by M2 and M3 (cf. Figures 2 and 3). Values shown refer to (*a*) semicircular and (*b*) nonsemicircular spectra obtained either under physiological conditions (control) or after application of EGTA; to discriminate semi- from nonsemicircular spectra the relation between apical and basolateral time constants was used, i.e., a value greater than five was taken as indicator of nonsemicircular shape.

IV. Discussion

A. Evaluation of Reference Methods

As expected, the naive reference methods M1 and M2 failed to provide precise estimations for C^{epi} on nonsemicircular spectra. Interestingly, however, estimations can also exhibit large relative deviations of more than 50 percent from the target value when applied to semicircular spectra. And while method M3 is also a rather rough estimation method with up to 80 percent relative deviation in the same task, less than 5 percent for quartiles of relative deviations indicate a remarkable specificity compared to M1 and M2. This becomes even more obvious with differences in prediction quality for the nonsemicircular spectra. In both domains, however, maximum relative deviations of M3 are by far too great to allow reliable predictions.

B. Comparison of Problem Representations

In the first step of this study, we compared learning from geometric data representations $S_{\perp}$ and $S_{\angle}$ to learning from the physical data representation $S_{\omega\Im}$. ANNs yielded better predictions than decision trees in both representations, as well as in both problem domains. In the semicircular, as well as in the nonsemicircular domain, median and average relative deviations of all three representations did not exceed one percent of the target value (in one case: 1.2 percent). On the maximum relative deviation, however, ANNs using $S_{\omega\Im}$ performed better than ANNs using $S_{\perp}$ and $S_{\angle}$. Usefulness of this representation is not surprising considering that estimations with reference method M3 already exhibited little inter-quartile distance. It is likely that this excellent performance is due to the immediate physical relation between the features of $S_{\omega\Im}$ and C^{epi}.

C. Clustering

In the second step of this study, we aimed to optimize ANN predictions based on the physical representation $S_{\omega\Im}$. As optimization criterion, we considered the predictiveness of the least predictive cluster, respectively; this was measured by decision trees and the relative deviation of predicted values form the target value. At the same time, the number of spectra per cluster was intended to be as large as possible; naturally, the number of spectra decreases with increase of the number of clusters. As can be seen in Figures 4 and 5, both goals are achieved by choosing *k-means* clustering with $k = 5$; in particular, this holds true for both semicircular and nonsemicircular spectra.

D. Cluster-based Estimations

While clustering analysis was carried out with decision trees, the optimal clustering ($k = 5$) was evaluated by ANNs afterwards to further improve predictions. As in the first step, evaluations were performed for semicircular and nonsemicircular spectra separately. ANN estimations yielded maximum relative deviations of less than 20 percent within all clusters. Moreover, in all cases median and average relative deviations were 0 percent and inter-quartile distance less than 0.5 percent points. Compared to ANN predictions on unclustered data (Tab. I and II), this is a notable improvement. Even more remarkable is the improvement compared to the two best performing reference methods M2 and M3 (Figure 6).

V. Conclusions

Impedance spectroscopy is a convenient method to determine the capacitance of an epithelial tissue. In practice, however, this clinically important parameter can only be roughly approximated from impedance data, as common estimation methods fail to provide reliable estimations. Here, we have shown that our approach of modeling cell properties and applying machine learning techniques is a fruitful approach for this task. For impedance spectra modeled after the epithelial cell line MDCK-I, we developed a cluster-based neural network approach that shows a maximum relative deviation from the

 ISBN: 978-1-61208-409-1

theoretical target of less than 20 percent. In future work, we will apply this approch to other epithelial cell lines, as well as native tissue and further optimize estimations.

Appendix

Impedance spectra for the cell line MDCK-I were calculated for two distinct cell states. Physiological conditions (control) and conditions after application of EGTA (EGTA) were modeled separately according to (1) using the parameter ranges in Table I. Note that for the given electric circuit $R_t = R_a + R_b$ and $C^{epi} = \frac{C_a \cdot C_b}{C_a + C_b}$.

