BIOTECHNO 2013

The Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-260-8

March 24 - 29, 2013

Lisbon, Portugal

BIOTECHNO 2013 Editors

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

Petre Dini, Concordia University, Canada / China Space Agency Center, China

BIOTECHNO 2013

Foreword

The Fifth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies [BIOTECHNO 2013], held between March 24 - 29, 2013 in Lisbon, Portugal, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

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

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, bio-nanotechnologies, drugs, or microbial propagation and immunity.

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2013 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 who dedicated much of their time and efforts to contribute to

BIOTECHNO 2013. 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 are grateful to the members of the BIOTECHNO 2013 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2013 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the fields of bioinformatics, biocomputational systems and biotechnologies.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the charm of Lisbon, Portugal.

BIOTECHNO 2013 Chairs:

Stephen Anthony, The University of New South Wales, Australia
Petre Dini, Concordia University, Canada / China Space Agency Center-Beijing, China

BIOTECHNO 2013

Committee

BIOTECHNO Advisory Chairs

Stephen Anthony, The University of New South Wales, Australia
Petre Dini, Concordia University, Canada / China Space Agency Center-Beijing, China

BIOTECHNO 2013 Technical Program Committee

Basim Alhadidi, Albalqa' Applied University - Salt, Jordan
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
Yang Cao, Virginia Tech – Blacksburg, USA
Cesar German Castellanos Dominguez, Universidad Nacional de Colombia - Manizales,Colombia
Yili Chen, Monsanto Company - St. Louis, USA
Eugen Czeizler, Aalto University,Finland
Rolf Drechsler, DFKI Bremen || University of Bremen, Germany
Lingke Fan, University Hospitals of Leicester NHS Trust, UK
Victor Felea, "Al.I. Cuza" University - Iasi, Romania
Jerome Feret, INRIA, France
Xin Gao, KAUST (King Abdullah University of Science and Technology), Saudi Arabia
Alejandro Giorgetti, University of Verona, Italy
Paul Gordon, University of Calgary, Canada
Radu Grosu, Vienna University of Technology, Austria
Jun-Tao Guo, The University of North Carolina at Charlotte, USA
Mahmoudi Hacene, University Hassiba Ben Bouali – Chlef, Algeria
Saman Kumara Halgamuge, University of Melbourne, Australia
Steffen Heber, North Carolina State University-Raleigh, USA
Elme Huang, Peking University, China
Asier Ibeas, Universitat Autònoma de Barcelona, Spain
Attila Kertesz-Farkas, International Centre for Genetic Engineering and Biotechnology, Trieste, Italy
Daisuke Kihara, Purdue University - West Lafayette, USA
DaeEun Kim, Yonsei University - Seoul, South Korea
Fatih Kurugollu, Queen's University - Belfast, UK
José Luis Oliveira, University of Aveiro, Portugal
Roger Mailler, The University of Tulsa, USA
Igor V. Maslov, EvoCo Inc. - Tokyo, Japan
Bud Mishra, NYU, USA

Copyright Information

Table of Contents

Assessment of Physiological States of Plants *in situ*

An Innovative Approach to the use of Electrical Impedance Spectroscopy

Elisabeth Borges, Mariana Sequeira, André F. V. Cortez, Helena Catarina Pereira, Tânia Pereira, Vânia Almeida, João Cardoso and Carlos Correia
Physics Department of the University of Coimbra
Instrumentation Center
Coimbra, Portugal
eborgesf@gmail.com

Teresa M. Vasconcelos, Isabel M. Duarte and Neusa Nazaré
Escola Superior Agrária de Coimbra of the Instituto Politécnico de Coimbra
Centro de Estudos de Recursos Naturais Ambiente e Sociedade
Coimbra, Portugal
tvasconcelos@esac.pt

***Abstract*— The fast spread of asymptomatic forest diseases, with no cure available to date, constitute a problem of economical and forestall huge proportions. Furthermore, there is a lack of equipments and systems able of assessing and characterizing the physiological state of plant organisms, both in the diagnosis of disease and as a mean for supporting physiological studies. It is known that electrical impedance measurements have been proving its value in the characterization of vegetal tissues. However, the available commercial solutions are expensive, unfeasible for *in vivo* and *in situ* measurements and unspecific for biological applications. Herein is proposed the usage of impedance techniques to assess the physiological state of plants. Emphasis is given to the assessment of the hydric stress level of plants and its relation with the disease condition. To accomplish the study, a portable electrical impedance spectroscopy system was designed attending the biological application purpose. In order to show the potential of this technique and system, the procedure and the results obtained for three different species (*Pinus pinaster* Aiton, *Castanea sativa* Mill and *Jatropha Curcas* L), with economical and/or forestall interest, is also presented.**

Keywords-plant disease; physiological state; hydric stress; biodiesel; impedance techniques

I. Introduction

Plant organisms, being a plant organism any kind of living plant (tree, bush, or other), are affected by numerous diseases promoted by biological agents (such as fungus, virus, bacteria, nematodes, insects, and others) and/or inhospitable environmental conditions (such as drought, fires, extreme heat, contamination of soil and air, and others) [1]. It is worth adding that the knowledge of the health state of the plant organisms is important, particularly when diseases affect crops with economic and/or forestall impact. Currently there are some diseases strongly affecting specific crops with significant economical relevance in specific countries or regions. Such cases include, for instance, the pinewood nematode, affecting mostly the *Pinus pinaster* Aiton specie, the ink disease in the chestnuts and the esca disease in the grapevines. Whether these diseases are caused by fungus, nematodes or other biotic or abiotic agents, they are mostly asymptomatic, exhibit fast spread rate and currently have no cure properly developed and commercialized [1, 2].

The standard method to diagnose diseases in plant organisms is the symptomatology visualization by skilled personnel [1, 2]. However, usually, the external symptoms are only able of being visually accessible during the terminal stages of the diseases [1]. The plant organism that is considered affected is cut and destroyed. In order to avoid the fast spreading, the neighbor plant organisms are also cut and burned, even if visual symptoms are not accessible [2]. This preventive act poses a problem: the deforestation, and the resulting economic losses, caused by the massif felling.

In a similar way, but in the perspective of marketing and consumption, the characterization of plant organisms, which are important in processes like, for instance, the production of biodiesel [3] and the physiological studies of plant organisms for new applications [3, 4], also lacks a detailed and similar technical analysis [4]. The techniques available for characterizing plant organisms require expensive laboratory equipments and materials, are time consuming and hard to implement [1, 5].

Hereupon, it can be said that, in general, there is a lack of equipments and systems able of assessing and characterizing the physiological state of plant organisms. This overall described panorama motivated the present work. Herein, the authors propose an Electrical Impedance Spectroscopy, EIS, system and the usage of impedance techniques to assess the physiological states of plant organisms. Emphasis was giving to the assessment of the hydric stress level and its relation with the disease condition.

The hydric stress refers the internal hydration condition of a plant organism and it is one of the most relevant parameters to assess physiological states [5]. This parameter takes special significance in the assessment of diseases, since water absorption by the plant is one of the physiological processes firstly and strongly affected during a biotic or abiotic disease condition [5].

ISBN: 978-1-61208-260-8

EIS has been proving efficacy and utility in a wide range of areas, from the characterization of biological tissues to living organisms [6]. The electrical impedance of a biological material is a passive electrical property that measures the opposition relatively to an alternating current flow applied by an external electric field. The current I, as it passes across a section of a material of impedance Z, drops the voltage V, established between two given points of the same section, yielding the well-known generalized Ohm's law: V=IZ, where V and I are complex scalars and Z is the complex impedance. Hence, the result of the EIS measurements is a set of complex (magnitude and phase) of impedance versus frequency.

Cell membranes, intracellular fluid (cytosol) and extracellular fluid are the major contributors of the impedance of biological tissues [6]. A commonly used circuit to represent biological tissues consists of a parallel arrangement between a resistor, simulating the extracellular fluid, and a second serial arrangement connecting a resistor, this one of the cytosol, and a capacitor, of the membrane [7, 8].

The model commonly used to represent impedance values is the Cole bioimpedance model, in which the bioimpedance spectrum is represented by means of a Cole-Cole plot that explores resistance versus reactance, allowing the determination of the ohmic values of the cytosol and the extracellular fluid [9].

The physiological changes, due to diseases and nutritional or hydration levels, have direct influence in the impedance spectrum. The phase angle and other interrelated indices, such as Z_0/Z_∞ [6] and Z_0/Z_{50} [10], have been used to extract information about the physiological condition of biological materials.

The nature of the impedance excitation signal varies depending on the application. It is possible to excite the sample with a current and measure a voltage or to do the exact opposite. The discussion on what source, voltage or current, is the most convenient remains. Current sources, CS, provide suitably controlled means of current injection [11] and present reduced noise due to spatial variation when compared with voltage sources, VS [12]. However, CS accuracy decreases with high frequency [13], especially due to their output impedance degradation [12]. Since the impedance measurements are limited to field strength where the current is linear with respect to the voltage applied [8], or vice-versa, CS need high-precision components [14] and a limited bandwidth operation range [13, 14] to overcome the stated limitation. On the other hand, VS, although producing less optimal electrical impedance spectroscopy, EIS, systems [14], can operate over a sufficient broad frequency range [13, 14] and are built with less expensive components [14].

Nowadays, instruments with high precision, high resolution and frequency ranges extending from some Hz to tens of MHz are commercially available [6]. However, in what concerns to the range of low or high frequencies (already above 100 kHz), the degradation of the excitation signal affects the accuracy of the measurements [6]. Besides, the typical solutions consist in impedance analyzers and LCR meters which are desktop instruments [6], unfeasible for in vivo [6] and in field applications.

The EIS system presented herein is able to perform AC scans within a selectable frequency range. The system implements the phase sensitive detection, PSD, method and can drive either a current or a voltage signal to excite a biological sample *in situ* or *in vivo*. The instrumentation was designed to be cost-effective and usable in several applications.

TABLE I. SUMMARY OF SPECIFICATIONS OF THE EIS SYSTEM

	Range	
Parameter	***Current Mode***	***Voltage Mode***
Measuring method	2 electrodes	
Frequency	1 kHz to 1 MHz	
Signal amplitude	25 uA	4.6 V
Impedance magnitude	100 Ω to 100 kΩ 1	1.5k Ω to 2.2 MΩ1
Impedance phase	-π rad to π rad	-π rad to π rad
Mean absolute magnitude error	1675.45 Ω	709.37 Ω
Mean absolute phase error	2.45 %	2.06 %
Mean distortion	0.29 %	0.48 %
Mean SNR	117.0 dB	118.8 dB
Calibration	Automatically calibrated by software	

In this paper it is also resumed the most relevant studies obtained for three different plant species, with economical and/or forestall impact: pine *(Pinus pinaster* Aiton), cheastnut (*Castanea sativa* Mill) and *Jatropha Curcas* L.

The following sections of this paper are: 1) *System Design*, where the developed EIS system is presented in detail; 2) *Assessment of the Hydric Stress Level*, which presents the method and the results obtained for the hydric stress assessment for the three studied species; 3) *Study of Disease Condition*, which presents the method and the results obtained for the study of the nematode disease in *Pinus pinaster* specie; and 4) *Conclusions*, which resumes the main obtained results.

II. SYSTEM DESIGN

A. General Description

The developed EIS system employs two electrodes and consists of three main modules: signal conditioning unit, acquisition system (PicoScope® 3205A) and a laptop for data processing (Matlab® based software).

The electrodes being used are beryllium cooper gold platted needles with around 1.02 mm in diameter. The bioimpedance measurement requires the most superficial possible penetration of the electrodes in order to reduce the dispersion of the needles surface current density [9], and also to reduce damage on the biologic sample.

The digital oscilloscope PicoScope® 3205A has dual functionality: 1) synthesizes and provides the excitation AC signal to the conditioning unit (ADC function); 2) digitizes both excitation and induction signals at high sampling rates (12.5 MSps) and transfers data to the computer via USB where it is stored. The signal conditioning unit receives the exciting AC signal, coming from the PicoScope®, and amplifies it to be applied, through an electrode, to the specimen under study. The induced AC signal is collected by a second electrode and is redirected to the conditioning unit where it is also amplified. Both excitation and induced signals are conduced to the PicoScope® to be digitized.

The features of both excitation modes are described below.

B. Design Specifications

The current mode circuit employs the current-feedback amplifier AD844 in a non-inverting ac-coupled CS configuration (see Figure 1), already studied by Seoane, Bragós and Lindecrantz, 2006 [15].

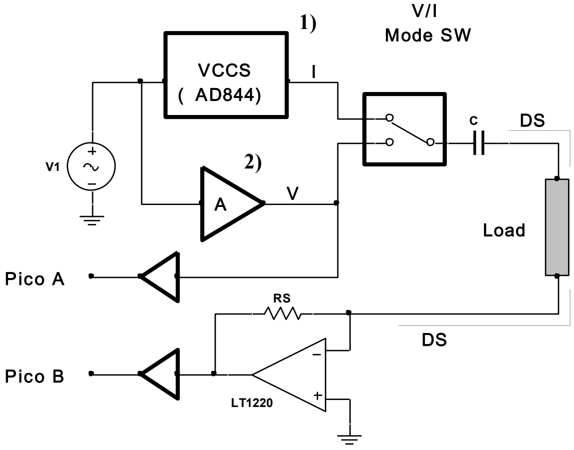


Figure 1. Schematic of the EIS system conditioning unit - 1) AC current source; 2) AC voltage source; 3) current/voltage sense.

A common problem inherent to bioimpedance measurements is the charging of the dc-blocking capacitor between the source and the electrode due to residual DC currents [15]. This effect lead to saturation of the transimpedance output of the AD844. The DC feedback of the implemented configuration maintains dc voltage at the output close to 0V without reducing the output impedance of the source. Subsequently, the output current, is maintained almost constant over a wide range of frequencies.

The high speed voltage-feedback amplifier LM7171 is employed in the voltage mode circuit (see Figure 1). This behaves like a current-feedback amplifier due to its high slew rate, wide unit-gain bandwidth and low current consumption. Nevertheless it can be applied in all traditional voltage-feedback amplifier configurations, as the one used. These characteristics allow the maintenance of an almost constant voltage output over a wide range of frequencies.

Current or voltage signals resulting from voltage or current excitation modes, respectively, are sensed by a high speed operational amplifier, LT1220 (see Figure 1), which performs reduced input offset voltage and is able of driving large capacitive loads.

Gain values of both current excitation source and voltage excitation source can be changed in order to extend the range of impedance magnitude. The transductance gain of the LT1220 is currently set to 5.1 kΩ and defines the gain of the system. Since the gain values are known and also the amplitude of the AC excitation signal, V_1, from the PicoScope®, the EIS system is calibrated automatically by software.

C. Cables Capacitance

For an optimized signal-to-noise ratio, coaxial cable must be used. Nevertheless, this type of cable is prone to introduce high equivalent parasitic capacitances, which translate in errors in the bioimpedance measurements, especially at high frequencies.

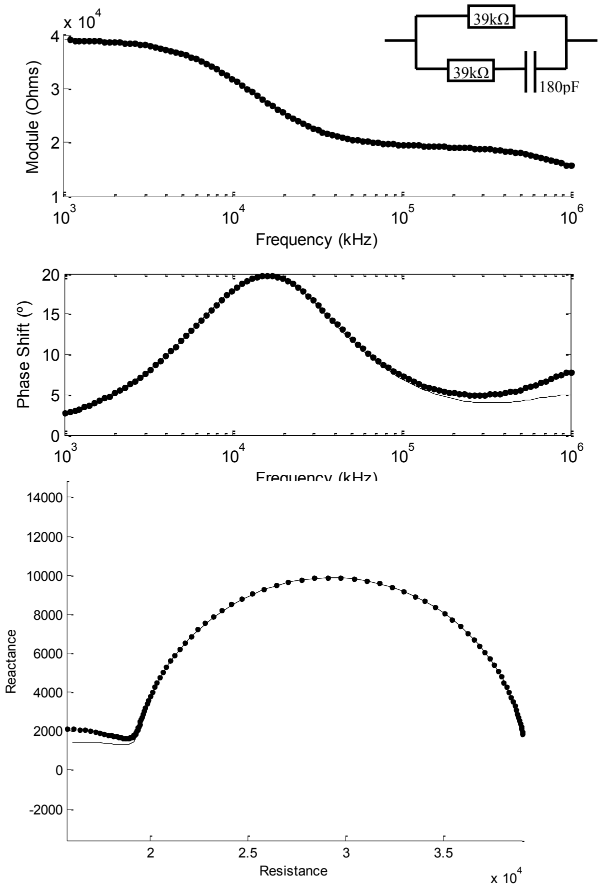


Figure 2. Bode and Cole-Cole diagram showing the reduction of cables capacitive effect by the application of the driven shield technique. The voltage mode excitation was used to analyze the circuit at the right top. The reduction is more noticeable at high frequencies where the capacitive effects have more influence.

When assessing bioimpedance, the capacitive effects from cables are not the only exerting influence. In fact, phase shift effects, perceptible especially in the high frequencies range, are introduced mainly by the amplifiers. The influence of phase shift errors has a cumulative effect that is translated, in the impedance spectra, as an inflexion that occurs at high frequencies (see Figure 2).

This behavior can be simulated by an equivalent circuit as it is like the system analyzes any load always in parallel with a capacitor.

The impedance magnitude, at high frequencies, is also affected. It presents a characteristic decline as the bode diagrams of the Figure 2 show. In the developed EIS system, the slight decline of the impedance magnitude is due to the loss of the product gain-bandwidth of the LT1220 for high frequencies.

Since stray capacitances are considered systematic errors of the system, thus affecting all the measurements, theirs influence doesn't directly affect the results. Although, it is convenient to have an approached sense of the real equivalent circuit (see Figure 3), in such a way that the effect of all the parasitic elements can be considered and/or discounted where justified.

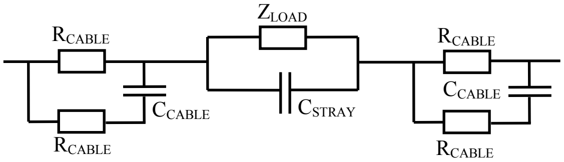


Figure 3. Equivalent electric circuit of all parasitic elements affecting impedance measurements of a load, Z_{LOAD}. The effect of the stray capacitances from cables, C_{CABLE}, is minimized by the driven shield. Other stray capacitance effect, C_{STRAY}, due primarily to the phase shift of amplifiers, can be minimized by software.

III. Assessment of Hydric Stress Level

A. Materials and Methods

Assessing physiological states of plants, using impedance techniques, implies the knowledge of the typical EIS profiles of the species under study, i.e., the EIS profiles for healthy individuals under controlled environment conditions. For this reason, the studies presented herein required an exhaustive EIS assessment and monitoring, performed over months, and extensive data analysis. In addition, the populations of different plant species were kept under controlled environment conditions (temperature, luminosity, soil content and watering), in order to reduce the quantity of variables that may change EIS profiles.

For assessing EIS profiles and studying the hydric stress level there were used: 1) eight young healthy pine trees (*Pinus pinaster* Aiton), with about 0,8 meters tall and 1 to 2 centimeters in diameter; 2) eight young healthy chestnuts trees (*Castanea sativa* Mill), with about 0,5 meters tall and 1 to 2 centimeters in diameter; and 3) four young healthy *Jatropha Curcas* L. trees, with about 0,2 meters tall and 3 to 4 centimeters in diameter.

The choice of the plant species under study is substantiated by the economic and/or forestalls relevance they have. Chestnuts and pines have a crucial economic impact in our country and are currently affected by uncontrolled diseases: the nematode disease, in the case of pine trees, and the ink disease, in the case of chestnut trees. *Jatropha curcas*, by other hand, is a tropical species, lacking physiological studies, which seed are used for biodiesel production.

To perform the EIS measurements, the electrodes were placed in the trunk of each tree, in a diametric position, and about 20 cm above the soil, in the case of the pine and chestnut trees, and about 10 cm above de soil, in the case of *Jatropha curcas*. It was used the portable EIS system version in the voltage mode of excitation and a frequency range between 1 kHz and 1 MHz. Routine acquisitions took place between 11 a.m. and 13 p.m. since it was already verified in previous studies that at this time period the trees impedance is higher and presents few variation (see Figure 4).

To study the hydric stress level variation, there was performed EIS monitoring over two months for one individual of each plant species. Plants were kept one month without watering and, during the remaining month, plants were watered regularly. This process was repeated three times for each case.

B. Results

For each obtained impedance spectrum, there were assessed several impedance parameters. Due to paper space limitation and also because it is a well-known impedance parameter, it will only be presented the results obtained for the ratio Z_1/Z_{50}. Note that it is used the index 1, that corresponds to the lowest analyzed frequency (1 kHz), instead of the index 0, as explained in the *Introduction* section.

The EIS measurements revealed that EIS profiles have a daily oscillation. To analyze this behavior it was calculated the R_1/R_{50} ratio (R represents module) for a period of 4 days.

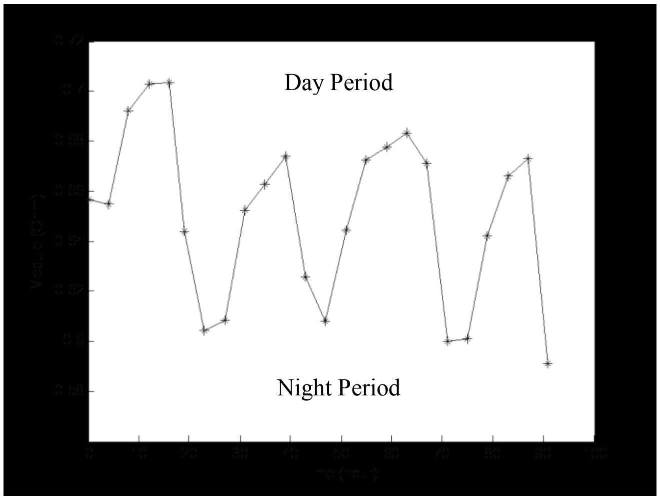


Figure 4. Variation of the R_1/R_{50} ratio during the monitoring of a healthy pine tree. The impedance values show a daily oscillation that is characteristic of the studied trees.

To confirm the daily oscillation it was calculated the Fast Fourier Transform of this ratio. A frequency of 11,57 μHz was clearly founded, which corresponds to a frequency of 24 h. The lower values of the ratio R_1/R_{50} correspond to the night period, while the higher values correspond to the day period where the temperature and luminance are higher (between 11 a.m. and 15 p.m.). Previous studies on plants also shown that, during the day period, the variation of impedance values is lower than the one observed at the night period.

Data obtained for the EIS monitoring and the study of the hydric stress level, revealed a consistent behavior for all the studied species. The Z_1/Z_{50} ratio tends to increase with higher values of hydric stress, for the pines and chestnut trees. After introducing regular watering, the same parameter progressively tended to the typical values of hydrated trees.

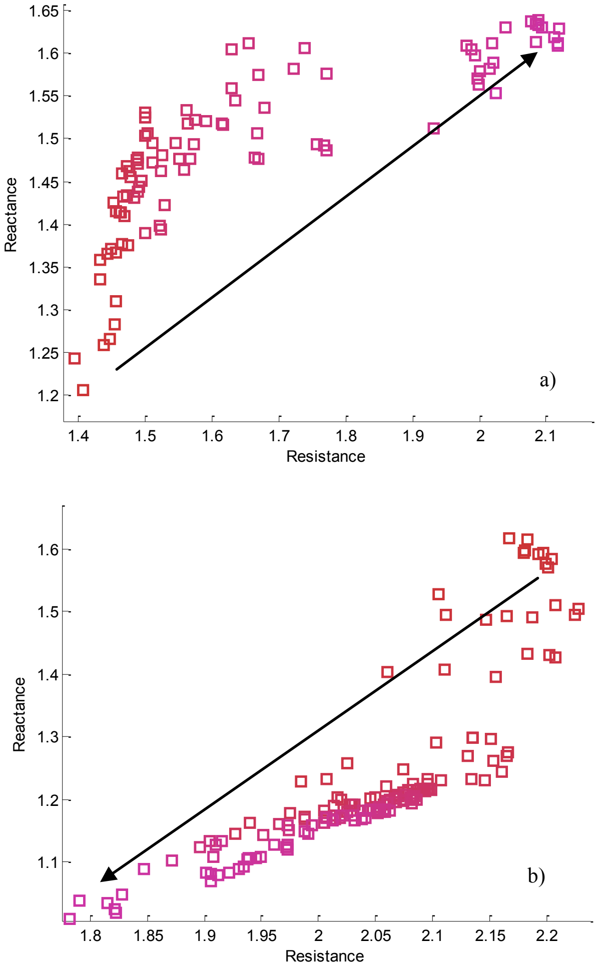


Figure 5. Evolution of the Z_1/Z_{50} ratio during the monitoring of a healthy chestnut tree a) while kept without watering and b) with regular watering. The arrows indicate the direction of the Z_1/Z_{50} ratio evolution.

In the case of *Jatropha curcas*, the impedance behavior with the hydric stress level variation was completely different. The found different results could be explained due the fact that this species contain latex vessels, while pine and chestnut trees do not. In this case, the Z_1/Z_{50} tend to decrease for higher values of hydric stress level and, after introducing regular watering, the ratio abruptly assumed the typical values.

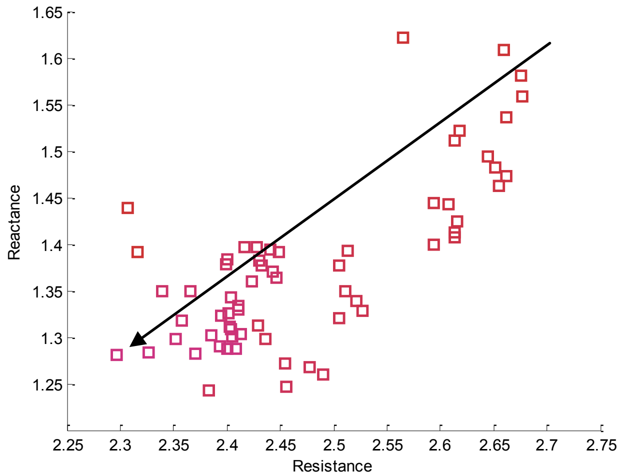


Figure 6. Evolution of the Z_1/Z_{50} ratio during the monitoring of a healthy *Jatropha curcas* tree while kept without watering. The arrow indicates the direction of the Z_1/Z_{50} ratio evolution.

IV. Study of Disease Condition

A. *Materials and Methods*

Twenty four healthy pine trees (*Pinus pinaster* Aiton), with about 2,5 meters tall and 2 to 3 centimetres in diameter, constituted the population for the conducted study. The pine trees were placed in vases in a controlled water environment at a greenhouse. Half of the tree population was watered during 5 minutes per day (~ 133,37 mL/day), while the other half were watered during only 2 minutes per day (~ 66,67 mL/day). This second half was less watered to maintain a relevant level of hydric stress.

After one month elapsed since the pine trees were placed in the greenhouse, the inoculations with pinewood nematode, PWN, (*Bursaphelenchus xylophilus* Nickle) and with the bark beetle (*T. destruens* Wollaston) were performed. Six pines were inoculated with PWN, other 6 pines were inoculated with bark beetles, other 6 pines were inoculated simultaneously with PWN and bark beetles, while the remaining 6 were kept under normal conditions, i.e., healthy. The position of the pines in the greenhouse was made so that each sub-group had the same number of pines with normal watering (5 min/day) and with reduced watering (2 min/day).

To perform the inoculations with bark beetles, callow adults were collected immediately after emergence. In each tree, a box containing 15 beetles were placed in the middle and the device was covered using Lutrasil tissue to avoid beetles escape.

The inoculation with the PWN followed an innovative approach. Firstly, three 2 x 2 cm rectangle of cork were removed from the first tiers of the trunk (about 1,80 m above the soil) and exposed phloem was erased with a scalpel in order to increase the adhesion of the PWN. Afterward, 0,05 mL of of a PWN suspension was placed on in each incision. In the total, 6000 nematodes were inoculated per tree. To finalize the task, the removed rectangle of cork was fixed in the respective place and wrapped with plastic tape.

Seventy days after the inoculations, the EIS measurements were performed in all the tree population. At this time, the pine trees inoculated with PWN presented some visually symptoms of the PWD. The decay of those trees, rounded 40 %. Two of the healthy pines died (decay of 100 %) due to hydric stress. All remaining individual appeared healthy.

To perform the EIS measurements, the electrodes were placed in the trunk of each tree, in a diametric position, and about 30 cm above the soil. It was used the portable EIS system version in the voltage mode of excitation and a frequency range between 1 kHz and 1 MHz. There were taken two measurements for each tree. The acquisitions also took place between 11 a.m. and 13 p.m.

In order to relate the EIS data with the PWD and the stage of the disease, the trunk of the pine trees inoculated with PWN were cut in three distinct regions to perform a count of nematodes. The cuts were executed: a) immediately below the inoculation incision (180 cm above the soil); b) 30 cm above the soil (where EIS measurements took place); and c) in the middle of the previous two cuts (approximately 80 cm above the soil).

After the EIS measurements, two healthy pines were monitored by two independent portable EIS systems. After a week of monitoring, the same pines were inoculated with PWN, and the measurements continued during 7 more weeks. The main purpose of this last experiment was to study the variation of the pine EIS profiles during the decay due to the PWD.

B. *Results*

In order to compare results between the different physiological states of the trees, several impedance parameters were assessed. The impedance parameter that showed better results was the Z_1/Z_{50} ratio.

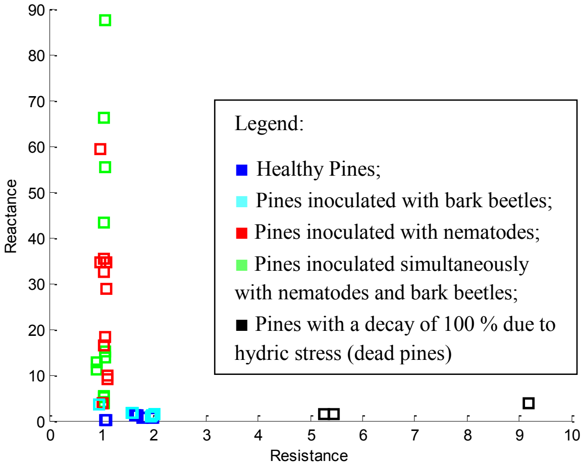


Figure 7. Values of the impedance parameter Z_1/Z_{50} for each of the 24 pine trees. Note that there are represented two values for each pine.

The analysis of the obtained results shown that the healthy pines and the pines inoculated with bark beetles have similar Z_1/Z_{50} values. In fact, the bark beetles doesn't damage the inner structure of the trees, therefore it was expected that the impedance profiles were similar between healthy pines and pines inoculated with bark beetles.

On the other hand, Z_1/Z_{50} values for the pines inoculated with nematodes and also, for those inoculated simultaneously with nematodes and bark beetles, locates in the same region, different from the previous one, of the graph of Figure 7. Those values present a relatively high dispersion in terms of reactance. It was later confirmed that higher reactance Z_1/Z_{50} values correspond to higher number of nematodes in the tree (see Figure 8).

The pines that died due to hydric stress (decay of 100%) were also studied and the Z_1/Z_{50} parameter present high resistance values in relation to all the other pines.

The counting of nematodes in the several cut sections revealed that the concentration of nematodes was higher in the cut sections b) and c) for the pines less watered (pines 1, 2 and 3) – see Table II. It is known that the nematodes move toward watered regions along the trunk. For this reason, the concentration of nematodes in the lower parts of the trunks was much higher for the pines with less watering than for those with regular watering (pines 4, 5 and 6).

TABLE II. NUMBER OF NEMATODES IN THE TRUNKS OF PINE TREES PER CUT SECTION

Tree	Cut Section	Number of nematodes in 0,05 mL
1	*a*	1
	b	0
	c	133
2	*a*	0
	b	43
	c	1
3	*a*	0
	b	0
	c	112
4	*a*	4
	b	20
	c	0
5	*a*	0
	b	17
	c	0
6	*a*	0
	b	0
	c	14

These results for the nematodes counting support the already referred results obtained for the Z_1/Z_{50} impedance parameter. In fact, it is observed a clear relation between the number of nematodes and the reactance dispersion for the Z_1/Z_{50} parameter, as Figure 8 shows. The higher the number of nematodes is, the higher is the reactance value of Z_1/Z_{50}. It is considered that the dispersion in terms of resistance is not significant when compared with values from pines in other physiological condition – see Figure 7.

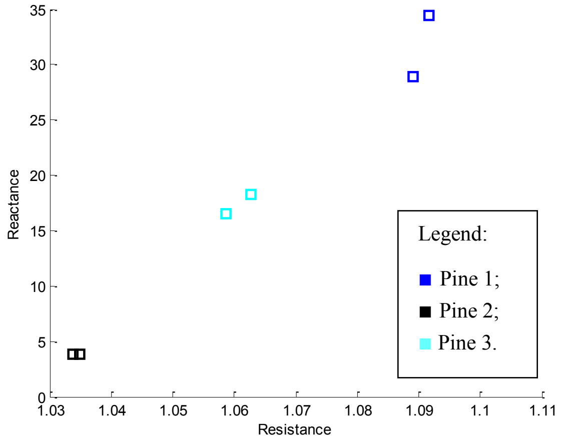


Figure 8. Values of the impedance parameter Z_1/Z_{50} for the pines inoculated with nematodes and with low watering (pines 1, 2 and 3 from the Table II). Note that there are represented two values for each pine.

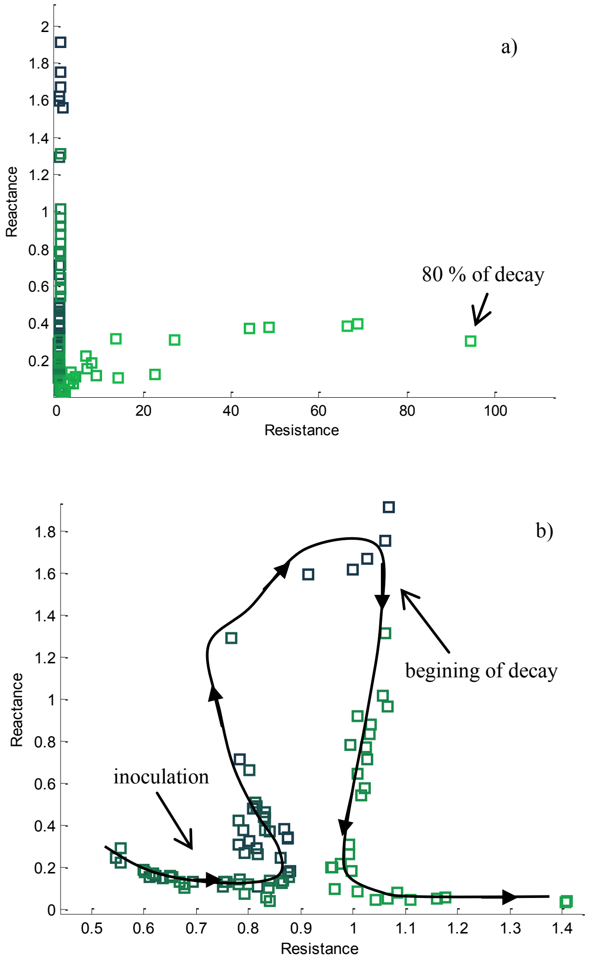


Figure 9. a) Evolution of the Z1/Z50 during the monitoring time (8 weeks). b) Closer view from the Z_1/Z_{50} evolution, showing a hysteresis-like behaviour.

There were monitored two healthy pines, one with low watering (2 min/day) and another with regular watering (5 min/day). After one week from the beginning of the monitoring, both pines were inoculated with nematodes. It was shown again a dispersion of the reactance values of the Z_1/Z_{50} parameter, as Figure 11 shows. As time passed the reactance values became higher. The higher values of reactance were achieved for the pine with less watering. According to the previous presented results, it was expected that the number of nematodes increase in the below part of the trunk for the pines with less watering; and consequently, to observe a higher rising of the reactance of the Z_1/Z_{50} parameter. After the 6th week, pines start to decay strongly and it was observed a relevant decrease of the reactance and a significant increase of the resistance for the same parameter – see Figure 9. The higher values of resistance were achieved for the pine with less watering, and also in a shorter period of time. At the end of the monitoring, the decay of the pines, evaluated by an expertise, was about 80 % for the pine with regular watering and 100 % for the pine with less watering.

From the Figure 9 b), that represents a closer view of the Z_1/Z_{50} values for the monitoring, it is possible to observe that the path followed during the period of nematodes population increasing is different from the path followed during the period of decay, i.e., it is observed an hysteresis-like behavior.

V. CONCLUSIONS

The EIS system was developed in order to ensure a robust, efficient and fast bioimpedance analysis. The adaptability to different biological applications, the portability and the usage of easily accessible and affordable components, were preferred aspects taken into account. In this manner, the system allows the user to choose the settings of the analysis that best fit to a specific application.

The system is able to perform AC scans within a frequency range from 1 kHz to 1 MHz. The frequency limits and the number of intervals of the scan can be selected at the user interface (developed with Matlab® tools). The type of signal used to excite de sample, voltage or current, can be preselected by an external switch. This allows the usage of the source with the best behavior in a concrete application.

To overcome problems inherent to stray capacitive effects from cables, a driven shield technique is applied. The maximum phase shift reduction is estimated at 20.4 % for the current excitation mode and at 35.8% for the voltage mode.

The biological application study aimed at discriminating between different physiological states of three plant species: pine *(Pinus pinaster* Aiton), cheastnut (*Castanea sativa* Mill) and *Jatropha Curcas* L.

The obtained results suggest that the implemented method may constitute a first innovative approach for the assessment of physiological states of plant organisms and to the early diagnosis of plant diseases. The consistency of the results obtained for the three studied species reveals the transversality of the method. Since EIS profiles are obtained for healthy individuals of a given plant organism it is possible to assess and study the physiological states.

EIS profiles showed a consistent behavior whit the hydric stress level of the three studied species. The Z_1/Z_{50} impedance parameter presents increased values of both reactance and resistance when the hydric stress is high, for the pines and chestnuts. In the case of *Jatropha curcas*, this parameter presented decreased values of both reactance and resistance when the hydric stress was high. This inverse behavior may be explained due to the presence of latex vessels in this species.

The evolution of the Z_1/Z_{50} impedance parameter may be used to predict risky hydric stress level of a plant organism.

In addition, the achieved impedance parameters allow discriminating three different physiological states for pine trees: healthy trees, trees with PWD and trees in hydric stress.

The trees with PWD present Z_1/Z_{50} ratio with high values of reactance, suggesting that the current flows preferably trough the cytosol. In fact, the action of the nematodes inside the tree may destroy cell membranes. This means that membranes capacitor effect becomes less significant in the impedance measurement.

It was also shown that the number of nematodes and Z_1/Z_{50} impedance parameter are related. The higher the number of nematodes is, the higher the reactance of the ratio is.

The action of bark beetles seems not to interfere, at least in measurable terms, in the level of hydric stress of pine trees.

Healthy trees, with high values of hydric stress (decays above 80 %), and also trees with PWD at advanced stages, revealed low reactance and high resistance for the same studied parameter. The high values of resistance are justified due to the water loss in the tree. Consequently, it means that for this specific case, the method cannot distinguish between trees with PWD or trees with high level of hydric stress but with no disease. However, it is known that advanced stages of PWD promote high levels of hydric stress. This means that both cases represent, in practical terms, the same situation, i.e., the tree presents high probability to die. In addition, in the stages where the method is able to distinguish between healthy trees and trees with PWD, the decay was determined to round the 40 %. Therefore, if a cure is available, this diagnosis could help to administrate a treatment and reverse the disease evolution.

Hence, the main conclusion of the developed study is that the implemented method could be used to assess physiological states (such as the hydric stress) of living trees, and that the Z_1/Z_{50} impedance parameter could be applied as a risk factor.

Acknowledgment

We acknowledge support from Fundação para a Ciência e Tecnologia, FCT (scholarship SFRH/BD/61522/2009).

References

[1] L. Martins, J. Castro, W. Macedo, C. Marques and c. Abreu, “Assessment of the spread of chestnut ink disease using remote sensing and geostatistical methods,” Eur J Plant Pathol, vol. 119, pp. 159–164, April 2007.

[2] CABI and EPPO, “Data sheets on quarantine pests – *Bursaphelenchus xylophilus*,” EPPO quarantine pests, contract 90/399003.

[3] W. Parawira, “Biodiesel production from *Jatropha curcas*: a review,” Scientific Research and Essays, vol. 5(14), pp. 1796–1808, July 2010.

[4] T. Repo, G. Zhang, A. Ryyppo and R. Rikala, “The electrical impedance spectroscopy of scots Pine (*Pinus sylvestris* L.) shoots in relation to cold acclimation,” Journal of Experimental Botany, vol. 51 (353), pp. 2095–2107, December 2000.

[5] A. Pirzad, M. K. Shakiba, S. Zehtab-Salmasi, S. A. Mohammadi, R. Darvishzadeh and A. Samadi, “Effect of water stress on leaf relative water content, chlorophyll, proline and soluble carbohydrates in *Matricaria chamomilla* L.,” Journal of Medicinal Plants Research, vol. 5(12), pp. 2483–2488, June 2011.

[6] L. Callegaro, “The metrology of electrical impedance at high frequency: a review,” Meas. Sci. Techno, vol. 20, 022002, February 2009.

[7] U. G. Kyle, I. Bosaeus, A. D. De Lorenzo, P. Deurenberg, M. Elia, J. M. Gómez, *et al*, “Bioelectrical impedance analysis – part I: review of principles and methods,” Clinical Nutrition, vol. 23(5), pp. 1226–1243, October 2004.

[8] U. Pliquett, “Bioimpedance: a review for food processing,” Food Engineering Reviews, vol. 2(2), pp. 74–94, June 2010.

[9] S. Grimnes and O. G. Martinsen, “Bioimpedance and bioelectricity basics,” 2nd Edition, Academic Press of Elsevier, 2008.

[10] T. Hayashi, M. Iwamoto and K. Kawashima, “Identification of irradiated potatoes by impedance measurements,” Biosci Biotechnol Biochem, vol. 56(12), pp. 1929–1932, December 1992.

[11] M. Rafiei-Naeini, P. Wright and H. McCann, “Low-noise measurement for electrical impedance tomography,” IFMBE Proceedings, vol. 17(10), pp. 324–327, 2007.

[12] A. S. Ross, G. J. Saulnier, J. C. Newell and D. Isaacson, “Current source design for electrical impedance tomography,” Physiol Meas, vol. 24(2), pp. 509–516, May 2003.

[13] P. J. Yoo, D. H. Lee, T. I. Oh and E. J.Woo, “Wideband bio-impedance spectroscopy using voltage source and tetra-polar electrode configuration,” Journal of Physics, vol. 224(1), pp. 224-228, 2010.

[14] G. J. Saulnier, A. S. Ross and N. Liu, “A high-precision voltage source for EIT,” Physiol Meas, vol. 27(5), pp. 221–236, May 2006.

[15] F. Seoane, R. Bragós and K. Lindecrantz, “Current source for multifrequency broadband electrical impedance spectroscopy systems – a novel approach,” Proceedings IEEE Eng Med Biol Soc, vol. 1, pp. 5121–5125, August 2006.

[16] T. Yamamoto, Y. Oomura, H. Nishino, S. Aou and Y. Nakano, “Driven shield for multi-barrel electrode,” Brain Research Bulletin, vol. 14(1), pp. 103-104, January 1985.

A Semantic-Based Similarity Measure for Human Druggable Target Proteins

Eduardo C. dos Santos, Marcelo M. Santoro
Marcos A. dos Santos, Julio C. D. Lopes
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
edu@edusantos.eti.br, santoro@icb.ufmg.br
marcos@dcc.ufmg.br, jlopes.ufmg@gmail.com

Bráulio R. G. M. Couto
Centro Universitário de Belo Horizonte UNIBH
Belo Horizonte, Brazil
coutobraulio@hotmail.com

***Abstract*—The target identification is the first step on drug discovery pipeline. Thus, techniques which address the selection of potential "druggable targets" and potential "therapeutical targets" are very relevant to the discovery of new drugs and therapies. Nowadays, public databases with drug target indication provide target similarity searching based on BLAST. We demonstrate that the current protein annotation terms may be used on the development of semantic-based measures to provide target similarity searching. This approach allows to predict target similarities based on known signatures of a given protein even without the knowledge of the whole sequence. Our method produces a semantic ordering of the drug targets and provides a tool for latent information retrieving and for clustering analysis. New candidates may be compared against the known targets in a reduced space vector defined by singular value decomposition.**

Keywords-drug targets; SVD; semantic similarity; clustering.

