IARIA

BIOTECHNO 2014

The Sixth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-335-3

April 20 - 24, 2014

Chamonix, France

BIOTECHNO 2014 Editors

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

Pascal Lorenz, Université de Haute Alsace, France

BIOTECHNO 2014

Foreword

The Sixth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2014), held between April 20 - 24, 2014 in Chamonix, France, 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 2014 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 2014. 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 2014 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2014 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 Chamonix, France.

BIOTECHNO 2014 Chairs:

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

BIOTECHNO 2014

Committee

BIOTECHNO Advisory Chairs

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

BIOTECHNO Industrial/Research Chairs

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

BIOTECHNO 2014 Technical Program Committee

Basim Alhadidi, Albalqa' Applied University - Salt, Jordan
Hesham H. Ali, University of Nebraska at Omaha, USA
Stephen Anthony, The University of New South Wales, Australia
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada
Siegfried Benkner, University of Vienna, Austria
Gilles Bernot, University of Nice Sophia Antipolis, France
Tom Bersano, University of Michigan, USA
Christian Blum, IKERBASQUE, Basque Foundation for Science - Bilbao, Spain
Razvan Bocu, University of Brasov, Romania
Magnus Bordewich, Durham University, UK
Sabin-Corneliu Buraga, "A. I. Cuza" University - Iasi, Romania
Eduardo Campos dos Santos, Universidade Federal de Minas Gerais (UFMG), Brazil
Yang Cao, Virginia Tech – Blacksburg, USA
Cesar German Castellanos Dominguez, Universidad Nacional de Colombia - Manizales,Colombia
Yili Chen, Monsanto Company - St. Louis, USA
Rolf Drechsler, DFKI Bremen || University of Bremen, Germany
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
Dong-Chul Kim, University of Texas at Arlington, USA
Danny Krizanc, Wesleyan University, USA
Fatih Kurugollu, Queen's University - Belfast, UK
Cedric Lhoussaine, Université Lille 1, France
Yaohang Li, Old Dominion University, USA
Yueh-Jaw Lin, University of Texas at Tyler, USA
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
José Manuel Molina López, Universidad Carlos III de Madrid, Spain
Giancarlo Mauri, University of Milano-Bicocca, Italy
Chilukuri K. Mohan, Syracuse University, USA
Julián Molina, University of Malaga, Spain
Victor Palamodov, Tel Aviv University, Israel
Sever Pasca, Politehnica University of Bucharest, Romania
Maria Manuela Pereira de Sousa, University of Beira Interior, Portugal
Nadia Pisanti, University of Pisa, Italy || Leiden University, The Netherlands
Clara Pizzuti, ICAR-CNR - Rende (Cosenza), Italy
Enrico Pontelli, New Mexico State University, USA
Ravi Radhakrishnan, University of Pennsylvania, USA
Robert Reynolds, Wayne State University, USA
Mauricio Rodriguez Rodriguez, Centro de Bioinformatica y Biologia Computacional de Colombia - CBBC, Colombia
Luciano Sanchez, Universidad de Oviedo, Spain
Steffen Schober, Ulm University, Germany
Sylvain Sené, Aix-Marseille University, France
Avinash Shankaranarayanan, Aries Greenergie Enterprise (P), Ltd., India
Patrick Siarry, Université Paris 12 (LiSSi), France
Anne Siegel, CNRS - Rennes, France
Raj Singh, University of Houston, USA
Zdenek Smékal, Brno University of Technology, Czech Republic
Takehide Soh, Kobe University, Japan
Bin Song, Oracle - Redwood shores, USA
John Spouge, National Center for Biotechnology Information /National Library of Medicine - Bethesda, USA
Ondrej Strnad, Masaryk University, Czech Republic
Andrzej Swierniak, Silesian University of Technology, Poland
Sing-Hoi Sze, Texas A&M University, USA
Sophia Tsoka, King's College London, UK

Marcel Turcotte, University of Ottawa, Canada
Ugo Vaccaro, Universita` di Salerno, Italy
Chun Wu, Mount Marty College - Yankton, USA
Boting Yang, University of Regina, Canada
Wang Yu-Ping, Tulane University, USA
Alexander Zelikovsky, Georgia State University, USA
Erliang Zeng, University of Notre Dame, USA

Copyright Information

Table of Contents

Impact of Population Size and Selection within a Customized NSGA-II for Biochemical Optimization Assessed on the Basis of the Average Cuboid Volume Indicator

Susanne Rosenthal, Markus Borschbach
University of Applied Sciences, FHDW
Faculty of Computer Science, Chair of Optimized Systems,
Hauptstr. 2, D-51465 Bergisch Gladbach, Germany
Email: {susanne.rosenthal, markus.borschbach}@fhdw.de

***Abstract*—The key part in the area of peptide design is the prediction of the peptides' molecular features. The performance of the drug design process depends on the identification of peptides that optimize several molecular properties at the same time. The synthesis of peptides for laboratory characterization is very cost-intensive. Therefore, drug development is a wide field of activity for multi-objective Genetic Algorithms (moGAs). A customized NSGA-II has been especially evolved for biochemical optimization with the focus on producing a great number of very different high quality peptides within a very low number of generations (under 20), termed early convergence. The main task of this paper is to verify empirically the effect of early convergence for this customized NSGA-II within a limited range of population size. Furthermore, an insight into the impact of the interdependence between the population size and the selection procedure is examined with the objective of giving a configuration rule for the selection parameter and the population size exemplary determined for a three-dimensional biochemical minimization problem. Although, this optimization problem is as generic as possible. The performance is assessed on the basis of a convergence indicator especially evolved for our preference of comparing the convergence behavior of populations with different sizes. Moreover, we propose a summarization of open source Java tools that are discussed regarding the potential of an easy implementation of the customized NSGA-II for biochemical optimization.**

***Index Terms*—multi-objective biochemical optimization; population size; average cuboid volume; open source Java tools.**

I. INTRODUCTION

Small peptides are of special interest in the area of drug design as they have some favorable features like conformational restriction, membrane permeability, metabolic stability and oral bioavailability [1]. Nevertheless, for this purpose these peptides have to optimize several molecular features at the same time. As both the synthesis and the laboratory characterization of peptides is very cost-intensive [23], moGAs provide an economical and robust method for peptide identification.

For this purpose, a customized Non-dominated Sorting Genetic Algorithm (NSGA-II) has been evolved and introduced in [2] with a considerable low number of generations and population size, termed early convergence. The NSGA-II is customized w.r.t. the encoding and the components mutation, recombination and selection. Different mutation and recombination methods have been evolved for this purpose and are introduced in [3][4]. These components and their parameter are not only inter related, but are also responsible for the performance of a GA. So far, less work has been done to gain an insight in the influence of the population size on the performance and in the interdependence with the selection operator and its parameters in the case of moGAs. The population size is an important topic in influencing the performance of evolutionary algorithms [5]. Small population sizes tend to result in poor convergence and large populations extend the computational complexity of a GA in finding high quality solutions [6]. Therefore, an adequate population size that results in good performance is challenging. Diverse results have been presented w.r.t. the choice and the handling of the populations size for single-objective GA: Yu et. al [7] study the connection between selection pressure and population size and ratify the concept of interdependence of parameters and operators in GA. The concept of self-adaption is used to overcome the problem of determining the optimal population size. Two forms of self-adaption are used: First, Bäck et al. [8] uses self-adaption as a previous setup and configuration step for evolutionary strategies. The population size then remains the same over all iterations. Second, Arabas et al. [9] introduces a GA with varying population size. The self-adaption of the population size is used throughout the whole GA run and depends among others on different parameters like the reproduction ratio. Eiben et al. [10] provide empirical studies that self-adaption of selection pressure and population size is possible and further rewarding w.r.t. algorithm performance. In this case study, the global parameters tournament size and population size are regulated.

The questions that we consider in this paper are: 1. Do large populations speed up the convergence behavior of the customized NSGA-II for a three-dimensional biochemical minimization problem? 2. Is there a predictable impact between population size and selection? 3. Is there a range of population size which is able to perform well?

These questions are answered in an empirical way: The performance of the customized NSGA-II is assessed w.r.t. its early convergence and a high diversity within the solutions. Some metrics have been proposed to evaluate the convergence behavior of a moGA [24]. These metrics, generally, measure the distance of non-dominated solution sets to the true Pareto front [24]. This makes a comparison of generations with different sizes impossible. Therefore, a convergence indicator is introduced especially for the comparison of the generations with different sizes based on the hypervolume. The favorable features of this indicator are also discussed. Furthermore, we will discuss available open source Java tools that allow an easy implementation of the customized NSGA-II to solve multi-objective biochemical optimization problems.

The remainder of this paper is organized as follows: Section II describes the components of the customized NSGA-II. Section III provides a comparison of open source Java frameworks focussed on a most simple implementation of the customized NSGA-II. Section IV introduces the new convergence metric and discusses the motivation for its evolution and the indicator features. Section V provides the performance results of the configurations with different population sizes. Section VI gives responses to the questions raised in this section.

II. The customized NSGA-II

In this section, the customized NSGA-II is described as used in the presented experiments. In the previous work [2][3], we have assessed the performance and interaction of different recombination and mutation operators. In these experiments, we have selected the optimal combination of recombination and mutation method that is used within the following experiments. Additionally, we have customized the encoding and selection for the purpose of peptide optimization. The procedure corresponds to the procedure of the traditional NSGA-II [3].

A. The encoding

The individuals are encoded as 20-character strings symbolizing the 20 canonical amino acids. This is the most intuitive way of peptides encoding. The individuals have a fixed length of 20 amino acids.

B. Three-dimensional biochemical minimization problem

We use three fitness functions predicting molecular features. Two fitness functions make use of the primary structure and the third works on the secondary structure. These fitness functions provide physiochemical properties that are used for drug design [11]. Moreover, this combination of fitness functions describes important peptide properties [25].The first fitness function is the calculation of the Molecular Weight (MW) that is an important peptide feature for the purpose of drug design [1]. This fitness function is selected from the open source library BioJava [12]. The second fitness function is the determination of the hydrophilicity (hydro) of a peptide. A hydrophilicity value is assigned to each peptide via the hydrophilicity scale of Hopp and Woods with a window size of the peptide length [13]. These two fitness functions work on the primary structure. The third fitness function determines the optimal global similarity score provided by the Needleman-Wunsch Algorithm (NMW) that is also part of the BioJava library. These three fitness functions act comparatively: individuals are compared to a predefined reference-solution. Therefore, these three objective functions have to be minimized.

C. The recombination operator

The $n-$point recombination operator is used, where n is determined by a linearly decreasing function:

$$x_R(t) = \frac{l}{2} - \frac{l/2}{T} \cdot t \qquad (1)$$

that depends on the peptide length l, the total number of the GA generations T and the index of the actual generation t. This operator results in combination with the following mutation in the - so far - best performance.

D. The mutation operator

An adaption of the deterministic dynamic operator of Bäck and Schütz is used to determine the number of mutation points.

$$p_{aBS} = (5 + \frac{l-2}{T-1}t)^{-1}, \qquad (2)$$

Again, l describes the peptide length, T the total generation number of the GA and t the index of the actual generation number.

E. The Aggregate Selection

The flow diagram in Figure 1 depicts the selection methods. The Aggregate Selection is tournament-based. From the tournament set individuals are chosen from the first front with a probability p_0 and with a probability $1-p_0$ the individuals are chosen via Stochastic Universal Sampling (SUS). The number N of pointers is the number of fronts and the segments are equal in size to the number of individuals in each front.

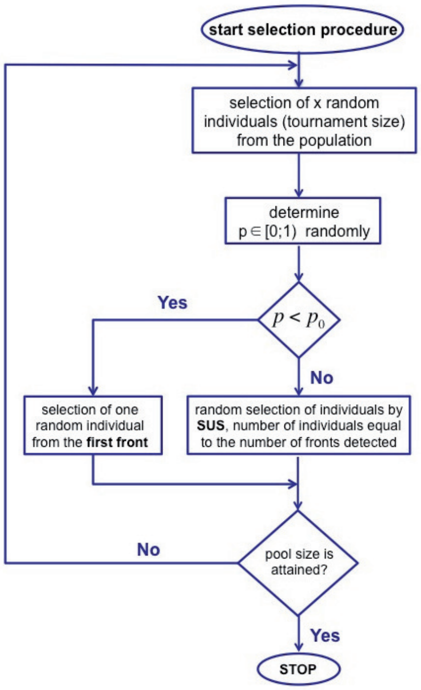

start selection procedure
selection of x random individuals (tournament size) from the population
determine p ∈ [0;1) randomly
$p < p_0$
Yes
No
selection of one random individual from the first front
random selection of individuals by SUS, number of individuals equal to the number of fronts detected
pool size is attained?
No
Yes
STOP

Fig. 1. Aggregate selection strategy

Therefore, the selection method has two parameters, the tournament size and the probability of choosing individuals from the first front. The tournaments size of 10 has proven to be an optimal choice. The parameter p_0 is challenging w.r.t. the population size.

III. Open source Java frameworks

In this section, we summarize and describe different open source Java tools that provide Genetic Algorithm implementations. The summarization is focussed on Java frameworks for a most simple implementation of BioJava, which provides several physiochemical properties via APIs. The main goal of this framework analysis is the selection of a tool which allows an easy implementation of the proposed customized NSGA-II.

The framework Java API for Genetic Algorithm (JAGA) in its current version 1.0 beta is a research tool developed and supported by the Computer Science Department of the University College London [19]. This tool does not include any moGAs, but it provides a protein string sequence encoding using 20 different characters symbolizing the 20 canonical amino acids. Among others, eight different scales like hydrophobic, aliphatic, aromatic and polar are pre-defined for each canonical amino acid. In addition, it contains for each genotype a parameter-depending crossover and mutation method and a elongation for amino acid patterns. The user interested in a moGA application has to extend this tool for this purpose, but the amino acid character encoding is a clear benefit. Other useful functions are the opportunity of creating a random initial population of protein sequences and the implementation of the Needleman-Wunsch or Smith-Waterman Algorithm.

The framework Metaheuristic Algorithms in Java (jMetal) in its current version 4.3 is an extensive and complex tool especially for moGA applications [20]. It contains beneath NSGA-II the moGA variants: Pareto Envelope-based Selection Algorithm (PESA), improved Strength Pareto Evolutionary Algorithm (SPEA2), improved PESA (PESA2), S-Metric Selection Evolutionary Multiobjective Evolutionary Algorithm (SMS-EMOA), Indicator-Based Evolutionary Algorithm (IBEA) and Multiobjective Evolutionary Algorithm based on Decomposition (MOEA/D). Further, different variation operators are implemented like single-, two- point, Simulated Binary Crossover (SBX) and polynomial, uniform and swap mutation. 'Ranking&crowding selection' is included as the traditional NSGA-II selection method as well as tournament and PESA2 selection. Additionally, jMetal provides several established metrics to evaluate the performance like the hypervolume, Inverse General Distance (IGD), General Distance (GD) and a measure for diversity. A definite advantage of jMetal is the intuitive and clear program construction, which allows an easy algorithmically extension. The disadvantage is a missing character or string encoding.

The framework Java-based Evolutionary Computation Research System (ECJ) in its current version 21 is comparable with jMetal in the issues functional complexity and potential extension. ECJ is developed at George Mason University's Evolutionary Computation Laboratory [21]. It includes the moGAs NSGA-II and SPEA2. Furthermore, different vector representations with corresponding variation operators are included as well as SUS and tournament selection, among others. Moreover, it proposes the potential to read populations from files. It does not show the intuitive and clear program structure of jMetal.

Evolutionary Algorithms workbench (EvA2) is a Java framework developed by the department of computer science at the Eberhard Karls University in Tübingen [22]. It is not only intended for research, but is also deployed for industrial applications and is available under LGPL license. Its specificity is its easy-to-use graphical user interface and provides a MATLAB interface to optimize functions in MATLAB with standard algorithm implementations in EvA2. It also has a client-server structure and provides NSGA-II, PESA and SPEA2 as moGA implementations. A string or character encoding is not implemented and an implementation afterwards is challenging, because encoding affects all parts of the tool box.

The framework Java Class library for Evolutionary Computation (JCLEC) in the current version 4 includes the evolutionary features NSGA-II and SPEA2. It proposes different encodings with various variation operators except string or character encoding, but provides an expendable program structure. Further, selection strategies like tournament and SUS based selection are also included.

Figure 2 gives an overview of the reviewed Java frameworks. These frameworks are compared under the aspects of: (i) configuration of a character or string encoding as an option, (ii) an implementation of NSGA-II, (iii) potential of a simple extension, and (iv) an intuitive program structure according to the moGA components.

TABLE I
Overview of the special framework aspects

	JAGA	jMetal	ECJ	EvA2	JCLEC
(i)	x				
(ii)		x	x	x	x
(iii)		x	x		x
(iv)		x			

Table I reveals that none of the open source Java frameworks attains all required aspects in an adequate level. As a consequence, the experiments are conducted with an user-specific implementation of this customized NSGA-II. In other cases, the open source tool jMetal is a possible alternatives for a user-specific implementation.

IV. Evaluation measures for convergence and diversity

Firstly, the convergence measure is introduced, which has been especially evolved to evaluate generations with different sizes. Subsequently, the features of this indicator are discussed followed by the presentation of measurement for diversity.

A. Introduction of the average cuboid volume

In the past, several metrics have been proposed to evaluate the convergence behavior of populations produced by a moGA.

Usually, they act on the distance of the non-dominated solution set of a generation to the true Pareto front. The hypervolume or the S-metric measures the overlapped space of the non-dominated solution set to a predefined anti-optimal reference point [14]. The hypervolume is a very established convergence metric with its favorable mathematical properties as one reason. Another convergence metric is the D-metric [24]. The D-metric makes use the hypervolume and calculates the coverage difference of two solution sets. A reference set is needed to assess the convergence to the true Pareto front. The C-metric is an appropriate measure to compare the dominance of two Pareto optimal sets [14]. The Error Ratio (ER) is a percentage measure for the number of solutions in a set that are to be found on the true Pareto front [24]. GD is a measure of the average distance between a Pareto optimal solution set to the true Pareto front [15]. It includes the minimal component-wise distance of a solution set to the nearest one on the true Pareto front. The convergence metric of Deb also measures the distance between a solution set and a reference set of the Pareto front [16]. It calculates the average normalized distance for all solutions in the solution set. A recently published convergence metric is the Averaged Hausdorff Distance Δ_p [17]. It is based on GD and the IGD [18].

The reasons for the evolution of a new convergence metric in this paper and in the scientific community in general are multiple: The disadvantage of the metrics D-metric, ER, GD, Δ_p and the convergence metric of Deb is their dependency on the true Pareto front or at least a reference set of Pareto optimal solutions that are usually unknown in the case of real-world MOPs. Furthermore, these metrics are not useful indicators for an entire ranking between generations of different sizes. However, the populations in moGAs are generally limited in size. From a more global point of view, the evaluation and comparison of the global convergence behavior of whole populations - not only the non-dominated solution set of a generation - is required with respect to the influence of the population size or the selection pressure.

For this purpose, a new metric is presented that reflects the convergence behavior of a whole population and is a 'fair' indicator for comparison of generations of different sizes. This Average Cuboid Volume (ACV) is evolved according to the model of the hypervolume. The motivation for the exploitation of the hypervolume model is to profit from its preferable properties as mentioned above. The benefit of this new metric compared to the hypervolume is the low computational complexity as no point ordering is required.

In the following, we assume that the underlying optimization problem is to minimize. The metric calculates the average cuboid volume of the cuboids spanned by the solution points to a pre-defined reference point r:

$$ACV(X) = \frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{k}(x_{ij}-r_j)\right), \qquad (3)$$

where n is the population size, k is the number of objectives, x_i are the solutions on the population X and x_{ij} is the $j-th$ component of a solution x_i. It holds $(x_{ij}-r_j) > 0$ as the pre-defined reference point is chosen as the theoretical minimal limit of the true Pareto front. The lower the indicator values the more positive is the global convergence behavior as the reference point is chosen as a theoretical optimal point.

B. Discussion of the average cuboid volume

The question regarding the suitability of a metric for evaluation depends on the intention of the investigation object and the preferences. ACV is intended to evaluate the global convergence behavior of a whole population with the ultimate aim of comparing solution sets of different sizes according to the proximity to the true Pareto front.

The first expectation that is important for the use of ACV is that the convergence quality shall not change if the number of equally solutions increases. ACV does not fulfill this averaging strategy: Let $x \in \mathbb{R}^k$ be a solution of the optimization problem and $X = \{x\}$. Further, $Y = \{x, ...x\}$ is a set of n equally copies of the solution x, then $ACV(Y) = ACV(X)$.

The second expectation is described by the following observation: An intuitive indicator reflecting the quality of approximation sets of different Pareto front refinements requires 'better' indicator values for the finest approximation set. This effect is demonstrated for ACV by an example also used in [10]:

Example 1: The Pareto front is the line segment between the points $y_1 = (0,1)$ and $y_2 = (1,0)$ meaning

$$PF_{true} = \{i \cdot y_1 + (1-i)\cdot y_2 | i \in (0,1)\}. \qquad (4)$$

We consider the following three approximation sets of increasing refinement of the Pareto front

$$Y_1 = \{(i\cdot 0.2, 1-i\cdot 0.2)|i \in \{1,2,3,4\}\}, \qquad (5)$$
$$Y_2 = \{(i\cdot 0.1, 1-i\cdot 0.1)|i \in \{1,2,...,9\}\}, \qquad (6)$$
$$Y_3 = \{(i\cdot 0.01, 1-i\cdot 0.01)|i \in \{1,2,...,99\}\}. \qquad (7)$$

The indicator values of ACV for the approximation sets with the reference point $(0,0)$ are: $ACV(Y_1) = 0.2$, $ACV(Y_2) = 0.183$ and $ACV(Y_3) = 0.167$.

The third preferable expectation of this indicator is the averaging effect. It is trivial that a dominating solution x yields better indicator values than the dominated one y, because $ACV(\{x\}) = \prod_{i=1}^{k}(x_j - r_j) < \prod_{i=1}^{k}(y_j - r_j) = ACV(\{y\})$. From this observation it can be interpreted that if one dominated solution x_1 in the solution set $X = \{x_1, x_2, ..., x_n\}$ is replaced by a dominating one $\bar{x}_1$, then $ACV(\{x_1, x_2, ..., x_n\}) > ACV(\{\bar{x}_1, x_2, ..., x_n\})$. The averaging effect of ACV is illustrated by the example which has also been used for Δ_p [17]:

Example 2: The true discrete Pareto front is described by $P = \{p_i | p_i = (0.1\cdot(i-1); 1-(i-1)\cdot 0.1)$ with $i = 1, ..., 11\}$. Two solution sets are given by $X_1 = \{x_{1,1}, p_2, ..., p_{11}\}$ and $X_2 = \{x_{2,1}, x_{2,2}, ..., x_{2,11}\}$ with the elements $x_{1,1} = (\epsilon, 10)$ and $x_{2,i} = p_i + (\frac{\epsilon}{2}, 5)$ with $i = 1, ..., 11$. For the outlier $x_{1,1}$

the values $\epsilon = 0.001$ is used for numerical evaluations. X_1 is a better approximation of the true Pareto front, but it contains the outlier $x_{1,1}$. On the other side, X_2 is close to the true Pareto front and the difference of each element to the Pareto front is less than the one of the outlier. As we are interested in an averaging effect, the indicator values of X_1 has to be better than the one of X_2. This is true for ACV as $ACV(X_1) = 0.15$ and $ACV(X_2) = 2.65$.

C. The diversity measure

The measure for diversity calculates the average distance of all pairs of solutions (see [4]):

$$\Delta = \sum_{i,j=1,i<j,i\neq j} \frac{|d_{ij} - \bar{d}|}{N} \qquad \text{with } N = \binom{n}{2}. \tag{8}$$

$d_{i,j}$ symbolizes the Euclidean distance of two solutions x_i and x_j, $\bar{d}$ is the mean of all measured distances and n is the population size.

V. Simulation onset and experiments

The test runs are performed for different configurations. The configurations are composed of a differing population size (30, 50, 70, 100, 130, 150) and the selection parameters $p_0 = 0\%, 30\%, 50\%$. These parameters have been emphasized by previous experiments. The selection parameter $p_0 = 0\%$ stands for SUS exclusively. Each multi-objective configuration is repeated 20 times until the 18th generation - for statistical reasons. The test runs are evaluated by the convergence indicator ACV and the diversity measure as introduced in the last section. ACV uses the theoretical minimal limit $(0/0/0)$ of the Pareto front as an optimal reference point. Therefore, a good performance is achieved if the ACV value is as low as possible and the diversity value is as high as possible. Boxplots are created for each configuration and for each objective of evaluation (Fig. 2 - Fig. 8). The values of ACV and diversity are scaled under the same criterion for a better graphical presentation. The figures are ordered according to the population size. The standard population size within the customized NSGA-II is 100 [2][3] (Fig. 5). Therefore, the results are discussed w.r.t. an increase and an decrease of this size: In general, a decrease of the population size to 70 and 50 results in an increase of the ACV values and a decrease of the diversity values (Fig. 3, Fig. 4). This means that the convergence and the spread within the solutions is reduced caused by decreasing the population size. The ACV values decrease for a population size of 30 (Fig. 2) independent of the choice of the selection parameter. Moreover, the diversity also decreases and results in the lowest diversity among all configurations. An increase of the population size to 130 results in a decrease of ACV and in an increase of the diversity, once more independent of the selection parameter (Fig. 6). A further increase of the population size to 150 and 200 results in a stagnation of the ACV and diversity values (Fig. 7, Fig. 8).

Further, the effect of the selection parameter is discussed: Varying the population size from 50 to 100 (Fig. 3- Fig. 6),

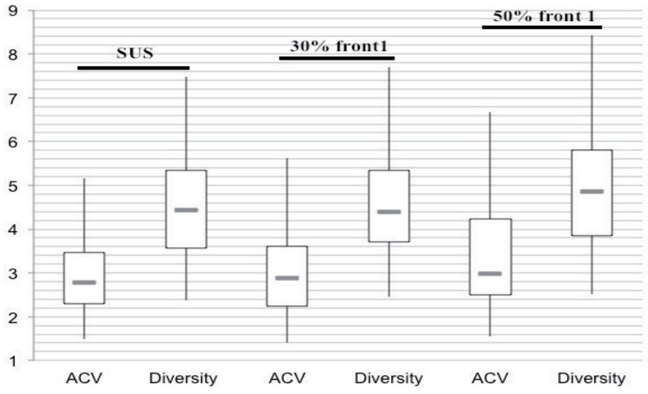


Fig. 2. Population size 30

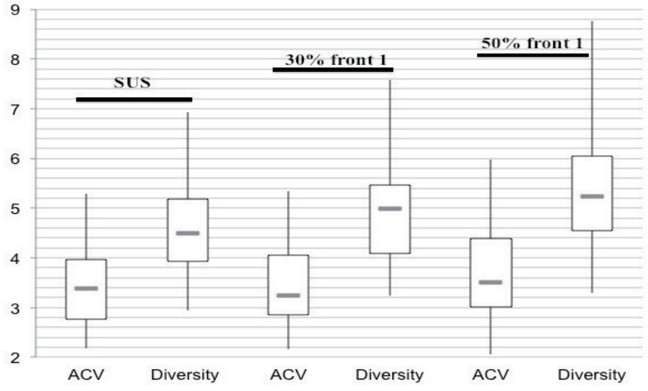


Fig. 3. Population size 50

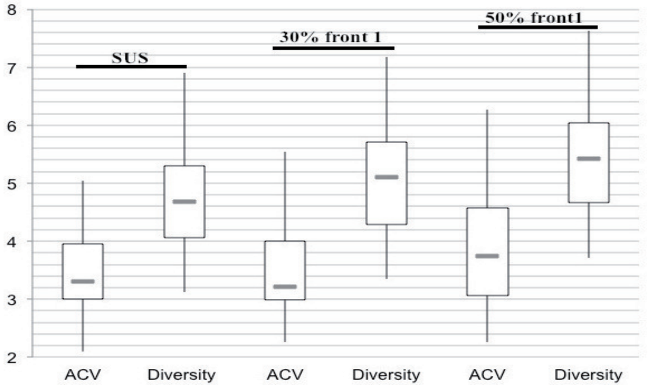


Fig. 4. Population size 70

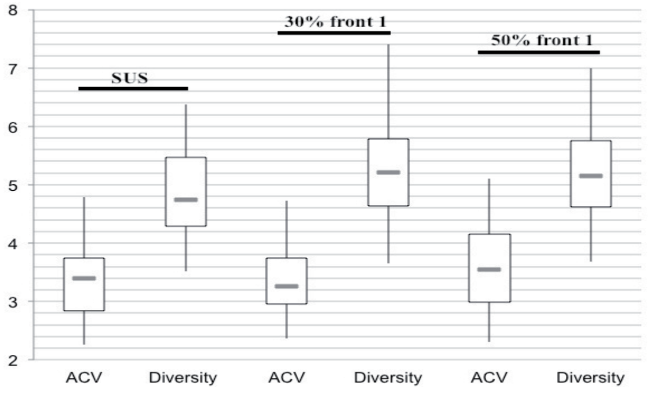


Fig. 5. Population size 100

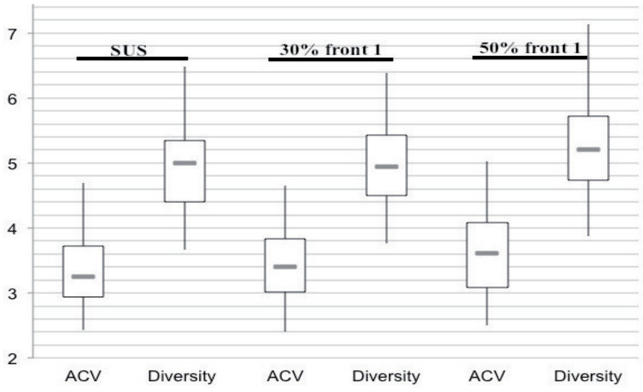


Fig. 6. Population size 130

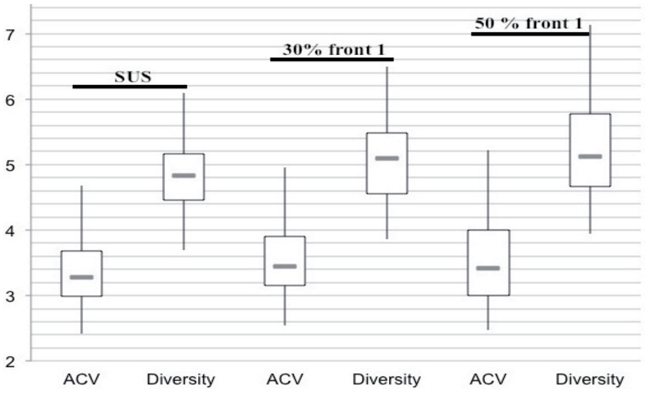


Fig. 7. Population size 150

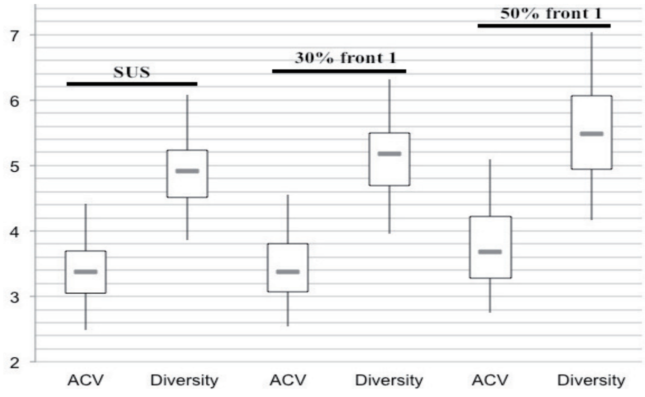


Fig. 8. Population size 200

the ACV values are comparable for $p_0 = 0\%$ (denoted as 'SUS' in the figures) and $p_0 = 30\%$ (denoted as '30% front 1' in the figures), though the diversity improves evidently for $p_0 = 30\%$ compared to SUS. Independent of the population size, $p_0 = 50\%$ (denoted as '50% front 1' in the figures) results in a remarkable increase of the ACV values and only a slight improvement of diversity compared to SUS and $p_0 = 30\%$. For the population sizes from 130 to 200, the influence of the selection parameter is reduced (Fig. 6- Fig. 8): There is only a slight improvement to report in diversity for $p_0 = 30\%$ compared to SUS. The convergence is remarkable reduced for $p_0 = 50\%$, though the diversity is improved.

The best performance of the configurations is received with a population size from 70 to 100 and a selection parameter of 30% as the values for ACV are at most low, whereas the diversity values are at most high. At least, the performance of the configurations with a population size from 50 to 100 with $p_0 = 30\%$ are comparable in convergence and diversity with the performance of the configuration population size of 130 and SUS. Concluding, the best configuration is expectable with a population size in the range from 70 to 100 and a selection parameter of $p_0 = 30\%$.

Regarding the questions presented in the introduction we conclude that an increase of the population size does not result in better performance. The customized NSGA-II provides good performance regarding convergence and diversity within a limited range of population size for the presented three-dimensional minimization problem. Empirically, there is no interdependence between population size and selection: The choice of $p_0 = 30\%$ usually results in the best performance independent of the population size. Therefore, it is not possible to speed up the convergence by increasing or decreasing of the population size and a suitable adaption of the selection parameter.

VI. Conclusion and Future Work

The interdependence of the population size and the selection parameter in this customized NSGA-II is exemplary examined on a generic three-dimensional biochemical minimization problem focused on three central questions: The first question is aimed at the influence of large populations on the convergence speed. Early convergence as a main goal of our moGA is defeated since an increase of the population size results in higher speed of convergence. The experiments show that the optimal population size w.r.t. convergence and diversity is in a limited range from 70 to 100. An increase of the population size above 100 results in a stagnation of the convergence behavior and the diversity. A population size lower than 50 does not provide a convincing diversity within the solutions. Our second question is focused on the impact of the population size and the selection parameter. A configuration rule for the selection parameter depending on the population size is necessary in the case of a large dependence of both. However, the experiments do not reveal an interdependence of the population size and the selection parameter. Though, the diversity of the configurations with a population size from 50 to 100 is remarkably improved with a selection parameter of 30% compared to $p_0 = 0\%$ (SUS). Higher values for p_0 are not advisable as the speed of convergence is reduced. The third question asks for a range of the population size providing the best performance: This range is fixed to a population size from 70 to 100 based on the evaluation of the experiments.

The convergence performance of the experiments is assessed via a newly introduced convergence indicator, which is especially evolved to compare the convergence behavior of populations with different sizes. It is based on the established hypervolume and calculates the average cuboid volume of the cuboids spanned by the solution points to a pre-defined reference point, which is chosen as a theoretical optimal point. The benefit of ACV compared to the hypervolume is the lower computational complexity and the choice of the reference point. It is easier to determine an optimal point than an anti-optimal one for real world applications. Furthermore, a comparison of Java frameworks is submitted as a guidance for a simple implementation of the customized NSGA-II. The frameworks are compared to the aspects of a character or string encoding, the implementation of NSGA-II and the potential of a most simple extension.

For future work, we currently work on the confirmation of these results on a four-dimensional biochemical optimization problem. Further, we will evolve a selection strategy based on ACV for ongoing improvements.

References

[1] D. J. Craik, D. P. Fairlie, S. Liras, and D. Price, ”The Future of Peptide-based Drugs,” Chemical Biology & Drug Design, 81(1), 2013, pp. 136-147.

[2] S. Rosenthal, N. El-Sourani, and M. Borschbach, ”Introduction of a Mutation Specific Fast Non-dominated Sorting GA Evolved for Biochemical Optimization,” SEAL 2012, LNCS 7673, 2012, pp. 158-167.

[3] S. Rosenthal, N. El-Sourani, and M. Borschbach, ”Impact of Different Recombination Methods in a Mutation-Specific MOEA for a Biochemical Application,” L. Vanneschi, W. S. Bush, and M. Giacobini (Eds.): EvoBIO 2013, LNCS 7833, 2013, pp. 188-199.

[4] S. Rosenthal and M. Borschbach, ”A Benchmark on the Interaction of Basic Variation Operators in Multi-Objective Peptide Design evaluated by a Three Dimensional Diversity Metric and a Minimized Hypervolume,” M. Emmerich et al. (eds.): EVOLVE - A Bridge between Probability, Set Oriented Numerics and Evolutionary Computation IV, 2013, pp. 139-153.

[5] J. T. Alander, ”On Optimal Population Size of Genetic Algorithms,” in Proceedings of the IEEE Computer Systems and Software Engineering, 1992, pp. 65-69.

[6] V. K. Koumousis and C. P. Katsaras, ”A Saw-Tooth Genetic Algorithm Combining the Effects of Variable Population Size and Reinitialization to Enhance Performance,” IEEE Transactions on Evolutionary Computation, vol. 10, no. 1, 2006, pp. 19-28.

[7] T.-L. Yu, K. Sastry, D. E. Goldberg, and M. Pelikan, ”Population sizing for entropy-based model building in genetic algorithms,” Illinois Genetic Algorithms Laboratory, University of Illinois, Tech. Rep., 2006.

[8] T. Bäck, A. Eiben, and V. der. Vaart, ”An empirical study on GAs without parameters,” in Proceedings of the 6th International Conference on Parallel Problem Solving from Nature, 2000, pp. 315-324.

[9] Z. M. Jaroslaw Arabas and J. Mulawka, ”GAVaPSa genetic algorithm with varying population size,” in Proceedings of the IEEE International Conference on Evolutionary Computation, 1995, pp. 73-78.

[10] A. E. Eiben, M. C. Schut, and A. R. Wilde, ”Is Self-Adaption of Selection Pressure and Population Size Possible? a Case Study,” in Parallel Problem Solving from Nature - PPSN IX, vol. 4193, 2006, pp. 900-909.

[11] T. Sovany et al., ”Application of Physiochemical Properties and Process Parameters in the Development of a Neural Network Model for Prediction of Tablet Characteristics,” AAPS PharmSciTech, vol. 14(2), 2013, pp. 511-516.

[12] BioJava: CookBook, URL: http://www.biojava.org/wiki/BioJava/ [retrieved: December, 2013].

[13] T. P. Hopp, K. R. Woods, ”A computer program for predicting protein antigenic determinants,” Mol Immunol, 20(4), 1983, pp. 483-489.

[14] E. Zitzler and L. Thiele, ”Multiobjective Optimization using Evolutionary Algorithms - a Comparative Case Study,” in A. E. Eiben, T. Bäck, M. Schoenauer, and H. P. Schwefel (EDS.), Fifth International Conference on Parallel Problem Solving form Nature (PPSN-V), 1998, pp. 292-301.

[15] D. A. Van Veldhuizen and G. B. Lamont, ”Multiobjective Evolutionary Algorithm Test,” in Proceedings of the 1999 ACM Symposium on Applied Computing, San Antonio, Texas, 1999, pp. 351-357.

[16] K. Deb, S. Jain, ”Running performance metrics for Evolutionary Multi-objective Optimization,” Kan GAL Report No. 2002004, Kanpur Genetic Algorithms Laboratory, Indian Institute of Technology Kanpur, 2002.

[17] O. Schütze, X. Esquivel, A. Lara, and C. A. Coello Coello, ”Using the Averaged Hausdorff Distance as a performance measure in evolutionary multiobjective optimization,” IEEE Transactions on Evolutionary Computation, vol. 16(4), 2012, pp. 504-522.

[18] C. A. Coello Coello and N. Cruz Cortis, ”Solving Multiobjective Optimization Problems using an Aritifical Immune System. Genetic,” Programming Evolvable Mach., vol. 6 (2), 2005, pp. 163-190.

[19] Java API for Genetic Algorithm (JAGA), URL: www.jaga.org/ [retrieved: January, 2014].

[20] Metaheuristic Algorithms in Java (jMetal), URL: www.jmetal.sourceforge.net/ [retrieved: January, 2014].

[21] Java-based Evolutionary Computation Research System (ECJ), URL: www.cs.gmu.edu/∼edab/projects/ecj/ [retrieved: January, 2014]

[22] Evolutionary Algorithms workbench (EvA2), URL: www.ra.cs.uni-tuebingen.de/software/EvA2/introduction.html/ [retrieved: January, 2014]

[23] N. Röckendorf, M. Borschbach, and A. Frey, ”Molecular Evolution of Peptide Ligands with Custom-tailored Characteristics,” PLOS Comput Biol 8(12), 2012

[24] G. Grosan, M. Oltean, and D. Dumitrescu, ”Performance Metrics for Multiobjective Evolutionary Algorithms,” Proceedings of Conference on Applied and Industrial Mathematics (CAIM), 2003

[25] D. Heider et al. ”A Computational Approach for the Identification of Small GTPases based on Preprocessed Amino Acid Sequences,” in Technol. Cancer Res. Treat 8, 2009, pp. 333-341.

Evaluation of Imputation Methods for Missing Data and Their Effect on the Reliability of Predictive Models

Xiao-Ou Ping, Feipei Lai, Yi-Ju Tseng
Graduate Institute of Biomedical Electronics and Bioinformatics, Department of Computer Science and Information Engineering
Dept. of Electrical Eng.
National Taiwan University
Taipei, Taiwan
pingxiaoou@gmail.com

Ja-Der Liang, Guan-Tarn Huang, Pei-Ming Yang
Department of Internal Medicine
National Taiwan University Hospital and National Taiwan University College of Medicine
Taipei, Taiwan
jdliang@ntuh.gov.tw

***Abstract* — In medical research, the problem of missing data occurs frequently. In this paper, eight imputation methods are evaluated based on accuracy and stability through a simulation experiment. The objective of this paper is to find appropriate methods for handling incomplete data sets during the development of predictive models which predict the recurrence status of liver cancer patients. Support vector machine (SVM) is employed for building predictive models. The data sources produced by different missing data handling methods (complete variable analysis and imputation method) are used for evaluating the impact on the development of the recurrence predictive model. Imputation methods show the potential benefit of features with missing values during the development of the recurrence predictive model.**

Keywords - incomplete data; missing value; predictive model; liver cancer

I. INTRODUCTION

According to a study reviewing 100 articles among seven cancer journals, up to 81 articles have evidence of missing data [1]. The problem of missing data occurs frequently. Therefore, how to handle incomplete data set is a crucial issue during data analysis. To handle incomplete data sets, several general handling methods are proposed [2]: (1) complete variable analysis: dropping the variables with missing data and analyzing only the variables without missing data, and (2) imputation method: estimating the missing values based on different methods. In the study, the complete variable data set, and the data sets imputed by different imputation methods are both employed for evaluating the impact of missing value handling methods for developing the predictive model. Performances of predictive models built based on these two types of data sets are compared for checking whether if the features with missing data have potential benefit for building the predictive model.

To estimate missing values in data sets, eight imputation methods are employed in this study and we design a simulation experiment for comparing the imputation performance on the stability and accuracy. Of eight imputation methods, six are single imputation (i.e., single imputed value for each missing value) and two are multiple imputations (i.e., multiple imputed values for each missing value). In this work, normalized root mean squared error (NRMSE) [3][4] is used for evaluating the accuracy of imputation methods; furthermore, the stability of imputation methods is also evaluated through repeated simulation experiments.

According to the global cancer statistics in 2011, liver cancer in men is the second most frequent cause of cancer death and in women, it is the sixth leading cause of cancer death. Hepatocellular carcinoma (HCC), as the most common primary liver cancer [5], has been the leading cause of cancer death in Taiwan.

For patients with early-stage HCC who are not suitable for surgical resection or liver transplantation, radiofrequency ablation (RFA) is the best alternative treatment [6]. In previous studies, researchers estimated that the cumulative 5 year recurrence rate is more than 70% for patients who received RFA [7]. The recurrence predictive models play a crucial role for physicians and patients in enabling the opportunities of supporting early prediction of a recurrence status.

In the work, support vector machines (SVM) [8][9] is employed as a classifier for developing the recurrence predictive model for newly diagnosed HCC patients receiving RFA treatment in one year. The predictive models built based on the data sources produced by different missing data handling methods (i.e., complete variable analysis and imputation method) are further evaluated and compared for presenting the impact of these methods. In the past few years, SVMs have been widely employed in medical specialties such as breast cancer [10] and liver diseases (fatty liver [11] and liver fibrosis [12]).

This study introduces a two-level approach to evaluating imputation methods and predictive models when the problem of missing data occurs. The performance of an imputation method (i.e., accuracy and stability) may have variance according to its parameter settings and different data sets (e.g., different data sets from patients with different diseases). In the first level, an evaluation of imputation methods assists researchers in selecting appropriate imputation methods and their parameter settings for a specific data set through a designed simulation experiment. Furthermore, this simulation experiment can further provide information for evaluating reliability of predictive models which are developed based on imputed data sets (i.e., missing values are imputed by an imputation method). In the second level, the performance of an imputation method is further evaluated based on each clinical feature with missing values. When predictive models employ clinical features with estimated values (estimating by an imputation method), the performance of an imputation method for these features can be regarded as factors in evaluating reliability of these predictive models. When a predictive model relies heavily on a specific clinical feature and the performance of an imputation method for this feature is not good, the predictive model may be not a model with good reliability. This study focuses on a case study which is to find appropriate methods for handling incomplete data sets during the development of predictive models which predict the recurrence status of liver cancer patients.

The construction of this paper is organized as follows. In Section II, an overview of our method is presented. Section III introduces the material about a specific data set used in this study. Section IV describes a designed simulation experiment, and introductions and evaluations of imputation methods. Section V contains a method of building predictive models and evaluations of these models. Sections VI and VII present results of imputation methods and predictive models. Section VIII discusses results and limitations of this study. Finally, conclusion and future work are given in Sections IX and X.

II. Method overview

A two-level approach is introduced to evaluating imputation methods and predictive models when the problem of missing data occurs. Fig. 1 shows the overview of this work. A simulation experiment is designed for evaluating performance of imputation methods and reliability of a predictive model.

Before performing the simulation experiment, cases that have missing values (MVs) are removed from original data sets for producing complete cases data set without MVs. In each simulation round, partial original data (ten percent data) in the complete cases data set are masked as missing values and then these missing values are imputed by imputation methods. After the process of data imputation, the original values included in the complete cases data set and the imputed valued estimated by imputation methods can be compared for evaluating performances of these imputation methods based on NRMSEs.

Two major data sources are used for developing the predictive model: (1) complete variables data set: dropping the variables (features) with missing data and using only the variables without missing data, and (2) imputed data sets: imputing the missing values included in the original data set and using the imputed data set. The predictive models are evaluated based on criteria such as sensitivities and specificities.