The parameter interval for R^{epi} was 10 to 2,000 Ωcm^2 and 10 to 200 Ωcm^2 for control and EGTA, respectively; all other intervals were chosen dynamically to yield ten values per range. By this, a total of 1,865,823 and 1,684,784 spectra were produced for control and EGTA, respectively.

TABLE V. Parameter ranges for a MDCK-I-equivalent circuit.

	R^{epi}	R_p [Ωcm^2]	R_t	C^{epi}	C_a [$\mu F/cm^2$]	C_b
Control	10–2000	10–10000	10–5000	0.5–5.0	1–5	1–75
EGTA	10–200	10–250	10–5000	0.5–5.0	1–5	1–75

From all of these spectra, 137,162 possessed a τ quotient less than five ($\sim$ semicircular) and 3,413,445 possessed a τ quotient larger than five ($\sim$ nonsemicircular). From these, a sample of 45,000 semicircular and a sample of 45,000 nonsemicircular spectra were randomly selected and analyzed.

To confirm correctness of our model, congruency with impedance measurements from laboratory experiments on MDCK-I cells was evaluated graphically as previously described [7]. During these experiments, R^{para} had been manipulated by the application of EGTA; 56 spectra were recorded before EGTA application, 49 after application. As reference data 25,000 modeled spectra from each condition were used.

References

[1] S. Zeissig et al., "Changes in expression and distribution of claudin 2, 5 and 8 lead to discontinuous tight junctions and barrier dysfunction in active Crohn's disease," Gut, vol. 56, no. 1, 2007, pp. 61–72.

[2] H. Troeger et al., "Structural and functional changes of the duodenum in human norovirus infection." Gut, vol. 58, no. 8, 2009, pp. 1070–1077.

[3] H. Troeger et al., "Effect of chronic Giardia lamblia infection on epithelial transport and barrier function in human duodenum," Gut, vol. 56, no. 3, 2007, pp. 328–335.

[4] H. Fricke, "The electric capacity of suspensions of red corpuscles of a dog," Physical Review, vol. 26, no. 5, 1925, pp. 682–687.

[5] J. R. Macdonald and W. B. Johnson, Fundamentals of Impedance Spectroscopy. John Wiley & Sons, Inc., 2005, pp. 1–26.

[6] D. Günzel et al., "From TER to trans-and paracellular resistance: lessons from impedance spectroscopy," Annals of the New York Academy of Sciences, vol. 1257, no. 1, 2012, pp. 142–151.

[7] T. Schmid, M. Bogdan, and D. Günzel, "Discerning apical and basolateral properties of HT-29/B6 and IPEC-j2 cell layers by impedance spectroscopy, mathematical modeling and machine learning," PLOS ONE, vol. 8, no. 7, 2013, p. e62913.

[8] T. Schmid, D. Günzel, and M. Bogdan, "Efficient prediction of x-axis intercepts of discrete impedance spectra," in Proceedings of the 21st European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2013, pp. 185–190.

[9] C. R. Gaush, W. L. Hard, and T. F. Smith, "Characterization of an established line of canine kidney cells (MDCK)." Proc Soc Exp Biol Med, vol. 122, no. 3, 1966, pp. 931–935.

[10] B. R. Stevenson, J. M. Anderson, D. A. Goodenough, and M. S. Mooseker, "Tight junction structure and ZO-1 content are identical in two strains of Madin-Darby canine kidney cells which differ in transepithelial resistance." J Cell Biol, vol. 107, no. 6 Pt 1, 1988, pp. 2401–2408.

[11] M. Furuse, K. Furuse, H. Sasaki, and S. Tsukita, "Conversion of zonulae occludentes from tight to leaky strand type by introducing claudin-2 into Madin-Darby canine kidney I cells," Journal of Cell Biology, vol. 153, no. 2, 2001, pp. 263–272.

[12] T. Schmid, D. Günzel, and M. Bogdan, "Automated quantification of the relation between resistor-capacitor subcircuits from an impedance spectrum," in Proceedings of the 7th International Conference on Bio-Inspired Systems and Signal Processing, 2014, pp. 141–148.

[13] K. S. Cole and R. H. Cole, "Dispersion and absorption in dielectrics I. alternating current characteristics," The Journal of Chemical Physics, vol. 9, no. 4, 1941, pp. 341–351.