I. Introduction

The drug discovery pipeline has the target identification and validation as the very first phases. Recently, known drug usage has been optimized addressing it to different targets [1], [2], [3]. The purpose is to use the known chemical properties and response of compounds with acceptable ADMET (administration, distribution, metabolism, excretion and toxicity) properties on developing new therapies (repurposing approved drugs) and/or on developing new lead compounds on a information-driven rational approach. The study of target similarity may be also helpful for predicting promiscuous binding sites and some kind of side-effects.

Public resources with drug target indication (as TTD [4] and DrugBank [5]) provide target similarity searching based on BLAST algorithm. But it is known that there are also important correlations (structural similarity and off-target similarity) even for low-similar sequences. Known signatures of targets (as annotated on GO, InterPro, Pfam, PROSITE and other resources) may be used for predicting correlations among different targets and/or among different target subsets. Indeed, 130 InterPro entries were identified on "druggable genome" searching [6]. It was also shown that Pfam annotation may be used for the same purpose [7].

The objective of this study was to evaluate whether semantic similarity across protein annotation terms can be used as an alternative to sequence alignment for predicting target similarities.

In general, semantics is the study of meaning. Semantic similarity is a concept whereby a metric is assigned to terms or documents in a set of terms or documents according to the likeness of their meaning in a pragmatic approach (i.e., considering how the context contributes to meaning). Broadly speaking, "two objects are semantically similar if they are related to similar objects" [8]. A semantic similarity measure may reveal new correlations, which are not possible by strictly direct queries onto relational databases. It is called *latent information retrieving*. Furthermore, semantic-based similarities may be determined over data hold in the form of annotation, which are more suitable for humans, and may be used to knowledge discovery exploring scientific data resources. Indeed, the use of semantic-based similarities across the Gene Ontology (GO) has been evaluated in the literature [9], [10].

Firstly, a protein drug target was represented by a binary column vector with m rows, each one representing the presence or absence of one InterPro signature in the sequence. A database with n protein drug targets is represented by a mxn binary matrix A, that is submitted to singular value decomposition [11] in order to develop a similarity measure among human protein drug targets.

The methodology can be expanded to incorporate different kinds of descriptors (e.g., MeSH terms) to discover more specific drug target relationships.

A. Singular Value Decomposition

The Singular Value Decomposition (SVD) establishes non-obvious but relevant relationships among clustered entities [11], [12], [13]. The rationale behind SVD is that a mxn matrix A can be represented by a set of derived matrices [13], which allows by a numerically different data representation without loss of semantic meaning.

Let A be any mxn matrix of ranking r. Then there exist a mxm matrix U_f, a nxn matrix V and a mxn matrix S for which:

$$A = U_f S V^T, \tag{1}$$

where:

- U_f is an mxm orthogonal matrix, which columns are the eigenvectors of the matrix AA^T;
- S is a mxn diagonal matrix with the **singular values** of A along its main diagonal in decreasing order;
- V is an nxn orthogonal matrix, which columns are the eigenvectors of the matrix A^TA.

These dimensions are for what is called the **full SVD**. Since all the elements of S below the n^{th} row are zero, partitioning the matrix U it can be taken the so called **thin SVD**[12]:

$$A = USV^T, \tag{2}$$

where:

- U is an mxn orthogonal matrix;
- S is a nxn diagonal matrix with the **singular values** of A along its main diagonal in decreasing order;
- V is an nxn orthogonal matrix.

Taking only the k most significant singular values of A, where $k < r$, the matrix A can be approximated by a low-dimensional matrix (A_k) given by:

$$A \approx A_k = U_k S_k V_k^T = \sum_{e=1}^{k} u_e s_e v_e^T, \tag{3}$$

where u_e and v_e are, respectively, the $column$ vector of U and the row vector of V both related to the e^{th} singular value in the decreasing order and k is the index of the highest relevant singular value.

The data approximation depends on how many singular values are used [14]. In this case, the k number of singular values is also the rank of the matrix A_k. The technique allows information extraction with less data. It is possible to compress/decompress data within a non-exponential execution time, and it make viable complex analysis across large amount of data [14]. A data set represented by a smaller number of singular values than the full size original data set has a tendency to group together certain data items that would not be grouped if the original data set is used [13]. This could explain why clusters derived from SVD can expose non-trivial relationships among the original data set items [15].

There are different methods to determine the $rank$ k of A_k. One of them is by the *scree test* [16].

A new entity represented by a column vector q, which is equivalent to ones of the original matrix A, may be compared with the entities represented in A in the smaller-dimensional space by a simple and low computing cost method. First, obtain the equivalent vector q_k in the reduced space vector. This can be made, as proposed by Lars Eldén [11], by computing:

$$q_k = q^T U_k. \tag{4}$$

Then apply some similarity metric (e.g., cosine measure or Euclidean distance) to compare q_k with the row vectors of $V_k S_k$. Thus, it is not necessary to compute the SVD factorization every time that a new target is introduced. It is only necessary to recompute the SVD factorization with the new query vector if it can not be represented by a combination of the vectors of the base. Otherwise, the new vector q_k may be incorporated to the matrix V_k.

B. Similarity measures

To assess the similarity between two entities, it can be used some similarity measure and evaluate its significance. There are different measures which may be tested, as the Euclidean distance; the cosine similarity; etc. In this paper, entities were represented in the low-dimensional space produced by SVD factorization and, after that, it was applied a cosine-based similarity measure, which is calculated as:

$$sim(c_i, c_j) = \cos(\alpha_{ij}) = \frac{c_i c_j'}{\sqrt{c_i c_i'}\sqrt{c_j c_j'}}, \tag{5}$$

where:
- c_i corresponds to the *i-th* row in $V_k S_k$ and;
- c_i' is the transpose of the vector row c_i.

II. Material and methods

A matrix with 1906 binary vectors was constructed, which represent protein drug targets retrieved from public databases (TTD [4], DrugBank [5] and KEGG-Drug [17]). Each protein representing vector is a set of 2700 binary descriptors. Each of these descriptors represent an InterPro annotation. It was used InterPro annotations of the following types: Family (F), Domain (D), Region (R), Active Site (A) and Binding Site (B). On considering every site-related annotation it was observed if the signature has occurred or not on a region of the sequence for which exists some annotation of F, D or G type. 365 of the 1906 targets were extracted randomly for training and validating purpose and the remaining 1541 were used to generate a representative vector space using SVD.

SVD factorization was applied to A and $k = 320$ factors were selected by *scree test* to determine the low-rank approximation A_k (Fig. 1).

The factorization provided a reduced dimensionality space in which relationships among the drug targets could be established. The similarity between any pair of drug targets was calculated as the cosine of the angle between the respective target representing vectors on the reduced space. Thus, the similarity measure of a pair of targets is equivalent to the dot product between the respective rows of the matrix $V_k S_k$ given by the (3).

The similarity relationships were analyzed by using clustering techniques implemented in the software named Multi-Experiment Viewer (MeV) [18], a freely available software

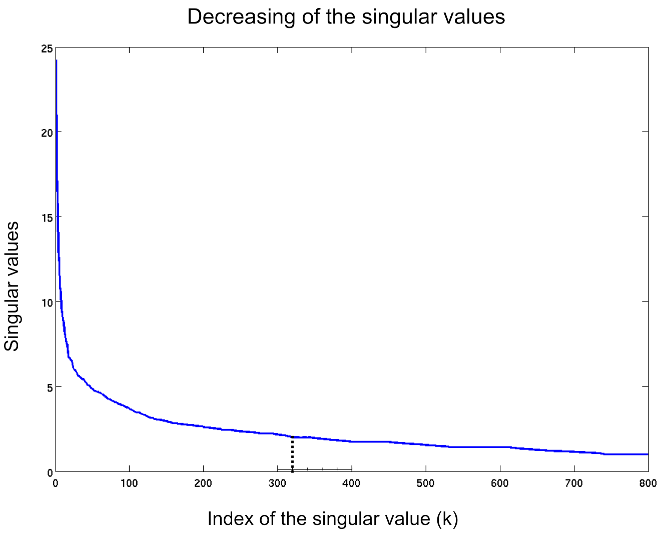


Figure 1. Singular values of A (as obtained by SVD factorization). The first 320 singular values (and respective orthogonal vectors) were selected by the *scree test*.

application that provides an extensive library of algorithms and visualization tools for integrative data analysis from a user-friendly interface.

III. Results

A similarity matrix was constructed from the values of the cosines computed as described on the previous section and used this matrix as input into the software named MeV. Then, it was applied the hierarchical clustering algorithm (HCL) implemented in MeV and it produced a heatmap with the targets semantically reordered (Fig. 2, 3, 5 and 6). Fig. 2 shows the heatmap for the whole ensemble. Fig. 3 shows in detail the region at the heatmap related to nuclear receptors (NR). The similarity measure was found to be efficient in discriminating the NR members in a second level grouping Peroxisome; Retinoid and Vitamin D receptors.

The Euclidean distance was also evaluated and let us to conclude that, referring to our application and data set, Euclidean distance and cosine angle measure provide similar clustering results (Fig. 4).

Similarly to the case of nuclear receptors, Fig. 5 shows a cluster (drug targets with NAD-P binding domain) larger than the NR cluster and with deeper hierarchy level. Fig. 6 illustrates the efficiency of the method to discover relationships hardly recognized by simple sequence similarity search – it shows an interesting relationship between Fibronectin type-III like folding and Immunoglobulin like folding – two domains that have low-similar sequence but high-similar structure and that co-occur in some protein-folding pathways [19].

The results were compared with the ones produced by sequence similarity (with BLASTALL) [20]. For the whole

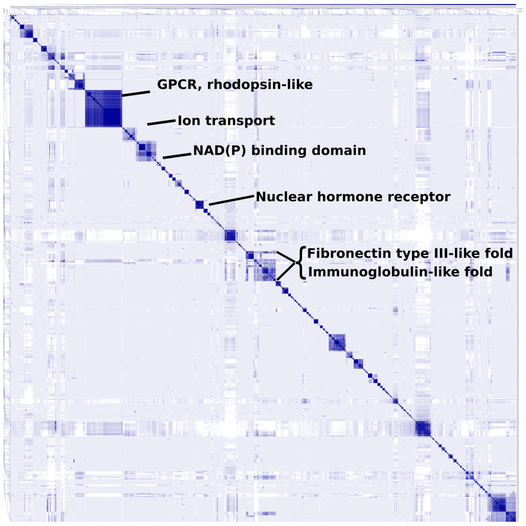


Figure 2. Heatmap with all 1541 drug targets reordered semantically by the hierarchical clustering algorithm. It was easy to identify various clusters as the GPCRs (the greatest group) and other cases showed in detail in other figures.

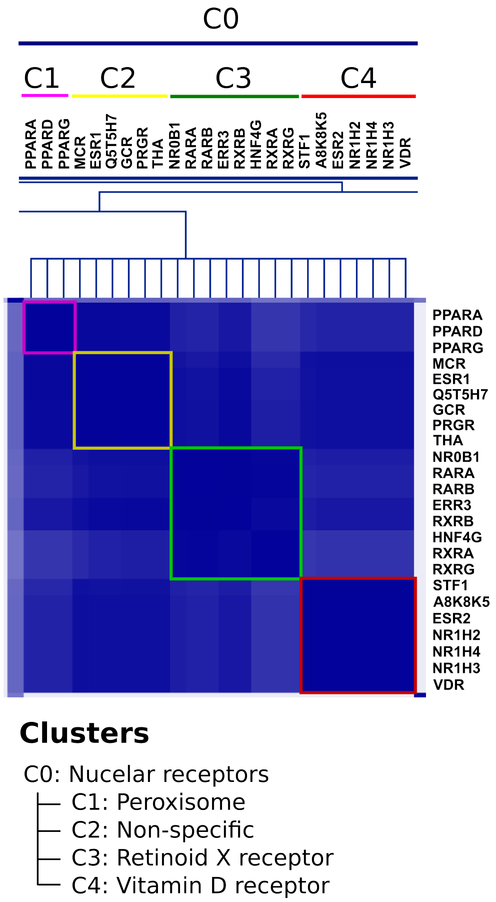


Figure 3. Zoom view of the region (of Fig. 2) related to hormone nuclear receptors. It is evident the in-depth consistency according to their additional annotations.

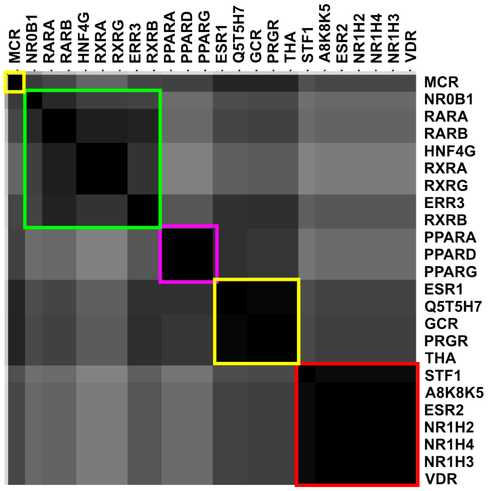


Figure 4. Zoom view of the region related to hormone nuclear receptors from the heatmap obtained from Euclidean distance. The clustering results provided by Euclidean distance for small samples were very similar to the ones provided by cosine measure.

data set, the BLAST bitscore did not provide good discrimination, but for a small sample the clustering results were very similar when using the SVD-based similarity score and the sequence similarity based score. It was performed two different clustering methods in this case: HCL and cluster affinity searching technique (CAST), both implemented in MeV. Fig. 7 illustrates the clustering results using semantic similarity and sequence alignment for 42 selected targets. Five non-unitary groups could be easily identified. One GPCR (PE2R3_HUMAN), left as orphan by the two clustering methods across the sequence similarity matrix. The same target was correctly grouped (in the context of biological annotations) with other GPCR by both, HCL and CAST, across the semantic-based similarity matrix.

To find potential “druggable” candidates, it was projected other proteins into the reduced space. As an example of interesting finding, the case of Kynurenine 3-monooxygenase (KMO) can be cited (Table I). The value of the distance-like coefficient is significantly low only for two known drug targets: ERG1 and SOX. ERG1 shares annotation with both KMO and SOX, but there are not shared annotation between SOX and KMO. So, the space transformation indicates a non *prima facie* relationship between KMO and SOX. That “secondary” relationship is not retrieved from the original data set or from the transformed space when it is added many factorized terms. The higher the number of terms of the factorization, the smaller the retrieval capability to discover hidden relationships (with many terms it is only possible to compute the coefficient for pairs whose members share some annotation, the remaining becomes equivalent to infinite). The kynurenine pathway is the main pathway for tryptophan metabolism and have been considered a pathway with a lot of potential sites for drug discovery in neuroscience [21].

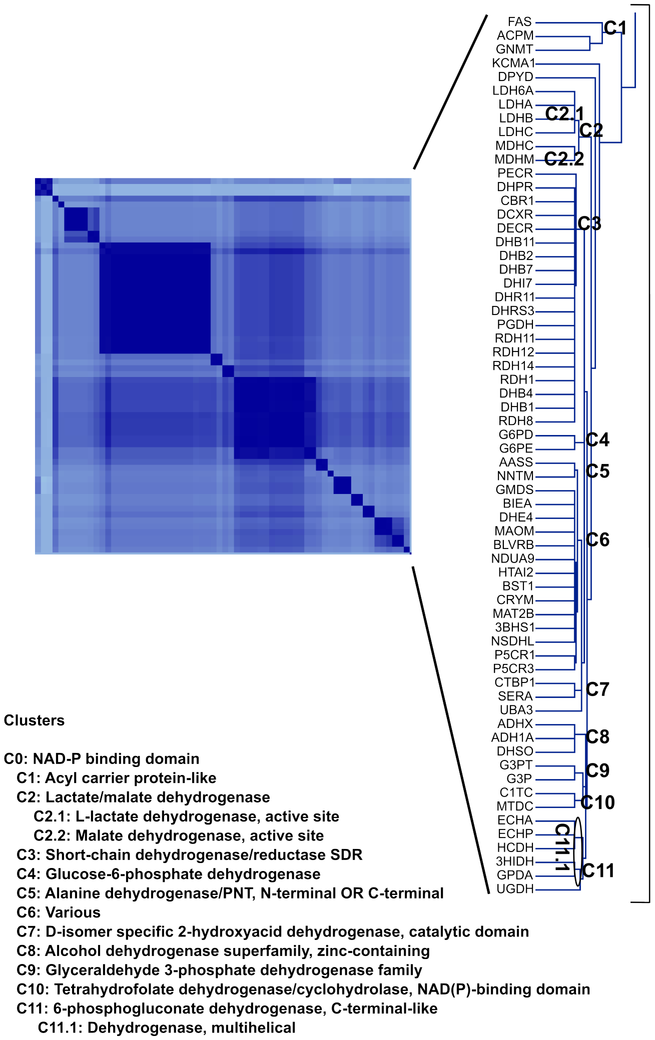


Figure 5. Zoom view of another region of the Fig. 2 related to drug targets with NAD-P binding domain. Again, it is evident the in-depth consistency – here, on a larger cluster than the one with nuclear receptors and showing deeper hierarchy level.

Particularly, KMO (a member of the kynurenine pathway) has the gene located in the chromosome region associated with schizophrenia [22]. On the other hand, it is known that glycine binds to SOX and it is used as an alternative therapy of schizophrenia [23], [24].

IV. Conclusion and Future Work

A semantic-based measure across the InterPro annotations of protein drug targets was developed. It was shown that this measure may be used for similar targets searching. Nowadays, public resources provide target similarity searching using a local BLAST algorithm. Our method has a fixed computational time consumption independently of the sequence size. New targets may be compared against the current set representing it by their biological annotations, projecting it on the U_k space and, then, computing the cosine

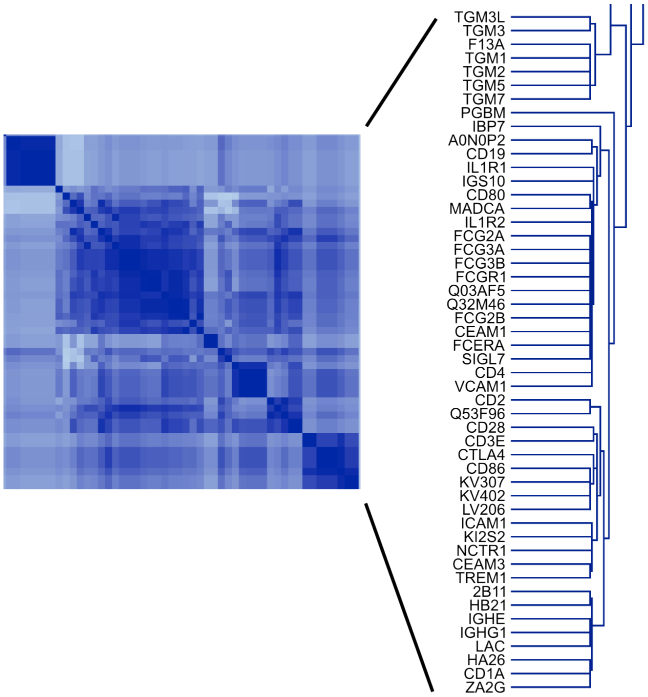


Figure 6. Zoom view of the region of the Fig. 2 showing targets with Fibronectin type III-like fold domain and/or Immunoglobulin-like fold domain. The correlation among these targets are shown as estimated [19].

Table I
RANKED LISTS FOR KMO_HUMAN OF SIMILAR TARGETS

rank	$k = 320$		$k = 800$		original	
	target	score	target	score	target	score
1	ERG1	0.0009	ERG1	0.0445	ERG1	0.2374
2	SOX	0.0022	-	∞	-	∞
3	CBPE	0.4656	-	∞	-	∞
4	SO1B1	0.5135	-	∞	-	∞
5	P85A	0.5512	-	∞	-	∞
6	DCK	0.5550	-	∞	-	∞

Each ranked list is given by the distance-like score computed from the k-dimensional space or from the original vector space (before apply SVD). The value considered infinite is 0.6931.

among the produced column vector with each row vectors of V_kS_k. The biological annotations of the new targets may be determined by InterProScan [25] over their sequence or may be inferred by the research by other experimental observations. Thus, it was shown that the effort exerted on annotation can be explored to order data semantically. The measure is consistent and complementary to BLAST-based sequence alignment approach allowing identification of similar and co-existent fold domains even for low-similar sequences. So, the measure can be potentially effective to discover hidden relationships that are hardly recognized by simple sequence similarity search. Furthermore, the methodology can be expanded to incorporate different kinds of descriptors (e.g., MeSH terms) to discover more specific drug target relationships.

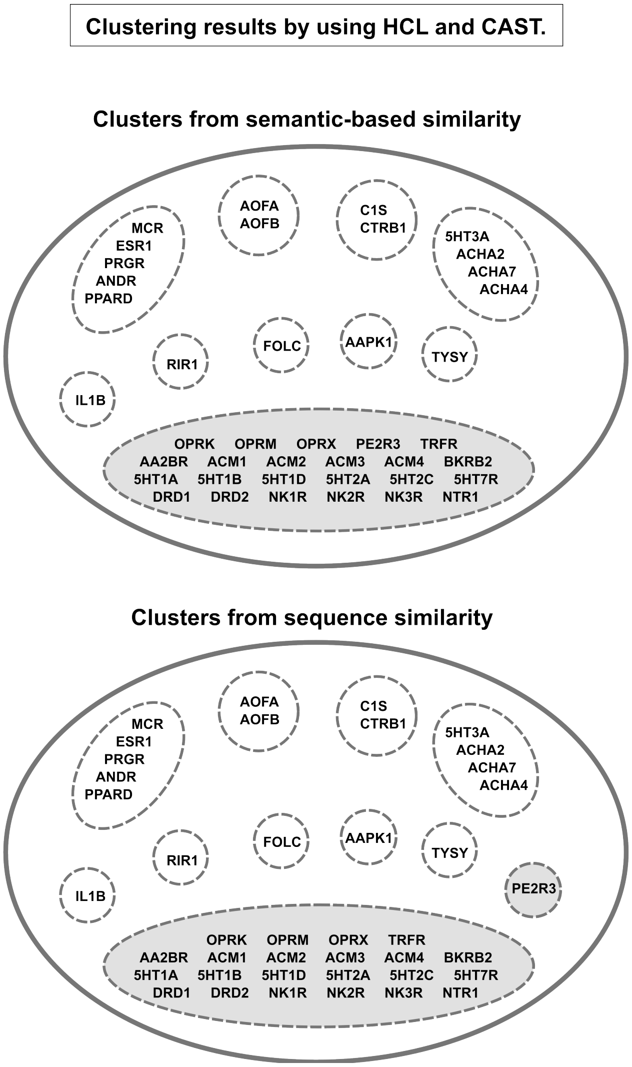

Clustering results by using HCL and CAST.
Clusters from semantic-based similarity
MCR ESR1 PRGR ANDR PPARD
AOFA AOFB
C1S CTRB1
5HT3A ACHA2 ACHA7 ACHA4
IL1B
RIR1
FOLC
AAPK1
TYSY
OPRK OPRM OPRX PE2R3 TRFR
AA2BR ACM1 ACM2 ACM3 ACM4 BKRB2
5HT1A 5HT1B 5HT1D 5HT2A 5HT2C 5HT7R
DRD1 DRD2 NK1R NK2R NK3R NTR1
Clusters from sequence similarity
MCR ESR1 PRGR ANDR PPARD
AOFA AOFB
C1S CTRB1
5HT3A ACHA2 ACHA7 ACHA4
IL1B
RIR1
FOLC
AAPK1
TYSY
PE2R3
OPRK OPRM OPRX TRFR
AA2BR ACM1 ACM2 ACM3 ACM4 BKRB2
5HT1A 5HT1B 5HT1D 5HT2A 5HT2C 5HT7R
DRD1 DRD2 NK1R NK2R NK3R NTR1

Figure 7. Clusters resulted by both HCL and CAST methods for a small sample with 42 targets. The algorithms were performed across the similarity matrices obtained from the semantic-based similarity measure and the BLAST bitscore based similarity. Clusters in grey denote some discrepancy between the two methods.

We are going to expand our work on:

- Optimizing the applied algorithm and parameters (clustering algorithm, rank determination, etc.);
- Ensemble correlations and other cross-correlation analysis;
- Predicting new potential drug target candidates and new possible therapy applications;
- Applying the method for non-human targets;
- Incorporating other types of annotations to the descriptors set (e.g., MeSH, OMIM, and UMLS);
- Comparing the performance of the method with approaches using other types of decomposition: PCA and NMF.

REFERENCES

[1] M. J. Keiser, V. Setola, J. J. Irwin, C. Laggner, A. I. Abbas, S. J. Hufeisen, N. H. Jensen, M. B. Kuijer, R. C. Matos, T. B. Tran, R. Whaley, R. A. Glennon, J. Hert, K. L. H. Thomas, D. D. Edwards, B. K. Shoichet, and B. L. Roth, "Predicting new molecular targets for known drugs," Nature, vol. 462, pp. 175–181, 2009.

[2] M. J. Keiser, B. L. Roth, B. N. Armbruster, P. Ernsberger, J. J. Irwin, and B. K. Shoichet, "Relating protein pharmacology by ligand chemistry," Nature Biotechnology, vol. 25, pp. 197–206, 2007.

[3] M. Campillos, M. Kuhn, A.-C. Gavin, L. J. Jensen, and P. Bork, "Drug target identification using side-effect similarity," Science, vol. 321, pp. 263–266, 2007.

[4] F. Zhu, B. Han, P. Kumar, X. Liu, X. Ma, X. Wei, L. Huang, Y. Guo, L. Han, C. Zheng, and Y. Chen, "Update of ttd: Therapeutic target database," Nucleic Acids Research – Database issue, vol. 38, pp. D787–D791, 2010.

[5] D. S. Wishart, C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali, "Drugbank: a knowledgebase for drugs, drug actions and drug targets," Nucleic Acids Research – Database issue, vol. 36, pp. D901–D906, 2008.

[6] A. L. Hopkins and C. R. Groom, "The druggable genome," Nature reviews. Drug discovery, vol. 1, no. 9, pp. 727–730, September 2002.

[7] A. P. Russ and S. Lampel, "The druggable genome: an update." Drug Discovery Today, vol. 10, no. 23-24, pp. 1607–1610, 2005.

[8] G. Jeh and J. Widom, "Simrank: A measure of structural-context similarity," in In KDD, 2002, pp. 538–543.

[9] P. Lord, R. Stevens, A. Brass, and C. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation," Bioinformatics, vol. 19, pp. 1275–1283, 2003.

[10] M. Chagoyen, P. Carmona-Saez, C. Gil, J. M. Carazo, and A. Pascual-Montano, "A literature-based similarity metric for biological processes," BMC Bioinformatics, vol. 7, pp. 363–375, 2006.

[11] L. Eldén, "Numerical linear algebra in data mining," Acta Numerica, vol. 15, pp. 327–384, 2006.

[12] L. Eldén, Matrix methods in data mining and pattern recognition. Fundamentals of Algorithms 4. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2007.

[13] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," University of Tennessee, Tech. Rep. UT-CS-94-270, 1995.

[14] D. del Castillo-Negrete, S. P. Hirshman, D. A. Spong, and E. F. D'Azevedo, "Compression of magnetohydrodynamic simulation data using singular value decomposition," J. Comput. Phys., vol. 222, pp. 265–286, March 2007.

[15] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," Journal of the American society for information science, vol. 41, no. 6, pp. 391–407, 1990.

[16] R. B. Cattell, "The Scree Test For The Number Of Factors," Multivariate Behavioral Research, vol. 1, no. 2, pp. 245–276, 1966.

[17] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "Kegg for representation and analysis of molecular networks involving diseases and drugs," Nucleic Acids Research, no. 38, pp. D355–D360, 2010.

[18] E. Howe, K. Holton, S. Nair, D. Schlauch, R. Sinha, and J. Quackenbush, "Mev: Multiexperiment viewer," Biomedical Informatics for Cancer Research, pp. 267–277, 2010.

[19] D. J. Leahy, "Implications of atomic-resolution structures for cell adhesion," Annual Review of Cell and Developmental Biology, vol. 13, pp. 363–393, November 1997.

[20] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," Journal of Molecular Biology, vol. 215, pp. 403–410, 1990.

[21] J. Rodgers, T. Stone, M. Barrett, B. Bradley, and P. Kennedy, "Kynurenine pathway inhibition reduces central nervous system inflammation in a model of human african trypanosomiasis," Brain, vol. 132, no. 5, pp. 1259–1267, May 2009.

[22] M. Holtze, P. Saetre, S. Erhardt, L. Schwieler, T. Werge, T. Hansen, J. Nielsen, S. Djurovic, I. Melle, O. A. Andreassen, H. Hall, L. Terenius, I. Agartz, G. Engberg, E. G. Jansson, and M. Schalling, "Kynurenine 3-monooxygenase (kmo) polymorphisms in schizophrenia: An association study," Schizophrenia Research, vol. 127, no. 1-3, pp. 270–272, 2011.

[23] J. Semba, "Glycine therapy of schizophrenia; its rationale and a review of clinical trials," Nihon Shinkei Seishin Yakurigaku Zasshi, vol. 18, no. 3, pp. 71–80, 1998.

[24] U. Heresco-Levy, M. Ermilov, P. Lichtenberg, G. Bar, and D. C. Javitt, "High-dose glycine added to olanzapine and risperidone for the treatment of schizophrenia." Biological psychiatry, vol. 55, no. 2, pp. 165–171, Jan. 2004.

[25] E. M. Zdobnov and R. Apweiler, "InterProScan – An integration platform for the signature-recognition methods in InterPro," Bioinformatics, vol. 17, no. 9, pp. 847–848, September 2001.

Compression- based Algorithms for Comparing Fragmented Genomic Sequences

Ramez Mina, Dhundy Bastola, Hesham H. Ali
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE, USA
Email: $\{$*rmina,dkbastola,hali*$\}$*@unomaha.edu*

***Abstract*—Sequence comparison is a fundamental tool in bioinformatics research since it helps to distinguish one sequence from another in terms of structure and function. Typically, methods such as global or local alignment are the preferred tools to measure a distance between sequence samples. Although they are often suitable tools for differentiation work, they could give erroneous results when the sequence data includes sequencing errors, gaps, repeats, and trans-locations which interfere with alignment methods. Next Generation sequence assembly tasks produce an enormous number of contigs and are reliant on alignment technologies to correctly place adjacent contigs together in the final sequence. If these alignment methods are confused by interruptions (i.e., fragmentation, gaps, mismatches or other blemishes) in the sequence data, then the assembly task may not be successful. We therefore suggest that sequence comparison can be successfully performed using alignment-free technologies and sequence compression methods which are less sensitive to inherent faults in sequencing tasks. In this paper, we evaluate different compression complexities and describe the use of compression algorithms for comparing biological sequence data. We analyze algorithm performance using protein sequence data and mitochondrial genomes with differing levels of interruption. Mitochondria is small dataset but is a well studied medium and is suitable to describe the effectiveness of the Lempel-Ziv complexity, Kolmogorov complexity using Lempel-Ziv-Welch, and Kolmogorov complexity using the Huffman coding schemes. We conclude our study by showing that sequence comparison via compression techniques is largely successful and could be a major help to high-throughput next-generation sequencing projects.**

***Keywords*-compression algorithms; Kolmogorov complexity; Lempel-Ziv complexity; tree path difference; next generation sequencing;**

I. INTRODUCTION

At the core of bioinformatics research is the comparison of sequence data. Since the 1970's, computational tools implementing local and global alignments were recommended methods to detect alterations between sequences. ClustalW and ClustalX [1] are example of such tools widely used in comparative work. However, they are not always appropriate for expansive orders of data. For this reason, heuristics such as Blast [2] and Blat [3] are alternative approaches.

Sequencing technology as well as the sequence assembly algorithms are continiously evolving. The alignment algorithms of complexity $o(n^2)$, which are used to determine the read placements for genome construction, are often slow to produce results [4]. Furthermore, as the sequencing technologies begin to produce longer reads, these algorithms may soon become obsolete and make way for other forms of sequence comparison. More importantly, the analysis by these alignment methods may be in accurate due to sequence noise such as, mutations, trans-locations and similar natural sequence altering factors [5].

Alignment-free methods are becoming increasing popular due to increased number of sequencing projects. These methods do not depend on base-by-base comparison but, instead, depend on the comparison of distributions of elemental frequencies in the sequences. For instance, the similarity of two sequences are determined by comparing frequency distributions. The generation of the frequencies and their comparison, typically a task of linear complexity, may easily take less time to run than to employ a traditional alignment algorithm of a $o(n^2)$ complexity. During a sequencing project where much data must be applied to alignment algorithms to determine adjacent reads in a genome, there is clearly a mounting demand to spend less time in the alignment bottleneck.

A. Next Generation Sequencing

Next generation sequencing, a major advancement of the Sanger sequencing technologies of the 1980's are able to generate as much data in 24 hours as several hundred Sanger-type DNA capillary sequencers [6]. They also produce a variety of different sizes of reads [7] [8]. When these reads are placed together in the correct order then a genome can be constructed. However, gaps often appear in the scaffolding that must be manually filled-in using reference sequences. This process can take a long time and could result in many inaccuracies in the completed genome.

Although recent research introduced alignment-based methods for the next generation sequencing as in Schatz M. C. [9], these new techniques did not eliminate the process of predicting the gaps between the fragments. Therefore the alignment-based method would still be a time-consuming and inaccurate approach. In particular, sequence alignment may fail to identify the distance between genomes, as the filled gaps are based on references that could be incorrect and produce inaccuracies. Therefore, alternative methods which are able to deal with reads of different sizes and orders are in demand such as alignment-free methods. These methods have been highlighted in the last two decades and have attracted much attention to address the abundance of data from bioinformatics research [10].

Compression-based techniques for comparing biological sequences would be a better method for comparing these reads. The fact that compression-based techniques generally run in linear time and are capable of identifying the distance between groups of reads by an analysis of elemental frequencies may be able to create more accurate results and help to expedite the completion of assembly projects. In this paper, we intend to support our hypothesis that compression-based techniques are comparable with alignment-based methods. We provide evidence from experimental results on mitochondrial data-sets that fragmented sequence data is able to be conveniently processed for sequence comparison by compression algorithms based on Lempel-Ziv and Lempel-Ziv-Welsh, Kolmogorov and nearest-neighbor clustering.

B. Alignment-free Methods

Earlier work has been done to evaluate compression-based techniques for the comparison of mitochondrial genomes, such as the works by [11] and [12].

However, to the best of our knowledge, comparing mitochondrial genomes with interrupted or incomplete data has not yet been addressed. Here we employ data from mitochondria because it is generally of a convenient size and is generally agreed upon to cover a large breadth of sequence structure and form which may encompass many of the kinds of obstacles encountered in nuclear DNA. This paper is based on the hypothesis that, compared to alignment-based techniques, compression-based techniques will provide a better measure to determine the relatedness between genomes, which are constantly being subjected to various natural events such as rearrangements, inversion, and trans-location. Additionally, there are many genome sequences that show sequence assembly errors, many sequences that are incomplete from their unordered fragments. Therefore these events, whether natural or simply associated with the sequencing technology, may seriously affect the development of software solutions used in the automation of the genome assembly and sequence comparison process.

Consequently, closing the genomic sequencing is one of the most time-consuming steps in the entire genome sequencing and annotation pipeline. Therefore, the need for computer algorithm(s) that accommodate the features of the data and help to overcome the limitations associated with the data is highly desirable. Alignment-free methods are suited for this work since they analyze and compare elemental frequencies across sequences. Their results can be conveniently described by trees of relatedness. Therfore, in the present study, we use a mathematical method to compare these trees, which allows for a comparison of results obtained with different alignment-free or compression-based techniques for the same data-set. We used a standard algorithm for tree comparison with a modified representation of the results in order to normalize them.

C. Background on Compression

The development of data compression techniques in computer science was motivated by the need to reduce network traffic when transmitting large amounts of data. In addition, storage was also a lending factor to this development. Compression methods from computer science became popular research topics in bioinformatics research when it was noted that DNA, appearing random, could not be easily efficiently compressed by Gzip or Bzip2 [13]. DNA has since been shown not to be as random as previously thought [14], and can be applied to compression techniques using only two bits. It was established that DNA had a syntax for coding genes [15] and furthermore that this information could be applied to compression techniques. These techniques derived elemental frequencies from the syntax to be compared with other kinds of sequence data.

Lempel and Ziv, along with Kolmogorov, introduced the concept of compression complexity, which later became the gateway for introducing the Lempel-Ziv compression technique in 1976. It is the complexity of a sequence that enables us to evaluate whether or not a particular compression algorithm is applicable. In [13] and [16] it was discussed that, in higher eukaryotes, biological sequences have tandem repeats and multiple copies of genes, which make them a good subject for compression techniques. In addition to these properties, DNA sequences are rich with other properties that are hidden within the sequences. These properties could be useful for compression since they include natural evolutionary events such as random mutations, translocation, cross-overs, and reversal events. In [16] it was discussed how compression would address such properties and take advantage of them to compress the sequences. The compression would then reflect the relatedness between the sequences. By concatenating two sequences we would be able to compress them effectively if they share common information.

D. Kolmogorov Complexity

For any two sequences x and y, we define conditional Kolmogorov complexity, $K(x|y)$, as the shortest binary program that computes x in terms of y [4]. Also, the Kolmogorov complexity of a sequence x we defined as $K(x)$ or $K(x|\lambda)$, where λ signifies an empty string. We also define the information distance ID between two sequences x and y as, $ID(x,y) = max\{K(x|y), K(y|x)\}$

The Kolmogorov complexity of a sample of information, such as text, is a measure of the computational resources needed to specify the sample. Kolmogorov theory is a concept more than a measure and does not offer a metric value that could be used in constructing a tree of relatedness. The Universal Similarity Metric (USM) was thus implemented to measure the complexity of Kolmogorov, where we represent the compression of sequence x by $C(x)$ and the compression of sequence x appended by sequence y by $C(yx)$. Three practical approximations of Kolmogorov were suggested, namely:

- Universal Compression Distance/Dissimilarity (UCD)
- Normalized Compression Distance/Dissimilarity (NCD)
- Compression Distance/Dissimilarity (CD).

In mathematical terms, we have the following,

$$UCD(x,y) = \frac{max\{|C(xy)| - |C(x)|, |C(yx)| - |C(y)|\}}{max\{|C(x)|, |C(y)|\}}$$

$$NCD_1 = \frac{\{|C(xy)| - min\{|C(x)|, |C(y)|\}\}}{max\{|C(x)|, |C(y)|\}}$$

$$CD(x,y) = \frac{min\{|C(xy)|, |C(yx)|, |C(x)| + |C(y)|\}}{|C(x)| + |C(y)|}$$

$$then,$$

$$NCD(x,y) = min\{NCD_1(x,y), NCD_1(y,x)\}$$

E. Lempel-Ziv Complexity

Consider the sequence $S = AACGTACC$. Some of its histories, or fashions of grouping and placing adjacent components of the sequence together as defined by [5], are the following:

1) $H(S) = A \cdot A \cdot C \cdot G \cdot T \cdot A \cdot C \cdot C$
2) $H(S) = A \cdot AC \cdot G \cdot T \cdot A \cdot C \cdot C$
3) $H(S) = A \cdot AC \cdot G \cdot T \cdot ACC.$

The exhaustive history, presented by the same authors, is defined as the history where no substring has a repetition, and no substring can be found in the whole sequence before this substring. This means that if a substring is chosen at the i^{th} position, then the sequence of characters before this position will not contain an occurrence of this substring. A mathematical representation for this concept could be the resulting number of components which making up an entire sequence, here called a *unique history*. Upon examining the histories in the previous example, we note that the first two cannot be exhaustive histories since 'A' and 'C' are repeated, but the third one is exhaustive. The Lempel-Ziv-complexity is defined as the least exhaustive history of a sequence and is noted as $C(sequence)$ implying the number of components in a an exhaustive history of a sequence. Consider the following three sequences:

- $S = AACGTACCATTG$
- $R = CTAGGGACTTAT$
- $Q = ACGGTCACCAA$

With the component words, separated by dots making up the entire sequence, the exhaustive histories for the sequences are the following:

- $HE(S) = A \cdot AC \cdot G \cdot T \cdot ACC \cdot AT \cdot TG$
- $HE(R) = C \cdot T \cdot A \cdot G \cdot GGA \cdot CTT \cdot AT$
- $HE(Q) = A \cdot C \cdot G \cdot GT \cdot CA \cdot CC \cdot AA$

The total number of components of these exhaustive histories are the following:

- $c(S) = c(R) = c(Q) = 7$

The exhaustive histories for SQ and RQ are:

- $HE(SQ) = A \cdot AC \cdot G \cdot T \cdot ACC \cdot AT \cdot TG \cdot ACGG \cdot TC \cdot ACCAA$
- $HE(RQ) = C \cdot T \cdot A \cdot G \cdot GGA \cdot CTT \cdot AT \cdot ACG \cdot GT \cdot CA \cdot CC \cdot AA$
- $c(RQ) = 12$ and $c(SQ) = 10$

This implies that S is closer to Q than R is to Q, which is evident by the following:

- $S = AACGTACCATTG$
- $Q = ACGGTCACCAA$
- $Q = ACGGTCACCAA$
- $R = CTAGGGACTTAT$

Lempel-Ziv complexity itself is not a distance measure between sequences. It is instead a form of distance measurement.

Distance measure 1:
$d(S,Q) = max\{c(SQ) - c(S), c(QS) - c(Q)\}$

Distance measure 2:
$d^{*}(S,Q) = \frac{max\{c(SQ)-c(S), c(QS)-c(Q)\}}{max\{c(S), c(Q)\}}$

Distance Measure 3:
$d_1(S,Q) = c(SQ) - c(S) + c(QS) - c(Q)$

Distance Measure 4:
$d_1^{*}(S,Q) = \frac{c(SQ)-c(S)+c(QS)-c(Q)}{c(SQ)}$

These distances would be the same as the scoring values of any sequence alignment method and would be used in building the tree of relatedness of the data-set. Notice that the shorter the numerical distance, the closer the pair sequences are to each other.

The rest of this paper is organized as follows:

- **Section 2**. Two different compression techniques are tested with different parameters, namely Kolmogorov complexity and Lempel-Ziv complexity, on a nucleic acid sequence (mitochondria) and on protein sequences from different species. Random incomplete genome fragments are then generated with different percentages, where these fragments could be ordered or disordered. Trees are then generated for both compression-based techniques and for multiple sequence alignment, a method of comparing similarity across more that two sequences.
- **Section 3**. These trees are compared against the standard tree, which serves as a reference for each data-set, while calculating the distance between the two trees.
- **Section 4**. We present our conclusions regarding the usefulness of the compression-based algorithms for sequence comparison.

II. Methodology

We start with the experimental design, then collect the data-sets, and finally apply the steps of each experiment to evaluate our hypothesis.

A. *Experimental Design*

The experiments consisted of three phases including data-set assembly, scoring matrices compilation and calculation, and evaluation of results.

The Lempel-Ziv-Welch and Huffman compression algorithms which rely on prefix coding, were the seeds for Kolmogorov complexity metrics. The Lempel-Ziv complexity has its own algorithm to measure the complexity before seeding it to the metrics, implemented with a modified algorithm published by Borowska et al. [17]. All four of the distance measures were calculated for Lempel-Ziv complexity.

B. *Dataset Collection*

Our data-sets varied according to the experiment. The first experiment used both protein and whole genome mitochondrial sequences (CK-36-PDB and AA-15-DNA). The other four experiments used only a mitochondrial data-set (AA-15-DNA). These two data-sets were used to test the viability of compression techniques in comparing biological sequences. This data has been previously used and consists of 36 protein domains in the amino acid sequence set and the genome data consists of complete DNA sequences 15 different organisms [4].