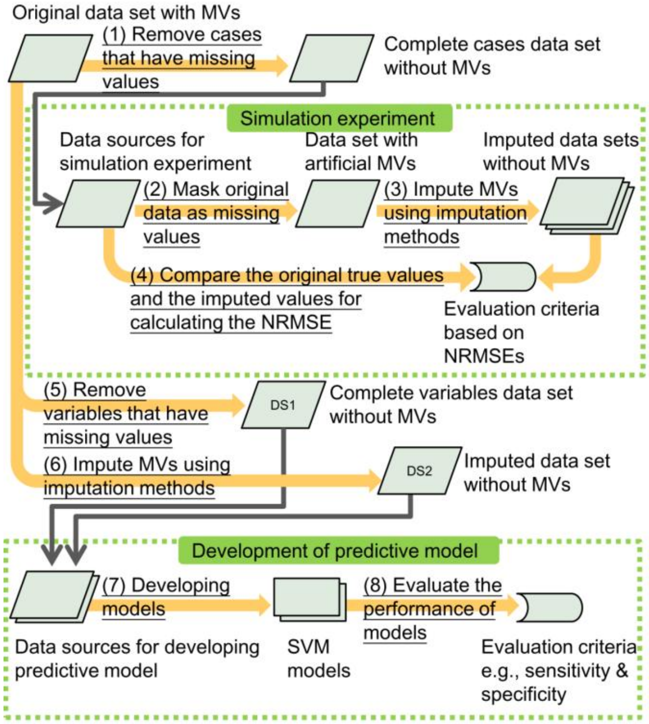

Original data set with MVs
(1) Remove cases that have missing values
Complete cases data set without MVs
Simulation experiment
Data sources for simulation experiment
Data set with artificial MVs
Imputed data sets without MVs
(2) Mask original data as missing values
(3) Impute MVs using imputation methods
(4) Compare the original true values and the imputed values for calculating the NRMSE
Evaluation criteria based on NRMSEs
(5) Remove variables that have missing values
DS1
Complete variables data set without MVs
(6) Impute MVs using imputation methods
DS2
Imputed data set without MVs
Development of predictive model
(7) Developing models
(8) Evaluate the performance of models
Data sources for developing predictive model
SVM models
Evaluation criteria e.g., sensitivity & specificity

Figure 1. Simulation experiments and development of a predictive model based on two different data sets (DSs).

III. Material

83 HCC patients received ultrasound guided RFA were included in this study. RFA is their first treatment for HCC in NTUH between 2007 and 2009. Of the 83 patients, 18 patients had recurrent HCCs in one year after the RFA treatment and 65 patients were not recurrent in one year. A total of 20 clinical features included in this study are as follows: age, gender, tumor number, the size of the maximal tumor, liver cirrhosis, Barcelona Clinic Liver Cancer (BCLC) staging classifications [13], and 14 serum laboratory tests, including prothrombin time (INR), albumin, aspartate aminotransferase (AST), alanine transaminase (ALT), total bilirubin, creatinine, platelet count, alpha-fetoprotein (AFP), HBsAg, anti-HCV, alkaline phosphatase (ALP), direct bilirubin, total protein, and Gamma-glutamyl transpeptidase (GGT). In each feature, one value that is before and closest to the RFA treatment is used. The last four features include missing values. The missing rates of these features are as follows. ALP has 8.43% missing values, direct bilirubin has 15.66%, total protein has 22.89%, and GGT has 27.71%. The complete data set is produced by dropping these four features with missing values. The imputed data sets are produced by estimating missing values of these four features based on eight imputation methods and different parameter settings.

IV. DATA IMPUTATION

A. Simulation experiment

There are two types of imputation methods, including single imputation (i.e., single imputed value for each missing value) and multiple imputation (i.e., multiple imputed values for each missing value). This study employed both types of imputation methods for comparing their influence on this specific data set in our study.

For applying the single imputation methods, the “pcaMethods” is employed. The “pcaMethods” is a Bioconductor package and proposed by Stacklies et al. [14]. In the package, they implement a collection of principal component analysis (PCA) based methods and a non-PCA based method for estimating the missing values of the incomplete data. This package contains six single imputation methods, including singular value decomposition based imputation method (SVDImpute), local least squares imputation method (LLSImpute), probabilistic PCA (PPCA), Bayesian PCA (BPCA), non-linear PCA (NLPCA), and non-linear estimation by iterative partial least squares PCA (Nipals PCA). For applying the multiple imputation methods, the multivariate imputation by chained equations (MICE) [15] and multiple imputation (MI) [16] packages in R are employed. In this work, these methods included in these packages are employed for estimating missing data of the clinical features (e.g., serum laboratory tests).

For each feature with missing values, the cases with missing values are removed and simulation experiment is performed based on the complete data set (Fig. 2). The data set is separated into ten parts randomly. To reduce the bias due to just one simulation, the simulation experiment is repeated ten times and each case has one chance to be masked as missing values. The concept is similar to ten-fold cross validation. In each round of simulation experiment, each 10% data of this feature are masked as artificial missing values and eight imputation and different parameter settings are used for estimating the missing values. Then the true original values and the estimated values are compared for evaluating the performance of imputation method.

For each imputation method and its corresponding parameter setting, the experiment is repeated 10 times and each case has one chance to be masked as missing values. The distribution of these 10 NRMSEs is analyzed rather than calculating one mean value of these NRMSEs. The average of first quartile, third quartile, and the median of these 10 NRMSEs is regarded as the imputation method selection criterion for comparing the imputation performance in terms of stability and accuracy. A low NRMSE score means few imputation errors and high accuracy. Low imputation method selection criterion means high stability and accuracy.

Eight imputation methods and their corresponding parameter settings (total 37 combinations) are employed for estimating missing values of each feature. For each feature, what combinations achieve top 10 leading imputation performances are analyzed (i.e., top 10 scores of imputation method selection criterion). For example, the SVDImpute have five combinations (i.e., five parameter settings). If its four combinations achieve top 10 leading imputation performances, then its rate of combinations achieving top 10 leading imputation performances is 80%. After simulations of all features are done. The overall rate of combinations achieving top 10 leading imputation performances for all feature can be calculated by averaging the rate of combinations achieving top 10 leading imputation performances of each feature.

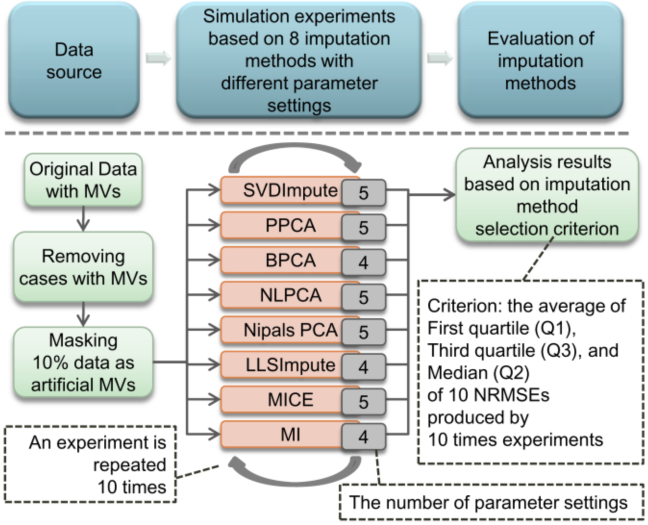

Data source
Simulation experiments based on 8 imputation methods with different parameter settings
Evaluation of imputation methods
Original Data with MVs
Removing cases with MVs
Masking 10% data as artificial MVs
SVDImpute 5
PPCA 5
BPCA 4
NLPCA 5
Nipals PCA 5
LLSImpute 4
MICE 5
MI 4
Analysis results based on imputation method selection criterion
Criterion: the average of First quartile (Q1), Third quartile (Q3), and Median (Q2) of 10 NRMSEs produced by 10 times experiments
An experiment is repeated 10 times
The number of parameter settings

Figure 2. The simulation experiment based on data set without missing values using eight imputation methods.

B. Imputation methods

In SVDImpute, singular value decomposition (SVD) is used for obtaining a set of mutually orthogonal expression patterns (e.g., eigengenes in their study) [17]. These patterns can be used to approximate the expression of all features in data sets based on the linear combination of these patterns. PCA is popular approach for data analysis and data processing (e.g., dimension reduction). PCA is not based on a probability model and PPCA [18] includes an expectation–maximization (EM) approach for PCA with a probabilistic model [14]. BPCA is based on three processes, including principal component (PC) regression, Bayesian estimation, and an EM-like repetitive algorithm [19]. NLPCA is regarded as a non-linear generalization of standard linear PCA [20]. Nipals PCA [21] is a method at the root of PLS regression [14]. The parameter settings of SVDImpute, PPCA, BPCA, NLPCA, and Nipals PCA all include the number of principal components.

LLSimpute [22] estimates missing values based on a linear combination of k selected similar variables. The k variables are selected by the Euclidean distance or by Pearson correlation coefficients. The optimal combination is found by local least squares (LLS) regression [14]. The parameter setting of LLSImpute is the number of variables selected for regression.

MICE is used for generating multiple imputations [15] and it contains different imputation functions, including the predictive mean matching (pmm), Bayesian linear regression (norm), linear regression ignoring model error (norm.nob), unconditional mean imputation (mean), and random sample from the observed values (sample). The parameter setting of MICE is based on the number of iterations and different imputation functions used for multiple imputations.

Multiple imputation (MI) is used for generating multiple imputations based on iterative regression imputation [16] and it contains different models for different variable types. For example, the “binary” regression model is used for binary data, and the “categorical” regression model is used for unordered categorical data. The parameter setting of MI is based on the number of iterations and the functions used for adding noise in multiple imputation procedure (e.g., reshuffling and fading).

In multiple imputation methods, MICE and MI, the number of imputed data sets is set as 2 in this study and two imputed data sets are generated. Both of these data sets are used for developing predictive models and a model with better performance is selected.

C. *Evaluation of imputation methods*

NRMSE is frequently applied for evaluating the performance of imputation methods. The root mean squared error (RMSE) is used for calculating the error between the estimated values of the missing entries and original true values in the complete data set. The RMSE is further normalized by the following constant: the range of the original true values over the missing entries [3][4].

$$NRMSE = \frac{\sqrt{mean\ [(y_{estimated} - y_{true})^2]}}{\max(y_{true}) - \min(y_{true})} \qquad (1)$$

The “*max()*” function denotes the maximum value of a listing numbers. The “*min()*” function denotes the minimum value of a listing numbers. The $y_{\text{estimated}}$ means the estimated values over the missing entries and the estimated values are imputed by different imputation methods. The y_{true} means the original true values of the missing entries and the true values are from the original complete data set.

V. Recurrence Predictive Model

In the work, the SVM is employed as a classifier for the prediction of the recurrence status of the patients with HCC after RFA treatment in one year.

A. *SVM for classification*

The SVM was proposed by Boser, Cortes, and Vapnik and it is widely used for solving classification problem [8][9]. The mapping function is used for mapping input feature vectors into higher dimensional feature space for SVM to find the linear separating hyperplane for separating the instances into two classes. In the work, radial basis function (RBF), is selected as the kernel function and the cost parameter *C* and parameter of kernel function gamma are the parameters that can be adjusted during the development of the SVM classification model. In this work, we perform SVM based on LIBSVM [23]. The grid search is employed for searching the appropriate parameters, *C* and gamma, of SVM models. In the work, the grid search based on 5-fold cross-validation (inner loop) is adopted for finding appropriate parameters [23][24]. The sensitivity is regarded as the criterion for finding parameters in grid search to achieve better sensitivity in our data set.

B. *Feature Selection*

In the work, the hybrid feature selection method was employed. We combine simulated annealing (SA) [25] with random forests [26]. First, SA was employed for selecting a subset of features from all features. Random forests (RF) was employed for assigning the weight of importance for the feature in the selected subset. The selected features were added stepwise as the input data to train our SVM model. SA was developed from the idea of annealing of metallurgy. The raw material can be heated for growing a crystal. The temperature is reduced until the crystal structure is frozen and the better results can be achieved through slower process of cooling [27]. In this study, the “better result” denotes the better subset of features. RF creates many classification trees. The importance of a feature is determined by permuting this feature and all others were reserved and then calculating the increased amount of prediction error [26]. In the work, SA is performed using R package, named “Subselect” [28] and RF was performed using R package, named “FSelector” [29].

Double five-fold cross-validation loop method is adopted. An inner loop five-fold cross-validation is performed on the training data set of an outer loop for finding appropriate parameter settings of SVM model. Then, the selected parameters are used for training the whole training data set of an outer loop and getting a trained SVM model and a training classification result. The selected features were added stepwise to train different SVM models. This training classification result is further used as criteria for selecting appropriate feature combinations. Finally, the average classification results of outer loop cross-validation are presented.

C. *Evaluation of the recurrence predictive model*

The sensitivity, specificity, accuracy, balanced accuracy (BAC), positive predictive value (PPV), negative predictive value (NPV), and area under the receiver operating characteristic (ROC) curve (AUC) are used as evaluated metrics for evaluating the predictive models. The definitions are as follows:

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \qquad (2)$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP}) \qquad (3)$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{FP} + \text{FN} + \text{TN}) \qquad (4)$$

$$\text{BAC} = ((\text{Sensitivity} + \text{Specificity}) / 2) \qquad (5)$$

$$\text{PPV} = \text{TP} / (\text{TP} + \text{FP}) \qquad (6)$$

$$\text{NPV} = \text{TN} / (\text{TN} + \text{FN}) \qquad (7)$$

TP (True Positive) is the patient predicted with recurrent HCC and the patient actually has recurrent HCC. TN (True Negative) is the patient predicted without recurrent HCC and the patient is actually without recurrent HCC. FP (False Positive) is the patient predicted with recurrent HCC but the patient is actually without recurrent HCC. FN (False

Negative) is the patient predicted without recurrent HCC but the patient actually has recurrent HCC. In this study, the ROC is created based on decision values of the SVM [30].

VI. PERFORMANCE OF SIMULATION EXPERIMENTS

The rate of imputation methods which has top 10 leading performances with different parameter settings for four features with missing values are shown in Fig. 3. SVDImpute and Nipals PCA perform well (in top 10 leading performances) in 80% of cases (different parameter settings and different features). The PPCA performs well in 55% of cases and the BPCA performs well in 44% of cases. Other imputation methods perform well under 30% of cases, especially the LLSImpute only performs well in 6% of cases.

For each imputation method, a specific parameter setting which has the best rate of top 10 leading performances for four features with missing values are selected and further analyzed, including ALP, direct bilirubin (DB), total protein (TP), and GGT. In TABLE I, the SVDImpute with the parameter value of 15, performs well (in top 10 leading performances) in all four features with missing values. The Nipals PCA with the parameter value of 15, also performs well in all four features. The PPCA, BPCA, and MICE can find parameters to perform well in three features (ALP, total protein, and GGT). The NLPCA can find parameters to perform well in two features (Total protein and GGT). The LLSImpute and MI methods can only find parameters performing well in a single feature (The LLSImpute can perform well in GGT, and MI can perform well in direct bilirubin).

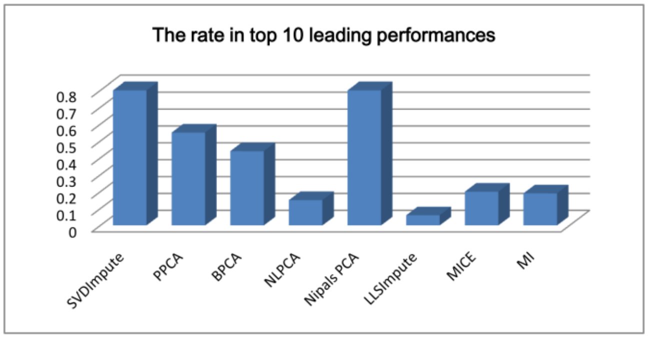


Figure 3. The rate of imputation methods which has top 10 leading performances with different parameter settings for four features with missing values.

TABLE I. THE PARAMETER SETTINGS WITH THE BEST RATE OF LEADING 10 PERFORMANCES FOR FOUR FEATURES WITH MISSING VALUES.

Method	Parameter	ALP	DB	TP	GGT
SVDImpute	15	V	V	V	V
PPCA	20	V		V	V
BPCA	1	V		V	V
NLPCA	1			V	V
Nipals PCA	15	V	V	V	V
LLSImpute	5				V
MICE	mean, 20	V		V	V
MI	fading, 100		V		

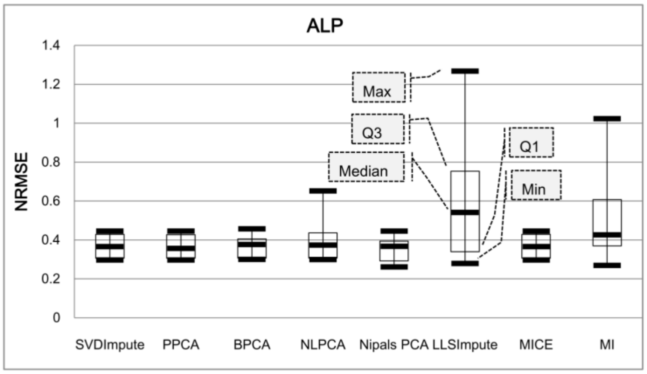


Figure 4. The experiment results based on eight imputation methods for ALP.

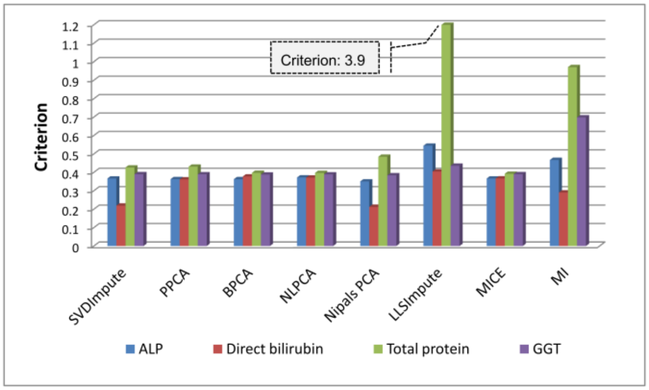


Figure 5. The imputation method selection criterion (averaging Q1, median and Q3) of eight imputation methods for four features.

The further information relevant to eight imputation methods with the parameter settings which have the best rate of top 10 leading performances are shown in Fig. 4 (using ALP as an example). The figure shows the maximum, first quartile (Q1), third quartile (Q3), median, and minimum of 10 NRMSEs in 10 repeated simulation experiments. In Fig. 4, the maximum, and the median in NLPCA, LLSImpute, and MI are larger than those of other imputation methods. Fig. 5 shows the performance of eight imputation methods for four features with missing values. For example, for total protein, the LLSImpute and MI did not perform better than the other six imputation methods.

VII. PERFORMANCE OF A PREDICTIVE MODEL

The performance of predictive models produced using complete data set and imputed data sets are shown in TABLE II. The average of performance in five-fold cross-validation is presented. The complete data set contains 16 clinical features with no missing values. The imputed data set contains 16 features as complete data set and other four features with missing values. These missing values are imputed using eight imputation methods. Therefore, there are eight different data sets named by the imputation methods.

Fig. 6 shows the used frequencies of four features with missing values which selected for building SVM models in five-fold cross-validation.

TABLE II. PERFORMANCE OF PREDICTIVE MODELS PRODUCED BY DIFFERENT DATA SETS

Data Set	Sen	Spe	BAC	Acc	PPV	NPV	AUC
Complete	66.67	85.64	76.16	81.91	68.57	90.12	69.06
SVDImpute	71.67	82.66	77.16	80.59	63.50	91.50	76.00
PPCA	73.33	79.93	76.63	78.31	52.48	91.50	75.14
BPCA	60.00	85.64	72.82	80.66	68.57	88.88	67.90
NLPCA	71.67	79.73	75.70	78.16	53.81	91.67	73.57
Nipals PCA	88.33	71.78	80.06	75.66	49.60	95.96	80.43
LLSImpute	78.33	78.06	78.20	78.16	52.48	92.87	72.70
MICE	66.67	86.23	76.45	81.98	65.33	90.84	66.52
MI	73.33	84.67	79.01	81.99	63.57	91.96	80.28

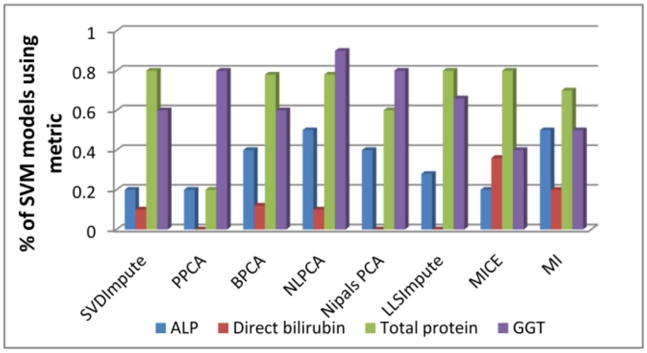


Figure 6. The used frequency of four features for predictive models.

VIII. DISCUSSION

In a study relevant to missing values proposed by Janssen et al., they apply logistic regression to modeling the risk of deep venous thrombosis (DVT). They conduct simple methods for dealing with missing data which will lead to misleading results [2]. Therefore, before building predictive model, the simulation experiments are performed firstly for evaluating the imputation methods in accuracy and stability of estimating missing values in our data sets. Through this simulation, the imputation methods with their parameter settings which are suitable for our data sets can be selected.

The performance of the predictive model with the complete data set is regarded as a reference. Of eight models imputing using different imputation methods, six of them can achieve higher sensitivity than that of complete data set. Especially, the model with Nipals PCA increases about 20% in sensitivity (comparing to the model of complete data set). Although their PPVs are lower than that of complete data set, the model with higher sensitivity can identify more patients with recurrent status. According to the results, the imputation methods reveal the potential of features with missing values in improving sensitivity.

The model with MICE has similar performance with complete data set. In our data set, MICE may not be a suitable method to impute missing values for developing predictive models. The model with BPCA cannot achieve better performance than that of the complete data set. In our data set, the BPCA may not be a suitable method to impute missing values for developing predictive model.

In Fig. 5, the LLSImpute has high imputation selection criterion (not accurate and not stable) in total protein (3.9). MI has high imputation selection criterion in total protein (0.97) and GGT (0.70). However, in Fig. 6, the predictive model with LLSImpute relies heavily on total protein (which appears about four times in five-fold cross-validation). The predictive model with MI relies on total protein and GGT (which appear about 3.5 times and 2.5 times in five-fold cross-validation). Because of above reasons, the reliability of the models with the LLSImpute and MI may not be as good as the models with other imputation methods in our data set. Several limitations of this work are listed in the following content.

Most previous studies concerning patients' recurrence statuses after RFA were focused on risk factors analysis, but not development of predictive models. For example, among these four studies concerning risk factors [7][31][32][33], sample sizes are 118, 124, 190, and 273, respectively. In these study, patients received RFA within specific ranges from four years to five years (e.g., within four years between 2003 and 2007). In our study, 83 patients are collected and they received RFA within a specific range (i.e., within two years between 2007 and 2009). Our sample size is smaller than theirs.

The relationship between patients with missing values and patients without missing values are not further analyzed and discussed in this study. However, we hope these predictive models can also predict patients' statuses when they have missing values. In this study, two ways for handling missing values, complete variable analysis and imputation method, do not remove the patients with missing values. For complete variable analysis, only the features (i.e., variables) with missing values are removed and the number of patients is still 83. For imputation method, features with missing values are reserved and the number of patients is also 83. These missing values are estimated before classification is performed. Therefore, predictive models developed based on these ways can also be used when patients have missing values.

Different feature selection methods may select different feature sets from the same data set which is handled by a specific imputation method. The selection of features may affect predictive models and classification results. In this study, predictive models which built by selected features have better performance than predictive models which built by all features (i.e., not performing this feature selection method). In this study, relationship between imputation methods and feature selection methods is not further discussed.

IX. CONCLUSION

Before imputation methods are employed for estimating missing values during data analysis (e.g., classification), the performance (i.e., accuracy and stability) of imputation methods for a specific data set can be evaluated through the first level evaluation and suitable imputation methods for this data set can be selected. In the second level evaluation, we can not only evaluate the correction of predictive models, but also the reliability of these models. A two-level evaluation method proposed in this study may be applied to other data sets and other predictive targets for finding appropriate imputation methods and providing information of evaluating the reliability of predictive models.

X. FUTURE WORK

In the data set of this study, a clinical feature only has single value. Actually, a clinical feature may have various values which are measured at different time points. Data analysis based on data sets with multiple measurements would be one of our future works. Because SVM can work on higher dimensional feature space, we select this algorithm for finding solutions in different feature spaces. Maybe the comparisons between different algorithms based on this data set can be regarded as one of our future works.

ACKNOWLEDGMENT

The authors acknowledge the members of the research team lead by F Lai of National Taiwan University for their contributions toward this study. F Lai received grants from the National Science Council, Taiwan (NSC 101-2221-E-002-203-MY3).

REFERENCES

[1] A. Burton and D. G. Altman, "Missing covariate data within cancer prognostic studies: a review of current reporting and proposed guidelines," British Journal of Cancer, vol. 91, Jul 5 2004, pp. 4-8.

[2] K. J. M. Janssen, et al., "Missing covariate data in medical research: To impute is better than to ignore," Journal of Clinical Epidemiology, vol. 63, Jul 2010, pp. 721-727.

[3] S. Delepoulle, F. Rousselle, C. Renaud, and P. Preux, "A comparison of two machine learning approaches for Photometric Solids Compression," in Intelligent Computer Graphics 2010, ed: Springer, 2010, pp. 145-164.

[4] R. Jagannathan and S. Petrovic, "Dealing with missing values in a clinical case-based reasoning system," in Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference on, 2009, pp. 120-124.

[5] A. Jemal, et al., "Global cancer statistics," CA Cancer J Clin, vol. 61, Mar-Apr 2011, pp. 69-90.

[6] H. B. El-Serag, "Hepatocellular carcinoma," N Engl J Med, vol. 365, Sep 22 2011, pp. 1118-1127.

[7] W. Y. Kao, et al., "Risk factors for long-term prognosis in hepatocellular carcinoma after radiofrequency ablation therapy: the clinical implication of aspartate aminotransferase-platelet ratio index," European Journal of Gastroenterology & Hepatology, vol. 23, Jun 2011, pp. 528-536.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proceedings of the fifth annual workshop on Computational learning theory, 1992, pp. 144-152.

[9] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, 1995, pp. 273-297.

[10] X. A. Zhao, et al., "A support vector machine (SVM) for predicting preferred treatment position in radiotherapy of patients with breast cancer," Medical Physics, vol. 37, Oct 2010, pp. 5341-5350.

[11] G. Li, et al., "Computer aided diagnosis of fatty liver ultrasonic images based on support vector machine," Conf Proc IEEE Eng Med Biol Soc, vol. 2008, 2008, pp. 4768-4771.

[12] Y. Sela, et al., "fMRI-based hierarchical SVM model for the classification and grading of liver fibrosis," Biomedical Engineering, IEEE Transactions on, vol. 58, 2011, pp. 2574-2581.

[13] J. M. Llovet, C. Bru, and J. Bruix, "Prognosis of hepatocellular carcinoma: the BCLC staging classification," Semin Liver Dis, vol. 19, 1999, pp. 329-338.

[14] J. Selbig, W. Stacklies, H. Redestig, M. Scholz, and D. Walther, "pcaMethods - a bioconductor package providing PCA methods for incomplete data," Bioinformatics, vol. 23, May 1 2007, pp. 1164-1167.

[15] S. Buuren and K. Groothuis-Oudshoorn, "MICE: Multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, 2011, pp. 1-67.

[16] Y.-S. Su, M. Yajima, A. E. Gelman, and J. Hill, "Multiple imputation with diagnostics (mi) in R: opening windows into the black box," Journal of Statistical Software, vol. 45, 2011, pp. 1-31.

[17] O. Troyanskaya, et al., "Missing value estimation methods for DNA microarrays," Bioinformatics, vol. 17, Jun 2001, pp. 520-525.

[18] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," Neural Computation, vol. 11, Feb 15 1999, pp. 443-482.

[19] S. Oba, et al., "A Bayesian missing value estimation method for gene expression profile data," Bioinformatics, vol. 19, Nov 1 2003, pp. 2088-2096.

[20] M. Scholz, F. Kaplan, C. L. Guy, J. Kopka, and J. Selbig, "Non-linear PCA: a missing data approach," Bioinformatics, vol. 21, Oct 15 2005, pp. 3887-3895.

[21] H. Wold, "Estimation of principal components and related models by iterative least squares," Multivariate analysis, vol. 1, 1966, pp. 391-420.

[22] H. Kim, G. H. Golub, and H. Park, "Missing value estimation for DNA microarray gene expression data: local least squares imputation," Bioinformatics, vol. 21, Jan 15 2005, pp. 187-198.

[23] C. C. Chang and C. J. Lin, "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, 2011, pp. 27:21-27:27.

[24] C. W. Hsu, C. C. Chang, and C. J. Lin, "A Practical Guide to Support Vector Classification," 2010.

[25] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," Science, vol. 220, 1983, pp. 671-680.

[26] A. Liaw and M. Wiener, "Classification and regression by randomForest," R News, 2002, pp. 18–22.

[27] Z. Michalewicz, M. Schmidt, M. Michalewicz, and C. Chiriac, Adaptive Business Intelligence: Springer, 2007.

[28] J. Cadima, J. O. Cerdeira, P. D. Silva, and M. Minhoto, "The subselect R package, Version 0.11," 2011.

[29] P. Romanski, "Selecting attributes, Package 'FSelector'," 2012.

[30] E. R. Delong, D. M. Delong, and D. I. Clarkepearson, "Comparing the Areas under 2 or More Correlated Receiver Operating Characteristic Curves - a Nonparametric Approach," Biometrics, vol. 44, Sep 1988, pp. 837-845.

[31] V. W. T. Lam, et al., "Risk factors and prognostic factors of local recurrence after radiofrequency ablation of hepatocellular carcinoma," Journal of the American College of Surgeons, vol. 207, Jul 2008, pp. 20-29.

[32] K. Shiozawa, et al., "Risk factors for the local recurrence of hepatocellular carcinoma after single-session percutaneous radiofrequency ablation with a single electrode insertion," Molecular Medicine Reports, vol. 2, Jan-Feb 2009, pp. 89-95.

[33] B. W. Yang, et al., "Risk factors for recurrence of small hepatocellular carcinoma after long-term follow-up of percutaneous radiofrequency ablation," European Journal of Radiology, vol. 79, Aug 2011, pp. 196-200.

The glymphatic system and Alzheimer's disease: possible connection?

Christina Rose Kyrtsos
Institute for Systems Research
University of Maryland
College Park, MD, US
ckyrtsos@hmc.psu.edu

John S. Baras
Institute for Systems Research
University of Maryland
College Park, MD, USA
baras@umd.edu

Abstract— **Alzheimer's disease (AD) is the most common cause of dementia in the elderly, accounting for 60-80% of all dementias. Symptoms of AD include memory loss (initially short term, later on long term as well), changes in mood and sun-downing. Magnetic resonance imaging (MRI) shows diffuse cortical atrophy and enlargement of the lateral ventricles due to neuronal cell loss. At the cellular level, neuroinflammation, activation of microglia, deposition of beta amyloid (Aβ) in the brain parenchyma and cerebral vasculature, and increases in the permeability of the blood-brain barrier (BBB) have all been observed as well. A newly developed theory purports that the glymphatic system of the brain may play a significant role in AD pathogenesis. This paper presents an initial mathematical model of the glymphatic system and studies how changes of transporter density and deposition of Aβ at the BBB changes clearance of Aβ. Changes in local neuronal cell density were also modeled. This represents one of the first attempts to computationally study the role of the glymphatic system in AD.**

Keywords-Alzheimer's disease, math model, glymphatic system, LRP1, Aβ clearance

I. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia and affects nearly 1 in 6 people over the age of 65 years. The most common symptom associated with AD is the loss of short term memory (memory forming capability), followed by loss of long term memories important to the identity of the patient [7], [8]. Clinically, AD is diagnosed via post-mortem histology, and is identified from the presence of diffuse Aβ plaques throughout the cortices, which are particularly dense in the hippocampi and amygdala. Beta amyloid is generated in response to several known stimuli, including: trauma, inflammation, infection and synaptic remodeling [13], [14], [15]. Enhancements in the capability of imaging modalities have demonstrated that many AD brains have diffuse cortical atrophy that is more severe than that seen for normal aging, as well as enlargement of the lateral ventricles believed to be due to neuronal cell loss. Vascular dysfunction such as cerebral amyloid angiopathy (deposits of beta amyloid in the brain vasculature) and decreased cerebral blood flow have also been observed [11], [12].

Other features seen on histology include an increased number of neurofibrillary tangles (NFTs) and neuroinflammatory markers. Beta amyloid is known to activate microglia and induce the expression of pro-inflammatory cytokines, including interleukin-1 (IL-1) and tumor necrosis factor alpha (TNFα) [1]. This inflammation triggers further production of Aβ and creates a positive feedback loop if the inflammation is not controlled by IL-6 inhibition and Aβ is not cleared effectively. Such a local inflammatory environment is quite neurotoxic and can lead to synaptic dysfunction and neuronal cell death over time.

Recent experimental research has shown that this paravascular space, an area located between the cerebral vasculature and astrocytes, functions as a "lymphatic" system, clearing cerebral waste products via active (membrane transporters) and passive (transport to cervical lymph nodes) transport [5]. Experimental results demonstrated that the majority (approximately 50%) of fluid in the paravascular space was transported via this pathway to the cervical lymph nodes [5]. The remainder of solutes are either cleared by receptors along the venule or at the arachnoid granulations.

Conversely, cerebrospinal fluid (CSF) is believed to drain along cranial nerves, predominantly cranial nerve 1 via the cribiform plate [6]. CSF is mainly generated at the choroid plexus and travels through the ventricular system. Experiments have shown that significant changes in the choroid plexus occur in AD, aging and normal pressure hydrocephalus, and include: flattening and loss of the secretory epithelium; thickening of the basement membrane; cyst formation; lipid accumulation with fibrosis and calcification; hyalination; and especially important to AD pathogenesis, Aβ deposition in choroidal vessels [2]. The exact role of these changes is currently not well understood, but may contribute to abnormal clearance of Aβ from the brain which is studied in more detail in this paper.

This paper studies the role of the glymphatic system in clearance of Aβ from the brain. The generation of CSF and interstitial fluid (ISF), the effect of decreased densities of a key transport receptor, LRP-1 (low density lipoprotein-related receptor protein 1) and the effect of increased Aβ generation rate were studied. Section II will discuss more details on the model background information and it's basis in currently available experimental data, as well as describe derivations of key equations that were used in modeling. Section III will describe the results of the model, while section IV will delve into the discussion of these results and how they relate to current ideas on the pathophysiology of the biological processes occurring in AD. This represents one of the first mathematical models of the brain's glymphatic system and applies it to an important clinical scenario (AD).

ISBN: 978-1-61208-335-3

II. Model background

In this model, the glymphatic system of the brain has been studied with respect to the clearance of Aβ. The brain has been roughly divided into three overlapping compartments: the blood-brain barrier (BBB), the brain parenchyma and the paravascular space that runs between the two. The CSF fills the cerebral ventricles and empties into the dural venous sinuses at the arachnoid granulations. Flow progresses from the choroid plexus inferiorly before splitting and heading either anteriorly towards the frontal lobe, or inferiorly, transversing around the cerebellum and occipital lobes until both pathways converge and drain to the superior sagittal sinus. This is the location of the majority and largest of the arachnoid granulations, sites where the arachnoid delves into the dura and an exchange of CSF from the subarachnoid space to the venous blood occurs.

The CSF serves several roles, including: buoyancy for the brain; acid-base buffering; and delivery of electrolytes, signaling molecules and micronutrients to cells within the brain parenchyma [4]. In addition to these roles, the ISF also has the task of clearing solutes and waste products from the intraparenchymal space. At the cellular level, the abluminal Na+/K+ ATPase is responsible for generating both CSF and ISF; CSF is generated at a rate of about 350-400 µL/min, while ISF is generated at about 0.17 µL/min [3], [4]. CSF is predominantly generated at the choroid plexus by secretory epithelium, while ISF is generated at the BBB. Rate of production is a function of the following:

$$\text{CSF Gen} = \text{f(diuretics, digoxin, } \beta \text{ blockers, longterm CSF } \mathbb{P}\text{, aging)}$$

It is important to note that CSF generation rates do *not* change with acute changes in CSF pressure; however, ISF generation rates may vary with the blood pressure of the involved vasculature and can be modeled according to Starling's law:

$$J_v = L_p S(\Delta P - \sigma \Delta \pi) \qquad (1)$$

where J_v is the net fluid flux, L_p is the hydraulic conductivity, S is the surface area through which fluid is being transferred across, ΔP is the hydrostatic pressure difference, σ is the osmotic reflection coefficient and Δπ is the oncotic pressure difference. The pressure differences (hydrostatic, oncotic) are taken across the BBB; that is, for ISF generation, the pressure gradient is the difference between the cerebral capillary and the interstitial space.

Once the ISF is generated, it flows with a predominantly convective nature as diffusion coefficients are so trivial compared to the fluid flow rate that they can be considered negligible. ISF generation at the BBB by astrocytes occurs at a constant rate k_5 (12 µL/hr) and is also dependent on the amount of CSF that is recycled into the interstitial space as described by the following equation:

$$\frac{dISF_a}{dt} = k_5 + k_2 CSF - d_1 ISF_a \qquad (2)$$

where ISF_a is the amount of ISF produced by astrocytes, k_2 is the rate of CSF recycled into the ISF, and d_1 is the amount degraded at each time step. The total amount of ISF within the brain parenchyma (ISF) is modeled as the sum of that generated by astrocytes in addition to the amount produced by neurons:

$$\frac{dISF}{dt} = ISF_a + k_1 N - d_3\, ISF_b \qquad (3)$$

where k_1 is the rate of ISF generation by neurons and d_3 is the degradation rate of ISF within the brain parenchyma (ISF_b). Beta amyloid is generated by neurons (k_7) and cleared via LRP-1 at the BBB (k_3), at brain endothelial cells at the venule and at the dural venous sinuses as defined by the following:

$$\frac{dAB_b}{dt} = k_7 N + 0.001 AB_{csf} - AB_b \qquad (4)$$

$$\frac{dAB_p}{dt} = 0.5 AB_b - k_3 AB_b LRP1 - AB_p \qquad (5)$$

$$\frac{dAB_{csf}}{dt} = AB_p\,(1 - k_4 LRP_{csf}) \qquad (6)$$

where AB_b is the amount of beta amyloid generated within the brain, AB_p is the total amount beta amyloid, ABCSF is the amount of beta amyloid in the CSF, and k_4 is the amount of Aβ transported to the CSF. LRP-1 levels at the venule and dural venous sinus were varied linearly to study the effect of decreased LRP-1 density on Aβ clearance from the brain. The system of equations was modeled using Matlab using a timestep of 1 day. The total time modeled was 5 years. For the simulations done here, the CSF concentration was held constant (blood pressure effect was not modeled here). The number of neurons was also modeled, dependent on the local concentration of Aβ such that neuron number decreased a specified percent if a level above a specified threshold (>1x10^{10}) was reached. This threshold level was held constant for all simulations such that only very high levels of beta amyloid would lead to neuronal cell death. The model did not take into account neuronal cell injury (and subsequent death) that can occur at lower levels of Aβ that occur for prolonged periods of time.

III. Results

An initial simulation was run to verify that the system was stable (Figure 1). The system of ODEs was simulated using Matlab and modeled over the course of a 5 year timespan. Results showed the absence of any perturbations and the simulation was stable over the time course and the range of rate constants used.

The highest concentration of Aβ is found within the brain, then within the dural venous sinuses and lastly in the CSF. Neuron number was constant over time and was dependent on the concentration of Aβ within the brain, as well as the nutrients from the ISF.

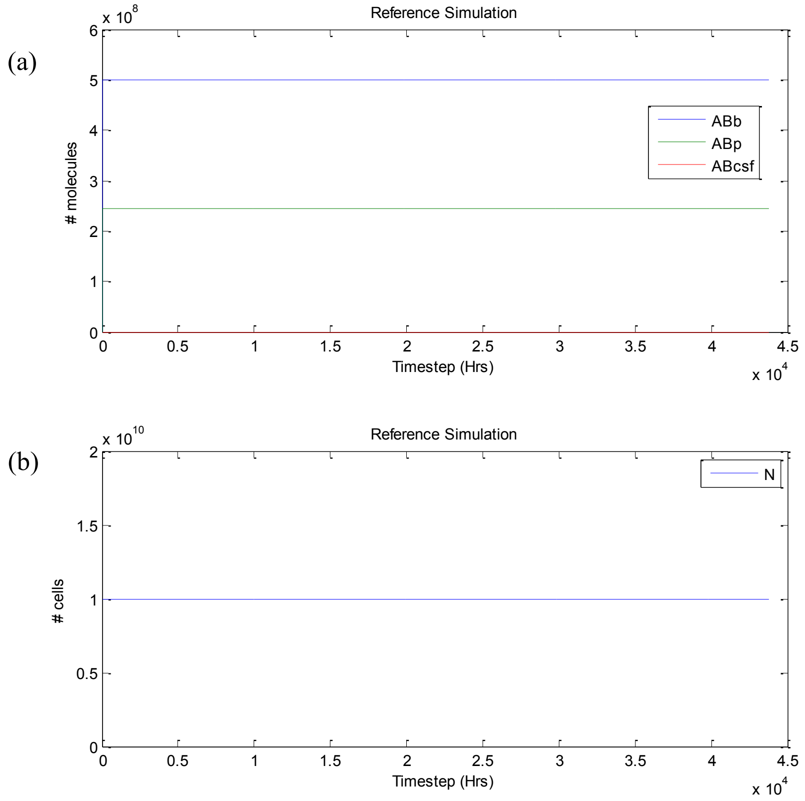


Figure 1: (a) Changes in Aβ levels in the brain, at the dural venous sinus and in the CSF. (b) Neuron number. There are no perturbations to the system.

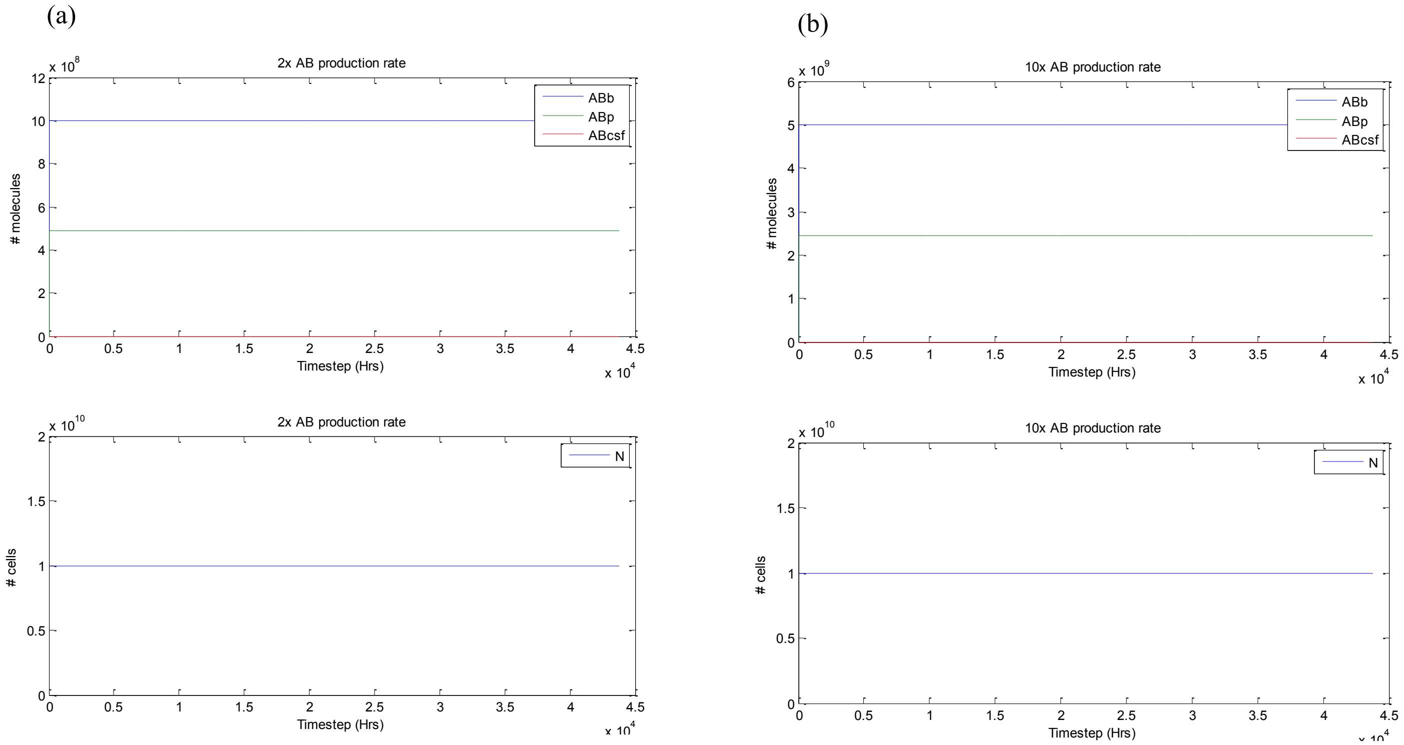


Figure 2: The generation rate of Aβ was increased to (a) 2x and (b) 10x the normal rates. In both cases, neither simulation showed any change in the neuronal density. Only the levels of Aβ increased, demonstrating that having increased generation rate alone is not sufficient to lead to Aβ levels that are high enough to lead to neuronal cell loss.

ISBN: 978-1-61208-335-3

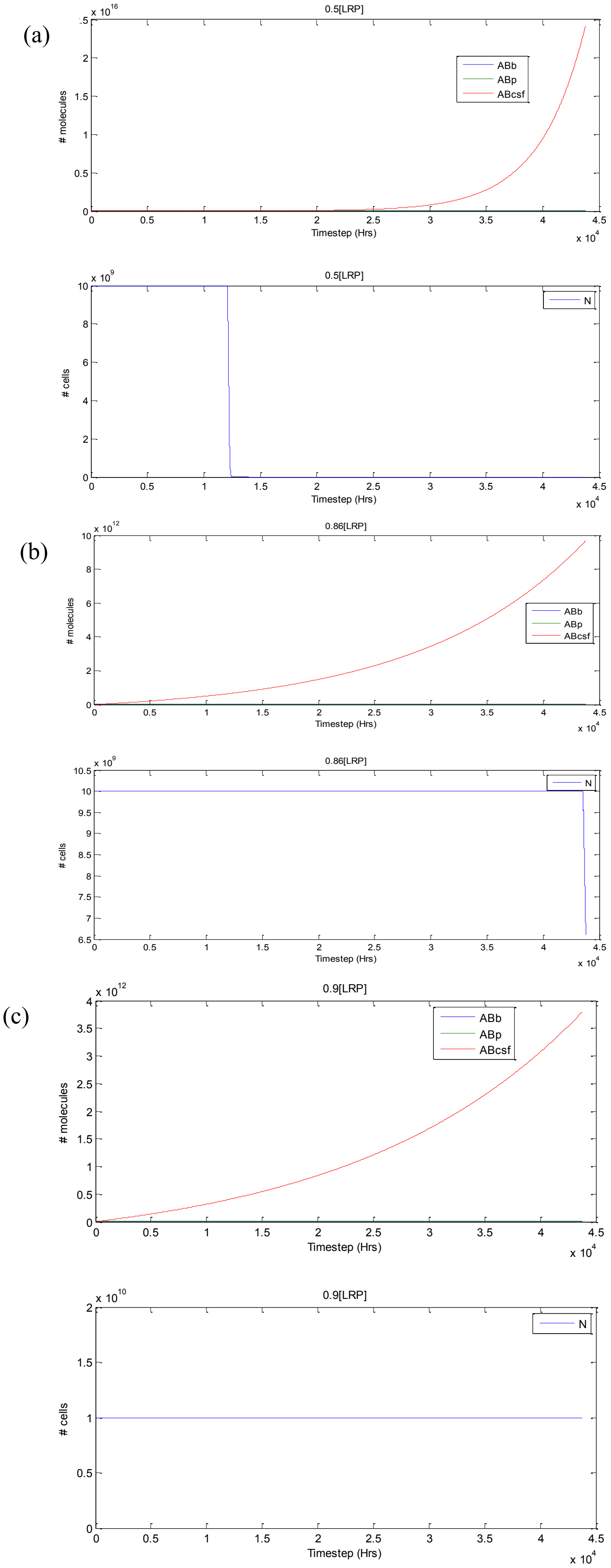


Figure 3: LRP1 levels at the venule and dural venous sinus were varied to study the effect of decreased expression LRP1 expression levels on the level of Aβ within the brain, CSF and paravascular spaces. (a) 50% normal LRP1 expression (b) 86% normal LRP1 expression; this is the transition point between a sustainable decrease in LRP1 and one that leads to neuron loss (c) 90% normal LRP1 expression.