[14] I. Kasa, "A circle fitting procedure and its error analysis," Instrumentation and Measurement, IEEE Transactions on, vol. 1001, no. 1, 1976, pp. 8–14.

[15] M. E. Orazem and B. Tribollet, Eds., Electrochemical Impedance Spectroscopy. Wiley, 2008, ch. Methods for Representing Impedance, pp. 309–331.

[16] M. K. Arras and K. Mohraz, FORWISS Artificial Neural Network Simulation Toolbox v.2.2, Bayerisches Forschungszentrum für wissensbasierte Systeme, Erlangen, Germany, 1996.

[17] A. Liaw and M. Wiener, "Classification and regression by randomforest," R news, vol. 2, no. 3, 2002, pp. 18–22.

[18] S. E. Fahlman, "Faster-learning variations on Back-propagation: An empirical study," in Proceedings of the 1988 Connectionist Summer School. San Mateo, CA: Morgan Kaufmann, 1988, pp. 38–51.

[19] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," Journal of Statistical Software, vol. 36, no. 11, 2010, pp. 1–13.

 ISBN: 978-1-61208-409-1

Protein-Protein Interaction in the Light of the Maximum Ordinality Principle

Corrado Giannantoni

ENEA's Researcher and Consultant of
Duchenne Parent Project Onlus
Rome, Italy
email: c.giannantoni@parentproject.it, corrado.giannantoni@tin.it

***Abstract* – The present paper is aimed at showing how it is possible to obtain Protein-Protein Interaction (PPI) in explicit formal terms, when this process is modeled by adopting the Maximum Ordinality Principle (MOP) as the basic reference criterion.**

Keywords-Protein-Protein Interaction (PPI); Maximum Ordinality Principle (MOP); Molecular Docking; Drug Design.

I. INTRODUCTION

Protein-Protein Interaction (PPI) decisively represents a fundamental process in Pharmacology. However, in spite of its recognized importance, PPI has not manifested all its related potentialities yet, mainly because of the intrinsic unsolvability, in explicit terms, of the famous "Three-body Problem", as demonstrated by H. Poincaré in 1889 [1].

This result represents a strong limitation, because it is also valid in *Protein Dynamics*. Not only (and especially) in Protein Folding, but also in PPI.

On the other hand, the research for a numerical solution often overcomes the computation capacities of the most powerful computers at present available (10 Petaflops).

Under such conditions, about 40 different approaches to PPI have been proposed in Literature. All of them, however, always introduce some (more or less) marked approximations. For example, the two interacting proteins are sometimes modeled as they were "rigid bodies". Consequently, apart from the solution to some particular (and specific) cases of PPI, such approximations do not always lead to satisfactory general solutions, when compared with experimental data.

This is because, in the absence of an explicit formal solution to PPI problem, all the various approaches adopted are also characterized by a correlative absence of an effective *predictive capacity*. In particular, when the latter is referred to the three-dimensional configuration of the final compound.

In the framework of such a "state of the art", the main aim of this paper is to bring out the possibility of obtaining Protein-Protein Interaction, in a *fast* and *reliable* way, as the *formal solution to an N-body interaction problem*, when the process is modeled on the basis of the Maximum Ordinality Principle (MOP).

This is because, after having obtained the solution to the "Three-body Problem" in terms of *fractional incipient derivatives* [2], previously introduced in [3][4], the extension to the case of N bodies was obtained in the contest of the mathematical formulation of the MOP [5].

This result immediately suggested its application to Protein Folding [6][7] and, now, to the case of PPI.

In this respect, Section 2 will preliminary present the input/out of the mathematical model adopted. Section 3 will illustrate the solution process through an ostensive example. The related informatics advances will be discussed in Section 4, while Section 5 will consider a possible extension of the same approach to other pharmacological fields. Section 6 (devoted to the conclusion) will reconsider all the previous aspects in the light the basic principles of self-organizing systems of *ordinal nature*.