The second experiment focused on comparing incomplete sequences, containing only 10 - 90 percent of total genome sequences, and the start positions varied as shown (Figure 1). The third experiment evaluated incomplete genomes made from separate segments, but the total length contained 10 - 90 percent of the whole genomes (Figure 2). The fourth experiment explored genomes that were 10 - 100 percent incomplete, containing several shuffled fragments combined together (Figure 3). The fragments were placed in random order using the Fisher-Yates algorithm, an algorithm which generates a random permutation of a finite set [18]. The fifth experiment dealt with sequences with variants. The variants were obtained with different percentages and reflected point-mutations seen in nature.

C. *Sequence Comparison*

For each experiment, multiple sequence alignment was used to analyze each data-set. To accomplish this, the MUSCLE [19] package, a software used to compute the multiple sequence alignment for protein and nucleotide sequences, was also employed.

The comparison between trees was accomplished by estimating the path-length-difference metric as described in Felsenstein [20]. For this, a matrix was constructed for each tree. The size of the matrix is m^2, where m is the number of tree leaves (the species), and each cell in the matrix has the number of branches that separates the species of the corresponding row and column. The squared difference was computed between each cell in a matrix and its representative in the gold standard tree matrix. The distance was then calculated by finding the square root of the sum of the cells where we took care not to include duplicate values. The distance was normalized by dividing the distance by the summation of the distances between each of the cells in the gold standard tree.

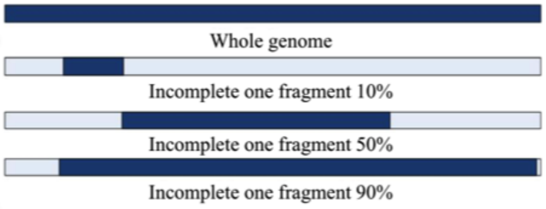

Whole genome
Incomplete one fragment 10%
Incomplete one fragment 50%
Incomplete one fragment 90%

Figure 1. Cartoon diagram depicting the imperfection in genome sequence generated by choosing different fragment lengths from the original whole genome sequence.

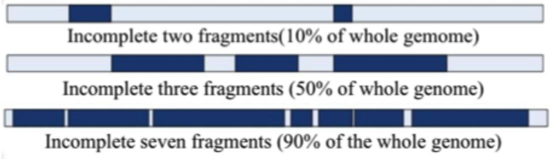

Incomplete two fragments(10% of whole gemome)
Incomplete three fragments (50% of whole genome)
Incomplete seven fragments (90% of the whole genome)

Figure 2. Cartoon diagram depicting the imperfection in genome sequence generated by choosing fragments from different regions of the whole genome.

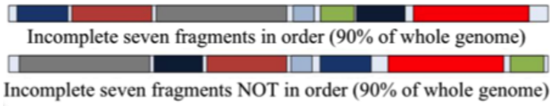
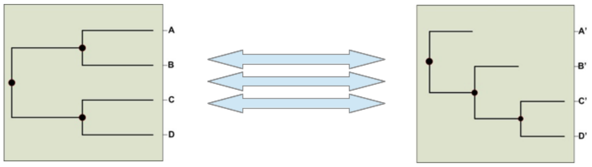

Incomplete seven fragments in order (90% of whole genome)
Incomplete seven fragments NOT in order (90% of whole genome)

Figure 3. Cartoon diagram depicting the imperfection in genome sequence generated by choosing fragments of different length and order in the whole genome



$$
\begin{aligned}
Distance &= \sqrt{(AB-A'B')^2+(AC-A'C')^2+(AD-A'D')^2+(BC-B'C')^2+(BD-B'D')^2+(CD-C'D')^2} \\
&= \sqrt{(2-3)^2+(4-4)^2+(4-4)^2+(4-3)^2+(4-3)^2+(2-2)^2} \\
&= \sqrt{1^2+0^2+0^2+1^2+1^2+0^2} = \sqrt{1+0+0+1+1+0} = \sqrt{3}
\end{aligned}
$$

Figure 4. Two hypothetical trees. We show the calculations to determine tree distance between reference and generated trees.

ISBN: 978-1-61208-260-8

Consider the two trees in Figure 4, where the tree on the left represents the gold standard tree (species A, B, C, and D), and the second tree on the right represents the output tree of an algorithm (species A’, B’, C’, and D’). The scoring matrices of Table I were calculated by summing the edges between two nodes in a tree. The distance between these two trees was calculated by finding the mean root square, noted in Figure 4. Their distance, $\sqrt{3}$, is normalized by division of the sum of the distances between the species in the gold standard tree. The calculation of this sum of distances is the following: (AB + AC + AD + BC + BD + CD) = (2 + 4 + 4 + 4 + 4 + 2) = 20. The normalized distance between the two trees is $\frac{\sqrt{3}}{20} \approx 8.66\%$.

Table I
SCORING MATRIX FOR TREES IN FIGURE 4

	A	B	C	D		A’	B’	C’	D’
A	0	2	4	4	A’	0	3	4	4
B	2	0	4	4	B’	3	0	3	3
C	4	4	0	2	C’	2	3	0	2
D	4	4	2	0	D’	4	3	4	0

III. RESULT AND ANALYSIS

The results show the performance of compression-based methods over different kinds of data-sets. To evaluate these methods over data-sets with different features that reflected imperfection in the input sequence data, we started with a data-set that was error-free (according to NCBI), then we picked a genomic data-set and manufactured data-sets with errors to incorporate imperfection in the sequence data.

Results are shown for the phylogenies generated from the compression-based methods from multiple sequence alignment (Table II). The purpose of having results from multiple sequence alignment is to evaluate whether compression-based methods were similar, worse, or better than multiple sequence alignment with data-sets of different quality. These results are the distances between the calculated trees and the gold standard tree. These distances reflect the quality of clustering for the species, based on the pair-wise distances generated from the methods, (i.e., the scoring matrices). The column labeled as *Variant*, lists the different distance measures which calculated for the first experiment. Table II shows the results from trees created by neighbor-joining methods, and also the UPGMA methods.

Shaded cells in Table II indicate the cases where compression-based algorithm performed better than multiple sequence alignment

A. *Analysis of Datasets with No Errors*

The first experiment determined the feasibility of using compression-based algorithms in phylogenetic analysis of sequence data. The goal was to test the algorithms against regular data-sets that are error-free and helps to evaluate whether these methods are capable of measuring the distances of normal data-sets. We compare the results obtained from various versions of compression-based sequence comparison with results obtained from multiple sequence alignment. In this experiment, two data-sets were used: a set of protein sequences and a set of complete mitochondrial genomes. The gold standard trees for both data-sets were available to provide the base line comparison. Tables II and III display the results for the first experiment. The shaded cells reveal the compression techniques that surpassed multiple sequence alignment. In the protein data-set (Table III), the consistently desirable results were derived from UPGMA clustering using the scoring matrices of both Kolmogorov and Lempel-Ziv complexities.

Table II
COMPARISON OF COMPRESSION ALGORITHMS AND MULTIPLE SEQUENCE ALIGNMENT FOR THE PROTEIN DATASET CK-36-PDB IN EXPERIMENT 1

.

Test		**Protein data-set CK-36-PDB**	
Algorithm	Variant	Neighbor-Joining	UPGMA
Kolmogorov using Huffman coding	CD	2.395244	3.169468
	NCD	2.328382	2.264505
	UCD	2.328382	2.264505
Kolmogorov using LZW compression	CD	2.176959	2.165911
	NCD	2.210704	2.215544
	UCD	2.305268	2.238781
Lempel - Ziv complexity	Dist 1	2.345943	2.280642
	Dist 2	2.330589	2.219562
	Dist 3	2.26719	2.287058
	Dist 4	2.272324	2.306048
Multiple Sequence Alignment		2.370071	1.937603

In the mitochondrial data-set (Table III), only Lempel-Ziv outperformed multiple sequence alignment. These results clearly indicate that compression-based sequence comparison provides a valid measure of similarity for biological sequences.

These measurements are comparable to the ones produced by multiple sequence alignment and outperform alignment in several instances. It is also clear that a careful selection of the clustering algorithm, compression methods, and associated distance measure can improve the overall results. As in Table II, shaded cells in Table III indicate outcomes that are better than multiple sequence alignment.

The purpose here was two-fold: first, to determine if the imperfection in the quality of sequence data and the choice of compression-based methods used impacted the outcome, and second, to determine which method would be a better solution for the type of imperfection in the data-sets. For this purpose, the mitochondrial genomes were used, incrementally removing percentages of genome and using

Table III
COMPARISONS OF COMPRESSION ALGORITHMS AND MULTIPLE SEQUENCE ALIGNMENT FOR THE MITOCHONDRIAL GENOME DATASET AA-15-DNA, IN EXPERIMENT 1

Test		Mitochondrial Genome data-set AA-15-DNA	
Algorithm	Variant	Neighbor-Joining	UPGMA
Kolmogorov using Huffman coding	CD NCD UCD	7.871585 7.871582 7.871582	7.871585 7.871582 7.871582
Kolmogorov using LZW compression	CD NCD UCD	3.034474 2.797647 2.878755	3.034474 2.797647 2.878755
Lempel-Ziv complexity	Dist1 Dist2 Dist3 Dist4	1.554705 1.554705 1.554705 1.554705	1.357058 1.357058 1.357058 1.357058
Multiple Sequence Alignment		1.5547053	1.878762

an algorithm to randomly choose the starting position of the remaining genome (refer back to Figure 1).

Upon examining the neighbor-joining method and UPGMA (Figures 5 and 6), Lempel-Ziv complexity surpassed multiple sequence alignment in all the trials (with both neighbor-joining and UPGMA clustering), except for in one case. Kolmogorov with Lempel-Ziv-Welch had viable results but was not competitive to Lempel-Ziv. These results showed that Lempel-Ziv complexity offered the most likelihood of revealing the similarities between the genomes. Despite the variations in the length of the genomes, Lempel-Ziv was able to address the dissimilarities between the sequences.

B. Analysis of Sequences Data with Incomplete Fragments that Are Not Continuous

This experiment was an expansion of the second experiment, where the genome was broken into several pieces, and the total size of the sequence was reduced to the same 10-90 percent but where each fragment was allowed to be a different random size (refer back to Figure 2).

Multiple fragments were then combined together and tested. The results obtained here with both the neighbor-joining method and UPGMA (Figure 7 and 8) mirrored the earlier results (in the second experiment) in that Lempel-Ziv complexity outperformed multiple sequence alignment in almost every percentile. Also, Kolmogorov using Lempel-Ziv-Welch compression and Kolmogorov, using Huffman coding, failed to perform better than multiple sequence alignment (results not presented).

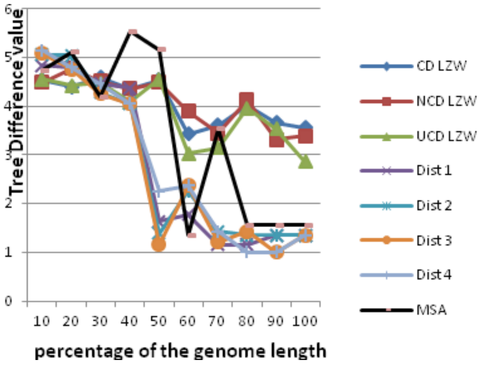


Figure 5. Analysis of the mitochondrial genomes using Neighbor-Joining.

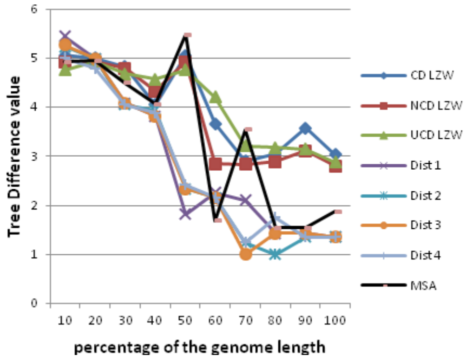


Figure 6. UPGMA clustering on the distances obtained with different algorithms.

C. Analysis of Datasets with Incomplete Fragments that Are Not Continuous and Not in Order

This experiment was designed to establish the goodness of fit of multiple sequence alignment and compression algorithms. The genomes for this experiment were cut into multiple fragments, randomly decreased in length to a total 10-100 percent of the original size, and then rearranged (refer back to Figure 3).

While the compression algorithms returned results similar to the previous experiments, multiple sequence alignment performed much worse (Figure 9). For the incomplete genomes less than 50 percent in length, Kolmogorov using Lempel-Ziv-Welch and Lempel-Ziv both surpassed multiple sequence alignment, but Kolmogorov was overtaken by multiple sequence alignment at 60 percent and above.

In this experiment where the data-set depicted translocation of DNA fragments, multiple sequence alignment performed very poorly and failed to detect the relatedness between the genomes. However, the compression-based method of Lempel-Ziv still detected the relationships among genomes and gave an accurate clustering. Even Lempel-Ziv-Welch was competitive with multiple sequence alignment in finding the right dissimilarities between the genomes.

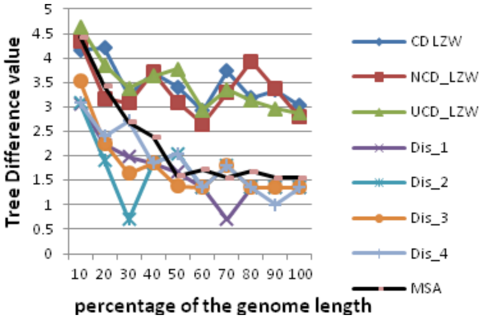


Figure 7. Analysis of sequence data of unequal length with data-set as shown in Figure 2 using neighbor-joining.

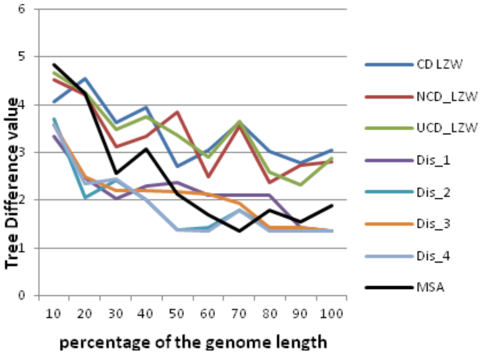


Figure 8. UPGMA clustering of the distances obtained with shown algorithms.

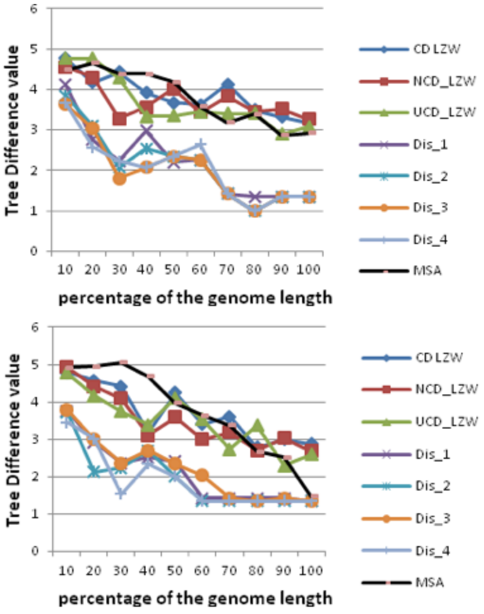


Figure 9. Analysis of sequence data of varying length with data-set as shown in Figure 3 using neighbor-joining (top figure) and UPGMA (bottom figure) clustering of the distances obtained with shown algorithms.

D. Analysis of Sequence Data-sets Containing Mutated Nucleotides

This experiment was designed to evaluate the performance of the compression-based methods on a mutated data-set. As the sequences mutate, it is difficult for methods like multiple sequence alignment to identify the relatedness among species. Mutations (point mutations) were taken with percentages of 1, 3, 5, and 7 percent. Comparison of the results to multiple sequence alignment was conducted in the same manner described earlier for sequence data with different fragment lengths and for measuring the distance between the resulting trees to the gold standard tree. The results obtained from this experiment are shown in Table IV.

As we can see, the shaded cells contain the results of Lempel-Ziv complexity, which performed relatively better than other compression or non-compression methods. With reasonable mutations, which typically would result in changes in the functions of the species but not in an evolution of the species itself, Lempel-Ziv complexity performed best and was able to detect the similarities among the species when compared to the multiple sequence alignment method. Kolmogorov complexity failed to detect similarities with this data-set.

IV. Conclusion

Compression-based techniques provide a viable alternative to multiple sequence alignment that is typically used to compare biological sequence data. In cases where the data-sets contained errors, gaps, or the arrangement of DNA fragments, compression-based techniques performed better than alignment in our experiments. Compression algorithms were also faster than alignment, particularly for large sequences. Of the three compression techniques examined in this study, Lempel-Ziv complexity performed the best in classifying the incomplete and highly imperfect data-sets.

To summarize these results, Lempel-Ziv complexity led in performance to the alignment-free techniques and even outperformed multiple sequence alignment in several cases. From the results obtained with the different experiments, we can see that compression techniques in general, and Lempel-Ziv in particular, were able to capture the relatedness among the input sequences and were less impacted by the incompleteness or rearrangement of the fragments.

V. Acknowledgment

The authors would like acknowledge the help by support staff in the UNO-Bioinformatics Core Facility, funded by the grants from the National Center for Research Resources (5P20RR016469) and the National Institute for General Medical Science (NIGMS) (8P20GM103427). We also like to thank Oliver Bonham-Carter for his help in editing and formatting of this manuscript.

Table IV
COMPARISON OF THE PERFORMANCE OF COMPRESSION AGAINST MULTIPLE SEQUENCE ALIGNMENT ON A MUTATED DATA-SET WITH MUTATION PERCENTAGES OF 1, 3, 5, AND 7 PERCENT.

		NJ (1 percent)		UPGMA (3 percent)		NJ (5 percent)		UPGMA (7 percent)	
Kolmogorov using Huffman coding	CD	7.184	7.872	7.184	7.872	7.184	7.872	7.184	7.872
	NCD	7.054	7.872	7.054	7.872	7.054	7.872	7.054	7.872
	UCD	7.054	7.872	7.054	7.872	7.054	7.872	7.054	7.872
Kolmogorov using Lempel-ZivW compressions	CD	3.201	3.266	3.443	3.152	3.643	3.097	3.696	3.009
	NCD	3.272	2.996	3.278	2.791	3.324	3.487	3.387	2.964
	UCD	3.41	3.041	3.128	2.99	3.278	2.707	3.537	3.003
Lempel and Ziv complexity	Dist1	1.357	1.357	1.357	1.774	1.858	2.101	2.054	2.276
	Dist2	1.357	1.357	1.357	1.357	1.357	1.53	1.357	1.357
	Dist3	1.357	1.357	1.357	1.774	1.357	2.7	1.159	2.276
	Dist4	1.357	1.357	1.357	1.542	2.017	1.426	1.357	1.357
Multiple Sequence Alignment		1.555	1.879	1.555	1.774	1.555	1.357	1.685	1.879

REFERENCES

[1] M. Larkin, G. Blackshields, N. Brown, R. Chenna, P. McGettigan, H. McWilliam, F. Valentin, I. Wallace, A. Wilm, R. Lopez *et al.*, "Clustal w and clustal x version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[2] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman *et al.*, "Basic local alignment search tool," *Journal of molecular biology*, vol. 215, no. 3, pp. 403–410, 1990.

[3] W. Kent, "Blatthe blast-like alignment tool," *Genome research*, vol. 12, no. 4, pp. 656–664, 2002.

[4] P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente, "Compression-based classification of biological sequences and structures via the universal similarity metric: experimental assessment," *BMC bioinformatics*, vol. 8, no. 1, p. 252, 2007.

[5] H. Otu and K. Sayood, "A new sequence distance measure for phylogenetic tree construction," *Bioinformatics*, vol. 19, no. 16, pp. 2122–2130, 2003.

[6] S. Schuster, "Next-generation sequencing transforms todays biology," *Nature*, vol. 200, no. 8, 2007.

[7] H. Li and N. Homer, "A survey of sequence alignment algorithms for next-generation sequencing," *Briefings in Bioinformatics*, vol. 11, no. 5, pp. 473–483, 2010.

[8] G. Zhang, I. Fedyunin, S. Kirchner, C. Xiao, A. Valleriani, and Z. Ignatova, "Fanse: an accurate algorithm for quantitative mapping of large scale sequencing reads," *Nucleic acids research*, vol. 40, no. 11, pp. e83–e83, 2012.

[9] M. Schatz, C. Trapnell, A. Delcher, and A. Varshney, "High-throughput sequence alignment using graphics processing units," *BMC bioinformatics*, vol. 8, no. 1, p. 474, 2007.

[10] S. Vinga and J. Almeida, "Alignment-free sequence comparisona review," *Bioinformatics*, vol. 19, no. 4, pp. 513–523, 2003.

[11] M. Li, J. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny," *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[12] D. Burstein, I. Ulitsky, T. Tuller, and B. Chor, "Information theoretic approaches to whole genome phylogenies," in *Research in Computational Molecular Biology*. Springer, 2005, pp. 992–992.

[13] X. Chen, S. Kwong, and M. Li, "A compression algorithm for dna sequences and its applications in genome comparison," in *Proceedings of the fourth annual international conference on Computational molecular biology*. ACM, 2000, p. 107.

[14] B. Behzadi and F. Le Fessant, "Dna compression challenge revisited: a dynamic programming approach," in *Combinatorial Pattern Matching*. Springer, 2005, pp. 85–96.

[15] Y. Neuman, "Meaning-making in language and biology," *Perspectives in biology and medicine*, vol. 48, no. 3, pp. 317–327, 2005.

[16] E. Rivalsy, O. Delgrangez, J. Delahayey, and M. Dauchety, "Compression and sequence comparison," 1994.

[17] M. Borowska, E. Oczeretko, A. Mazurek, A. Kitlas, and P. Kuc, "Application of the lempel-ziv complexity measure to the analysis of biosignals and medical images," *Annual proceedings of Medical Science*, 2005.

[18] R. Fisher, F. Yates *et al.*, "Statistical tables for biological, agricultural and medical research." *Statistical tables for biological, agricultural and medical research.*, no. Ed. 3., 1949.

[19] R. Edgar, "Muscle: multiple sequence alignment with high accuracy and high throughput," *Nucleic acids research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[20] J. Felsenstein *et al.*, *Inferring phylogenies*. Sinauer Associates Sunderland, 2004, vol. 2.

Controllability of a Model of Treatment Response to Combined Anticancer Therapy

Andrzej Świerniak
Department of Automatic Control
Silesian University of Technology
Gliwice, Poland
e-mail: andrzej.swierniak@polsl.pl

Jerzy Klamka
Department of Automatic Control
Silesian University of Technology
Gliwice, Poland
e-mail: jerzy.klamka@polsl.pl

***Abstract*—The controllability of a combination of antiangiogenic treatment and chemotherapy for cancer is considered in the paper. The treatment is modeled as a two-dimensional control action in the second order dynamical system described by a model belonging to a class proposed so far. Sufficient conditions of local constrained controllability are found and verified for the model and their biological interpretation is presented.**

Keywords-controllability; dynamics; semilinear systems; biomathematical modelling; cancer therapy.

I. INTRODUCTION

Controllability is a qualitative property of dynamical control systems and its meaning, roughly speaking, is the following: a dynamical system is controllable if it is possible to steer it from an arbitrary initial state to an arbitrary final state using the set of admissible controls. In the existing literature there are many different definitions of controllability strongly depending on the class of dynamical control systems (see, e.g., [1], [2] and references therein). In the present paper, we consider constrained local controllability problems for second-order finite-dimensional semilinear stationary dynamical systems described by a set of two ordinary differential state equations. More precisely, we discuss the control properties of a model belonging to a class proposed in [3] to which two control variables describing two treatment modalities have been introduced. The line of reasoning is similar to our previous study [4] in which however only antiangiogenic therapy was considered, in other words only one control variable was used. The results are based on theorems proved in [2]. The idea of the theorems is that under suitable assumptions the constrained global relative controllability of a linear first-order associated approximated dynamical system implies constrained local relative controllability near the origin of the original semilinear second-order dynamical system. The Hahnfeldt *et al.* model [3] is based on the assumption that tumor growth with an incorporated vascularization mechanism can be described by a Gompertz-type or logistic-type equation with variable carrying capacity which defines the dynamics of the vascular network. The main idea of this class of models is to incorporate the spatial aspects of the diffusion of factors that stimulate and inhibit angiogenesis into a non-spatial two-compartmental model for cancer cells and vascular endothelial cells. The control properties of such models in the context of combined therapy were discussed among others in [5], [6] and [7]. In [5], following the line of reasoning proposed by d'Onofrio and Gandolfi in [8], conditions for asymptotic tumor eradication by constant and periodic therapy were given. Moreover, in [5] and [6], the necessary conditions for optimal treatment protocols in a given finite time were considered. The interesting finding is that for the d'Onofrio-Gandolfi version of the model [8] the optimal trajectory does not contain singular arcs. This property has been found previously for a sub-class of the models of this class for antiangiogenic therapy [9], but for the remaining models from this class the existence of intervals of singular optimal control has been proved rigorously by Ledzewicz and Schattler [6], [10], [11]. All the considerations related to finite time control are however conditioned on the concept of controllability of the dynamical systems discussed which, to our knowledge, has not been analyzed by other authors except in our previous paper [4]. This is a major motivation for the present study.

In the second section we present the most important biological information related to the topic of our study. The mathematical model and its properties are presented and discussed in the third section. In section 4 we define a class of semilinear systems and we present some results related to controllability of such systems. Section 5 contains the most important results of our study dealing with controllability of the models of the combined anticancer therapy. Final remarks and conclusions are given in the section 6.

II. BIOLOGICAL BACKGROUND

Tumors, like normal tissues, have physiological constraints on growth, such as access to oxygen and nutrients for metabolism. The diffusion of oxygen in tissues is limited to a distance of about 150 μm, thus tissue growth is restricted to a few cubic millimeters if no new vasculature is formed. For vascularization to occur, the nearest vessel or capillary needs to become destabilized so that the endothelial cells lining the vessel can loosen from their neighbors and migrate through the extracellular matrix towards the tumor. Only after a tumor has recruited its own blood supply it can expand in size. Tumors do this via the production of angiogenic factors secreted into local tissues and stroma, a process termed the angiogenic switch. The angiogenic switch is a discrete step in tumor development that can occur at different stages in the tumor-progression pathway, depending on the nature of the tumor and its microenvironment. Since in normal healthy adults the process of angiogenesis is very limited, it should, at least in theory, be possible to inhibit tumor angiogenesis without affecting normal tissues. Antiangiogenic therapies proposed by Folkman in the early seventies of the previous century [12] have become one of the most promising approaches in anti-cancer drug development. Successful preclinical research data lead to clinical trials based on different strategies. Approaches currently under evaluation for inhibiting angiogenesis may either be direct (targeting cell surface bound proteins/receptors) or indirect (targeting growth factor molecules) [13]. The genetic instability and high mutation rate of tumor cells is responsible, in part, for the frequent emergence of acquired drug resistance to conventional cytotoxic anticancer therapy. In contrast, vascular endothelial cells, like bone marrow cells, are genetically stable and have a low mutation rate. Therefore Kerbel [14] proposed the hypothesis that antiangiogenic therapy would be a strategy to bypass drug resistance. In [15] the gap between preclinical (mouse models – localized primary tumor) and clinical testing (late-stage metastatic tumor) is noted; anti-angiogenic agents are not efficient at the level suggested by preclinical trials and different results have been observed depending on the disease stage. Biologists suggest that anti-angiogenic therapy might become an essential component of multidrug cancer therapy [16], [17], especially when combined with chemotherapy. One possible strategy is using angiogenesis inhibitors to normalize the abnormal vasculature and thereby to facilitate drug delivery [18], [19]. Some results from clinical studies of such combination therapy are shown in [16]; a dose of antiangiogenic agent (Bevacizumab 5 mg/kg) showed a significantly different (higher) median survival than chemotherapy alone, and a larger dose (10 mg/kg) even increased survival compared to chemotherapy alone. Several clinical trials of combined therapy have been discussed recently, and some examples are presented in [20]. Continuous treatment with angiogenic inhibitors ultimately leads to a decrease in tumor blood flow and a decreased tumor uptake of co-administrated cytotoxic drugs. In periodic therapy the main goal of anti-angiogenic agents is to normalize tumor vasculature which might facilitate recovery of tumor cells from cytostatic agents [16].

This is why when formulating objectives of the combined therapy mathematically, one should take into account final states which could be reached by the admissible control actions. The problem of the reachability could be solved by the respective conditions of controllability of the model.

III. MODEL OF COMBINED THERAPY AND ITS PROPERTIES

Hahnfeldt *et al.* [3] proposed a model of vascularized tumor development described by a self limiting growth mechanism (e.g. a Gompertz- or logistic-type equation) with a variable carrying capacity defining the dynamics of the vascular network. They proposed to treat the carrying capacity constraining the tumor growth as a varying tumor volume sustainable by the vessels and roughly proportional to the vessel volume. The complete model requires an additional equation describing changes of the volume of the vessels, and the equation below expresses Gompertz-type growth:

$$\frac{dN(t)}{dt} = -\beta N(t) \ln \frac{N(t)}{K(t)} \qquad (1)$$

where *N* represents tumor volume as the size of the cancerous cell population, *K* describes the maximum tumor volume sustainable by the supporting vascular network, and β is a growth parameter.

The models considered in the present study are based on that proposed by Hahnfeldt *et al* who have developed and biologically validated a two-dimensional model of ordinary differential equations for interactions between primary tumor volume and the carrying capacity of the vasculature network which in turn is proportional to the square of the tumor diameter. For simplification, it was necessary to assume spherical symmetry of the tumor mass. Therefore the expression for *K* has the following form:

$$\frac{dK(t)}{dt} = \gamma N(t) - \lambda N(t)^{\frac{2}{3}} K(t) - \mu K(t) \qquad (2)$$

where γ represents the effect of the stimulation, λ the effect of the inhibition, and μ the natural cell death. Taking into account that tumor growth is relatively slow compared to the rate of release of pro- and anti- angiogenic factors, it is possible to assume that parameters γ, λ, μ are constant. The model (1), (2) may be modified by introducing a logistic-type growth equation instead of the Gompertz-type one and by changing the ratio between stimulating and blocking angiogenic factors [8]. This leads to a set of models which although behaving similarly when uncontrolled, may have different control properties [9]. For example, all the models have the same equilibrium point which is both locally and globally asymptotically stable:

$$N^* = K^* = \left(\left(\gamma - \mu\right)/\lambda\right)^{3/2} \qquad (3)$$

On the other hand, conditions of tumor eradication under periodic therapy are both sufficient and necessary for all the models, except for the original Hanhfeldt model for which they are only necessary. Similar differences are observed when optimal antiangiogenic treatment protocols are considered. The original Hahnfeldt model contains singular arcs in optimal trajectories which are absent in other models [9], [10], [11]. To focus attention we consider the modification of the Hahnfeldt model proposed in [8]:

$$\frac{dK(t)}{dt} = \gamma K(t) - \lambda N(t)^{\frac{2}{3}} K(t) - \mu K(t) \qquad (4)$$

This model is strongly nonlinear, but by a logarithmic change of variables and some scaling transformation we are able to transform it into the semilinear form. More precisely, by the transformation:

$$\begin{aligned} & x = \ln N / N^*, y = \ln K / K^* \\ & x^* = y^* = 0, \tau = \beta t, \vartheta = (\gamma - \mu) / \beta \qquad (5) \\ & x' = dx / d\tau, y' = dy / d\tau \end{aligned}$$

we are led from model (1), (4) to the following semilinear system:

$$\begin{aligned} x'(t) &= y(t) - x(t), \\ y'(t) &= \vartheta(e^{(2/3)x(t)} - 1) \end{aligned} \qquad (6)$$

Application of antiangiogenic therapy can be incorporated in the model by a factor increasing multiplicatively the rate of loss of the vessels, which leads to the following equation:

$$\frac{dK(t)}{dt} = \gamma K(t) - \lambda N(t)^{\frac{2}{3}} K(t) - \mu K(t) - \eta K(t) u(t) \qquad (7)$$

where $u(t)$ denotes the dose of the agent scaled to its effect on the vascular network, and η is a constant parameter and plays the role of a control variable. For the constant dose U, the equilibrium points take the form:

$$N^* = K^* = \left(\left(\gamma - \mu - \eta U\right) / \lambda\right)^{3/2} \qquad (8)$$

which, according to the conditions of stability given in [8], leads to the conclusion that:

$$U = (\gamma - \mu) / \eta \Rightarrow K^*, N^* = 0 \qquad (9)$$

In other words, the vascular network and in turn the tumor can be eradicated, a conclusion which is crucial for the philosophy of the entire analysis. It is enough to ensure that the population of endothelial cells responsible for angiogenesis behaves in the required way because the size of the tumor population in some sense tracks the same transients. A similar line of reasoning could be applied in the case of combined antiangiogenic and chemotherapy when two control variables are present. The main difference is that chemotoxic agents kill both cancer and critical normal tissues including endothelial cells:

$$\frac{dN(t)}{dt} = -\beta N(t) \ln \frac{N(t)}{K(t)} - \psi v(t) \qquad (10)$$

$$\frac{dK(t)}{dt} = \gamma K(t) - \lambda N(t)^{\frac{2}{3}} K(t) - (\mu + \eta u(t) + \xi v(t)) K(t) \qquad (11)$$

where $v(t)$, the second control variable, denotes the dose of the chemotherapy scaled to its effect on tumor and normal tissues, and ξ and ψ are constant scaling parameters. Of course, the additional chemotherapy supports the effect of antiangiogenic therapy. Moreover the effect of tumor eradication may be achieved more easily and faster, although the theoretical results based on the theory of stability still have an asymptotic form. For constant doses of antiangiogenic and chemotoxic agents (denoted by U and V respectively), the equilibrium point is given by :

$$\begin{aligned} N^* &= ((\gamma - \mu - \eta U - \xi V) / \lambda)^{3/2} \\ K^* &= N^* e^{\xi V / \beta} \end{aligned} \qquad (12)$$

In this case the equilibrium point is not the same for both populations, but it is related very closely, and it can be easily seen that the conditions for both its local and global asymptotic stability are similar to those given above. The main difference is that now both control actions "collaborate" in conditions for convergence of solutions of the model equations to 0. More precisely, the condition (9) should be substituted by:

$$U + \xi V / \eta = (\gamma - \mu) / \eta \Rightarrow K^*, N^* = 0 \qquad (13)$$

The use of the previously considered transformation of variables leads to the following semilinear model of the combined anticancer therapy:

$$\begin{aligned} & x'(t) = y(t) - x(t) - \varepsilon v(t), \\ & y'(t) = \vartheta(1 - e^{(2/3)x(t)}) + \sigma u(t) + \varsigma v(t), \qquad (14) \\ & \sigma = -\eta / \beta, \varepsilon = \psi / \beta, \varsigma = -\xi / \beta \end{aligned}$$

which will be used in further analysis. The main problem with these results is, however, their asymptotic character. In practice only a finite therapy horizon could be considered, which leads to the problem of the system's controllability.

IV. Semilinear systems and their controllability

In this section, we study the general form of the semilinear stationary finite-dimensional control system described by the following ordinary differential state equation:

$$\underline{x}'(t) = A\underline{x}(t) + F(\underline{x}(t), \underline{u}(t)) + B\underline{u}(t) \qquad (15)$$

with zero initial conditions: $\underline{x}(0) = 0$, where the state $\underline{x}(t)\in R^n$ and the control $\underline{u}(t)\in R^m$, A is $n\times n$ dimensional constant matrix, B is $n\times m$ dimensional constant matrix. Moreover, let us assume that the nonlinear mapping $F: X\times U\rightarrow X$ is continuously differentiable near the origin and such that $F(0,0)=0$, and X and U denote state and control spaces, respectively.

In practice, admissible controls are always required to satisfy certain additional constraints. Generally, for arbitrary control constraints it is very difficult to give easily computable criteria for constrained controllability. However, for some special cases of the constraints it is possible to formulate and prove simple algebraic constrained controllability conditions. Therefore, we assume that the set of values of controls $U_c\subset U$ is a given closed and convex cone with nonempty interior and vertex at zero. Then the set of admissible controls for the dynamical control system (15) has the following form:

$$U_{ad}=L_\infty([0,T],U_c). \tag{16}$$

For the semilinear dynamical system (15), it is possible to define many different concepts of controllability. In the sequel we shall focus our attention on the so-called constrained controllability in the time interval $[0,T]$. In order to do this, first of all let us introduce the notion of the attainable set at time $T>0$ from zero initial conditions, denoted shortly by $K_T(U_c)$ and defined as follows:

$$K_T(U_c) = \{\underline{x}\in X : \underline{x} = \underline{x}(T,\underline{u}),\ \underline{u}(t)\in U_c\ \} \tag{17}$$

where $\underline{x}(t,u)$, $t > 0$ is the unique solution of the differential state equation (15) with zero initial conditions and a given admissible control $\underline{u}\in U_{ad}=L_\infty([0,T],U_c)$. Under the assumptions stated for the nonlinear term F, such a solution always exists. Now, using the concept of the attainable set, we recall the well-known definitions of constrained controllability in $[0,T]$ for a semilinear dynamical system.

Definition 1: The dynamical system (15) is said to be U_c-locally controllable in $[0,T]$ if the attainable set $K_T(U_c)$ contains a neighborhood of zero in the space X.

Definition 2: The dynamical system (15) is said to be U_c-globally controllable in $[0,T]$ if $K_T(U_c) = X$.

Now, we shall introduce certain notations and present some important facts from the general theory of nonlinear operators. Let U and X be given spaces and $g(\underline{u}):U\rightarrow X$ be a mapping continuously differentiable near the origin 0 of U. Let us suppose for convenience that $g(0)=0$. It is well known from the implicit-function theorem that if the derivative $Dg(0):U\rightarrow X$ maps the space U onto the whole space X, then the nonlinear map g transforms a neighborhood of zero in the space U onto some neighborhood of zero in the space X. In the more general case when the domain of the nonlinear operator g is Ω, U_c denotes a closed and convex cone in U with vertex at 0. In the sequel, we shall use for controllability investigations a property of the nonlinear mapping g, which is a consequence of a generalized open-mapping theorem. This result seems to be widely known, but for the sake of completeness we shall present it here, though without proof and in a slightly less general form sufficient for our purpose.

Lemma 1: Let X, U, U_c, and Ω be as described above. Let $g:\Omega\rightarrow X$ be a nonlinear mapping and suppose that on Ω nonlinear mapping g has derivative Dg, which is continuous at 0. Moreover, suppose that $g(0) = 0$ and assume that linear map $Dg(0)$ maps U_c onto the whole space X. Then there exist neighborhoods $N_0 \subset X$ about $0\in X$ and $M_0\subset\Omega$ about $0\in U$ such that the nonlinear equation $\underline{x}=g(\underline{u})$ has, for each $\underline{x}\in N_0$, at least one solution $\underline{u}\in M_0\cap U_c$, where $M_0\cap U_c$ is a so-called conical neighborhood of zero in the space U. Using lemma 1 we study constrained local controllability in $[0,T]$ for a semilinear dynamical system (15) using the associated linear dynamical system.

$$\underline{z}'(t) = C\underline{z}(t) + D\underline{u}(t) \quad \text{for} \quad t\in[0,T] \tag{18}$$

with zero initial condition $z(0)=0$, where

$$C = A + F_x(0,0) \qquad D = B + F_u(0,0) \tag{19}$$

are n×n-dimensional and n×m-dimensional constant matrices, respectively. The main result is the following sufficient condition for constrained local controllability of the semilinear dynamical system (15) which will be used to study controllability of the model of combined anticancer therapy:

Theorem 1 [2]. Suppose that (i) $F(0,0) = 0$, (ii) $U_c\subset U$ is a closed and convex cone with vertex at zero, (iii) the associated linear control system (17) is U_c-globally controllable in $[0,T]$.
Then the semilinear stationary dynamical control system (17) is U_c-locally controllable in $[0,T]$.

In practical applications of Theorem 1, the most difficult problem is to verify the assumption (iii) about constrained global controllability of the linear time-invariant dynamical system. In order to overcome this difficulty, we may use the following Theorem.

Theorem 2 [2]: Suppose the set U_c is a cone with vertex at zero and nonempty interior in the space R^m. Then the associated linear dynamical control system (17) is U_c-globally controllable in $[0,T]$ if and only if:

(1) it is controllable without any constraints, i.e.

$$\text{rank}[D,CD,C^2D,...,C^{n-1}D] = n \tag{20}$$

(2) there is no real eigenvector $w\in R^n$ of the matrix C^{tr} satisfying inequalities

$$w^{tr}D\underline{u} \le 0, \text{ for all } \underline{u}\in U_c. \tag{21}$$

These theorems could be proved using the generalized open mapping theorem.

V. Controllability of the Model of Therapy

Now, let us consider the constrained local controllability of the model of combined anticancer therapy described by the semilinear differential state equations (14) defined in a given time interval [0,*T*].

In this case the state vector $\underline{x} = [x, y]^T$, the control vector $\underline{u} = [u, v]^T$, and $\underline{z}$ is the state of the associated linear system. Taking into account the general form of the semi-linear dynamic system we have:

$$A = \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix}$$

$$F(x,y,u,v) = \begin{bmatrix} 0 \\ -\vartheta(e^{(2/3)x} - 1) \end{bmatrix}$$

$$B = \begin{bmatrix} 0 & -\varepsilon \\ \sigma & \zeta \end{bmatrix} \qquad (22)$$

Hence, we have:

$$F(0{,}0{,}0{,}0) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$F_x(0{,}0{,}0{,}0) = \begin{bmatrix} 0 & 0 \\ -\vartheta\frac{2}{3} & 0 \end{bmatrix}$$

$$C = A + F_x(0{,}0{,}0{,}0) = \begin{bmatrix} -1 & 1 \\ -\vartheta\frac{2}{3} & 0 \end{bmatrix} \qquad (23)$$

In order to consider the controllability of dynamical system (14) we use the Theorems presented in the previous section. The admissible controls are assumed to be positive, hence the set of admissible controls is a positive cone U_c in the space R^2.

The characteristic polynomial for matrix C^{tr} has the form:

$$P(s) = \det(sI - C^{tr}) = \det\begin{bmatrix} s+1 & \frac{2}{3}\vartheta \\ -1 & s \end{bmatrix} = \qquad (24)$$
$$= s^2 + s + \tfrac{2}{3}\vartheta$$

Therefore the discriminate of the characteristic polynomial is : $\Delta = 1 - \frac{8}{3}\vartheta$

and the characteristic equation *P*(*s*) = 0 has two roots.

It is necessary to consider the following three cases:

I. $\Delta < 0$, for $\vartheta > \frac{3}{8}$

In this case, we have two complex eigenvalues

$$s_1 = 0.5(-1 - j\sqrt{\Delta}) = 0.5(-1 - j\sqrt{1 - \tfrac{8}{3}\vartheta})$$

and when the eigenvalues are complex, then the system is constrained controllable.

II. $\Delta = 0$, for $\vartheta = \frac{3}{8}$

In this case, we have one real eigenvalue

$s_{12} = -0.5$ with multiplicity 2.

Therefore, to verify controllability it is necessary to use Theorem 2. In order to do that we first find the eigenvector w of the matrix C^{tr}. From the spectral equation

$$C^{tr}w = -0.5w \qquad (25)$$

the real eigenvector has the following form:

$$w = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$$

thus

$$w^{tr}B\underline{u} = 2\sigma u + (\varepsilon + 2\xi)v > 0 \qquad (26)$$

for all positive controls. Therefore, there is no real eigenvector satisfying (21). Hence, taking into account Theorem 2 the system is controllable with positive admissible controls.