A. Alteration of Aβ generation rates

Alteration of Aβ generation rates by two and ten times the basal rate demonstrated no change in neuron number. Only an increase in the level of Aβ proportionate to the increased generation rate was seen; however, these levels never crossed the threshold to lead to neuronal cell death (Figure 2).

B. Alteration of LRP-1 levels

LRP-1 levels were varied from 50-90% of normal levels to study whether decreased levels had a significant effect on the clearance of Aβ from the brain (Figure 3). By decreasing LRP-1 levels by just 10%, the amount of Aβ within the CSF started to increase exponentially but not to a level that caused neuronal cell death. At 86% of the normal density of LRP-1, there was a late occurring, catastrophic loss of neurons within the local region modeled as Aβ levels climbed above the threshold that lead to neuronal cell death. Decreasing the LRP-1 density further simply shifted the point at which neuronal loss occurred to earlier time points.

IV. Discussion & Conclusion

In this paper, the role of the glymphatic system in the clearance of Aβ has been studied. Reference simulations showed that the basic model developed here was stable over a total time course equivalent to 5 years and for a range of rate constants. Altering the Aβ generation rate demonstrated that although Aβ levels increased in response to the elevated generation rate, levels never reached the threshold where neuronal cell death started to occur. This suggests that an increased Aβ generation rate alone is not sufficient enough to lead to the pathology seen in AD. This conclusion, however, is dependent on the threshold value set for where acute increases in Aβ lead to neuronal cell death. An alternative approach to this situation would be to study how chronic increases in Aβ at lower concentrations can have a negative effect on neuronal cell density.

Decreasing the density of LRP-1 receptors at the venule and dural venous sinuses did, however, have a significant effect on the neuronal cell number and Aβ level. A decrease of just 10% led to an exponential rise in Aβ over the duration of the simulation. This suggests that this clearance method is important in maintaining net balance of Aβ levels within the brain. When the LRP-1 density was further decreased to 86% of the normal levels, neuronal cell death occurred. Decreasing the density further to half of normal levels simply caused neuronal cell death to occur at an earlier time point. Interestingly, Aβ levels continued to rise after neuronal cell death as inflammation was rampant at this point and neurons died, releasing their Aβ into the interstitial fluid to be cleared. This is in accordance with the experimental literature which has shown that decreased

LRP-1 levels are common in AD and may be a significant contributing factor to the buildup of Aβ within the brain [9], [10]. The dramatic changes seen in neuronal cell number with this simulation suggest an important role for LRP-1 at the dural venous sinuses in maintaining normal Aβ levels within the brain, which has been demonstrated experimentally [9].

The mathematical model presented here is one of the first iterations that has been used to study the interaction between the glympahtic system and AD pathogenesis. The model has been developed to study how changes in key steps of clearance can affect the Aβ levels within the brain, and has focused here on the importance of LRP-1 in Aβ clearance. The model here is just a starting point and will be expanded to study the roles that alterations in clearance to the cervical lymph nodes play; how chronic hypertension and hypotension can play a role in decreased Aβ clearance; how Aβ deposition in the cerebral vasculature changes the ability of Aβ to be cleared as well as induces the breakdown of the BBB and subsequently causes a local inflammatory response; and how inflammation can change the clearance of Aβ. This work promises to be important in understanding how Aβ clearance mechanisms, especially how the glymphatic system, plays a significant role in the pathogenesis of AD and could be helpful in developing possible treatment modalities in the future.

References

[1] R. O. Weller, D. Boche, and J.A.R. Nicoll. "Microvasculature changes and cerebral amyloid angiopathy in Alzheimer’s disease and their potential impact on therapy." Acta neuropathologica, 118(1), Feb 2009, pp. 87-102.

[2] E. K. Agyare et al, “ Traffic jam at the blood-brain barrier promotes greater accumulation of Alzheimer’s disease amyloid β proteins in the cerebral vasculature,” Mol Pharmaceutics, 10(5), Dec 2012, pp. 1557-1565.

[3] H. S. Sharma and J. Westman, “ Blood-Spinal Cord and Brain barriers in Health and Disease,” Elsevier Academic Press, 2004, pp. 92.

[4] G. D. Silverberg, M. Mayo, T. Saul, E. Rubenstein, and D. McGuire, “Alzheimer’s disease, normal-pressure hydrocephalus, and senescent changes in CSF circulatory physiology: a hypothesis,” Lancet Neurol, 2(8), 2003, pp. 506-511.

[5] Iliff et al, “A paravascular pathway facilitates CSF flow through brain parenchyma and the clearance of interstitial solutes, including beta amyloid,” Sci. Transl. Med, 4(147), 2012, pp. 1299-1309.

[6] Battal et al, “Cerebrospinal fluid flow imaging using phase contrast MR technique,” Br J Radiol, 84(1004), 2011, pp. 758–765.

[7] R. B. Maccioni, J. P. Munoz, and L. Barbeito, "The molecular bases of Alzheimer's disease and other neurodegenerative disorders." Arch Med Res 32(5), 2001, pp. 367-81.

[8] D. J. Selkoe, "Soluble oligomers of the amyloid beta-protein impair synaptic plasticity and behavior." Behav Brain Res 192(1), 2008, pp. 106-13.

[9] R. A. Fuentealba, Q. Liu, T. Kanekiyo, J. Zhang and G. Bu, “Low-density lipoprotein receptor-related protein 1 (LRP1) promotes anti-apoptotic signaling neurons by activating AKT survival pathway,” J. Biol. Chem. 284, 2009, pp. 34045-34053.

[10] D. E. Kang et al, “Modulation of amyloid β-protein clearance ad Alzheimer’s disease susceptibility by the LDL receptor-related protein pathway.” J. Clin. Invest. 106, 2000, pp. 1159-1166.

[11] B. V. Zlokovic, “Neurovascular pathways to neurodegeneration in Alzheimer’s disease and other disorders.” Nature Rev. Neurosci. 12, 2011, pp. 723-738.

[12] I.V. J. Murray, J.F. Proza, F. Sohrabji, J. M. Lawler, “Vascular and metabolic dysfunction in Alzheimer’s disease: a review.” Exp. Biol. Med 236(7), 2011, pp. 772-782.

[13] N. Marklund et al, “Monitoring of β-amyloid dynamics after human traumatic brain injury.” J. Neurotrauma 31(1), 2014, pp. 42-55.

[14] J. W. Lee et al, “Neuro-inflammation induced by lipopolysaccharide causes cognitive impairment through enhancement of beta-amyloid generation.” J. Neuroinflam. 5(37), 2008.

[15] M. S. Parihar and G. J. Brewer, “Amyloid beta as a modulator of synaptic plasticity.” J. Alzheimers Dis. 22(3), 2010, pp. 741-763.

Acknowledgement

This work was supported by a grant from the Hynet Center at the University of Maryland.

A Framework for Inverse Virtual Screening

Large-Scale Protein Targets Identification

R. Vasseur[1, 2], S. Baud[1], L. A. Steffenel[1], X. Vigouroux[2], L. Martiny[1], M. Krajecki[1], M. Dauchez[1]
1-UFR Sciences Exactes et Naturelles, University of Reims (URCA), Reims, FRANCE
2-Bull SAS, Education & Research, Echirolles, FRANCE

romain.vasseur@etudiant.univ-reims.fr
stephanie.baud@univ-reims.fr
luiz-angelo.steffenel@univ-reims.fr

xavier.vigouroux@bull.net
laurent.martiny@univ-reims.fr
michael.krajecki@univ-reims.fr
manuel.dauchez@univ-reims.fr

***Abstract*—Molecular docking are widely used computational technics that allow studying structure-based interactions complexes between biological objects at the molecular scale. The purpose of the current work is to develop a framework that allows performing inverse virtual screening to test at a large scale a chemical ligand docking on a large dataset of proteins, which has several applications in the field of drug research. We developed different strategies to distribute the docking procedure, as a way to efficiently exploit the computational performance of multi-core and multi-machine (cluster) environments. This tool has been tested on 24 protein-ligand complexes taken from the Kellenberger dataset to show its ability to reproduce experimentally determined structures and binding affinities.**

Keywords—Protein-Ligand docking; inverse docking; ranking methods; distributed computations; HPC experiments.

I. INTRODUCTION

In the field of drug discovery or drug design, molecular docking is focused on protein-ligand complexes to study how the chemical ligand that is a drug will bind the target protein receptor. The prediction of the binding mode of a ligand into a protein target cavity, the structure of the complex and the estimation of the binding affinity between both partners is crucial to find new therapeutic compounds to cure life threatening diseases. Molecular docking represents a virtual alternative to costly and time-consuming systematic wet biological experiments such as High Throughput Screening (HTS) processes and/or Nuclear Magnetic Resonance (NMR)-based screening. Then, it is called Virtual Ligand Screening (VLS) or *in silico* ligand screening and has become a method of choice for rational drug design, hits identification and hits to leads optimization [1][2][3]. At present, several applications are available for virtual screening, such as PLANTS [4], DOCK Blaster [5], GOLD [6], AutoDock [7][8], FlexX [9], Glide HTVS [10], ICM [11] and LigMatch [12].

VLS tries to predict probable bindings of a huge number of ligands (to the order of millions) to a unique target receptor and is linked to multiple ligand dockings. Such methods require knowledge of the three dimensional structure of a receptor alone or associated with its experimental ligand. Many chemical databases and libraries provide millions of compounds, among which we can cite some public and free ones such as the PDBbind database [13] or the ZINC database [14], some with fees access as the Cambridge Structural Database [15] and several private pharmaceutical collections. Protein structures are obtained from the Research Collaboratory for Structural Biology (RCSB) Protein Data Bank (PDB) [16], an open source database that collects all public experimental data on tridimensional biological structures. For a large number of proteins, X-ray crystallography and NMR provide experimental structural data. In November 2013, the number of protein structures publicly available in the Protein Data Bank is over 85,000 the number of nucleic acids structures is about 2,500 and the number of structures of nucleic acids-protein complexes is about 4,000. The total number of structures available in the PDB increased on average by 6,500 structures per year during the last decade [16]. Yet, it is important to highlight that these statistics do not include the large number of proprietary structures as described above held by pharmaceutical companies that dispose of their own private structures databanks. To use non-resolved structures for a protein of interest, 3D prediction models can be built *de novo* [17] or based on partially known fragments by homology modelling [18][19].

The purpose of the current work is to develop a new virtual screening tool that allows performing large-scale structure-based inverse docking. The main idea of this approach is to test at a large scale a chemical ligand on a large dataset of proteins. In the fields of drug design and structural biology, inverse docking methodology would find several applications. It can be used to search for additional uses of new drugs, by searching for interactions with protein groups outside the usual research field. Inverse docking can also be used to identify potential side effects of new drugs or to help choosing the less harmful treatment for a disease. Several problems arise when performing inverse docking, as we are no longer targeting a single protein but thousands. One of the main concerns is the computation time, which represents a clear obstacle when dealing with a large number of different proteins. For instance, even with

ISBN: 978-1-61208-335-3

the use of multicore processing we shall not restrain the inverse docking to a single computer but rely on multiple computational environments such as clusters and grids. In order to effectively use wide computational resources, however, we cannot simply launch a batch of docking computations but we must rethink docking in terms of task distribution, of pipelining, as well as load balance and fault tolerance. Recently, in number of works, several implementations to performed massively parallel ligand screening are reported in the literature with Message Passing Interface (MPI or openMP) only [20] or combined with multi-threading programming [21], with cloud-computing to treat Full Flexible Receptors (FFR) models [22][23] or even with FPGAs or GPUs accelerators [24].

In this work, docking simulations were performed with the AutoDock4.2 software [25] and we developed a set of Python scripts to reverse the docking process. We also developed a Python framework embedding different strategies to distribute the docking procedure, as a way to efficiently exploit the computational performance of multi-core and multi-machine (cluster) environments. Data presented in this paper result from the testing described hereafter. The experiment was conducted to compare the docked poses obtained with our tool for a set of chemical ligands on their experimental target to the determined structure of the complex obtained by X-ray crystallogrophy. The rest of this paper is structured as follows: Section II presents the different strategies we developed to decompose the docking computation, the description of the test set and methods we used to generate and to rank the docking poses. In Section III, docking poses given by these strategies are compared to the native ones (X-ray structures). Finally, all results are afterward discussed in Section IV.

II. Methods

A. Parallel Decomposition

To obtain a better implication of the computational resources, we must imperatively improve task parallelism when conducting large-scale inverse docking. If decomposing a docking job in parallel task may trigger a better utilization of the computational resources through pipelining and load balance, it also contributes to the fault tolerance aspects since only a small segment of the execution is lost in the case of a computer crash or execution failure. For this, we developed two methods to decompose the docking computation and improve tasks distribution and fault tolerance.

The first strategy to distribute docking computations aims at the reduction of the exploring space through the multiplication of the number of small 3D boxes. For instance, the "single grid" used in a blind docking experiment and describing the whole protein volume is arbitrary split into several grids. Each grid is a sub-volume of points covering a piece of the protein. Assuming a regular decomposition, we define a geometrical Arbitrary Cutting method (AC) as 12-part decomposition scheme, i.e., 3x2x2 (3 on the longest axis of the protein). We also tested multiple space cuttings of the whole-space to find a suitable decomposition ratio in prior experiment and the 12-part scheme showed better quality docking results than other geometrical cuttings into multiple subspaces as *n*-part schemes where *n* = 8 (2x2x2), 27 (3x3x3) or 64 (4x4x4) [26]. Indeed, a large number of 3D boxes may improve parallelism but the number of subspaces is also dependent on the size and shape of the protein. So, having too small 3D boxes may limit the movement of the ligand and impact the success of the ligand docking. Hence, the choice of decomposition must be carefully tuned and the number of generated chunks must be precisely balanced. Moreover, the several subspaces are overlapping each other to explore the entire protein surface and overcome the presence of the 3D boxes edges. Indeed, one of the constraints imposed by AutoDock is that the ligand cannot bind outside of the box. The overlapping is inherently dependent on the ligand size, so in our experiments we set two ranges for the partial overlapping: a third of the juxtaposed boxes if the ligand size is inferior to it, or the size of the ligand if the ligand is larger than that.

This decomposition strategy is simple to implement and the subspace grids can be easily generated from the coordinates of the protein. By multiplying the number of 3D boxes we can deploy the docking over different processors in order to be computed in parallel. One drawback of this strategy, however, is that it does not check the protein surface for cavities (which are potential docking sites), and may therefore "cut" right in the middle of a potential cavity, making it less interesting. Another drawback of this method is that only ligands inside the grid can be evaluated. Indeed, any atom of the ligand outside the 3D box will not be treated and will eliminate the pose of the conformer during the sampling process, which may prevent the detection of potential bindings when part of the ligand crosses the boundaries of the 3D box. So, to overcome boundaries problems, we also use a more rational knowledge-based method.

This second method to perform space cutting consists in predicting upstream pockets and cavities on the surface receptor with additional programs and carry out dockings only on these pockets [27][28]. For this Pocket Search method (PS), we used the Fpocket program [29] that screens pockets and cavities using a geometrical algorithm based on Voronoï tessellations. The second version of the software (Fpocket2) is compatible with a multiprocessing parallel use. Only pockets that show a long side superior to a third of the whole protein longest side and inferior to the half of the whole protein longest side are conserved as to limit the number of generated jobs and to avoid multi-exploration of the same space. One advantage of the pocket strategy is to refocuse the docking algorithm exploration zones only on predictive biological sites of interest (potential binding sites). As only these interesting zones are included in the docking procedure they can drastically improve the overall inverse docking performance. At the opposite side, the pocket search is a predictive method and as such it may exclude some potential zones, which should not be overpassed by the AC method described above.

B. *Preparation of the Test Set*

The test set used in this study is constructed from the Rognan's group [30] set of 100 protein-ligand complexes. To be able to perform accurate High Definition (HD) docking only proteins structures with a long side inferior to 60 Angstroms are conserved. Twenty-four complexes have passed this process and are included in the final test set (see TABLE I). Molecular weights of ligand molecules range from 114 to 659 Daltons, number of atoms in the ligand range from 10 to 52 and number of rotatable single bonds (rotors) in ligand molecules range from 0 to 23. All ligands molecules bind to their target protein non-covalently. Structures files and coordinates of all the complexes are downloaded from the Structural Chemogenomics Group website [30]. For the convenience of computation, each complex file was split into a protein molecule file in PDB format and a ligand molecule file, which is saved in Mol2 format. All preparation settings are available in the work from Kellenberger *et al.* [31]. The program automatically generates all docking parameters files and each complex is then subjected to an exhaustive conformational sampling procedure with AutoDock.

C. *Conformationnal Sampling Procedure*

The AutoDock program (version 4.2) is used to generate an ensemble of docked conformations for each ligand molecule. This program utilizes a Lamarckian Genetic Algorithm (LGA) for conformational sampling [32]. Each LGA run outputs a single docked conformation as a final result. For the AC method and the PS method 50 individual LGA runs are performed to generate 50 docked conformations for each ligand. All AutoDock docking experiments were performed with the default parameters of the Lamarckian algorithm for initial population size (*ga_pop_size = 150*), maximal number of energy evaluation (*ga_num_evals = 2500000*) and maximal number of generations (*ga_num_generations = 27000).* The protein structure is kept fixed during docking.

D. *Ranking the Best Ligand Pose*

AutoDock needs to compute an affinity grid for each atomic type to pre-evaluate the binding energy. The affinity grid is contained in a 3D box that frames the protein surface. The binding energy is evaluated with a tri-linear interpolation of the eight-grid points affinity value surrounding each atom of the ligand. For the scoring step, computation time will only depend of the number of atoms in the ligand and will be independent of the protein volume. The free energy of binding ΔG is computed with the AutoDock4 scoring function (AD4) [33]. The AD4 scoring function is composed by several energy terms of classical physics force fields. The free energy of bonding is expressed by the sum of molecular mechanics components such as a dispersion-repulsion term, a term for the hydrogen bonding, a term for the electrostatics contribution, a term descripting the energy associated to bond lengths, bond angles and associated restriction entropy loss and a term for the desolvation energy (equation (1)).

$$\Delta G = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{tor} + \Delta G_{solv} \quad (1)$$

TABLE I. THE 24 EXPERIMENTAL PROTEIN-LIGAND COMPLEXES

PDB code	Res. (Å)	Protein	Ligand
1azm	2.0	Carbonic Anhydrase I	5-Acetamido-1,3,4-Thiadiazole-2-Sulfonamide
1cbs	1.8	Cellular Retinoic-Acid-Binding Protein Type II	Retinoic Acid
1ebp	2.1	Epididymal retinoic acid binding protein	Retinoic Acid
1fkg	2.0	Fk506 Binding Protein	(1R)-1,3-Diphenyl-1-Propyl(2S)-1-(3,3-Dimethyl-1,2-Dioxopentyl)-2-Piperidinecarboxylate (Rotamase Inhibitor)
1fki	2.2	Fk506 Binding Protein	(21S)-1-Aza-4,4-Dimethyl-6,19-Dioxa-2,3,7,20-Tetraoxobicyclo Pentacosane
1glp	1.9	Glutathione S-Transferase Yfyf	Glutathione Sulfonic Acid
1glq	1.8	Glutathione S-Transferase Yfyf	S-(P-Nitrobenzyl) Glutathione
1hfc	1.5	Fibroblast Collagenase	(N-(2-Hydroxymatemethylene-4-Methyl-Pentoyl)Phenylalanyl)Methyl Amine
1icn	1.7	Intestinal Fatty Acid Binding Protein	Oleate (Oleic Acid)
1lic	1.6	Adipocyte Lipid-Binding Protein	Hexadecanesulfonic Acid
1lmo	1.8	Mucopeptide N-Acetylmuramylhydrolase	Di-N-Acetylglucosamine
1mcr	2.7	Immunoglobulin delta Light Chain Dimer	N-Acetyl-L-His-D-Pro-Oh
1mmq	1.9	Matrilysin	Hydroxamate Inhibitor
1mup	2.4	Major Urinary Protein Complex	2-(Sec-Butyl) Thiazoline
1nco	1.8	Holo-Neocarzinostatin	Apo-Carzinostatin chromophore
1poc	2.0	Phospholipase A2	1-O-Octyl-2-Heptylphosphonyl-SN-Glycero-3-Phosphoenolamine
1rob	1.6	Ribonuclease A	Cytidylic Acid
1srj	1.8	Streptavidin	Naphthyl-Haba
1stp	2.6	Streptavidin	Biotin
1tng	1.8	Trypsin	Aminomethylcyclohexane
1tnl	1.9	Trypsin	Tranylcypromine
1ukz	1.9	Uridylate Kinase	Adenosine-5'-Diphosphate
3ptb	1.7	*beta*-Trypsin	Benzyldiamine
8gch	1.6	*gamma*-Chymotrypsin	Gly-Ala-Trp (peptide)

The best ligand poses obtained by AC and PS methods are discriminated using the best energy of binding for each method with the AD4 function. In addition, the localization of best energy docked poses is compared to the experimental pose with the measurement of the Euclidian Distance (ED) between the two ligands geometrical mass centers. When ligands are in the same binding cavity as the experimental one and the ED is lower than 2.5 Angstroms, the ligand pose is considered similar to the crystallographic pose and is called X-pose. When ligands are partially docked in the experimental cavity or able to dock in a juxtaposed cavity and ED is included between 2.5 and 8.5 Angstroms, the ligand pose is called J-pose (for Juxtaposed-pose). Beyond this value, we checked that any ligand is localized in another binding area than the experimental structure. In this case, the wrong ligand pose is called W-pose. (All of these ligands poses were checked by hand and visualized with VMD [34]). Thus, ligand pair Root Mean Square Deviation (RMSD) computation evaluates the shift between the binding conformation of the best-docked ligands and the crystallographic conformation. The RMSD corresponds to the measure of the average distance between atomic positions of two structures expressed in Angstroms as it shows in equation (2).

(2)

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2+(v_{iy}-w_{iy})^2+(v_{iz}-w_{iz})^2)}$$

III. Results

As described above, our methodology was tested on 24 experimental protein-ligand complexes available in the PDB. Both AC and PS methods were used individually and in a combined procedure to evaluate their ability to re-dock an experimental ligand on its native protein target receptor.

For the PS method, the experiment shows that for this size of proteins (see II.A), the Fpocket algorithm found at the most five or six different well-sized pockets. TABLE II gives the volume of the three first pockets found for each experimental complex. For all proteins of the set (100%), one pocket at least is detected, for nineteen proteins in the set (19/24, 79%) two pockets are detected and for 14/24 (58%) three pockets are detected. If structures displaying at least 4 pockets are selected, the ratio of the set falls down to 9/24 (37.5%) and decreases even more when considering a higher number of pockets. Thus, it appears that for each protein-ligand complex selecting only the first pocket found by the Fpocket algorithm is enough to consider the whole set; the results point that selecting at most the three first pockets should refine the search. In addition, the number of jobs launched partly depends on the number of pockets that will be explored. Thus, the number of jobs launched is precisely defined for each complex.

A fixed number of jobs can be very interesting to monitor the speed-up and the scalability of the program over a variant number of available cores. In theory, the optimal load balance should be reached if the number of available cores is superior or equal to the number of launched jobs. So, to optimize the computation time we should set the best ratio jobs/cores and to do this a fixed number of jobs is necessary. For example, this set of complexes generates a pool of maximum 360 jobs (24 complexes x (12 AC method boxes + 3 pockets boxes from the PS method at the most)). So, the best energy structure of the ensemble of the twelve boxes is conserved for the AC method and the best energy structure of each of the first three pockets is conserved for the PS method. Finally, four docked poses at the most are obtained for each complex, which will be compared with the experimental ligand pose of the crystallographic ligand-protein complex. Previously, we define that the re-docking is successful if an X-pose or a J-pose were obtained for the ligand (see II.D).

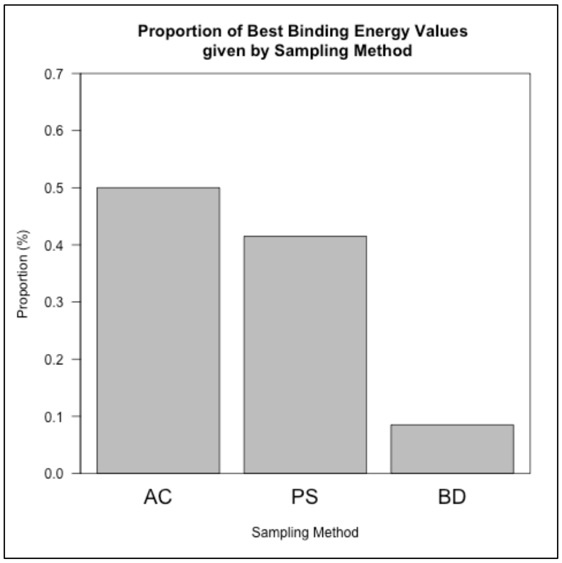


Figure 1. Proportion of Best Binding Energy Values given by the Sampling Method (AC: 62.5%, PS: 29%, BD: 8.5%).

Firstly, the results for AC and PS methods are compared with the corresponding Blind Docking experiment (BD). Blind Docking was introduced to detect possible binding sites and ligands binding modes by scanning the entire surface of protein targets [35][36]. This represents the “naïve” approach to dock ligands on unknown targets but is barely parallelizable. In fact, for each complex the AutoDock software will launch only one infrangible docking task with the whole volume to explore. Depending on the shape of each receptor, a large number of runs/generations is required in order to systematically cover the entire protein surface and consequently to obtain good docking results.

For twelve experiments out of the set (12/24, 50%) the best energy score was obtained by the PS method, for 10/24 (41.5%) it was obtained by AC method and only for 2/24 (8.5%), it was obtained by BD experiment (Figure 1). Moreover, the combined results of AC method and PS method give a better energy of docking for 22/24 (91.5%) compared to BD. Furthermore, for 54% of the cases the combined methods gave a RMSD between the experimental structure and the best docking pose lower than 5 Angstroms and a RMSD lower than 10 Angstroms for 23/24 (96%) versus only one for BD (4%) in both case (TABLE III). These results highlight that our methods perform better exploration of the protein surface. Indeed, the ratio (volume/number of runs) explored in the case of our methodology is better optimized than in the case of BD. Both methods ensure a better conformational sampling and a better quality of docking pose than using the BD.

TABLE II. NUMBER OF POCKETS DETECTED FOR EACH PROTEIN AND THEIR VOLUMES

	Pocket 1 (PS1)	**Pocket 2 (PS2)**	**Pocket 3 (PS3)**
PDB	***Volume (Å^3)***	***Volume (Å^3)***	***Volume (Å^3)***
1azm	833	786	244
1cbs	1626	378	557
1ebp	1262	370	616
1fkg	549	**N/A**	**N/A**
1fki	576	756	**N/A**
1glp	1307	370	640
1glq	607	637	686
1hfc	762	683	485
1icn	1655	**N/A**	**N/A**
1lic	978	927	**N/A**
1lmo	1306	143	561
1mcr	676	192	**N/A**
1mmq	409	276	548
1mup	479	583	756
1nco	350	**N/A**	**N/A**
1poc	1016	504	642
1rob	654	576	686
1srj	408	**N/A**	**N/A**
1stp	367	**N/A**	**N/A**
1tng	647	610	**N/A**
1tnl	602	466	512
1ukz	600	1072	**N/A**
3ptb	549	328	529
8gch	765	619	383

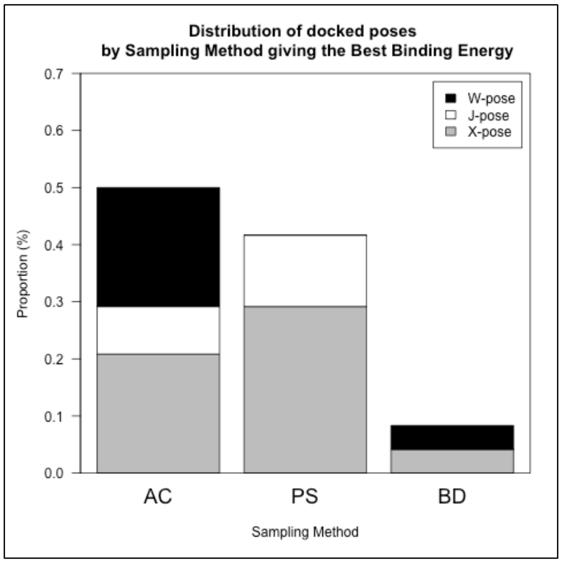


Figure 2. Distribution of docked poses (X-pose in grey, J-pose in white and W-pose in black) by Sampling Method giving the Best Binding Energy.

The distribution of docked poses depending on the sampling method associated with the best energy is presented in Figure 2. For 18/24 (75%) the sample methods that give the best free energy of binding give also the best docking poses (X-pose or J-pose) distributed as follows: 7/18 (39%) for AC method and 10/18 (55%) for the PS method and 1/18 (6%) for the BD experiment. Among these complexes, the combined method that gives the best free energy of binding gives also the best docking pose for 17 (94.5%) versus only one for BD (5.5%). From TABLE III, we can extract the following correlation: comparing the docked poses at rank 1 of Euclidian distance and rank 1 for the lowest RMSD value, there is a match for 6/24 (25%) in the case where a J-pose is observed and for 14/24 (58.5%) in the case where an X-pose is observed. So, at rank 1 for the two previous criteria, the ligand docked poses (X-poses and J-poses) give the lowest RMSD value for 18/24 (75%). Comparing the docked poses at rank 1 of Euclidian distance and rank 1 and 2 for the lowest RMSD value the proportion reach 22/24 (91.5%). The match ratio is distributed by sampling method as follows: The AC method gives the X-pose for 3 complexes with a mean RMSD value equal to 2.32 Angstroms compared to the experimental structures (1glp, 1mup, 1tnl). The AC method gives also a J-pose for 3 complexes (1hfc, 1icn, 1rob) and an associated RMSD value equal to 7.82 Angstroms compared to the experimental structures. Nevertheless, it is important to mention that for 1hfc and 1icn poses are reverse poses that is to say the ligand acquires a head to tail conformation compared to the experimental one so the RMSD increases. The PS method gives the X-pose for 11 complexes (1azm, 1cbs, 1ebp, 1fkg, 1fki, 1mmq, 1nco, 1stp, 1tng, 3ptb, 8gch). In these cases, the mean RMSD with the experimental structure is 2.93 Angstroms. The PS method gives a J-pose for 4 complexes (1lic, 1mcr, 1poc, 1ukz) and an associated mean RMSD value with the experimental structure of 5.54 Angstroms (Figure 3). The BD method gives an X-pose for 1srj with a RMSD value of 2.47 Angstroms. If the rank 2 for the Euclidian distance is also considered, the PS method is able to replace the ligand for 1srj in an X-pose with 2.35

Angstroms of RMSD. So, the combined method with these evaluation criterions gives the best pose for 22/24, 91.5% of the cases of the total set.

TABLE III. EVALUATION CRITERIONS OF THE SAMPLING METHODS

	Energy (kcal/mol)		**RMSD (Angstroms)**				**Gravity Centers Euclidian Distance (Angstroms)**					
	Rank 1		***Rank 1***		***Rank 2***		***Rank 1***			***Rank 2***		
PDB	*Method*	*Value*	*Method*	*Value*	*Method*	*Value*	*Method*	*Value*	*Pose*	*Method*	*Value*	*Pose*
1azm	AC	-5.15	PS1	1.95	**N/A**	**N/A**	PS1	1.12	X-pose	**N/A**	**N/A**	**N/A**
1cbs	PS2	-6.84	PS2	2.24	AC	8.86	PS2	1.17	X-pose	PS1	1.59	X-pose
1ebp	PS2	-8.68	PS2	2.00	AC	2.73	PS2	0.77	X-pose	PS1	1.23	X-pose
1fkg	PS1	-5.96	PS1	5.49	AC	8.22	PS1	1.43	X-pose	AC	3.98	J-pose
1fki	PS1	-10.49	PS1	0.60	PS2	1.75	PS1	0.59	X-pose	PS2	1.00	X-pose
1glp	AC	-4.46	AC	2.71	PS1	5.42	AC	0.76	X-pose	PS1	2.74	X-pose
1glq	BD	-3.66	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
1hfc	AC	-4.78	AC	8.75	**N/A**	**N/A**	AC	5.43	J-pose	**N/A**	**N/A**	**N/A**
1icn	AC	-3.97	AC	8.80	**N/A**	**N/A**	PS1	3.49	J-pose	AC	3.66	J-pose
1lic	PS1	-4.65	PS1	5.75	**N/A**	**N/A**	PS1	3.63	J-pose	AC	4.23	J-pose
1lmo	AC	-3.26	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
1mcr	AC	-4.03	PS2	4.41	**N/A**	**N/A**	PS2	2.81	J-pose	**N/A**	**N/A**	**N/A**
1mmq	AC	-6.31	AC	3.97	PS1	4.16	PS1	0.79	X-pose	AC	1.59	X-pose
1mup	AC	-4.23	AC	2.59	PS1	4.04	AC	1.55	X-pose	PS1	2.02	X-pose
1nco	PS1	-7.19	PS1	7.83	**N/A**	**N/A**	PS1	2.10	X-pose	AC	8.22	J-pose
1poc	PS1	-1.91	PS1	6.71	**N/A**	**N/A**	PS1	3.95	J-pose	**N/A**	**N/A**	**N/A**
1rob	PS2	-5.29	AC	5.91	PS1	9.89	AC	5.32	J-pose	PS2	8.05	J-pose
1srj	BD	-7.48	PS1	2.35	BD	2.47	BD	0.45	X-pose	PS1	1.23	X-pose
1stp	PS1	-6.10	PS1	1.34	AC	2.42	PS1	0.37	X-pose	AC	0.55	X-pose
1tng	PS1	-5.87	PS1	1.05	AC	1.53	PS1	0.63	X-pose	AC	0.83	X-pose
1tnl	AC	-5.96	AC	1.68	PS1	2.44	AC	0.35	X-pose	PS1	0.41	X-pose
1ukz	AC	-6.74	PS1	5.31	**N/A**	**N/A**	PS1	3.39	J-pose	**N/A**	**N/A**	**N/A**
3ptb	AC	-5.52	PS1	1.52	AC	2.07	PS1	0.19	X-pose	AC	0.23	X-pose
8gch	AC	-5.00	PS1	4.32	**N/A**	**N/A**	PS1	1.03	X-pose	**N/A**	**N/A**	**N/A**

a. **N/A: Non Applicable** data – RMSD or Euclidian Distance > 10 Angstroms

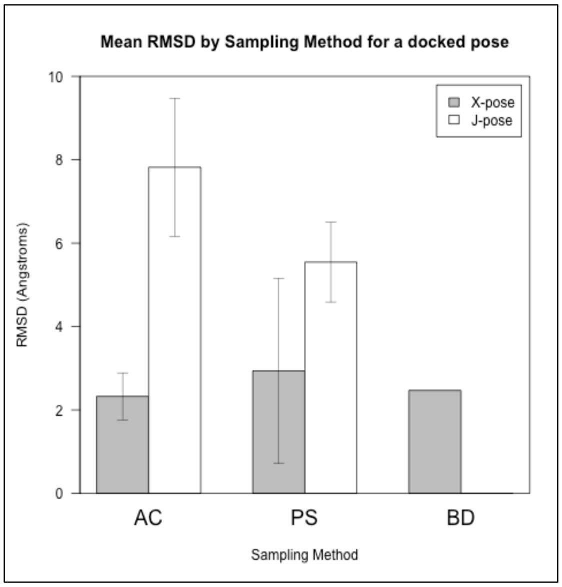


Figure 3. Mean RMSD (in Angstroms) for an X-pose (in grey) and a J-pose (in white) by Sampling Method at rank 1 of Euclidian distance and rank 1 and 2 of RMSD.

Figure 4 shows the results obtained with the AC method for the experimental complex 1stp. An X-pose with an Euclidian distance between ligands geometrical mass centers of 0.55 Angstroms (rank 1) with a RMSD value of 2.42 Angstroms (rank 1) is observed. As we can see on Figure 4 and Figure 5 with two different types of protein representations, the re-docked ligand reached successfully the experimental cavity of binding and adopts a similar conformation compared to X-ray structure. On Figure 4, the New Cartoon style represents only the secondary structure of the backbone skeleton of the protein whereas on Figure 5, all amino-acids side chains are included to build the protein surface thanks to the MSMS algorithm. The local structure of side chains creates reliefs and since some of them display specific chemical properties, they can arrange themselves in binding cavities. The ligand pose and conformation in the binding site will be related to the cavity geometry. As we can see in Figure 4, a good ligand pose implies a chemical conformation that precisely place the chemical groups implied in Hydrogen bonds in an appropriate range of distance (around 2.0 Angstroms). Hydrogen bonds are strong dipole-dipole interactions between electro-negative atoms, and according to local chemical composing they are partially in charge of ligand docking in a binding pocket. For ligands from seven complexes, there is a match between RMSD and mass centers distance but not between both and the best binding energy. In all cases the pose giving the best energy is localized in different from cavities that the crystallographic ones. These results can be explained by several settings of using decomposing method (Figure 5). For 1azm, the best energy is obtained with the box-11 of the AC method (-5.15 kcal/mol) whereas best RMSD with an X-pose is obtained by the PS method (PS1). The AC pose is localized in a different cavity from the crystallographic one. The box-11 dimensions do not allow to include the crystallographic area and they do not permit to refind the experimental pose. On the other side, the PS1 box dimensions do not allow to refind the AC pose cavity neither. The experimental cavity (S1) is included in an another AC box, box-7. The ligand pose obtained with this box is localized in the same cavity as the previous AC box (S2) and presents a better energy than PS1 pose. If we set the dimensions of a tuned box able to include the two binding sites S1 and S2, the ligand pose obtained binds into S1 with even better energy of -5.26 kcal/mol. Finally, to maximise the number of energy evaluations and the conformationnal sampling, we carried out a 256 runs on the previous tuned box and anew the crystallographic cavity is obtained with a poorer energy compared to S1 of -4.49 kcal/mol. So, just the box boundaries presence is not enough to conclude, 1azm complex may wrong prepared or this case shows the limits of the AutoDock force-field.

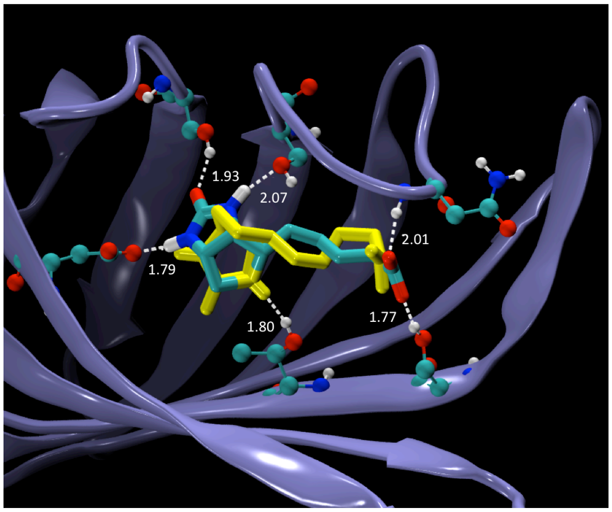


Figure 4. 1STP -- Streptavidin (New Cartoon, in purple)/Biotin (Licorice, X-ray in cyan, X-pose in yellow) protein-ligand complex stabilized by hydrogen bonds in the binding site.

Crystallographic pose refinding may be precluded by boxes boundaries but it is also impacted by protein shape specifications. In fact, for 1ukz, the cavity is closer to a funnel with a long and slight pipe that sinks into the protein structure. The experimental ligand is housed at the bottom of the pipe in a burried area in the protein core. Fpocket detects the left large extremity as part as a full binding pocket (PS2) and the hidden area as an another binding pocket (PS1). The AC box (giving the best energy) only takes in the funnel cavity and does not include the burried site (like PS2 does) and inversely PS1 includes the crystallographic cavity but does not take in the large surface cavity. It explains why there is no match between the AC method that gives the better energy and the PS1 X-pose.

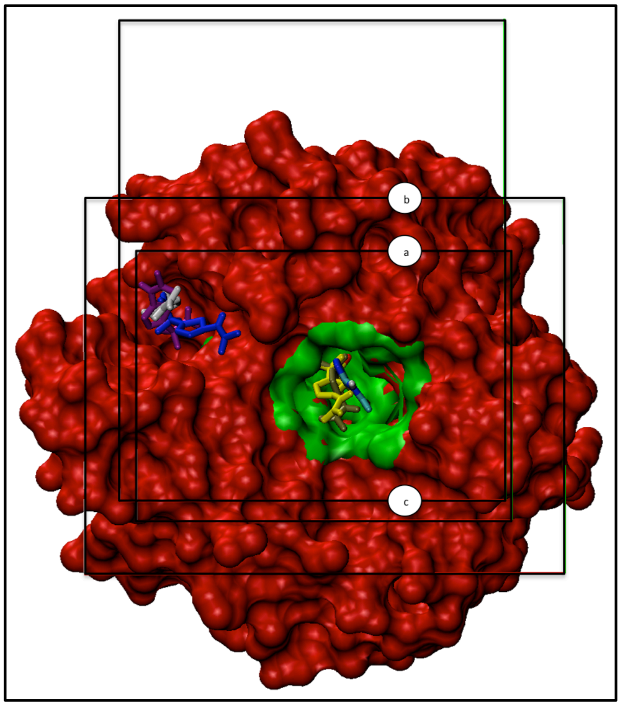


Figure 5. 1azm ligands in the crystallographic cavity (MSMS, in green): X-ray pose (in cyan), PS1 pose (in tan), Tuned box-256R pose (in yellow) and 1azm ligands in another cavity: AC box-11 pose (in blue), AC box-7 pose (in purple), Tuned box-50R pose (in purple) bound on the whole Carbonic Anhydrase I protein receptor with a: PS1 pocket box, b: Tuned box, c: AC box-7

Failed dockings can be explained by protein shape specifications but also by ligand chemical structure. Some ligands as 1lmo or 1rob are very exposed in large valleys at the protein surface, which are correctly identified by the Fpocket program as a binding pocket but the docking program could fail to place correctly the ligand on a planar surface. Else, the chemical nature of ligands could increase the docking process weakness: 1lmo ligand is a big flexible di-saccharide and 1rob ligand is an ADN nucleoside both containing –ose residues hard to treat with the Autodock force field.

Only for 1glq in the test set, the best energy value is given by the blind docking experiment (-3.66 kcal/mol). The ligand pose is neither in the crystallographic cavity neither in any pocket cavity and binds on a relative open cavity. However, 1glq and 1glp are two crystallographic structures of the same protein with about the same degree of resolution complexed with two similar ligands (see TABLE I). For all that, Fpocket is not able to find precisely the same pockets in 1glq so the boundaries are not exactly at the same place and do not allow to retrieve the experimental pose with the PS method. The AC method does but the energy of binding is worse than for the pose obtained by the blind docking. Nevertheless, if we launch multiple blind docking experiments, this artefact binding mode should not be retrieved several times.

1stp and 1srj are two crystallographic structures of the same protein (see TABLE I) with a large variation in resolution neatness. In fact, if a structure with a higher resolution than 2.0 Angstroms is available it is assumed that a structure with a lower resolution degree is a worse structure. In this case the two structures do not show remarkable difference of structure of the binding site. For the binding site in 1stp, the protein-ligand complex is the well known Streptavidin/Biotin complex in which the protein have a β-barrel secondary structure. This complex is one of a strongest non-covalent interactions known in nature. It is used extensively in molecular biology as a marker. The ligand fits perfectly in the binding site and the interactions are stabilized through a complex network of Hygrogens bonds. For 1stp, the experimental ligand was well replaced by the PS1 method (0.37 Angstroms) and AC method (0.55 Angstroms) with the best binding energy equal to -6.10 kcal/mol for PS1. For 1srj the ligand is Naphtyl-Haba docked in the same cavity as Biotin. It was well re-placed by the blind docking experiment (0.45 Angstroms) and PS1 method (1.23 Angstroms) with the best binding energy equal to -7.48 kcal/mol for BD. This results could be explained by the asymetric shape of the protein that confers a geometry less spherical than a regular globular protein. Consequently, the long axis of the protein takes a high value and imposes the same grid spacing as the others proteins. But in this case, the surface to explore included in the blind docking box is less important and the majority of the grid points are not on the protein surface. So, the ratio volume/runs is very high and the algorithm explore much more precisely the binding pocket and leads to the best energy pose with the maximum goodness.

IV. Discussion/Conclusion

In order to be able to treat many hundred proteins computations on High Performance Computing (HPC) architectures, we developed a set of methods to parallelize the treatment of each protein, as well as to distribute the tasks among a given set of machines as a way to speed up the overall execution of the inverse docking. For this, we developed a framework that can embed the AC and the PS method to explore as best as possible the protein surface and rationally dock the ligand into the binding cavity.

Our results show that the methods we are developing perform better volume exploration with a better ratio volume/runs than a classical blind docking experiment. In fact, to perform an accurate high definition docking we have to deal with coherent grid spacing. By default, AutoDock builds affinity grids with a spacing of 0.375 Angstroms that corresponds to a quarter of the bond between two atoms of Carbone. We defined a spacing interval between 0.375 and 0.450 in which we consider the accuracy of the simulation as a HD docking. The main drawback of this method is that AutoDock is able to build and also explore a 3D box of 126 x 126 x 126 points at the most. So, only a protein whose long axis is lower than 60 Angstroms can fit into the grid box.

Considering this kind of protein for a blind docking experiment, the AutoDock program is also limited in the number of simulations runs, that is to say in the number of times the initial LGA is reinterred (256 runs max.). So, AC method considers the BD box volume cut into 12 sub-boxes with a partial overlap. Each sub-box is explored by the LGA with 50 runs of simulations that roughly correspond to the half of the ratio volume/runs for the BD. Whereas the ratio is more difficult to precisely estimate, it is even better with PS method, which explains the effectiveness of the program to perform better exploration and to obtain better docking quality results than BD experiments.