II. INPUT/OUPUT OF THE MATHEMATICAL MODEL

The formal enunciation of the MOP, with specific reference to biological problems, was presented in [6][7]. Such a formulation is able to facilitate the solution to the PPI problem because, when the amino acids of a protein are modeled as they were "atoms" of a macromolecule, the 3D structure of the protein can be obtained without necessarily knowing its primary structure (that is, the specific linear sequence of its amino acids).

In fact, it is sufficient:

i) to know the *total number* of amino acids (*N*);

ii) to assign three parameters $(\Sigma_{12}, \Phi_{12}, \Theta_{12})$ that define, in polar coordinates, the reciprocal positions of two *arbitrary* amino acids, understood as being *one sole* "*isolated" entity*. This is also the reason why the latter is referred to its own internal reference system;

iii) to assign, in addition, six appropriate parameters $(\varepsilon_1, \varepsilon_2, \varepsilon_3, \psi_1, \psi_2, \psi_3)$, that define the internal *Relation Space* (RS) of the protein analyzed.

 ISBN: 978-1-61208-409-1

More specifically: $(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ characterize the spatial orientation of the protein (understood as a whole), with respect to its internal reference axes; whereas (ψ_1, ψ_2, ψ_3) define the periodicities (along the three basic axes) of the mathematical solutions which "emerge" from the MOP.

These solutions are precisely those that give the positions of all the amino acids with respect to the internal axes of the considered protein. In this way, the afore-mentioned solutions characterize any considered protein as a *unique*, *specific* and *irreducible* entity.

Under such conditions, each protein, precisely because modeled as a "self-organizing" system of *ordinal nature* (see Section 6), is also characterized by its own specific *self-organizing capacity*, whose *activity* can faithfully be represented by its associated "virtual work", defined (in polar coordinates) as

$$W = \sum_{j=2}^{N} \{(\rho_{1j}) + (\rho_{1j}\varphi_{1j}) + (\rho_{1j}\vartheta_{1j})\} \qquad (1)$$

where the subscript 1j indicates the couples of amino acids successively considered in the sum.

In the case of PPI, when there exists a given affinity between the interacting proteins, the resulting compound generally shows a "virtual work" W_3 that "exceeds" the sum of the "virtual works" W_1 and W_2 pertaining to the interacting proteins. Consequently, the *ratio* between such an excess of "virtual work"

$$\delta W = \{W_3 - (W_1 + W_2)\} \qquad (2)$$

and the sum of virtual works of the interacting proteins, that is

$$\delta W / (W_1 + W_2) \qquad (3)$$

can be assumed as a "measure" of the reciprocal *affinity* between the interacting proteins or, equivalently, as their *elective propensity* to realize a *stable* compound.

III. AN OSTENSIVE EXAMPLE

The example deals with diabetic therapy. It is well-known that human insulin has a reduced affinity with blood albumin, so that the subcutaneously injected insulin cannot efficiently be conveyed by blood albumin in the various parts of the body.

The therapy then consists in adopting a modified form of insulin, which presents a higher affinity with blood albumin. The modified form of insulin usually adopted is insulin detemir, also termed as levemir.

Figure 1 represents the three-dimensional structure of human insulin (51 amino acids), obtained by means of an appropriate simulator, run on a simple PC (10^9 Flops), in less than 1 s.

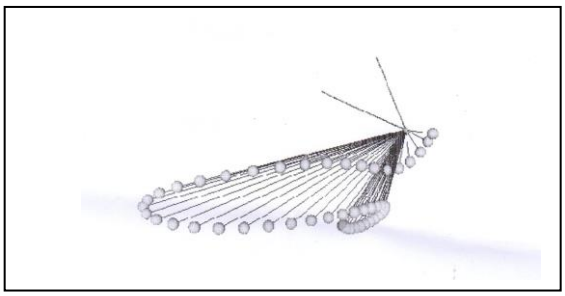

Figure 1 - Three-dimensional structure of human insulin (51 amino acids: 21 in subunit A and 30 in subunit B)

The simulator was termed as Emerging Quality Simulator (EQS), precisely because based on the MOP and its corresponding "emerging solutions" (see later on). The 3D structure so obtained can easily be modified (if needed) by means of slightly variations of some parameters of the RS.