III. $\Delta > 0$, for $\vartheta < \frac{3}{8}$

In this case, we have two different real eigenvalues. Hence, to verify controllability we use Theorem 2. The real eigenvalues have the following form:

$$s_1 = 0.5(-1 - \sqrt{1 - \tfrac{8}{3}\vartheta}) < 0$$
$$s_2 = 0.5(-1 + \sqrt{1 - \tfrac{8}{3}\vartheta}) < 0$$

Therefore, the corresponding real eigenvectors are

$$w_1 = \begin{bmatrix} -1 \\ -s_1^{-1} \end{bmatrix} \quad \text{and} \quad w_2 = \begin{bmatrix} -1 \\ -s_2^{-1} \end{bmatrix}$$

Thus,

$$w_1^{tr}B\underline{u} = -s_1^{-1}\sigma u + (\varepsilon - s_1^{-1}\xi)v > 0$$

$$w_2^{tr} B\underline{u} = -s_2^{-1}\sigma u + (\varepsilon - s_2^{-1}\xi)v > 0$$

for all positive controls.

Therefore, there is no real eigenvector satisfying inequality (21). Hence, taking into account Theorem 2 the system is controllable with positive admissible controls. Summarizing, the semilinear dynamical system (14) is constrained controllable in a given time interval [0,*T*] with positive controls. From the biological point of view, this means that if the size of the tumour and its vascular network is not too large then there exists a combination of antiangiogenic therapy and chemotherapy which enables eradication of the tumour. The important finding is that this property does not depend on the parameters of the model, whose estimation may be difficult. In the existing literature, e.g., [3], [8] one can find some estimates for the parameters, but their accuracy is of course very low. This may be not true if only one modality (e.g., antiangiogenic therapy) is used. As proved in [4], local constrained controllability of the model of antiangiogenic therapy is guaranteed only when its parameters satisfy additional conditions related to oscillatory behavior in the untreated case.

VI. Conclusion

In this study, we have shown how, by using quite simple models, we can analyze and design therapy protocols of combined antiangiogenic and chemotherapy of tumors. This type of cancer treatment is still in experimental and clinical trials. The results are promising, but knowledge of the processes behind the evolution of cancer vascular networks, the equilibrium between stimulatory and inhibitory factors, different forms of antiangiogenic therapy, its side effects. and the results of combined use of different treatment modalities is still far from being complete. The important finding presented in this paper is that sufficient conditions of local constraint controllability for the simple model of combined therapy are satisfied independent of its parameters, which is not true for the model of antiangiogenic therapy alone [4]. A more realistic model should take into account drug resistance of the cancer cell population caused by cytotoxic agents (see, e.g., [7]). Of course the situation *in vivo* is more complicated than the two-compartment models considered in this paper but, in our opinion, it may be treated similarly and may lead to similar qualitative results. The results will not change if linear pharmacokinetics of antiangiogenic and/or cytotoxic drugs is included in the model. Qualitatively, the controllability problem will change if delays in the dynamics of tumor growth and vascular network development are taken into account, and such a model was proposed in [21] and analyzed without control terms in [22]. We hope that its controllability could be also examined using theorems presented in [23] based on the similar mathematical engine.

Acknowledgment

This work was partially supported by Polish National Centre of Science grants no. DEC-2011/03/B/ST6/04384 in year 2012 (AS) and no. NN514513340 in year 2012 (JK).

References

[1] J. Klamka, Controllability of Dynamical Systems, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1991.

[2] J. Klamka, “Constrained controllability of nonlinear systems”. J. Math. Anal. Appl., vol. 201, 1996, pp. 365-374.

[3] P. Hahnfeldt, D. Panigraphy., J. Folkman and L. Hlatky “Tumor development under angiogenic signaling: A dynamic theory of tumor growth, treatment response and postvascular dormacy”, Cancer Res. vol. 59, 1999, pp. 4770-4778.

[4] A. Swierniak and J. Klamka, “Control properties of models of antiangiogenic therapy”, in: Advances in Automatics and Robotics (K. Malinowski and R. Dindorf R. Eds.), Monograph of Committee of Automatics and Robotics PAS, vol.16, 2011 Kielce, pt..2, pp.300-312.

[5] A. Swierniak, “Direct and indirect control of cancer populations”, Bull. PAS Tech. Sci. vol. 56, 2008, pp. 367-368

[6] U. Ledzewicz, H. Schättler and A. d’Onofrio, “Optimal Control for Combination Therapy in Cancer”, Proc. 47th IEEE Conference on Decision and Control, 1537-1542, 2008

[7] A. Swierniak, Control problems related to three compartmental model of combined anticancer therapy, Proc. 20 IEEE Mediterenian Conf. Autom. Contr. MED 12, Barcelona, 2012, pp. 1428-1433

[8] A. d’Onofrio and A. Gandolfi, “Tumour eradication by antiangiogenic therapy analysis and extensions of the model by Hahnfeldt et al (1999)”, Math. Biosci vol. 191, 2004, pp. 159-184

[9] A. Swierniak, “Comparison of six models of antiangiogenic therapy”, Applicationes Mathematicae, vol. 36, 2009, pp. 333-348

[10] U. Ledzewicz and H. Schattler, “Anti-angiogenic therapy in cancer treatment as an optimal control problem”, SIAM J. Contr. Optim, 46, 2007, pp.1052-1079.

[11] U. Ledzewicz and H. Schattler, “Analysis of mathematical model for tumor anti-angiogenesis”, Optim. Contr. Appl. Meth., vol. 29, 2008, pp. 41-57.

[12] J. Folkman, “Tumor angiogenesis: therapeutic implications”, N. Engl. J. Med., vol. 295, 1971, 1182-1186.

[13] R.K. Jain, “Molecular regulation of vessel maturation”, Nature Medicine, vol. 9, 2003, 685-693.

[14] R.S. Kerbel S., A cancer therapy resistant to resistance, Nature, vol. 390 , 1997, 335-340.

[15] J.M.L. Ebos and R.S. Kerbel, “Antiangiogenic therapy: impact on invasion, disease progression, and metastasis”, Nature Rev. Clin. Oncol., vol. 8, 2011, pp.:210-221.

[16] J. Ma and D.J. Waxman, “Combination of anti-angiogenesis with chemotherapy for more effective cancer treatment”, Mol. Canc. Ther., vol. 7, 2010, 3670-3684.

[17] T. Li-Song., J. Ke-Tao, H. Kui-Feng, W. Hao-Hao, C. Jiang and Y. De-Cao, “Advances in Combination of Antiangiogenic Agents Targeting VEGF-binding and Conventional Chemotherapy and Radiation for Cancer Treatment”, J. Chin. Med. Ass., vol. 73, 2010, pp. :281–288,

[18] R.K. Jain, “Normalization of tumor vasculature: an emerging concept in antiangiogenic therapy”, Science, vol. 307, 2003, pp. 58-62. Article in a conference proceedings:

[19] A. d’Onofrio and A. Gandolfi., “Chemotherapy of vascularised tumours: role of vessel density and the effect of vascular ‘pruning”, J. Theor. Biol., vol. 264, 2010, 253-265.

[20] US National Institutes of Health, *Clinical Trials* [online], www.clinicaltrials.gov. , 2012

[21] A. d’Onofrio and A. Gandolfi, “A family of models of angiogenesis and anti-angiogenesis anti-cancer therapy”, Math. Med. Biol., vol. 26, 2009, pp. 63-95.

[22] M. J. Piotrowska, U. Foryś, “Analysis of the Hopf bifurcation for the family of angiogenesis models”, J. Math. Anal. Appl., Vol. 382, 2011, pp. 180–203.

[23] J. Klamka, “Constrained controllability of semilinear systems with multiple delays in control”. Bull. PAS, Techn. Sci., vol. 52 , 2004, pp. 25-30.

Design and Development of an Information System to Manage Clinical Data about Usher Syndrome Based on Conceptual Modeling

Verónica Burriel, M. Ángeles Pastor, Matilde Celma, J. Carlos Casamayor, Laura Mota
Centro de Investigación en Métodos de Producción de Software PROS
Universitat Politècnica de València
Valencia, Spain
<vburriel, mapastor, mcelma, jcarlos, lmota>@pros.upv.es

***Abstract* — The inefficient management of clinical data in many research environments is a problem which slows down the service provided to patients. The benefits of an Information System created following the conceptual modeling rules have been proved in multiple environments with data management difficulties. The main hurdle to overcome is the large gap between the language and concepts employed by informaticians and the ones used by biologists. The work described in this paper shows how these technologies can also be applied to the clinical domain, after a long period of mutual approaching in order to understand each other. The research clinical data of an expert research group on Usher syndrome have been studied, analyzed and redesigned using conceptual modeling, helping this group to offer a better service.**

Keywords-information system; usher syndrome; database; conceptual modeling.

I. INTRODUCTION

Usher syndrome is a condition characterized by hearing loss or deafness and progressive vision loss. The loss of vision is caused by an eye disease called *retinitis pigmentosa* (RP) which affects the layer of light-sensitive tissue, at the back of the eye (the retina). Vision loss occurs as the light-sensitive cells of the retina gradually deteriorate. Usher syndrome is thought to be responsible for 3 percent to 6 percent of all childhood deafness and about 50 percent of deaf-blindness in adults. This disease is estimated to occur in at least 4 per 100,000 people.

Mutations in the CDH23, CLRN1, GPR98, MYO7A, PCDH15, USH1C, USH1G, and USH2A genes cause Usher syndrome. These genes provide instructions for proteins that play important roles in normal hearing, balance, and vision. They have influence in the development and maintenance of hair cells, which are sensitive cells in the inner ear that help to transmit sound and motion signals to the brain. In the retina, these genes are also involved in determining the structure and function of light-sensitive cells called rods and cones. In some cases, the exact role of these genes in hearing and vision is unknown. Most of the mutations responsible for Usher syndrome lead to a loss of hair cells in the inner ear and a gradual loss of rods and cones in the retina. Degeneration of these sensitive cells causes hearing loss, balance problems, and vision loss characteristic of this condition.

For research groups related to this disease, having data properly stored and classified is very important in order to access them easily. Furthermore, it is possible to detect relationships between the data if it is correctly structured and linked. This would help to improve the understanding of the disease.

However, despite the advances in information technologies, nowadays there are still research groups in health which store their data in spreadsheets or simply sheets of paper. The most advanced groups use simple databases to store information about their patients, but most of these databases are focused on the solution-space. Despite that it is not the most appropriate solution, these tools are a first approximation to a solution, but they lack a previous conceptual scheme. This problem entails poor management of the involved data and the consequent loss of quality of information and simplicity of operation. The conceptual modeling ensures the adequacy of the stored data, improving their usefulness and maintenance, aspects currently seldom taken into account in the clinical domain. The main difficulty for solving this deficiency, following our experience, is the enormous distance between the concepts and languages used by the scientists from the life areas in front of these used by the technics in informatics. This problem has been highlighted every time when our group has shared a work with biologists or doctors.

The Genoma research group of PROS (*Centro de Investigación en Métodos de Producción de Software*) is a research group with expertise in Information Systems and Bioinformatics, which is working for a while in the design of a model to represent all the knowledge acquired so far about the genomic domain using conceptual modeling technologies. This is the main step to create a Genomic Information System with a database capable of storing comprehensive genomic information [1-3]. This knowledge about genomics and Information Systems is the perfect environment to solve the problem described above.

In following sections, the information system developed will be explained in detail. In section II, similar solutions to different departments will be exposed. In section III, the current system of data storage used by the clinicians before

ISBN: 978-1-61208-260-8

installing this information system is detailed. The advantages of having an information system modeled by a conceptual scheme are reasoned in section IV. The conceptual scheme created to develop this information system is deeply explained in section V. In section VI and VII, the database implemented and the loading and managing processes are defined. Finally, the conclusions and future work are described in detail in section VIII of this paper.

II. State of the Art

Despite being well-known that Health Information Systems have a great potential to improve quality of clinical services and reduce costs, only about 17 percent of doctors and 8 percent to 10 percent of U.S.A. hospitals use electronic medical records. Most medical institutions do not want to risk installing these systems due to lack of formal evaluations and evidence regarding its successful implementation [4] .

However, there are some groups that have implemented Information Systems for a clinical environment with satisfactory results. In University Medical Centre Ljubljana (Slovenia) a clinical information system is used to support medication process (prescribing, ordering, dispensing, administration and monitoring) and offer participating medical teams real time warnings and key information regarding medications and patient status, thus reducing medication errors [5].

Other example of the success of Health Information Systems is installed in the University Cancer Centre in Frankfurt (Germany). The hospital information system gives physicians the possibility to access to all patient information in a hospital. Furthermore, a special query and reporting tool has been integrated in the health information system to recognize patients with a specific disease and with basic inclusion and exclusion criteria for a specific clinical trial [6].

Instead, this progress of health information systems has not been so relevant in Information Systems with genomic information. Predefined online databases have been used to store information about the patients in these situations, instead of personalized health information systems. One example of this situation is the locus-specific database for mutations in GDAP1 gene created in the INSERM of Angers (France). It helps in the analysis of genotype-phenotype correlations in Charcot-Marie-Tooth diseases type 4A and 2K [7]. A complete health information system can offer much more services than a simple database and could be more useful in this purpose.

Due to the different ways of working in every health and research centre, a specific information system adapted to every center or centers with similar needs will be the easier way to introduce information systems in these domains without changing too much their current working methods.

III. Current System of Data Storage

As an example of this problem, the storage of data from a center for research in neurosensory diseases has been analyzed [8-10]. This center studies patients with Usher Syndrome coming from every hospital in Spain, gathering together a lot of data from this disease. To obtain all these data, they use several methods: questionnaires, which are given to the patients and relatives to be fulfilled to obtain certain data about the severity of the disease, and genetic analysis in order to obtain some objective data from the blood sample of a patient.

When a patient arrives to the hospital with signs and symptoms of suffering Usher Syndrome, a first diagnosis is done by the doctor, who creates a diagnosis report with these signs and symptoms. During the same visit, a questionnaire is given to him or, in case he is not able to fill it, to his parents. Next, a sample of blood is extracted from the patient in order to be analyzed genetically. Sometimes, some genetic analyses are also performed to the relatives, looking for the disease genetic heritage. These samples are analyzed using a sequencer which directly produces the results in electronic support. All the information obtained from the diagnosis report, questionnaires and genetic analyses are used by the clinicians to make the definitive diagnosis of the disease suffered by the patient.

The information extracted from these processes is stored using different methods. The questionnaires and diagnosis reports are filed in physical folders sorted by families and following an alphabetical order. On the other hand, the genomic data is saved in Excel files and one Access table without any kind of structure behind it.

Once the questionnaires have been filled by the patients and their relatives on a set of sheets of paper, introducing all this information into an information system is a hard work that requires a lot of time, and the clinicians are too busy to do it. If these questionnaires had been done by computer directly, the information would have been introduced instantly into a database. The same problem appears with the diagnosis reports. This solution would help a lot to the clinicians and would save a lot of time and space.

Sorting the information by families is a peculiar classification which is very common when information about a genetic disease is stored. It is very important and relevant to know the genetic profile of the members of the family of the patient, because they have a high probability of suffering the same disease or being carriers of it.

Another problem found among the data is the lack of formalism. This means that information is repeated in multiple places creating redundancy. Often some data are missed or represented using cryptic codes without clear semantics.

This case study is not an isolated case. There are many research groups or clinical departments that do not store their data in an information system based on conceptual models.

IV. Advantages of Conceptual Modeling

Nowadays, the most advanced approximations of Information Systems development, which are oriented towards producing quality systems, propose the use of conceptual model-based methodologies[11]. Conceptual modeling is widely used in the Information Systems field because it helps developers in the understanding and description of the problem domain before implementation. In this way, conceptual modeling improves the developed system helping to manage its evolution in the future [11, 12]. For a long time, conceptual modeling techniques have been used successfully to build IS in many different domains [13].

Quality clinical practice must necessarily be based on a data set of maximum reliability and accurate interpretation. To make this possible, it is necessary to develop conceptual schemas that characterize the information to be stored, used and modified to ensure its continued updating.

To create this conceptual schema, a strong knowledge of the domain is strictly necessary. To deal with this problem we were helped by the clinicians from the Center for Research in Neurosensory Disease, who explained us the domain in great detail, enabling us to understand every piece of stored information and validating our interpretation of the domain. After the understanding of the domain, all the information has been correctly represented in a conceptual schema by our group..

From the conceptual schema, the corresponding database schema will be generated. This schema, then, will be used to create a database to correctly store the data ensuring their quality. This will provide a system to store and access data in a much easier and faster way, allowing much more complex queries on these data than before. Having all the information linked into this information system, the possibility of discovering the relation between data and improving the treatments for the patient is more feasible.

The new Information System will improve the quality of information stored by the clinicians and consequently, the service provided to the patient, as well as the time saving for doctors in the management and recovery of the information.

V. Conceptual Scheme Description

The conceptual schema shown in Figure 1 represents all the information handled by the Center for Research in Neurosensory Group. To understand it easily, this section explains in detail its contents. The information about the persons studied in its laboratory is represented in this schema by *Person* class. There is a relevant attribute of *Person* to take into account, *family_id*, which connects the relatives. This attribute is very useful to analyze samples of new members of the same family, who can have the same variations. Other relevant attribute is *origin_code*, where the code from the department that sends the sample is stored. Basic data for a person registration as *name, surname* and *date* of registration in included in this class too. A person can be related with two persons through the associations *father* and *mother* which represent the parents. Being one of the few hospitals in Spain experts in this disease, samples of every Spanish hospital are received in the laboratory. The provenance of these samples is conveniently stored in the *Provenance* class.

In addition, a person can be a patient or not, depending if this person suffers this disease or not. This is represented by the *Patient* class. It is also important to take into account if studies about *segregation* and *consanguinity* have been done to the patient. Sometimes, the diagnosis of the patient is not totally clear and this lack of sureness is represented in *diagnosis_reliability* attribute.

The common use of markers to identify which genes

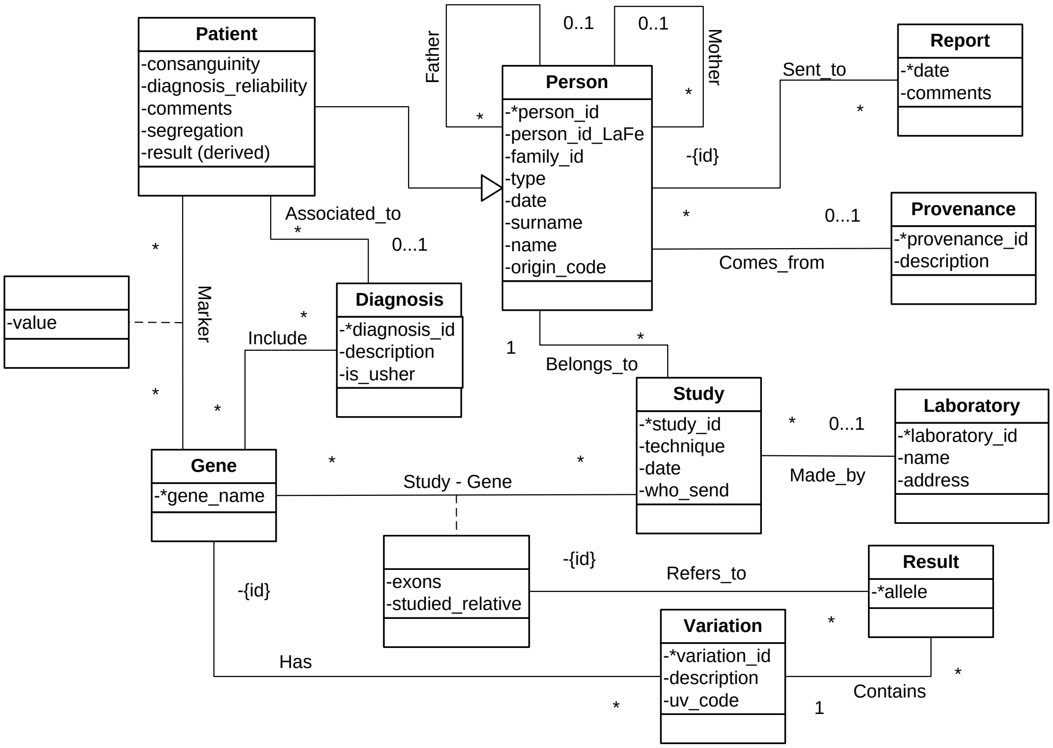


Figure 1. Information System Conceptual Schema

(represented by *Gene* class and its *gene_name* attribute) are inherited from the patient's mother and father has promoted the inclusion of a *Marker* class in this conceptual scheme. These markers help to exclude the non-mutated genes from the study, knowing previously the non-mutated genes of the parents. The value obtained from the marker is stored in *value* attribute, which is referred only to one patient and one gene.

As in any clinical service, a *Diagnosis* is associated to a patient and one or more genes. In this case, the *diagnosis_id* content represents one of the three cataloged types of the disease (Usher I, Usher II or Usher III), a combination of them if it is not clear, to the Usher Syndrome if it is not possible be more specific, or any other disease if it is necessary.

The process to analyze a sample of a person to define if he suffers the disease or if he is a carrier of it, is defined by the *Study* class. Knowing the person *who_send* the sample to an external laboratory to analyze it and the *date* of the sending is very interesting to value if the study is precise enough. To ensure the reliability of the study, storing the *technique* used to analyze the genes is very important too.

The studies are done gene by gene (represented by the *Study-Gene* class), so if the variation is not found in that gene, a new study has to be done. It is also important to store the studied *exons* (transcriptable DNA within a gene), defining which exons have been analyzed when the gene has not been completely analyzed, and to notify if a previous study of that gene in a relative has been done, which is expressed by the *studied_relative* attribute.

As mentioned before, every study can be done in a different laboratory so, it is important to store basic information about the laboratory, such as *name* and *address*, where the study has been done using the *Laboratory* class.

Obviously, after a study has been done, some results, which are represented by *Result* class, have been found. These results are related only to one *allele* and they can contain some variations which may have been found during the study.

The variations found into the DNA are represented by the *Variation* class and are related to a gene. The HGVS code of the variation is included in the *variation_id* attribute, complementing this information with a *description* attribute. Additionally, the standard *uv_code* is used to express the pathogenicity of the variation.

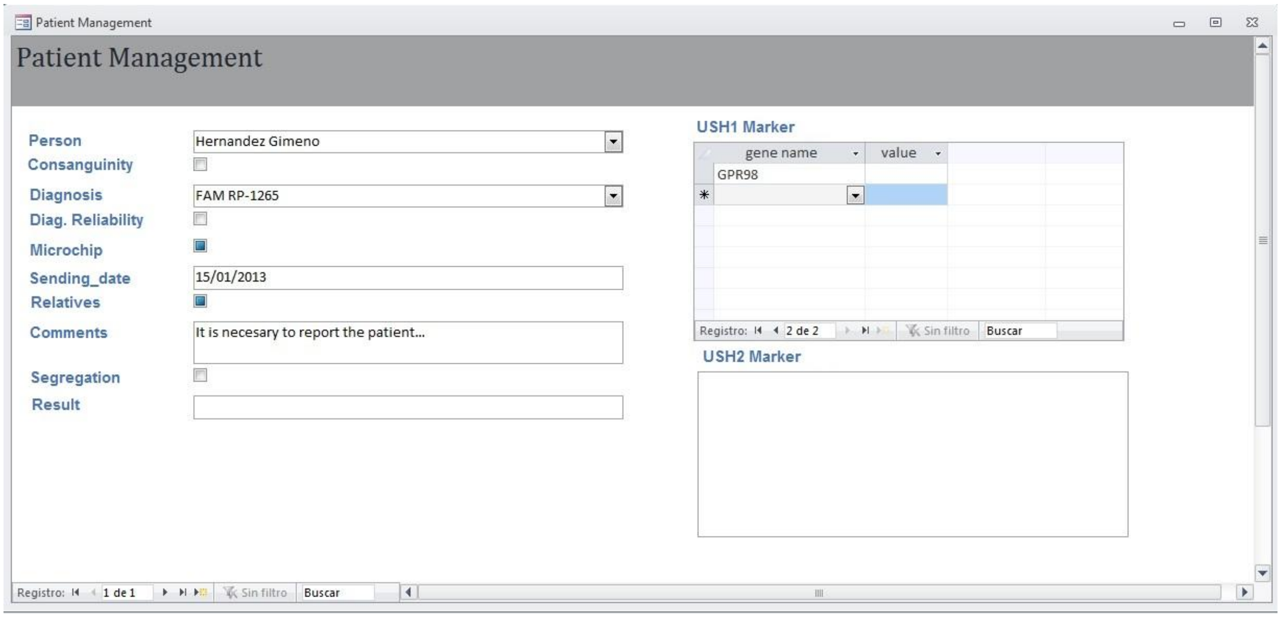
Finally, after making a complete genetic study of the person, a report has to be presented to inform the person about the results. The *date* and the *comments* related to this report are represented by the *Report* class.

VI. DATABASE AND LOADING PROCESS

From the conceptual scheme explained below, a database, which is able to store data from the sources mentioned below, has been created. The conceptual scheme, on which this database is based on, ensures the correct structure of data and the efficiency of this database. It has been produced following the Conceptual Modeling rules, ensuring the quality of information and the efficiency of storage of genomic information.

Furthermore, this database has been designed and implemented using Microsoft Access technologies. This tool used to develop the database, has been chosen taking into account that it is a well-known environment used by the clinicians. Another aspect that was taken into account is that the quantity of data handled in this research centre is not too large, so it is not necessary to use a more sophisticated System Management Database. This choice will make the information system easy to be used by clinicians.

Once the database has been developed, loading the information into the database is the next step. This process



Patient Management

Person	Hernandez Gimeno
Consanguinity	☐
Diagnosis	FAM RP-1265
Diag. Reliability	☐
Microchip	☑
Sending_date	15/01/2013
Relatives	☑
Comments	It is necesary to report the patient...
Segregation	☐
Result	

USH1 Marker

gene name	value
GPR98	

USH2 Marker

Figure 2. Patient Management Form

can't be done directly due to the peculiar form of the information in the original sources. As explained in section II, the data is saved in sheets of paper, Excel files and some Access tables without any kind of structure behind it. Storing different data into the same cell of the spreadsheet or expressing the same information in different formats, are some of the encountered problems. Facing this situation together with the clinicians trying to clarify the information found in the sources is the essential step to start the loading process.

Some loading modules have been developed in order to introduce the information. These modules have been implemented analyzing the information contained in the sources, extracting the information, transforming it to the new format and loading it into the database. The "transforming" step is essential to introduce the information with the appropriate format and structure into the new Information System, avoiding the mentioned above format mistakes.

VII. Managing the data

To achieve an efficient use and avoid inaccurate management of the created Information System, a user interface has been implemented. Specific Microsoft Access forms have been developed to introduce new data and manage the database. There are five different forms that can be accessed by a *Main Menu* form.

These five forms help users correctly enter data into the information system and make queries to extract the information previously introduced. Information about *Diagnosis, Gene, Study, Patient* and *Person* is correctly stored thank to the corresponding forms.

VIII. Conclusions and Future Work

The implementation of developed information system has allowed the clinicians to overcome previous drawbacks as the waste of time looking for physical papers, the manual work, the use of different sources of information, the personal use of identification symbols to describe the data, etc. as we mentioned in section III. With this information systems-based policy, the quality of the stored data has significantly improved. The ambiguity of their data has disappeared and they can access to all the information faster and easier. It has also increased the usefulness of their data because the data can be more precisely searched and multiple parameters can be used, and even the information that previously was stored in folders can be searched through. This information system has helped this clinical research group to properly manage their data, but the problem is not an isolated case and an information system based on conceptual models can resolve problems like this in many other research centers or medical departments. Conceptual modeling technique has been used in many areas to ensure a correct management of the data, but in the clinical area, most information systems lack a conceptual model base. This work has shown that conceptual modeling is also effective in the clinical field. This work has reaffirmed us in the idea that the main difficulty for achieving this improvement in quality is the long time consuming meetings that are needed in order to experts from the informatics area understand the experts from the life sciences area.

As future work, connecting all the information about this disease would be very useful to the research about Usher Syndrome. The information system explained in this paper can help to achieve this future goal. If this information system was installed into every centre of research in Usher Syndrome, connecting these information systems to share the information between them would be possible. This big information system about Usher Syndrome would join all the information about the disease, making easier and faster the research about it.

Acknowledgment

This work has been developed under support of the director of PROS, Dr. Oscar Pastor, and jointly with Dr. José María Millán, Dr. Elena Aller and Dr. Teresa Jaijo, who belong to the *Centro de Investigación en Enfermedades Neurosensoriales* of the *Instituto de Investigación Sanitaria La Fe* in Valencia (Spain).

In addition, it is also important to highlight that this work has been done under the framework of the *Cátedra Tecnologías para la Salud* of the *Universitat Politècnica de València* financed by INDRA Systems.

References

[1] O. Pastor, "Conceptual Modeling Meets the Human Genome," Conceptual Modeling-ER 2008, pp. 1-11, 2008.

[2] O. Pastor, A. Levin, J. Casamayor, M. Celma, L. Eraso, et al., "Enforcing Conceptual Modeling to improve the understanding of human genome," 2010, pp. 85-92.

[3] M. Pastor, V. Burriel, and O. Pastor, "Conceptual Modeling of Human Genome Mutations: A Dichotomy Between What we Have and What we Should Have," in BIOSTEC Bioinformatics, Valencia, 2010, pp. 160-166.

[4] J. C. Goodman, "Health Information Technology: Benefits and Problems."

[5] A. Cufar, A. Droljc, and A. Orel, "Electronic medication ordering with integrated drug database and clinical decision support system" Studies in health technology and informatics, vol. 180, p. 693-697, 2012.

[6] M. Koca, G. Husmann, J. Jesgarz, M. Overath, C. Brandts, et al., "A special query tool in the hospital information system to recognize patients and to increase patient numbers for clinical trials" Studies in health technology and informatics, vol. 180, p. 1180-1181, 2012.

[7] J. Cassereau, A. Chevrollier, D. Bonneau, C. Verny, V. Procaccio, et al., "A locus-specific database for mutations in GDAP1 allows analysis of genotype-phenotype correlations in Charcot-Marie-Tooth diseases type 4A and 2K" Orphanet Journal of Rare Diseases, vol. 6, p. 87-94, 2011.

[8] F. P. M. Cremers, W. J. Kimberling, M. Külm, A. P. de Brouwer, E. van Wijk, et al., "Development of a genotyping microarray for Usher syndrome" Journal of medical genetics, vol. 44, pp. 153-160, 2007.

[9] J. M. Millán, E. Aller, T. Jaijo, F. Blanco-Kelly, A. Gimenez-Pardo, et al., "An update on the genetics of usher syndrome" Journal of ophthalmology, 2010.

[10] C. Nájera, M. Beneyto, J. Blanca, E. Aller, A. Fontcuberta, et al., "Mutations in myosin VIIA (MYO7A) and usherin (USH2A) in Spanish patients with Usher syndrome types I and II, respectively" Human Mutation, vol. 20, pp. 76-77, 2002.

[11] A. Olivé, Conceptual modeling of information systems, Springer, 2007.

[12] O. Pastor and J. C. Molina, Model-driven architecture in practice: a software production environment based on conceptual modeling, Springer, 2007.

[13] E. D. Falkenberg, W. Hesse, P. Lindgreen, B. E. Nilsson, J. L. H. Oei, et al., "A framework of information systems concepts" in IFIP WG, 1998.

Antinociceptive Activity of ST36 Acupoint Stimulation by Low-Level Light Therapy (LLLT) in Mice

Vanessa Erthal, Percy Nohama
Rehabilitation Engineering Laboratory/CPGEI/UTFPR
Curitiba, Brazil
e-mail: acupuntura_vane@yahoo.com.br;
percy.nohama@gmail.com

Maria Fernanda de P. Werner, Cristine H. Baggio
Pharmacology department/UFPR
Curitiba, Brazil
Email: mariafernandaw@gmail.com;
crisbaggio@hotmail.com

Abstract— **Laser acupuncture is a modality of low-level light therapy used as an alternative to needling for the past three decades. The ST36 (Zusanli) acupoint is used to treat inflammatory processes, pain and gastrointestinal disturbs. For this reason, the aim of the present study was to evaluate the antinociceptive effect of Laser acupuncture (830 nm, 3 J/cm2) on ST36 acupoint model of acute nociception. To contribute to this understanding was investigated the involvement of adrenergic and adenosinergic systems in the antinociceptive effect of ST36 acupoint stimulation in mice model of acute nociception induced by glutamate (20 µmol per paw). Our results demonstrate that ST36 laser acupuncture inhibited the glutamate-induced nociceptive behavior in 44%. Moreover, the antinociceptive activity of laser irradiation in the glutamate test was significantly reversed by pre-treatment of yohimbine (0.15 mg/kg, i.p.) and caffeine (3 mg/kg, i.p.) but not by prazosin (0.15 mg/kg, i.p). Collectively, these results demonstrated that ST36 photonic stimulus showed antinociceptive effect in acute model of nociception in mice, and that this effect is mediated by activation of the α2 adrenergic and adenosinergic systems.**

Keywords: Laser acupuncture; antinocicpetion; Zusanli

I. INTRODUCTION

Laser acupuncture is a type of low-level laser therapy (LLLT) and consists of a non-invasive stimulation of the traditional acupuncture points. Its clinical application is becoming widespread being used to treat pain and inflammation, especially with regard to the control of chronic pain states, avoiding the use of analgesics and anti-inflammatory (NSAIDs) [1-2]. This form of phototherapy has been shown to modulate various biological processes and may be used as a supplementary or alternative treatment of several symptoms [3]. The ST36 (Zusanli) acupoint has been used to treat inflammatory processes, pain and gastrointestinal disturbs [4]. Previous pre-clinical data showed that ST36 stimulation with LLLT, during 2, 6 and 10 min, inhibited the nociception induced by formalin in mice [5]. Besides, a clinical study with LLLT set to 830 nm and 30 mW, applied on ST36 and IG4 acupoints in children reduced the migraine [6]. However, there are few evidences about its efficacy and mechanisms of action. For this reason, the aim of the present study was to evaluate the antinociceptive effect of LLLT on ST36 acupoint and its mechanism of action using models of acute nociception in mice. Although the work is in a preliminary phase, several conclusions can be longer designed specially involving the mechanisms of action of LLLT. The paper was divided into four parts: (I) Introduction, in which we present the scientific fundaments involved in the antinociceptive effect of Laser acupuncture and the goal of the experimental study proposed (II) Material and methods performed on this research, as well as the main experimental models involved; (III) Results related to the application of the experimental protocols and discussion about the main results; (IV) Conclusion, where we highlight the most important findings of the developed study.

II. MATERIAL AND METHODS

A. Animals

Experiments were conducted using female Swiss mice (25–35 g), housed at 22 ± 2 °C under a 12/12 h light/dark cycle (lights on at 06:00 h) and with free access to food and water. All experimental protocols were performed after approved by the Committee of Animal Experimentation of the Federal University of Paraná (CEUA/BIO - UFPR, protocol 514) and were carried out in accordance with the ethical guidelines for investigations of experimental pain in conscious animals [7]. The number of animals and intensities of noxious stimuli set were the minimum value necessary to demonstrate the consistent effects of the treatment.

B. LLLT treatment procedures

For the experiments, a low-intensity AsGaAI laser equipment was used. Its main parameters were: wavelenght of 830 nm (in continuous-mode), fluence of 3 J/cm², power of 30 mW, irradiation area reached 6 mm², duration of 6 s on the acupoint. The animals were randomly divided in three groups (n=8): (1) Control group, which was not treated; (2) Laser group, which was treated with unilateral ST36 laser acupuncture; and (3) Off group, in which laser device was turned off but holding the probe in contact with ST36 acupoint. ST36 (Zusanli) acupoint is located between the tibia and the fibula, approximately 5 mm lateral to the anterior tubercle of the tibia [8].

C.. Nociception induced by glutamate

The procedure used was similar to that described previously [9]. A volume of 20µl of glutamate (20 µmol per paw) prepared in isotonic saline solution, with pH adjusted to 7.4 by the addition of NaOH, according to Meotti et al. (2006) [10], was injected intraplantarly (i.pl.) in the ventral surface of the right hindpaw. Animals were observed individually for 15 min following glutamate injection. The amount of time they spent licking the injected paw was recorded with a chronometer and considered as indicative of nociception as Fig. 1. Animals were treated with unilateral ST36 laser acupuncture 30 min before glutamate injection.

D. Involvement of α_1-adrenergic system

To investigate the participation of the α_1-adrenergic system in the antinociceptive effect of ST36 laser acupuncture, mice were pre-treated with prazosin (0.15 mg/kg, i.p, a nonselective α_1-adrenergic receptor antagonist) 30 min before application of unilateral ST36 laser acupuncture, phenylephrine (10 mg/kg, i.p., α1-adrenergic receptor agonist) or vehicle (0.1 ml/10 g, i.p.). Another group of mice was pre-treated with vehicle and after 30 min, received unilateral ST36 laser acupuncture, phenylephrine or vehicle, 30 min before the glutamate injection. The nociceptive response was evaluated as previously reported.

E. Involvement of α_2-adrenergic system

To assess the possible participation of the α_2-adrenergic system, mice were pretreated with yohimbine (0.15 mg/kg, i.p, a nonselective α_2-adrenergic receptor antagonist) 30 min before application of unilateral ST36 laser acupuncture, clonidine (0.1 mg/kg, i.p., α_2-adrenergic receptor agonist) or vehicle (0.1 ml/10 g, i.p.). Another group of mice was pre-treated with vehicle and after 30 min, received unilateral ST36 laser acupuncture, clonidine or vehicle, 30 min before the glutamate injection. The nociceptive response was evaluated as previously reported.

F. Involvement of adenosinergic system

To investigate the involvement of adenosinergic system in the antinociception caused by ST36 laser acupuncture, mice were pre-treated with caffeine (3 mg/kg, i.p., a nonselective adenosine receptor antagonist) 30 min before application of unilateral ST36 laser acupuncture or vehicle (0.1 ml/10 g, i.p.). After 30 min, the nociceptive response to the glutamate intraplantar injection was recorded as previously reported.

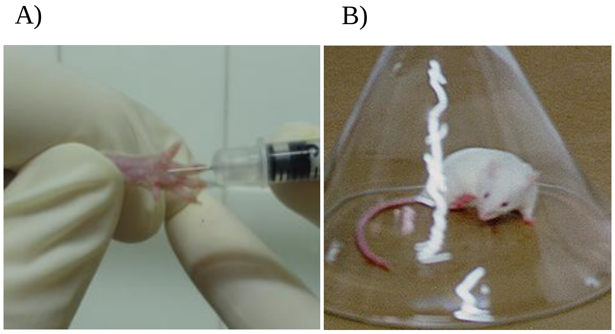


Figure 1. (A) Application intraplantar (ipl.) the glutamate in ventral surface of the right hind paw and (B) observation in the mice glutamate injection.

G. Statistical analyses

Data are presented as means ± standard error of the mean (S.E.M.). Comparisons between experimental and control groups were performed by one-way analysis of variance (ANOVA) followed by Newman Keul's test when appropriate. *P* values less than 0.05 were considered as indicative of significance.

III. RESULTS AND DISCUSSION

The results presented in Fig. 2 demonstrate that ST36 laser acupuncture significantly inhibited nociception induced by glutamate in mice, with inhibition values of 44%. The treatment with laser device turned off was not able to reduce nociception induced by glutamate injection when compared with control group.

In order to identify mechanisms by which it promotes its antinociceptive acupoint, alpha-adrenergic and adenosinergic systems were tested using pre-treatments antagonists of such systems, to evaluate whether some of them would modified the results found in the model of nociception induced by glutamate.

Involved with antinociception, α-adrenergic pathways descendants are recruited from stimuli in brain structures such as Periaqueductal Gray substance (PAG) [11]. The antinociceptive properties of agonists of α2-adrenergic receptors modulation reflect the influence of excitatory neurons in primary afferents projection [10]. Evidence also suggests that α1-adrenergic receptors modulate nociceptive processing in the dorsal horn of the spinal cord [11].

The involvement of adrenergic system on the antinociceptive action of Laser was demonstrated through the pre-treatment of animals with the non-selective adrenergic receptor.

Next, the involvement of the α1 and α2 adrenergic systems on ST36 laser acupuncture was investigated.

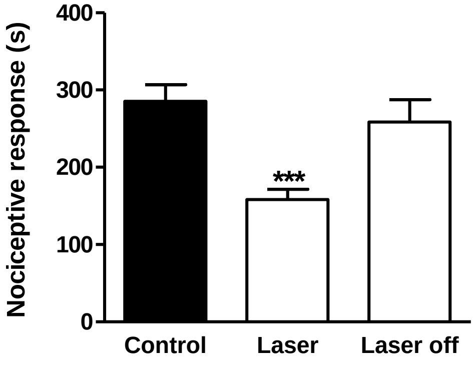


Figure 2. Effect of ST36 laser acupuncture on glutamate induced nociceptive behavior in mice. Each group represents the mean of 8 animals, and the vertical lines indicate the S.E.M. ***p<0.001 when comparing with control group (one-way ANOVA followed by Newman-Keuls Multiple Comparison Test).

Results presented in Fig. 3A show that prazosin pre-treatment completely reverted the antinociceptive effect of both phenylefrina and the effect antinociceptive of ST36 laser irradiation was not reverted that prazosin group in the glutamate test.

Although the results in Fig. 3B show the effect antinociceptive of ST36 laser irradiation was reverted that Yomhibine. Using bee venom injection in acupoint CV12 (Zhongwan), Kwon and colleagues (2001) [12] demonstrated the involvement of this receptor α2-adrenergic antinociception. Park et. al. (2012) [13] observed that α2-adrenoceptor antagonist and β-adrenoceptor antagonist inhibited the analgesic effect of eletroacupuncture (EA) on the inflammatory pain in a rat model of collagen-induced arthritis, but α1-adrenoceptor did not inhibit the analgesic effect of EA, as in our results [13].

Despite being an ancient technique in the east, acupuncture is still a relatively recent scientific study in the west. In fact, there are few studies showing the mechanisms of action and the difference among acupoints in the scientific literature. Another important factor to evaluate is the technique used, since most of the researches employs electroacupuncture, while few of them use photonic stimulation (known as laserpuncture).

Laser light is a good alternative to metal needles for stimulation of acupuncture points, and it has been used successfully for several decades. However, there are only few studies proving the effectiveness of this kind of acupuncture. Most publications focus on red or infrared laser stimulation, and there are several relevant studies [14–17].

The systemic pre-treatment of mice with caffeine in Fig. 4 did significantly reverse the antinociception caused by Laser and against glutamate induced nociception. The findings of Goldman et al. (2010) [18] indicate that adenosine is central to the mechanical actions of acupuncture, as the results demonstrated in our experiments in laser acupuncture.

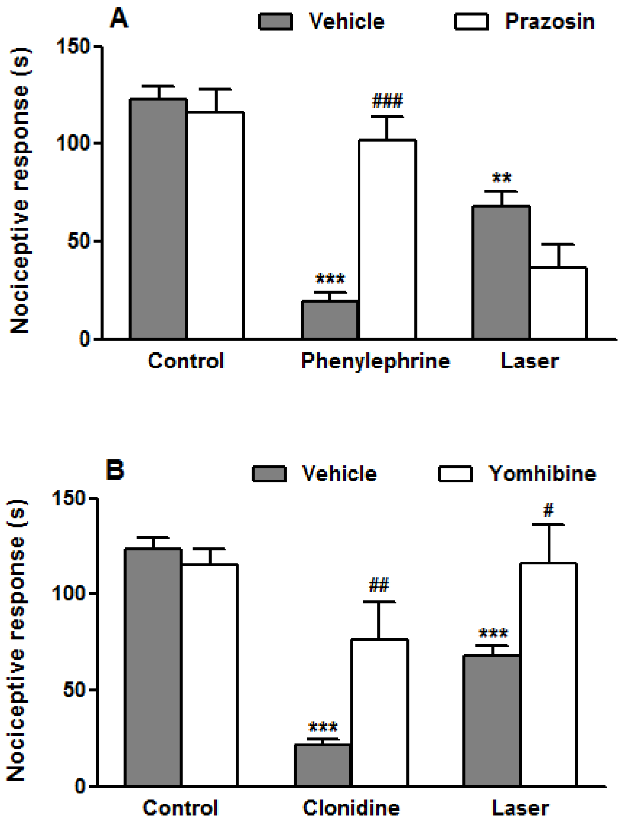


Figure 3. Effect of pre-treatment with prazosin (Panel A), yohimbine (Panel B) on the antinociceptive effect of ST36 laser acupuncture, phenylephrine, vehicle or clonidine. Each group represents the mean of 8 animals, and the vertical lines indicate the S.E.M.in the ***p<0.001 when comparing to control group; $^{\#\#}$p<0.01 and $^{\#}$p<0.05 when comparing antagonist vs antagonist + agonist or laser treatment. (one-way ANOVA followed by Newman-Keuls Multiple Comparison Test).