As many docking programs [37][38], we have shown that our framework is a successful tool to re-place correctly the ligand into the active site of the target receptor in a non-covalent manner. Furthermore, it is also able to predict accurate ligands bindings independently of active site knowledge [39]. For this, we evaluated a good docking pose using three criteria: free energy of binding, Euclidian distance between mass centers and RMSD of the re-docked ligand with respect to the crystallographic ligand. Combinations of these criteria are able to discriminate right docking poses from experimental data. The combination between the binding energy and the RMSD (rank1 and 2) is able to discriminate 66.5% of the test set and the one between the mass center distance (rank1 and 2) and the RMSD (rank1 and 2) is able to discriminate 91.5% of the test set. On the other side, the ratio is 75% for the combination of binding energy and center of mass distances (rank1 and 2) and 71% for the combination of the triad. This is explained by the nature of the evaluation criteria. RMSD and mass centers distances are implicitly correlated because they both describe a space position. Mass centers distances describe a space position for the entire ligand whereas RMSD describe a space position for each atom of the ligand, both always in respect to the experimental structure. In fact the RMSD reflects the ligand structure in a local environment, its capacity to adapt itself to the binding cavity. Consequently, taking into account the numbers of atoms implied both in the binding site and in the ligand structure and the number of torsions available for the ligand, the probability to obtain a low RMSD in a different cavity than the crystallographic binding site is close to zero. This is well shown in TABLE III, for 8 cases out of 9 if the RMSD is higher than 10 Angstroms the corresponding mass center distance is higher than 10 Angstroms too (N/A data). That explains the good ratio for these criteria combination. On the other side, the space position adopted by the ligand in the binding site translated by the RMSD value impacts the chemical match between chemical groups able to make non-covalent interactions (Hydrogen bonds, van der Waals forces and electrostatics) with atoms in the binding cavity. These forces represent a major contribution into the energy function that is used to evaluate the free energy of binding (see II.D). So, the ratio of the combination of RMSD and energy of binding can be explained partially by this relationship.

Nonetheless, in this experiment we have shown that we reproduce ligands experimental poses with our framework. As the references are experimental data, we dispose of comparison elements (RMSD and mass centers distances). The results obtained in this study (distances determining X-pose and J-pose and associated RMSD) are good enough to validate the method for detecting workable binding sites. To identify already known binding sites or new ones the aim of this program is to perform predictive experiments on large sets of proteins for a given ligand of interest. For these, we will only dispose of the free energy of binding to discriminate good docking poses. For 7/24 there is no match between the binding energy and the geometric criterions. In some remarkable cases we have shown previously, only the free binding energy computation does not allow to retrieve similar poses to the crystallographic ones. That is demonstrating that the evaluation of the binding energy is not an absolute reference. To reduce the unsuccessful ratio we have to reinforce the ranking evaluation process by adding other scoring methods able to make up rare cases of force field failures. However, in most of the cases we have seen that the PS method strongly performs to detect druggable cavities on a protein receptor. In fact some proteins present multiple binding sites well described in enzymology allosteric phenomena especially. The advantage of using multiple pockets search is to identify well differentiated multiple sites on the fly during a unique docking simulation. That allows us to consider ligand repositioning experiments and also second targets and off-targets hunting. In addition the AC method is able to overcome the PS method failures with adding search completeness and not excluding planar binding surfaces such as protein-protein binding area in particular. So, we demonstrate that the combination of the two methods is an accurate strategy to identify new protein targets for a given ligand.

We developed an effective tool to perform large-scale inverse virtual screening works on both HPC hardware and personal computer able to identify proteins targets for a chemical ligand of interest. Originally developed for and with AutoDock4.2, the framework will embed a version with AutoDock Vina [40] as docking engine that supports multithreading natively but does not allow fine-grain control of algorithm parameters contrary to the previous AutoDock software.

ACKNOWLEDGMENT

The experiments presented on this paper were carried out on the ROMEO Computing Center (https://romeo.univ-reims.fr) and the Multi-scale Molecular Modelling Platform (https://p3m.univ-reims.fr). Romain Vasseur is founded by a Bull SAS CIFRE grant. We thank Romulo Gadelha de Moura Lima, for his great contribution to format the code.

REFERENCES

[1] R. Abagyan and M. Totrov, “High-throughput docking for lead generation,” Curr. Opin. Chem. Biol., vol. 5, no 4, 2001, pp. 375-382.

[2] D. Giganti *et al.*, “Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment”, J. Chem. Inf. Model., vol. 50, no 6, 2010, pp. 992-1004.

[3] G. Klebe, “Virtual ligand screening: strategies, perspectives and limitations,” Drug Discov. Today, vol. 11, no 13-14, 2006, pp. 580-594.

[4] O. Korb, T. Stützle, and T. E. Exner, “Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS,” J. Chem. Inf. Model., vol. 49, no 1, 2009, pp. 84-96.

[5] J. J. Irwin *et al*., “Automated Docking Screens: A Feasibility Study,” J. Med. Chem., vol. 52, no 18, 2009, pp. 5712-5720.

[6] G. Jones, P. Willett, and R. C. Glen, “Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation,” J. Mol. Biol., vol. 245, 1995, pp. 43-53.

[7] S. Cosconati *et al*., “Virtual screening with AutoDock: theory and practice,” Expert Opin. Drug Discov., vol. 5, no 6, 2010, pp. 597-607.

[8] S. Zhang, K. Kumar, X. Jiang, A. Wallqvist, and J. Reifman, “DOVIS: an implementation for high-throughput virtual screening using AutoDock,” BMC Bioinformatics, vol. 9, no 1, 2008, pp. 126.

[9] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, “A fast flexible docking method using an incremental construction algorithm,” J. Mol. Biol., vol. 261, no 3, 1996, pp. 470-489.

[10] R. A. Friesner *et al*., “Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes,” J. Med. Chem., vol. 49, no 21, 2006, pp. 6177-6196.

[11] Y. Y. Li, J. An, and S. J. Jones, “A computational approach to finding novel targets for existing drugs,” Plos Comput. Biol., vol. 7, no 9, 2011, pp. e1002139.

[12] S. L. Kinnings and R. M. Jackson, “LigMatch: A Multiple Structure-Based Ligand Matching Method for 3D Virtual Screening,” J. Chem. Inf. Model., vol. 49, no 9, 2009, pp. 2056-2066.

[13] R. Wang, X. Fang, Y. Lu, and S. Wang, “The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures,” J. Med. Chem., vol. 47, no 12, 2004, pp. 2977-2980.

[14] J. J. Irwin and B. K. Shoichet, “ZINC-a free database of commercially available compounds for virtual screening,” J. Chem. Inf. Model., vol. 45, no 1, 2005, pp. 177-182.

[15] F. H. Allen, “The Cambridge Structural Database: a quarter of a million crystal structures and rising,” Acta Crystallogr. B, vol. 58, no 3, 2002, pp. 380-388.

[16] RCSB Protein Data Bank - http://www.rcsb.org/pdb

[17] P. Bradley, K. M. Misura, and D. Baker, “Toward high-resolution *de novo* structure prediction for small proteins,” Science, vol. 309(5742), 2005, pp. 1868-1871.

[18] J. Moult, “A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction,” Curr. Opin. Struct. Biol., vol. 15, no 3, 2005, pp. 285-289.

[19] N. Eswar *et al*., “Comparative protein structure modeling using Modeller,” Curr. Protoc. Bioinforma., 2006, pp. 5-6.

[20] A. P. Norgan, P. K. Coffman, JP. A. Kocher, D. J. Katzmann, and C. P. Sosa, “Multilevel parallelization of AutoDock 4.2,” J. Cheminf., vol. 3(1), no 1, 2011, pp. 1-9.

[21] X. Zhang, S. E Wong, and F. C. Lightstone, “Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines”, J. Comp. Chem., vol. 34, no 11, 2013, pp. 915-927.

[22] R. De Paris, F. A. Frantz, O. Norberto de Souza, and D. A. Ruiz, “wFReDoW: A Cloud-Based Web Environment to Handle Molecular Docking Simulations of a Fully Flexible Receptor Model”, BioMed Res. Inter., 2013, pp. 1-12.

[23] T. Y. Tsai, K. W. Chang, and C. Y. Chen, ”iScreen: world’s first cloud computing web server for virtual screening and de novo drug design based on TCM database@Taiwan,” J. Comput. Aided Mol. Des., vol 25, 2011, pp. 525-531.

[24] I. Pechan and B. Fehér, “Molecular docking on FPGA and GPU platforms”, International Conference on Field Programmable Logic and Applications, 2011.

[25] G. M. Morris *et al*., “AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility,” J. Comput. Chem., vol. 30, no 16, 2009, pp. 2785-2791.

[26] R. Vasseur *et al*., ”Parallel strategies for an inverse docking method,” Proc. PBio: International Workshop on Parallelism in Bioinformatics, EuroMPI User’s Group Meeting (EuroMPI 2013), ACM, Sept. 2013, pp. 253-258.

[27] D. Ghersi and R. Sanchez, “Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites,” Proteins Struct. Funct. Bioinforma., vol. 74, no 2, 2009, pp. 417-424.

[28] C. Hetényi and D. van der Spoel, “Toward prediction of functional protein pockets using blind docking and pocket search algorithms,” Protein Sci., vol. 20, no 5, 2011, pp. 880-893.

[29] V. Le Guilloux, P. Schmidtke, and P. Tuffery, “Fpocket: An open source platform for ligand pocket detection,” BMC Bioinformatics, vol. 10, no 1, 2009, pp. 168.

[30] Structural Chemogenomics Group - http://bioinfo-pharma.u-strasbg.fr

[31] E. Kellenberger, J. Rodrigo, P. Muller and D. Rognan, “Comparative evaluation of eight docking tools for docking and virtual screening accuracy,” PROTEINS: Struct. Funct. Bioinf., vol.57, 2004, pp. 225-242.

[32] G. M. Morris *et al*., ”Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function,” J. Comp. Chem., vol. 19, no 14, 1998, pp. 1639-1662.

[33] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, “A semiempirical free energy force field with charge-based desolvation,” J. Comput. Chem., vol. 28, no 6, 2007, pp. 1145-1152.

[34] W. Humphrey, A. Dalke and K. Schulten, K., “VMD - Visual Molecular Dynamics,” J. Molec. Graphics, vol. 14, 1996, pp. 33-38.

[35] C. Hetényi and D. van der Spoel, “Blind docking of drug-sized compounds to proteins with up to a thousand residues,” Febs Lett., vol. 580, no 5, 2006, pp. 1447-1450.

[36] B. Iorga, D. Herlem, E. Barré, and C. Guillou, “Acetylcholine nicotinic receptors: finding the putative binding site of allosteric modulators using the blind docking approach,” J. Mol. Model., vol. 12, no 3, 2005, pp. 366-372.

[37] I. W. Davis, K. Raha, M. S. Head, and D. Baker, “Blind docking of pharmaceutically relevant compounds using RosettaLigand,” Protein Sci., vol. 18, no 9, 2009, pp. 1998-2002.

[38] A. Grosdidier, V. Zoete, and O. Michielin, “Blind docking of 260 protein-ligand complexes with EADock 2.0,” J. Comput. Chem., vol. 30, no 13, 2009, pp. 2021-2030.

[39] R. Vasseur *et al*., “Parallel strategies for an inverse docking method,” J. Parall. Comput., special issue on Parallelism in Bioinformatics, in press.

[40] O. Trott and A. J. Olson, “AutoDock Vina: Improving the speed and accuracy od docking with a new scoring function, efficient optimization, and multithreading,” J. Comput. Chem., vol 31, 2010, pp. 455-461.

ISBN: 978-1-61208-335-3

The Anti-inflammatory Effects of Laser Acupuncture at ST36 (Zusanli) Acupoint in the Model of Acute Inflammation Induced by Carrageenan in the Paw of Mice

Vanessa Erthal, Percy Nohama*
Rehabilitation Engineering Laboratory/CPGEI
Federal Technological University of Paraná UTFPR
Curitiba, Brazil
e-mail: acupuntura_vane@yahoo.com.br

Maria Fernanda de P. Werner, Cristiane H. Baggio
Pharmacology Department
Federal University of Paraná UFPR
Curitiba, Brazil
e-mail: crisbaggio@gmail.com

Abstract— **Low-intensity Laser acupuncture (LA) has been applied as an alternative to needling along the past three decades. The ST36 (Zusanli) acupoint is used to treat inflammatory processes, acute pain and gastrointestinal disturbs. For this reason, the aim of the present study was to evaluate the anti-inflammatory effect of Laser acupuncture (830 nm, 4 J/cm^2) on ST36 acupoint on paw edema induced by carrageenan in mice, a model of acute inflammation. Mice were treated with LA 30 min before intraplantar injection of carrageenan (300 µg/paw). The formation of edema was assessed using a digital micrometer and temperature analyzed through thermography. The results indicate that ST36 LA significantly inhibited the paw edema induced by carrageenan and reduced the temperature on skin plantar surface. In conclusion, these results demonstrated that ST36 photonic stimuli have anti-inflammatory effect in acute model of inflammation in mice.**

Keywords: Laser acupuncture; inflammation; edema.

I. INTRODUCTION

Acupuncture is an effective procedure for pain relief, nausea and vomiting, bronchial asthma, musculoskeletal disorders and inflammatory conditions [1]. There are different acupuncture techniques, which traditionally use needle puncture [2]. However, laser acupuncture (LA) is a noninvasive and noninfectious method that can avoid pain and psychological fear promoted by the insertion of needles [2-5]. Indeed, LA is a form of phototherapy at acupoint similar to needle acupuncture, differing in the type of stimulus [6].

The ST36 (Zusanli) acupoint has been used to treat inflammation, acute pain and gastrointestinal disturbs [7]. Previous studies showed that ST36 acupoint stimulated with Low Level Laser Therapy (LLLT), during 2, 6 and 10 min, inhibited the nociceptive response induced by formalin in mice [8]. Recently, Erthal et al. [4] also demonstrated that ST36 stimulation with Gallium Aluminium Arsenide (GaAlAs) laser elicited significant antinociceptive effect against acetic acid- and formalin-induced nociception in rats, with participation of opioidergic and serotonergic systems. Moreover, it has been reported that the application of ultra low level laser therapy (ULLLT) on ST36 and TB5 acupoints also reduced acute and chronic inflammation induced by carrageenan and complete Freund's adjuvant, respectively (for review see Baratto et al. [9]). Interestingly, a clinical study with LA, set to 830 nm and 30 mW, applied on ST36 and IG4 acupoints reduced significantly the migraine in children [10].

Carrageenan is a substance widely used for induction of inflammation in animal models, and is a test employed to assess the effects of alternative methods, as LA, for inflammation and pain control [11]. It is well known that temperature can be a parameter in models of inflammation and several studies with acupuncture use the local measuring of temperature to evaluate its effect on inflammatory processes. Sanchez et al. [12] demonstrated that the thermal imaging technology (thermography) is a rapid, highly reproducible method to quantify the degree of inflammation in rat models of general inflammation.

However, further studies must be done to deeply assess its clinical efficacy and to investigate the molecular mechanisms involved in its effect. Despite the stimulation of ST36 acupoint being used to treat inflammatory conditions, there are few evidences demonstrating the effectiveness of LA for reducing edema due to inflammation. For this reason, the aim of the present study was to evaluate the anti-inflammatory effect of LA on ST36 acupoint using the acute inflammatory model in mice.

The paper was divided into four parts: (I) Introduction, in which we present the scientific fundamentals involved in anti-inflammatory 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; and (IV) Conclusion, where we highlight findings on the performed 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 they were approved by the Committee of Animal Experimentation of the Federal University of Paraná (CEUA - UFPR, protocol number 514).

B. LLLT treatment procedures and body location

For the experiments, a low-intensity GaAlAs laser equipment was used. Its main parameters were: wavelength of 830 nm (in continuous-mode), fluence of 4 J/cm², power of 30 mW, irradiation area reached 6 mm², duration of 8 s on the acupoint. The animals were randomly divided into four groups (n=8 animals per group): (1) Control group, which was not treated; (2) Laser on group, which was treated with unilateral ST36 laser acupuncture; and (3) Laser off group, in which laser device was turned off but holding the probe in contact with ST36 acupoint; (4) Dexamethasone group [DEXA, 0.5 mg/kg, intraperitoneal (i.p.)], a positive control of the test. ST36 (Zusanli) acupoint is located between the tibia and the fibula, approximately 5 mm lateral to the anterior tubercle of the tibia [4].

C. Acute inflammation induced by carrageenan

The animals were treated with laser and DEXA and, after 30 min, an intraplantar (i.pl.) injection of carrageenan (300 µg/paw, 20 µl) was administered into the right hind paw of the mice. The thickness of the paw was measured using a digital micrometer (Great, MT-045B) before the induction of edema (B: basal) and at different time points after the injection of the phlogistic agent. All of the assessments were performed by the same investigator in order to reduce any potential inter-operator differences.

D. Thermografic analysis

Temperature measurements of the hind paw´s plantar surface were obtained by an Infrared Camera, model A325 (FLIR Systems, Inc.). The main parameters of the thermographic camera are: acquisition frequency of 60 Hz with 16 bits-resolution, 320 x 240 pixels image resolution, detecting wavelengths from 7.5 up to 13 µm, lens incorporating autofocus, temperature measurement in the range of -20 to +120 °C, with 2% accuracy, thermal resolution of 0.08 °C and 0.1 mm of spatial resolution. The software used for thermographic images acquisition, storage and analysis was the ThermaCAM Researcher Pro 2.9, developed by FLIR Systems, Inc. Skin temperature was measured after leaving each animal at least one hour to acclimate with the laboratory temperature. Animals were lightly anesthetized with sodium pentobarbital (30-40 mg/kg, i.p., Cristália, Brazil) to suppress the righting reflex while preserving the withdrawal reflex. Anesthetized mice were gently placed on the box, and the dorsal surface of the hind paw was fixed. The hind paws were positioned, and the heat emitted from the plantar region was measured using the infrared camera.

E. Statistical analysis

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

III. Results and Discussion

Inflammation is the body’s immediate response to the tissue damage and defined by vasodilation, exudation of fluid and cell migration. In 1962, the carrageenan-induced inflammatory response was described for the rat paw, and in 1969 for mice [13]. Since that time, edema on mice´s paw has been increasingly used to test new anti-inflammatory drugs and treatments [14]. Then, we used this model of acute inflammation induced by carrageenan to evaluate the anti-inflammatory effect of LA treatment. In our experiments, it was observed a rapid onset of paw edema in the control group after i.pl. injection of carrageenan. Interestingly, the results depicted in Fig. 1 indicate that ST36 LA significantly inhibited the paw edema at 2 and 3 h after phlogistic agent injection, with inhibitions of 13 and 18%, respectively. However, the treatment with laser device turned off was not able to reduce the edema when compared to the control group. DEXA, a steroidal anti-inflammatory drug and positive control of the test, also reduced the paw edema at 2 and 3 h (Fig. 1). It is known that laser therapy activates both local microcirculation and cellular metabolism, and produces anti-inflammatory, analgesic and regenerative effects [15], suggesting that these factors could be involved in our LA treatment. In accordance with our findings, another type of LA, the ULLLT, is also able to inhibit the paw edema induced by carrageenan [16]. Lee et al. [17] showed that electroacupuncture, applied on ST36 and SP6 acupoint, on different frequencies such as 2, 15 and 120 Hz, produced relevant anti-edema effects compared with control group. For laser therapy, a range of wavelengths 633.8 up to 904 nm can be applied [18]. Besides, according to the Arndt–Schultz law for biostimulation, anti-inflammatory and analgesic effects occur at doses between 0.05 and 10 J/cm^2 [19]. In this study, LA with wavelength of 830 nm and radiant exposure of 4 J/cm^2 showed significant effects on the inflammatory model. In previous experimental studies in our laboratory, the applied dose of 3 J/cm^2 demonstrated best responses on inflammatory and nociception models. In studies applying eletroacupuncture on ST36 point, Park and colleagues [20]

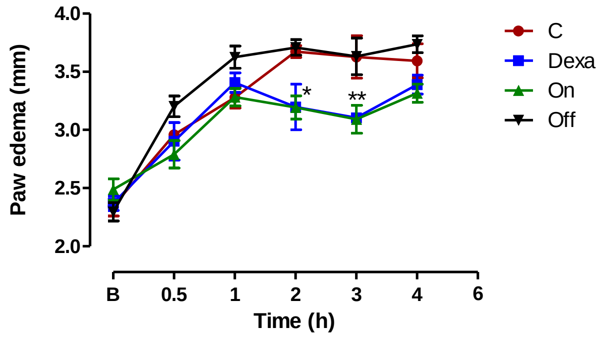


Figure 1. Effects of LA or DEXA on carrageenan-induced edema. Edema of the carrageenan-inflamed hindpaw was determined by measuring paw thickness with micrometer (n=8), indicate the S.E.M. **p<0.01.

reported effectiveness of this acupoint for anti-inflammatory and anti-arthritic on collagen-induced arthritis (CIA), via suppressing autoimmunity and modulating immune abnormality. Kim et al. demonstrated that EA stimulation of the Zusanli acupoints produced significant suppressive effect on carrageenan-induced paw inflammation and hyperalgesia [21]. Thermography has been useful for diagnosis of inflammatory processes because it can assess variations in skin surface temperature. It is highly sensitive and noninvasive, capable of detecting very small alterations on skin temperature [22]. The thermal image analysis involves measurements of the surface temperature the body using an array of infrared sensors installed inside the camera. This image allows the simultaneous measurement of temperatures of multiple points on the skin [23]. The images and the graphic illustrated in Fig. 2 show the skin temperature change for the plantar surface of the hind paw. Changes were measured through the infrared camera adjusted to the range of 25 to 37 °C. Fig. 2c shows that the inflammation promoted by carrageenan increased the temperature of plantar surface in 14% (Naive: 26.5 ± 0.5 °C). However, the treatment of animals with LA on acupoint ST36 reduced the temperature in 12% when compared with control group (Control: 30.1 ± 1.0 °C). By means of the thermographic analysis, we can conclude that LA on the ST36 acupoint has an anti-inflammatory action.

IV. Conclusion

Several years of research have produced a steady stream of laser acupuncture studies; however, the objective assessment and reproducibility of results are difficult because of the lack of information about the main physical parameters set. However, recent studies have shown that laser therapy when administered by a specific emission mode may elicit significant biological effects

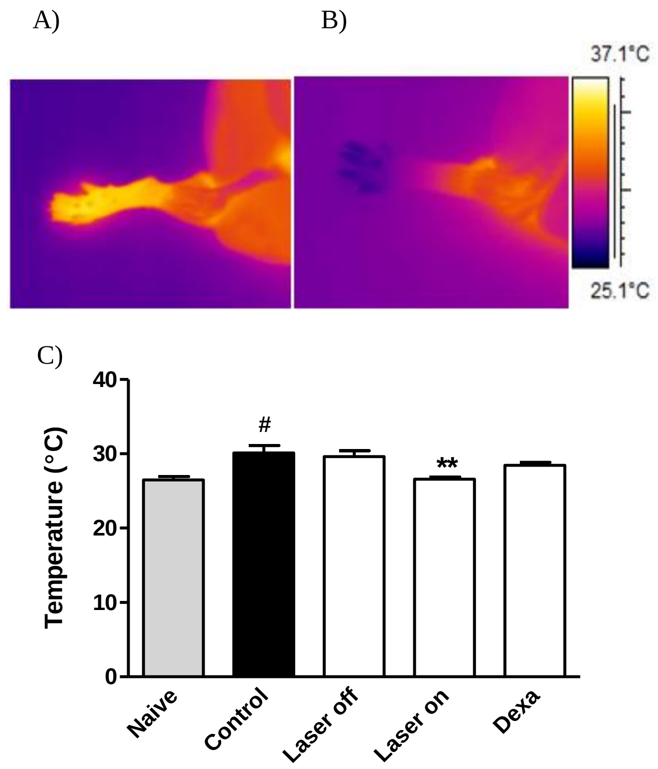


Figure 2. Digitized images of hind paw plantar surface skin temperature, the control (A) and laser on (B) groups. (C) Effect of ST36 laser acupuncture or DEXA in carrageenan-induced paw in mice. Each group represents the mean of 8 animals, and the vertical lines indicate the S.E.M. **p<0.01 and #p<0.05 when comparing with control group.

Finally, this study allows us to conclude that stimulation of the acupoint ST36 with LLLT produces relevant suppression of carrageenan-induced paw edema. Further studies have been and will be carried out in our laboratory to understand the effects of LA on inflammation.

Acknowledgment

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

References

[1] Y. Yim; et al., “Electro-acupuncture at acupoint ST36 reduces inflammation and regulates immune activity in Collagen-Induced arthritic mice,” eCAM, , vol. 4(1), pp. 51-57, March 2007.
[2] I. Quah-Smith, C. Smith, J. Crawford, and J. Russell, “Laser acupuncture for depression: A randomized double blind controlled trial using low intensity laser intervention, Journal of Affective Disorders, vol. 148, pp. 179-87, June 2013.
[3] 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, vol. 88, pp. 159-65, 2010.

[4] 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., vol. 28, pp. 1345-51, Sep. 2013.

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

[6] M. L. Lin, et al., “Evaluation of the effect the laser acupuncture and cupping with Ryodoraku and visual analog scale on low-back pain,” eCAM, vol. 2012, 2012, doi:10.1155/2012/521612.

[7] G. Maciocia, “Os Fundamentos da Medicina Chinesa,” Editora Roca, 1996.

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

[9] L. Baratto, et al., “Ultra-low-level laser therapy,” Lasers Med Sci., vol. 26, pp. 103–112, Jan. 2011.

[10] S. Gottshling, et al., “Laser acupuncture in children with headache: a double-blind, randomized, bicenter, placebo-controlled trial,”. Pain, vol. 10, pp.1-8, Jul. 2008.

[11] S. R. Barretto, et al., “ Evaluation of anti-nociceptive and anti-inflammatory activity of low-level therapy on temporomandibular joint inflammation in rodents,” Journal of Photochemistry and Photobiology B: Biology, vol. 129, pp. 135–142, Dec. 2013

[12] B. M. Sanchez, et al., “Use of a portable thermal imaging unit as a rapid, quantative method of evaluating inflammation and experimental arthritis”, Journal of pharmacological and toxicological methods, vol. 57, pp. 169-175, May-Jun. 2008.

[13] J. C. Castardo, et al., “Anti-inflammatory effects of hydroalcoholic extract and two biflavonoids from *Garnicia gardneriana* leaves in mouse paw oedema,” Journal Ethnopharmacology, vol. 118, pp. 405-411, Aug. 2008.

[14] B. S. Wang, G. J. Huang, Y. H. Lu and L. W. Chang, “Anti-inflammatory effects of an aqueous extract of Welsh onion green leaves in mice”, Food Chemistry, vol. 138 , pp. 751–756, Jun. 2013.

[15] P. Avic, et al.,”Low-Level Laser (Light) Therapy (LLLT) in skin: stimulating, healing, restoring”, Semin Cutan Med Surg, vol. 32, pp. 41-52, 2013.

[16] A. Giuliani, et al., “Very low level laser therapy attenuates edema and pain in experimental models’,. Int J Tissue React, vol. 26, pp. 29–37, 2004.

A. C. Guimarães, et al., “Low-level laser therapy (LLLT) reduces the COX-2 mRNA expression in both subplantar and total brain tissues in the model of peripheral inflammation induced by administration of carrageenan”, Lasers in Medical Science, February 2014, DOI 10.1007/s10103-014-1543-2.

[17] J. H. Lee, K. J. Jang, Y. T. Lee, Y. H. Choi and B. T. Choi, “Eletroacupuncture inhibits inflammatory edema and hyperalgesia through regulation of ciclooxygenase synthesis in both peripheral and central nociceptive sites,” The American Journal of Chinese medicine, vol. 34, pp. 981-988, 2006.

[18]. M. Artés-Ribas, J. Arnabat-Dominguez and A. Puigdollares, “Analgesic effect of a low-level laser therapy(830 nm) in early orthodontic treatment,” Laser Med Sci, vol. 28, pp. 335-41, Jan. 2013

[19] W. Yu, et al., “Effects of photostimulation on wound healing in diabetic mice,” Laser Surg Med, vol. 20, pp. 56-63, 1997.

[20] D. S. Park, B. K. Seo and Y. H. Baek, “Analgesic effect of eletroacupuncture on inflammatory pain in collagen-induced arthritis rats: medition by alpha-2 and beta-adrenoceptors,” Reumatol Int., vol. 33, , pp. 309-14, Feb. 2013.

[21] H. Kim, et al., “Low-frequency eletroacupuncture suppress carrageenan-induced paw inflammation in mice via sympathetic post-ganglionic neurons, while high-frequency EA suppression is mediated by the sympathoadrenal medullary axis,” Brain Research Bulletin, vol. 75, pp. 698-705, Mar. 2008.

[22] V. Z. Sacharu, et al., “Thermographic evaluation of hind paw skin temperature and functional recovery of locomotion after sciatic nerve crush in rats,” Clinics, vol 66, pp. 1259-1266, 2011.

[23] G. Litscher, “ Integrative Laser Medicine and High-Tech acupuncture at the medical university of Graz, Austria, Europe,” Laser in medicine, , vol. 2012, pp. 1-21, Jan. 2012.

Detection of AV Impulse Frequency and Verification of Pacemaker Battery Status

Ivana Gálová, Michal Gála
Department of Electromagnetic and Biomedical Engineering
University of Žilina
Žilina, Slovak Republic
ivana.galova@fel.uniza.sk, michal.gala@fel.uniza.sk

Martin Augustýnek, Martin Černý, Marek Penhaker
Department of Cybernetics and Biomedical Engineering
VSB – Technical University Ostrava
Ostrava, Czech Republic
martin.augustynek@vsb.cz, martin.cerny@vsb.cz,
marek.penhaker@vsb.cz

***Abstract* — Pacemaker therapy has irreplaceable position in therapy of many cardiac diseases. The lifetime of implantable devices is several years and functionality depends on the battery condition. The verification of pacemaker battery is possible using portable pacemaker programmers manufactured for specific pacemakers. This option is costly for the general public. An alternative approach is to analyze the patient's electrocardiogram. This article deals with possibilities and methods of pacemaker battery state verifying using stimulation pulse width detection. Future plans include the construction of a small and embedded device, enabling general physicians and patients to obtain on-demand information about the battery status of the pacemaker.**

Keywords - electrocardiography; single chamber pacemaker; dual chamber pacemaker; pulse width; battery.

I. Introduction

Pacemakers play an important role in the treatment of cardiac diseases. This is especially true in cases where medication is no longer sufficient and patients require a pacemaker while waiting for a heart transplant. The other group of patients includes less severe heart-related problems which can be satisfactorily solved by implanted pacemakers and thus preserving patient vital functions. The lifetime of such devices is several years. Guaranteed operation is currently limited by the battery life of the device. When the battery capacity decreases, the device must be replaced. This end, manufacturer-specific portable pacemaker programmers must be used to verify battery properties. These are however expensive and usually available only to cardiologists or internal physicians performing invasive cardiology. An alternative approach is to evaluate the state of the pacemaker battery by analyzing the patient electrocardiogram (ECG) recording. This article discusses the possibilities and means to verify the battery status of implantable stimulation devices and it proposes future plans for the construction of a small and embedded device, enabling general physicians and patients to obtain on-demand information about the battery status of the pacemaker [1][2][3][4].

Section 2 deals with lifetime of the power source according to norm EN 45502-2-2:2008. Section 3 describes a system for measurement and analysis of simulated ECG signal. This section describes the steps of algorithm for distance detection between the AV impulses that stimulate the atria and ventricles. In Section 4, we test the proposed algorithm. Testing is realized on simulated data with different kinds of noises and on real patient data. After AV impulse detection, the algorithm evaluates battery state. Section 5 deals with conclusion and future work.

II. Pacemaker battery states

According to standard EN 45502-2-2:2008, active implantable medical devices must include means to warn in advance of low battery state. The warning interval (during normal use of the device) must be comparable to regular patient checks at the ambulance. The conditions of replacement are model and manufacturer specific. In Figure 1, the following device lifetime phases are defined based on the residual battery capacity of implantable medical devices:

- Beginning of life (BOL) - implantable device is first approved by the manufacturer and certified for marketing,
- Elective replacement time (ERT) - the power supply reaches a pre-determined threshold capacity of the manufacturer, in which case the device replacement is recommended. This indicates the beginning stages of prolonged service period,
- Prolonged service period (PSP) - time period for the recommended replacement, during which the device continues to operate as specified by the manufacturer,
- End of life (EOL) - expiration of the extended service life, which is the end of the original manufacturer's specified functionality of the device, power saving mode until complete failure.

More frequent patient monitoring is recommended after crossing the elective replacement near (ERN) threshold [1][9].

ISBN: 978-1-61208-335-3

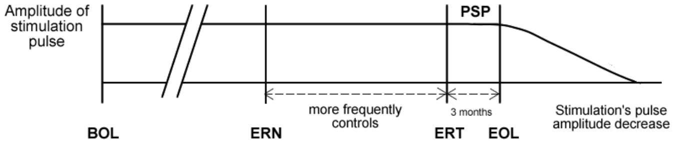

Amplitude of stimulation pulse
PSP
more frequently controls
3 months
Stimulation's pulse amplitude decrease
BOL
ERN
ERT
EOL

Figure 1. Lifetime of the power source.

The mentioned norm does not define dependence of the device lifetime phases on frequency or on magnitude, but most of the manufacturers especially indicate the dependence on frequency (see Table I). The frequency is measured in asynchronous mode which is activeted by permanent magnet with induction of more than 1 mT [1][9].

TABLE I. THE PACING PULSE FREQUENCIES IN ASYNRONOUS MODE FOR BOL AND ERT STAGES

Manufacturer	BOL [1/min]	ERT [1/min]
Biotronik	90	80
Boston Scientific	100	85
ELA	96	80
Medtronic	85	65
MEDICO	100	70
St. Jude Medical	98.6	86.3
Vitatron	100	86

A. Low battery

Figure 2 shows that the battery status of a single chamber pacemaker is determined by measuring the frequency (f = 1/T, f – frequency, T – period) and pacing pulse width (label - d) in asynchronous mode. In general the pacing pulse width can be from 0.05 ms to 2 ms [1].

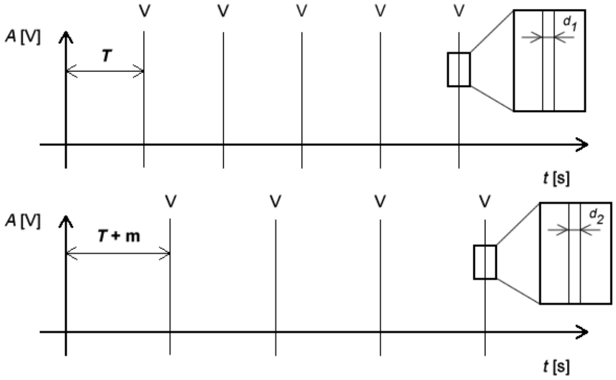


Figure 2. The frequency of single chamber pacemaker stimulation in asynchronous mode: a) the proper function, b) low battery condition.

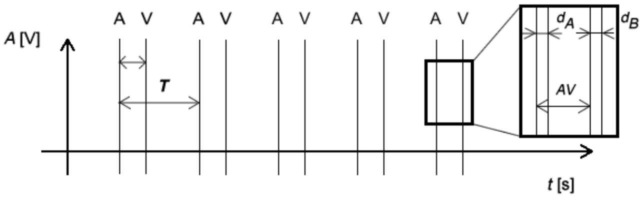


Figure 3. The stimulation frequency in asynchronous mode during correct functioning of a dual chamber pacemaker.

Figure 3 shows that dual chamber pacemakers can also determine the distance between the AV impulses that stimulate the atria and ventricles [1].

III. CONSTRUCTION OF SYSTEM FOR SENSING AND EVALUATION OF THE BATTERY STATUS

To evaluate the pacemaker battery, we must first use a permanent magnet to switch the pacemaker into asynchronous mode and then record the ECG signal. To capture the signal, we used the MP36 unit manufactured by Biopac which we connected to the patient simulator FLUKE MPS450. In Figure 4 we can see the mentioned test system.

Using the simulator, we tested asynchronous mode of single chamber (Figure 5) as well as of dual chamber (Figure 6) pacemaker. The pacemaker in the mentioned mode generates periodic pacing pulses regardless of the actual heart electrical activity. The recorded signal enables us to obtain information about the frequency and pulse width of the stimulus, or the duration of AV delay in dual chamber pacemaker. Based on the measured parameters, it is possible to evaluate the condition of the battery. Stimulation frequency decreases in time and the pacing pulse widens. The proposed algorithm is designed only for these conditions (frequency, pulse width and AV delay) and it does not give a solution for the dependence of the device lifetime phases on magnitude. The future work will be focused on the solution for this problem.

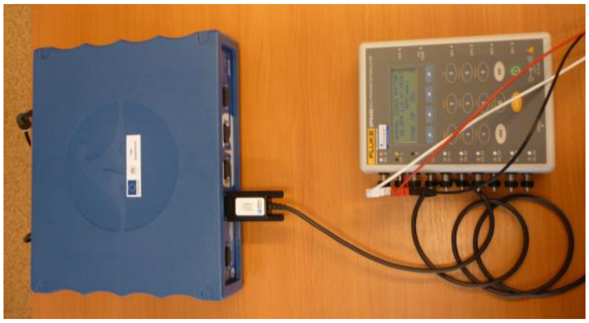


Figure 4. The Test system (measuring unit and MP36 patient simulator MPS450).

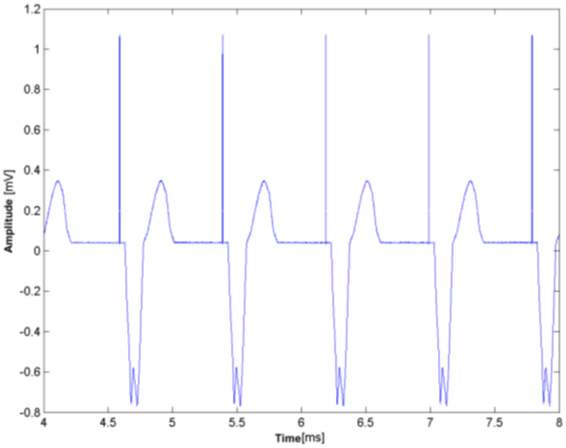


Figure 5. The simulated ECG signal of a patient with a single chamber pacemaker in asynchronous mode.

The design and implementation of the algorithm was carried out in MATLAB. The algorithm development was divided into several steps, each addressing possible drawbacks so as to achieve the desired accuracy and speed of the proposed algorithm. The first step consisted in the analysis and evaluation of the original recording without any modification. The algorithm takes into account only samples above the pre-selected threshold (red line, Figure 7). By comparing the distance between successive samples and known pacing pulse width (information obtained from the simulator), pacing pulses were detected and the required parameters were calculated.

However, detection was not accurate because the signal contained undesired artifacts. To eliminate these artifacts, different filter types were applied (differential filter, averaging filter, band-stop filter) [5][8]. However, after their application, the processed signal lost the information necessary to determine the pulse width (after applying the differential filter) and the evaluation time significantly increased (due to the averaging filter). In the end, we decided to use simple mathematical operations, as one of the main criteria was to design a fast and simple algorithm to enable use thereof in a microcontroller.

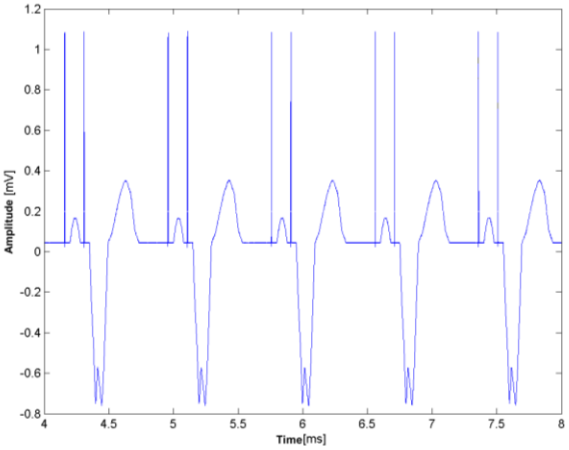


Figure 6. The simulated ECG signal of a patient with a dual chamber pacemaker in asynchronous mode.

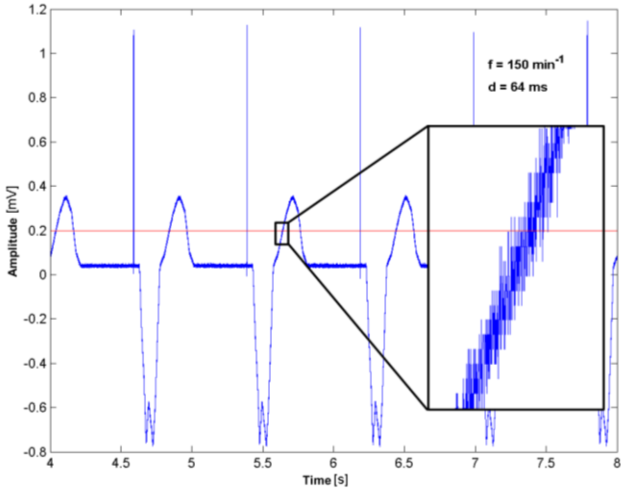


Figure 7. Noisy simulated ECG of a patient with a single chamber pacemaker in asynchronous mode and detail of noisy signal at the boundary (f - frequency of 75 min^{-1}, d - 2 ms pulse width).

Each recorded sample was amplified by its square in order to “smooth out” the recording (amplify the stimulation pulses). Boundaries for pacing pulse detection were first determined intuitively (empirically). At later stages the algorithm has then been modified so as to determine these boundaries based on individual recordings. Subsequent analysis of samples exceeding the threshold value enabled the evaluation of the pulse width, frequency and, in case of dual chamber pacemaker, the duration of the AV delay.

Figure 7 shows that the search for local maxima was performed in order to confirm the correct identification of stimulation pulses and to discard possible noise-related artifacts [6][7].

IV. TESTING OF THE PROPOSED ALGORITHM

The proposed algorithm has been tested on several simulated signals. Individual signals were degraded by different types of artifacts (AC 50 or 60 Hz noise and various breathing motion artifacts). The following section details the individual signals along with the detection and indication of stimulation pulses. Figure 8 shows a sample signal heavily degraded by noise from muscle activity and Figure 9 shows the same signal after processing. In Figure 10, it is shown the detected pacing pulses of a dual chamber pacemaker.

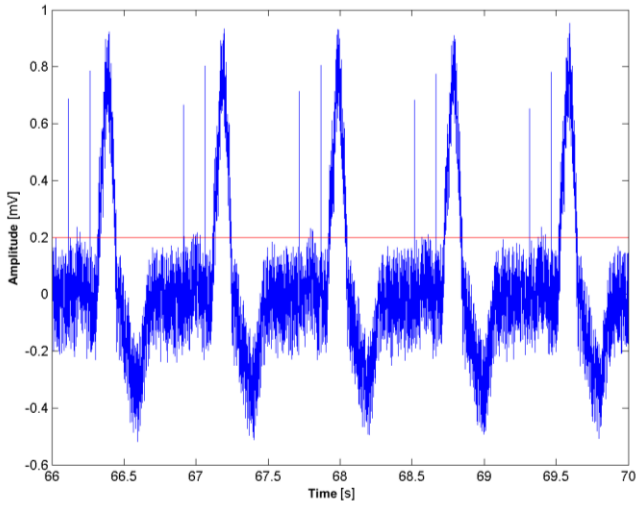


Figure 8. The captured signal containg noise from muscle activity.

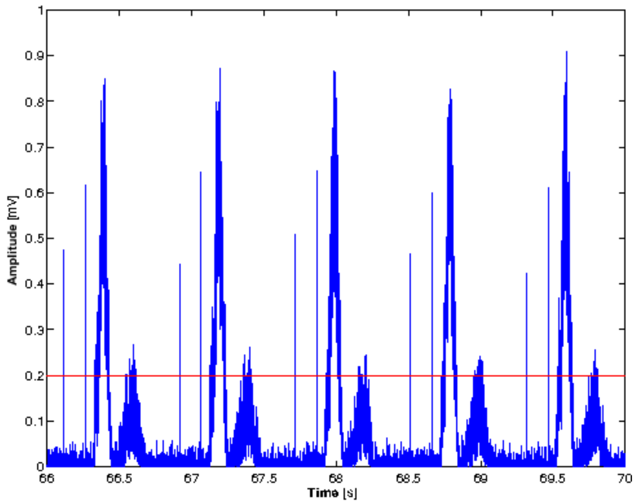


Figure 9. Signal after amplification.

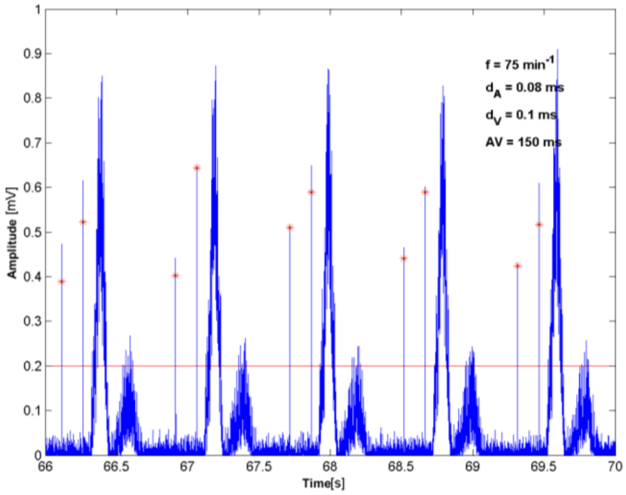


Figure 10. Detection of stimulus pulses (red points).

The developed algorithm was tested with multiple one hour long recordings and the reliability of pacing pulse detection was above 98%. We also tested real-life patient recordings obtained from the Faculty hospital in Zilina. The recording was measured from the first "standard" bipolar limb lead. The patient pacemakers were manufactured by Medtronic. The algorithm efficiency was once again confirmed and the results were discussed with a cardiologist (Dr. Jan Lehotsky) and compared with those obtained from portable pacemaker programmers manufactured by Medtronic. In Figure 11, it is shown the final evaluations of battery status obtained from real patient data. Figure 11a represents the single chamber pacemaker and Figure 11b the dual chamber pacemaker.

V. CONCLUSION AND FUTURE WORK

The algorithm has been tested on real life data from patients with implanted pacemakers (Faculty Hospital Zilina) as well as on simulated recordings containing different types of artifacts (50 Hz, 60 Hz, breathing activity, muscular activity and random noise) and the success rate was above 98%. Current development is focused on hardware design of a small and portable ECG logger to measure and consequently evaluate the status of the battery pacemaker.

The mentioned algorithm is designed only for detection of frequency, pulse width and AV delay and it does not solve the dependence of the device lifetime phases on magnitude. The future work will be therefore focused on solution of this question.

ACKNOWLEDGMENT

This paper is supported by the following project: University Science Park of the University of Zilina (ITMS: 26220220184) supported by the Research & Development Operational Program funded by the European Regional Development Fund.