This allows us to achieve a more accurate comparison, not only with the spatial configurations available in Literature (e.g., at the level of secondary structure), but also, and especially, with X-Ray Crystallography and/or Nuclear Magnetic Resonance (NMR) images available in qualified Protein Data Banks. This is because the output of the simulator, apart from the 3D structure, also gives the associated coordinates of all the amino acids, together with some other important indicators. Among others, and in particular, the corresponding "virtual work" associated to the protein.

Figure 2, in turn, represents the three-dimensional structure of blood albumin, made up of 585 amino acids. This spatial configuration was also obtained by means of the same simulator, run on the same PC, in a computation time of about 1 s.

As in the previous case, such a 3D structure can easily be compared with the corresponding spatial configurations available both in Literature and in Protein Data Banks.

At this stage, if we consider the interaction process between the two afore-mentioned proteins, we obtain that: i) insulin and albumin result as being characterized by virtual works whose values, expressed in the scale units usually adopted in EQS, are W_1 = 88.38 and W_2 = 587.66, respectively; ii) whereas the virtual work associated to the resulting compound is W_3 = 683.65.

Consequently, the corresponding ratio (3) gives

$$\delta W / W_3 = 0.0112 \qquad (4)$$

 ISBN: 978-1-61208-409-1

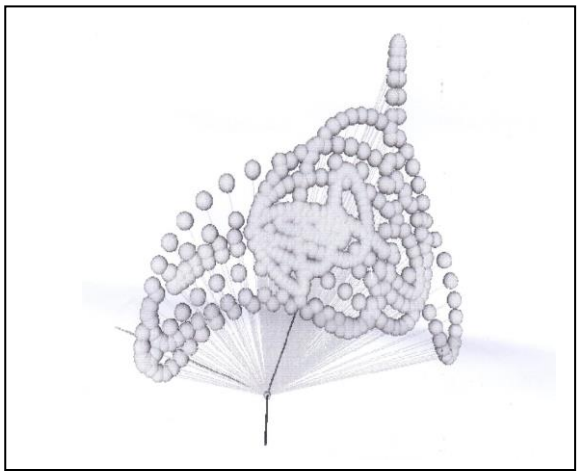

Figure 2 - Three-dimensional structure of blood albumin (585 amino acids)

This result clearly shows that human insulin has a very reduced affinity with albumin (about 1%). At the same time, it also explains why human albumin is usually modified in the form of levemir, in order to achieve a higher affinity.

Levemir insulin differs from human insulin in that the amino acid in position B30 is omitted, and a C14 fatty acid chain (termed as myristic acid) is attached to the amino acid B29.

Figure 3 represents the 3D structure of levemir, whose virtual work now results as being W_1^* = -29.95.

The *negative* value obtained simply indicates that the modified protein has an *inverse chirality* with respect to its primary form of insulin. This aspect generally favors the interaction process. In fact, the virtual work associated to the resulting compound now becomes W_3^* = 667.29.

Consequently, the interaction process between levemir and albumin (obtained by means of the same simulator in less than 2 s) gives origin to a final compound characterized by a higher "excess" of virtual work (2).

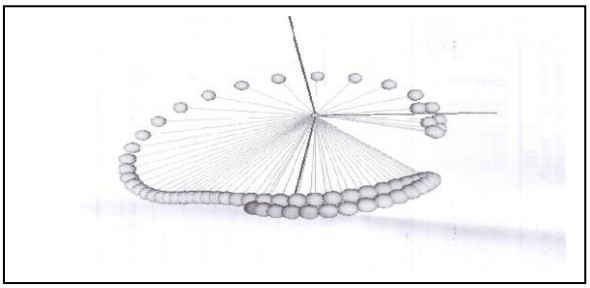

Figure 3 - Three-dimensional structure of levemir (50 amino acids plus the chain of 14 atoms of C)

Correspondingly, ratio (3) now gives

$$\{W_3^* - (W_1^* + W_2)\}/(W_1^* + W_2) = 0.1965 \quad (5)$$

This result clearly shows that such a modified form of human insulin presents an affinity of about 20% with respect to blood albumin. A value that allows levemir to be conveyed by albumin, without preventing, however, its subsequent release in the various parts of the body.