IV. CONCLUSION

LLLT, one of the most recent and promising treatment therapies in the acupuncture, has been shown to reduce to relieve pain significantly. In summary, the present study demonstrated that stimulation of the acupoint E36 shows antinociceptive activity in nociception model with glutamate and that this action seems to be involved in the activation of α2 adrenergic and adenosinergic systems.

ACKNOWLEDGMENT

This study was supported by grants from Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

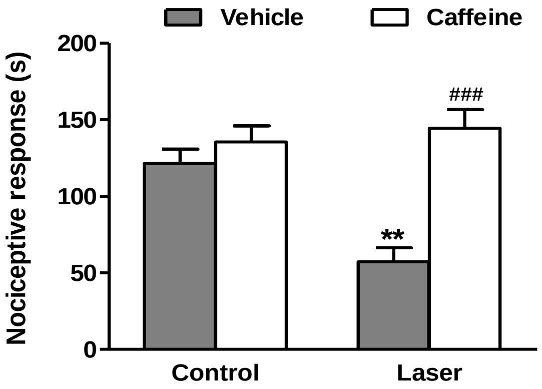


Figure 3. Effect caffeine pre-treatment on the antinociceptive effect of ST36 laser acupuncture in the glutamate test. Each group represents the mean of 8 animals, and the vertical lines indicate the S.E.M. **p<0.01 when comparing to control group; $^{###}$p<0.001 when comparing treatment vs caffeine+ treatment (one-way ANOVA followed by Newman-Keuls Multiple Comparison Test).

REFERENCES

[1] L. Lorenzini, A. Giuliani, L. Giardino, and L. Calzà, "Laser acupuncture for acute inflammatory, visceral and neuropathic pain relief: An experimental study in the laboratory rat," Res Vet Sci, 2009, vol. 88(1), pp. 159-65.

[2] V. Erthal, M. D. da Silva, F. J. Cidral-Filho, A. R.S. dos Santos, and P. Nohama, "ST36 laser acupuncture reduce pain-related behavior in rats: involvement of the opiodergic and serotonergic systems", Lasers Med Sci., 2013, DOI 10.1007/s10103-012-1260-7.

[3] P. Whittaker, "Laser acupuncture: past, present, and future," Lasers Med Sci., 2004, vol. 19(2), pp 69-80.

[4] G. Maciocia, "Os Fundamentos da Medicina Chinesa," Editora Roca, 1996.

[5] P. Y. Limansky, Z. Tamarova, and S. Gulyar, "Suppression of pain by exposure of acupuncture points to polarized light" Pain Res Manage, 2006, vol. 11(1), pp. 49-57.

[6] S. Gottshling, et.al., "Laser acupuncture in children with headache: a double-blind, randomized, bicenter, placebo-controlled trial,". Pain, 2007, vol. 10, pp.1-8.

[7] M. Zimmermann, "Ethical guidelines for investigations of experimental pain in conscious animals," Pain, 1983, vol. 16(2), pp. 109-10.

[8] M. A. Medeiros, N. S. Canter, D. Suchecki, and L. Mello, "c-Fos expression induced by electroacupuncture at the Zusanli point in rats submitted to repeated immobilization,". Braz J Med Biol Res, 2003, vol. 36(12), pp. 1673-84.

[9] A. Beirith, A. R. S. Santos, and J. B. Calixto, "Mechanisms underlying the nociception and paw edema caused by injection of glutamate into the mouse paw," Brain Research, 2002, vol. 924, pp. 219–228.

[10] F. C. Meotti, A. P. Luiz, C. Pizzolatti, C. Kassuya, J. B. Calixto, and A. R. S. Santos, "Analysis of the antinociceptive effect of the flavonoid myricitrin: evidence for a role of the l-arginine–nitric oxide and protein kinase C pathways," The Journal of Pharmacology and Experimental Therapeutics, 2006, vol. 316, pp. 789-796.

[11] M. J. Millan, "Descending control of pain," Progress in Neurobiology, 2002, vol. 66, pp. 355–474.

[12] Y. B. Kwon, M. S. Kang, H. J. Han, A. J. Beitz, and J. H. Lee, « Visceral antinociception produced by bee venom stimulation of the Zhongwan acupuncture point in mice: role of α2 adrenoceptors," Neuroscience Letters, 2001, vol. 308, pp. 133-137.

[13] D. S. Park, B. K. Seo, and Y. H. Baek, " Analgesic effects of eletroacupuncture on inflammatory pain in CIA rats: mediation by alpha-2 and beta-adrenoceptores," Reumatol Int., 2012, DOI 10.1007/s00296-012-2369-5.

[14] G. D. Baxter, C. Bleakley, and S. McDonough, "Clinical effectiveness of laser acupuncture: a systematic review," J Acupunct Meridian Stud., 2008, vol. 1(2), pp. 65-82..

[15] G. Heller P. H. Langen PH, and J. Steffens, "Laser acupuncture as third-line therapy for primary nocturnal enuresis. First results of a prospective study," Urologe A., 2004, vol. 43(7), pp. 803-6.

[16] I. Quah-Smith, et al., "The brain effects of laser acupuncture in healthy individuals: an FMRI investigation," PLoS One, 2010, vol. 5(9), pp.1261-9.

[17] G. Xin-Yan, G. Litscher, K. Liu, and B. Zhu, "Sino-European Transcontinental Basic and Clinical Higth-tech Acupuncture studies-Part 3: Violet Laser Stimulation in Anesthetized Rats," Evid Based Complement Alternat Med., 2012, doi:10.115.

[18] N. Goldman, et.al., "Adenosine A1 receptors mediate local antinociceptive effects of acupuncture," Nature Neuroscince, 2010, doi:10.1038/nn.256.

Simultaneous Ultrasonic Measurement of Vascular Flow-mediated Dilation and Quantitative Wall Shear Stress for Endothelium Function Assessments

Naotaka Nitta
Human Technology Research Institute
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Japan
n.nitta@aist.go.jp

***Abstract*— For early detection of arteriosclerosis, the evaluation of endothelium function has attracted attention in recent years. Flow-mediated dilation (FMD) is the blood vessel dilation due to the smooth muscle relaxation, which is caused by the effect that the wall shear stress (WSS) on the endothelium induces the nitric oxide (NO) production. Therefore, for accurately evaluating the endothelial function, the quantitative %FMD and WSS must be measured simultaneously. In this study, with the aim of assessing the endothelium function accurately, a novel ultrasound system for simultaneous measurements of %FMD and quantitative WSS is presented. A feature of this system is to combine the %FMD and the WSS obtained by considering the ultrasonically-estimated blood viscosity. The system performance was verified through in vitro experiment using bovine blood and in vivo measurement for healthy volunteers. These results revealed the effectiveness of developed system.**

Keywords - flow-mediated dilation; wall shear stress; blood viscosity; ultrasound; endothelium function

I. Introduction

For early detection of arteriosclerosis, the evaluation of endothelium function has attracted attention in recent years. One method for evaluating endothelium function is the flow-mediated dilation (FMD) measurement, as shown in Fig. 1(a). In the typical FMD measurement, blood vessel diameters dilating after avascularization during 5 minutes are measured (Fig. 1(c)), and compared with that before avascularization (resting phase; Fig. 1(b)). This dilating rate is referred to as %FMD, which is an evaluation index for the endothelium function. Here, one of stimulus sources for activating endothelium function is the wall shear stress (WSS). The WSS applied on the endothelium induces nitric oxide (NO) production, and the NO relaxes smooth muscle. Thus, the WSS is an important parameter for evaluating endothelium function. However, conventional equipments for the %FMD measurement measure only the blood vessel dilation.

Therefore, for accurate evaluation of endothelial function, quantitative WSS and %FMD must be measured simultaneously. In this study, a novel ultrasonic evaluation system of endothelium function based on simultaneous measurements of %FMD and WSS is presented. The system performance was verified through in vitro experiment using bovine blood and in vivo measurement for healthy volunteers.

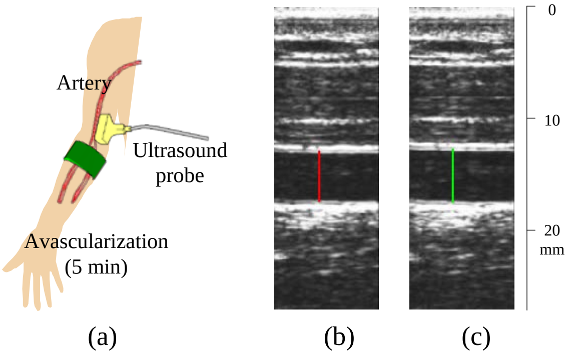


Figure 1. Ultrasonic FMD measurement. (a) A scheme of ultrasonic FMD measurement, (b) blood vessel diameter in resting phase, and (c) blood vessel diameter after releasing avascularization.

II. A Developed System

In the developed system, it is important to simultaneously measure the %FMD and WSS on the central sagittal section of blood vessel. Therefore, we used a H letter-typed ultrasound probe connected to the ultrasound diagnosis equipment (UNEXEF, UNEX, Japan), as shown in Fig. 2. This probe has center frequencies of 8 MHz for blood flow measurement and 10 MHz for vessel imaging and diameter measurement. Moreover, in order to identify the central sagittal section, three apertures for two short axis views and one sagittal view are mounted on the probe. By this configuration, precise positioning of the probe on the central sagittal section, and accurate measurements of %FMD and WSS are available. In the %FMD measurement, forearm artery parallel to the probe surface is often the target for measurement. Therefore, the oblique-incidence beam is available for blood flow measurement in the forearm artery, by using the beam steering technique. Here, the beam steering angle is fixed to 15 degrees.

In the following subsections, methods for measuring %FMD and WSS are described.

A. %FMD measurement

Blood vessel diameter is detected as the inner diameter of blood vessel. Fig. 3 shows the echo signals from the vessel walls. Thus, the vessel walls can easily be detected as echoes with high intensity in the ultrasound image. Therefore, by tracking the high intensity part, the temporal trend of blood vessel diameter can be measured and %FMD is calculated.

ISBN: 978-1-61208-260-8

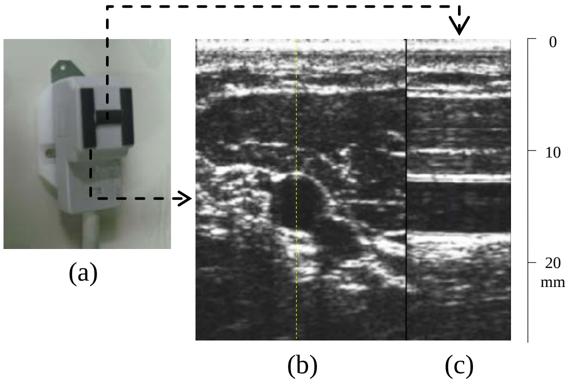


Figure 2. Ultrasound probe for FMD diagnosis. (a) H letter-typed ultrasound probe, (b) short axis view and (c) sagittal view of forearm artery.

B. Wall shear stress measurement

Shear stress is defined as a product of viscosity and shear rate. Therefore, the viscosity and shear rate are necessary for obtaining the quantitative shear stress.

1) SV curve: Fig. 4 shows an example of shear rate-viscosity (SV) curve in whole blood [1]. Since the whole blood is the non-Newtonian fluid, viscosity is higher in the lower range of shear rate and lower in the higher range of shear rate, as shown in Fig. 4. Although a set of methods for obtaining SV curve has presented in the past works [2-4], the methods are described briefly as follows. The 2-D velocity vector distribution is obtained by Doppler measurement and incompressible condition, and the viscosity coefficient is estimated by substituting the velocity vector distribution into the Navier-Stokes equations. In this methodology, non-Newtonian property of blood should be considered. First, to consider this property, intravascular area is divided into several ROIs. Next, in each ROI, kinematic viscosity coefficient is calculated, based on Navier-Stokes equations [2].

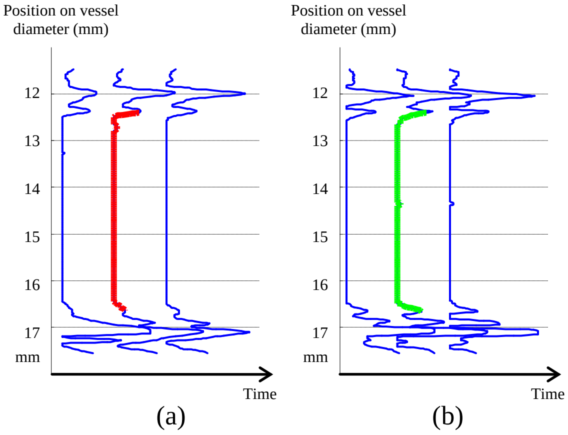


Figure 3. Inner diameter measurements of forearm artery using echoes. (a) Thick solid line indicates the diameter in resting phase. (b) Thick solid line indicates the maximum diameter after releasing avascularization.

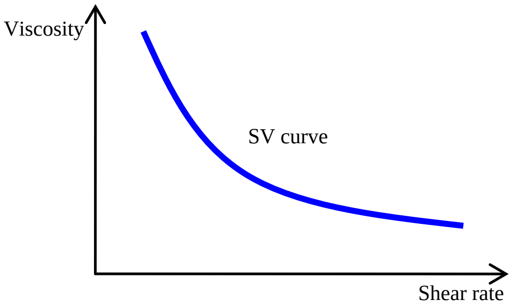


Figure 4. SV curve for obtaining quantitative wall shear stress.

Thirdly, in each ROI, shear rate is also calculated by spatially-differentiating the 2-D velocity vector distribution in blood flow [3]. Finally, shear stress is calculated in each ROI. Viscosity, shear rate and shear stress can be estimated by using only the 2-D velocity vector distribution in each ROI. Therefore, for gathering all ROIs, a SV curve is reconstructed [4].

2) Wall shear stress calculation using SV curve: The SV curve can be modeled as follows.

$$\mu = \alpha \cdot e^{\beta} \qquad (1)$$

Here, μ and e indicate viscosity and shear rate. α and β are regarded as constants determined by only the hematocrit. Based on Eq.(1), WSS σ_{wall} can be obtained as follows.

$$\sigma_{wall} = \alpha \cdot (e_{wall})^{\beta+1} \qquad (2)$$

where e_{wall} indicates the wall shear rate (WSR), that is, the shear rate on the vessel wall. If α and β have already calculated by modeling the SV curve using viscosity and shear rate data in resting phase, the WSS trend can be calculated by measuring the only WSR e_{wall}. Therefore, the WSS can be obtained in real time, if Eq.(2) is used.

III. In Vitro Experiments

Fig. 5 shows an in vitro experimental setup for verifying the system performance. Circulation system using silicone tube with a diameter of 4 mm was constructed. Mean velocity of flow is adjusted by a pump (31.6 cm/s). Fresh bovine blood with non-Newtonian property was circulated in the silicone tube at 37 degrees C. Hematocrit of the blood was 25 %.

Fig. 6(a) shows the comparison between results of ultrasonic measurement and viscometer in the SV curve reconstructions. Both plots exhibited similar property. Fig. 6(b) shows the comparison between the reference (true) shear stress and the measured shear stress. Here, the reference shear stress was obtained by multiplying viscosity by shear rate measured by the viscometer.

As the result, the shear stress obtained by the developed system coincides well with that obtained by the viscometer.

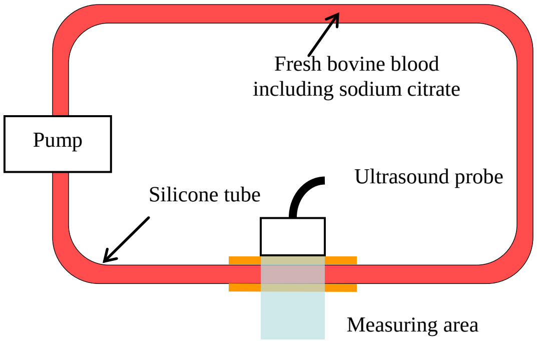
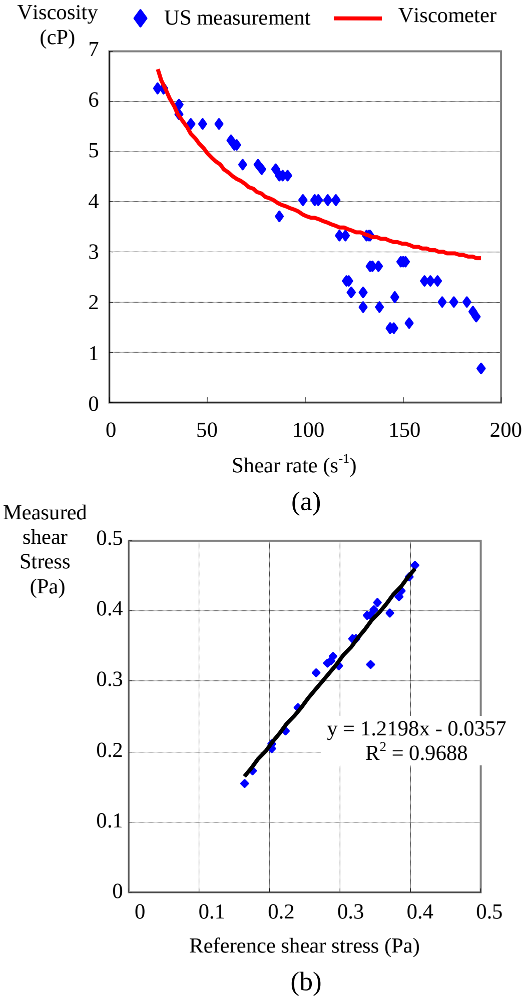


Figure 5. Experimental setup using fresh bovine blood.



(a)

(b)

Figure 6. Experimental results using bovine blood. (a) Reconstructed SV curve, and (b) comparison between measured and reference shear stresses.

IV. In Vivo Measurement

As the next verification, in vivo measurements with the healthy male subjects were conducted. The echo data were acquired by using the developed system, and the temporal trends of blood vessel diameter and WSS were obtained.

A. *Protocol*

The measurement protocol is described as follows.

The simultaneous trends of the vessel diameter and WSS were acquired during 10 seconds in resting phase, and then the trends were acquired during 5 minutes at the avascularization phase. Finally, the trends were acquired during about 2 minutes after releasing the avascularization.

B. *Results*

Fig. 7 shows the intravascular imaging of shear rate, viscosity, and shear stress, respectively. Fig. 7(a)-(c) indicate images in resting phase and Fig. 7(d)-(f) indicate images after releasing avascularization. While the shear rate exhibits the maximum values near the tube wall, the viscosity at the tube center shows the maximum value because the shear rate at the tube center is close to zero. In addition, obviously, the WSS after releasing avascularization is larger than that before avascularization.

Fig. 8(a) and (b) show the examples of inner diameter and WSS trends, respectively. Left plots indicate the trends in resting phase, and right plots indicate the trends after releasing the avascularization. After releasing avascularization, the vessel diameter gradually expands and exhibits the maximum value after about 60 seconds in the post-avascularization phase. On the other hand, the WSS exhibits the maximum value immediately after releasing avascularization.

In order to characterize these properties, the following %FMD and IR_{SS} were introduced on a trial basis.

$$\%\mathrm{FMD} = 100 \times \frac{\Delta d}{d_{rest}} \tag{3}$$

$$\mathrm{IR}_{\mathrm{SS}} = \frac{\mathrm{SS}_{\mathrm{max}}}{\mathrm{SS}_{\mathrm{rest}}} \tag{4}$$

where Δd indicates the difference between the mean diameter d_{rest} in resting phase and the maximum diameter in the post-avascularization phase. The IR_{SS} (Increasing rate of shear stress) means the ratio of the maximum value SS_{max} in the post-avascularization phase to the mean value SS_{rest} in resting phase.

The correlation between %FMD and IR_{SS} for 4 subjects is presented in Fig. 9. High correlation between them was observed. Similarly, IR can also be calculated for mean velocity (MV), maximum velocity (MAXV), flow volume (FV), wall shear rate (WSR), and viscosity (VSC). Therefore, the result of each IR is also presented in Fig. 9. Based on this comparison, the highest correlation between the %FMD and WSS is suggested. Although this suggestion is still preliminary, it is expected that the quantitative WSS is available based on the SV curve in resting phase, and a quantitative evaluation of endothelial function is possible based on the relationship between the flow-mediated dilation and wall shear stress measured by the developed system.

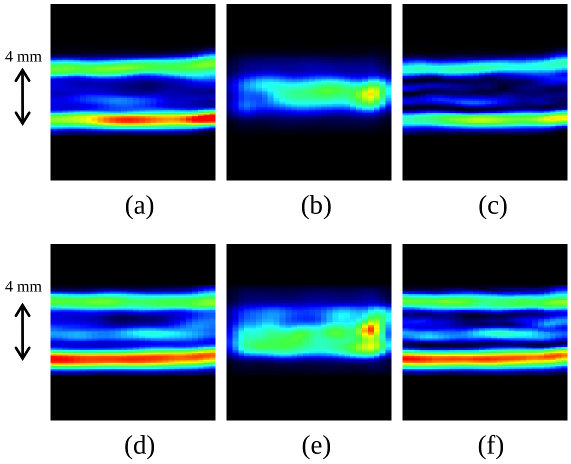


Figure 7. Intravascular imaging of shear rate ((a) and (d)), viscosity ((b) and (e)), and shear stress ((c) and (f)), respectively. (a)-(c) indicate images at rest phase and (d)-(f) indicate images after releasing avascularization.

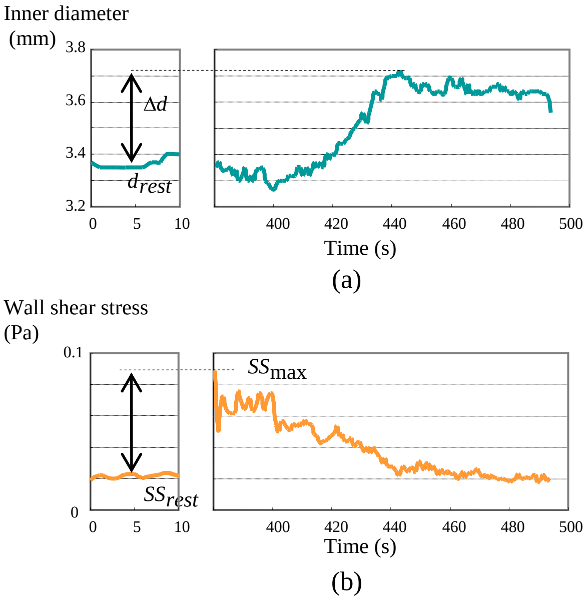


Figure 8. Examples of (a)vessel diameter and (b)WSS trends.

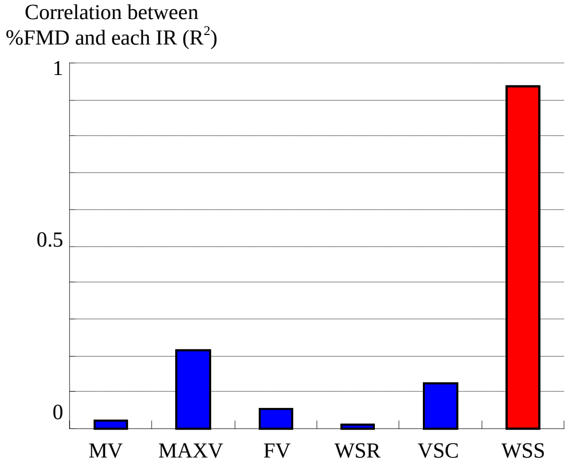


Figure 9. Correlation between %FMD and each IR.

V. Conclusions

In this study, a novel ultrasound system for evaluating endothelial function based on simultaneous measurements of flow-mediated dilation and wall shear stress was developed. In vitro experiment using fresh bovine blood and in vivo measurement for healthy volunteers revealed the effectiveness of the developed system.

In future work, the feasibility in clinical use will be investigated by increasing the number of cases.

Acknowledgment

This work was supported by a Grant-in-Aid for Scientific Research C (22500449) from Japan Society for the Promotion of Science.

References

[1] D. E. Brooks, J. W. Goodwin, and G. V. F. Seaman, “Interactions among erythrocytes under shear,” J. Appl. Physiol., vol. 28, pp.172-177, 1970.

[2] N. Nitta, and K. Homma, “Ultrasonic Measurement of Fluid Viscosity for Blood Characterization,” Jpn. J. Appl. Phys., vol. 44, pp. 4602-4608, 2005.

[3] N. Nitta, and K. Homma, “Real-Time Estimation of Intravascular Shear Stress Distribution Using an Ultrasound Technique,” Trans. Jpn. Soc. Med. Biol. Eng., vol. 44, pp. 190-198, 2006.

[4] N. Nitta, H. Masuda, and H. Suzuki, “Hematocrit Evaluation Based on Ultrasonic Estimations of Shear Rate and Viscosity in Blood Flow,” Proc. IEEE Ultrasonics Symp., vol. 1, pp. 1349-1354, 2010.

Analyzing Meiotic DSB Interference by Combining Southern Blotting and Microarray Analysis

Hiroshi Toyoizumi
Graduate School of Accouting, Waseda University,
Department of Applied Mathematics, Waseda University,
Tokyo, Japan
e-mail: toyoizumi@waseda.jp

Hideo Tsubouchi
MRC Genome Damage and Stability Centre,
University of Sussex,
Brighton, United Kingdom
e-mail: H.Tsubouchi@sussex.ac.uk

***Abstract*—It is well-known that crossover formation events on a chromosome interfere with each other during meiosis, and this interference affects the distribution of genetic exchanges on a chromosome in sexual reproduction. However, due to the technical difficulties, it is unknown if meiotic double strand break (DSB) formation, the initiating event of meiotic recombination, shows interference. We discuss a method that employs probability theory of survival analysis in conjunction with: chromosome fragment distribution, detected by Southern blotting; and genome-wide DSB intensity maps, obtained by microarray analysis. We show that this method is a promising tool to analyse DSB interference.**

***Keywords*-meiosis; DSB; Southern blotting; crossover interference; non-homogenous Poisson process; survival analysis.**

I. Introduction

Meiosis is a specialized cell cycle essential for producing gametes in sexual reproduction [1]. During meiosis, DNA double-strand breaks (DSBs) are programmed to be formed and induce homologous recombination. The homologous recombination mechanism facilitates recognition of homologous chromosomes and establishes physical connections between them via crossovers. The study of meiotic DSB formation and homologous recombination is important because these events are not only central to the life cycle of sexually reproducing organisms, but they are also a driving force for the production of genetic diversity.

Crossover interference is a phenomenon that is known to influence the distribution of crossovers such that the presence of a crossover reduces the likelihood of another crossover forming nearby [2]–[4]. However, it is not known if there is a similar interference mechanism operating at the level of DSB formation, since unlike the case of crossovers, DSB interference can not be observed directly [2].

The formation of meiotic DSBs and crossovers is controlled by many complex biological processes and the mechanism has been intensively studied using various methods [5]–[7]. There are two popular ways to analyse DSB formation: Southern blotting and whole genome mapping obtained by microarray analysis. Both methods fall short of analysing potential DSB interference, but by combining both methods we can analyze the strength of DSB interference. Recently, by assuming that there is no strong DSB interference, we derived an algorithm to estimate the number of DSBs from the experimental results of Southern blotting [8]. In this paper, we show that, by reversing the logic, we can check for the presence of DSB interference. In a large part of past linkage analysis, genetic recombination, which is the consequence of DSB formation, has been assumed to be a non-homogenous Poisson process [3], [9]. In the context of survival analysis with the partial observation, the non parametric Nelson-Aalen estimator has been intensively used to estimate the cumulative hazard rate from censored data [10]–[12]. By simply comparing the two DSB intensities, the one obtained from microarray analysis and the other based on the Nelson-Aalen estimator, which will coincide if DSB formation follows a non-homogenous Poisson process, we can analyze the DSB interference on a chromosome.

II. Southern Blotting and Whole Genome Mapping by Microarray Analysis

A. Southern Blotting and Distribution Function of First Break

A common molecular biology technique called Southern blotting (see Fig. 1) enables detection of one unique location in the genome, making it possible to examine DSB formation per given chromosome. DSB formation can then be studied in mutants of interest, such as those that form but do not repair DSBs [13], [14]. In Southern blotting, the total genomic DNA prepared from cells introduced into meiosis is separated according to size by gel electrophoresis. The separated DNA molecules are then transferred to a nylon membrane on which broken chromosome fragments carrying one end of a chromosome to be examined are probed with a radioactive oligonucleotide. Thus, when a chromosome is intact (i.e., no DSBs are formed), only a single band appears at the location corresponding to the size of the whole chromosome. Once DSBs are formed, chromosomal DNA is broken and smaller molecular pieces appear accordingly, producing numerous bands below the position of the intact chromosome. Although this method is suitable for determining the location of DSBs along chromosomes, the strength of the signal at a given location

ISBN: 978-1-61208-260-8

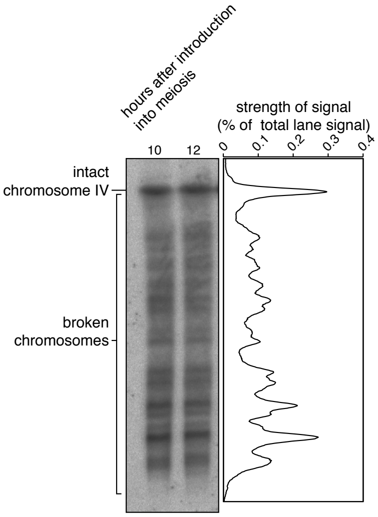


Figure 1. An example of DSB detection by Southern blotting. Mutant budding yeast diploid cells (the *sae2* mutant) were introduced into meiosis and cells were harvested at 10 and 12 hours after introduction into meiosis. The Southern blotting was used to detect chromosome IV. Lane profiles of 10 and 12 hours in each mutant background were normalized and averaged to obtain the profiles shown on the right [13], [14]. This figure is adopted from [8].

does not necessarily correlate with the actual number of DSBs formed there. This is because, when two or more DSBs are formed on a chromosome, only the DNA fragment carrying the chromosome end hybridised to the probe is detected, while others are invisible. In other words, we can only detect the first DSB from one end of the chromosome using this approach.

Let T_1 be the size of the first DSB from the left end of a chromosome. Define the probability distribution of T_1 as $F(x) = P\{T_1 \leq x\}$, which is assumed to be continuous and differentiable for simplicity. Here, the probability measure P can be regarded as the sample ratio of Southern blotting. Thus, by using these sample ratio values obtained from Southern blot analysis, we can estimate the distribution $F(x)$.

B. Microarray Analysis and DSB intensity

DSBs formed during meiosis are exonucleolytically resected from their ends, producing 3'-ended single-stranded DNA (ssDNA). These ssDNA molecules can be selectively recovered by using Benzoyl naphthoyl DEAE (BND) cellulose. Based on microarray analysis of break-associated ssDNA isolated by BND cellulose enrichment, the intensity and distribution of meiotic DSBs were measured [15].

Let t be the position from left end point on a chromosome and $N(t)$ be the number of breaks in this region $(0, t)$. The intensity of DSBs is defined by

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{E[N(t + \Delta t) - N(t)]}{\Delta t}, \tag{1}$$

where the expectation can be regarded as the sample averages of the signal obtained from the microarray analysis. This method is very effective and precise about the location of DSBs, and is used to locate the hot spots and cold spots on chromosomes. However, this analysis cannot reveal which DSBs occurred on a particular chromosome, thus we cannot test for DSB interference.

III. Interference Function and Hazard Rate Function

Interference is often measured by the coincidence function [3], [9]. Let $E(t_1, t_2)$ be the event that a break is in the interval $[t_1, t_2)$ on a chromosome. The coincidence function is then defined by

$$c(s,t) = \frac{P\{E[s, s + \Delta t], E[t, t + \Delta t]\}}{P\{E[s, s + \Delta t]\}P\{E[t, t + \Delta t]\}}, \tag{2}$$

for some small positive Δt. The formation of DSBs is said to have positive interference when a break constrains the formation of other breaks. Thus, $P\{E[s, s + \Delta t], E[t, t + \Delta t]\}$ is smaller than the product probability $P\{E[s, s + \Delta t]\}P\{E[t, t + \Delta t]\}$ and $c(s,t) < 1$. If there is no interference and any two breaks occur independently, the numerator and denominator of (2) coincide, and $c(s,t) = 1$. It is well-known that, in the case of crossovers, $c(s,t) < 1$ and there is positive interference among crossovers.

However, the coincidence function is not useful for analysing DSB interference, because neither Southern blotting nor whole genome mapping can identify two breaks occurring on a single chromosome. Hence, $P\{E[s, s + \Delta t], E[t, t + \Delta t]\}$ cannot be estimated.

Here, we propose a similar but different indicator for DSB interference based on the intensity of DSBs and the hazard rate function of the first break. Let $h(x)$ be the hazard rate function of the first break T_1, which is defined as

$$h(x)dx = P\{x \leq T_1 < x + dx | T_1 \geq x\} = \frac{f(x)dx}{1 - F(x)}, \tag{3}$$

where $f(x) = dF(x)/dx$ is the density of the random variable T_1. We define a new function $\tilde{c}(t)$ by

$$\tilde{c}(t) = \frac{h(t)}{\lambda(t)} = \frac{f(t)}{\lambda(t)(1 - F(t))}. \tag{4}$$

Roughly speaking, the coincidence function $c(s,t)$ measures the tendency of double breaks around t and s, while $\tilde{c}(t)$ measures the tendency of breaks at the position t given no breaks in $[0, t)$. When DSB formation shows positive interference, $E[N(t + dt) - N(t) | N(t) > 0] < E[N(t +$

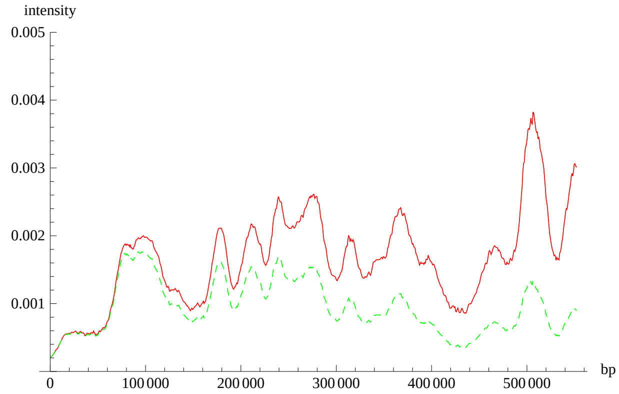


Figure 2. The hazard rate function $h(t)$ (red line) and the size density $f(t)$ (green dashed line) deduced from Southern blotting on chromosome no. 11. The data is adopted from [13].

$dt) - N(t)]$. Thus, we have

$$\begin{aligned}\lambda(t)dt &= E[N(t+dt) - N(t)|N(t) = 0]P\{N(t) = 0\} \\ &\quad + E[N(t+dt) - N(t)|N(t) > 0]P\{N(t) > 0\} \\ &\quad < h(t)dt(1 - F(t)) + \lambda(t)dtF(t),\end{aligned}$$

and $\lambda(t) < h(t)$, and thus $\tilde{c}(t) > 1$. The hazard rate function $h(t)$ coincides with the intensity $\lambda(t)$ and $\tilde{c}(t) = 1$ when there is no DSB interference and $N(t)$ is a non-homogeneous Poisson process (see [8] for example).

IV. Preliminary Results

Using existing data of Southern blotting and genome wide DSB intensity maps, we have obtained some preliminary results about DSB interference. Fig. 2 depicts the hazard rate function $h(t)$ deduced from the DSB size distribution of Southern blotting [13] (see [12] for the estimation method based on Nelson-Aalen estimator). Also, Fig. 3 is the intensity $\lambda(t)$ from the microarray analysis of [15]. Since the two datasets come from completely different experiments, the signal levels are not adjusted. Here, we use a simple normalization to rescale the intensity $\lambda(t)$ to satisfy

$$\int_0^c h(t)dt = \int_0^c \lambda(t)dt, \tag{5}$$

which should be valid when there is no DSB interference, since in that case, $h(t) = \lambda(t)$. The resulted rescaled intensity $\lambda(t)$ is compared with the hazard rate function $h(t)$ in Fig. 4. We use the moving average of 20 signals of the original data of $\lambda(t)$ for noise filtration.

The two curves $h(t)$ and $\lambda(t)$ are agreeable with the locations of hot spots and cold spots. However, $h(t)$ is smaller on the left hand side of the chromosome, and especially at around 400 Kbp, where we see a significant difference between $\lambda(t)$ and $h(t)$. Note that $f(t)$ corresponds to the intensity for the strongest positive DSB interference and only one break exists on the chromosome. The interference

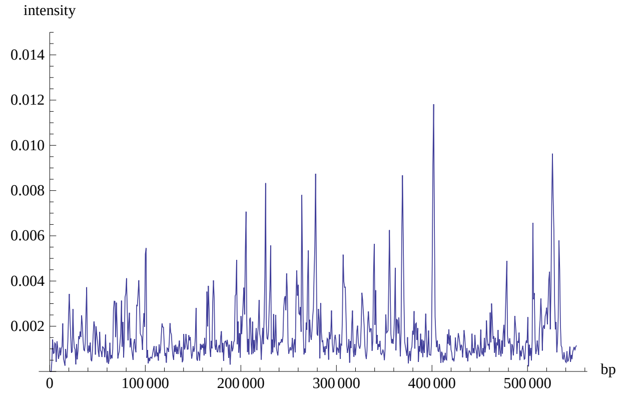


Figure 3. The intensity function $\lambda(t)$ on chromosome no. 11 deduced from whole genome analysis. The data is adopted from [15].

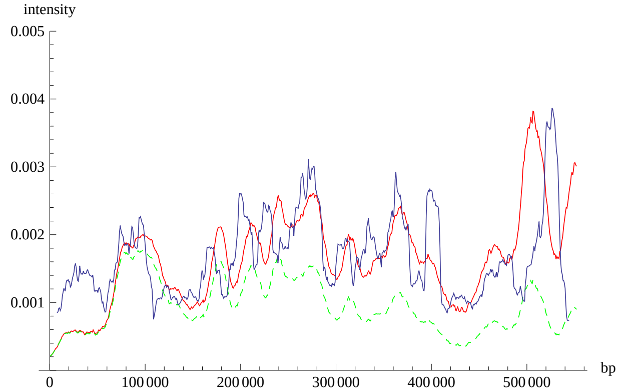


Figure 4. The hazard rate function $h(t)$ (red), the smoothed and scaled intensity function $\lambda(t)$ (blue) and the size density function $f(t)$ (green dashed).

function $\tilde{c}(t)$ in Fig. 5 also shows no or very small positive interference (less than $\tilde{c}(t) < 1.5$). Most of the variance of $\tilde{c}(t)$ can be explained by the poor calibration of two datasets obtained from different experiments. Thus, we can conclude that there is no or very weak positive DSB interference on chromosome no. 11 of budding yeast in the sae2 mutant background.

V. Conclusion

In this paper, we discuss the possibility of analyzing DSB interference by combining the results from Southern blotting and microarray analysis. We also included some preliminary results about DSB interference, but we should point out that the results obtained by applying our method are not final and we need further analysis to discuss the existence of DSB interference.

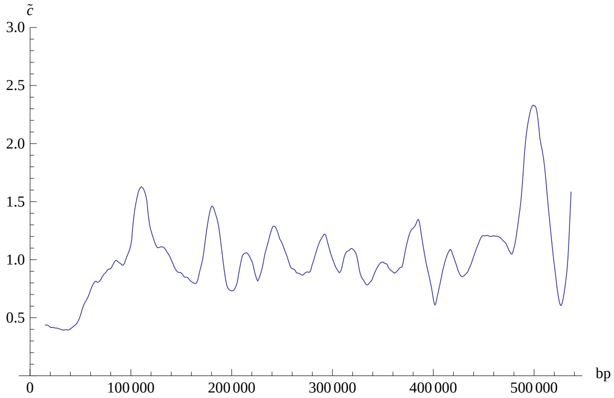


Figure 5. The interference function $\tilde{c}(t)$. $\tilde{c}(t)$ is supposed to be 1 when there is no interference.

References

[1] M. Petronczki, M. F. Siomos, and K. Nasmyth, "Un ménage àquatre: The molecular biology of chromosome segregation in meiosis," *Cell*, vol. 112, no. 4, pp. 423–440, 2 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0092867403000837 Accessed March 2013

[2] L. Berchowitz and G. Copenhaver, "Genetic interference: don't stand so close to me," *Current Genomics*, vol. 11, no. 2, p. 91, 2010.

[3] M. S. McPeek and T. P. Speed, "Modeling interference in genetic recombination," *Genetics*, vol. 139, no. 2, pp. 1031–1044, 1995. [Online]. Available: http://www.genetics.org/content/139/2/1031.abstract Accessed March 2013

[4] S. Karlin and U. Liberman, "A natural class of multilocus recombination processes and related measures of crossover interference," *Advances in Applied Probability*, vol. 11, no. 3, pp. 479–501, 1979.

[5] S. Keeney, *Mechanism and control of meiotic recombination initiation*. Academic Press, 2001, vol. Volume 52, pp. 1–53. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0070215301520086 Accessed March 2013

[6] W. Crismani, C. Girard, N. Froger, M. Pradillo, J. L. Santos, L. Chelysheva, G. P. Copenhaver, C. Horlow, and R. Mercier, "Fancm limits meiotic crossovers," *Science*, vol. 336, no. 6088, pp. 1588–1590, 06 2012. [Online]. Available: http://www.sciencemag.org/content/336/6088/1588.abstract Accessed March 2013

[7] A. Lorenz, F. Osman, W. Sun, S. Nandi, R. Steinacher, and M. C. Whitby, "The fission yeast fancm ortholog directs non-crossover recombination during meiosis," *Science*, vol. 336, no. 6088, pp. 1585–1588, 06 2012. [Online]. Available: http://www.sciencemag.org/content/336/6088/1585.abstract Accessed March 2013

[8] H. Toyoizumi and H. Tsubouchi, "Estimating the number of double-strand breaks formed during meiosis from partial observation." *J Comput Biol*, vol. 19, no. 12, pp. 1277–1283, Dec 2012. [Online]. Available: http://dx.doi.org/10.1089/cmb.2012.0209 Accessed March 2013

[9] J. Haldane, "The combination of linkage values and the calculation of distances between the loci of linked factors," *J. Genet*, vol. 8, no. 29, p. 309, 1919.

[10] P. K. Andersen, Ø. Borgan, N. L. Hjort, E. Arjas, J. Stene, and O. Aalen, "Counting process models for life history data: A review [with discussion and reply]," *Scandinavian Journal of Statistics*, vol. 12, no. 2, pp. 97–158, 01 1985. [Online]. Available: http://www.jstor.org/stable/4615980 Accessed March 2013

[11] P. Andersen, *Statistical models based on counting processes*. Springer Verlag, 1993.

[12] T. Fleming and D. Harrington, *Counting processes and survival analysis*. Wiley Online Library, 1991, vol. 8.