REFERENCES

[1] D. Korpas, Pacemaker technology, Praha: Mladá fronta, 2011, ISBN 978-80-204-2492-1.

[2] Z. Labza, D. Korpas, and M. Penhaker, "Determination of the Electric Parameters of Dual-Chamber Cardiostimulator," Proc. 9th IFAC Workshop on Programmable Devices and Embedded Systems (PDES 2009), Roznov pod Radhostem, Czech Republic, Feb. 2009, pp. 290-293, ISBN 978-3-902661-41-8, ISSN 1474-6670, doi:10.3182/20090210-3-CZ-4002.xxxx.

[3] J. J. Bax, T. Abraham, S. S. Barold, O. A. Breithardt, J. W. Fung, S. Garrigue, J. Gorcsan, D. L. Hayes, D. A. Kass, J. Knuuti, C. Leclercq, C. Linde, D. B. Mark, M. J. Monaghan, P. Nihoyannopoulos, M. J. Schalij, C. Stellbrink, and C. M. Yu, "Cardiac resynchronization therapy: Part 1--issues before device implantation," J Am Coll Cardiol, Dec. 2005.

[4] D. Korpas, "Psychological Intolerance to Implantable Cardioverter Defibrillator," Biomedical Papers-Olomouc, vol. 152, no. 1, June 2008, pp. 147-149, ISSN 1213-8118.

[5] M. Penhaker, T. Stula, and M. Augustynek, "Long-term heart rate variability assessment" The 5th Kuala Lumpur International Conference on Biomedical Engineering (BIOMED 2011), Held in Conjunction with the 8th Asian

Pacific Conference on Medical and Biological Engineering, (APCMBE 2011), Springer-Verlag Berlin Heidelberg, 20 - 23 June 2011, pp. 532-535, ISSN 16800737, ISBN 978-364221728-9, doi:10.1007/978-3-642-21729-6_134.

[6] B. Babusiak and M. Gala, “Detection of Abnormalities in ECG” Information Technologies in Biomedicine (ITIB 2012), Springer-Verlag Berlin Heidelberg, 2012, pp. 161-171, ISBN 978-3-642-31195-6, ISSN 0302-9743.

[7] Š. Borik and I. Čáp, “Investigation of Pulse Wave Velocity in Arteries” The 35th International Conference on Telecommunications and Signal Processing (TSP 2012), Brno: University of Technology, 2012, pp. 562-565, ISBN 978-1-4673-1116-8.

[8] M. Penhaker, R. Hajovsky, and D. Korpas, “Measurement and Analysis EMC Parameters of Implantable Pacemaker,” Przegląd Elektrotechniczny (Electrical Review), vol. 87, no. 5, 2011, pp. 265-268, ISSN 0033-2097.

[9] EN 45502-2-2:2008, Active implantable medical devices - Part 2-2: Particular requirements for active implantable medical devices intended to treat tachyarrhythmia (includes implantable defibrillators), May 2007.

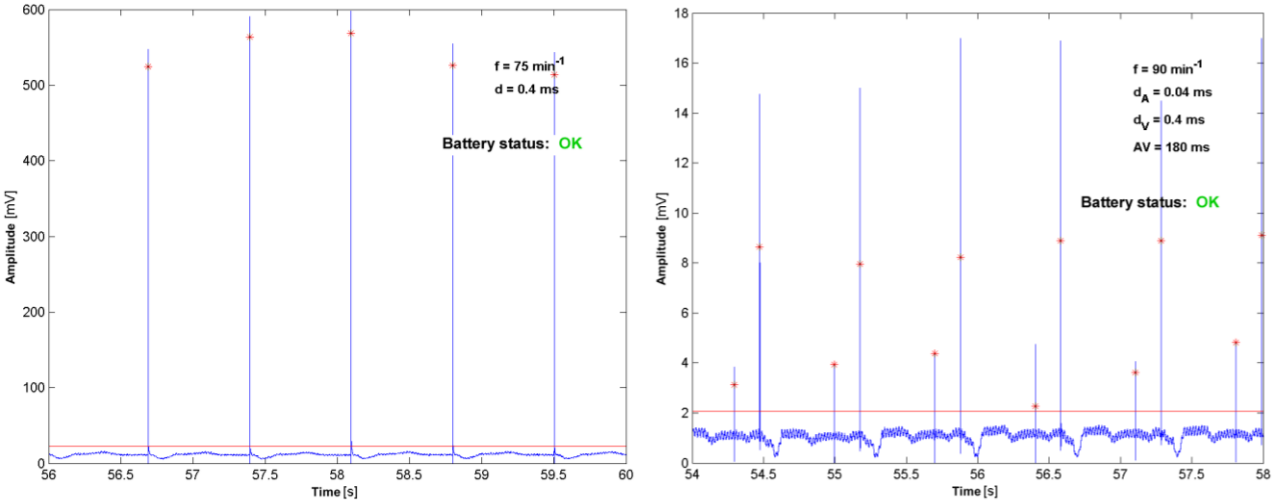


Figure 11. Data analysis of patient with: a) single chamber pacemaker, b) dual chamber pacemaker.

SymRec - Web Tool for Planned Parenthood and Hormonal Therapy

Michal Gála, Ivana Gálová, Zuzana Judáková, Branko Babušiak
Department of Electromagnetic and Biomedical Engineering
University of Žilina
Žilina, Slovak Republic
michal.gala@fel.uniza.sk, ivana.galova@fel.uniza.sk, zuzana.judakova@fel.uniza.sk, branko.babusiak@fel.uniza.sk

***Abstract* — The Symptothermal method is based on monitoring, recording and evaluation of fertility-related symptoms. The mentioned method can help prevent unintended conception, assist achieve conception, help identify various gynecological diseases and timing of hormonal treatment. The Symptothermal method is the preferred method for planned parenthood to hormonal contraception because it has no adverse side effects. Regular paper tables used in gynecology wards require significant user knowledge. However, the digitized data tables require only simple data entry from the user. The developed web interface allows for automatic analysis, evaluation and data archiving. Table records in electronic form significantly save user time, simplify the workflow and include long-term data pertaining to patient health status, thus playing an important role in the early diagnosis of diseases and successful treatment.**

Keywords - Basal temperature; cervical mucus; diagnosis of gynecological diseases; hormonal therapy; cervix; planned parenthood; symptothermal method.

I. INTRODUCTION

The transition from traditional paper-based health records to their electronic counterparts is currently a top priority in many countries. Electronic health records offer a simplified and rapid access to patient data under various circumstances (e.g., heath risk, early diagnosis) The reproducibility and transferability of patient data allows remote consultation with medical staff (telemedicine, home care) [1][2]. This paper deals with implementation of our web tool for users in Slovak Republic.

Symptothermal records used in planned parenthood are also considered to be health records since they offer additional information when diagnosing various gynecologic diseases, including newly detected or long-term health states and may also be used in case of infertility-related treatments [3]. The use of the mentioned records can help unnecessary medical therapy in cases where incorrectly timed sexual intercourse is the cause of infertility. They can also be used for setting-up hormonal therapy and advanced treatment monitoring [4]. Longer records (several years) can help diagnose even minimal menstrual cycle deviations pertaining associated with possible health risks and thus allow for the early diagnosis (cervix cancer) and complete patient recovery.

This paper consists of several parts. Section 2 of the paper describes symptothermal method in general. All rules, which are used by the symptothermal method in our web tool, are described in Section 3. Sections 4 and 5 describe the structure of SymRec web tool. These sections deal with user interface, parts of SymRec web tool window and security of stored data. Section 6 describes the future work.

II. SYMPTOTHERMAL METHOD

The symptothermal method is a universal approach consisting of monitoring, recording and evaluation of fertility symptoms to either prevent or facilitate conception. The mentioned method is natural (without medication) and has no adverse side effects. The process includes the monitoring of basal temperature changes, cervix and quality of cervical mucus. Basal body temperature is the lowest temperature attained by the body during rest (usually during sleep). Lowest values are attained early morning and then continue to slowly increase each half an hour by approximately 0.05°C. Monitored symptoms are recorded in the record table. In Figure 1, it is shown one record table, which is used to store menstrual cycle data (42 day limit) and offers overview of fertile days and cycle changes.

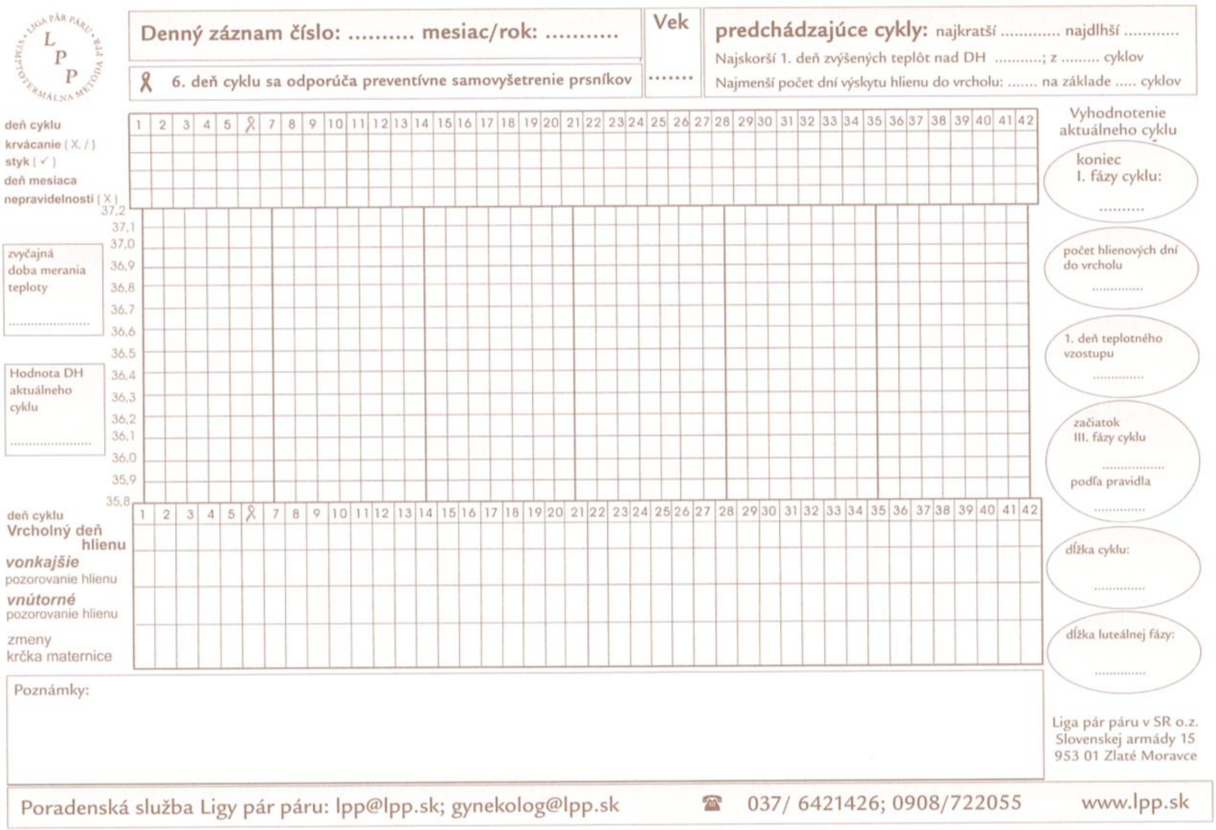
The cycle is divided into four phases:

1. Menstruation,
2. 1st phase – pre-ovulation infertility,
3. 2nd phase – fertile days,
4. 3rd phase – post-ovulation infertility.

The mentioned phases are not identical with physiological changes of the endometrium (proliferative phase, ovulation, secretive phase) except for menstruation, because they include the mutual pair fertility (life time of sperm) [5][6].

III. EVALUATION OF THE 3RD PHASE OF POST-OVULATION INFERTILITY

Post-ovulation infertility commences after the egg reaches its lifetime and begins to degrade within 48 hours post-ovulation. This phase can also be described as completely non-fertile because progesterone generated by the yellow body suppresses any possible ovulation. Monitoring of fertility symptoms does not allow the precise determination of ovulation, however it does allow precise post-ovulation infertility phase. Numerous rules are used to


LIGA PÁR PÁRU • SYMPTOTERMÁLNA METÓDA PÁR
L
P
P

Denný záznam číslo: .... mesiac/rok: ....

Vek

....

predchádzajúce cykly: najkratší .... najdlhší ....

Najskorší 1. deň zvýšených teplôt nad DH ....; z .... cyklov

Najmenší počet dní výskytu hlienu do vrcholu: .... na základe ..... cyklov

🎗 6. deň cyklu sa odporúča preventívne samovyšetrenie prsníkov

deň cyklu 1 2 3 4 5 🎗 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

krvácanie (X, /)

styk (✓)

deň mesiaca

nepravidelnosti (X)

37,2 37,1 37,0 36,9 36,8 36,7 36,6 36,5 36,4 36,3 36,2 36,1 36,0 35,9 35,8

zvyčajná doba merania teploty

....

Hodnota DH aktuálneho cyklu

....

deň cyklu 1 2 3 4 5 🎗 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42

Vrcholný deň hlienu

vonkajšie pozorovanie hlienu

vnútorné pozorovanie hlienu

zmeny krčka maternice

Vyhodnotenie aktuálneho cyklu

koniec I. fázy cyklu: ....

počet hlienových dní do vrcholu ....

1. deň teplotného vzostupu ....

začiatok III. fázy cyklu .... podľa pravidla ....

dĺžka cyklu: ....

dĺžka luteálnej fázy: ....

Poznámky:

Liga pár páru v SR o.z.
Slovenskej armády 15
953 01 Zlaté Moravce

Poradenská služba Ligy pár páru: lpp@lpp.sk; gynekolog@lpp.sk ☎ 037/ 6421426; 0908/722055 www.lpp.sk

Figure 1. Regular (paper) record in Slovak language.

determine this phase, combining various temperature curves and mucus changes. The mentioned rules are based on various observed situations in monitored cycles.

The following rules are currently used:

1. R, based on works published by Dr. Jozef Rötzer,
2. B, based on works published by Dr. John Billings,
3. K, based on works published by Konald Prem and John Kippley.

All mentioned rules are based on the assumption that the greater the temperature increase, the fewer the number of days of mucus drying necessary for accurate prediction of non-fertile phase. Higher accuracy can be obtained by adding one day to the 3rd cycle phase determined by these rules.

A. R rule

The phase of post-ovulation infertility begins in the evening of the 3rd day of temperature rise after maximum mucus day, if the mentioned day is also the 3rd day of mucus drying. The following conditions must be fulfilled for the temperature rise – all temperatures are valid, follow after each other, the first or second temperature is at least 0,1°C higher than lower boundary and third temperature has reached or surpassed the upper boundary. Valid temperatures within one cycle are measured each morning, within the same time range (± 30 minutes) and the same place (e.g., vagina, mouth or rectum), during complete body rest and healthy state.

B. B rule

The phase of post-ovulation infertility begins in the evening of the 3rd day of temperature rise after maximum mucus day, if the mentioned day is also the 4th day of mucus drying. The following conditions must be fulfilled for the temperature rise – all temperatures are valid, they are at least 0.05°C higher than the lower boundary, no temperature decrease has occurred and one temperature has reached or surpassed the upper boundary. The temperatures are measured analogous to R rule, during complete body rest and healthy state. The temperatures in the temperature rise may follow after each other and one temperature can be missing between them [7].

C. K rule

The phase of post-ovulation infertility begins in the evening of the 3rd day of full temperature rise if the mentioned day is also at least the 2nd day of mucus drying. A full temperature rise comprises three temperatures after each other (without interruption), all of which must be at the upper boundary or beyond.

In case only one fertility symptom is available for a certain non-standard situation, we can use rules based on monitoring only one fertility symptom. Insufficient control with additional symptom(s) is compensated by adding an additional day. However, these rules are unable to resolve certain abnormal manifestations of menstrual cycle during the female fertile period and thus are less effective than previous rules. The mentioned rules include the 4T rule (four temperatures), 5T rule (five temperatures) and Marshall Rule [3].

IV. Rule Implementation within the Web Interface

Based on obtained information and rules we created a record table integrated into a web interface. The mentioned application may be used for controlling or supporting female fertility. Another possible use is the early diagnosis of gynecological diseases and proper timing of hormonal therapy. Conventional paper records require significant user knowledge of the underlying evaluation rules. By digitizing the mentioned data we facilitate this process by automatic computer analysis, evaluation and archiving of entered data and automatic prediction of 3rd phase of post-ovulation infertility. The area is defined based on the determined local minima and maxima [8]. The developed interface evaluates a single symptom – basal body temperature.

V. Web Interface of SymRec Webtool

The electronic record table is used for daily data entry and evaluation. Data must be entered periodically each day and thus rely either on new user registration or existing user data. Data are saved in a database. We used MySQL open source database which is running on our server. The data are secured by name and password. Cryptographic hash function SHA1 is used.

Upon providing valid credentials the user is presented with a record table. Figure 2a shows that the first application screen contains information about the user account including the user name, record history, help, printing and user log-out.

In Figure 2, it is shown the page which consists of (ordered from top to bottom):

a) user account information,

b) header,

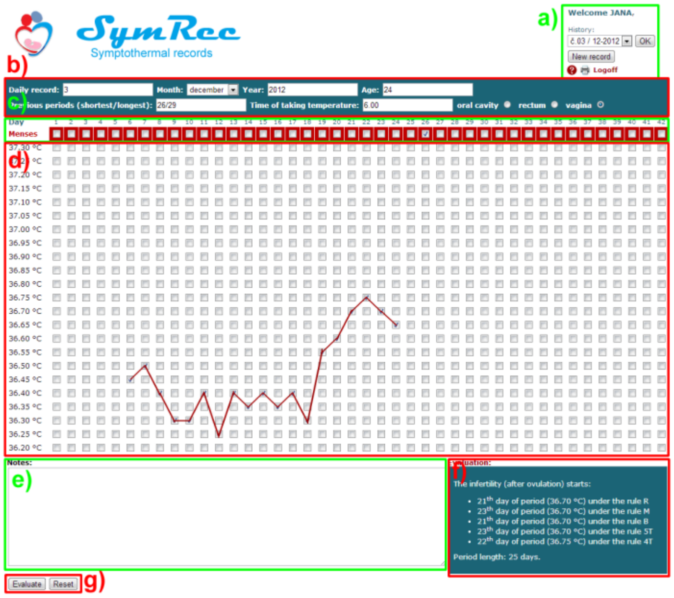
c) menstrual bleeding records,

d) basal temperature array,

e) notes,

f) evaluation box,

g) Evaluate and Reset buttons



Welcome JANA,

History: č.03 / 12-2012 OK

New record

Logoff

Daily record: 3 Month: december Year: 2012 Age: 24

Previous periods (shortest/longest): 26/29 Time of taking temperature: 6.00 oral cavity ○ rectum ○ vagina ◉

Notes:

Evaluation:

The infertility (after ovulation) starts:

- 21th day of period (36.70 °C) under the rule R
- 23th day of period (36.70 °C) under the rule M
- 21th day of period (36.70 °C) under the rule B
- 23th day of period (36.70 °C) under the rule 5T
- 22th day of period (36.75 °C) under the rule 4T

Period length: 25 days.

Figure 2. Webinterface of SymRec webtool.

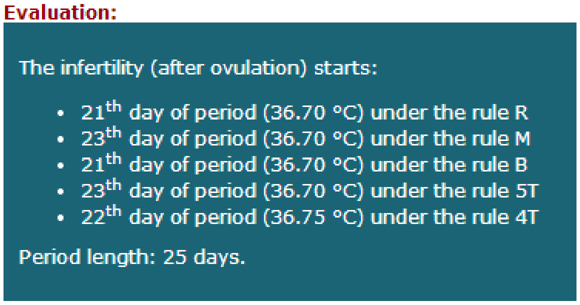
The header contains record information for single cycle, including record number, month, year, age and the length of shortest and longest previous cycle, time and location of measurement.

The first day of menstrual bleeding must be recorded for the current cycle and also on the first day of the successive cycle. This will determine the cycle length calculated by the program and shown in the "Evaluation" box (Figure 2f, bottom right).

The temperature array consists of rows and columns. Rows represent the basal temperature from 36.2°C to 37.3°C and columns represent cycle days. The basal temperature during one cycle must be evaluated under identical conditions and only valid values must be entered into the table. The temperature is measured each morning, same timeframe within one cycle and same place. Measurement must be performed using a thermometer with two decimal precision for more exact evaluation of fertile days. No physical activity has to be performed 1 hour prior to measurement. Temperature measured during illness is not considered valid and thus not recorded that day.

Notes or cycle irregularities may be noted in the "Notes, irregularities" box.

Figure 2g shows that the table header (daily record, month and year), basal temperatures and menstrual bleeding are confirmed using the "Evaluate" button. Entered data are stored in the database, analyzed and results are written to the "Evaluation" box. In Figure 3, it is shown the evaluation box. In evaluation box, it is written when the infertility starts by different rules and how long the period takes.


Evaluation:

The infertility (after ovulation) starts:

- 21th day of period (36.70 °C) under the rule R
- 23th day of period (36.70 °C) under the rule M
- 21th day of period (36.70 °C) under the rule B
- 23th day of period (36.70 °C) under the rule 5T
- 22th day of period (36.75 °C) under the rule 4T

Period length: 25 days.

Figure 3. Detail of "Evaluation" box.

Evaluation of older records allows monitoring of minimum cycle deviations. The application is thus appropriate not only for fertility management but also for early diagnosis of gynecologic diseases and hormonal treatment timing. The web application also includes extensible help detailing all application controls. In Figure 4, it is shown the sample of extensible help. There is written about registration, correct measurement of the temperature and correct notation of the symptoms. User can also find information about records history and evaluation there.

VI. Conclusion and Future Work

The developed web interface fulfills the requirements of the symptothermal method and additionally allows storing, analysis, evaluation and archiving of entered data. Evaluation calculates the third phase of post-ovulation infertility and predicts the cycle length. The interface may be used for controlling or promoting female fertility. Moreover, early diagnosis of gynecological diseases and hormonal treatment timing is also possible. Future development will add other evaluation criteria (calculation of the first phase of pre-ovulation infertility, mucus quality and cervix state) in order to obtain more precise calculation of the third phase of post-ovulation infertility. This time is the web tool available only in Slovak language however internationalization work is underway to include English.

Acknowledgment

This paper is supported by the following project: University Science Park of the University of Zilina (ITMS: 26220220184) supported by the Research&Development Operational Program funded by the European Regional Development Fund.

References

[1] K. Hayrinen, K. Saranto, and P. Nykanen, "Definition, structure, content, use and impacts of electronic health records: A review of the research literature," International Journal of Medical Informatics, vol. 77, no. 5, May 2008, pp. 291-304, doi:10.1016/j.ijmedinf.2007.09.001.

[2] M. Penhaker, M. Cerny, and J. Floder, "Embedded Biotelemetry System for Home Care monitoring," Proc. International Congresses on Bioelectromagnetism (ICBEM 2007), Aizu Wakamatsu, Japan, Oct. 2007, pp. 122-126, ISBN 978-4-9903873-0-3.

[3] J. Predáč and S. Predáčová, Handbook of symptothermal method PPR, 2nd ed., Olomouc: Matice cyrilometodějská, 2006, pp. 125, ISBN 80-7266-244-9.

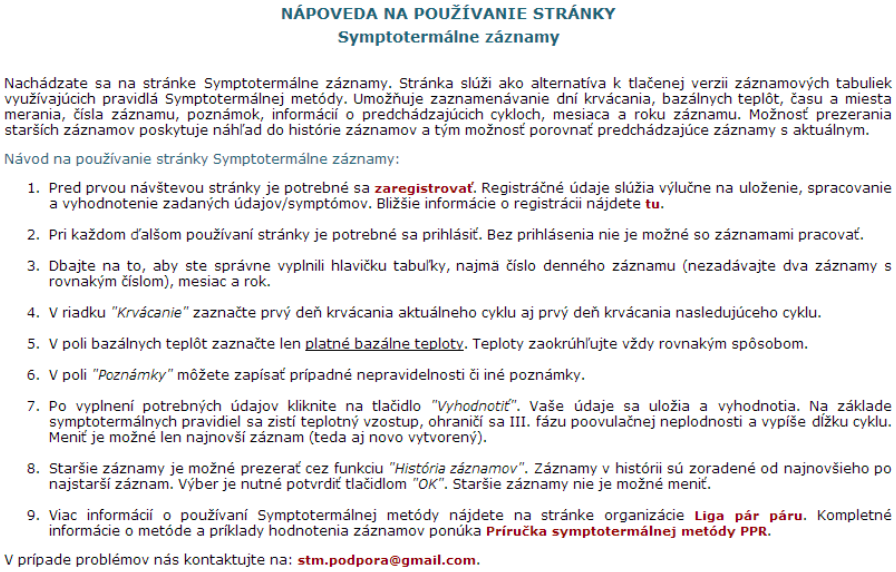
[4] W. L. Larimor and J. B. Standford, "Postfertilization Effects of Oral Contraceptives and Their Relationship to Informed Consent," in Arch Fam Med, vol. 9, Feb. 2000, pp. 126-133.

[5] A. Roztočil, Modern gynecology, Praha: Grada Publishing, 2011, pp. 508, ISBN 978-80-2472-832-2.

[6] L. Rob, A. Martan, and Karel Citterbart, Gynecology, Praha: Galen, 2008, pp. 319, ISBN 978-80-7262-501-7.

[7] E. L. Billings, J. J. Billings, and M. Catarinich, Atlas of the ovulation method, Trnava: Spolok svätého Vojtecha, 2011, pp. 103, ISBN 978-80-7162-859-0.

[8] Š. Borik and I. Čáp, "Investigation of Pulse Wave Velocity in Arteries," Proc. 35th International Conference on Telecommunications and Signal Processing (TSP 2012), Brno, Czech Republic, July 2012, pp. 562-565, ISBN 978-1-4673-1116-8.


NÁPOVEDA NA POUŽÍVANIE STRÁNKY
Symptotermálne záznamy

Nachádzate sa na stránke Symptotermálne záznamy. Stránka slúži ako alternatíva k tlačenej verzii záznamových tabuliek využívajúcich pravidlá Symptotermálnej metódy. Umožňuje zaznamenávanie dní krvácania, bazálnych teplôt, času a miesta merania, čísla záznamu, poznámok, informácií o predchádzajúcich cykloch, mesiaca a roku záznamu. Možnosť prezerania starších záznamov poskytuje náhľad do histórie záznamov a tým možnosť porovnať predchádzajúce záznamy s aktuálnym.

Návod na používanie stránky Symptotermálne záznamy:

1. Pred prvou návštevou stránky je potrebné sa **zaregistrovať**. Registráčné údaje slúžia výlučne na uloženie, spracovanie a vyhodnotenie zadaných údajov/symptómov. Bližšie informácie o registrácii nájdete **tu**.
2. Pri každom ďalšom používaní stránky je potrebné sa prihlásiť. Bez prihlásenia nie je možné so záznamami pracovať.
3. Dbajte na to, aby ste správne vyplnili hlavičku tabuľky, najmä číslo denného záznamu (nezadávajte dva záznamy s rovnakým číslom), mesiac a rok.
4. V riadku *"Krvácanie"* zaznačte prvý deň krvácania aktuálneho cyklu aj prvý deň krvácania nasledujúceho cyklu.
5. V poli bazálnych teplôt zaznačte len platné bazálne teploty. Teploty zaokrúhľujte vždy rovnakým spôsobom.
6. V poli *"Poznámky"* môžete zapísať prípadné nepravidelnosti či iné poznámky.
7. Po vyplnení potrebných údajov kliknite na tlačidlo *"Vyhodnotiť"*. Vaše údaje sa uložia a vyhodnotia. Na základe symptotermálnych pravidiel sa zistí teplotný vzostup, ohraničí sa III. fázu poovulačnej neplodnosti a vypíše dĺžku cyklu. Meniť je možné len najnovší záznam (teda aj novo vytvorený).
8. Staršie záznamy je možné prezerať cez funkciu *"História záznamov"*. Záznamy v histórii sú zoradené od najnovšieho po najstarší záznam. Výber je nutné potvrdiť tlačidlom *"OK"*. Staršie záznamy nie je možné meniť.
9. Viac informácií o používaní Symptotermálnej metódy nájdete na stránke organizácie **Liga pár páru**. Kompletné informácie o metóde a príklady hodnotenia záznamov ponúka **Príručka symptotermálnej metódy PPR**.

V prípade problémov nás kontaktujte na: **stm.podpora@gmail.com**.

Figure 4. Help for users (This time is the help available only in Slovak).

Patient-centric Data Warehouse Design

An Empirical Study Applied in Diabetes care

Lichin Chen
Graduate Institute of Biomedical Electronics and Bioinformatics
National Taiwan University
Taipei, Taiwan
e-mail: d98945012@ntu.edu.tw

Chiou-Shiang Wang, I-Ching Wang, Hui-Chu Yu
Department of Nursing
National Taiwan University Hospital, National Taiwan University
Taipei, Taiwan

Hui-Yu Peng, Hui-Chuen Chen
Department of Dietetics
National Taiwan University Hospital, National Taiwan University
Taipei, Taiwan

Chia-Hsiun Chang, Tien-Jyun Chang, Yi-Der Jiang, Lee-Ming Chuang
Department of Internal Medicine
National Taiwan University Hospital, National Taiwan University
Taipei, Taiwan

Feipei Lai
Graduate Institute of Biomedical Electronics and Bioinformatics,
Department of Computer Science and Information Engineering,
Department of Electrical Engineering
National Taiwan University
Taipei, Taiwan

***Abstract* — There are many healthcare settings dedicated for different diseases and treatments. As the information of the patients is collected in different healthcare settings, there is an opportunity to reveal further knowledge and understanding of the patients when the information is integrated. However, the services and evaluation are often isolated. The aim of this research is to design a data warehouse based on a patient-centric vision and to integrate the data to enable inner knowledge exploration. The design is initially applied in the case of diabetes. This study initially integrated the data of regular visits and telehealthcare program of diabetes care, and the data of 61 patients, who participates in both service settings, are collected. The results show that the data inputted by the patients outside the medical institutes are reliable and do represent patient's conditions. The potential value of the data warehouse is promising, and it is valuable to integrate the data across healthcare settings based on a patient-centric vision.**

Data warehouse; diabetes; diabetes self-management education; data integration.

I. INTRODUCTION

Diabetes is a chronic disease, which refers to a person who has high blood sugar, either because the pancreas does not produce enough insulin, or because cells do not respond to the insulin that is produced [1]. Diabetes control relies heavily on patient self-management and life style adjustment. The treatments of diabetes are aimed to postpone and prevent the development of complications, including medications, regular screening, self-management skill education, and long-term follow-ups. Traditionally, patients visit the medical institutes every three month for laboratory tests, diabetes self-management education (DSME), and receive complication screening annually. The data are commonly collected and recorded in information systems and the patient outcomes are evaluated based on the three-month interval data. Recently, telehealthcare has become popular that is designed to enhance self-care behaviors for diabetic patients outside the medical institutes. It combines the information communication technology and the commercialized biometric sensor devices to address disease management at a distance and facilitates longitudinal health status monitoring [2]-[5]. Patients are to record their daily activities on the online self-management information system, and monitor their glucose and other vital signs with the glucometer or other biometric sensors. The online self-management information system is often integrated with the biometric sensors, which enable data uploaded automatically. Patients often record their data on a daily or weekly time interval. Nevertheless, the services mentioned above are usually isolated and rarely connected to one another, and the information systems and the databases are commonly scattered. When it comes to evaluate patient's performances and outcomes, the two services are commonly done individually.

ISBN: 978-1-61208-335-3

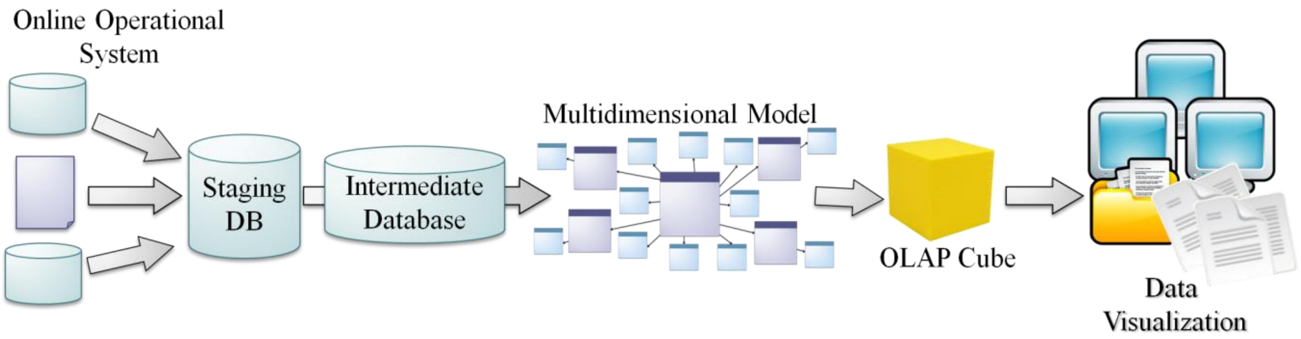


Figure 1 Process flow of data warehousing

However, integrating data from different perspectives offers an opportunity to further explore and understand patient conditions. Meanwhile, data warehousing has become prevalent in the healthcare industry because of the provision of manipulating large quantities of data and the decisions made based on the data [6]. It represents the process of data centralization, and duplicates the data from the online operational systems and organizes them into the analytical data structures. Data warehousing has the potential to integrate various data among applications, support massive data analysis, and improve the ability of future data mining, big data manipulation, and knowledge discovery [6]-[11]. The aim of this article is to design a data warehouse and offers an opportunity to integrate patient data and further explore patient conditions.

This article will first start with three hypotheses to evaluate patient condition and further introduce the data processing flow of the presented data warehouse in the method section. Based on the hypotheses, the result section presents the evaluation outcomes of patient conditions, and the result is interpreted in the discussion section, and concludes in the conclusion section.

II. METHODS

The aim of this research is to design a data warehouse, which is based on a patient-centric perspective, and further to explore the knowledge of diabetes through the integration of the data from regular visits and telehealthcare program of diabetes care. This study is applied in an educational hospital in northern Taiwan, which has highlighted the values of health information technology, and developed information systems for many clinical practices.

The hospital has implemented a shared care program for diabetes, which supports the patients with regular physician visits, DSME, laboratory tests, and routine follow-ups, and has additionally provided a telehealthcare program recently. The telehealthcare program allows the patients to interact with an online information system and upload daily glucose through an off-the-shelf, 3G glucometer [12]. Patients may participate in one of the programs, shared care program and telehealthcare program or both. The information generated from the services of diabetes patients is shown in Table I. It can be observed that the data are integrated from three healthcare settings and three heterogeneous databases.

After the construction of the data warehouse, this research initiates three hypotheses to evaluate patient's conditions.

1. In order to validate the reliability of the data given by the patients themselves. It is meaningful to find out how does daily glucoses correspond to the HbA_1c tests results?
2. Does telehealthcare program strengthen the skill of patients in choosing appropriate food ingredient after participating telehealthcare program?
3. Does telehealthcare program induce more performances in the frequency of self-monitoring of blood glucose (SMBG)?

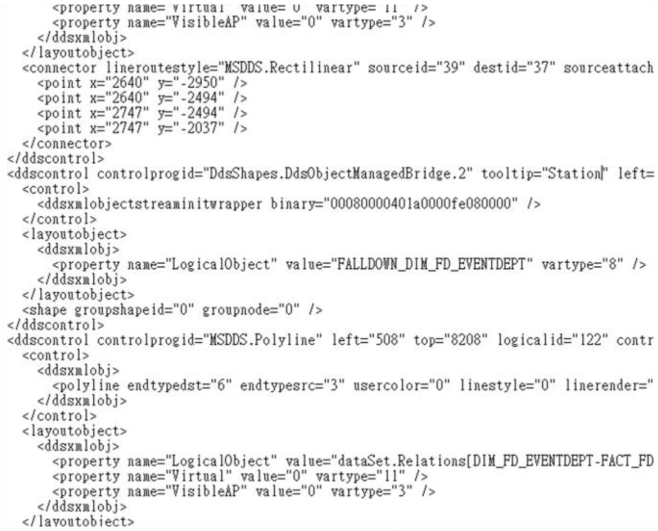
This research focused in analyzing the data of twenty-one-month duration, from September, 2011 to June, 2013, and compared patient's skills evaluated by the certificated diabetes educators (CDEs). The daily glucose monitoring data are grouped based on a three-month interval, which matched to the corresponding HbA_1c results. The skill evaluations discussed here consist of the eating behavior and the performing of SMBG. The eating behaviors of patients are evaluated through four aspects of food ingredient intake, including fiber, good fat, high fat, and sodium & desserts, the questionnaires are shown in Table II. Each question scores one point for achieving good behavior, and become the scoring of each ingredient intake. The frequencies of patients performing SMBG were recorded to see the effect of telehealthcare on patients performing SMBG. The frequency is calculated into weekly frequency and daily frequency. The first skill evaluation (T1) acts as the baseline of the self-care ability of patients, and the baseline is compared to the second (T2), and third (T3) evaluations if available.

A. Process of data warehousing

The data warehousing centralizes the data from the online operational systems and integrates them based on the analytical requirements. The data warehouse represents as a central data repository, and also consists of various kinds of defined analysis dimension tables, which enables the reusability among different analyses. The data warehouse applications commonly consist of three stages of data processing, including data transcription stage, data manipulation stage, and data visualization stage, as shown in Fig. 1. The data transcription stage is done through the use of Extract, Transform and Load (ETL) tools to duplicate the data from the online systems to a staging database in its original format. This is to ensure that the work of data warehousing does not interfere with the online systems.

TABLE I. INFORMATION IN DIABETES TREATMENT.

Information	Service	Source System	Database	Frequency
Diagnosis	Outpatient setting	Outpatient information system	Oracle	Every 3 month
Medication	Outpatient setting	Outpatient information system	Oracle	Every 3 month
Laboratory results	Outpatient setting	Laboratory information system	Oracle	Every 3 month
Diabetes education and assessment	Shared care program	Disease management information system	Sybase	Every 3 month
Self-management records	Telehealthcare	Online self-management information system	SQL server	Daily
Glucose and other vital signs monitoring	Telehealthcare	Online self-management information system	SQL server	Daily



```
      <property name="Virtual" value="0" vartype="11" />
      <property name="VisibleAP" value="0" vartype="3" />
    </ddsxmlobj>
  </layoutobject>
  <connector lineroutestyle="MSDDS.Rectilinear" sourceid="39" destid="37" sourceattach
    <point x="2640" y="-2950" />
    <point x="2640" y="-2494" />
    <point x="2747" y="-2494" />
    <point x="2747" y="-2037" />
  </connector>
</ddscontrol>
<ddscontrol controlprogid="DdsShapes.DdsObjectManagedBridge.2" tooltip="Station" left=
  <control>
    <ddsxmlobjectstreaminitwrapper binary="00080000401a0000fe080000" />
  </control>
  <layoutobject>
    <ddsxmlobj>
      <property name="LogicalObject" value="FALLDOWN_DIM_FD_EVENTDEPT" vartype="8" />
    </ddsxmlobj>
  </layoutobject>
  <shape groupshapeid="0" groupnode="0" />
</ddscontrol>
<ddscontrol controlprogid="MSDDS.Polyline" left="508" top="8208" logicalid="122" contr
  <control>
    <ddsxmlobj>
      <polyline endtypedst="6" endtypesrc="3" usercolor="0" linestyle="0" linerender="
    </ddsxmlobj>
  </control>
  <layoutobject>
    <ddsxmlobj>
      <property name="LogicalObject" value="dataSet.Relations[DIM_FD_EVENTDEPT-FACT_FD
      <property name="Virtual" value="0" vartype="11" />
      <property name="VisibleAP" value="0" vartype="3" />
    </ddsxmlobj>
  </layoutobject>
```

Figure 2 Example of XMLA

During data manipulation stage, the data are integrated, calculated, and manipulated in the intermediate database based on the topic and analysis requirements, which is known as the data marts. It is restructured into a Multi-dimensional Data Model (MDM) format based on the analytical requirements. Based on the MDM, the data are aggregated into an XML for Analysis (XMLA) format through on-line analytical processing (OLAP). The XMLA is also known as the OLAP data cube (shown as Fig. 2), which is designed to pre-calculate all the defined numerical values to facilitate instantaneous data queries and multi-factor comparisons [13]-[14].

Finally, the data visualization stage simply visualizes the data cube through a viewer. The users are free to drag the analyzing aspects into the analyzing column to obtain instantaneous queries and explore the data freely. The numbers of analyzing aspect are not limited. The MDM is usually interpreted into the structure of star schemas or snowflake schemas, which consists of a fact table in the middle and numerous dimension tables surrounding it. Each dimension represents a defined analysis aspect. Among the dimensions, few of the characters consisted of multiple-choice items that could refer to more than one option, such as medication. The star schema is suitable for singular items, and the snowflake schema is able to aggregate data in a normalized way. The relationships, attributes, and hierarchies between dimensions and fact tables are also defined during this stage.

TABLE II. QUESTIONNAIRES OF THE SKILL EVALUATIONS.

Ingredient	Questionnaire
Fiber	At least 2 meals consist of grain crops and rice each day.
	At least 1.5 bowl of vegetable a day.
	Eating fruit every day.
Food fat	Not taking in fat meat weekly.
	Not using pig fat at home.
	Not drinking complete fat milk.
	Not having saturated, trans fat and cholesterol-rich foods twice a week.
High fat	Not having deep fry more than 3 times a week.
	Not having 7 meals or more outside every week.
Sodium & desserts	Not having 2 times or more unplanned biscuits or dessert weekly.
	Not having pickled food 3 times or more weekly.

TABLE III. DEMOGRAPHIC INFORMATION OF PATIENTS.

		n	%
Diabetes Type	Type I	19	31.1
	Type II	42	68.9
Age	< 40	14	23.0
	40 ~ 65	38	62.3
	> 65	9	14.8
Gender	Male	30	49.2
	Female	31	50.8
Insulin	Yes	45	73.8
	No	16	26.2
Oral hypoglycemic agents	Yes	40	65.6
	No	21	34.4

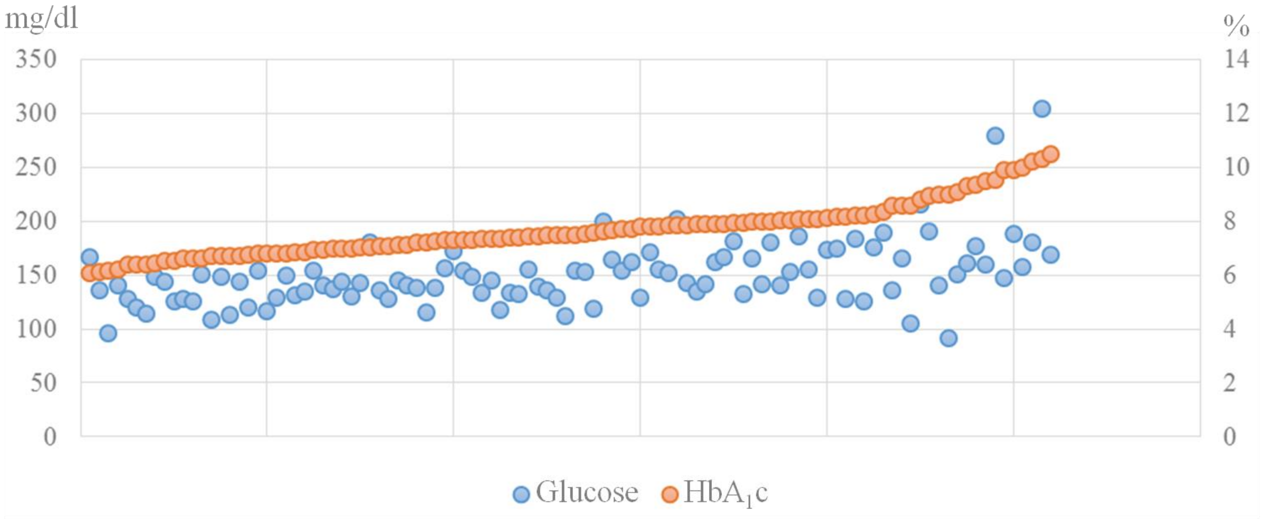


Figure 3 Relationship between daily glucose and HbA_1c

III. Results

A total of sixty-one patients participate in both service settings, and Table III shows the demographic information of the patients recruited. Fig. 3 illustrates the correlation of the daily glucose and the HbA_1c results, and the patients are arranged and listed according to their HbA_1c level. It can be observed that the two values are highly related and the daily glucose recorded by the patients does represent their outcomes.

TABLE IV. Scores of healthy eating.

	Ingredient (Score)	Mean	SD
T1 (n = 60)	Fiber score (3)	1.40	0.89
	Good fat score (4)	1.79	1.61
	High fat score (2)	0.69	0.74
	Sodium score (2)	0.51	0.75
T2 (n = 26)	Fiber score (3)	1.65	1.09
	Good fat score (4)	1.04	1.51
	High fat score (2)	0.38	0.64
	Sodium score (2)	0.27	0.60
T3 (n = 7)	Fiber score (3)	1.29	1.11
	Good fat score (4)	1.71	1.70
	High fat score (2)	0.57	0.79
	Sodium score (2)	0.71	0.95
T4 (n = 2)	Fiber score (3)	2.00	1.41
	Good fat score (4)	1.00	1.41
	High fat score (2)	0.50	0.71
	Sodium score (2)	0.50	0.71

Table IV shows the food ingredient score of the patients. Ninety-eight percent (n=60) of the patients were educated with healthy eating, but only 43% (n=26) of the patient were educated the second time, and even less were educated the third time. Comparing to T1, fiber score has increased in T2 and T4, good fat score and high fat score have decreased in T2, T3, and T4, and sodium score has decreased in T2 and T4. Table V shows the frequency of patients performing SMBG. 100% (n=61) of the patients were educated with SMBG, and 95% (n=58) of the patients were educated the second time. The frequency of SMBG has increased in T2 and T4.

TABLE V. Scores of blood glucose monitoring.

		Mean	SD
T1 (n = 61)	Weekly	7.42	9.62
	Daily	1.06	1.37
T2 (n = 58)	Weekly	8.34	10.13
	Daily	1.19	1.45
T3 (n = 38)	Weekly	7.30	11.07
	Daily	1.04	1.58
T4 (n = 8)	Weekly	8.25	5.87
	Daily	1.18	0.84
T5 (n = 2)	Weekly	3.50	4.95
	Daily	0.50	0.71

IV. Discussion

The reuse of data and data warehousing has become more prevalent due to the large quantities of data stored and the amount of decisions based on the data. The data warehouse offers the opportunity of obtaining better information, which results in better quality of care. The data warehouse supports dynamic comparison among multiple factors and provides instantaneous query, and is a powerful tool in the development of protocols for treatments and further knowledge exploration of diseases. The cleaning and the verifying of data require a lot of efforts. When the original design of the online information system offers free text input to provide the convenience of data input, it is common to obtain data that are easy for human to interpret but difficult for the machine to recognize, such as inputting "Y" for yes and the entire word "YES" means the same, but requires extra definitions for the system to recognize and identify them as the same meaning.