IV. INFORMATICS ADVANCES

The improvements here considered are directly referable to the *formal properties* that are intrinsic to the mathematical models adopted. In fact, any system modeled on the basis of the MOP, always presents *explicit solutions* in terms of *Incipient Differential Calculus* (see [3] and [4]).

This means that the method proposed has the *capacity* of *predicting the 3D structure of the resulting compound* essentially because the latter is understood as a self-organizing system of *ordinal nature*, and thus as *intrinsically* "*irreducible*" to functional relationships between its parts.

This correlatively also means: i) *a reduced number of computations*; ii) *a reduced need* of *High Performance Computing* (HPC); iii) *a reduced incidence* of *special numerical methods* to be adopted to get the corresponding solution.

What's more, the explicit solutions so obtained can also be termed as "*emerging solutions*" (see [5] and [8]), because *they always show an ordinal information content which is much higher than the corresponding content of the initial formulation of the problem.*

This is because the MOP is specifically finalized to describe "self-organizing" systems according to a *holistic approach*, in which, as is well-known, "*the whole is much more than the sum of its parts*".

V. BIO-INFORMATICS IN THE LIGHT OF THE MAXIMUM ORDINALITY PRINCIPLE

The method of solution previously illustrated with specific reference to Protein-Protein Interaction is also applicable to the majority of biological problems usually dealt with through informatics methods. In this sense, PPI only represents an ostensive example.

The same approach, in fact, has previously been adopted to improve the efficiency of the *exon skipping* method, usually used in Duchenne Muscular Dystrophy [9].

In such a case, the method enabled us to select the most appropriate Antisense Oligo-Nucleotides (AONs) with reference to four specific Exons: 51, 48, 44 and 39.

The pertinent experimental tests (in vitro and in vivo) are still in progress at LUMC (Leiden University Medical

 ISBN: 978-1-61208-409-1

Center) and the corresponding final results will be available at the end of next May.

This would indicate that the methodology here proposed could also be adopted in the case of Molecular Docking and Drug Design. In fact, it allows us to choose the *optimal ligand*, that is the one which is characterized by the most appropriate researched affinity (3), as in the case of exon skipping in DMD, previously mentioned.

Consequently, when considered from a more general point of view, the paper would intend to show that, in the light of the MOP, it is possible to realize mathematical models of several biological Systems, with very significant related advantages.

VI. CONCLUSION

The methodology here proposed seems to be able to give a significant contribution to Pharmacology. This is because the results previously shown indicate that the dynamic evolutions of a wide variety of biological processes can adequately be described by adopting the same reference principle (namely, the MOP).

Consequently, the various biological processes to be analyzed, when modeled by means an appropriate simulator, can be run on a simple personal computer and, in addition, in a computation time of a few seconds (or one minute, at the most).

This means that, by adopting the afore-mentioned approach, any researcher would be able to analyze the dynamic behavior of any biological process of interest by means of his/her own PC, simply sitting at his/her own desk.

The solutions obtained, in fact, will always describe a System whose parts are related to each other according to *ordinal relationships*. In other words, according to the same "relationships" that precisely take origin from *generative processes*, such as, for instance, the *genesis of two brothers*.

"Brothers", in fact, are properly defined as such, not because of their *direct* relationships. That is: because they respect each other or they love each (in fact, they might also hate each other). They are "brothers", *in essence*, because generated by the same father (or the same mother, or both). That is, because of their *direct relationship* with the *generative cause* of their being born.

Such a *genetic* relationship represents in fact something that is *unique*, *specific* and *irreducible*. Consequently, they cannot simply be accounted for as "two" (1+1), but as *one sole entity* (that is, as a whole), in spite of their clear reciprocal distinction. Consequently, the proper meaning of "*brothers*" refers to a clear "*irreducible extra*".

Precisely that represented by *their specific relationship* with the *same genetic* principle.

In accordance with such a concept of *ordinal relationship*, in the case of a given protein the *direct* relationship between two *any* amino acids is considered as being of the second order.

The *first order* relationship, in fact, is that which relates all the amino acids to the *same generative activity* of the protein, always understood as a *whole*.

This is why the explicit solutions that "emerge" from the MOP immediately give the positions of all the amino acids with respect to the internal axes of the protein (see Introduction, in particular, points i) and ii)).