[13] S. Farmer, E.-J. E. Hong, W.-K. Leung, B. Argunhan, Y. Terentyev, N. Humphryes, H. Toyoizumi, and H. Tsubouchi, "Budding yeast pch2, a widely conserved meiotic protein, is involved in the initiation of meiotic recombination," *PLoS ONE*, vol. 7, no. 6, p. e39724, 06 2012. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pone.0039724 Accessed March 2013

[14] S. Farmer, W.-K. Leung, and H. Tsubouchi, "Characterization of meiotic recombination initiation sites using pulsed-field gel electrophoresis," *Methods in Molecular Biology*, vol. 745, pp. 33–45, June 2011. [Online]. Available: http://www.springerprotocols.com/Abstract/doi/10.1007/978-1-61779-129-1_3 Accessed March 2013

[15] C. Buhler, V. Borde, and M. Lichten, "Mapping meiotic single-strand dna reveals a new landscape of dna double-strand breaks in saccharomyces cerevisiae," *PLoS Biol*, vol. 5, no. 12, p. e324, 12 2007. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pbio.0050324 Accessed March 2013

Modeling Patterns of microRNA:mRNA Regulation Through Utilization of Cryptographic Algorithms

Harry C. Shaw
NASA/Goddard Space Flight Center
Telecommunication Networks & Technology Branch
Greenbelt, MD, USA
Harry.c.shaw@nasa.gov

***Abstract*—The properties of microRNA mediated messenger RNA regulation (miRNA:mRNA) of expression are explored by use of an experimental cryptographic representation of miRNA and mRNA structures. The cryptographic approach allows differentiation of identical miRNA:mRNA bonding sequences. Hamming coding, Huffman coding and Rivest, Shamir, Adleman (RSA) encryption techniques are incorporated in the modeling process. Regulation is evaluated as a function of vector projections of mRNA sequences onto a miRNA. The model has a calibration function based upon the use of vector projections onto miRNA complementary and anti-complimentary sequences.**

Keywords-miRNA, mRNA; secondary structure; RSA algorithm; cryptography; hash code

I. INTRODUCTION

A model of microRNA:mRNA regulation and bonding from the standpoint of a cryptographic problem is being developed. miRNA:mRNA bonding exhibits a wide range of regulation. A single miRNA can regulate expression of many mRNAs and a single mRNA can be regulated by many miRNA. The mechanisms that determine specificity are not fully understood. The methodology of the work described in this paper is to treat miRNA:mRNA interaction as a short hash code authentication problem wherein multiple messages with different meanings may generate similar or identical codes. The underlying concept is that miRNA sequences can be treated as short messages. These messages appear to be identical, but are not. In fact, they are uniquely authenticated for a given mRNA target under conditions and conformations that are not totally apparent or understood. The challenge of this experimental modeling approach is that no assumptions are made about the seed and target sequences such as assumptions about evolutionary conservation of sequences, length of seed sequence, number and location of Watson-Crick pairs within seed:target locus, or any of the other commonly used (and often well-validated) assumptions. The goal is to determine if miRNA:mRNA regulation can be modeled within the context of a cryptographic approach.

The paper is structured as follows: A description of the motivation for the research and the potential benefits is provided as well as the challenges in analyzing microRNA (miRNA) mediated regulation. This is followed by a section summarizing miRNA regulation pathway. This is followed by a section summarizing the steps in coding and modeling of the miRNA:mRNA problem. This is followed by a detailed discussion of the coding and modeling of the miRNA:mRNA problem and the resulting output data. The paper concludes with a discussion of model validation, future work and conclusions.

A. Potential benefits to using cryptography to analyze problems in molecular biology

The process of encryption and decryption can be used to accommodate uncertainty and lack of complete information with regards to biochemical pathways. Unknown information that would be essential in creating an accurate physics-based model can be accommodated in a cryptographic model. The cryptographic process allows for wide differentiation between objects that appear to be similar or identical. In the case of microRNA, two similar sequences of 22 bases may have drastically different patterns of post-transcriptional regulation of mRNA expression. By using techniques from coding theory, identical genomic or proteomic sequences can be differentiated by specific coding for secondary structure, tertiary structure or other differentiating characteristics. The cryptographic process permits a hierarchy of coding such that an iterative modeling process can be performed.

B. Advantages of using a cryptography-based process

One major advantage of this approach is that a properly operating authentication process incorporates all of the behavior of the system, including behavior that is currently not well understood or modeled. This includes lack of accurate knowledge of all relevant steric and electrostatic forces, concentration dependencies, and understanding of all relevant molecular interactions (protein-protein, nucleic acid-nucleic acid, protein-nucleic acid, etc.). In this paper, a specific implementation of the protocol will be described that only includes the miRNA to mRNA interaction. New information about a process can be incorporated by expanding the model. The extensible nature of the model will allow for adding details such as RNA Induced Silencing Complex (RISC)-miRNA interaction coding as a future enhancement.

ISBN: 978-1-61208-260-8

II. Challenges in analyzing microRNA mediated regulation

It is known that sequence information alone is insufficient to predict miRNA:mRNA regulation. A single miRNA can regulate numerous mRNA. A single mRNA can be regulated by numerous miRNA. The presence of the seed-target matches does not guarantee downregulation of mRNA. There are structural considerations such as single or double strandedness of the miRNA, and protein bonding domain interaction with miRNA are factors in successful miRNA translational regulation. Both cis- and trans-regulatory effects exist in miRNA regulation as well as cooperative bonding effects on multiple miRNAs amplifying the regulation of a mRNA target. Location of the target within the 3' Untranslated Region (UTR) of the mRNA is a factor, but the target sequence can appear with an open reading frame, and appearance of the target within the 3' UTR is not a necessary and sufficient condition for mRNA regulation.

A. *miRNA Regulation Pathway*

The miRNA gene is transcribed by RNA Polymerase II to produce a primary miRNA transcript (pri-miRNA). Due to the charged nature of RNA, it is never found free and is always complexed with a protein. The pri-miRNA transcript is processed by a complex of DROSHA and DGC8 proteins to a ds-miRNA hairpin pre-miRNA transcript. The pre-miRNA transcript is transported from the nucleus to the cyotoplasm by Exportin 5 (XPO5) that can only bind pre-miRNA in the presence of the RAN-GTPase cofactor [1]. TRBP recruits the RNAse III DICER complex [2] and DICER cleaves the transcript into a ds-miRNA guide and passenger strand. The guide strand complexes with Ago2 and other proteins required for miRNA silencing, e.g., GW182) to form a RNA Induced Silencing Complex. RISC and mRNA targets that associate with sufficient complementarity and energetically favorable structures downregulate (or upregulate [3]) mRNA translational expression .

B. *miRNA:mRNA coding and modeling process*

Each miRNA and mRNA sequence is coded into a 15-bit vectors that represent the nucleotide sequence and assignment of secondary structure. There are 22 rows; one row for every nucleotide in the sequence. The 15 bit binary coded vectors are compressed into two sets of integer vectors (one integer equaling the sum of the 7 most significant bits and the other equaling the sum of the 8 least significant bits) for each miRNA, mRNA, a fully complementary miRNA sequence, e.g., let7d_comp) and a fully anti-complementary, e.g., let7d_anti). The following basic steps are performed:

- Each vector is encrypted using a Rivest, Shamir and Adleman (RSA algorithm) key pair (e,n).
- The encrypted vectors are decrypted with the correct private key and 12 other off-nominal keys in a process called detuning.
- For each 7-bit integer vector and each 8-bit integer vector, a linear vector projection is performed on mRNA vector to the miRNA vectors, the miRNA_comp vectors and miRNA _anti vectors.
- The projections of mRNA vectors onto the miRNA vector are compared to the projection of the mRNA vectors onto the miRNA_comp vector and the mRNA vector onto the miRNA_anti vector
- A regulation of expression score is computed on the basis of these comparisons.

III. Materials and Methods

A. *MATLAB Prototype*

A prototype model built on the MATLAB platform was developed for this research. In this particular model, only nucleic acid interactions at the primary and secondary structure levels are captured. The prototype utilizes the RSA algorithm for encryption. Although the algorithm was not intended for this application, it performs well enough to demonstrate the principles behind the operation of the model. Future versions of the model will utilize a new key pairing algorithm that has been optimized for this application.

B. *Coding of RNA bases*

A binary representation of the four RNA bases is generated from a series of four 11-bit vectors. Each base is represented by a (15, 11) Hamming code. A (15, 11) hamming code codes a 11 bit data field in 15 bits and provides 4 check bits.

C. *Coding of Secondary Structure Dictionary*

RNA folding is necessary for proper functioning of biological activities [4]. The model requires that a secondary structure code be applied to every sequence (mRNA or miRNA), even if the secondary structure under evaluation is modeled as linear, single stranded. The secondary structure coding allows two identical sequences to be differentiated. Therefore the secondary sequence coding provides for an authentication capability. The prototype model has 5 structural categories (classes of symbols) each with a probability mass function that is user-defined. Table I summarizes the categories. The location of each base within the sequence has a probability of being in a given structural category at its location in the nucleotide sequence. The base at that location is assigned a value from the Huffman code dictionary. The dictionary is applied via a language that provides shorthand for performing the secondary structure coding. There are 409 codes assigned to each category.

TABLE I. Secondary Structure Categories

Alpha	Description
X	Double strand, unpaired bulge
P	Double strand, Watson-Crick pair
W	Wobble pair
L	Loop
S	Ss, unpaired

ISBN: 978-1-61208-260-8

D. Secondary Structure Coding

A Huffman code generator produces a set of 2K x 15 bit secondary structure code dictionary to be applied as a hash code to each base in the RNA sequence. A Huffman code dictionary, which associates each data symbol with a codeword, has the property that no codeword in the dictionary is a prefix of any other codeword in the dictionary. The statistical frequency of a given secondary structure can be correlated to its Huffman code. Let *N* equal a finite field of a secondary structure space partitioned into sets of five members, with q = 409:

$$N_k= \{X_i,P_i,W_i,L_i,S_i\},\ 1 \le i \le q,\ 1 \le k \le 22, \tag{1}$$

$$X=\{SS_1^x,SS_2^x,\ldots,SS_q^x\}, \tag{2}$$

$$P=\{SS_1^p,SS_2^p,\ldots,SS_q^p\}, \tag{3}$$

$$W=\{SS_1^w,SS_2^w,\ldots,SS_q^w\}, \tag{4}$$

$$L=\{SS_1^l,SS_2^l,\ldots,SS_q^l\}, \tag{5}$$

$$S=\{SS_1^s,SS_2^s,\ldots,SS_q^s\} \tag{6}$$

Every base in the sequence falls into one of the above sets. The hamming distance, d, between nearest neighbors in any given set is:

$$d(SS_i^x,SS_{i+1}^x) = d(SS_i^p,SS_{i+1}^p) = 3 \tag{7}$$

$$d(SS_i^w,SS_{i+1}^w) = d(SS_i^s,SS_{i+1}^s) = 3 \tag{8}$$

$$d(SS_1^l,SS_{i+1}^l) = 6 \tag{9}$$

RNA structures have been experimentally shown to have a diverse free energy landscape and structural diversity [5]. The use of a space of Hamming codes to represent the multiplicity of energy and mechanical constraints on each molecular structure permits a dynamic modeling approach in which a progression of codes can be modeled to represent the diversity in the free energy states and structure. It remains to be seen if this approach is better than a conventional molecular simulations that rely on two-body additive force field equations or assumptions that involve finding the lowest energy conformation of a RNA sequence.

The use of Hamming distances as a metric to evaluate RNA secondary structures and folding has been utilized other modeling approaches [6], however the taxonomy of the secondary structures being utilized in this approach is very different as well as the method of developing the secondary structure codes.

E. Encryption, Decryption and Detuning

The Asymmetric Encryption Algorithm was developed by Rivest, Shamir and Adleman from MIT to facilitate development public key infrastructure based security. Originally a classified algorithm, it passed into the public domain in September 2000 [7]. The RSA algorithm is a convenient mechanism for generating public/private key pairs for evaluation of secondary structure hash codes. In this model three sets of RSA public/private key pairs are used to eliminate any possible bias that might be introduced by a single set of keys. The vectors representing each sequence will be decrypted with the corresponding private key and with 12 other prime numbers in a process called detuning. The operation of decryption with keys of successively greater distances from the primary decryption key should be traceable to thermodynamic states of greater energy, i.e., lower stability and lower probability of occurrence) for a given sequence. For encryption key *E* = 5, the decryption keys of 53, 47, 43, 41, 37, 31, 29, 23, 19, 17, 13, 11, 7 are used. The key pair of (5, 299) & (53,299) satisfies the RSA algorithm; the other decryption keys are for detuning. Similarly, for *E* = (1657, 24811) & *D* = (73, 24811), *D* = 73, 1559, 1483, 1367, 1013, 1009, 691, 601, 509, 443, 439, 349, 337 are used. For *E* = (89, 5629) & *D* = (233, 5629), *D* = 233, 229, 227, 223, 157, 137, 109, 101, 97, 61, 53, 47, 43 are used. The vector is encrypted with three RSA key pairs, in this case (5, 53), (1657, 24811) and (89, 5629).

The data is formed by the set of error vectors of the decrypted vector projected onto the reference vector. The RSA algorithm uses a public key (e, x) and a private key (d, x). There exist two prime numbers, p and q, p ≠ q. The following relationships apply:

$$x = p * q, \tag{10}$$

$$\phi(x) = (p-1) * (q-1), \tag{11}$$

$$e < \phi(x) \text{ and relatively prime to } \phi(x), \tag{12}$$

$$de=1 \bmod \phi(x)\ . \tag{13}$$

From the calculated key pairs, additional *d's* that do not meet the de=1mod ϕ(x) are utilized as detuning keys. The resulting encrypted vector is decrypted with the decryption key specified by the algorithm and 12 additional keys that are different from the decryption key. The decryption key pairs for (5, 53) are (53, 299), (47,299), (43, 299), (41,299), (37,299), (31,299), (29,299), (23,299), (19,299), (17,299), (13,299), (11,299), and (7,299). The decryption key pairs

ISBN: 978-1-61208-260-8

for (1657,24811) are (73,24811), (1559,24811), (1483,24811), (1367,24811), (1013,24811), (1009, 24811), (691, 24811), (601,24811), (509,24811), (443,24811), (439, 24811), (349,24811) and (337,24811). The decryption key pairs for (89, 5629) are (233,5629), (229, 5629), (227,5629), (223,5629),(157,5629), (137,5629), (109,5629), (101,5629), (97,5629), (61, 5629), (53,5629), (47,5629) and (43, 5629). The data is formed by the set of error vectors described in the following paragraph.

The working hypothesis is that the (d_m, x), (m=1,2,..,13) combination represents a projection of miRNA:mRNA with successively lower probabilities of occurrence as d_m diverges from d_1, e.g., higher minimum free energy for a successful regulation event (either up or down regulation). Each decryption key pair is associated with a probability mass function (pmf), P:

$$P = \{p_1,p_2,\ldots p_m\}=\{\ 2^{-1},2^{-2},\ldots,2^{-13}\},\ m=1,..,13 \tag{14}$$

$$\sum_{m}^{13} p_m \cong 1\ . \tag{15}$$

F. Projection of vector column spaces and generation of error vectors

Each sequence is now represented by a series of decrypted column vectors. The modeling of miRNA association to a mRNA is simulated by the projection of the column space of a decrypted mRNA onto the column space of the decrypted miRNA under evaluation. It is the error vector of the projections that form the data, not the projection vectors. Each representation of a sequence consists of a column space of two vectors, one derived from the 7-bit portion and one derived from the 8-bit portion. Fig. 1 details the vector projection algorithm. Each mRNA:miRNA simulation has 6 sets of error vectors calculated from the error vectors of projections of:

a) 7-bit mRNA on miRNA for all 3 (e,x), (d_m,,x) combinations

b) 8-bit mRNA on miRNA for all 3 (e,x), (d_m,,x) combinations

c) 7-bit mRNA on miRNA complementary sequence for all 3 (e,x), (d_m,,x) combinations

d) 8-bit mRNA on miRNA complementary sequence for all 3 (e,x), (d_m,,x) combinations

e) 7-bit mRNA on miRNA anti-complementary sequence for all 3 (e,x), (d_m,,x) combinations

f) 8-bit mRNA on miRNA anti-complementary sequence for all 3 (e,x), (d_m,,x) combinations

The miRNA complementary sequence and anti-complementary sequence used in this paper is shown in table IV.

$$c_n = \sqrt{E_n^2 + {d_m}^2}, n = 1,2,3 \tag{16}$$

$$\sin\theta_n = E_n / c_n, n = 1,2 \quad \sin\rho = E_3 / c_3 \tag{17}$$

An example of the origin of the variables is shown in table III. E_1, E_2, and E_3 are the value of the error projection for mRNA GATM at the 22nd base position in (starting at position 289 referenced to the 5'end) when projected onto let-7d miRNA, let-7d miRNA complement and let-7d miRNA anticomplement at the same position. The projection is for the encryption key pair (89,5629) and the decryption key pair (43,5629). This leads to geometric relationship shown in Fig. 2.

G. Significance of the relationship between θ and ϕ

θ represents the angular distance between the mRNA:miRNA projection to a perfectly complimentary sequence projected on the miRNA (let-7d comp:let-7d). The smaller the angular distance the greater the similarity between the state of basepairing (bp) in mRNA:miRNA bp and the corresponding mRNA_comp:miRNA. ϕ represents the angular distance between the mRNA:miRNA projection to a perfectly anti-complimentary sequence projected on the miRNA (let-7d anti:let-7d). The smaller the angular distance the greater the similarity between the error vector projection in mRNA:miRNA and the corresponding mRNA_anti:miRNA. It is postulated that there exists an optimum θ and ϕ that represent a maximization of the probability of stabilizing the sequences such that regulation (either up or down) is maximized. The scoring criteria is to retain all scores that satisfy $\sin\theta < r^*\sin\phi$, where $r \leq 1$. For this paper, r = 1. The following scoring logic applies:

$$s_{j,m} = [(\sin\theta_{j,m})/(\sin\phi_{j,m})]^*p_m\ , \tag{18}$$

where j = {1,…,22} nucleotides and m = {1,…13} decryption key pairs.

IV. SCORING

The evaluation of six mRNAs against a single miRNA yields a dataset of up to 20,592 values of ϕ and θ. Large datasets such as this are good candidates for single value decomposition (SVD) analysis is used to extract the principal values and score the data. SVD is widely used to identify the significant elements in large data sets [8]. It is widely used in analysis of gene expression. The entire set for each mRNA:miRNA combination is combined into matrix of the form $X = USV^T$. The score is the product of the most significant members of *U* and *S*. The software allows the user to bound the maximum and minimum scores. Widening the scoring range expands the predicted levels of up or down regulation. The scoring range in table II was (-0.5, +0.5). The score is a direct correlation to the predicted level of post-transcriptional regulation, B_t, by:

$$\mathrm{Score}_t = (U_{1,1} * S_{1,1}),\ B_t = 2^{\mathrm{Score}_j} \tag{19}$$

There are two scores, t = 1 and t = 2 representing the 7-bit and 8-bit integer scores respectively. The two scores provide an upper and lower bound of modeled regulation. The scoring range can span upregulation and downregulation predicted outcomes.

A. *mirSVR Scoring.*

The mirSVR algorithm is a powerful scoring system with a high degree of “ground truth” in its scoring methodology [9]. mirSVR scoring has been calibrated to be linear with log expression change. The mirSVR algorithm was trained on data from mRNA expression data from a panel of microRNA transfection experiments. Target sites are represented by miRNA:mRNA features, and local and global contextual features such as the composition of AU flanking sequences around the presumed mRNA target site. mirSVR uses features derived from the miRanda-predicted miRNA:mRNA sites, the extent of 3’ binding and local features such as AU composition flanking the target site and miRSVR defined secondary structure accessibility score. Other global features of the mirSVR scoring include length of UTR, relative position of target site from UTR ends, and conservation level of the block containing the target site. The downregulation scores from mirSVR correlate linearly with the extent of downregulation. Genes with multiple target sites can be scored by addition of the individual target scores.

B. *Cryptographic Scoring*

The cryptographic model provides a range of two scores. In this comparison, a let-7d sequence and mRNA sequences (from microRNA.org, with modifications to fit the 22 base sequence length model requirement) are programmed as single stranded RNA molecules without any internal Watson-Crick or G:U wobble pairing. The error vectors for the mRNA projected onto the let-7d vector are scored as previously described. Table II summarizes some early results. The scores for multiple target sites can also be summed to provide total mRNA regulation change.

TABLE II. SCORING EXAMPLES

mRNA	Test sequence	start position from 5’ end	mirSVR	Cryptographic high score	Cryptographic low score
C14ORF28	UUUUUUAUUUAUAUGUACCUCA	261	-0.3232	-0.21855	-0.04203
C14ORF28	UGUAUUUCUUUGCCCUACCUCA	423	-0.859	-0.16676	-0.13331
C14ORF28	ACAAUGGAACUUACCUACCUCA	1596	-0.7795	-0.05975	-0.0286
C14ORF28	AAAAAAACAUUUUUCUACCUCU	1676	-1.0895	-0.13636	0.031029
DNA2	CUAUCCUCCCUUACUAUCCUCC	894	-1.2921	-0.07659	-0.05211
HMGA2	UAAAAUUUUUAUUUCUACCUCA	8	-0.3281	0.111411	-0.31562
HMGA2	CAACGUUCGAUUUUCUACCUCA	1244	-0.994	0.083544	0.070764
HMGA2	CACUACUCAAAUUACUACCUCU	1604	-0.0989	0.014682	-0.0218
HMGA2	UACCCUCCAAGUCUGUACCUCA	1655	-0.2446	0.134442	0.05026
HMGA2	GACUUUGCAAAGACCUACCUCC	2213	-0.002	0.411992	0.118015
HMGA2	GUUUCAAAGGCCACAUACCUCU	2507	-0.8811	0.040069	0.001315
HMGA2	AUCAAAACACACUACUACCUCU	2526	-0.1048	0.156676	-0.24195
SMARCAD1	UCUUAAGUCCCAGUAUACCUCA	1498	-0.6971	0.005753	-0.08738
FIGNL2	UCAUGUGUUAAAUACUACCUCC	2227	-0.1242	-0.1562	-0.0833
CTPS2	GCCUAGGUGGGCACCUACCUCA	1315	-0.9344	0.031346	-0.021
FIGN	CAAAACCCAUACUACUACCUCA	635	-0.1957	-0.32631	-0.09044
FIGN	UUGUGAUUUGUACAGUACCUCA	724	-0.3286	-0.21063	0.057325

V. VALIDATION PLAN AND FUTURE WORK

The concepts in this paper require a high degree of validation. The goal is to team with a partner with the laboratory capabilities in running miRNA:mRNA regulation assays to:

- Validate the results with through carefully selected and executed assays.
- Evaluate all aspects of the coding and scoring algorithms.
- Determine the limits of the model to provide an accurate up or down regulation assessment.
- Increase the fidelity of the model by adding ribonucleoprotein-RNA interactions and RISC-RNA interactions.
- Development of a rule set to facilitate automated coding and analysis of miRNA:mRNA model interactions.

VI. CONCLUSIONS

A methodology for using tools from cryptography and information theory to analyze the regulation of mRNA:mRNA interactions has been presented. The work in progress preliminary model results support the concept of operations such that additional work will be conducted to optimize the model and validate the model results.

ACKNOWLEDGMENT

This paper is developed from a dissertation submitted to The George Washington University in partial fulfillment of the requirements for the Ph.D. degree.

ISBN: 978-1-61208-260-8

REFERENCES

[1] R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen, “Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs”, Genes & Development, vol. 17, Oct. 2003, pp.3011–3016, doi/10.1101/gad.1158803.

[2] T. P. Chendrimada, R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura and R. Shiekhattar, “TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing”, Nature, vol. 436, Aug. 2005, pp. 740-744, doi:10.1038/nature03868

[3] W. Ye, F. Qin, J. Zhang, R. Luo, and H. Chen, “Atomistic Mechanism of MicroRNA Translation Upregulation via Molecular Dynamics Simulations”, PLoS One, vol. 7, Aug. 2012, ,issue 8, pp. 1-11, doi: 10.1371/ journal.pone.0043788

[4] E. Mahen, P. Y. Watson, J. W. Cottrell, and M. J. Fedor,“mRNA Secondary Structures fold Sequentially but exchange rapidly in vivo”, PLoS Biology, vol. 8, Feb. 2010, issue 2, pp.1-14, doi:10.1371/journal.pbio.1000307

[5] M. Khalili, W. Kasprzak, M. H. Farris, J. Arroyo, and B. A. Shapiro, “Thermodynamic Studies of RNA Secondary Structure with Genetic Algorithm”, Mitre Corporation Report 09-5182, May 2010.

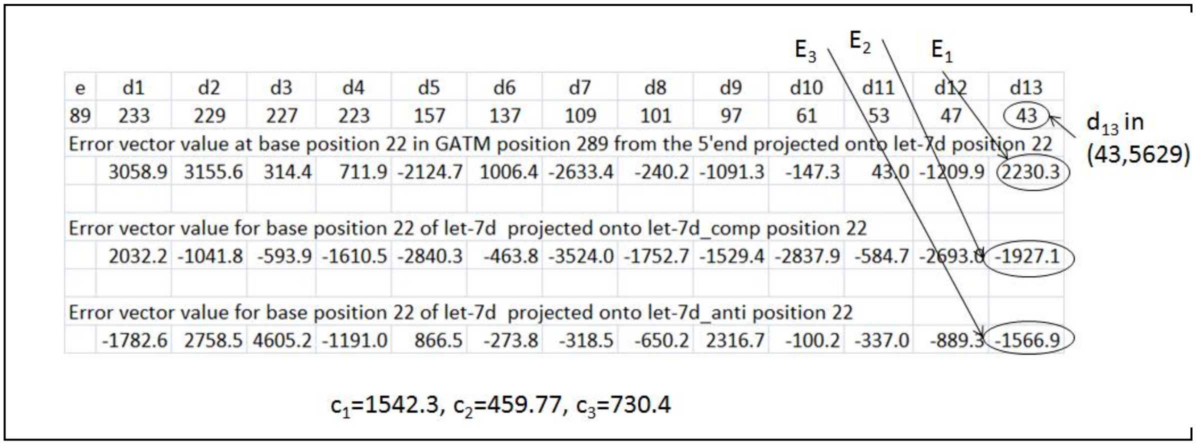
[6] P. Schuster, W. Fontana, P. F. Stadler, I. L. Hofacker, “From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures”, Proceedings: Biological Sciences, vol. 255, Mar. 1994, no. 1344, pp. 279-284.

[7] W. Stallings, Cryptography and Network Security, 4th edition, Upper Saddle River, NJ: Pearson Prentice-Hall, 2006

[8] M. E. Wall, A. Rechtsteiner, and L. M. Rocha, “Singular value decomposition and principal component analysis”, A Practical Approach to Microarray Data Analysis” D.P. Berrar, W. Dubitzky, M. Granzow, eds. pp. 91-109, LANL LA-UR-02-4001, 2003.

[9] D. Betel, A. Koppal, P. Agius, C. Sander, and C. Leslie, “Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites”, Genome Biology, vol. 11:R90, Aug. 2010, pp. 1-14, doi:10.1186/gb-2010-11-8-r90

TABLE III. EXAMPLE OF THE ORIGIN OF THE DATA FOR VARIABLES, E_1, E_2, E_3, d_m

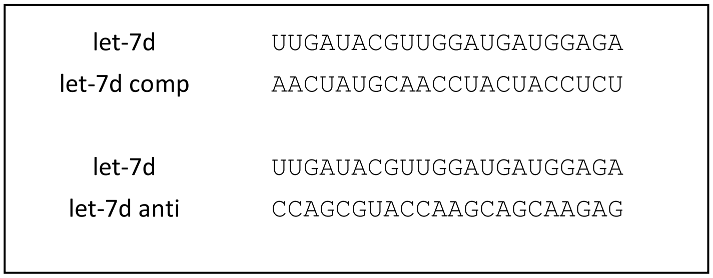


E_3
E_2
E_1
d_{13} in (43,5629)

e	d1	d2	d3	d4	d5	d6	d7	d8	d9	d10	d11	d12	d13
89	233	229	227	223	157	137	109	101	97	61	53	47	43
Error vector value at base position 22 in GATM position 289 from the 5'end projected onto let-7d position 22													
	3058.9	3155.6	314.4	711.9	-2124.7	1006.4	-2633.4	-240.2	-1091.3	-147.3	43.0	-1209.9	2230.3
Error vector value for base position 22 of let-7d projected onto let-7d_comp position 22													
	2032.2	-1041.8	-593.9	-1610.5	-2840.3	-463.8	-3524.0	-1752.7	-1529.4	-2837.9	-584.7	-2693.0	-1927.1
Error vector value for base position 22 of let-7d projected onto let-7d_anti position 22													
	-1782.6	2758.5	4605.2	-1191.0	866.5	-273.8	-318.5	-650.2	2316.7	-100.2	-337.0	-889.3	-1566.9

c_1=1542.3, c_2=459.77, c_3=730.4

TABLE IV. MIRNA CALIBRATION VECTORS



let-7d	UUGAUACGUUGGAUGAUGGAGA
let-7d comp	AACUAUGCAACCUACUACCUCU
let-7d	UUGAUACGUUGGAUGAUGGAGA
let-7d anti	CCAGCGUACCAAGCAGCAAGAG

Copyright (c) The Government of USA, 2013. Used by permission to IARIA. ISBN: 978-1-61208-260-8

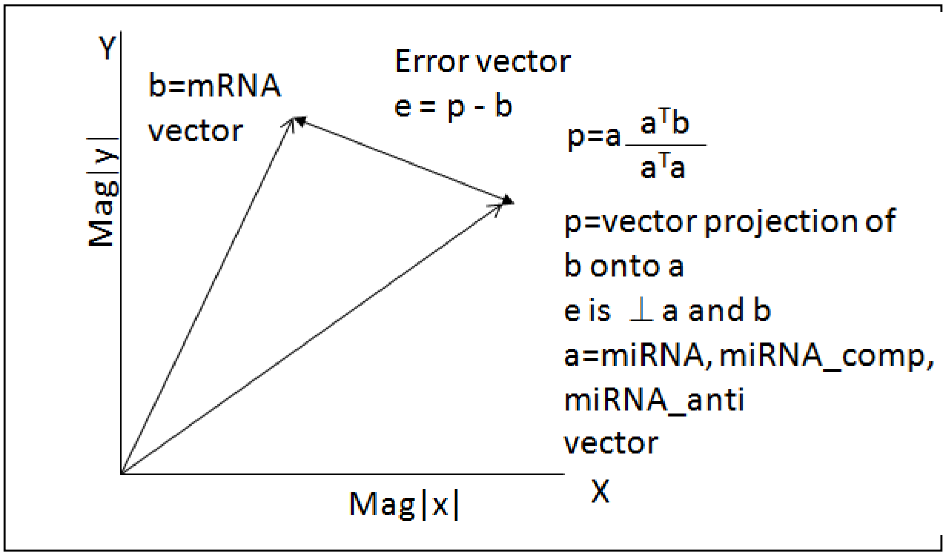


Figure 1. Vector Projection and Error Vector determination algorithms

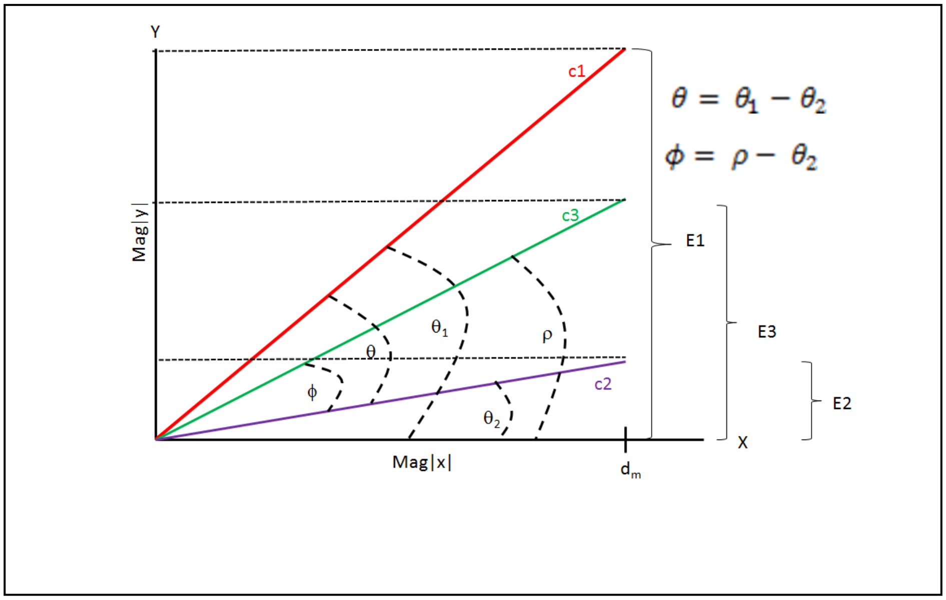


Figure 2. Geometric Relationships in (21) and (22) for example case.

ISBN: 978-1-61208-260-8

Chronological Prediction of Certainty in Recall Tests using Markov Models of Eye Movements

Noaya Takahashi and Minoru Nakayama
Human System Science, The Graduate School of Decision Science & Technology
Tokyo Institute of Technology
Meguro, Tokyo 152-8552 Japan
Email: ntakahashi@nk.cradle.titech.ac.jp, nakayama@cradle.titech.ac.jp

***Abstract*—To predict chronologically certainty levels of understanding of question statements during recall tests using observed eye movements, the hidden Markov model was employed. The feasibility of predicting the accuracy of responses was examined using optimization and simulation of the model together with experimental eye movement data. Chronological prediction accuracy could be calculated using this model, and the accuracy decreased with task difficulty. The highest accuracies were observed at 550 msec. after onset of stimuli. The certainty of correct responses was calculated using the probability of transition. This certainty was the highest during the 100-250 msec. after stimuli onset, and decreased with the duration of the response. These results provide the possibility of estimating the progress of understanding.**

***Keywords*-Eye Movements; Hidden Markov Model; Certainty; User Intention**

I. Introduction

Most web sites ask viewers to provide responses to questions about whether they understand or are interested in the content they are viewing, such as pressing the “Like” button. Even in the human-computer interaction environment, various systems ask users if they understand the specific situation or prefer the service the system provides. For example, many web sites often invite users to make responses such as “yes”, “no” or “Like”, to confirm their understanding of the context of an idea or explanation.

According to the item response theory for tests, response correctness or validity is related to the certainty of the response [1]. Using a different approach, users can make correct responses or provide satisfactory reactions acceptable to the application software, when the level of certainty is high. The theory uses eye movements which occur during the reading of statements to estimate certainty or levels of understanding of statements [2], [3]. It has been proposed that the user’s intent can be predicted using eye fixation behavior [4]. Recently, many studies deal with predicting the viewer’s intent using eye movement features and machine learning procedures, and have achieved a high level of accuracy [5].

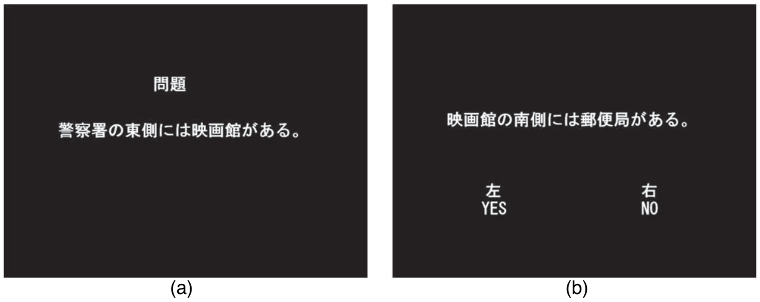
These approaches use all of the data from onset of stimulus to response, and the progress of decision making is not a concern. To investigate the decision making process as the reason why users accept the statements or services offered, a dynamic or time series analysis is required. The process of reading questions has been studied using eye movements to analyze the ways of understanding information which is read [6]. When some mental states are defined, transitions of states can be generated in time series, using behavioral data. For examples, the transition of cognitive states of drivers of automobiles was analyzed using the hidden Markov model (HMM) [7] using the pattern of eye movements [8]. The transition between states, which occurs when reading documents to determine their relevancy was established using HMM and features of eye movements [9]. According to previous studies, it may be possible to create a model to predict the level of certainty and the accuracy of responses using HMM while eye movements during the reading of question statements are observed. As mentioned above, the model may provide the possibility of interpolation for decision making. Analysis of the transition of states using eye movements during the early stages of reading statements will reveal the visual information process. This information may contribute to the improvement of usability of human interfaces and other HCI issues.

In this paper, the hidden Markov model was used to determine the levels of certainty of understanding by observing eye movements. The novelty of this approach is to present a possibility of conducting a dynamic or time series analysis to investigate the decision making process. The feasibility of predicting the accuracy of responses and estimating the levels of certainty were examined while the model was applied to recall tests [3].

Section 2 will summarize the related works. Section 3 will describe the experimental procedure and measurements. Section 4 will explain the development procedure of hidden Markov model for this experiment. Section 5 and 6 will evaluate and discuss the results. Section 7 will summarize our findings.

II. Related work

As mentioned in the introduction section, many studies about assessing usability [10] and predicting user intent [4] using eye movements exist. The ways of approaching the



問題

警察署の東側には映画館がある。

(a)

映画館の南側には郵便局がある。

左
YES

右
NO

(b)

Figure 1. The screen on the left (a) shows a sample of a definition statement: "A theater is located on the east side of the police station." The screen on the right (b) shows a sample of a question statement: "There is a post office on the south side of the theater."

matter can be divided into two categories, discrimination analysis, which uses all of the data, and state transition, which uses time series data. Mental states are defined as explicit in the former and implicit in the latter.

A previous study used two interface operations (zooming-in and zooming-out), which indicate a user's intent. Intent was predicted using a liner discrimination function with eye gaze data [4]. As machine learning studies advance, smaller sizes and sparser data sets can be used to create classifiers, though more specific features of eye movements are required. Various features which are related to a mental status, which is explicit may contribute to performance, as the extraction of features is a key issue in the development of the model. The appropriate combination of features can be used to perform discriminate analysis with a high level of accuracy, therefore the selection of these features is often discussed in detail [3], [5].

Another approach is the analysis of eye movement data as sequential data. The conventional approach is to employ scanpaths of eye movements [11]. Both fixations and saccades of eye movements are analyzed as temporal and spatial behavior data. However, analysis of scanpaths of eye movements is not useful sometimes. Markovian analysis of eye movements has been introduced to estimate user intent. In particular, HMM is preferred for the analysis of sequential data, such as speech recognition, hand writing recognition, analysis of biological sequence and others [7]. Liu (1999) defined some implicit cognitive states for drivers of cars, where the model parameters for left-to-right HMM were estimated from a data set of eye movements. This approach enabled recognition of driver's intentions from the eye movement data [8]. HMM was also applied to an information retrieval task, which consisted of two implicit states such as relevance and irrelevance while eye movements were observed [9]. As a result, the model could provide implicit relevance feedback by making inferences from the eye movement data.

According to the previous studies, implicit states may be flexibly defined if they are related to eye movement behavior. There are a variety of possible applications. Also, sequential behavior analysis of viewers or users may be possible by using their eye movement data. In this paper, a contextual understanding and transition among internal mental states has been employed, and both understanding and the decision making processes are analyzed.

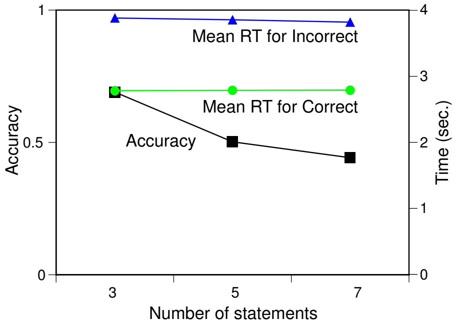


Figure 2. Correct rate and mean reaction times for responses

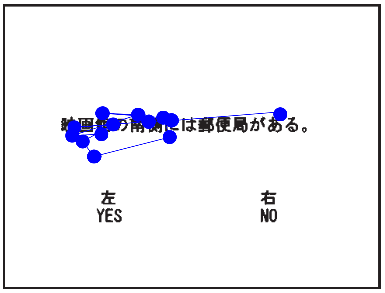


Figure 3. An example of scan-paths

III. Experiment

A. Experimental task

First, participants were asked to understand and memorize a number of definition statements which describe locational relationships between two objects (Figure 1a). Each definition statement was presented for 5 seconds. Second, ten questions were given in statement form, to determine the

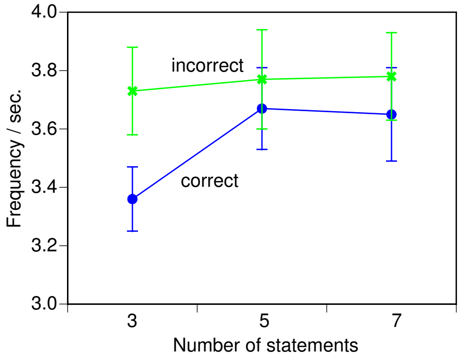


Figure 4. Frequency of saccadic eye movements by response correctness.

degree of understanding. These questions asked participants to select one of two choices about whether each question statement was “Yes (True)” or “No (False)” as quickly as possible (Figure 1b). Each question statement was shown for 10 seconds. All texts were written using Japanese Kanji and Hiragana characters, and the texts were read from left to right. This is a Prolog-like test for human subjects [12]. If the participant can understand the context of the definition statements, they can make the correct decision. No feedback about response accuracy was given. At definition statements 3, 5 and 7 the difficulty of the definition statements increased because the amount of information to be memorized increased. Again, during the experimental session, a set of definition statement containing a number of statements (3, 5, or 7), is presented as shown in Figure 1a, and the 10 question statements in Figure 1b are presented after that. This experimental condition was assigned randomly to participants, to prevent any sequential effects. Five sets of statements were created for each of the three task levels. In total, there were 150 data responses divided into 15 task sets. The subjects were 6 male university students ranging from 23 to 33 years of age. They had normal visual acuity for this experiment.

The correct rates and mean reaction times of responses for question statements are summarized in Figure 2 [3]. As shown in the figure, the response accuracy decreases with the number of definition statements, and mean reaction times depend on whether the responses are correct or incorrect.

B. Eye-movement measuring

The task was displayed on a 20 inch LCD monitor positioned 60 cm from the subject. During the experiment, participant’s eye movements were observed using a video-based eye tracker (EMR-8NL). Eye-movement was tracked on a 640 by 480 pixel screen at 60 Hz, and was recorded on a PC as time course data, while participants read and understood the content of each statement.

The tracked eye movement data was extracted for the period of time participants viewed each question statement before they pressed the mouse button. The tracking data was converted into visual angles according to the distance between the participants and the display. Differences between viewing positions were calculated using the recorded time course data, and eye movements were divided into saccades and fixations using a threshold of 40 degrees per second [13].

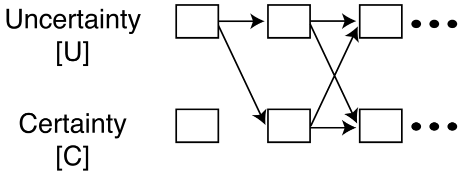


Figure 5. Trellis diagram of a transition.

In Figure 3, eye movements including fixation and saccade are superimposed on a question statement. The dots indicate fixation points, and the lines indicate saccade paths. The appearance of fixations and saccades are related to the correctness of the answers. The frequency of saccades per seconds are summarized in Figure 4. The horizontal axis indicates the number of definition statements, and the vertical axis indicates the saccade frequency. The frequencies are summarized by correct responses and incorrect responses. As shown in the figure, there is a significant difference between answer correctness [14]. Also, selected features of eye movements, such as saccadic eye movements, have been discussed previously [2], [3]. The eye movements are observed at a sampling rate of 60 Hz, and features of eye movement (saccades or fixations) appear in every sample. In this paper, we suppose that the viewer has certainty about an answer when the response is correct. Otherwise, the viewer has uncertainty when the response is incorrect. Though the viewer’s certainness about an answer may be fixed at the time of responding, the certainness is not fixed during the thinking about the answer. According to Figure 3, viewers repeatedly read the question statements. Therefore, their certainness about the answer may change before their final response. Here, two states of certainness about answers can be defined as “Certainty” and “Uncertainty”, and the level of certainness frequently moves between the two states. As Figure 4 shows that eye movements reflect answer correctness, eye movements may also suggest the state of certainty during the reading of question statements.