The result shows that while there are still debates on the reliability of the data that are inputted by the patients outside the medical institutes, the result in this study shows that the data do correspond to the results of the blood tests, and implies that the data are reliable. The eating behavior and the performing of SMBG are basic skills in self-management for diabetes patients. The result shows an unstable variation but a moderately positive outcome for patients in the behavior of healthy eating and SMBG performing. Generally, fiber score has increased and the other scores have decreased. Less desirable outcomes appear in T3 for eating behaviors and SMBG performing, the causes of such condition require further study.

Missing data and incomplete records are common when collecting data from online information systems. Meanwhile, there are seven skills to be enhanced in self-care behaviors. It is unlikely for CDEs to educate individual skill repeatedly or over three times in a short period of time, and it is difficult see the changes of the patients without repeated evaluation. Also, the evaluations were done by four different CDEs, and the results may differ from one to another, which has become a limitation of this study. More work is needed to further explore the behavior and outcome of patients. It would be promising in revealing more information by adding the laboratory results into the analysis.

V. Conclusion

As the services and treatments of healthcare become more and more advanced, and the involvement of the information communication technology increases, the integration of healthcare services has become essential in future service development.

The integration the data across healthcare settings based on a patient-centric vision is valuable and is supportive in the development of the research field of integrated healthcare. This study initially explored the integration of the data of diabetes care and applied in the validation of three hypotheses about patient conditions. The result shows that daily glucoses do correspond to the HbA_1c tests results, and the data inputted by the patients are reliable. The result shows a moderately positive encouragement for patients in the behavior of healthy eating and SMBG performing. However, the causes of the changes require further study. More work is required to specify the role of data warehouse in the healthcare industry and patient care.

References

[1] L. Chen, H. C. Yu, H. C. Li, Y. V. Wang, H. J. Chen, I. C. Wang, C. S. Wang, H. Y. Peng, Y. L. Hsu, C. H. Chen, L. M. Chuang, H. C. Lee, Y. Chung, and F. Lai, "An architecture model for multiple disease management information systems," Journal of Medical Systems, 37(2): pp. 9931, 2013.

[2] H. Y. Chiu and C. M. Chen, "Telenursing: The Integration of Information Technology and Community Health Nursing," Yuan-Yuan Nursing, 4 (2), pp. 5-10, 2010.

[3] L. Heinemann, "Measuring glucose concentrations: daily practice, current and future developments," Journal of Diabetes Science and Technology, 2 (4), pp. 710-717, 2008.

[4] E. H. Wagner, "The role of patient care teams in chronic disease management," British Medical Journal, 320 (7234), pp. 569-572, 2000.

[5] S. Y. Liu, J.L. Hsiao, J. L. Shen, and H. Q. Li, "A Research for Information Integration in Case Management System of Diabetes Mellitus," The Journal of Nursing, 4 (2), pp. 169-179, 2006.

[6] M. Silver, T. Sakata, H. C. Su, C. Herman, S. B. Dolins, and M. J. O Shea, "Case study: how to apply data mining techniques in a healthcare data warehouse," Journal of Healthcare Information Management, 15 (2), pp. 155-164, 2001.

[7] T. T. Lee, C. Y. Liu, Y. H. Kuo, M. E. Mills, J. G. Fong, and C. Hung, "Application of data mining to the identification of critical factors in patient falls using a web-based reporting system," International Journal of Medical Informatics, 80 (2), pp. 141-150, 2011.

[8] M. de Mul, P. Alons, P. van der Velde, I. Konings, J. Bakker, and J. Hazelzet, "Development of a clinical data warehouse from an intensive care clinical information system," Computer methods and programs in biomedicine, 105, pp. 22-30, 2012.

[9] M. J. Ball, C. Weaver, and P. A. Abbott, "Enabling technologies promise to revitalize the role of nursing in an era of patient safety," International Journal of Medical Informatics, 69 (1), pp. 29-38, 2003.

[10] M. C. Tremblay, R. Fuller, D. Berndt, and J. Studnicki, "Doing more with more information: Changing healthcare planning with OLAP tools," Decision Support Systems, 43 (4), pp. 1305-1320, 2007.

[11] J. H. Lubowitz and P. A. Smith, "Current Concepts in Clinical Research: Web-Based, Automated, Arthroscopic Surgery Prospective Database Registry," Arthroscopy: The Journal of Arthroscopic & Related Surgery, pp. 425-428, 2011.

[12] L. Chen, L. M. Chuang, C. H. Chang, C. S. Wang, I. C. Wang, Y. Chung, H. Y. Peng, H. C. Chen, Y. L. Hsu, Y. S. Lin, H. J. Chen, T. J. Chang, Y. D. Jiang, H. C. Lee, C. T. Tan, H. L. Chang, and F. Lai, "Evaluating Self-Management Behaviors of Diabetic Patients in a Telehealthcare Program: Longitudinal Study Over 18 Months," Journal of medical Internet research, 15 (12), pp. e266, 2013.

[13] M. C. Tremblay, R. Fuller, D. Berndt, and J. Studnicki, "Doing more with more information: Changing healthcare planning with OLAP tools," Decision Support Systems, vol. 43, pp. 1305-1320, 2007.

[14] N. Prat and J. Akoka, "A UML-based data warehouse design method," Decision Support Systems, 42, pp. 1449-1473, 2006.

Parallel Implementations of Numerical Simulation of the Vascular Solid Tumour Growth Model under the Action of Therapeutic Agents (Chemo- and Antyangiotherapy).

Damian Borys, Krzysztof Psiuk-Maksymowicz, Sebastian Student and Andrzej Świerniak
Institute of Automatic Control
Silesian University of Technology
Gliwice, Poland
damian.borys@polsl.pl,
krzysztof.psiuk-maksymowicz@polsl.pl,
sebastian.student@polsl.pl,
andrzej.swierniak@polsl.pl

***Abstract*—The authors' main interest was to develop vascular solid tumour growth model, implement efficient numerical methods for simulations towards finding a solution of the model and trying to optimise the influence of different types of therapies. A system of partial differential equations was introduced in order to simulate the growth of tumour and normal cells as well as the dynamics of the diffusing nutrient and anti-angiogenic or chemotherapeutic factors within the tissue. We have implemented finite difference time-domain (FDTD) method, which was formerly shown to produce numerical stable solutions. In order to make calculations in larger space, which includes a complex three-dimensional structure of capillaries, a single processing unit is not sufficient. Hence, there is the need for using high computing power in order to obtain the results at reasonable time. Furthermore, over some computing space limit, the amount of memory required to compute the solution extends the capacity of single computing machine, making computer cluster is the only choice. We are comparing the implementation of the numerical method for multi-computer system (cluster) using the message passing programming (MPI) paradigm with massively parallel computing implementation using graphic computing accelerators. The code was written in C++ and compared with Matlab implementation with appropriate toolboxes (Parallel Computing Toolbox and Distributed Computing Server). In all cases, the use of parallel implementation speedups the simulation time in comparison to the standard implementation on a single processor computer. Our results showed that we can reduce the simulation time significantly, when we use parallel computing written in C++. The speedup depends on the size of the computation domain, available memory size, the type of processors used and realization accuracy. Parallelisation of the code allows to perform optimisation of therapeutic protocols included in the model.**

Keywords - *tumour growth model*; *parallel computations*; *message passing interface*; *CUDA*

I. Introduction

Solid tumour progression is inseparably connected with vascular network, surrounding its volume [1], [2]. Tumour needs to grow oxygen and nutrition factors, which will be delivered through the vascular network. That is why, considering the network as well, as its dynamics is crucial in more realistic models. In literature, we can find many aspects of solid tumour models, based on cellular automation [3], structured models [4], single cell-based models [5] and models based on physical mass and momentum equations [6]. It is possible to distinguish different phases of the tumour growth. There are many models which focus on one particular phase, for example on hiperplastic growth phase [7], tumour growth *in situ* [8], invasion [9], angiogenesis [10], or process of metastasis [11]. The microvascular network plays crucial role in development of the solid tumours. It constitutes a source of the nutrient for the tumour and enables its continuous growth. However, due to fast metabolism of the tumour cells hypoxic regions may occur causing creation of tumour necrosis sites. The phenomenon of hypoxia is important because it may lead to the process of angiogenesis and additionally is a reason of lower efficiency of different therapies. The model taking into consideration processes mentioned above was developed and its numerical solution has been performed. Independently of the type of mathematical model, calculation of its solution is always time and resources demanding (computations time or computer memory) [12]. Presented here, the model of vascular tumour growth is described by set of partial differential equations. We have implemented FDTD method which was already shown to produce numerical stable solutions. In order to make calculations in larger space, which include complex three-dimensional structure of capillaries, single processor computers are not sufficient. Hence, there is need to use more computing power to obtain the results in a reasonable time. We are comparing the implementation of the numerical method for multi-computer system (cluster) with the message passing programming paradigm [13] with massively parallel computing implementation using graphic computing accelerators (Nvidia CUDA) [14].

The structure of this article is divided into this introduction section, next the description of materials and methods used for simulations, mathematical model section and sections for results presentation and, at the end, for discussion of presented results.

II. Materials and methods

In order to find a solution for the mathematical model, appropriate numerical methods have had to be used. Among the explicit numerical methods, one-step Lax-Wendroff method [15] was chosen for transport equations, and the standard forward time centered space was chosen for the diffusion equations. Numerical simulations have been done on the basis of synthetic micro environment created to reflect real environment in the tissue. Except for the normal cells fraction, tumour cells fraction and ECM (see the model description in the next section), syntetic vascular network has been included in the environment. It creates the distribution pathways for nutritients, oxygen and therapeutic agents. Parameters of the model has been based on a literature.

Computations have been performed in Mathworks Matlab for testing purposes (finding optimal and stable numerical method, non-parallel implementation) and in C++ language for parallel version using MPI (Message Passing Interface) libraries and C language for CUDA with thurst, CUBLAS and STL libraries. Each implementation details are presented in Fig. 1. For CUDA entire computational domain is processed by graphic accelerator processing units. For MPI, domain is divided for some subdomains and sent to workers (slaves) to compute new subdomain. After that results are sent to master node and new domain (for next time step) is merged. The code is based on dynamic task allocation, so the number of workers (S) is lower than the number of subdomains. This technique keeps the balance of workers load. For both implementations parallelisation is done only in one time step. Next time step is dependent from the previous one, that is reason for which it has to be calculated sequentially.

Calculations were carried out using the computer cluster Ziemowit [16] funded by the Silesian BIO-FARMA project No. POIG.02.01.00-00-166/08 in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre in the Silesian University of Technology. Every node used for MPI calculations has 2 six-cores Intel Xeon CPUs and 36GB RAM. Computer for CUDA computing was equipped with Nvidia Tesla C2075 graphic accelerator and Intel Xeon processor.

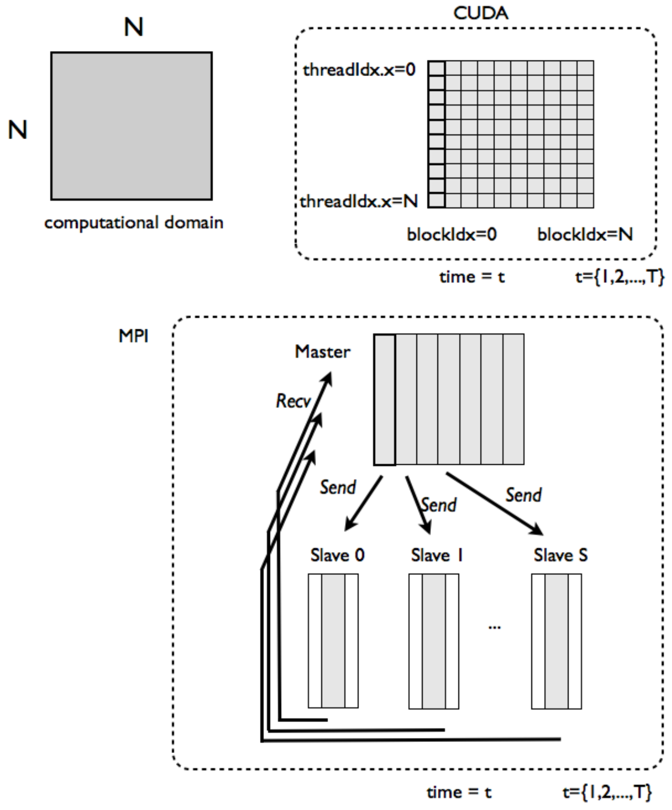


Figure 1. Diagrams of CUDA and MPI implementations for 2D computational domain. For CUDA whole computational domain is processed by graphic accelerator processing units. For MPI domain is divided for some subdomains and is sent to workers (slaves) to compute new domain.

III. Mathematical model

A set of partial differential equations was introduced in order to simulate growth of tumour and normal cells as well as the dynamics of the nutrient, anti-angiogenic and chemotherapeutic particles diffusing within the tissue. For modelling the tumour growth, different approaches are used. Unlike in [17], [18] we do not distinguish proliferative, quiescent and apoptotic cells. Cell behaviour is determined by the oxygen concentration in the tissue. The equations for the cell dynamics originate from the multiphase theory [19], [20], [21]. The main constituents of the multiphase part of the model are normal cells, tumour cells and extracellular matrix (ECM), thus variable n denotes volume fraction of normal cells, a denotes volume fraction of tumour cells, and m denotes volume fraction of the ECM. For simplicity volume fraction of ECM is assumed to be homogeneous and constant. The models in which the dynamics of the ECM is investigated can be found in works by Chaplain et al. [22] or by Psiuk-Maksymowicz [21]. The overall volume fraction occupied by the cells spread on the ECM must satisfy the inequality $\psi = n + a + m \leq 1$. In order to close the model, the porous media assumption is applied [23]. In order to provide physiological picture, the heterogeneity of concentration of the nutrient and xenobiotics is ensured. Mathematical model consists of five partial differential equations (PDE) :

$$\begin{cases} \dfrac{\partial n}{\partial t} = \nabla \cdot (n\,K\,\Sigma'\,\nabla\psi) + nF(c-c_P)[\alpha_n(1-\psi) - k_n d_{ch}] - \gamma_n nF(c_A - c), \\ \dfrac{\partial a}{\partial t} = \nabla \cdot (a\,K\,\Sigma'\,\nabla\psi) + aF(c-c_P)[\alpha_a(1-\psi) - k_a d_{ch}] - \gamma_a aF(c_A - c), \\ \dfrac{\partial c}{\partial t} = D_c \nabla^2 c - (k_{n_P} n + k_{a_P} a)F(c-c_P) - (k_{n_Q} n + k_{a_Q} a)F(c_P - c)F(c-c_A) + S_1(e), \\ \dfrac{\partial d_a}{\partial t} = D_{da}\nabla^2 d_a + S_2(e) - k_{da} d_a e - \lambda_{da} d_a, \\ \dfrac{\partial d_{ch}}{\partial t} = D_{dch}\nabla^2 d_{ch} + S_3(e) - k_{dch}(n+a)F(c-c_P) - \lambda_{dch} d_{ch}. \end{cases}$$

where K is a coefficient related to the permeability of the medium, Σ is a stress function, c stands for oxygen concentration, d_a stands for the concentration of anti-angiogenic treatment agent, d_{ch} stands for the concentration of chemical treatment agent, e stands for the binary function denoting occurrence of blood capillaries. Growth of the cells is of logistic type, where α_n and α_a stands for growth rate for normal and tumour cells, respectively. Normal and tumour cells undergo apoptosis with γ_n and γ_a rates, respectively. Growth and degradation of the cells is dependent on the oxygen availability, therefore in both terms sigmoid function $F(\cdot)$ is present. In growth terms it is dependent on the proliferation oxygen concentration c_P, and in degradation terms it is dependent on the apoptotic oxygen concentration c_A. In reaction-diffusion equations D_c, D_{da}, D_{dch} are present

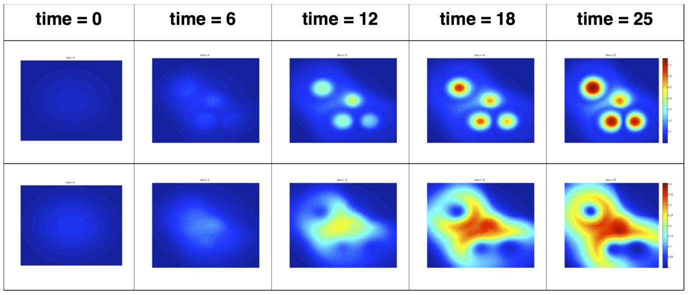


Figure 2. Spatial changes of the cellular density due to anti-angiogenic therapy (top row) and chemotherapy

denoting oxygen, anti-angiogenic agent and chemotherapeutic agent diffusion coefficients, respectively. Source terms are denoted by $S_i(\cdot)$, $i \in \{1, 2, 3\}$ functions dependent on the position of the blood vessels. Different types of stress-volume ratio relations can be taken into account. The simplest feature characterising stress is that below the value ψ_0 it vanishes, and increases for $\psi > \psi_0$ and tends to infinity as $\psi - 1$, e.g.

$$\Sigma(\psi) = E(1 - \psi_0) \left(\frac{\psi - \psi_0}{1 - \psi} \right)_+ , \qquad (1)$$

where $(f)_+$ denotes the positive part of f and E is the value of the derivative in$\psi = \psi_0$, a sort of Young's modulus for moderate compressions.

IV. Results

Example of model result with the drugs acting on healthy and tumour cells are presented in pictures collected in Fig. 2. The colors correspond to density of the cells after anti-angiogenic therapy (top row) and after chemotherapy (bottom row). For the top row, the higher density of the cells corresponds also with localisation of the vessels network.

The main results of our work present a comparison of the speedup of parallel implementation with the basic Matlab computations (Fig. 3). We have compared the speedups of MPI implementations with different domain sizes (100x100 and 400x400 points). The speedup is increasing up to about 12 cores then is slightly lower when the number of processing units increases (Fig. 3). This is caused by the architecture - single computing machine has 12 physical cores and, when increasing this number, we are causing that processes needs to communicate through the computer network which is always slower than shared memory architecture (even for Infiniband QDR connection). When spatial computational domain was increased 16 times the speedup increased up to 8 and the absolute computation times ratio increased maximum to about 10 times (Fig. 4). Comparing MPI (with 11cores) and CUDA implementation (Fig. 5) we can see that the speedup is higher for smaller domains but when increased the performance is significantly lower.

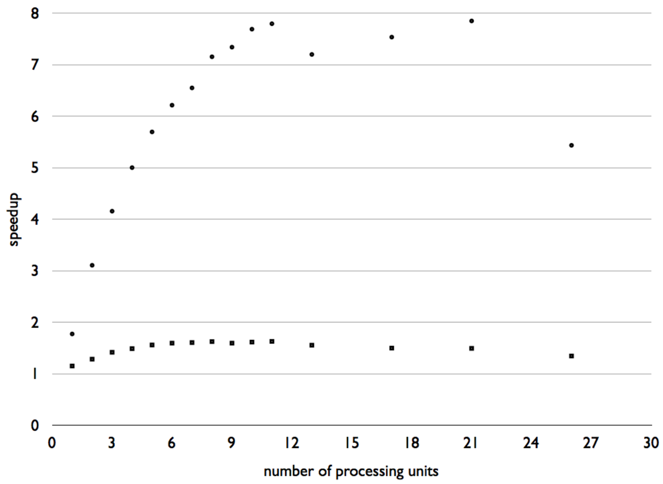


Figure 3. Speedup of the calculations time in dependence with the number of processing units. Two series are compared - with smaller spatial domain (100x100, boxes) and with large domain (400x400, circles).

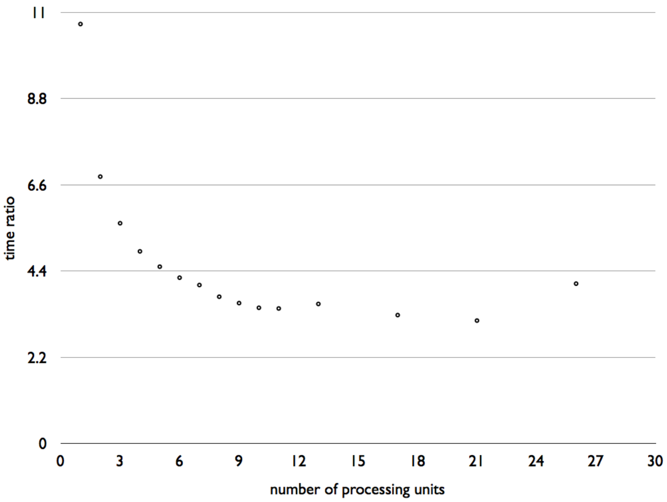


Figure 4. Time ratio after 16 times increase of the spatial domain (T400/T100) depending of the number of processing units.

V. Discussion

The presented results show that when the computing problem is relatively small, parallelisation using MPI technique and the usage of big cluster architecture is not the best choice, as long as the speedup is figurative. However, using CUDA architecture we can obtain very interesting results. With the growth of the size of the problem CUDA application meets its limitations related to memory bandwidth limits and hence MPI implementation seems to be a reasonable choice. We could also observe that, if the problem can be solved using single multi-core machine it will give us slightly better performance than using more machines.

Single simulation, having relatively small data domain (as in our case 400x400), is possible to compute using single computing machine and it takes about an hour (or less) to compute. However, switching the space domain to the third dimension only MPI implementation should be considered.

Parallelisation of presented numerical simulations serves us not only to study different methods of parallelisation performance, but it is a crucial step toward trying to find optimal therapeutic protocols of implemented chemo- and antyangiotherapy. To complete any optimization algorithm it is required to perform

ISBN: 978-1-61208-335-3

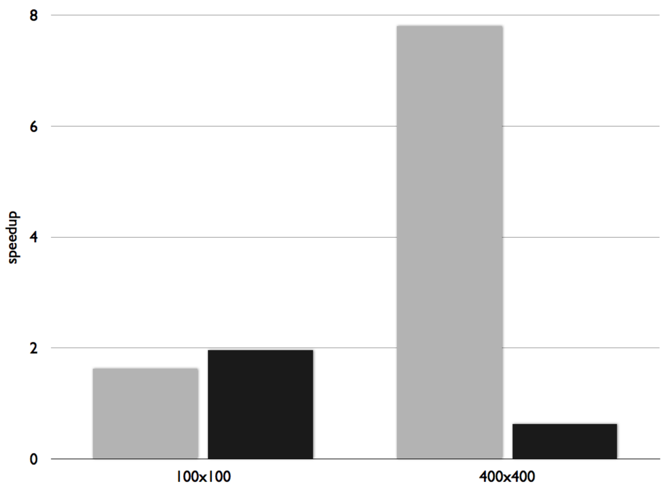


Figure 5. Speedup of MPI with 11cores (light grey colour) and CUDA (black) implementation for different spatial domain sizes (100x100 and 400x400).

thousands of repetitions of model simulation. Even for relatively small 2D domains and using a single computing machine the computation time is unacceptably long without using the parallelisation.

Further works will include implementation of meta-heuristic methods as simulated annealing, genetic algorithms, ant colony optimisation and others to find the optimal solution. However, this methods are inherently connected with multiple model simulation so even apparently small speedup of execution time, multiplicated during optimisation step, will contribute significanlty to the overall execution time.

ACKNOWLEDGMENT

This work was supported by the National Science Centre (NCN) in Poland under Grant No. N- N519-647840.

REFERENCES

[1] R. P. Araujo and D. L. S. McElwain, A history of the study of solid tumour growth: the contribution of mathematical modelling. Bull Math Biol, 66, 2004, pp:1039-1091.

[2] H. P. Greenspan. Models for the growth of a solid tumor by diffusion. Studies in Applied Mathematics, 4, 1972, pp:317-340.

[3] S. Dormann and A. Deutsch, Modeling of self-organized avascular tumor growth with a hybrid cellular automaton. In Silico Biology, 2, 2002, pp:393-406.

[4] F. Billy, J. Clairambault, F. Delaunay, C. Feillet and N. Robert, Age-structured cell population model to study the influence of growth factors on cell cycle dynamics. Math Biosci Eng, 10, 2013, pp:1-17.

[5] K. A. Rejniak, A single-cell approach in modeling the dynamics of tumor microregions. Math Biosci Eng, 2, 2005, pp:643-55.

[6] J. S. Lowengrub, et al., Nonlinear modelling of cancer: bridging the gap between cells and tumours. Nonlinearity, 23, 2010, pp:R1-R9.

[7] E. Mamontov, A. Koptioug and K. Psiuk-Maksymowicz, The minimal, phase-transition model for the cell-number maintenance by the hyperplasia-extended homeorhesis. Acta Biotheor, 54, 2006, pp:61-101.

[8] A.3rd Bankhead, N. S. Magnuson and R. B. Heckendorn, Cellular automaton simulation examining progenitor hierarchy structure effects on mammary ductal carcinoma in situ. J Theor Biol, 246, 2007, pp:491-498.

[9] Y. Kam, K. A. Rejniak, and A. R. Anderson, Cellular modeling of cancer invasion: integration of in silico and in vitro approaches. J Cell Physiol, 227, 2012, pp:431-438.

[10] M. A. Chaplain and A. R. Anderson, Mathematical modelling, simulation and prediction of tumour-induced angiogenesis. Invasion Metastasis, 16, 1996, pp:222-234.

[11] J. A. Sherratt, Predictive mathematical modeling in metastasis. Methods Mol Med, 57, 2001, pp:309-315.

[12] G. Evans, J. Blackledge, and P. Yardley, Numerical Methods for Partial Differential Equations. Springer London, 1999.

[13] The Message Passing Interface (MPI) standard, http://www.mcs.anl.gov/research/projects/mpi/, [retrieved 03.2014]

[14] Nvidia CUDA Toolkit, https://developer.nvidia.com/cuda-toolkit, [retrieved 03.2014]

[15] S. Larsson and V. Thomee, Finite Difference Methods for Hyperbolic Equations. Springer Berlin Heidelberg, 2003.

[16] Ziemowit cluster web page, http://www.ziemowit.hpc.polsl.pl, [retrieved 03.2014]

[17] F. Billy, et al., A pharmacologically based multiscale mathematical model of angiogenesis and its use in investigating the efficacy of a new cancer treatment strategy. J Theor Biol, 260, 2009, pp:545-62.

[18] P. Macklin and J. Lowengrub, Nonlinear simulation of the effect of microenvironment on tumor growth. J Theor Biol, 245, 2007, pp:677-704.

[19] H. M. Byrne, J. R. King, D.L.S. McElwain, and L. Preziosi, A two-phase model of solid tumour growth. Appl Math Lett, 16, 2003, pp:567-573.

[20] H. Byrne and L. Preziosi, Modelling solid tumour growth using the theory of mixtures. Math Med Biol, 20, 2003, pp:341-366.

[21] K. Psiuk-Maksymowicz, Multiphase modelling of desmoplastic tumour growth. J Theor Biol, 329, 2013, pp:52-63.

[22] M.A.J. Chaplain, L. Graziano, and L. Preziosi, Mathematical modelling of the loss of tissue compression responsiveness and its role in solid tumour development. Math Med Biol, 23, 2006, pp:197-229.

[23] D. Ambrosi and L. Preziosi, On the closure of mass balance models for tumor growth. Mathematical Models and Methods in Applied Sciences, 12, 2002, pp:737-754.

Tracking Action Potentials of Nonlinear Excitable Cells Using Model Predictive Control

Md. Ariful Islam, Abhishek Murthy,
Tushar Deshpande, Scott D. Stoller
and Scott A. Smolka

Department of Computer Science
Stony Brook University
Stony Brook, New York 11794
Email: {mdaislam,amurthy,tushard,
stoller,sas}@cs.sunysb.edu

Ezio Bartocci and Radu Grosu

Department of Computer Engineering
Vienna University of Technology
Vienna, Austria
Email: {ezio.bartocci,radu.grosu}
@tuwien.ac.at

***Abstract*—We present explicit and online Model Predictive Controllers (MPCs) for an excitable cell simulator based on the nonlinear FitzHugh-Nagumo model. Despite the plant's nonlinearity, we are able to formulate the model predictive control problem as an instance of quadratic programming, using a PieceWise Affine (PWA) abstraction of the plant. The speed-versus-accuracy tradeoff for the explicit and online versions is analyzed on various reference trajectories. Our MPC-based approach, enabled by the PWA abstraction, presents a framework for designing automated *in silico* biomedical control strategies for excitable cells, such as cardiac myocytes and neurons.**

***Keywords*–Biocomputing; Model Predictive Control; Excitable Cells.**

I. Introduction

Excitable cells, like neurons and cardiac myocytes, are building blocks of mammalian organ systems like the nervous and the cardiovascular systems. They exhibit characteristic cyclical responses to electrical stimuli, which could be provided externally or by neighboring cells via diffusion. The response is observed in terms of the change in their transmembrane potential in time and is called the Action Potential (AP). The cells are arranged contiguously to form the corresponding tissue. The periodic electrical excitation and diffusion at the cell-level leads to emergent patterns of electrical-wave propagation at the tissue-level [1]. Anomalous patterns at the tissue-level are associated with potentially fatal disorders like epilepsy and cardiac arrhythmias. For example, *reentry*, which corresponds to spiral waves in the cardiac tissue, is a precursor of Atrial Fibrillation (AFib) [2].

Controlling the cell-level response is critical for countering abnormal patterns at the tissue level. Excitable cells have the following distinguishing features that pose challenges to designing effective control strategies.

- **Nonlinearity**: The state space models for neurons and cardiac myocytes have highly nonlinear vector fields, which leads to multiple time scales.
- **Noise**: An actuator controlling a biological entity receives noisy readings corresponding to the state of the plant. Thus, robustness is critical while designing a control law.
- **Dimensionality**: Excitable cells are large dynamical systems and many state variables could be non-observable.

Model predictive control, a widely used process-control strategy, is well suited for biomedical applications involving excitable cells. It involves solving a finite horizon open-loop optimal control problem subject to the dynamics of the *plant*, which is the system to be controlled. Based on the measurements obtained at time T, the controller predicts the dynamic behavior of the system over a prediction horizon (T_p) and optimizes the control input over a control horizon ($Tc < T_p$) such that a predetermined open-loop performance objective function is minimized [3]. The objective function usually measures the plant's divergence from a prescribed reference trajectory, and thus is minimized by the controller. Disturbances and model mismatch constrain the controller's performance. The optimization can either be performed online (for accuracy) or can be done offline (for speed).

We present implementations of both the online and offline strategies for model predictive control of a neuron. Specifically, MPCs for a nonlinear model-based simulator of an neuron are presented. Fast and accurate model predictive control of excitable cells can be used for in-silico testing of biomedical control strategies, where a control law is designed and tested in software before fabrication. The biological entity being controlled is modeled using a simulator and the control law is tested on it, in software. Authors in [4] and [5] present novel strategies controlling and managing anomalous behaviors of neurons (epilepsy) and cardiac myocytes (ventricular tachycardia) respectively.

Model predictive control of plants with nonlinear dynamics, such as neurons, has garnered interest in the community, due to its unique challenges and wide-ranging applicability [3], [6]. The nonlinearity of the neuron dynamics results in an instance of nonlinear optimization to be solved during MPC. In general, nonlinear optimization is NP hard [7] and thus the implementation of online MPC is computationally expensive. Explicit MPC for nonlinear systems was proposed in [8], and has limited tool support. To circumvent these issues, *we adopt an approximately equivalent PieceWise Affine (PWA) abstraction of the nonlinear neuron model for both the online and explicit MPCs.* In [9], the Mixed Logical Dynamical (MLD) formalism was introduced for modeling systems whose state variables evolve continuously in time subject to logical constraints. The MPC problem for MLD systems was shown to be an instance of Mixed-Integer Quadratic Programming (MIQP). Later in [10], it was shown that PWA systems are equivalent to MLD systems. Thus, converting the nonlinear neuron model to an approximately equivalent PWA form transforms the corresponding MPC problem from an instance of nonlinear optimization to one of MIQP. Also, the PWA abstraction enables the design of explicit MPC by using the Multi-Parametric Toolbox (MPT) [11] in MATLAB.

Next, we outline the architecture of the controllers, see Fig. 1. The plant simulates an excitable cell and outputs the AP corresponding to the input stimuli provided by the MPC. The MPC's goal is to compute optimal inputs such that the plant tracks, in discrete time, a reference trajectory that consists of a nominal sequence of APs.

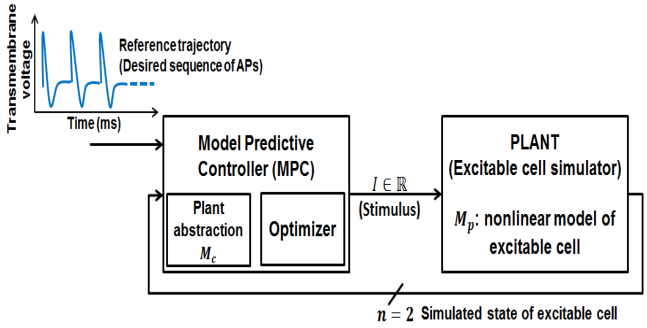

Transmembrane voltage
Time (ms)
Reference trajectory
(Desired sequence of APs)
Model Predictive Controller (MPC)
Plant abstraction M_c
Optimizer
$I \in \mathbb{R}$
(Stimulus)
PLANT
(Excitable cell simulator)
M_p: nonlinear model of excitable cell
$n = 2$ Simulated state of excitable cell

Figure 1. Architecture for tracking action potentials of nonlinear excitable cells using MPCs.

1) The plant uses a nonlinear model M_p of an excitable cell for its simulation. We use the FitzHugh-Nagumo (FHN) model [12], described in Section II, as M_p.
2) The plant outputs the n-dimensional state of M_p as the state of the underlying excitable cell. For the FHN model $n = 2$, and one of the state variables is the dimensionless transmembrane potential, which tracks the reference trajectory.
3) The MPC uses a PWA abstraction, M_C, of the plant model, to predict the behavior of the cell under simulation. We use a modified version of the hybrid model proposed in [13], henceforth referred to as the Dumas-Rondepierre (DR) model, as M_c. The PWA abstraction is used to cast the MPC's optimization problem as an instance of MIQP.
4) MPC is also equipped with an optimizer to compute the optimal stimulus input I, such that the observed state of the plant tracks a pre-defined reference trajectory.

The following simplifying assumptions are made in our implementation, and justified in the appropriate sections of the paper:

1) The plant's state is completely visible to the MPC. Ideally, only the membrane potential is measurable and the internal state is hidden.
2) No exogenous inputs (noise) are considered in the current implementation.
3) The only mismatch between the plant and its model, M_p, used by the MPC is due to the PWA abstraction.

We summarize our contributions below before outlining the remaining sections.

1) MPCs for a nonlinear model-based neuron have been designed by using a PWA abstraction. The resulting MIQP optimization instance is solved using both online and explicit approaches.
2) The PWA abstraction is used to enable the design of explicit MPC in MPT. The toolbox has been extended to track moving reference trajectories by augmenting the state vectors and thus making the penalty matrices time-varying.
3) The online and explicit approaches to the PWA abstraction-based MPC are compared using several test cases to analyze the tradeoff between accuracy and speed.

The remainder of the paper is organized as follows. The next section introduces the two models M_p and M_c in detail. We formulate the MPC problem for the FHN model-based plant in Section III. The implementation details follow in Section IV. Then, we compare and contrast the online and explicit strategies in Section V. We discuss related work in Section VI before concluding with directions for future work in Section VII.

II. Physiological Background

As mentioned in the previous section, excitable cells are characterized by their response to an external electric current, called the stimulus. Nonlinear Differential Equation Models (DEMs) capture the behavior of excitable cells in terms of the change in the transmembrane potential in time, as the cell oscillates between depolarization and repolarization in response to the stimulus.

The FHN model [12] is a two-dimensional system of differential equations, representing the dynamics of a neuron:

$$\dot{v} = v(1-v)(v-a) - w + I(t), \tag{1a}$$

$$\dot{w} = bv - cw, \tag{1b}$$

where v is the dimensionless transmembrane potential, w is a dimensionless recovery variable, I is the magnitude of the stimulus current and the parameters a, b and c are given in Table I.

TABLE I. Parameters of the FHN model (M_p) used by the plant to simulate an excitable cell.

Parameter	a	b	c
Value	0.20	0.05	0.01

The MPC uses a Modified DR (MDR) model [13], a PWA version of the FHN model, to predict the plant's behavior (simulation of the excitable cell). The cubic term in (1a) is linearized to obtain the PWA dynamics (2):

$$\dot{v} = \tilde{p}(v) - w + I(t), \tag{2}$$

where

Copyright (c) IARIA, 2014. ISBN: 978-1-61208-335-3

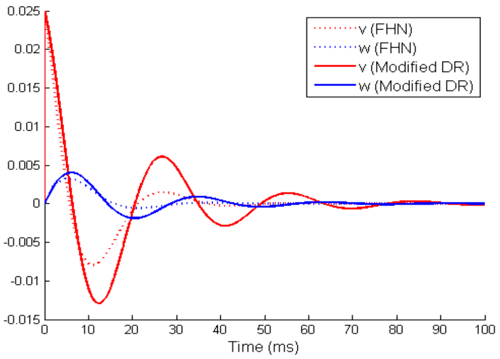


Figure 2. Simulations of the FHN and the MDR model. Maximum absolute L1 error for v = 0.0056 and w = 0.0013.

$$\tilde{p}(v) = \begin{cases} \frac{p(v_-)}{v_-}v & v < v_- \\ \left[\frac{p(v_+)-p(v_-)}{v_+-v_-}\right]v + \left[p(v_+) - \frac{p(v_+)-p(v_-)}{v_+-v_-}v_+\right] & v_- \leq v \leq v_+ \\ \frac{p(v_+)}{1-v_+}(1-v) & v > v_+. \end{cases} \tag{3}$$

The constants v_+ and v_- are given by

$$v_- = \frac{a + (1 - \sqrt{a^2 - a + 1})}{3}, \text{and} \tag{4a}$$

$$v_+ = \frac{a + (1 + \sqrt{a^2 - a + 1})}{3}. \tag{4b}$$

The function $p(.)$ is given by

$$p(v_-) = v_-(1 - v_-)(v_- - a), \text{and} \tag{5a}$$

$$p(v_+) = v_+(1 - v_+)(v_+ - a). \tag{5b}$$

The MDR model model represents M_c used by the controller to predict the plant's behavior and compute optimal stimuli values. It can be viewed as a hybrid model consisting of three modes and can be written in the following format:

$$\dot{\mathbf{x}} = A_i\mathbf{x} + B_i\mathbf{u} + \mathbf{f_i} \tag{6}$$

$$\mathbf{y} = C_i\mathbf{x} + D_i\mathbf{u} + \mathbf{g_i} \tag{7}$$

where

- i = index of the mode (piece),
- $\mathbf{x(t)}$ = $\mathbb{R}^n$ state vector,
- $\mathbf{u(t)}$ = $\mathbb{R}^m$ input vector,
- $\mathbf{y(t)}$ = $\mathbb{R}^p$ output vector,
- $A_i = n \times n$ the *Dynamics matrix* for mode i,
- $B_i = n \times m$ the Input matrix for mode i,
- $C_i = p \times n$ the Output matrix for mode i,
- $D_i = p \times m$ the Feed-through matrix,
- f_i = non-homogeneity in dynamics - real vector of size $n \times 1$ and
- g_i = real vector of size $p \times 1$.

In the MDR model, we have:

1) Three modes, i.e., $1 \leq i \leq 3$.
2) Two states v and w, i.e., $n = 2$.
3) One input, Stimulus I, i.e., $m = 1$.
4) Two outputs v and w produced by plant, i.e., $p = 2$ (all the states are observable).

The MPC works in discrete time. MATLAB's *c2dm* function was used, with a sampling rate of 0.01 sec/sample and the *zero-order hold (zoh)* method, to generate the discrete time version of the MDR model. We obtain the following matrices based on the parameters in Table I. The superscript "d" denotes the corresponding matrix in the discrete time version.

1) Mode 1, i = 1
   a) Invariant: $v < 0.0945$.
   b) $A_1 = \begin{bmatrix} -0.0955 & -1 \\ 0.05 & -0.01 \end{bmatrix}$, $A_1^d = \begin{bmatrix} -0.9990 & -0.01 \\ 0.0005 & 0.9999 \end{bmatrix}$.
   c) $B_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $B_1^d = \begin{bmatrix} 0.01 \\ 0 \end{bmatrix}$.
   d) $f_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $f_1^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
   e) $C_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $C_1^d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
   f) $D_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $D_1^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
   g) $g_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $g_1^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
2) Mode 2, i = 2
   a) Invariant = $0.0945 \leq v \leq 0.7055$.
   b) $A_2 = \begin{bmatrix} 0.1867 & -1 \\ 0.05 & -0.01 \end{bmatrix}$, $A_2^d = \begin{bmatrix} 1.0019 & -0.01 \\ 0.0005 & 0.9999 \end{bmatrix}$.
   c) $B_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $B_2^d = \begin{bmatrix} 0.01 \\ 0 \end{bmatrix}$.
   d) $f_2 = \begin{bmatrix} -0.0267 \\ 0 \end{bmatrix}$, $f_2^d = \begin{bmatrix} -0.0267 \\ 0 \end{bmatrix}$.
   e) $C_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $C_2^d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
   f) $D_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $D_2^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
   g) $g_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $g_2^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.
3) Mode 3, i = 3
   a) Invariant = $v > 0.7055$.
   b) $A_3 = \begin{bmatrix} -0.3566 & -1 \\ 0.05 & -0.01 \end{bmatrix}$, $A_3^d = \begin{bmatrix} 0.9964 & -0.01 \\ 0.0005 & 0.9999 \end{bmatrix}$.
   c) $B_3 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $B_3^d = \begin{bmatrix} 0.01 \\ 0 \end{bmatrix}$.
   d) $f_3 = \begin{bmatrix} 0.3566 \\ 0 \end{bmatrix}$, $f_3^d = \begin{bmatrix} 0.3566 \\ 0 \end{bmatrix}$.

e) $C_3 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $C_3^d = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.

f) $D_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $D_3^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

g) $g_3 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $g_3^d = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

Fig. 2 compares the FHN model and the MDR model in continuous time. A stimulus consisting of a spike of height 2.5 at the first time step was used to excite the model. The simulation was performed in MATLAB using the Euler method using a time step of 0.01 ms till 100 ms. Initial conditions were $v = 0$ and $w = 0$.

III. MPC Problem Formulation

Based on the $n \times 1$ state measurement (assuming that the complete state of the plant is observable: a state estimator would be needed in case of partial observability.) $\mathbf{x}(t)$ obtained at time t, the MPC predicts the dynamic behavior of the system and optimizes the control inputs such that the objective function in (8) is minimized:

$$\begin{aligned}
&\underset{U=[\mathbf{u}(t),\ldots,\mathbf{u}(t+N-1)]}{\text{minimize}} \quad J(U,\mathbf{x}(t)) = \\
&\sum_{k=1}^{N}[(\mathbf{x}(t+k)-\mathbf{x}_{ref}(t+k))'\lambda^k.Q(\mathbf{x}(t+k)-\mathbf{x}_{ref}(t+k)) \\
&\quad +(\Delta\mathbf{u}(t+k-1))'R(\Delta\ \mathbf{u}(t+k-1))] \\
&\text{subject to:} \\
&\mathbf{x}(t+k+1) = A_i^d\mathbf{x}(t+k) + B_i^d\mathbf{u}(t+k) + \mathbf{f}_i^d, \\
&\mathbf{y}(t+k) = C_i^d\mathbf{x}(t+k) + D_i^d\mathbf{u} + \mathbf{g}_i, \\
&\text{where} \\
&\Delta\mathbf{u}(t+k-1) = \mathbf{u}(t+k-1) - \mathbf{u}(t+k-2) \\
&0 \le k \le N-1 \text{ and } 1 \le i \le 3.
\end{aligned} \tag{8}$$

Optimization is performed over a finite horizon of length N. Q is an $n \times n$ identity matrix and $0 < \lambda \le 1$ is a parameter that assigns exponentially receding weights to the predicted deviations, $(\mathbf{x}(t+k) - \mathbf{x}_{ref}(t+k))$, over the horizon. Thus, the scheme is also called receding horizon control. R is a positive definite matrix that determines the penalty on differences between consecutive inputs.

The optimization problem is solved at time t and the inputs are calculated for the next N time steps. Only the next input is passed on to the plant, before repeating the MPC process.

IV. Implementation of Model Predictive Controllers for the FHN model-based Plant

Explicit and online MPCs were implemented for the FHN model-based plant using the MDR model for prediction purposes. The receding horizon parameter λ was fixed at 0.8 and $R = \begin{bmatrix} 10^{-3} \end{bmatrix}$ was the input penalty matrix. Next, we describe the implementation aspects of the online and explicit MPCs.

A. Online MPC

Online MPC involves solving (8) at every time step in runtime. The constrained nonlinear optimizer *fmincon* [14] was used to implement online MPC in MATLAB. At each time step t, the current state $\mathbf{x}(t)$ of the plant and the reference input $[\mathbf{x}_{ref}(t+1), ..., \mathbf{x}_{ref}(t+N)]$ over the finite horizon N were provided to the controller which then computed the optimal input for the FHN plant. An interior point algorithm was used for optimization. The FHN plant was then simulated using Euler method for one time step by applying the optimal input. This process was repeated for the whole simulation duration.

B. Explicit MPC

In explicit MPC, the optimization problem of (8) is cast as an instance of multi-parametric quadratic programming (mpQP) and solutions are computed offline for possibly overlapping polyhedral partitions, also known as *coverings*, of the state space. As shown in Fig. 3, the result of this one-time computation is a table of control laws corresponding to the partitions. At runtime, the current state sample is tested for membership in the list of partitions. The state may lie in more than one region due to possible overlap. In this case, the control law resulting in the most optimal value of the objective function is applied. The process is then repeated for the next state sample.

We implemented an explicit MPC for the FHN model-based neuron simulator using the MDR model as its PWA abstraction. The Multi Parametric Toolbox (MPT) [11] was used to implement the MDR-model based explicit MPC in MATLAB. The current implementation of MPT supports time-varying reference trajectories, but it considers constant reference at every time step of prediction horizon. We extended the tool to overcome this limitation. In the remaining sections, we elaborate on these modifications.