The same concept is evidently valid for any *self-organizing* system, when described on the basis of the MOP.

The formal enunciation of this principle, in fact, first given in [5], is nothing but the reformulation of the Maximum Em-Power Principle, proposed by H.T. Odum in [10][11][12], understood as an updated version of the Fourth Thermodynamic Principle, first enunciated by Boltzmann and afterwards by Lotka, in [13] and [14], respectively.

Odum's enunciation, in fact, after having received an appropriate mathematical formulation *under dynamic conditions* in [15], was reformulated in more general *terms* in [5], by means of *a new concept of derivative*, the "*incipient*" *derivative*, whose mathematical definition was first introduced in [3] and further developed in [4].

The corresponding verbal enunciation of the MOP then became: "*Every system tends to maximize its own ordinality, including that of the surrounding habitat*".

REFERENCES

[1] H. Poincaré, "Les Méthodes Nouvelles de la Mécanique Céleste", 1889. Ed. Librerie Scientifique et Technique A. Blachard. Vol. I, II, III, Paris, 1987.

[2] C. Giannantoni, "From Transformity to Ordinality, or better: from Generative Transformity to Ordinal Generativity". Proceedings of the 5th Emergy Conference. Gainesville, Florida, USA, January 31-February 2, 2008, pp. 581-598.

[3] C. Giannantoni, "The Problem of the Initial Conditions and Their Physical Meaning in Linear Differential Equations of Fractional Order". Applied Mathematics and Computation 141, 2003, pp. 87-102.

[4] C. Giannantoni, "Mathematics for Generative Processes: Living and Non-Living Systems". Applied Mathematics and Computation 189, 2006, pp. 324-340.

[5] C. Giannantoni, "The Maximum Ordinality Principle. A Harmonious Dissonance". Proceedings of the 6th Emergy Conference. Gainesville, USA, January 14-16, 2010, pp. 55-72.

[6] C. Giannantoni, "Protein Folding, Molecular Docking, Drug Design. The Role of the Derivative "Drift" in Complex Systems Dynamics". Proceedings of the Third International Conference on Bioinformatics. Valencia, Spain, January 20-24, 2010, pp. 193-199.

 ISBN: 978-1-61208-409-1

[7] C. Giannantoni, "Bio-Informatics in the Light of the Maximum Ordinality Principle. The Case of Duchenne Muscular Dystrophy". Proceedings of the 4th International Conference on Bioinformatics. Rome, January 26-29, 2011, pp. 244-250.

[8] C. Giannantoni, "The Relevance of Emerging Solutions for Thinking, Decision Making and Acting. The case of Smart Grids". Ecological Modelling 271, 2014, pp. 62-71.

[9] A. Aartsma-Rus, W. E. Kaman, R. Weij, J. T. Den Dunnen, G. J. van Ommen, J. C. van Deutekom, "Exploring the Frontiers of Therapeutic Exon Skipping for Duchenne Muscular Dystrophy by Double Targeting within One or Multiple Exons". Molecular Therapy, Vol. 14, 2006, pp. 401-407.

[10] H. T. Odum, "Ecological and General Systems. An Introduction to Systems Ecology". Re. Edition. University Press Colorado, 1994.

[11] H. T. Odum, "Environmental Accounting". Environmental Engineering Sciences. University of Florida, 1994.

[12] H. T. Odum, "Self Organization and Maximum Power". Environ. Engineering Sciences. University of Florida, 1994.

[13] L. Boltzmann, "Der zweite Haupsatz der mechanishen warme Theorie". Almanach der K. Acad. Wiss Mechanishe, Wien, 1905, Vol. 36, pp. 225-299 (printing of a lecture given by Boltzman in 1886).

[14] A. J. Lotka, "The Law of Evolution as a Maximal Principle". Human Biology, a record of research. September 1945, Vol. 17, n. 3.

[15] C. Giannantoni, "The Maximum Em-Power Principle as the basis for Thermodynamics of Quality". Ed. S.G.E., Padua, 2002, ISBN 88-86281-76-5.

 ISBN: 978-1-61208-409-1