IV. Hidden Markov modeling

According to the correct and incorrect responses in the task results, it is supposed that there are two internal states used when making choices. As mentioned in the previous section, these two internal states are defined as “Certainty” and “Uncertainty”. The state transition is illustrated as a Trellis diagram in Figure 5. Initially viewers are in a state of “Uncertainty”. Then a transition to “Certainty” occurs after

ISBN: 978-1-61208-260-8

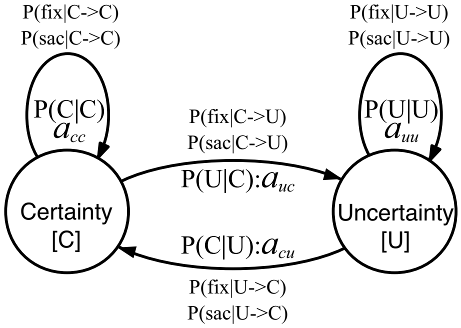


Figure 6. HMM diagram.

that. Also, eye movements may be related to the transition, therefore, specific eye movements such as outputs happen when these state transitions occur.

The outputs are simply saccades (sac) and fixations (fix), as mentioned in the above section. The series of observed eye movements, as the set of time series data, O can be defined using the following formula:

$$\begin{aligned} O &= \{o_1, o_2, o_3, \cdots, o_t\} \\ o_t &= \text{"}sac\text{"} : saccade \vee \text{"}fix\text{"} : fixation \qquad (1) \\ t &= T \times 60 - 1, T : sampling\ time(sec.) \\ & \quad 60 : samplingfrequency(Hz) \end{aligned}$$

In this paper, a hidden Markov model (HMM), which has two states of certainty for contextually understanding question statements, is employed as a dynamic arithmetic model [7]. A diagram of the model is illustrated in Figure 6. The two states indicate levels of high or low certainty for two responses know as Certainty and Uncertainty. When the internal state changes during eye movement, the probability of a correct response is high when the internal state stays high ("Certainty"), and the probability of such a response is low when the state stays low ("Uncertainty"). When the question statements are displayed, the certainty must be low, thus the first transition is movement from a low to high state of certainty. The transition is illustrated as a trellis diagram in Figure 5. The transition happens for every sample, as shown in the figure.

As a result, the model λ can be defined as follows:

$$\begin{aligned} \lambda &= \{S, Y, A, B, \pi\} \\ S &= \{s_i | \text{"}C\text{"} : certain \vee \text{"}U\text{"} : uncertain\} \\ Y &= \{fix, sac\} \qquad (2) \\ A &= \{a_{ij}, i, j = C, U | a_{cc}, a_{cu}, a_{uu}, a_{uc}\}, \sum_j a_{ij} = 1 \\ B &= \{b_{ij}(k), i, j = C, U, k = fix, sac\}, \sum_k b_{ij}(k) = 1 \\ \pi &= \{\pi_C = 0, \pi_U = 1\} \end{aligned}$$

The HMM λ can be measured using a series of output symbols O, as mentioned above. A set of parameters, $\theta \equiv (A, B)$ is optimized using experimental data. The Baum-Welch algorithm, which uses a likelihood function, is employed as shown in (3) [7].

$$\log P(S|O, \theta) \qquad (3)$$

The Forward algorithm provides a series of state transitions S, which maximize the likelihood function [7].

The performance of predicting the states and responses was evaluated using the leave-one-out technique. One set of responses, such as a data set for one level of difficulty for one subject, was used as a test set, while the remainder of the data was assigned as training data, to optimize the set of parameters in θ. Using the model and a series of output symbols of eye movements, state transitions were simulated. As previously stated, correct and incorrect responses were predicted, and the estimation accuracy was also evaluated during the time series.

In Figure 6, circles represent two states, which are levels of high and low certainty. The probability (Pr) of remaining in one state can be defined as c for "certain" and u for "uncertain", while $Pr(c) + Pr(u) = 1$. According to the features of the Markov transition, the transitional probability can be calculated as shown in (4). Then, the probability of certainty can be stated as a time series.

$$\lim_{m \to \infty} \begin{bmatrix} a_{cc} & a_{uc} \\ a_{cu} & a_{uu} \end{bmatrix}^m \times \begin{bmatrix} c^{ini.} \\ u^{ini.} \end{bmatrix} \to \begin{bmatrix} c \\ u \end{bmatrix} \qquad (4)$$

V. Results

The results of the simulation were summarized as a contingency table in Table I, for three of the definition statements, at 550 msec. (0.55 sec.) after onset of the question statement. Here, when a transition remains in a state of certainty at that time, the response may be correct. When the state of uncertainty occurs at that time, the response may be incorrect. According to this hypothesis, the accurate predictions are certainty-correct and uncertainty-incorrect. The prediction accuracy is given as 83.3% ((242+8)/300). The accuracy of the time series is calculated using the same procedure. The results for each condition between 0 and 1 second of the commencement of the statement presentation are summarized in Figure 7. In the figure, "3 tasks" means the prediction accuracy when three definition statements are used.

The accuracy when three definition statements are used is almost always the highest. The accuracy decreases as the number of definition statements increases from 3 to 7. The accuracy of the responses coincides with the level of

Table I
PREDICTION ACCURACY OF 3 STATEMENTS AT 550 MSEC.(ACCURACY=83.3%)

	Prediction	
	Certainty	Uncertainty
Correct	242	41
Incorrect	9	8

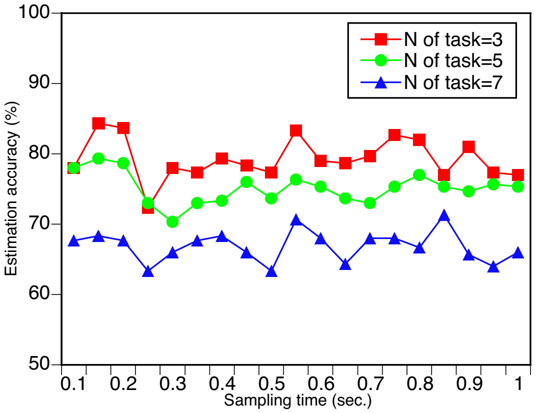


Figure 7. Temporal changes in prediction accuracy.

prediction accuracy. The procedure can not be concerned with the rate of correct responses because the training data consists of a series of equation symbols representing eye movements. According to the changes in prediction accuracy during observation, the accuracy for three statements is high from 0-250 msec. Also, the accuracy for all conditions is high around 550 msec. The highest accuracies, observed at 550 msec., are 83.3% for three statements, 76.3% for five statements, and 70.7% for seven statements. This 550 msec. produces a local accuracy maximum.

The accuracy performance is equivalent to the rates of estimation using all eye-movement data from before the moment when participants responded [3]. Time series prediction produces many response accuracy prediction benefits. During the reading process, accuracy depends on the duration of the decision making process. When eye movements are observed in real time, the response accuracy can be estimated.

According to (4), the probability of remaining in a state of certainty ($Pr(c)$, at $m = t$) can be calculated using the time sequence as well as the sampling rate (t in (1)). Again, the possibility of a correct response is also high when the probability is high at the time. The average probabilities for three levels of task difficulty in the time course are illustrated as bar graphs in Figure 8. The times (t) are between 0 and 1 seconds, and also at 2, 3, and 4 seconds. As shown in Table I, the distributions of both responses and predictions shift to the certain state. Therefore, all probabilities are higher than 0.7, but the changes in probability indicate the sensitivity of correct responses. The probability is high from 100-250 msec., and decreases with the duration. In particular, the rates decrease gradually after 1 second. The mean reaction times during the experiment are around 3 seconds, and the results indicate that the probability of certainty decreases along with the process of reading the statements. These results may suggest that understanding and decision making are occurring at an early stage, prior to decision responses.

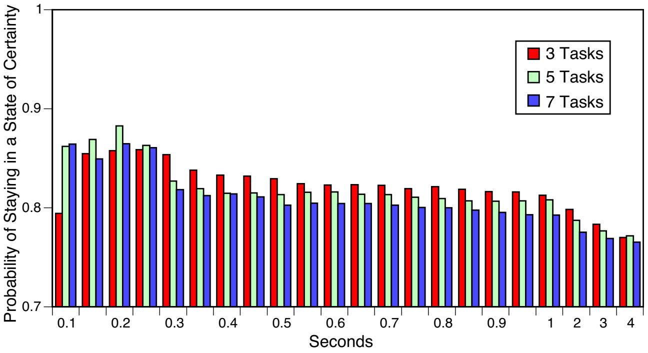


Figure 8. Change in probability of staying a state of certainty by durations.

VI. DISCUSSION

Both prediction accuracy and the probability of remaining in a state of Certainty are affected by the high rate of correct responses. The distribution of training data sometimes shifts to the Certainty state. However, the results for the condition of 7 tasks showed a similar tendency, though the correct rate was smaller than 50%. Since the deviations of the estimation accuracy and the probability along the time course are observed, the model can simulate the transitions of internal states.

The temporal changes in prediction accuracy and the probability of certainty may be concerned with the process of reading statements, because this task requires participants to understand question statements and to recall knowledge memorized from the definition statements. The reading process is tracked using eye movements [6], but the analysis is performed step by step. A typical analysis of the process of reading statements uses event related potentials (ERP). The negative peak is at around 400 msec. (N400) after statement presentation has been affected by the context [15]. This result suggests that statement reading requires more than 400 msec. The local maximum peak of prediction accuracy, concentrated around 550 msec., is included in the same period as the N400 peak. The results of both character perception and eye movement during reading suggest that 150-200 msec. is required to understand the first parts of the statements [6], [16].

According to the evidence, it is possible to suppose that contextual understanding is made between 150 and 400 msec. and that correct or incorrect responses are made after that. When the response is delayed, the level of certainty decreases with the duration, because the viewer can not make a certain decision or hesitates to respond. Further examination of these processes will be a subject of our further study.

In this paper, a recall test was employed to measure transitions from an internal state of “Certainty” to “Uncertainty” and back, using eye movements. This basic approach can be applied to general user interface issues. If the internal states are defined as sets of “Good usability” and “Poor usability”, then “Like” and other selectable icons, eye movements may indicate a change in user’s intentions while the user is interfacing with an application. The examination of this possibility will also be a subject of our further study.

VII. Conclusion

The hidden Markov model, which consists of two hidden states, has been created to predict the correctness of answer choices in response to questions in recall tests. The model was optimized using observations of participant’s responses and their eye movements. The prediction accuracy was calculated using the sequence of eye movement data, and the accuracy decreased with the difficulty of the task. The highest accuracy was observed at 550 msec. after stimuli onset. The probability of choosing the correct response while remaining in a state of certainty was calculated using the probability of transition. The certainty was the highest from 100-250 msec. after stimuli onset, and decreased as the duration increased. The results of the simulation coincided with the experimental observations. These results provide evidences that the model presents a possibility of conducting a dynamic or time series analysis.

Acknowledgement

This research was partially supported by the Japan Society for the Promotion of Science (JSPS), Grant-in-Aid for Challenging Exploratory Research (23650528: 2011-2012).

References

[1] Y. Zhang, *An Item Response Model on Probability-Testing (in Japanese)*. Tokyo, Japan: University of Tokyo Press, 2007.

[2] M. Nakayama and Y. Takahasi, “Estimation of certainty for responses to multiple-choice questionnaires using eye movements,” *ACM TOMCCAP*, vol. 5, no. 2, p. Article 14, 2008.

[3] M. Nakayama and Y. Hayashi, “Estimation of viewer’s response for contextual understanding of tasks using features of eye-movements,” in *Proceedings of ACM Symposium on Eye-Tracking Research & Applications (ETRA2010)*, A. Hyrskykari and Q. Ji, Eds. New York, USA: ACM, 2010, pp. 53–56.

[4] J. H. Goldberg and J. C. Schryver, “Eye-gaze determination of user intent at the computer interface,” in *Eye Movement Research, Volume 6: Mechanisms, Processes and Applications (Studies in Visual Information Processing)*, J. Findlay, R. Walker, and R. Kentridge, Eds. Elsevier, 1995, pp. 491–502.

[5] R. Bednarik, H. Vrzakova, and M. Hradis, “What do you want to do next: A novel approach for intent prediction in gaze-based interaction,” in *Proceedings of ETRA 2012: ACM Symposium on Eye-Tracking Research & Appliations*, J. B. Mulligan and P. Qvarfordt, Eds. New York, USA: ACM, March 2012, pp. 83–90.

[6] K. Rayner, “Eye movements in reading and information processing: 20 years of research,” *Psychological Bulletin*, vol. 124, no. 3, pp. 372–422, 1998.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, USA: Springer Science+Business Media, 2006.

[8] A. Liu, “What the driver’s eye tells the car’s brain,” in *Eye Guidance in Reading and Scene Perception*, G. Underwood, Ed. Elsevier, 1998, pp. 431–452.

[9] Y. Puolamäki, J. Salojärvi, E. Savia, J. Simola, and S. Kaski, “Combining eye movements and collaborative filtering for proactive information retrieval,” in *Proceedings of ACM-SIGIR 2005*, A. Heikkil, A. Pietik, and O. Silven, Eds., ACM. New York, USA: ACM Press, 2005, pp. 145–153.

[10] R. J. K. Jacob and K. S. Karn, “Eye tracking in human–computer interaction and usability research: Ready to deliver the promises,” in *The Mind’s Eye: Cognitive and Applied Aspects of Eye Movement Research*, Hyona, Radach, and Deubel, Eds. Oxford, UK: Elsevier Science BV, 2003.

[11] J. H. Goldberg and X. P. Kotval, “Computer interface evaluation using eye movements: methods and constructs,” *Industrial Ergonomics*, vol. 24, pp. 631–645, 1999.

[12] W. Takita and M. Nakayama, “Influence of audio-visual presentation style for sentences on their understanding,” *IEICE Technical Report*, vol. 104, no. 132, pp. 17–22, 2004.

[13] Y. Ebisawa and M. Sugiura, “Influences of target and fixation point conditions on characteristics of visually guided voluntary saccade,” *The Journal of the Institute of Image Information and Television Engineers*, vol. 52, no. 11, pp. 1730–1737, 1998.

[14] M. Nakayama and Y. Hayashi, “Feasibility study for the use of eye-movements in estimation of answer correctness,” in *Proceedings of COGAIN2009*, A. Villanueva, J. P. Hansen, and B. K. Ersboell, Eds., 2009, pp. 71–75.

[15] M. Kutas and S. A. Hillyard, “Reading senseless sentences: Brain potentials reflect semantic incongruity,” *Science*, vol. 207, pp. 203–208, 1980.

[16] H. Abe and M. Nakayama, “Event-related potential study of Kanji perception process,” *The Journal of the Institute of Image Information and Television Engineers*, vol. 60, pp. 397–404, 2006.

Clustering MicroRNAs from Sequence and Time-Series Expression

Didem Ölçer, Hasan Oğul
Department of Computer Engineering
Baskent University
Ankara,Turkey
e-mail: dtokmak@baskent.edu.tr, hogul@baskent.edu.tr

***Abstract*— Inferring co-operative actions of microRNAs is crucial for analyzing large-scale gene regulatory networks. We introduce here a probabilistic generative model to cluster microRNAs from their mature sequences and time-series expression profiles. Sequence model is defined over the distribution of k-mers, all possible k-length substrings from RNA alphabet. The expression model is built upon a spline-basis function over a Gaussian assumption. Two models are integrated to form a single likelihood. Cluster enrichment analysis has shown that the data integration over a Bayesian framework could improve the clustering ability and produce biologically more plausible patterns.**

Keywords- microRNA expression; microRNA regulation; graphical model; data integration; time-series data analysis

I. Introduction

Post-transcriptional regulation of genes is mainly directed by small non-coding RNAs called microRNAs (miRNAs). It has been shown that they are abundantly found in many organisms and affiliated with several biological processes such as development, aging and apoptosis [4][5][9][12][17]. It is proven that various diseases are associated with the abnormal behaviors of specific miRNAs [13][14][18]. Recent studies have shown that miRNAs usually operate in a co-operative manner to perform their activities [1]. This suggests that some miRNAs can form context-specific modules, i.e., cluster of entities, while regulating gene expression. Since the elucidation of gene regulatory networks comprising all actors is one of the ultimate goals of systems biology, which miRNAs are functionally similar in a certain context is high-potential knowledge for the researchers and clinicians working in this domain. Here, a functional similarity refers a common regulatory behavior in a certain context, e.g., a specific disease condition or temporal response to a stimuli.

Several features can be employed to infer functional miRNA clusters. An obvious indicator for miRNA's regulatory function is its expression profile. Its differentiation usually results with a consequential change in the expression of its target genes, thus in relevant regulatory pathways. On the other hand, a similarity between the expression profiles of two miRNAs does not necessarily imply a similarity about their genome-wide functions. Several other factors may affect the regulation of miRNAs, and therefore they may arbitrarily express in a similar way. Sequence information can also unveil the structural similarity between miRNAs since the target selection process is usually mediated by a complementarity between mature miRNA sequence and its target mRNA sequence [4]. We can easily argue that two miRNAs having similar sequences will have similar binding preferences, which lead them in target mRNA regulation. However, it was shown that a miRNA may not always be active in a certain context although its binding affinity is very high [6]. Hence, sequence information alone is not expected to give reliable results in miRNA functional similarity associations. In this study, we propose to use both information in a single model to obtain functional miRNA clusters. While designing our model, we were inspired from Kundaje et al. [11] where they combined the promoter sequence motifs with gene expression profiles to infer transcriptional modules. We adopt their model for miRNA expression profiles and propose a novel approach to integrate mature miRNA sequence into overall framework. The framework is built upon a probabilistic graphical model, which simultaneously integrates sequence and time-series expression data to infer coherent miRNA clusters. It enables to adjust and understand the contribution of each information to final cluster assignments. Experimental validation on a real biological data set demonstrates that the integration can improve the clustering ability and produce biologically more plausible patterns.

The rest of paper is organized as follows. We explain methods in Section II. A description of analysis and results can be found in Section III, followed by conclusion in Section IV.

II. Methods

A. Probabilistic Graphical Model

The problem is to learn the functional clusters of miRNAs where their similarity is explained by a common regulatory mechanism at the transcriptional level and consequential regulatory effect in post-transcriptional level. We define a probabilistic framework which assigns the miRNAs to clusters based on two types of data for each miRNA i: its time-series expression profile, i.e., a set of temporal expression values, E_i, and a set of features representing its mature sequence, S_i. We let the variable Z_i refer to the cluster assignment of miRNA i. Since we assume that both sequence and expression of a miRNA is conditioned on its cluster assignment, following graphical model can be used:

$$S \leftarrow Z \rightarrow E \tag{1}$$

The joint probability distribution for a single miRNA can be written as $P(E_i, S_i, Z_i) = P(E_i|Z_i)P(S_i|Z_i)P(Z_i)$, where E_i and S_i are assumed to be conditionally independent for given cluster assignment Z_i. Having the joint probability model, the task is then to learn the model parameters that maximize the likelihood of input data for a given set of cluster assignments. Since the expression and sequence data have different characteristics in nature, two independent sub-models are provided to define their conditional probabilities.

B. *Sequence Model*

Mature miRNA sequences might be of different lengths, usually between 22-24 nucleotides. For the probabilistic model defined above, the sequence is needed to be modeled by a fixed number of numerical features, which potentially represent its regulatory behavior. We select 3-mer model for this representation. In k-mer model, defined over RNA alphabet *A={'A','G','U','C'}*, a sequence $s_1s_2...s_m$ of miRNA *i* is represented by $S_i = \{n_{i1}, n_{i2}, \dots, n_{iP}\}$ where n_{ij} denotes the count of *j*th *k*-length substring among all possible substrings composed by the same alphabet *A*, and *P* is the number of such substrings. In our case, *3*-mer representation involves $P=4^3$ distinct count values of all possible 3-length substrings from *A*. This scheme is able to consider the content of the miRNA sequence as well as the order of residues inside the sequence, which is one of the major determinants of miRNA binding. Similar representations have been successfully applied in several domains [15].

For model integration, we represent the mature sequence for miRNA i as the sparse vector S_i of count of k-mers that it contains, where S_i is indexed by all possible k-mers: $S_i = \{n_{i1}, n_{i2}, \dots, n_{iP}\}$. We let $n_i = \sum_{p=1\dots P} n_{ip}$ be the total count of k-mers observed in miRNA i. For each cluster j, we define another vector of k-mer frequencies observed in the miRNAs of same cluster; $\theta_j = (\theta_{j1}, \theta_{j2}, \dots, \theta_{jp})$, where $\sum_{p=1\dots P} \theta_{jp} = 1.$ The sub-model for miRNA sequence then becomes a multinomial model, defined by the following conditional probability;

$$P(S_i|Z_i = j, \theta_j) = \frac{n_i!}{n_{i1}!n_{iP}!} \prod\nolimits_{p=1P} \theta_{j_p}^{n_{ip}} \tag{2}$$

With the assumption that observation of each k-mer is independent from each other, the model parameter to be evaluated here is θ_j.

C. *Expression Model*

In the expression model, we define each cluster by a Gaussian distribution over spline parameters that model the common time-course behavior of its member miRNAs. This model was originally proposed by Bar-Joseph et al. [3] and successfully applied for inferring temporal gene regulatory mechanisms [3]. In our framework, each miRNA expression profile is represented by a spline curve. More formally, for each miRNA i assigned to cluster j, its expression profile is given as a function of time as $(f_1(t) \dots f_q(t))(\mu_j + \gamma_{ij})$, where $f_1(t), \dots, f_q(t)$ are spline basis functions. Here, μ_j denotes the mean of coefficients for cluster j, q denotes the number of spline control points used and γ_{ij} is the miRNA specific variation of coefficients, which is treated as a latent variable. γ_{ij} is assumed to be normally distributed with mean 0 and covariance matrix Γ_j . $\epsilon \sim N(0, \sigma^2)$ is the random Gaussian noise. If we have m time points of observation denoted by $t_1, \dots, t_m$, the expression profile is given as:

$$E_i = \begin{pmatrix} f_1(t_1) & \cdots & f_q(t_1) \\ & \ddots & \\ f_1(t_m) & \cdots & f_q(t_m) \end{pmatrix} \left[\begin{pmatrix} \mu_j^1 \\ \vdots \\ \mu_j^q \end{pmatrix} + \begin{pmatrix} \gamma_{ij}^1 \\ \vdots \\ \gamma_{ij}^q \end{pmatrix} \right] + \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_m \end{pmatrix} \tag{3}$$

The probability of miRNA expression conditioned on any cluster assignment then becomes:

$$\begin{aligned} &P\left(E_{i,\gamma_{ij}} \middle| Z_i = j, \mu_j, \Gamma_j, \sigma^2\right) \\ &= (2\pi)^{-(m+q)/2} |\Gamma_j|^{-\frac{1}{2}} \sigma^{-m}. e^{-1/2\sigma^2 \left(E_i - f(\mu_j + \gamma_{ij})\right)^t (E_i - f(\mu_j + \gamma_{ij})} \\ &\quad e^{-1/2\gamma_{ij}^t \Gamma_j^{-1} \gamma_{ij}} \end{aligned} \tag{4}$$

As suggested by Bar-Joseph et al. [3], natural cubic splines are used where the size q of spline basis is equal to the number of evenly-spaced knots. Optimal value of q and number of clusters are selected using 2-fold cross-validation to maximize the likelihood computed from the sum of the log loss function at each fold.

Parameter estimation and cluster assignment are done using a set of alternating Expectation and Maximization (EM) steps. Initial clusters are obtained by k-means algorithm using only expression data. At each E-step, the algorithm calculates $p(i|j)$, the probability of gene i being in cluster j, and the expectations for latent variable γ_{ij}. M-step updates the parameters based on the expected values. These alternating iterations repeat until the likelihood convergences. The results of last M-step reports the final miRNA clusters and corresponding parameters inferred.

III. RESULTS

We perform our experiments in a recently released miRNA expression data for transcriptome analysis on ovarian cancer [7]. The data set involves the Affymetrix measurements for expression profiles of several miRNA probes at six time points in three replicates. In the experiment, they studied the pathways and growth properties of cultured human ovarian cancer cells that are expressing luteinizing hormone receptor (LHR). Their particular interest was to understand the changes in the expression as a result of the activation of receptor by its cognate ligand, gonadotropin (LH). They used SKOV3 ovarian cancer cell line stably transfected with LHR, and investigated the response of these cells in culture following exposure to LH. They chosen the parent SKOV-3 ovarian cancer cell line, which did not express LHR, as a control in the experiments and observed the alterations in gene expression elicited by LH. Resulting data set is composed of six groups of SKOV-3 cells: LHR- (parent cell line), LHR+ (just after transfection), and LHR+ incubated with LH in four time points: 1, 4, 8, and 20 h. To pre-process the data, we average over three replicates and calculate differential expression all miRNAs with respect to LRH- condition. For integrative analysis, we remove the miRNAs with no sequence information. This gives us a dataset of differential expression profiles of 80 miRNAs at five different time points.

To assess the functional homogeneity of clusters, we use the enrichment of Gene Ontology (GO) terms [2]. We extract a collection of confirmed miRNA targets from the TarBase [16], miRecords [19], miRTarBase_MTI [10] and circuitDB [8] databases. For each cluster, we build a target set by taking corresponding miRNAs and setting the union of their targets. We then evaluate the functional enrichment of GO terms in each target set based on biological process category, using a Bonferroni-corrected hypergeometric test with an original p- value of 0.01.

Two-fold cross-validation on expression data set suggests us an optimal model with *q*=4 (number of spline points) and *c*=5 (number of clusters). We run both single model and integrated model with these parameters. To see the effect of different number of clusters, we also compile the same setup for *c*=10. At the end of each run, we ignore the outliers by removing the clusters having less than three miRNAs from final cluster set.

GO-enrichment test results are shown in Table 1 for sequence model, expression model and integrative model for *c*=5 and *c*=10. For each run, we report the number final clusters obtained and the percentage of clusters with at least one GO-term enriched for targets of more than two miRNAs.

According to Table 1, each single model can achieve a fairly well percentage of GO-enrichments in resulting cluster set. This implies that coherent clusters can be obtained by using either sequence or expression data. On the other hand, the table demonstrates that the integration of two different data types can remarkably increase the number of clusters enriched with significant functional GO-terms. This result obviously suggests that the data integration can improve the clustering ability and helps to obtain biologically meaningful patterns.

TABLE I. COMPARISON OF THE CLUSTERING ABILITY OF MODELS WITH REGARD TO MIRNA TARGET GENE FUNCTIONAL ENRICHMENT

	***c*=5**			***c*=10**		
	Sequence	**Expression**	**Combined**	**Sequence**	**Expression**	**Combined**
Number of clusters	5	4	4	9	8	8
Number of GO-enriched clusters	3	2	4	5	4	6
Percentage of GO-enriched clusters (%)	60	50	100	56	50	75

IV. CONCLUSION

Inferring similar miRNAs can provide valuable information for understanding regulatory mechanisms behind gene expression. Mature miRNA sequence can explain the post-transcriptional regulation of miRNAs, but it cannot give any information about how miRNA itself is regulated. Their expression values can provide some clues about how they are regulated but not about how they regulate since their differential expression might occur due to several random effects. Therefore, the integration of two information sources is essential to discover context-dependent functional miRNA clusters. This study introduces an integrative model to combine two data sources over a probabilistic framework. Two independent models are designed for each type of information, which can also be compiled to obtain only transcriptional (using expression data solely) or only post-transcriptional (using sequence data solely) miRNA groups. The experiments performed on real biological data sets reveals that employing both information can improve the explanatory power of final clusters obtained.

Our study is ongoing to validate our model on larger datasets. Assessing the effects of different parameter selections, such as *k* in *k*-mer analysis, will be another future issue.

ACKNOWLEDGMENT

This study was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) under the Project 110E160.

REFERENCES

[1] A.V. Antonov, S. Dietmann, P. Wong, D. Lutter, and H.W. Mewes, “GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists,” Nucleic Acids Res, 37, 2009, pp. W323-W328.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., "Gene Ontology: Tool for the Unification of Biology," Nat. Genet., 25, 2000, pp.25-29.

[3] Z. Bar-Joseph, G. Gerber, D.K. Gifford, T.S. Jaakkola, and I. Simon, "A New Approach to Analyzing Gene Expression Time Series Data," Proc. 5th RECOMB Conf., 2002, Canada

[4] D.P. Bartel, "MicroRNAs: genomics, biogenesis, mechanism, and function," Cell, 116, 2004, pp. 281-297.

[5] D.P. Bartel, "MicroRNAs: target recognition and regulatory functions," Cell, 136, 2009, pp. 215-233.

[6] C. Cheng and L.M. Li, "Inferring microRNA activities by combining gene expression with microRNA target prediction," PLoS ONE, 3, 2008, pp. 1-9.

[7] J. Cui, J.B. Eldredge, Y. Xu, and D. Puett, "MicroRNA expression and regulation in human ovarian carcinoma cells by luteinizing hormone," PLoS One, 67, e21730, 2011.

[8] O. Friard, A. Re, D. Taverna, M. De Bortoli, and D. Cora, "CircuitDB: a database of mixed microRNA/transcription factor feed-forward regulatory circuits in human and mouse," BMC Bioinf., 11, 435, 2010.

[9] R.C. Friedman, K.K. Farh, C.B Burge, and DP. Bartel, "Most mammalian mRNAs are conserved targets of microRNAs," Genome Res, 19, 2009, pp. 1-11.

[10] S-D. Hsu, F-M. Lin, W-Y. Wu, C. Liang, W-C. Huang, W-L. Chan, W-T. Tsai, G-Z. Chen, C-J. Lee, C-M. Chiu, and et al., "miRTarBase: a database curates experimentally validated microRNA–target interactions," Nucleic Acids Res., 39, 2011, pp. D163-D169.

[11] A. Kundaje, M. Middendorf, F. Gao, C. Wiggins and C. Leslie, "Combining Sequence and time series expression data to learn transcriptional modules," IEEE/ACM Transactions on Computational Biology and Bioi nformatics, 2, 2005, pp. 194-202.

[12] R.C. Lee, R.L. Feinbaum, and V. Ambros, "The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14," Cell, 75, 1993, pp. 843-854.

[13] J. Lu, G. Getz, E.A. Miska, E. Alvarez-Saavedra, J. Lamb, D. Peck, A. Sweet-Cordero, B.L. Ebert, R.H. Mak, A.A. Ferrando and et al. "MicroRNA expression profiles classify human cancers," Nature, 435, 2005, pp. 834–838.

[14] S.F. Madden, S.B. Carpenter, I.B. Jeffery, H. Bjorkbacka, K.A. Fitzgerald, L.A. O'Neill, and D.G. Higgins, "Detecting microRNA activity from gene expression data," BMC Bioinf. 11, 257, 2010.

[15] H. Oğul and E. Mumcuoğlu, "A Discriminative Method for Remote Homology Detection Based on n-peptide Compositions with Reduced Amino Acid Alphabets," Biosystems, 87, 2007, pp. 75-81.

[16] G.L. Papadopoulos, M. Reczko, V.A. Simossis, P. Sethupaty, and A.G. Hatzigeorgion, "The database of experimentally supported targets: A functional update of TarBase," Nucleic Acids Res., 37, 2009, pp. D155-D158.

[17] X. Peng, Y. Li, K.A. Walters, E.R. Rosenzweig, S.L. Lederer, L.D. Aicher, S. Proll and M.G. Katze, "Computational identification of hepatitis c virus associated microRNA-mRNA regulatory modules in human livers," BMC Genomics, 10, 373, 2009.

[18] P.M. Voorhoeve, "MicroRNAs: Oncogenes, tumor suppressors or master regulators of cancer heterogeneity," Bioc Bioph Acta, 1805, 2010, pp. 72-86.

[19] F. Xiao, Z. Zuo, G. Cai, S. Kang, X. Gao, and T. Li, "miRecords: an integrated resource for microRNA-target interactions," Nucleic Acids Res., 37, 2009, pp. D105-D110.

Real-time Visualization of Protein Empty Space with Varying Parameters

Ondřej Strnad, Vilém Šustr, Barbora Kozlíková and Jiří Sochor
Faculty of Informatics
Masaryk university, Brno, Czech Republic
Email: xstrnad2@fi.muni.cz, xsustr@fi.muni.cz, xkozlik@fi.muni.cz, sochor@fi.muni.cz

***Abstract*—Exploration of the empty space inside protein structures is playing a crucial role in protein engineering and drug design. This empty space inside proteins can be utilized for the design of protein mutations. The importance of this empty space is also based on its ability to accept a small ligand molecule which can react with the protein. The product of such a reaction can form the basis of new medications. Many algorithms enabling computation of these empty spaces, often marked as voids, have been published and their results were evaluated by protein engineers to confirm their chemical relevance. However, not all voids of a protein can be considered as a target point of ligand binding. Thus, the following examination and assessment of all voids must be performed. In this phase the visual representation of voids is very valuable and substantially decreases time of this evaluation phase.**

In this paper, we introduce a novel algorithm for the visualization and further evaluation of these voids in real-time. This user-driven approach enables to compute and display empty space that satisfies the input parameters instantly. Basically, these parameters include setting of minimal desired width of the voids. The values of these parameters can be changed by the user anytime and the changes are immediately displayed and prepared for further exploration.

***Keywords*-protein, empty space, void, visualization, real-time, cavity**

I. Introduction

Long-term research in the area of protein analysis proved the importance of an empty space situated inside the macromolecular structure. This empty space can be further qualified according to various criteria and marked as a cavity, pocket, tunnel, channel, pore or other specific structure (see Fig. 1) . Inner cavities can serve as the destination for a small ligand molecule that can follow a pathway from the outside environment of the protein. The specific cavity accepting ligands is marked as an active site and in such cavity the chemical reaction between protein and ligand can take place. Products of such a reaction can serve as the basis for new drugs or various chemical compounds. A pocket can be defined as a hollow space on a protein's surface. This means that if the ligand is small enough the structure can be reached directly.

Channels and pores are specific pathways crossing the whole protein. They can be used for the transport of substrates, products, water molecules and other compounds through the protein. The distinction between channels and pores lies in their shape – a channel can have various curvatures whereas pores pass straight through the protein.

Our research in this field was concerned mainly with the detection of tunnels. These structures represent a path leading from a specific protein cavity (an active site) to the molecular surface. Thus for tunnel detection it is necessary to analyze and evaluate protein cavities to detect the active site in advance. Small substrate molecules entering the active site determine the minimal properties of computed tunnels, such as their width or curvature. Tunnel analysis and their further evaluation enhances the workflow of protein engineers or drug designers.

The derivation of empty space from the 3D structure of a protein's amino-acid sequence introduces a very complex task. Proteins with previously detected positioning in 3D space are stored in the well-known PDB database. This archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. Some of the PDB structures were thoroughly analyzed and active sites which were discovered were subsequently stored in the CSA (Catalytic Site Atlas) database. However, active sites of most of the structures still have not been revealed or published. This situation creates the necessity of using other semi-automated tools or even manually-detecting of the active site.

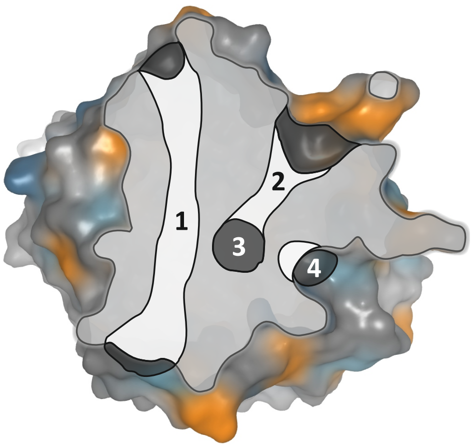


Figure 1. Illustration of a channel (1), a tunnel (2) leading from the active site (3) and a cavity (4).

ISBN: 978-1-61208-260-8

As mentioned above, the active site is placed inside a protein cavity. The first enhancement leads to the detection of all cavities inside the molecule. However, when operating with large protein complexes or even ribosomes, computation of all cavities in such molecule is very time and memory consuming. We introduce a novel method for detection and visualization of inner cavities focusing on minimizing the memory and time requirements. This technique is designed to operate in real-time, enabling users to interactively change the inner and outer size of the spherical probe utilized for empty space detection.

Without proper visualization it is complicated for biochemists to evaluate the computed cavity. Powerful visualization tool enabling not only displaying of detected voids but also allowing real-time alternations of parameters of detected voids is crucial for biochemists to be able to select the proper active site. When combining this method with other techniques, such as determination of partial charges, users can promptly recognize the possibility of occurrence of the desired chemical reaction. More specifically, when the surroundings of the cavity has neutral or small partial charge, this cavity probably will not be considered as an active site.

II. Related work

Detection and classification of the empty space inside proteins has been in the scope of biochemists for the last decades, and number of algorithms have been published in this field. Although the aim of the method described in this article is to, above all, detect and visualize cavities, in this section we will introduce related research focusing generally on the detection of empty space. These techniques can be adapted for the computation of tunnels, channels etc.

Algorithms detecting empty space inside protein macromolecules share similar principle; they are all based on computational geometry using protein geometry (positions and radii of atoms) as the input. These algorithms can be divided into two groups according to their approach to space representation. The first group is based on a grid approach while the second one utilizes a Voronoi diagram and Delaunay triangulation. The main difference lies in their precision, speed and memory consumption.

The outer environment can be considered a void as well, thus the protein has to be encapsulated in a bounding object. The void detection process inside protein structures is highly influenced by the protein surface, which gives an overview of the protein's compactness. In most cases, the empty space is detected only within the volume that is defined by the surface. The construction of the molecular surface will be presented in the second part of the related work.

A. Detection of empty space

1) Grid method: The entire protein is enclosed in an axis aligned bounding box subsequently sampled regularly to a voxel grid. Each vertex of the voxel grid is classified according to its collision with an atom. Non-colliding voxels form the empty space used for construction of cavities, tunnels and other structures. The quality of results is strongly influenced by the sampling density. Too sparse sampling can lead to a situation where all vertices of voxels are colliding with an atom and no empty space is detected. On the other hand, too dense a sampling causes an enormous increase in time and memory demands. The main advantage of this approach is its simplicity; the disadvantage, as already mentioned, follows from computational complexity $\mathcal{O}(n^3)$ with n depending on the sampling density.

The grid approach was adopted for tunnel computation in CAVER 1.0 [1]. Another tool using the grid approach for computation of specific cavities (pores) inside proteins is called CHUNNEL [2]. Each voxel is marked according to its distance to the nearest atom. Onto this structure the Dijkstra algorithm is used and the tunnel with highest voxel values (the widest tunnel) is detected.

Kleywegt et al. [3] presented their grid approach applied to the detection of cavities. Their implementation is presented in the VOIDOO application. The first step of the algorithm maps the protein onto a 3D grid with a spacing between 0.5 and 1.0 Ångströms. Each point of the grid is noted by the zero value. Then, each grid point is processed and when the distance to the nearest atom is less than the sum of the atomic radius and probe radius, its value is set to one. This method is also known as the flood-fill algorithm. Finally, the points inside cavities still have a zero value, so they can be easily detected and their volume can be measured.

2) Voronoi diagram and Delaunay triangulation: Another approach to protein 3D space inspection is based on the Voronoi diagram (VD) and its dual structure - the Delaunay triangulation (DT). The benefit of this approach is the division of the space without any dependency on user defined variables, which overcomes the main disadvantage of the previous grid approach. A detailed description of VD construction can be found in [4]. The dual structure to VD, the Delaunay triangulation (tetrahedrization in the 3D case), can be constructed by connecting neighboring points sharing the Voronoi edge (see Fig. 3). Tetrahedra of the Delaunay tetrahedrization fulfills the condition that no point is presented inside the circumsphere of any tetrahedra.

Voronoi diagrams and Delaunay tetrahedra were utilized by various software tools for tunnel and channel computation, such as CAVER 2.0 [5], MolAxis [6] or MOLE [7].

Another approach to cavity detection using Delaunay triangulation and the alpha complex was implemented in the CAST application [8] (CASTp is its online version). It is also able to analytically measure the area and volume of cavities as well.

In [9], Voronoi diagrams were extended to the Additively weighted Voronoi diagrams (AVD). AVDs were originally designed for environments containing non-uniform objects. They can be used to geometrically analyze protein structures consisting of many atoms with different radii. Compared to traditional VDs, AVDs gain the more adequate space subdivision through the specification of weight w attached to each site point. According to their weight values the respective points attract ($w > 1$) or repel ($w < 1$) the corresponding Voronoi edges. Resulting Voronoi edges have curvilinear shapes. AVD construction is more complex in comparison to traditional VD and thus the time and space complexity increases substantially. AVD were used in the protein visualization tool called Voroprot [10].

B. Protein surface

Detection and visualization of surfaces play an important role not only in the case of detection of voids in proteins but in many other fields as well. Thus, surface detection has been in researchers' scope for decades, and many approaches have been proposed. Two main groups of existing algorithms employ either analytical or numerical approach.

1) Analytical surface construction: The input set contains objects that should be encapsulated by the surface. The analytical approach describes the surface using a set of mathematical equations. For protein exploration there are two basic analytical approaches to generation of surfaces. The reduced surface [11] is constructed by rolling a probe of specific radius over the protein outer boundary. Inwards facing parts of the probe surface combined with parts of atoms' surfaces on the boundary create the resulting solvent-accessible surface. The second approach is based on the alpha-shapes theory [12]. The main disadvantage of the analytical representation of the surface comes from its complexity. Thus its utilization on large datasets (e.g. macromolecular structures) cannot be performed in real-time or can even fail.

2) Numerical surface construction: The accuracy of numerically based algorithms is strongly dependent on initial user settings. The basic principle is the division of the scrutinized space into a uniform voxel grid. Each voxel is classified according to its intersection with objects in space. Subsequently, the marching cubes algorithm [13] can be utilized for the detection of the surface. The marching cubes method was designed primarily for a simple and fast construction of iso-surfaces in volume data sets. This approach is widely used e.g. in MRI or other medical applications.

3) Visualization: The combination of surface detection with its visualization is not only crucial for the exploration of protein shapes. Protein surfaces play important role in many chemical simulations and many methods for the visualization of proteins geometry and their chemical properties have been designed. They include both analytical or numerical representations, such as [14] or the LSMS algorithm [15].

```
Algorithm Real-time visualization of protein voids
Require: set of atoms
1: compute Delaunay triangulation
2: convert it to a graph
3: while user is changing parameters do
4:   determine center point of the bounding box
5:   select empty space inside the protein
6:   visualize selected empty space
7: end while
```

Figure 2. Overview of the algorithm.

III. Real-time visualization of protein voids

In comparison with existing algorithms for highlighting empty voids, our approach does not require additional time for their re-computation when the input parameters change. The empty space corresponding to changed parameters is visualized immediately. In the rest of this section, the basic principle of our novel algorithm will be described. As noted above, the main aim of our approach is the real-time visualization of inner voids. Firstly, such voids must be computed. We utilize the standard Voronoi diagram, which omits the differences in radii of atoms (contrary to AVD approach) since our priority is the speed of the algorithm. From our experience, VD does not provide users with as precise results as AVD does, but the difference is not crucial for our purposes. The main difference between the results obtained by VD and AVD is in the exact representation of the surface of voids. However, the set of detected cavities is equal.

In the visualization phase, the algorithm is divided into five basic steps (see Fig. 2). Steps on lines 1 and 2 represent preprocessing and they are performed only once during the initialization phase. Steps on lines 4 to 6 (represented by subsections C to E) are iteratively repeated for any change of input parameters and are considered as the main contribution of this paper.

A. Construction of Delaunay triangulation

input: set of atoms A
output: Delaunay triangulation T

The input set A consists of all atoms of protein. Since we do not take into account the different radii, atomic centers were selected as representatives of atoms. The atomic centers then form the input set of points marked as P, which is subsequently processed. For the set P, the Delaunay triangulation T is constructed using the QuickHull 4D algorithm described in [16].

The triangulation T is afterward refined so that all tetrahedra intersecting the molecular surface of the protein are removed. In other words, all surface tetrahedra that are accessible from the outside by a probe with radius 2.8Å(double the van der Waals radius of oxygen) are removed from T

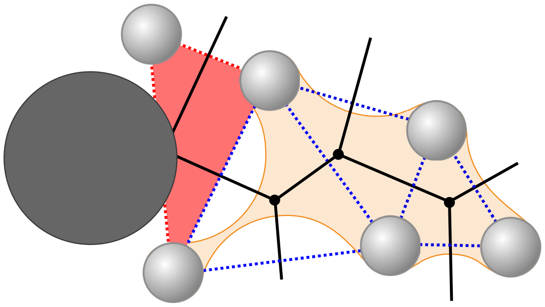


Figure 3. Tetrahedron (red) accessible by a probe (dark gray) is removed from the triangulation (blue dotted). The molecular surface defined by the probe is highlighted (orange).