The key idea for incorporating time-varying reference trajectories is as follows. The reference trajectory over the prediction horizon, $\mathbf{x}_{ref}$ is considered to be to be a sequence of unknown variables. Then these unknown variables are used to augment the state vector $\mathbf{x}$. Then, the dynamics of the augmented system is reformulated in a Δu-form, as the input necessary to keep the states at the reference are also not generally known. In this formulation, the input at time k is $\Delta u(k)$, where $u(k-1)$ is an additional state in the dynamical model. So, the original system input can be obtained as $u(k) = u(k-1) + \Delta u(k)$. The state update equation is then given by (9).

$$\begin{pmatrix} x(k+1) \\ u(k) \\ x_{ref}(k+1) \end{pmatrix} = \begin{pmatrix} A_i & B_i & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} x(k) \\ u(k-1) \\ x_{ref}(k) \end{pmatrix} + \begin{pmatrix} A_i \\ B_i \\ 0 \end{pmatrix} \Delta u(k), \tag{9}$$

As the state vector is augmented with new state variables, the penalty matrix Q needs to be augmented, too. The newly augmented penalty matrix is given by

$$\begin{pmatrix} Q & 0 & -Q \\ 0 & 0 & 0 \\ -Q & 0 & Q \end{pmatrix}$$

In our modification scheme, we consider time-varying reference at all steps of the prediction horizon. We augment x with the reference state vector for all steps of the horizon. The

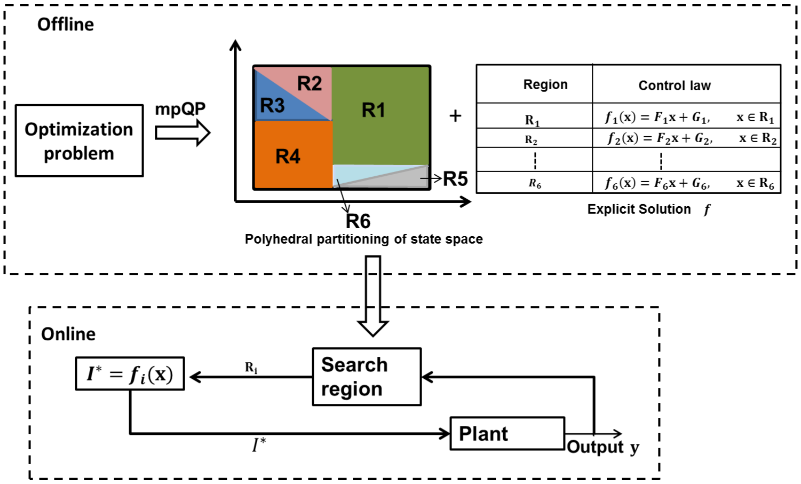


Figure 3. Workflow of Explicit MPC.

newly modified state update equation is given by (10)

$$\begin{pmatrix} x(k+1) \\ u(k) \\ x^1_{ref}(k+1) \\ \vdots \\ x^N_{ref}(k+1) \end{pmatrix} = \begin{pmatrix} A_i & B_i & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix} \begin{pmatrix} x(k) \\ u(k-1) \\ x^1_{ref}(k) \\ \vdots \\ x^N_{ref}(k) \end{pmatrix} + \begin{pmatrix} A_i \\ B_i \\ 0 \end{pmatrix} \Delta u(k), \tag{10}$$

where N is the number of steps in the horizon and $x^j_{ref}(k)$ is the reference vector in j^{th} time step of the horizon.

Due to the receding horizon principle, the penalty matrix will be different for each step of prediction horizon. Current implementation of MPT is amenable to add varying penalty matrix. The general form of the penalty matrix is given by

$$\begin{pmatrix} Q(k) & 0 & -Q(k) & 0 \\ 0 & 0 & 0 & 0 \\ -Q(k) & 0 & Q(k) & 0 \end{pmatrix},$$

where $Q(k) = \lambda^k Q$ for $k-th$ step of the prediction horizon.

The positions of $-Q(k)$ in the first row and $Q(k)$ in the third row are adjusted based on k.

V. Results

We conducted two sets of experiments on the online and the explicit MPCs to compare and contrast them.

Experiment Set 1: *Speed vs. Accuracy Tradeoff for Different Reference Trajectories*

The online and the explicit MPCs were tested against the same reference trajectory to study the speed vs. accuracy tradeoff. Two reference trajectories $[v^i_r(t), w^i_r(t)]$, $i = 1,2$ were generated by simulating the FHN model. The protocols used to generate them were as follows.

S1 Protocol for generating $[v^1_r(t), w^1_r(t)]$

1) Initial conditions: $v = 0$, $w = 0$ (rest conditions).
2) Time step used in the simulation: 0.1 ms.
3) Total duration of simulation: 240 ms (2400 time steps).
4) Stimuli pattern: One time-step-long (0.1ms) supra-threshold stimulus pulses of intensity (height) 1.5 were applied every 80 ms, to produce three APs in the simulation. Thus, the pacing frequency was 12.5 Hz.

S2 Protocol for generating $[v^2_r(t), w^2_r(t)]$

1) Initial conditions: $v = 0$, $w = 0$ (rest conditions).
2) Time step used in the simulation: 0.1 ms.
3) Total duration of simulation: 240 ms (2400 time steps).
4) Stimuli pattern: 10-steps-long (1 ms) supra-threshold stimulus pulses of intensity (height) 1.5 were applied every 80 ms, to produce three APs in the simulation. Thus, the pacing frequency was again 12.5 Hz.

The simulation was carried out using the Euler's method of numerical integration in MATLAB. Both the MPCs were tested against the S1 and S2 reference trajectories using a 3-step lookahead horizon. Their performance was compared using the following two metrics:

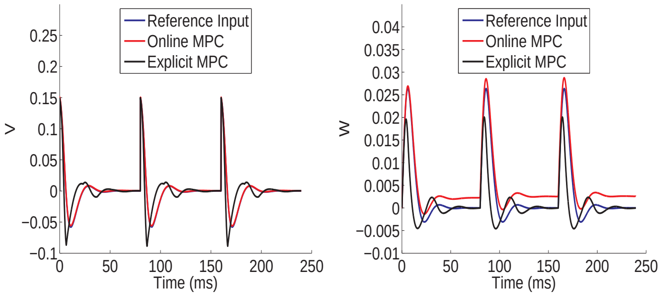
1) **Accuracy of the plant's operation** (μ^v_{l2}, μ^w_{l2}): measured using the mean L2 error between the reference trajectory and the output of the simulation carried out by the plant.
2) **Timeliness constraint on the MPC**: dictates that the working of the MPC must be fast enough to cope with the plant's operation. The degree to which the two MPCs met this constraint was measured as follows. The time taken by the Euler method-based simulation for producing the reference trajectory was noted, say t_1 secs. The plant + MPC combination was run on a single thread in a lock-step fashion, i.e., the plant was halted till the MPC finished its computation and provided the stimuli value for the next time step. The total time taken for tracking the reference trajectory was noted, say t_2 secs. Then, $t_{12} = (t_2 - t_1)$ provided an estimate of the time taken by the MPC to compute the stimuli. Ideally, $t_{12} < t1$, which ensures that the MPC's computation runs faster than the rate at which the plant evolves (simulates the FHN model).

Table II provides performance metrics for the two MPCs. Fig. 4 and Fig. 5 plot the evolution of v and w for protocols

S1 and S2, respectively.

TABLE II. PERFORMANCE METRICS FOR ASSESSING THE SPEED VS. ACCURACY TRADEOFF ACROSS DIFFERENT REFERENCE TRAJECTORIES.

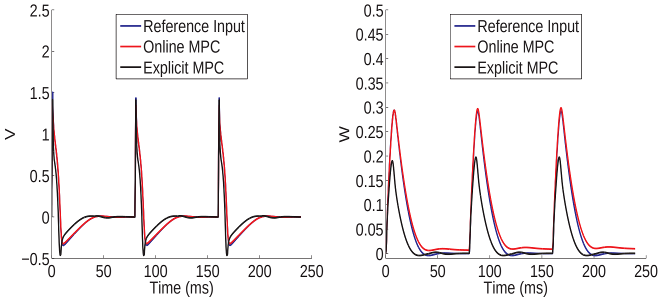
Protocol	Controller	μ^v_{l2}	μ^w_{l2}	$t_1(s)$	$t_2(s)$
S1	**Online MPC**	2.8×10^{-5}	4.9×10^{-5}	0.023	136.8
	Explicit MPC	4.3×10^{-4}	1.1×10^{-4}	0.023	88.8
S2	**Online MPC**	1.8×10^{-4}	2.2×10^{-4}	0.023	147.3
	Explicit MPC	2.7×10^{-2}	1.1×10^{-2}	0.023	88.3



(a) Evolution of v.

(b) Evolution of w.

Figure 4. Performance of the online and explicit MPCs on spike-shaped stimuli produced by the S1 protocol.



(a) Evolution of v.

(b) Evolution of w.

Figure 5. Performance of the online and explicit MPCs on rectangular pulse-shaped stimuli produced by the S2 protocol.

Discussion

Following inferences can be made from the results shown in the preceding paragraphs:

1) *The accuracy of the fmincon-based online MPC is better than the MPT-based explicit MPC.* This can be attributed to an accurate solution to the optimization problem, found by fmincon at run time, for the specific current state of the plant. The explicit MPC on the other hand, partitions the state space and finds a common control law for the whole partition. Accuracy is lost in this process.
2) *The fmincon-based online MPC is much slower than the MPT-based explicit MPC.* FMINCON exhaustively explores the whole state space at every time step of the plant's operation. This leads to its slow operation. The most time consuming step for the explicit MPC is searching for the partition corresponding to the current state, and this is achieved much faster than the online MPC's operation.

Experiment Set 2: *Effect of Horizon Length on Explicit MPC* Explicit MPC is enabled for time-varying reference trajectories by augmenting the state vectors and reformulating the dynamics. State space augmentation leads to an exponential increase in the number of polyhedral partitions. Table III compares the build-time and the number of partitions for different horizon lengths. Increasing the horizon length N is expected to improve the predictive accuracy of MPC. In the case of MPT-based design of explicit MPC, we observed that accuracy did not improve considerably on changing the horizon length from 2-step to 3-step lookahead. For both cases, the mean L2 errors for v and w were recorded to be around 0.027 and 0.001 respectively (for S2-type stimuli).

TABLE III. EFFECT OF N ON EXPLICIT MPC DESIGN IN MPT. HORIZON LENGTH OF 0, WHICH CORRESPONDS TO 0-STEP LOOKAHEAD, IS SPECIFIED FOR COMPARISON PURPOSES.

Horizon length	Build-time in MPT (secs.)	Number of partitions
0	0.45	3
1	2.78	30
2	51.36	277
3	576.41	2581

Having a smaller horizon leads to significant reduction in the search space, while searching for the partition corresponding to a given state sample. This reduction is critical as the search operation is performed at every time step during operation. In our case, t_2 reduced from 88.3 secs. to 11.2 secs when the horizon length was changed from 3 to 2 for the S2 protocol.

VI. RELATED WORK

MPC has been widely used in many domains like the chemical, food-processing, automotive, and aerospace industries. An exhaustive survey of both the theoretical and the practical aspects can be found in [15]. Explicit MPC, which is relatively new, has been surveyed in [16]. Recently MPC has found interesting biological and biomedical applications. In [17], [18], a platform for in silico realtime closed-loop control of gene expression in yeast has been proposed. It uses MPC to perturb inducible promoters in a systematic way to gain insights about gene expression. MPC has been successfully applied to devise therapeutic strategies in [19], [20], [21], [22]. In [23] MPC is used for functional electrical stimulation to estimate stimulation patterns for muscles that have been paralyzed due to spinal cord injury. Controllers for tracking neuron APs are designed in [24], [25] and thus are closest to our work. We compare and contrast each of them with our MPC-based approach below.

In [24], an adaptive input-output feedback linearization controller is presented to track a nominal AP using the FHN model. In contrast, MPC is a feed-forward control technique. Its predictive capability allows the controller to quickly adapt to a model mismatch caused due to the degradation/aging of excitable cells. Parameter estimation and tuning the model are the only steps involved in adapting to the changes, whereas a feedback controller needs complete redesign. Also, explicit MPC can track arbitrary fast-changing APs whereas the controller in [24] is designed for a nominal AP.

In [25], the authors present a sophisticated controller for a neuron based on the Hodgkin-Huxley (HH) model [26]. The HH model is augmented with random variables to capture stochastic behavior and external disturbances. Membrane potential is treated as the only observable and a state-estimator is employed for the other hidden variables of the HH model. On the other hand, we focus on comparing the online and explicit approaches to MPC in the case of a nonlinear plant being modeled using a PWA abstraction. The realistic setting of [25] is complementary to our work and provides directions for extending our scheme. Also, the FHN and the MDR models used in our work are order-reduced versions of the HH model.

VII. Conclusion and Future Work

Explicit and online MPCs were presented for tracking a reference sequence of APs using an FHN model-based neuron simulator. The MPCs employ a PWA abstraction of the nonlinear plant, thus enabling a QP formulation of the model predictive control optimization problem. The speed versus accuracy tradeoff was assessed using several test cases. The online approach provides excellent accuracy, but fails to satisfy the timeliness constraint. Offline MPC on the other hand, satisfies the timeliness constraint for a limited set of reference trajectories, but provides relatively lower accuracy than the online version.

We plan to pursue a combined approach that uses the best features of both the explicit and online MPCs to achieve high accuracy while satisfying the timeliness constraint. We also plan on adding noise to our implementation to test the robustness of the controller. Better QP solvers and search techniques will be explored to speed up the explicit MPC. The combined approach will then be examined for closed-loop stability and computational efficiency. In a realistic setting, the transmembrane potential is the only observable and state estimators would be added on the lines of [25]. We would like also to investigate real implementations of the proposed controllers using field-programmable gate arrays or other embedded microcontrollers. Finally, the MPC-based approach would be applied to complex excitable cells such as cardiac myocytes. In particular, we aim to investigate the use of piecewise multi-affine approximations of [27] to design explicit and online MPCs for cardiac cells.

Acknowledgments: We would like to thank the anonymous reviewers for their valuable comments. Research supported in part by NSF grants CCF-0926190 and CCF-1018459, and by AFOSR grant FA0550-09-1-0481.

References

[1] A. T. Winfree, “Heart muscle as a reaction - diffusion medium: The roles of electric potential diffusion, activation front curvature, and anisotropy,” International Journal of Bifurcation and Chaos, vol. 7, no. 3, March 1997, pp. 487–526.

[2] A. Kléber, “The fibrillating atrial myocardium. What can the detection of wave breaks tell us?” Journal of cardiovascular research, vol. 48, no. 2, August 2000, pp. 181–184.

[3] R. Findeisen and F. Allgöwer, “An introduction to nonlinear model predictive control,” in Proceedings of the 21st Benelux Meeting on Systems and Control, Veidhoven, 2002, pp. 1–23.

[4] K. N. Fountas et al., “Implantation of a closed-loop stimulation in the management of medically refractory focal epilepsy, a technical note,” Stereotactic and Functional Neurosurgery, vol. 83, no. 4, 2005, pp. 153–158.

[5] S. Luther et al., “Low-energy control of electrical turbulence in the heart,” Nature, vol. 475, 2011, pp. 235–239.

[6] S. J. Qin and T. A. Badgwell, “An overview of nonlinear model predictive control applications,” in Nonlinear Predictive Control. Verlag, 2000, pp. 369–392.

[7] S. O. Krumke, “Nonlinear optimization,” Lecture Notes, 2004.

[8] S. Summers, D. M. Raimondo, C. N. Jones, J. Lygeros, and M. Morari, “Fast explicit nonlinear model predictive control via multiresolution function approximation with guaranteed stability,” in 8th IFAC Symposium on Nonlinear Control Systems, 2010, pp. 533–538.

[9] A. Bemporad and M. Morari, “Control of systems integrating logic, dynamics, and constraints,” Automatica, vol. 35, no. 3, 1999, pp. 407 – 427.

[10] W. Heemels, B. D. Schutter, and A. Bemporad, “Equivalence of hybrid dynamical models,” Automatica, vol. 37, no. 7, 2001, pp. 1085 – 1091.

[11] “Multi-Parametric Toolbox (MPT),” 2004, URL: http://control.ee.ethz.ch/~mpt [accessed: 2014-03-25].

[12] E. M. Izhikevich, Dynamical Systems in Neuroscience: The Geometry of Excitability and Bursting. The MIT Press, 2007.

[13] J. G. Dumas and A. Rondepierre, “Modeling the electrical activity of a neuron by a continuous and piecewise affine hybrid system,” in Proceedings of the 6th international conference on Hybrid systems: computation and control. Berlin, Heidelberg: Springer-Verlag, 2003, pp. 156–171.

[14] “MATLAB Optimization Toolbox,” 2014, URL: http://www.mathworks.com/help/toolbox/optim [accessed: 2014-03-25].

[15] C. E. Garcia, D. M. Prett, and M. Morari, “Model predictive control: theory and practice a survey,” Automatica, vol. 25, no. 3, 1989, pp. 335–348.

[16] A. Alessio and A. Bemporad, “A survey on explicit model predictive control,” Nonlinear Model Predictive Control, Lecture Notes in Control and Information Sciences, vol. 384, 2009, pp. 345–369.

[17] J. Uhlendorf et al., “Long-term model predictive control of gene expression at the population and single-cell levels,” Proceedings of the National Academy of Sciences, vol. 109, no. 35, 2012, pp. 14 271–14 276.

[18] J. Uhlendorf, P. Hersen, and G. Batt, “Towards real-time control of gene expression: *in silico* analysis,” in Proceedings of the IFAC World Congress, vol. 18, 2011, pp. 14 844–14 850.

[19] H. Chang, A. Astolfi, and H. Shim, “A control theoretic approach to venom immunotherapy with state jumps,” in Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2010, pp. 742–745.

[20] T. Chen, N. F. Kirkby, and R. Jena, “Optimal dosing of cancer chemotherapy using model predictive control and moving horizon state/parameter estimation,” Computer Methods and Programs in Biomedicine, vol. 108, no. 3, 2012, pp. 973 – 983.

[21] S. L. Noble, E. Sherer, R. E. Hannemann, D. Ramkrishna, T. Vik, and A. E. Rundell, “Using adaptive model predictive control to customize maintenance therapy chemotherapeutic dosing for childhood acute lymphoblastic leukemia,” Journal of Theoretical Biology, vol. 264, no. 3, 2010, pp. 990 – 1002.

[22] Jeffry A. Florian Jr., J. L. Eiseman, and R. S. Parker, “Nonlinear model predictive control for dosing daily anticancer agents using a novel saturating-rate cell-cycle model,” Computers in Biology and Medicine, vol. 38, no. 3, 2008, pp. 339 – 347.

[23] S. Mohammed, P. Poignet, P. Fraisse, and D. Guiraud, “Toward lower limbs movement restoration with input-output feedback linearization and model predictive control through functional electrical stimulation,” Control Engineering Practice, vol. 20, no. 2, 2012, pp. 182 – 195.

[24] R. Naderi, M. J. Yazdanpanah, A. Azemi, and B. Roaia, “Tracking normal action potential based on the FHN model using adaptive feedback linearization technique,” in Proceedings of the IEEE International Conference on Control Applications (CCA), 2010, pp. 1458–1463.

[25] B.S. Chen and C.W. Li, “Robust observer-based tracking control of hodgkin-huxley neuron systems under environmental disturbances,” Neural computation, vol. 22, no. 12, 2010, pp. 3143–3178.

[26] A. L. Hodgkin and A. F. Huxley, “A quantitative description of membrane current and its application to conduction and excitation in nerve,” Journal of Physiology, vol. 117, 1952, pp. 500–544.

[27] R. Grosu et al., “From cardiac cells to genetic regulatory networks,” in Proceedings of Computer Aided Verification, ser. Lecture Notes in Computer Science, vol. 6806. Springer Berlin/Heidelberg, 2011, pp. 396–411.

Magnetic Resonance Angiogram Processing and Modelling of the Cerebral Vascular Network

Krzysztof Psiuk-Maksymowicz
Institute of Automatic Control
Silesian University of Technology
Gliwice, Poland
Email: krzysztof.psiuk-maksymowicz@polsl.pl

Jarosaw Śmieja
Institute of Automatic Control
Silesian University of Technology
Gliwice, Poland
Email: jaroslaw.smieja@polsl.pl

Damian Borys
SInstitute of Automatic Control
Silesian University of Technology
Gliwice, Poland
Email: damian.borys@polsl.pl

***Abstract*—The main aim of the work is to develop an algorithm for creation of three-dimensional model of the cerebral vascular network on the basis of medical images. Two types of non-contrast enhanced magnetic resonance angiograms were used as input data. First of them, Time of Flight angiography, enabled visualization of the arterial tree, whereas the next one, Phase Contrast angiography, enabled visualization of both arterial and venous trees. Medical images stored in Digital Imaging and Communications in Medicine (DICOM) file format required pre-processing. This process included resizing of the images, segmentation of the brain area and morphological white top-hat transformation in case of Time of Flight images, for noise reduction and gamma transformation necessary for enhancement of the vessels. The main core of the algorithm consists of segmentation of the vessel sections, selection of the vessels centroids, and construction of graph structure storing branches and nodes of the network. Three types of segmentation algorithms were analysed: algorithms with automatic and manual threshold segmentation (where threshold is based on image histogram); and binarization with hysteresis. The vascular trees obtained from two types of images were compared with each other. The developed algorithm allows for creation of three-dimensional model of cerebral vascular network and is potentially useful for the diagnosis of various vascular system abnormalities/diseases of the brain , as well as for the scientific simulations of blood flow or prediction of the drug distribution in the brain.**

***Keywords*-Medical image processing; image segmentation; MRI angiography; cerebral vasculature.**

I. Introduction

Contemporary Magnetic Resonance Imaging (MRI) technique allows for angiographic studies without the use of enhancing contrast agent. Cardiovascular imaging is crucial in diagnosis of vascular pathologies, *e.g.*, cerebral aneurysm. The most common sites of intracranial saccular aneurysms are anterior and posterior communicating arteries of the Circle of Willis.

The main aim of this study is to create an efficient algorithm for creation of the model of the cerebral blood vessels trees from three-dimensional medical MRI angiograms that reflects real structure of both venous and arterial vessels. The developed algorithm can be used not only for visualisation of the cerebral vascular system but also can be useful in detection of geometric deformations or abnormal narrowing of the blood vessels.

MRI images are stored in standardized DICOM file format [1]. Apart from the pre-processing improving the quality of the input data, the problem of extraction of vascular tree comes down to the problem of segmentation. There are many methods for medical image segmentation (see Chapter II in [2]). Segmentation techniques can be divided into classes in many ways, depending on classification scheme. The most commonly used segmentation techniques can be classified into two broad categories: region segmentation, and edge-based segmentation techniques. The most common region segmentation method is method of thresholding (*e.g.*, global or local (adaptive) thresholding) [2], [3]. Other region segmentation techniques include clustering, region growing, and watershed algorithms. Whereas, the edge-based segmentation algorithms include graph searching and contour following. The more sophisticated techniques use fuzzy clustering, neural networks, or deformable models.

Different segmentation techniques have different applications. Lesage *et al.* [4] review the techniques applicable for the vessel lumen segmentation. They also raise the topic of post-processing since initial extraction results may be lacking in different aspects: surface information may be missing or inaccurate, the vessels topology may be incorrect, non-vessel regions may be included, vessel segments may be disconnected or missing. An example of post-processing is skeletonization [5], which is used to determine the centerline of the vessel. In practice, image noise and the inherent limitations of 3D thinning algorithm may result in a skeleton that contains cycles and spurious spurs. Thus, it is necessary to perform skeleton pruning [6] in order to preserve unit-width skeleton without any bridges.

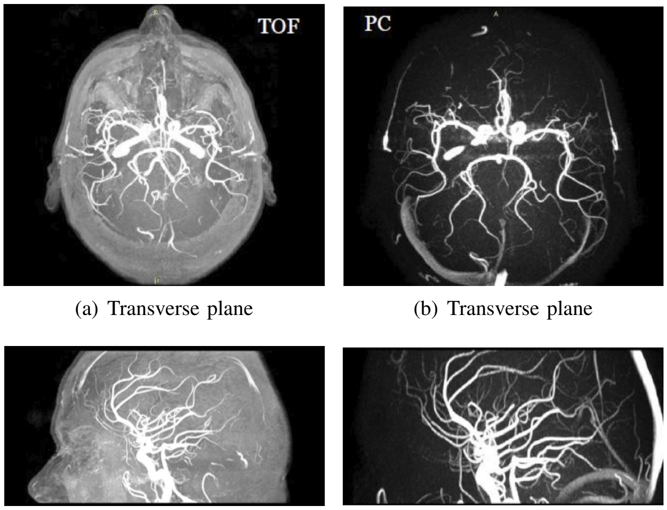
The rest of the paper is organized as follows: First, the research methodology is described in Section II. In Section III, we show the results of both image processing and algorytmic part of the work. Finally, in Section IV, we present the concluding remarks.

II. Materials and methods

Two types of non-contrast enhanced magnetic resonance angiograms were the input data of the developed algorithm. First of them, Time of Flight (TOF) angiography, enabled visualization of the arterial tree (see Fig. 1(a) and 1(c)), whereas

ISBN: 978-1-61208-335-3

the next one, Phase Contrast (PC) angiography, enabled visualization of both arterial and venous trees (see Fig. 1(b) and 1(d)). An advantage of the TOF and PC techniques over the tradicional MRI angiography is that there is no need to use contrast agent, which may have adverse effects. Maximum Intensity Projection (MIP) of both types of images is presented in Fig. 1. The quality of the PC image data was higher then of the TOF images. The PC images were of size 1140 x 1140 voxels, whereas TOF were of size 256 x 256 voxels.



(a) Transverse plane

(b) Transverse plane

(c) Sagittal plane

(d) Sagitall plane

Figure 1: MIP projection of magnetic resonance images performed by means of TOF (a), (c) and PC (b), (d) techniques.

Medical images are often subjected to pre-processing, for increasing the quality of images. It enables for a more effective segmentation. In order to reduce the noise of the images, frequently median or convolution filters are applied. In this work, median filters ware used. Second step in pre-processing was application of the morphological White Top-Hat (WTH) transformation [7]. The WTH transformation is defined as the difference between the input image and its opening by some structuring element. The resulting images highlights the sections of blood vessels; since, their intensities have higher values than the image background. All images were also subjected to nonlinear gamma correction [8] in order to improve the contrast of the images.

Additionally, for the set of TOF images, it was necessary to remove the skull oval.

For the purpose of segmentation, either thresholding segmentation or binarization with hysteresis were applied. The thresholds were chosen experimentally or automatically by means of the image histograms.

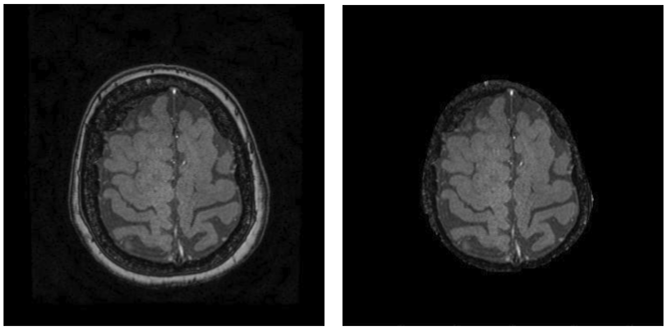
A significant part of the work was devoted to creation of object-oriented model of the vascular network structure. This work required:

- Selection of the vessels centroids on each slice,
- Iterative creation of branches from centroids, which subsequent cross-sections overlapped with each other,
- Saving the contours of subsequent sections for each branch,
- Capturing bifurcations, and storing it as node objects,
- Recursive attaching branches to nodal points.

The process of branches creation was complemented by a procedure of joining branches for those branches, which marginal cross-sections did not overlapped but were in close proximity. In such a case, changes in vessel circumference of subsequent cross-sections were analysed.

III. Results

All of the calculations were carried out in Matlab environment. Input TOF and PC images were subjected to successive image transformations. Median filtering of the images is not shown but the best results were obtained for 5x5 size of the mask. Figure 2 presents the outcome of the process of skull elimination from TOF image.

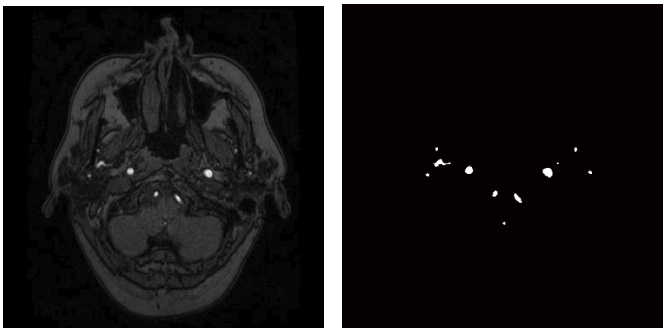


(a) Original MRI image.

(b) Image after skull removal.

Figure 2: Process of elimination of the skull from the TOF images.

In the case of gamma correction, the best results for TOF images were obtained for $\gamma = 1.2$, whereas for PC images it was $\gamma = 1.4$. For the white top-hat transformation different structural elements were analysed. The best element turned out to be a square element with 30 pixels side length.



(a) Original MRI image.

(b) Segmented vessels.

Figure 3: Process of vessel segmentation from TOF images with applied method of binarization with a hysteresis.

Different types of segmentation methods for vessel extraction were analysed. Those implemented were compared with methods implemented in graphical programs, *e.g.*, Yen method [9]. Nevertheless, those more advanced methods did not show any significant difference in comparison to the methods described in Section II. For the TOF images the best results were obtained for the binarization with hysteresis (see Fig. 3). While for the PC images the best segmentation results were obtained for binarization with threshold based on the image histogram.

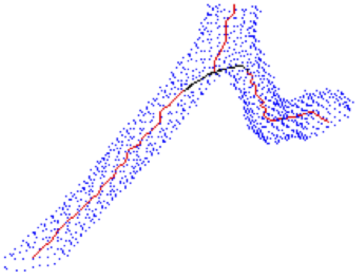
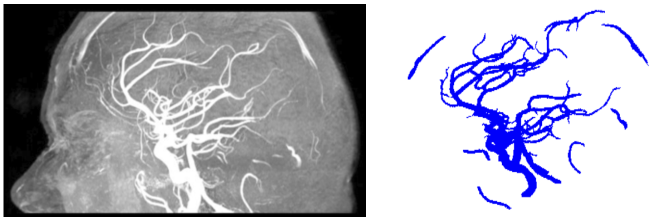


Figure 4: Visualization of two branching vessels with visible skeleton (in red) of the vessel.

The implemented algorithm for creating the object-oriented model of the vascular network structure is working properly. An example of two enlarged network segments is present on Fig. 4. One can see three-dimensional visualisation of the branch objects in red colour connected with nodal points by black lines, together with successive vessel contours presented in blue colour.



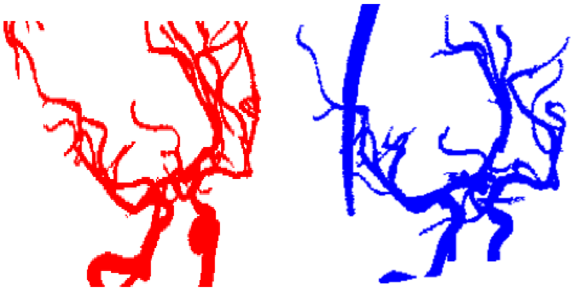
(a) MIP projection from TOF images.

(b) Visualization of the blood vessel network.

Figure 5: Comparison between input TOF data (a) and the result of developedalgorithm (b).

The comparison between input TOF data and visualisation of the model of arterial network for the whole brain is presented in Fig. 5. The selected method for the segmentation was binarization, with the threshold based on the image histogram. The advantage of the model is that one can selectively visualize only a part of the branches of the vascular network (or distinguish it with a different colour). Thanks to the model, the statistical analysis of the network parameters may be significantly reduced.

The comparison between the structure of the vessel network obtained by means of two different techniques is presented in Fig. 6. Images complement each other. Interestingly, despite the TOF images are of much lower resolution and contain more noise, this angiography technique gives more dense arterial network.



(a) Vessel structure obtained from TOF MRI.

(b) Vessel structure obtained from PC MRI.

Figure 6: Comparison between two vessel network obtained from TOF (a) and PC (b) MRi angiographies.

IV. Conclusion

The developed algorithm allows for creation of three-dimensional model of cerebral vascular network and is potentially useful for the visualisation and diagnosis of various vascular system diseases of the brain, as well as for the scientific simulations for blood flow or prediction of the drug distribution in the brain. The best segmentation results for TOF images were obtained by means of the binarization with hysteresis, whereas, the best segmentation results for the PC images were obtained for binarization with threshold based on the image histogram. Beyond the image processing, the work required the development of the algorithm of creation of blood vessel skeleton. The way of merging centroid points in the branches and bifurcations was no always trivial (in particular due to vessel twisting). The present work is still under development, it requires a quantitative analysis of the results. Application of two types of non-contrast MRI imaging is performed in order to extract the veins, which are present only at PC images together with the arteries. Therefore, the goal is not to improve the visualization of the entire vessel network but correct estimation of both 3D structures. The quality of the final model of cerebral vascular network depends strongly on the quality of the input data provided. Pre-processing and choice of segmentation method play an important role, however breaks in branches are mostly caused by imperfection of the acquisition method.

Applicability of the method is very broad. Visualization of the three-dimensional vascular network may have clinical applications, and the same network model can be used in further research. The obtained vascular tree structure of the brain can complement for example mathematical model describing vascular growth of tumours, mathematical modelling of blood flow, or models describing creation and growth of the blood vessels (*e.g.*, angiogenesis).

Acknowledgment

This work was supported by the National Science Centre (NCN) under Grant No. 2011/03/B/ST6/04384.

References

[1] Official www website of the Association of Electrical Equipment and Medical Imaging Manufacturers describing the DICOM standard http://medical.nema.org/standard.html, [retrieved: 03, 2014].

[2] I. N. Banckman, Handbook of Medical Imaging. Processing and Analysis. United States of America: Academic Press, 2000.

[3] M. Sezgin, and B. Sankur, ”Survey over image thresholding techniques and quantitative performance evaluation”, Journal of Electronic Imaging, vol. 13, Jan. 2004, pp. 146-165, doi: 10.1117/1.1631316.

[4] D. Lesage, E. D. Angelini, I. Bloch, and G. Funka-Lea, ”A review of 3D vessel lumen segmentation techniques: Models, features and extraction schemes,” Medical Image Analysis, vol. 13, Dec. 2009, pp. 819-845, doi:10.1016/j.media.2009.07.011.

[5] L. Verscheure, L. Peyrodie, A. S. Dewalle, and N. Reyns, ”Three-dimensional skeletonization and symbolic description in vascular imaging: preliminary results,” International Journal of Computer Assisted Radiology and Surgery, vol. 8, Mar. 2012, pp. 233-246, doi:10.1007/s11548-012-0784-4.

[6] Z. Chen and S. Molloi, ”Automatic 3D vascular tree construction in CT angiography,” Comuterized Medical Imaging and Graphics, vol. 27, Nov. 2003, pp. 469-479, doi:10.1016/S0895-6111(03)00039-9.

[7] P. Soille, Morphological Image Analysis. Berlin: Springer, 2004.

[8] Y. Shi, J. Yang, and R. Wu, ”Reducing Illumination Based on Nonlinear Gamma Correction,” Image Processing, 2007. ICIP 2007. IEEE International Conference on, vol.1, pp.I - 529,I - 532, Sept. 16 2007-Oct. 19 2007, doi:10.1109/ICIP.2007.4379008.

[9] J. C. Yen, F. J. Chang, and S. Chang, ”A new criterion for automatic multilevel thresholding,” Image Processing, IEEE Transactions on, vol.4, pp.370-378, Mar. 1995, doi:10.1109/83.366472.

Experimental Verification of the Quality of Clusterings Produced by Hard Clustering Algorithms After the Removal of Unstable Data Elements

Wim De Mulder*
Department of Electrical Energy,
Systems and Automation
University of Ghent
Ghent, Belgium
Email: wim.demulder@ugent.be

Zahra Zavareh*,
Konika Chawla
and Martin Kuiper
Department of Biology
Norwegian University of Science and Technology
Trondheim, Norway
Email: martin.kuiper@ntnu.no
*These authors contributed equally

***Abstract*—Many different clustering algorithms have been developed to detect structure in data sets in an unsupervised way. As user intervention for these methods should be kept to a minimum, robustness with respect to user-defined initial conditions is of crucial importance. In a previous study, we have shown how the robustness of a hard clustering algorithm can be increased by the removal of what we called unstable data elements. Although robustness is a main characteristic of any clustering tool, the most important feature is still the quality of the produced clusterings. This paper experimentally investigates how the removal of unstable data elements from a data set affects the quality of produced clusterings, as measured by the mutual information index, using three biological gene expression data sets.**

Keywords-hard clustering; cluster quality; unstable elements; mutual information context; microarray data.

I. Introduction

A. Introduction to cluster analysis

Clustering is an important approach for the analysis of large-sized data sets. Clustering partitions data sets into groups or clusters, such that data elements within the same group share a higher similarity than data elements that are member of different groups. Similarity is typically expressed in terms of a user-defined distance measure.

For large data sets, cluster analysis is often a necessary preprocessing step, since it organizes the data in manageable subsets. The field of bioinformatics widely uses clustering approaches for the analysis of often huge biological data sets with millions of data elements which today can be easily, cheaply and accurately measured with advanced functional genomics technologies (e.g., inferring the expression levels of all genes of an organism on a microarray [1]). However, these so-called high throughput technologies confront the biologist with the daunting challenge of analyzing massive data sets. Yet, for researchers in cluster analysis these data sets offer an interesting alternative to the low dimensional toy data sets that all too often are used to perform cluster analysis experiments on and to validate a certain hypothesized behavior of a clustering algorithm. Here, we focus on gene expression data sets, for which cluster analysis is more often than not a necessity before other specific biological analysis tools can be applied [2].

Since cluster analysis is an unsupervised method, the values for the parameters are typically chosen by simple rules of thumb. Examples of clustering algorithm parameters include the initial cluster centers, in case of k-means [3]; the maximal number of maintained edges, in case of the Memory Constrained-Unweighted Pair Group Method with Arithmetic Mean (MC-UPGMA) clustering algorithm [4]; the fuzzifier, in case of fuzzy c-means [5], etc. Since it is hard to accept that the intrinsic structure of a data set depends on some hit or miss values for these parameters chosen by a user, robustness with respect to these values is of crucial importance. In previous work [6], we introduced the concept of 'unstable data element' and showed that removing such data elements from data increases the robustness in terms of the measure called instability, introduced in the same paper. This previous work is shortly discussed in Section I-B, for convenience. A question we did not consider is how the quality of the result of a clustering algorithm is influenced by the removal of unstable data elements. The theorems we have proven only show that robustness is increased when the most unstable data elements are removed, but they do not exclude the possibility that as a side-effect the quality of the produced clusterings is adversely affected. In informal words, removing unstable data elements implies that a data set can be better clustered by a clustering algorithm, but it is possible that the better separable clusters are worse in terms of cluster validation measures. In this paper, we compare the quality of clusterings, produced by k-means, before and after the removal of unstable data elements, using three relatively large biological data sets that describe the activity of genes from an organism. The quality is measured using the mutual information index, a theoretically well-founded measure that is often used as cluster validation measure if external labels are available [7], [8], [9], [10].

The paper is organized as follows. Section I-B outlines our previous work. In Section II, we describe the three biological data sets that are used to investigate the research question mentioned above. Section II-C explains how the mutual information index can be used as cluster validation measure if gene annotations are available. In Section II-D, we recall from our previous work what we mean by the most robust clustering from a given sample of clusterings. Section III contains the experiments where we compare the quality of the most robust clustering of the data set after removal of unstable genes to clusterings of the data set before the removal of such genes.

B. Previous work

1) Introductory notions: The concepts and methods discussed in our previous work [6] apply to hard clustering algorithms. Such algorithms produce hard clusterings, meaning that every element is member of exactly one cluster, to full degree. K-means is the best known example of such an algorithm.

Given a data set $D = \{g_1, \ldots, g_n\}$, any hard clustering can be represented as a matrix C with elements $C(j,k), j = 1...n, k = 1...n$:

$$\begin{aligned} C(j,k) &= 1 \quad \text{if } g_j \text{ and } g_k \text{ are placed in different clusters} \\ &= 0 \quad \text{if } g_j \text{ and } g_k \text{ are placed in the same cluster} \end{aligned}$$

We defined the expected clustering $E[C]$ as the matrix that contains as elements $E[C](j,k) = E[C(j,k)]$, where the expected value is taken over all hard clusterings of the data produced by a given hard clustering algorithm and where randomness arises from the random selection of initial conditions. This matrix can be considered as independent of any specific choice of initial conditions, and thus maximally robust, since it is the uniquely defined probability-weighted sum over all possible clusterings generated by the given clustering algorithm. It is clear that the expected clustering is only a theoretical concept, i.e., it cannot be determined in practice. In practice, a sample of clusterings $\{C_1, \ldots, C_N\}$ is generated and the expected clustering is approximated by the average clustering $\bar{C}$ with elements $\bar{C}(j,k) = \frac{1}{N}\sum_{i=1}^{N} C_i(j,k)$.

2) Instability: We introduced the instability of a data element g_k:

$$\mu(g_k) = \frac{1}{n-1}\Big(\sum_{j=1}^{k-1} \sigma(\bar{C}(j,k)) + \sum_{j=k+1}^{n} \sigma(\bar{C}(k,j))\Big) \tag{1}$$

with

$$\begin{aligned} \sigma(a) &= 1 - a \qquad 0.5 \leq a \leq 1 \\ &= a \qquad\qquad 0 \leq a < 0.5 \end{aligned}$$

for $a \in [0,1]$.
We define the instability of a given clustering algorithm for a data set, as

$$\mu = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j<k\leq n} \sigma(E[C](j,k)) \tag{2}$$

In practice, the instability is approximated using the average clustering $\bar{C}$ corresponding to a sample of clusterings generated by the given clustering algorithm, for a data set, as follows:

$$\hat{\mu} = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j<k\leq n} \sigma(\bar{C}(j,k)) \tag{3}$$

For convenience, we will write μ instead of $\hat{\mu}$. The intuition behind the instability of a given clustering algorithm, for a given data set, is that it represents a measure for the difference between clusterings generated with different initial conditions. As such, it is an inverse measure for robustness. The concept of instability is more extensively described in [6].
It was proven that the instability of a clustering algorithm equals the average instability of the data elements:

Theorem 1.

$$\frac{1}{n}\sum_{k=1}^{n} \mu(g_k) = \mu$$

3) Cluster stability variance: We extended the variance of a random variable taking values on $\mathbb{R}$ to the variance of a hard clustering algorithm C, for a given data set: $\sigma^2(C) = E[d(C, E[C])^2]$ where $d(C, E[C])$ denotes the 'distance' from C to $E[C]$ which we defined as:

$$d(C, E[C]) = \frac{2}{n(n-1)} \sum_{j=1}^{n-1} \sum_{j<k\leq n} |C(j,k) - E[C](j,k)| \tag{4}$$

Randomness arises from the random choice of initial conditions. In practice, the variance is approximated by what we called the cluster stability variance (CSV) using a sample of clusterings $\{C_1, \ldots, C_N\}$:

$$CSV = \frac{1}{N-1}\sum_{i=1}^{N} d(C_i, \bar{C})^2 \tag{5}$$

The CSV is at least zero, and it is only zero when the produced clusterings are independent of the choice of initial conditions. The larger the CSV, the more dependent on initial conditions the produced clusterings are, for a given data set. In other words, the CSV is also an inverse measure for robustness.

4) Relationship between instability and cluster stability variance: We proved the following relationship between instability and CSV, given a sample of N clusterings:

Theorem 2.

$$CSV \leq \frac{N}{N-1}\mu$$

5) Reducing the instability: We showed that the instability is reduced by removing the most unstable data element from the data set:

Theorem 3.

$$\mu(g_l) = \max\{\mu(g_k) | 1 \leq k \leq n\} \Rightarrow \Delta\mu_l \leq 0$$

where $\Delta\mu_l$ represents the change in instability after removing g_l.
In other words, the instability of a clustering algorithm on any data can be increased by removing the most unstable data element. Due to Theorem 2 the CSV is also possibly reduced as a side effect. This process of removing the most unstable

data element can be repeated until the CSV attains a minimum or has stabilized.

6) Novelty of the described method: Our work has introduced two concepts, namely the instability of a data element and the cluster stability variance. The instability of a data element refers, loosely speaking, to the uncertainty about the cluster to which this element should be assigned. Stated another way, a data element has a high instability if the considered clustering algorithm is not able to reliably assign it to a cluster. From the instability of a data element, we have defined the instability of a given hard clustering algorithm. The cluster stability variance can be interpreted as an inverse measure for the robustness of a clustering algorithm, for a given data set.

We have introduced theorems that show how the instability of a given hard clustering algorithm can be reduced, i.e., by removing appropriate unstable data elements.

II. Methods

A. Data sets

Three biological data sets with measurements of the level of expression of genes are used for our experimental study. Briefly, genes are segments of the genome of an organism, encoding the functional components (most often proteins) of that organism. The first step in the decoding of this information is the production of messenger ribonucleic acid (mRNA) molecules from these genes, the quantity of which provides information about the activity of said genes. Measuring the quantity of all mRNAs of all genes is a common approach in modern biology, and the ensuing data is called 'gene expression data set'. In our assessment of clustering performance we chose gene expression data sets that contain a series of measurements (essentially constituting a data vector for genes) covering a certain time range (time series experiment), usually spanning several hours after a particular stimulus, with measuring points some minutes or hours apart. Experiments of this type usually include a control group not undergoing the stimulus, allowing to express the observed expression as a ratio relative to a control. We preprocessed these data to get unique genes with significant expression values.

1) Rat data set: The first data set represents the gene expression response to a stimulus by the stomach hormone gastrin, a data set produced by Selvik et. al. [11], and available at the Gene Expression Omnibus (GEO) database [12] (accession ID GSE32869). This data set concerns a time series experiment on a cell line obtained from rats. Following treatment with gastrin, cells were sampled at 11 intervals during a 14 hour period to record how expression responses evolve over time. The experiment was done twice to obtain more robust data (average of replicates). The time series data was analyzed using an extended dimension reduction framework for significance analysis of gene expression data as described [13]. The extended framework uses partial least squares regression (PLS) in combination with a priori defined time curves based on hypothesized network motifs (unpublished data). This resulted in 2292 genes that were differential expressed in response to gastrin, while also displaying a relatively smooth profile. This data set is further referred to as the Rat data set.

2) Human data set: The second dataset is a time series gene expression analysis based on human breast cancer cells stimulated by the growth hormone epidermal growth factor [14]. These data were obtained from the GEO database (accession ID GSE13009). Although the data covers a 72 hour period, we considered the first 14 time points covering 24 hours. The data was preprocessed to select genes that were significantly affected by the hormone stimulus. This involved normalization by the Robust Multi-array Average (RMA) method [15] and filtering for high variation using the Genefilter package [16]. This resulted in 2194 genes representing the most significant variation in the data set. We refer to this data set as the Human data set.

3) Yeast data set: The third data set was produced for identification of genes showing activity changes during the process of cell division in the yeast Schizosaccharomyces pombe [17]. This data set is freely available [18]. We chose a subset of these data that had the lowest number of missing values, named 'elutriation 3'. We could link 374 of the 407 cell cycle related genes in this set through their systematic IDs but filtered out an additional 118 genes because they had a high incidence of missing values in their data vectors. The result is a data set containing expression profiles for 256 genes covering 20 time points. This data set is referred to as the Yeast data set.

B. Clustering algorithm

Throughout this paper k-means is used as clustering algorithm. The reason is that the application of our method is restricted to hard clusterings, as described in Section I-B1. In a hard clustering it holds that any two different elements either belong to the same cluster or belong to different clusters. K-means is the natural choice to produce such clusterings. Since our method does not refer to the hard clustering algorithm that produces the hard clusterings, it can equally well be applied to hard clusterings generated by other algorithms (e.g., for example, hard clustering algorithms based on multiple dissimilarity matrices [19]). We leave the application to hard clusterings produced by other algorithms than k-means to future research.