(see Fig. 3). This ensures that the tunnel throat will not contain excessive boundary spheres.

B. Construction of the graph G

input: Delaunay triangulation T
output: evaluated graph G

For each tetrahedron $t_i \in T$ a node N_i is inserted into a newly constructed graph G. An edge e_{jk} connecting nodes N_j and N_k is added into G if their referenced tetrahedra t_j and t_k share a face f_{jk}. For every edge $e_{jk} \in G$, we define its center point $c(e_{jk})$ and width $w(e_{jk})$ as follows. The center point $c(e_{jk})$ is defined as a point in f_{jk} where sphere with maximal possible radius not intersecting any atom from t_j or t_k can be placed. The width $w(e_{jk})$ is then defined by the radius of such a sphere. The evaluation process is illustrated in Fig. 4.

C. Selection of center point

input: Delaunay triangulation T
output: center point $\mathbf{C}$

The algorithm was designed to operate with large macromolecules. In this case, computation and visualization of all inner voids usually leads to complex and ambiguous results, which the biochemist cannot properly explore, thanks to the huge amount of visualized data. In order to avoid this situation, we allow computing inner voids from a starting point $\mathbf{C}$ that represents the center of the bounding sphere. The empty space is then visualized only inside this bounding sphere - the area of interest. The point $\mathbf{C}$ set by the user can be determined in two ways. The user can enter the space coordinates of the point directly. In most cases the binding site is loaded from the CSA (Catalytic Site Atlas) database [17]. Once the center point $\mathbf{C}$ is set, it can be stored for further iterations of the algorithm.

D. Selection of relevant edges

input: graph G, point $\mathbf{C}$, distance d, parameter w_{min}
output: set of filtered edges E

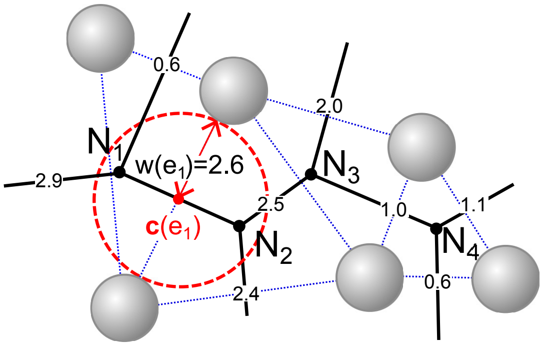


Figure 4. Illustration of a part of a graph G. Thick lines represent Voronoi edges. Every edge is evaluated by the value representing its distance to the nearest atom.

In this phase, the iteration process is started. The goal is to select a set E of edges which satisfy the condition of thickness (driven by the parameter w_{min} representing the minimal width of the edge) and proximity (parameter d defining the bounding sphere radius). For remark, every edge e_{jk} connecting two nodes N_j and N_k is evaluated by a width $w(e_{jk})$.

The set of filtered edges E consists of all edges having the $w(e_{jk})$ greater or equal to w_{min} and with the distance to $\mathbf{C}$ lower than d. More formally, let G_E is the set of all edges from G. The set of filtered edges is then $E = \{e_{jk} \in G_E | dist(\mathbf{C}, c(e_{jk})) < d \wedge w_{min} \leq w(e_{jk})\}$.

E. Visualization

input: set of edges E, selected visualization method(s)

Firstly, the set E has to be transformed into geometrical objects, which are possible to render. Every edge e_{jk} is transformed into a sphere s_{jk} with center in $c(e_{jk})$ and with radius equal to $w(e_{jk})$. The set S of all such spheres is then prepared as an input for selected visualization method(s). For our case of protein visualization, we utilized two basic methods effectively describing the empty space inside macromolecules.

1) Rendering of spheres: represents the most intuitive visualization method, and also the fastest one (see Fig. 5). It simply displays all spheres of the set S. From the construction introduced above, all spheres fill the empty space inside the molecule and do not intersect with any atom. Using this method, the empty space is highlighted, but it looks distracting and for user it can sometimes be difficult to distinguish between an atom of the molecule and a sphere highlighting the empty space.

2) Grid sampling: enables users to visualize a continual surface of voids, which gives more intuitive and user friendly results. To construct such a surface, all spheres from the set S are enclosed into an axis aligned bounding-box. This bounding-box is then regularly sampled with a user defined $density$. It is obvious that a higher density leads to a

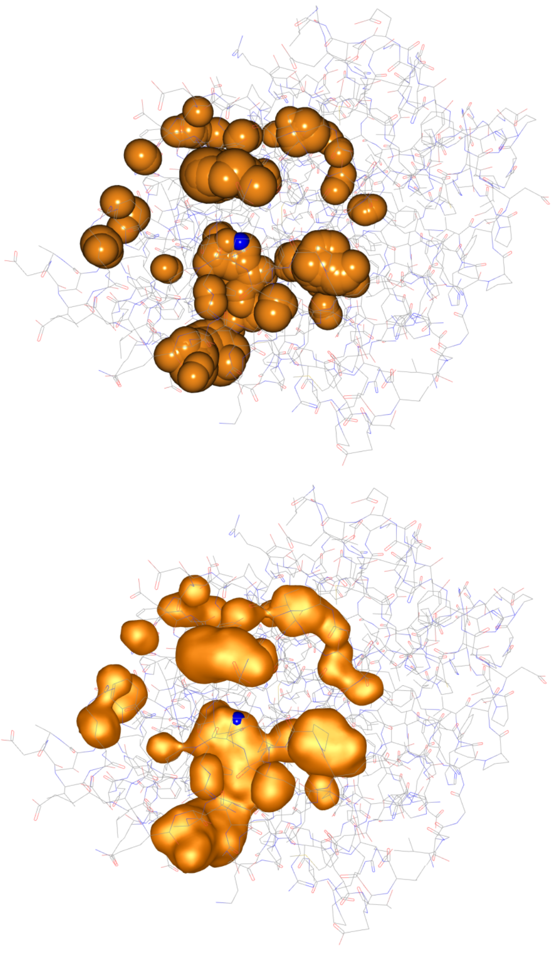


Figure 5. Empty space visualized as a set of spheres (top) or a surface (bottom).

more precise surface. On the other hand, the number of samples directly influences the memory and time complexity of the computation. We found out that for exploring of local neighborhood the empirically obtained $density = 200$ (i.e. grid 200x200x200) is optimal. Subsequently, every vertex of each cell in the grid is evaluated according to its intersection with any sphere from S. When all vertices are processed, the fully evaluated grid serves as the input for the marching cubes algorithm. For a notice, this algorithm operates with a predefined set of configurations, thus it its very straightforward and fast when constructing the resulting surface (see fig. 6).

IV. Results and discussion

In this paper we presented a novel method for real-time visualization of empty space inside macromolecules which concentrates on user-driven evaluation of computed voids. The method is not limited by the size of the molecule (the number of atoms) as the encapsulation of displayed voids into a bounding sphere allows to restrict the amount of processed data respectively. The implementation does not demand any special hardware or software, the algorithm was implemented in 32-bit Java environment. The performance was tested on a common single-threaded 2.66GHz computer. Both rendering strategies, as well as various types of macromolecules (ranging from proteins to ribosomes) underwent this test. For the sphere rendering strategy and for the surface strategy where only the closer neighborhood ($|E| < 10000$) is visualized, the algorithm operates in real-time. On larger structures, where there is a necessity to process huge amount of edges, the interaction is not fluent but still operable. The examples of tested combinations are summarized in table I.

Table I
Performace of the algorithm meassured in FPS (Frames Per Seconds). Parameter d represents a radius of a bounding sphere, $w(e)_{min}$ the minimal value of an edge.

d	$w(e)_{min}$	**1CQW (≈3k atoms)**		**1AON (≈60k atoms)**	
		spheres	surface	spheres	surface
15Å	1.4Å	>100	26	73	16
max	1.4Å	>100	19	28	2

For testing purposes haloalkane dehalogenase with approx. 3000k atoms (PDB ID 2HAD) and GroEL-GroES-(ADP)7 chaperonin complex with approx. 60k atoms (PDB ID 1AON) were chosen. To illustrate the real use of our algorithm, Fig. 5 visualizes the empty space in the 1CQW. The figure shows voids computed with user settings $w_{min} = 1.4\text{Å}$, $d = 15\text{Å}$ and **C**$= 25; 95; 35$.

To evidence the relevance of empty space detected and visualized using our approach, we performed a comparison with results obtained by the well acknowledged CAVER algorithm. CAVER was designed for the detection of tunnels inside proteins and the results were thoroughly tested by the community of protein engineers [18]. Thus, to manifest the relevance of voids detected by our new approach, the computed voids must contain all detected structures such as tunnels or cavities. We verified that tunnels detected by the CAVER algorithm lead through the empty space highlighted by our method (see Fig. 6). All visualized tunnels can be subsequently compared with protein empty spaces in their neighbourhood simply by changing few visualization parameters.

V. Future

The first extension of our implementation should lead to the parallelization of the marching cubes algorithm on the modern graphic cards [19]. Such implementation would substantially increase the performance of the rendering phase. We expect to be able to apply the surface method on large macromolecules in real-time.

Acknowledgment

This work was supported by The Grant Agency of the Czech Republic, Contract No. P202/10/1435.

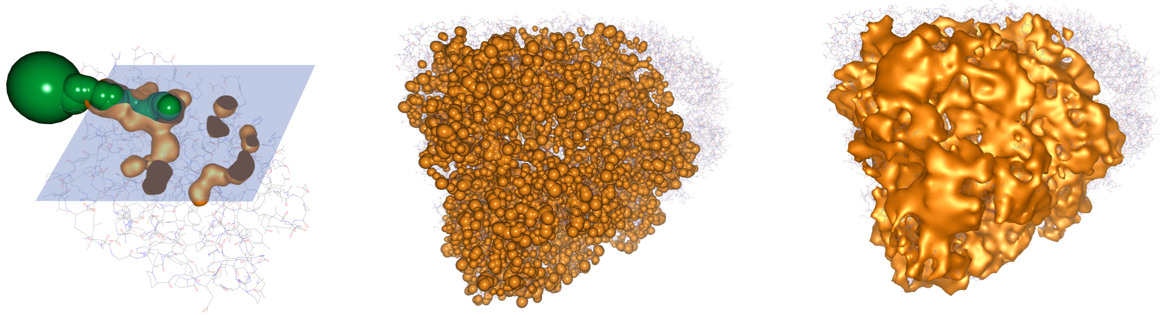


Figure 6. Left: Tunnel (green) detected by the CAVER algorithm lies inside the surface (bronze). The molecule is cut by a clipping plane (blue). Rendering of large protein structure, 1AON (approx. 60k atoms) by spheres (middle), surface rendering mode (right).

References

[1] M. Petřek, M. Otyepka, P. Banáš, P. Košinová, J. Koča, and J. Damborský, "Caver: A new tool to explore routes from protein clefts, pockets and cavities," *BMC Bioinformatics*, vol. 7, p. 316, 2006.

[2] R. G. Coleman and K. A. Sharp, "Finding and characterizing tunnels in macromolecules with application to ion channels and pores," *Biophysical Journal*, vol. 96, no. 2, pp. 632 – 645, 2009.

[3] G. J. Kleywegt and T. A. Jones, "Detection, delineation, measurement and display of cavities in macromolecular structures," *Acta Crystallographica Section D*, vol. 50, no. 2, pp. 178–185, Mar 1994. [Online]. Available: http://dx.doi.org/10.1107/S0907444993011333

[4] F. Aurenhammer, "Voronoi diagrams a survey of a fundamental geometric data structure," *ACM Comput. Surv.*, vol. 23, no. 3, pp. 345–405, sep 1991. [Online]. Available: http://doi.acm.org/10.1145/116873.116880

[5] P. Medek, P. Beneš, and J. Sochor, "Computation of tunnels in protein molecules using delaunay triangulation," *Journal of WSCG*, vol. 15(1-3), pp. 107–114, 2007.

[6] E. Yaffe, D. Fishelovitch, H. J. Wolfson, D. Halperin, and R. Nussinov, "MolAxis: efficient and accurate identification of channels in macromolecules." *Proteins*, vol. 73, no. 1, pp. 72–86, oct 2008. [Online]. Available: http://dx.doi.org/10.1002/prot.22052

[7] M. Petřek, P. Košinová, J. Koča, and M. Otyepka, "Mole: a voronoi diagram-based explorer of molecular channels, pores, and tunnels," *Structure*, vol. 15, no. 11, pp. 1357 – 1363, 2007.

[8] J. Liang, H. Edelsbrunner, and C. Woodward, "Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design," *Protein science : a publication of the Protein Society*, vol. 7, pp. 1884–1897, sep 1998.

[9] H. Edelsbrunner, *Algorithms in combinatorial geometry*. New York, NY, USA: Springer-Verlag New York, Inc., 1987.

[10] K. Olechnovič, M. Margelevičius, and v. Venclovas, "Voroprot," *Bioinformatics*, vol. 27, no. 5, pp. 723–724, mar 2011. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btq720

[11] M. F. Sanner, A. J. Olson, and J. C. Spehner, "Reduced surface: an efficient way to compute molecular surfaces," *Biopolymers*, vol. 38, pp. 305–320, Mar 1996.

[12] H. Edelsbrunner, D. G. Kirkpatrick, and R. Seidel, "On the shape of a set of points in the plane," *IEEE Transactions on Information Theory*, vol. 29, pp. 551–559, jul 1983.

[13] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *SIGGRAPH Comput. Graph.*, vol. 21, no. 4, pp. 163–169, Aug. 1987. [Online]. Available: http://doi.acm.org/10.1145/37402.37422

[14] M. Totrov and R. Abagyan, "The contour-buildup algorithm to calculate the analytical molecular surface," *J Struct Biol*, no. 1, pp. 138–43, 1995.

[15] T. Can, C. Chen, and Y. Wang, "Efficient molecular surface generation using level-set methods," *Journal of Molecular Graphics & Modelling*, vol. 25, pp. 442–454, 2006.

[16] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, dec 1996. [Online]. Available: http://doi.acm.org/10.1145/235815.235821

[17] C. T. Porter, G. J. Bartlett, and J. M. Thornton, "The catalytic site atlas: a resource of catalytic sites and residues identified in enzymes using structural data," *Nucleic Acids Research*, vol. 32, no. Database-Issue, pp. 129–133, 2004.

[18] E. Chovancová, A. Pavelka, P. Beneš, O. Strnad, J. Brezovský, B. Kozlíková, A. Gora, V. Šustr, M. Klvaňa, P. Medek, L. Biedermannová, J. Sochor, and J. Damborský, "Caver 3.0: A tool for the analysis of transport pathways in dynamic protein structures," *PLoS Comput Biol*, vol. 8, no. 10, p. e1002708, 2012. [Online]. Available: http://dx.doi.org/10.1371/journal.pcbi.1002708

[19] C. Dyken, G. Ziegler, C. Theobalt, and H.-P. Seidel, "High-speed marching cubes using histopyramids," *Computer Graphics Forum*, vol. 27, no. 8, pp. 2028–2039, 2008. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2008.01182.x

Applied Multi-Expert Decision Making Issue Based on Linguistic Models for Prostate Cancer Patients

H. Zettervall, E. Rakus-Andersson
Department of Mathematics and Science
Blekinge Institute of Technology
Karlskrona, Sweden
e-mail: Hang.Zettervall@bth.se
e-mail: Elisabeth.Andersson@bth.se

Janusz Frey
Department of Surgery and Urology
Blekinge County Hospital
Karlskrona, Sweden
e-mail: janusz.frey@ltblekinge.se

***Abstract*—In this paper, two models, one is called the probabilistic model and the other is known as the model of 2-tuple fuzzy linguistic representations, are applied to solve multi-expert decision making issues (MEDM). A MEDM problem is considered, in which a group of physicians are independently asked about assessing the effectiveness of a set of treatment therapies for a prostate cancer patient. The objective of this paper is to find the most common judgment by means of these two models. Moreover, fuzzy linguistic terms are used to express the experts' opinions and *s*-parametric membership functions are designed to depict the fuzzy linguistic terms.**

Keywords-multi-expert decision making; group decision making; fuzzy group decision making; linguistic modeling; linguistic choice function; 2-tuple fuzzy linguistic representation model; computing with words (CW).

I. INTRODUCTION

Multidisciplinary team conferences or multidisciplinary cancer conferences play a very important role in decision-making process in modern treatment of cancer patients. In the Urology Department of Blekinge County Hospital, Karlskrona, the Multidisciplinary Team Conference (MDT) is a forum of health care providers including medical oncologists, urologists, urology sub-specialized nurses, radiologists and pathologists. The aim of the conference is to establish assessments and treatment decisions for particular patients with a spectrum of problematic urological conditions that cannot be easily solved by the means of available resources. Our long term aim is also to discuss the best and available treatment modalities of all newly diagnosed cases of prostate cancer. Quite often the decision making process is very clear and straight forward, but some cases lay outside the frames of guidelines and recommendations. Obviously the final choice of a treatment is also on discretion of the patient. This approach has two pitfalls. One of them is when there is a discrepancy between forum members and the other one is when the patient is not interested in the treatment modality chosen by the panel. The best solution is to obtain a method for solving discrepancy and simultaneously to find a method that shows the panel's results as a treatment recommendation grade range between strongly recommended and contraindicated. Such approach should be very helpful in such diseases as a prostate cancer, which has a broad spectrum of treatment methods that can be tailored to the particular patient's needs and requirements.

In real life, we often are in such situations that we need to evaluate some information by means of numerical values. But when the numerical values are no longer available, then the linguistic approach [1] can be seen as a good alternative. Especially, in medical community, the information often is characterized vaguely and imprecisely, which makes it hard to be evaluated by numerical values. For example, the expressions such as "very painful", "slightly painful", "medium" and "not very painful" are just some examples of the linguistic evaluations of ache degrees of postoperative pain. Also in group decision making cases, when the experts assess the effectiveness of treatment therapies for prostate cancer patients, the semantic terms such as "contraindicated", "doubtful", "acceptable", "possible", "suitable", "recommended" and "strongly recommended" can be used. Comparing to the numerical quantity, the linguistic approach is regarded by [2-3] as a more realistic, intuitionistic and natural method. Due to the advantages of the linguistic approach, an extensive application has been presented in the references [4-6].

By applying two models, namely, the probabilistic model and the 2-tuple fuzzy linguistic representations [7] we wish to select the most consensual treatment therapy for a prostate cancer patient in a multi-expert decision making (MEDM) problem. Thus, the entire process will be defined in the linguistic framework.

The construction of this paper is organized as follows. In Section II the preliminaries are presented. In Section III a practical study about how these methods are applied for a prostate cancer patient is provided and the results are presented. Finally, conclusions and discussion are given in Section IV and V, respectively.

II. PRELIMINARIES

In this section, some preliminary items are presented. We start with the detailed description of the probabilistic model.

In reference [7], a general property of a MEDM problem is considered as the introduction of a finite set of experts denoted by $E = \{e_1, \cdots, e_p\}$ who are asked for selecting assessments stated in another finite set of alternatives $A = \{a_1, \cdots, a_n\}$. The assessments are expressed by semantic words in an order structured linguistic term set $S = \{s_0, \cdots, s_g\}$, such that $s_k < s_l$ if and only if $k < l$. An example of the ordered structured linguistic term set S is given below.

Example 1: Suppose that we determine a linguistic term set $S = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6\}$ consisting of s_0 = "contraindicated" = CI, s_1="doubtful"= D, s_2= "acceptable" = A, s_3= "possible" = P, s_4= "suitable" = S, s_5= "recommended" = R and s_6= "strongly recommended" = SR.

A. The Probabilistic Model

According to [7], the probability model mainly contains four steps:

- In the first step all the assessments are collected in a judgment table as shown in TABLE I. Here each judgment $L_{ij}, i = 1, \cdots, n$ and $j = 1, \cdots, p$ is expressed by the linguistic term selected from the linguistic term set S.
  We should emphasize that each linguistic term is associated with a general s-parametric membership function [8-10] given by

$$\mu_{s_l}(z) = \begin{cases} 2\left(\frac{z-((z_{\min}-h_z)+h_z l)}{h_z}\right)^2 \text{ for} \\ (z_{\min}-h_z)+h_z l \le z \le \left(z_{\min}-\frac{h_z}{2}\right)+h_z l, \\ 1-2\left(\frac{z-(z_{\min}+h_z l)}{h_z}\right)^2 \text{ for} \\ \left(z_{\min}-\frac{h_z}{2}\right)+h_z l \le z \le z_{\min}+h_z, \\ 1-2\left(\frac{z-(z_{\min}+h_z l)}{h_z}\right)^2 \text{ for} \\ z_{\min}+h_z l \le z \le \left(z_{\min}+\frac{h_z}{2}\right)+h_z l, \\ 2\left(\frac{z-((z_{\min}+h_z)+h_z l)}{h_z}\right)^2 \text{ for} \\ \left(z_{\min}+\frac{h_z}{2}\right)+h_z l \le z \le (z_{\min}+h_z)+h_z l\,, \end{cases} \quad (1)$$

  where $z \in [0,1]$ is a symbolic reference set of effectiveness, $z_{\min} = 0$, and h_z is defined as the distance of the peaks between two adjacent fuzzy sets. If we set $z_{\min}$ and h_z as fixed values when choosing $l = 0, \cdots, g$, then we will obtain the membership functions for $s_0, \cdots, s_g$.

TABLE I. THE JUDGMENT TABLE

Alternatives	Experts		
	e_1	⋯	e_p
a_1	L_{11}	⋯	L_{1p}
a_2	L_{21}	⋯	L_{2p}
⋯	⋯	⋯	⋯
a_n	$L_{\text{n}1}$	⋯	L_{np}

- X_{a_i} is assumed as a random preference value for each alternative $a_i, i = 1, \cdots, n$, with associated probablility distribution P defined by [11] as

$$P(X_{a_i} = s_l) = P_E(\{e_j \in E | z_{ij} = s_l\}). \quad (2)$$

  It is worth highlighting that the statement of random preference X_{a_i} is a crucial procedure in the approach of probability. Since each X_{a_i} is stochastically independent of each other, it will make the comparisons of any two random preferences to be possible.

- The choice value $V(a_i)$ for each alternative $a_i, i = 1, \cdots, n$, is computed by the choice function implemented by

$$V(a_i) = \sum_{i \neq j} P\left(X_{a_i} \ge X_{a_j}\right)$$
$$= \sum_{i \neq j} \sum_{s_l \in S} \left[P(X_{a_i} = s_l) \sum_{\substack{z_{ij} \in S \\ s_l \ge z_{ij}}} P\left(X_{a_j} = z_{ij}\right) \right], \quad (3)$$

  where the quantity $P\left(X_{a_i} \ge X_{a_j}\right)$ could be interpreted as the probability of "the performance of a_i is as least as good as that of a_j".

- Finally, by ranking the choice values obtained by the former step, we can select the optimal one by (4)

$$a_{\text{optimal}} = \max_{a_i \in A}(V(a_i)). \quad (4)$$

B. The Model of 2-tuple Fuzzy Linguistic Representation

In this model, the physicians' judgments of the treatments are represented by 2-tuples of the form of (s_l, α), where $s_l \in S$ is a fuzzy semantic term and $\alpha \in [-0.5, 0.5)$ is defined as a numerical value.

A 2-tuple fuzzy linguistic representation model presented in [11] composes the following steps:

- Each judgment which is expressed by a fuzzy semantic word in TABLE I is changed into a 2-tuple fuzzy linguistic representation as (s_l, α). If $s_l \in S$, then $(s_l, 0)$ will reflect s_l. Next, $x_{a_i} = \{(s_l, \alpha)\}$ is defined as a finite set that consists of judgments of the 2-tuple fuzzy linguistic representations for each alternative $a_i, i = 1, \cdots, n$.

- Two transformations are used in this model.

The first transform Δ^1 maps a 2-tuple fuzzy representation (s_l, α), which belongs to the space of $S \times [-0.5, 0.5)$, into a numerical value $\beta \in [0, g]$. Here s_l has the closest index label to β, and $[0, g]$ represents the interval consisting of the semantic label indices in the linguistic term set $S = \{s_l\}$, $l = 0, \cdots, g$. The action of Δ^1 is formalized by

$$\Delta^1: \quad S \times [-0.5, 0.5) \rightarrow [0, g]$$
$$(s_l, \alpha) \rightarrow \beta = l + \alpha.$$

Example 2: Let $S = \{s_0, \cdots, s_6\}$ and $\beta \in [0,6]$. In TABLE II the assessment of a_1, given by expert e_3, is expressed by the fuzzy semantic term s_2 = "acceptable" =A. By the 2-tuple fuzzy representation we can employ the judgment (A, 0) for s_2 = "acceptable" =A and $\alpha = 0$.

Due to the first transformation, the 2-tuple fuzzy representation of (A, 0) can be performed as a numerical value $\beta = l + \alpha = 2 + 0 = 2$, which belongs to the interval [0, 6].

TABLE II. THE DECISION TABLE OF THE JUDGMENTS

Alternatives	Experts			
	$\boldsymbol{e_1}$	$\boldsymbol{e_2}$	$\boldsymbol{e_3}$	$\boldsymbol{e_4}$
a_1	s_0	s_1	s_2	s_3
a_2	s_2	s_0	s_1	s_4
a_3	s_3	s_4	s_5	s_1
a_4	s_2	s_1	s_2	s_0

The second transformation Δ^2 can be regarded as an inverse of the first one, i.e., it maps a numerical value $\beta \in [0, g]$ into a 2-tuple (s_l, α) by

$$\Delta^2: \quad [0, g] \rightarrow S \times [-0.5, 0.5)$$
$$\beta \rightarrow (s_l, \alpha).$$

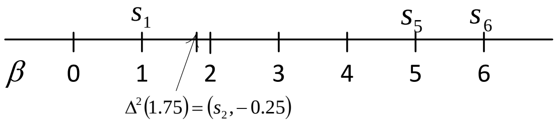


Figure 1. The 2-tuple fuzzy linguistic representation of $\beta = 1.75$

Example 3: Let $S = \{s_0, \cdots, s_6\}$ and $\beta \in [0,6]$. Suppose that $\beta = 1.75 \in [0, 6]$. Since 1.75 is closer to s_2 than to s_1, then we choose s_2 as the semantic word. The difference between 1.75 and 2 is 0.25, and 1.75 lies to the left of 2. Therefore, we choose –0.25 to be the value of α. By means of the second transformation, $\Delta^2(1.75) = (s_2, -0.25)$, which is depicted in Fig 1.

- The third step contains the computation of the arithmetic mean of $\bar{x}^e_{a_i}$ of 2-tuples for each alternative $a_i, i = 1, \cdots, n$. This is based on transformations Δ^1 and Δ^2 involved in

$$\bar{x}^e_{a_i} = \Delta^2\left(\sum_{l=0}^{g} \frac{1}{n} \Delta^1(s_l, \alpha)\right). \tag{5}$$

Since the arithmetic means, supplied from the previous step, are presented by 2-tuples, a computational technique to compare the arithmetic mean for each alternative proposed in [11] is given as follows.

- Let (s_k, α_1) and (s_l, α_2) be two 2-tuples fuzzy linguistic representations, with each one representing a counting of information as follows:

  1) if $k < l$, then (s_k, α_1) is smaller than (s_l, α_2).

  2) if $k = l$, then
  if $\alpha_1 = \alpha_2$, then (s_k, α_1) and (s_l, α_2) represents the same information.
  if $\alpha_1 < \alpha_2$, then (s_k, α_1) is smaller than (s_l, α_2).
  if $\alpha_1 > \alpha_2$, then (s_k, α_1) is greater than (s_l, α_2).

- At last, by comparing the arithmetic values with each other and ranking the alternatives, the optimal alternative(s) will be obtained.

III. A PRACTICAL STUDY

In this section we want to present a practical study in medical group decision making task. The members of a physician group are asked for providing the opinions on some treatment schemes for a prostate cancer patient. The methods of probabilistic model and the 2-tuple fuzzy linguistic model are applied and the results are presented.

A. The Probabilistic Model

Let us suppose that $E = \{e_1, e_2, e_3, e_4\}$ denotes a collection consisting of four physicians. And another set

$A = \{a_1, a_2, a_3, a_4, a_5, a_6\}$ contains sex types of treatment schemes for a prostate cancer patient, where a_1= “wait and see”, a_2= “active monitoring”, a_3= “symptom based treatment”, a_4 = “brachytherapy”, a_5 = “external beam radiation therapy” and a_6= “radical prostatectomy”. Also, $L = \{s_0, s_1, s_2, s_3, s_4, s_5, s_6\}$ includes seven linguistic terms, in which s_0= “contraindicated”, s_1= “doubtful”, s_2= “acceptable”, s_3= “possible”, s_4 = “suitable”, s_5= “recommended” and s_6= “strongly recommended”.

By inserting $z_{\min} = 0$, $h_z = 0.167$ and l = 0 in (1), we obtain the function for s_0 = “contraindicated” expanded by

$$\mu_{s_0}(z) = \begin{cases} 2\left(\frac{z+0.167}{0.167}\right)^2 & \text{for} \quad -0.167 \le z \le -0.0835, \\ 1-2\left(\frac{z}{0.167}\right)^2 & \text{for} \quad -0.0835 \le x \le 0, \\ 1-2\left(\frac{z}{0.167}\right)^2 & \text{for} \quad 0 \le z \le 0.0835, \\ 2\left(\frac{z-0.167}{0.167}\right)^2 & \text{for} \quad 0.0835 \le x \le 0.167. \end{cases} \tag{6}$$

By following the same procedure for l =1, 2, 3, 4, 5 and 6 we generate membership functions

$$\mu_{s_1}(z) = \begin{cases} 2\left(\frac{z}{0.167}\right)^2 & \text{for} \quad 0 \le z \le 0.0835, \\ 1-2\left(\frac{z-0.167}{0.167}\right)^2 & \text{for} \quad 0.0835 \le z \le 0.167, \\ 1-2\left(\frac{z-0.167}{0.167}\right)^2 & \text{for} \quad 0.167 \le z \le 0.2505, \\ 2\left(\frac{z-0.334}{0.167}\right)^2 & \text{for} \quad 0.2505 \le z \le 0.334, \end{cases} \tag{7}$$

$$\mu_{s_2}(z) = \begin{cases} 2\left(\frac{z-0.167}{0.167}\right)^2 & \text{for} \quad 0.167 \le z \le 0.2505, \\ 1-2\left(\frac{z-0.334}{0.167}\right)^2 & \text{for} \quad 0.2505 \le z \le 0.334, \\ 1-2\left(\frac{z-0.334}{0.167}\right)^2 & \text{for} \quad 0.334 \le z \le 0.4175, \\ 2\left(\frac{z-0.501}{0.167}\right)^2 & \text{for} \quad 0.4175 \le z \le 0.501, \end{cases} \tag{8}$$

$$\mu_{s_3}(z) = \begin{cases} 2\left(\frac{z-0.334}{0.167}\right)^2 & \text{for} \quad 0.334 \le z \le 0.4175, \\ 1-2\left(\frac{z-0.501}{0.167}\right)^2 & \text{for} \quad 0.4175 \le z \le 0.501, \\ 1-2\left(\frac{z-0.501}{0.167}\right)^2 & \text{for} \quad 0.501 \le z \le 0.5845, \\ 2\left(\frac{z-0.668}{0.167}\right)^2 & \text{for} \quad 0.5845 \le z \le 0.668, \end{cases} \tag{9}$$

$$\mu_{s_4}(z) = \begin{cases} 2\left(\frac{z-0.501}{0.167}\right)^2 & \text{for} \quad 0.501 \le z \le 0.5845, \\ 1-2\left(\frac{z-0.668}{0.167}\right)^2 & \text{for} \quad 0.5845 \le z \le 0.668, \\ 1-2\left(\frac{z-0.668}{0.167}\right)^2 & \text{for} \quad 0.668 \le z \le 0.7515, \\ 2\left(\frac{z-0.835}{0.167}\right)^2 & \text{for} \quad 0.7515 \le z \le 0.835, \end{cases} \tag{10}$$

$$\mu_{s_5}(z) = \begin{cases} 2\left(\frac{z-0.668}{0.167}\right)^2 & \text{for} \quad 0.668 \le z \le 0.7515, \\ 1-2\left(\frac{z-0.835}{0.167}\right)^2 & \text{for} \quad 0.7515 \le z \le 0.835, \\ 1-2\left(\frac{z-0.835}{0.167}\right)^2 & \text{for} \quad 0.835 \le z \le 0.9185, \\ 2\left(\frac{z-1.002}{0.167}\right)^2 & \text{for} \quad 0.9185 \le z \le 1.002, \end{cases} \tag{11}$$

and

$$\mu_{s_6}(z) = \begin{cases} 2\left(\frac{z-0.835}{0.167}\right)^2 & \text{for} \quad 0.835 \le z \le 0.9185, \\ 1-2\left(\frac{z-1.002}{0.167}\right)^2 & \text{for} \quad 0.9185 \le z \le 0.1002, \\ 1-2\left(\frac{z-1.002}{0.167}\right)^2 & \text{for} \quad 1.002 \le z \le 1.0855, \\ 2\left(\frac{z-1.169}{0.167}\right)^2 & \text{for} \quad 1.0855 \le z \le 1.169. \end{cases} \tag{12}$$

We sample all functions (6)–(12) in a family of fuzzy set restrictions, which are plotted in Fig 2.

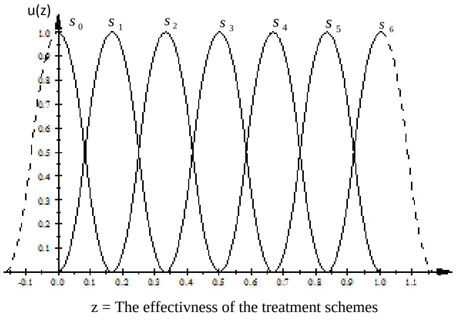


Figure 2. S-parametric membership functions for linguistic fuzzy sets

By using the probabilistic model, we collect all the experts’ judgments in TABLE III, whereas the random preference value of each judgment is given in TABLE IV.

By using (3), we calculate the choice value for a_1 as the following structure

TABLE III. THE COLLECTION OF THE JUDGMENTS

Alternatives	Experts			
	e_1	e_2	e_3	e_4
a_1	s_0	s_0	s_0	s_0
a_2	s_6	s_6	s_5	s_5
a_3	s_0	s_0	s_0	s_0
a_4	s_3	s_2	s_4	s_4
a_5	s_3	s_1	s_3	s_4
a_6	s_4	s_5	s_4	s_5

TABLE IV. THE AGGREGATION OF RANDOM PREFERENCE

	Random Preference						
	s_0	s_1	s_2	s_3	s_4	s_5	s_6
X_{a_1}	1	0	0	0	0	0	0
X_{a_2}	0	0	0	0	0	0.5	0.5
X_{a_3}	1	0	0	0	0	0	0
X_{a_4}	0	0	0.25	0.25	0.5	0	0
X_{a_5}	0	0.25	0	0.5	0.25	0	0
X_{a_6}	0	0	0	0	0.5	0.5	0

$$
\begin{aligned}
V(a_1) &= \sum_{1\neq j} P\left(X_{a_1} \geq X_{a_j}\right) \\
&= \sum_{1\neq j}\sum_{s_l\in S}[P\big(X_{a_1} = s_l\big) \sum_{\substack{z_{ij}\in S \\ s_l\geq z_{ij}}} P(X_{a_j} = z_{ij})] \\
&= P\big(X_{a_1} \geq X_{a_2}\big) + \cdots + P\big(X_{a_1} \geq X_{a_6}\big) \\
&= 0 + 1 + 0 + 0 + 0 = 1.
\end{aligned}
$$

For other a_i, i = 2,3,4,5,6, $V(a_i)$ are calculated in the similar way as

$V(a_2) \ = 1 + 1 + 1 + 1 + 1 = 5,$
$V(a_3) \ = 1 + 0 + 0 + 0 + 0 = 1,$
$V(a_4) = 1 + 0 + 1 + 0.75 + 0.25 = 3,$
$V(a_5) = 1 + 0 + 1 + 0.5 + 0.125 = 2.625,$
and
$V(a_6) \ = 1 + 0.25 + 1 + 1 + 1 = 4.25.$

The collection of choice values for each $a_i, i = 1, \cdots, 6$ is aggregated in Table V.

We choose the optimal therapy alternative by means of (4) as

TABLE V. THE COLLECTION OF CHOICE VALUES

The Collection of Choice Value for Each Alternative					
$\boldsymbol{V(a_1)}$	$\boldsymbol{V(a_2)}$	$\boldsymbol{V(a_3)}$	$\boldsymbol{V(a_4)}$	$\boldsymbol{V(a_5)}$	$\boldsymbol{V(a_6)}$
1	5	1	3	2.625	4.25

$$
\begin{aligned}
a_{\text{optimal}} &= \max_{a_i\in A}\{V(a_i)\} = \max\{1, 5, 1, 3, 2.625, 4.25\} \\
&= 5 = V(a_2).
\end{aligned}
$$

The value of 5 indicates the choice value of a_2 to be maximal. This means that the second therapy alternative is the most efficacious.

We want to confirm the result by applying the model of 2-tuple fuzzy linguistic representations.

B. The Model of 2-tuple Linguistic Representation

According to the algorithm for the model of 2-tuple fuzzy representation, the judgment which is transformed into 2-tuples is given in TABLE VI.

TABLE VI. THE JUDGMENTS EXPRESSED IN THE 2-TUPLES REPRESENTATION MODEL

Experts	Alternatives					
	a_1	a_2	a_3	a_4	a_5	a_6
e_1	(CI, 0)	(SR, 0)	(CI, 0)	(P, 0)	(P, 0)	(S, 0)
e_2	(CI, 0)	(SR, 0)	(CI, 0)	(A, 0)	(H, 0)	(R, 0)
e_3	(CI, 0)	(R, 0)	(CI, 0)	(S, 0)	(P, 0)	(S, 0)
e_4	(CI, 0)	(R, 0)	(CI, 0)	(S, 0)	(S, 0)	(R, 0)

We calculate the arithmetic mean for the first alternative a_1 by means of (5).

$x_{a_1} = \{(\text{CI}, 0), (\text{CI}, 0), (\text{CI}, 0), (\text{CI}, 0)\}$ is a finite set consisting of four 2-tuple linguistic representations for the alternative a_1. By adopting (5), the arithmetic means value for a_1 is calculated as:

$$\bar{x}_{a_1}^e = \Delta^2\left(\frac{1}{4}(0 + 0 + 0 + 0)\right) = \Delta^2(0) = (s_0, 0.5).$$

For the second alternative the arithmetic means value is given as follows:

$$\bar{x}_{a_2}^e = \Delta^2\left(\frac{1}{4}(6 + 6 + 5 + 5)\right) = \Delta^2(5.5) = (s_5, 0.5).$$

By the same reasoning, when setting i = 3,4,5,6 in (5), we implement

$$\bar{x}_{a_3}^e = \Delta^2\left(\tfrac{1}{4}(0 + 0 + 0 + 0)\right) = \Delta^2(0) = (s_0, 0),$$

$$\bar{x}_{a_4}^e = \Delta^2\left(\frac{1}{4}(3 + 2 + 4 + 1)\right) = \Delta^2(2.5) = (s_2, 0.5),$$

$$\bar{x}_{a_5}^e = \Delta^2\left(\frac{1}{4}(3 + 1 + 3 + 4)\right) = \Delta^2(2.75) = (s_3, -0.25),$$

and

$$\bar{x}^e_{a_6} = \Delta^2\left(\frac{1}{4}(4+5+4+5)\right) = \Delta^2(4.5) = (s_4, 0.5).$$

We present the collection of the arithmetic values for all alternatives in TABLE VII.

TABLE VII. TABLE OF THE ARITHMETIC VALUES

The Collection of the Arithmetic Values					
$\bar{x}^e_{a_1}$	$\bar{x}^e_{a_2}$	$\bar{x}^e_{a_3}$	$\bar{x}^e_{a_4}$	$\bar{x}^e_{a_5}$	$\bar{x}^e_{a_6}$
$(s_0, 0)$	$(s_5, 0.5)$	$(s_0, 0)$	$(s_2, 0.5)$	$(s_3, -0.25)$	$(s_4, 0.5)$

According to the computational technique presented earlier, we compare the above 2-tuples which represent the arithmetic values for all the alternatives. We obtain the result presented as $a_2 > a_6 > a_5 > a_4 > a_1 = a_3$, which shows that alternative a_2 is the most efficacious treatment scheme. This result converges to the previous result from "the probabilistic model".

IV. CONCLUSION

In this paper two models, like the probabilistic model and the model of 2-tuple fuzzy linguistic representations, have been applied in a MEDM problem to find the most consensual treatment scheme for a prostate cancer patient. The convergence results from both of the approaches verify the high reliability of adopting the linguistic approach in solving group decision making problems. Moreover, the independent assumed preferences of each alternative make the computation of comparing the probabilities easy to be performed. Especially, the use of the 2-tuple fuzzy linguistic representation model prevents the loss of information and makes the result more precise. At last but not at least, the use of *s*-parametric membership functions not only makes the fuzzy sets intuitionistic, but also increases the accuracy rate of the comparative analysis. Having discovered the hierarchy of therapies we also wish to utilize the formulas of membership functions to assign the group decision efficacy of each treatment to an expression from the list suggested. We treat this query as a challenge in our future research.

V. DISCUSSION

From the medical point of view, we found both methods very interesting in decision-making process when panelists were not unanimous. The results seem to be reasonable. The process of sampling the data by filling the questionnaires was easy and quickly accomplished. We hope to introduce one of the models in our clinical practice to assess the method in a real life conditions. Hopefully, this approach can allow us to find better treatment strategies and to give prostate cancer patients more flexibility concerning the treatment options. This should be a great complement to the current guidelines and scientific society recommendations.

ACKNOWLEDGMENT

The authors are very thankful for the support and suggestions provided by physicians from the Urology Department of Blekinge County Hospital in Karlskrona, Sweden.

REFERENCES

[1] L. A. Zadeh, "The concept of a linguistic variable and its applications to approximate reasoning," Inform. Sci., pt. I, no. 8, 1975, pp.199–249.

[2] M. Delgado, J. L. Verdegay, and M. A. Vila, "Linguistic decision making models," Int. J. Intell. Syst., vol. 7, 1992, pp. 479–492.

[3] F. Herrera and J. L. Verdegay, "Linguistic assessments in group decision," Proceedings of the 1st European Congress on Fuzzy and Intelligent Technologies, Aachen, 1993, pp. 941–948.

[4] S. J. Chen and C. L. Hwang, Fuzzy Multiple Attribute Decision Making –Methods and Applications, Springer, Berlin Heidelbeg, 1992.

[5] J. Kacprzyk and M. Fedrizzi, Multiperson Decision Making Models Using Fuzzy Sets and Possibility Theory, Kluwer Academic Publishers, Dordrecht, 1990.

[6] M. Tong and P.P. Bonissone, "A linguistic approach to decision making with fuzzy sets," IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-10, 1980, pp. 716–723.

[7] V. N. Huynh and Y. Nakamori, "Multi-expert decision-making with linguistic information: a probabilistic-based model," Proceedings of the 38th Hawaii International Coference on System Sciences, 2005.

[8] H. J. Zimmermann, Fuzzy Set Theory and Its Applications. Kluwer Academic Publishers, Dortrecht, 2001.

[9] E. Rakus-Andersson and L. Jain, "Computational intelligence in medical decisions making," Recent Advances in Decision Making, Springer, Berlin Heidelberg, 2009, pp. 145-159.

[10] E. Rakus-Andersson, H. Zettervall, and M. Erman, "Prioritization of weighted strategies in the multi-player games with fuzzy entries of the payoff matrix," Int. J. of General Systems, vol, 39, Issue 3, 2010, pp. 291-304.

[11] F. Herrera and L. Martinez, "A 2-tuple fuzzy linguistic representation model for computing with words," IEEE Trans. Fuzzy Syst., vol. TFS-8, 2000, pp. 746–752.