C. Mutual information index as cluster validation measure

1) Attribute matrix: The mutual information index is often used to validate clusterings, provided that external labels for qualifying the cluster elements are available. For gene expression data sets, we can rely on publicly available databases containing gene ontology (GO) annotations to be used as the external labels [20]. Gene ontology annotations are standardized terms that biological experts use to functionally describe various qualities of a gene such as their molecular function, biological process and cellular location that can be attributed to them. These qualifications represent attributes that can be collected for many genes through the BiomaRt package [21] of the Bioconductor analysis software [22]. Associations of the genes with these unique attributes are then represented by an attribute matrix T [23] such that if gene i is annotated with attribute j, we define $T(i, j) = 1$, otherwise $T(i, j) = 0$. We built attribute matrices for each data set.
Subsequently, the filtering method described in [24] was

applied. More concretely, the following type of genes and attributes were removed from the data:

- Unannotated genes, since they were not associated with any attribute (GO term).
- Attributes (GO terms) associated with fewer than 10 genes as these were considered to be not informative enough.

This procedure reduced the Rat data set to 1776 genes with a total of 257 different attributes, the Human data set to 1818 genes with 517 attributes and the Yeast data set to 253 genes with 30 attributes.

2) Calculation of mutual information index for clusterings of gene expression data sets: Given the attribute matrix, the mutual information index for a clustering C, containing clusters $C_1, \ldots, C_m$, is calculated in several steps.
First, we calculate the entropy $H(C)$:

$$H(C) = -\sum_{i=1}^{m} p(C_i) \log p(C_i) \tag{6}$$

where $p(C_i)$ denotes the probability of a gene belonging to C_i, which we approximate by the number of genes belonging to cluster C_i divided by the total number of clustered genes.
Secondly, the entropy of all attributes A_j is calculated. This is defined similarly as $H(A_j) = -p(A_j) \log p(A_j)$, where $p(A_j)$ is now estimated as the number of genes g_i for which $T(i,j) = 1$ divided by the total number of clustered genes.
Thirdly, we calculate the joint entropy $H(C, A_j)$ between the clustering C and each of the attributes A_j, defined as $H(C, A_j) = -\sum_{i=1}^{m} p(C_i, A_j) \log p(C_i, A_j)$, where $p(C_i, A_j)$ is the probability that a gene has attribute A_j and belongs to cluster C_i, estimated as the fraction of genes belonging to C_i and at the same time having attribute A_j, i.e., such that $\sum_i \sum_j p(C_i, A_j) = 1$.
Fourthly, the mutual information between C and each of the attributes A_j is defined as

$$I(C, A_j) = H(C) + H(A_j) - H(C, A_j) \tag{7}$$

Finally, the mutual information index is calculated as $\sum_j I(C, A_j)$.
The intuition behind the mutual information index is that it measures the degree to which the clustering of the genes is consistent with the information known about these genes as represented by the attribute matrix.

D. Most robust clustering from a sample of clusterings

In Section I-B1, and especially in our previous work [6], it was argued that the expected clustering can be considered as maximally robust. However, the expected clustering is only a theoretical concept and cannot be determined in practice. The average clustering is an approximation for the expected clustering, and thus can be considered as very robust, provided that the corresponding sample of clusterings is large enough. The problem is that the average clustering $\bar{C}$ is not a hard clustering, because it typically contains elements $\bar{C}(j,k)$ different from 0 and 1. This implies that the mutual information index cannot be calculated for $\bar{C}$, since this requires to calculate $H(\bar{C})$, which in turn requires to determine the number of genes belonging to each cluster. However, this number can only be determined for a hard clustering. A good candidate robust clustering is the clustering from the given sample of clusterings that is closest to the average clustering, in terms of the distance measure (4). That is, we define as most robust clustering, given a sample $\{C_1, \ldots, C_N\}$, the clustering C for which it holds that

$$C = \arg \min_{1 \le i \le N} \{d(C_i, \bar{C})\} \tag{8}$$

Evidently, this clustering is hard and it is legitimate to consider it as the most robust one among all clusterings belonging to the given sample.

III. Discussion

A. Experimental setup

1) Determination of the parameters: K-means was used with each data set to produce 500 clusterings with randomly chosen initial centers. From these samples of clusterings, the average clustering could be calculated, and this in turn allowed to calculate the instability of each gene, the instability of the clustering algorithm and the CSV (see Section I-B). The similarity measure was chosen as the correlation distance, since it has been argued that this measure is well suited to measure the coexpression between genes [25]. The number of clusters produced with k-means for the different Rat, Human and Yeast data sets was set to 9, 6 and 11, respectively. Determining the optimal number of clusters for a biological data set is not an exact science; many cluster analysis experts even argue that there does not exist something as 'the' optimal number of clusters. Rather, determining the optimal number of clusters is the subject of a continuing debate, and different optimization methods typically declare different numbers as being the optimal number of clusters. Therefore, we used a heuristic approach consisting of a visual inspection of the types of genes that are distributed over different clusters, while imposing a range of cluster numbers on k-means. This allowed us to set the above numbers based on the approximate distribution of genes over different clusters *vis a vis* the types forced together within single clusters, essentially checking the biological plausibility of separating genes or lumping them together. It is important to note, however, that this paper is not about defining the optimum numbers of clusters, but rather about the quality of clusterings before and after removal of unstable elements. The above mentioned numbers of clusters were kept fixed for all 500 clusterings of the different data sets.

2) Research topic: comparison of the quality of clusterings of the original data set with the quality of the most robust clustering of the reduced data set: As outlined above, our goal is to compare the quality of clusterings, as measured by the mutual information index, before and after the removal of unstable genes. This is done as follows. First, 500 clusterings are generated using k-means and their mutual information index is calculated as described in Section II-C2. Secondly, unstable genes are detected and removed from the data set until the CSV appears to reach a minimum or has stabilized (see Section I-B5). Thirdly, k-means is now applied on the reduced data set to produce again 500 clusterings and the mutual information index is calculated for the most robust clustering (see Section II-D). The obtained 501 mutual information indices are plotted in a histogram, and the research question considered is how the mutual information index of the most robust clustering (of the reduced data set) compares to the mutual information indices of the clusterings of the original data set. This procedure is repeated for each of the three data sets.

B. *Experimental results*

For each data set we plot the mutual information indices as outlined above, and the CSV and instability after elimination of the most unstable gene (repeating this procedure until the CSV appears to have reached a minimum or has stabilized).

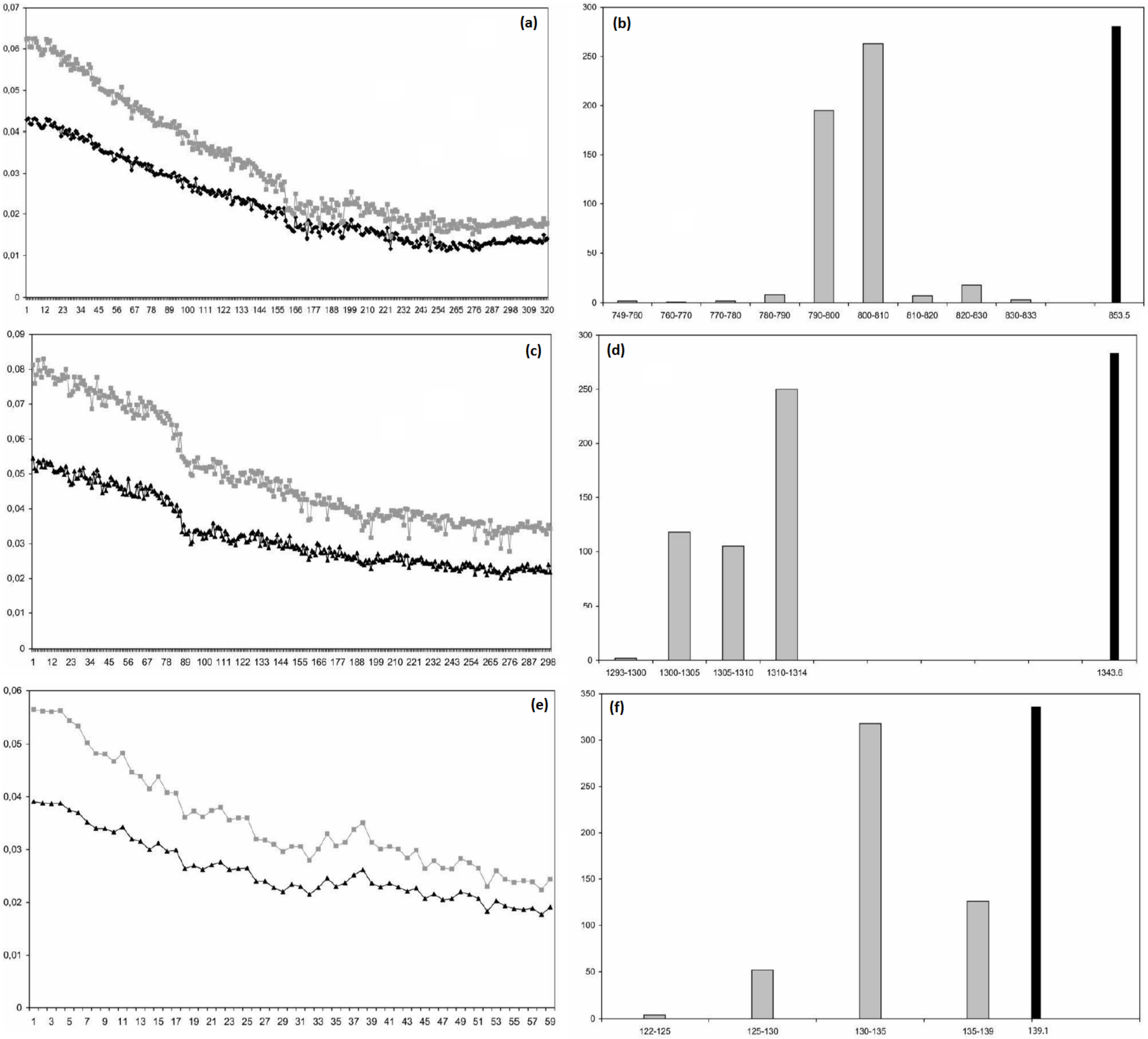


Figure 1: Left column (a,c,e): CSV (black) and instability (grey), the x-axis denotes the number of removed genes; Right Column (b,d,f): Mutual information indices for clusterings of original data set (grey) and of reduced data set (black), Y-axis denotes the number of clusterings, de x-axis denotes the MI; (a,b) Rat data set; (c,d) Human data set; (e,f) Yeast data set.

It will be noticed that the CSV is smaller than the instability, in accordance with Theorem 2. Theorem 3 states that the instability decreases after the removal of an unstable gene, although in the figures below one will see that although the general trend of the instability is decreasing, the instability sometimes *increases* after the current most unstable gene has been deleted. The explanation for this phenomenon is simply that practical restrictions in terms of time and memory force us to work with samples of clusterings rather than with the set of all possible clusterings as implicitly assumed by the theorems above.

1) Rat data set: The CSV, the instability and the mutual information indices for the Rat data set are shown in Fig 1a and Fig 1b. The number of genes that we removed was chosen rather heuristically, since we have to find a compromise between the desire to make the clustering algorithm as robust as possible (i.e., removing all unstable genes) and to limit the number of removed genes, since we want to end up with a clustering that contains a significant number of the genes from the original data set. We decided to remove 174 genes, since the sharpest decrease in the CSV appears up to this number of genes. This amounts to removing about 10% of all genes. Fig 1b shows the mutual information indices. The quality of the most robust clustering of the reduced data set (i.e., after removing the 174 genes) is significantly higher than the quality of any clustering of the original data set.

2) Human data set: The results for the Human data set are shown in Fig 1c and Fig 1d. We chose to remove 270 genes, since the CSV appears to stabilize after that number. This amounts to about 15% of all genes. The difference in quality between the most robust clustering of the reduced data set and the qualities of the clusterings of the original data set is striking.

3) Yeast data set: Fig 1e and Fig 1f display the results for the Yeast data set. Thirty-one unstable genes are removed, since the sharpest decrease in the CSV appears up to this number. In relative terms, about 12% of all genes are taken out of the given data set. The mutual information index of the most robust clustering of the reduced data set is higher than that of any of the clusterings of the original data set.

IV. Conclusion and future work

In previous work, we introduced the concepts 'instability' and 'cluster stability' variance as inverse measures of the robustness of a hard clustering algorithm, with respect to initial conditions. We showed how removing unstable data elements from a data set increases the robustness of the clustering algorithm. A question we did not consider is how the removal of such unstable data elements affects the quality of the produced clusterings. Although the reduced data set can be better clustered by a clustering algorithm in the sense that this clustering algorithm is able to recognize a more definite structure in the data set, irrespective of the initial conditions, it is possible that the produced clusterings are of a lower quality, meaning that the recognized structure is further away from the real structure. In this paper, the quality of clusterings of the original data set is compared to the quality of the most robust clustering of the reduced data set, i.e., after removing unstable genes, performed on three authentic biological gene expression data sets, where the quality is measured in terms of the mutual information index. Although our hope was that the quality of the most robust clustering of the reduced data set would be higher than that of most clusterings of the original data set, to our surprise it turned out that the robust clustering of the pruned data set significantly outperforms *all* clusterings of the original data set. The main conclusion that we therefore draw from this and our previous work is that it is beneficial to detect and remove unstable data elements from a data set, both in terms of robustness of the clustering algorithm with respect to initial conditions and in terms of the quality of the generated clusterings.
As future work, we plan to apply our method to hard clusterings produced by other hard clustering algorithms than k-means.

Acknowledgment

The authors would like to thank Endre Anderssen, Arnar Flatberg and Torunn Bruland from NTNU for their help in processing and filtering the Rat data set. These data were provided by the Genomics Core Facility (GCF), of the Norwegian University of Science and Technology (NTNU). GCF is funded by the Faculty of Medicine at NTNU and the Central Norway Regional Health Authority.

References

[1] M. Schena, D. Shalon, R.W. Davis, and P.O. Brown, ”Quantitative monitoring of gene expression patterns with a complementary DNA microarray,” Science, vol. 270, pp. 467-470, Oct. 1995, doi: 10.1126/science.270.5235.467.

[2] D. Jiang, C. Tang and A. Zhang, ”Cluster analysis for gene expression data: a survey,” Transactions on Knowledge and Data Engineering, vol. 16, pp. 1370-1386, Nov. 2004.

[3] A. Alrabea, A.V. Senthilkumar, H. Al-Shalabi and A. Bader, ”Enhancing K-means algorithm with initial cluster centers derived from data partitioning along the data axis with PCA,” Journal of Advances in Computer Networks, vol. 1, pp. 137-142, Jun. 2013, doi: 10.7763/JACN.2013.V1.28.

[4] Y. Loewenstein1, E. Portugaly, M. Fromer and M. Linial, ”Efficient algorithms for accurate hierarchical clustering of huge datasets: tackling the entire protein space,” Bioinformatics, vol. 24, i41-i49., Jul. 2008, doi: 10.1093/bioinformatics/btn174.

[5] R. Winkler, F. Klawonn and R. Kruse, ”Fuzzy clustering with polynomial fuzzifier function in connnection with M-estimators,” Applied and Computational Mathematics, vol. 10, pp. 146-163, 2011.

[6] W.D. Mulder, M. Kuiper and R. Boel, ”Clustering of gene expression profiles: creating initialization-independent clusterings by eliminating unstable genes,” Journal of Integrative Bioinformatics, vol. 7, Mar. 2010, doi: 10.2390/biecoll-jib-2010-134.

[7] J.M. Buhmann, ”Information theoretic model validation for clustering,” International Symposium on Information Theory, pp. 1398-1402, Jun. 2010, arXiv:1006.0375.

[8] S.A. Fattah, C.-C. Lin and S.-Y. Kung, ”A mutual information based approach for evaluating the quality of clustering,” IEEE International Conference on Accoustics, pp. 601-604, May 2011, ISSN: 1520-6149.

[9] A. Strehl and J. Ghosh, ”Cluster ensembles - a knowledge reuse framework for combining multiple partitions,” Journal of Machine Learning Research, vol. 3, pp. 583-617, Apr. 2003.

[10] N.X. Vinh, J. Epps and J. Bailey, ”Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance,” Journal of Machine Learning Research, vol. 11, pp. 2837-2854, Oct. 2010.

[11] L.-K. Selvik, C. Fjeldbo, A. Flatberg et al., ”The duration of gastrin treatment affects global gene expression and molecular responses involved in ER stress and anti-apoptosis,” BMC Genomics, vol. 14, Jun. 2013, doi:10.1186/1471-2164-14-429.

[12] T. Barrett and R. Edgar, ”Gene expression omnibus: microarray data storage, submission, retrieval, and analysis,” Methods Enzymol., vol. 411, pp. 325-369, Oct. 2006, doi: 10.1016/S0076-6879(06)11019-8.

[13] L. Gidskehaug, E. Anderssen, A. Flatberg, and B.K. Alsberg, ”A framework for significance analysis of gene expression data using dimension reduction methods,” BMC Bioinformatics, vol. 8, Sept. 2007, doi:10.1186/1471-2105-8-346.

[14] Y. Saeki, T. Endo, K. Ide et al., ”Ligand-specific sequential regulation of transcription factors for differentiation of MCF-7 cells,” BMC Genomics, vol. 10, Nov. 2009, doi:10.1186/1471-2164-10-545.

[15] R.A. Irizarry, B. Hobbs, F. Collin et al., ”Exploration, normalization, and summaries of high density oligonucleotide array probe level data,” Biostatistics, vol. 4, pp. 249-264, Apr. 2003.

[16] http://www.bioconductor.org/packages/2.3/bioc/html/genefilter.html

[17] G. Rustici, J. Mata, K. Kivinen et al., ”Periodic gene expression program of the fission yeast cell cycle,” Nature Genetics, vol. 36, pp. 809-817, Aug. 2004.

[18] http://www.bahlerlab.info/projects/cellcycle/

[19] F. de A.T. de Carvalho, Y. Lechevallier and F.M. de Melo, ”Partitioning hard clustering algorithms based on multiple dissimilarity matrices,” Pattern Recognition, vol. 45, pp. 447-464, Jan. 2012.

[20] F.D. Gibbons and F.P. Roth, ”Judging the quality of gene expression-based clustering methods using gene annotation,” Genome Research, vol. 12, pp. 1574-1581, Oct. 2002.

[21] S. Durinck, Y. Moreau, A. Kasprzyk et al., ”BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis,” Bioinformatics, vol. 21, pp. 3439-3440, Aug. 2005.

[22] R.C. Gentleman, V.J. Carey, D.M. Bates et al., ”Bioconductor: open software development for computational biology and bioinformatics,” Genome Biol., vol. 5, Sept. 2004.

[23] M. Ashburner, C.A. Ball, J.A. Blake et al., ”Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” Nature Genetics, vol. 25, pp. 25-29, May 2000.

[24] R. Steuer, P. Humburg and J. Selbig, ”Validation and functional annotation of expression-based clusters based on gene ontology,” BMC Bioinformatics, vol. 7, Aug. 2006, doi: 10.1186/1471-2105-7-380.

[25] L.J. Heyer, S. Kruglyak and S. Yooseph, ”Exploring expression data: identification and analysis of coexpressed genes,” Genome Research, vol. 9, pp. 1106-1115, 1999, doi: 10.1101/gr.9.11.1106.

Smartphone Application for Post-operative Gastric Patients: Surgery Diary

Jin-Ming Wu
Department of Surgery of National Taiwan University Hospital
Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University
Taipei, Taiwan
e-mail: kptkptkpt@yahoo.com.tw

Te-Wei Ho, Xing-Yu Su
Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University
Taipei, Taiwan
e-mail: skbaskba@gmail.com, aipeople0513@gmail.com

Ming-Tsan Lin
Department of Surgery of National Taiwan University Hospital
Taipei, Taiwan
e-mail: linmt@ntu.edu.tw

Feipei Lai
Graduate Institute of Biomedical Electronics and Bioinformatics of National Taiwan University
Department of Computer Science and Information Engineering of National Taiwan University
Department of Electrical Engineering of National Taiwan University
Taipei, Taiwan
e-mail: flai@ntu.edu.tw

***Abstract*—Gastric cancer is one of the most common gastrointestinal diseases around the world. In general, surgical resection is the only intervention. Surgical gastric cancer patients may suffer from malnutrition, which can be associated with gastrointestinal complications, surgical stress, and cancer cachexia. Malnourished patients typically have poor oncological outcomes and a decreased quality of life. To hasten the recovery of post-gastrectomy patients, we developed a smartphone application (app) that implements the functions of nutrition monitoring, medical information management, follow-up of drains, and wound care. This app is written using objective-C and the iOS (previously iPhone OS) 6.1 SDK (Software Development Kit), which provides iOS 5.1 and later compatibility. To integrate users' records, mySQL was used for data management and computing. Moreover, this app includes clinical rule support that informs patients of severe body weight loss or possible internal bleeding. Using this app, patients are able to monitor their general condition, wound, and nutritional status by themselves. Based on the preliminary analysis from users, this app is considered a useful tool that users can use it to reduce the impact on body weight loss significantly. More importantly, this app acquires a high degree of support with 93.3 percent from users.**

Keywords-smartphone; application; gastric cancer; surgery

I. INTRODUCTION

Gastric cancer is the fourth most common cancer in the world [1]. In general, radical surgical resection is the only potentially curative treatment. However, most patients are usually diagnosed at an advanced stage owing to occult clinical presentation, and the five-year survival rate is less than 10 percent. Moreover, patients diagnosed with gastric cancer have a higher proportion of malnutrition (60 to 85 percent) [2], which may be related to a cancer-inducing gastrointestinal obstruction, cancer cachexia, or perioperative stress. Surgical patients may suffer from wound pain, anorexia, malaise, and gastrointestinal digestion disorders [3], which can also result in malnutrition. Reduced gastric functioning is associated with both diminished food intake and malabsorption of vitamins, fats, and proteins [4]. If the malnourished status is not recognized, patients will have inferior clinical outcomes and a decreased quality of life compared to those with suitable nutritional status [5][6].

With the popularity of smartphones, people have changed their method of accessing information. There is an increased interest in smartphone applications (apps) as a tool for medical professionals and patients to deliver medical information. It is reported that the market for mobile health apps for smartphones and tablets will be US $26 billion by 2017 [7]. As of 2013, there are an estimated forty thousand medical apps available on the market. It has been indicated by previous studies that smartphone apps for patient self-monitoring are feasible [8][9]. Hence, for the purpose of the high-quality care, the aim of this study was to create a smartphone app, including perioperative medical education, long-term follow-up of body weight implemented with clinical rule support, and wound care for post-operative gastric cancer patients. In addition, we evaluate the utility of this app and assess the feasibility of it.

The construction of this paper is organized as follows: 1) *System Architecture*, where the development of the application system is presented in detail; 2) *Result*, which presents the demographics of users, and the results obtained for the assessment for the app; and 3) *Conclusion and Future Work*, which resumes this paper by summarizing our contributions and discussing directions for future work.

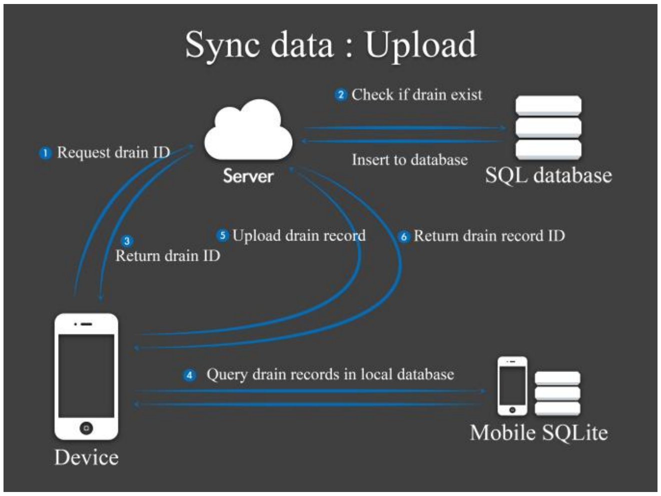


Figure 1. The architecture of application system

II. SYSTEM ARCHITECTURE

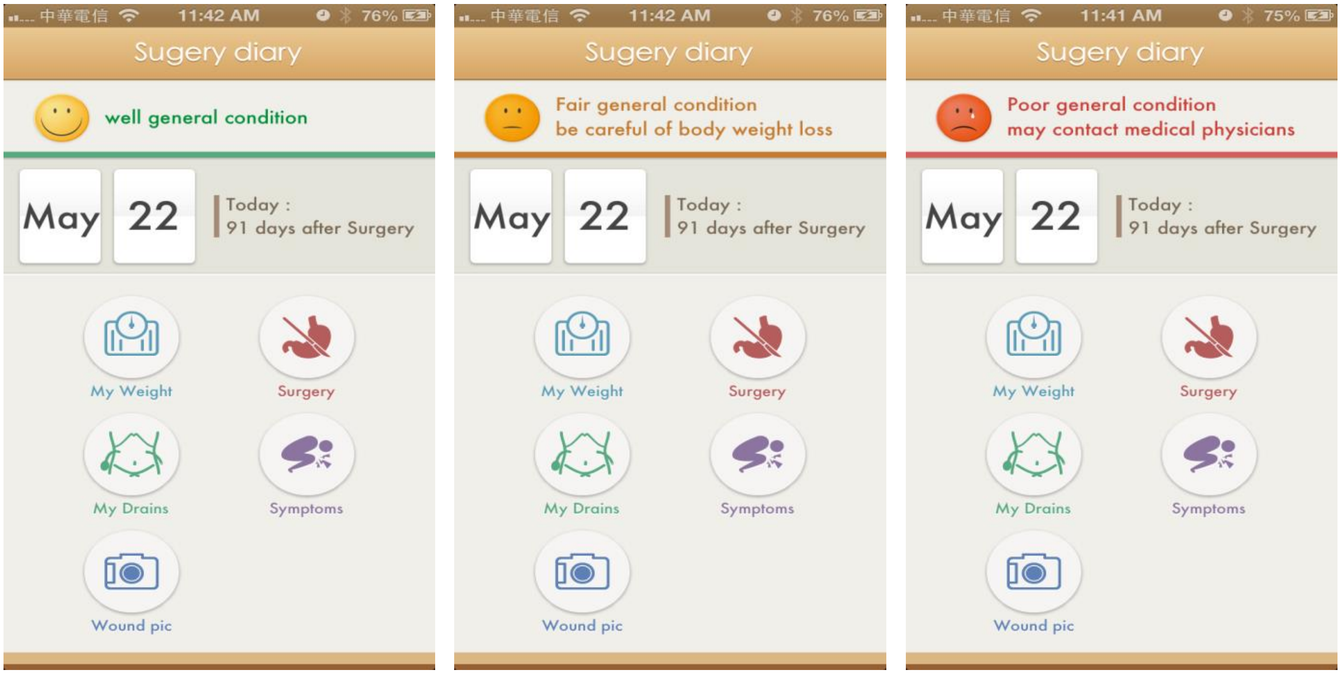
This app was developed using the objective-C language and iOS (previously iPhone OS) 6.1 SDK (Software Development Kit), which provides iOS 5.1 and later compatibility. To integrate users' records, mySQL was used for data management and computing. Besides, the app automatically synchronizes data with a server-side database implemented by the C# language in a .NET framework. To provide web services, a Microsoft SQL (Structured Query Language) server is utilized for data storage in the study. Figure 1 shows the architecture of the application system.

A. *Interfaces Implementation*

We divided the major functions of this app into six interfaces, such as home page, my weight, my drains, wound pictures, surgery, and symptoms.

1) Home page: The interface of the home page (Figure 2), which is grouped from two parts. The first part shows the summary of patient's health status and the days after surgery. According to the last uploaded data from patients, we divide the summary into three categories by the rules of physicians' suggestions. Hence, the categories are defined as "well general condition", "fair general condition", and "poor general condition", respectively. The well general condition means normal in both weight and drain. The fair general condition means abnormal in body weight decreasing by 5 to 10 percent. The poor general condition means either the last body weight decreasing more than 10 percent or the last drain fluid turning red. With these alarms, the patient could contact with case managers, and receive appropriate suggestions from them. Besides, we employ a corresponding facial expression icon to easily represent a condition that the patient has. Another part of the home page represents the shortcut of each function, such as my weight, my drains, symptoms, surgery, and wound pictures.

2) My weight: For the purpose of continual care, we recommend patients to upload the body weight each day. The interface will illustrate the weight with a trend by one week and three months (Figure 3; left), then we can easily undertake observation in the variance of body weight on it. Besides using the clinical rule support, the app

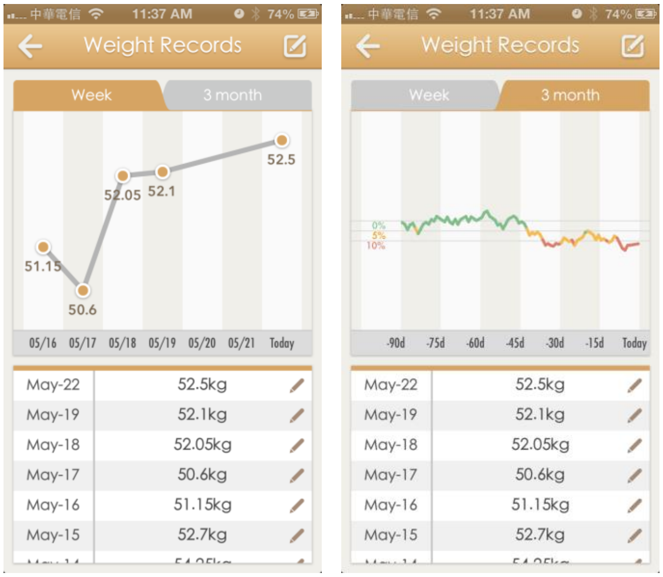

well general condition

Fair general condition
be careful of body weight loss

Poor general condition
may contact medical physicians

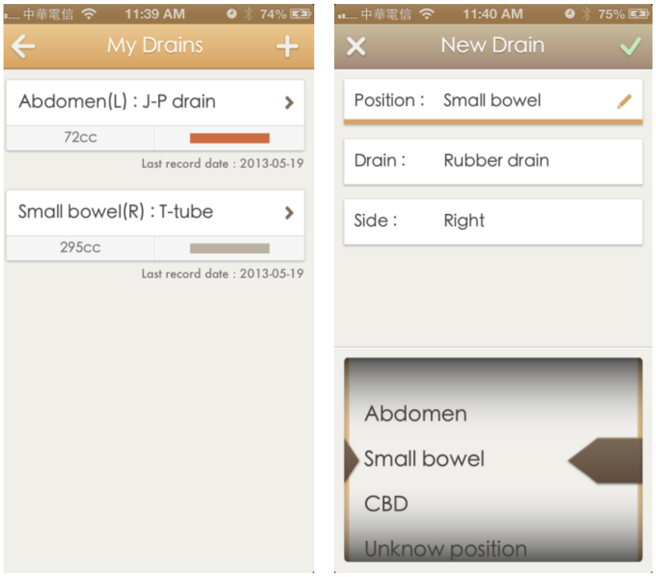
Today :
91 days after Surgery

Figure 2. The home page of smartphone application


11:37 AM
Weight Records
Week
3 month
52.5
52.05
52.1
51.15
50.6
05/16 05/17 05/18 05/19 05/20 05/21 Today
0%
5%
10%
-90d -75d -60d -45d -30d -15d Today

May-22	52.5kg
May-19	52.1kg
May-18	52.05kg
May-17	50.6kg
May-16	51.15kg
May-15	52.7kg

Figure 3. Trend of weight records by 1 week and 3 months



Abdomen(L) : J-P drain
72cc
Last record date : 2013-05-19

Small bowel(R) : T-tube
295cc
Last record date : 2013-05-19

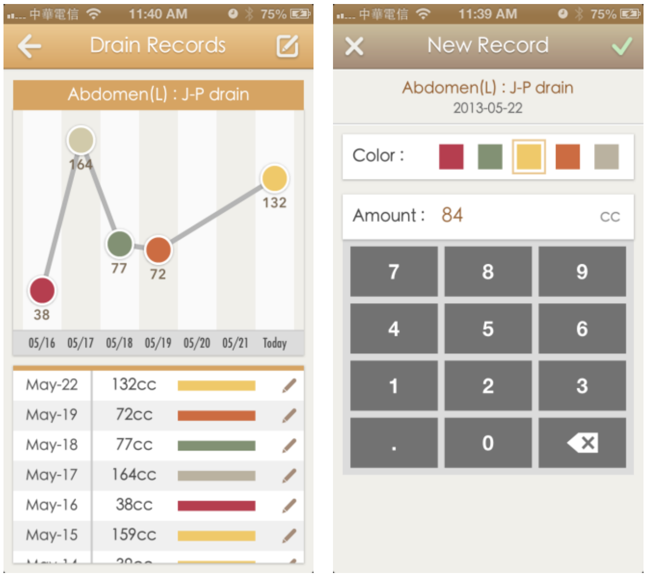
Position : Small bowel
Drain : Rubber drain
Side : Right

Abdomen
Small bowel
CBD
Unknow position

Figure 4. Select the characteristic of drains

automatically calculates the change in body weight over the past three months or the date of operation. Specifically, we define the orange sign to express the patient's body weight decreases by 5 to 10 percent, more than 10 percent for red and normal range (less than 5 percent) for green, respectively (Figure 3; right). Likewise, patients could receive the alert message on the homepage due to the decline in body weight. Through these manners, it may be able to give patients assistance to know their body weight condition much easier.

3) *My drains:* The gastrectomy is a complex operation, which may be associated with complications such as leakage. Previously, the patient recorded the data on paper, which was not convenient or reliable. With the app, the patient can easily record the drain information by himself/herself. First, the patient needs to select the feature of drains at first time. What is more, the patient must choice the kind of drains by their own selection with the position, type and body side. Also, he/she can add a new drain with the same setting format about last time (Figure 4). Next, the patient could input the daily color and volume of any drain by color charts, including red, orange, green, yellow, and gray (Figure 5; right). Generally, these colors are the most commonly used for recording of drain fluids. Meanwhile, we draw the trend and the spot according to the historical records (Figure 5; left). The convenience of this graph is similar to that of weight graph. The patient can readily gain insight into the trend and the color points. It is worth noting that if the drain fluid turns red, it might be the symbol of internal bleeding, then both the patient and medical practitioners could receive alarm message from the system.



May-22	132cc
May-19	72cc
May-18	77cc
May-17	164cc
May-16	38cc
May-15	159cc

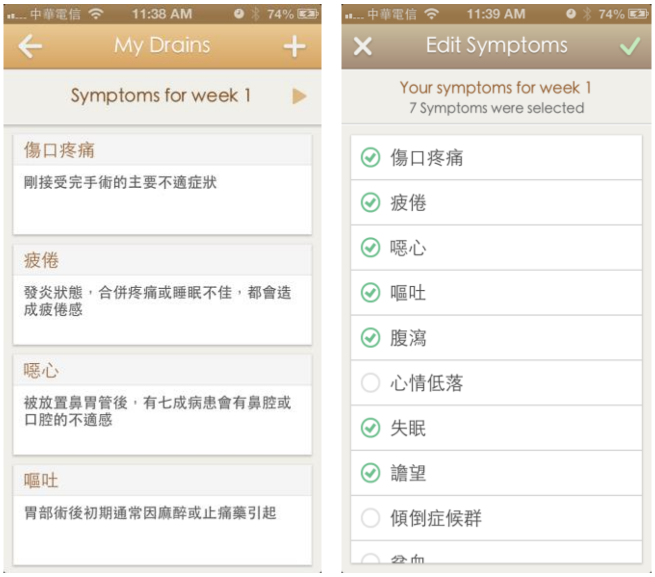
Abdomen(L) : J-P drain
2013-05-22
Color : 🟥 🟩 🟨 🟧 ▨
Amount : 84 cc

Figure 5. The interface of drain records

4) *Wound pictures:* The patient can take pictures of the wound and drains if necessary. The pictures are sent to the web server, where the medical staff can check the condition of the wound. This function is also integrated with the telecare center at the National Taiwan University Hospital (NTUH). If the wound is infected, the medical staff can call the patient and pursue medical treatment.

5) *Surgery:* This section provides the prescriptions of surgery that the patient has. The patient can understand the detail information about the disease and surgery.

6) *Symptoms:* To give appropriate suggestions of symptoms that patients may encounter after the gastrectomy, the app automatically calculates the interval between the date of the operation and the present day. Next, according to



Symptoms for week 1

傷口疼痛

剛接受完手術的主要不適症狀

疲倦

發炎狀態，合併疼痛或睡眠不佳，都會造成疲倦感

噁心

被放置鼻胃管後，有七成病患會有鼻腔或口腔的不適感

嘔吐

胃部術後初期通常因麻醉或止痛藥引起

Your symptoms for week 1
7 Symptoms were selected

- ✓ 傷口疼痛
- ✓ 疲倦
- ✓ 噁心
- ✓ 嘔吐
- ✓ 腹瀉
- ○ 心情低落
- ✓ 失眠
- ✓ 譫望
- ○ 傾倒症候群
- ○ 貧血

Figure 6. The interface of symptoms

the interval days, the function provides patients the information regarding typical weekly discomfort after the surgery (Figure 6; left). In addition, we offer more than ten general symptoms such as wound pain, fatigue, diarrhea, insomnia, and so forth. Patients can simply add or remove their specific symptoms with the detail information through the interface (Figure 6; right). Thus, the system would count the number of symptoms over time.

B. *Web Service and Auto Notification Service*

We designed and developed a platform for patients and medical staff. On the platform, they can review follow all of the historical upload record, such as weight, drain, symptoms, and wound pictures. The patient can access the platform to input their post-operative data if a smartphone is not available (not shown here). In addition, the web service will automatically send a cell phone message to the medical staff when the patient uploads wound pictures. With this service, medical professionals can immediately check the wound condition, and notify the patient if the wound infection should be investigated.

C. *Assessment of the App Performance*

We conducted a simple questionnaire to evaluate the usefulness of this app from users at the National Taiwan University Hospital. Besides, In order to compare with the app users group, we retrospectively collected subjects undergoing gastrectomy as a control group. More importantly, for the assessment of prognosis in both groups, we also collected the medical record in terms of outpatient clinic (OPC) visits, re-admission visits, and emergency room visits. In order to make a comparison between the app group and the control group, all values were expressed as mean with their associated standard deviations or frequency.

TABLE I. CHARACTERISTICS OF PATIENTS BETWEEN GROUPS

Variable	**App group (n=15)**	**Control group (n=15)**	***P* value**
Gender (male: female)	7:8	8:7	0.910
Age (years)	61.6±12.1	60.7±18.1	0.876
Body mass index	23.4±3.6	22.7±3.9	0.611
Comorbidity	8	4	0.135
Indication of gastric surgery Gastric cancer Gastric tumor	 12 3	 12 3	0.999
Pathological staging Stage Ⅰ Stage Ⅱ Stage Ⅲ Stage Ⅳ	 5 6 4 0	 6 5 4 0	0.912
Pre-operative chemotherapy	0	0	0.999
Post-operative chemotherapy	2	3	0.921

The Mann-Whitney U test was used for continuous variables, and the Fisher's exact test was used for categorical variables. Data analyses were performed using SPSS software version 15.0. A 2-sided *P* value of less than 0.05 was considered statistically significant.

III. RESULTs

This app had been available since late June, 2013 on Apple App Store. There were 15 consecutive patients at NTUH accessing this app (app group). Besides, we retrospectively collected 15 cases undergoing gastrectomy as a control group. The demographics of both groups are shown in Table 1. Mean age of the app group was 61.6±12.1 years, and that of the control group was 60.7±18.1 years. The mean body mass index of app group and control group were 23.4±3.6 and 22.7±3.9, respectively. Even though this study design did not use randomization, the two groups were similar in all variables with non-significant difference. For the clinical results (Table 2), the app group had the less proportion of body weight loss percentages compared to control group during six-month follow up (4.6±0.5 vs. 11.4±1.2, P＜0.01). However, the patients of app group had more out-patient clinic (OPC) visits than the control group (10.8±1.4 vs. 8.3±1.7, P＜0.01). With this application, the patients developing body weight loss more than 5 percent three months after operation could receive the warning from the app, which also inform the medical staff to do nutritional assessment and intervention for the malnourished cases, such as consultant of dietician. After all, nearly 93.3 percent of users are willing to recommend this app to others. As a result, the patients of app group visited the OPC more times to undergo the nutritional evaluation to achieve early diagnosis and early intervention.

TABLE II. THE CLINICAL OUTCOMES BETWEEN TWO GROUPS

Variable	**App group (n=15)**	**Control group (n=15)**	***P* value**
Percentages of body weight loss (compared to pre-operative status)	4.6±0.5	11.4±1.2	<0.01
Number of outpatient clinic (OPC) visit	10.8±1.4	8.3±1.7	<0.01
Re-admission	2 (13.3%)	3 (20.0%)	0.598
Emergency room visit	0 (0.0%)	1 (6.7%)	0.999
Would you recommend this App to others? Yes No	 14 (93.3%) 1 (6.7%)	-	-

IV. CONCLUSIONS AND FUTURE WORK

Based on the preliminary analysis from users, this app is considered a useful tool that users can use it to reduce the impact on body weight loss significantly. More importantly, this app acquires a high degree of support with 93.3 percent from users. Smartphones have become increasingly popular with both medical professionals and patients in recent years. For this reason, the number of associated applications is dramatically increasing. The quality of the rapidly growing amount of information in these medical apps; however, could be inconsistent. It is well-documented that only 55.8 percent of the apps are associated with scientifically validated data, with the best quality information being developed by healthcare professionals or organizations [10]. To develop this app, we worked with the medical surgical staff and dieticians at the NTUH. The information, which consists of not only the educational data but also the clinical rules, is considered accurate and reliable.

In this fast-changing arena, human communication and healthcare information-gathering methods have changed. With diverse online social networking services such as Facebook and YouTube, people are beginning to access these services to manage their health condition [11][12]. In the near future, combined services (the app and the platform) may become the leading trend in health care.

However, there are several limitations or special considerations in this study. First, this app is not regulated by the medical authorities in Taiwan. The US Food and Drug Administration (FDA) has set some regulations to oversee medical apps that contain medical device accessory functionality or transform mobile communications into regulated medical devices [13]. In the future, wishing to augment the care provider in the healthcare process decisions, further studies are needed to integrate the medical record and communication services to the platform. Additionally, additional time and cases are required to prove if this app has clinical benefits such as improved quality of life or improved nutritional status.

This app is implemented by information technology specialists, and it is also co-designed by medical staff with clinical rule support. It does not, even so, take the place of clinical intervention and judgment. In addition, patients input and store their information in the app, which is then transferred to a web server. The protection of patient data, which may be vulnerable to intentional or unintentional attack, is very important. In this study, we designed and implemented three secure layers to protect the privacy of the patients and their data on the web server as described previously [14].

This app is a feasible software solution for gastric cancer patients to record their post-operative information and as an alternative tool for self-care. However, the long-term clinical value must be validated in a future study, and number of patients are considered to be necessary to properly evaluate its effectiveness.

ACKNOWLEDGMENTS

The two first authors Jin-Ming Wu and Te-Wei Ho contributed equally to this work and should be considered co-first authors. The authors would like to thank National Taiwan University Hospital for its financial support. The Grant No. is UN102-018.

REFERENCES

[1] T. J. Price et al., “Management of advanced gastric cancer,” Expert. Rev. Gastroenterol. Hepatol, vol. 6, pp. 199–209, Apr 2012.

[2] C. Mariette, M. L. De Botton, and G. Piessen, “Surgery in esophageal and gastric cancer patients: what is the role for nutrition support in your daily practice,” Ann. Surg. Oncol, vol. 19, pp. 2128-2134, Jul 2012.

[3] F. A. Moore, “Effects of immune-enhancing diets on infectious morbidity and multiple organ failure,” JPEN J. Parenter. Enteral Nutr, vol. 25, pp. S36-42, Mar 2001.

[4] P. Senesse et al., “Nutritional support during oncologic treatment of patients with gastrointestinal cancer: who could benefit,” Cancer Treat. Rev, vol. 34, pp. 568-575, Oct 2008.

[5] V. Catalano et al., “Gastric cancer,” Crit. Rev. Oncol. Hematol, vol. 54, pp. 209-241, Jun 2005.

[6] E. Van Cutsem and J. Arends, “The causes and consequences of cancer-associated malnutrition,” Eur. J. Oncol. Nurs, vol. 9, pp. S51-63, 2005.

[7] T. K. Burki, “Cancer apps,” Lancet Oncol, vol. 14, pp. 580-1, Jun 2013.

[8] A. Rao, P. Hou, T. Golnik, J. Flaherty, and S. Vu, “Evolution of data management tools for managing self-monitoring of blood glucose results: a survey of iPhone applications,” J. Diabetes Sci. Technol, vol. 4, pp. 949-957, Jul 2010.

[9] B. A. Rosser and C. Eccleston, “Smartphone applications for pain management,” J. Telemed. Telecare, vol. 17, pp. 308-312, 2011.

[10] A. Pandey, S. Hasan, D. Dubey, and S. Sarangi, “Smartphone apps as a source of cancer information: changing trends in health information-seeking behavior,” J. Cancer Educ, Vol. 28, pp. 138-42, Mar 2013.

[11] T. Sahama, J. Liang, and R. Iannella, “Impact of the social networking applications for health information management for patients and physicians,” Stud. Health Technol. Inform, vol. 180, pp. 803-807, 2012.

[12] A. N. Vyas, M. Landry, M. Schnider, A. M. Rojas, and S. F. Wood, “Public health interventions: reaching Latino adolescents via short message service and social media,” J. Med. Internet Res, vol. 14, pp. e99, Jul 2012.

[13] T. J. Orchard et al., “Haptoglobin Genotype and the Rate of Renal Function Decline in the Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions and Complications Study,” Diabetes, vol. 62, pp. 3218-3223, Jun 2013.

[14] H. J. Yu et al., “A sharable cloud-based pancreaticoduodenectomy collaborative database for physicians: Emphasis on security and clinical rule supporting,” Comput. Methods Programs Biomed, vol. 111, pp. 488-497, Aug 2013

Identification of Short Motifs for Comparing Biological Sequences and Incomplete Genomes

Ramez Mina and Hesham H. Ali
College of Information Science and Technology
University of Nebraska at Omaha
Omaha, NE 68116, USA
hali@unomaha.edu

***Abstract* — Sequence comparison remains one of the main computational tools in bioinformatics research. It is an essential starting point for addressing many problems in bioinformatics; including problems associated with recognition and classification of organisms. Although sequence alignment provides a well-studied approach for comparing sequences, it has been well documented and reported that sequence alignment fails to solve several instances of the sequence comparison problem, particularly for those sequences that contains errors or those that represent incomplete genomes. In this work, we propose an approach to identify the relatedness among species based on whether their sequences contain similar short sequences or signals. We cluster species based on biological signals such as restriction enzymes or short sequences that occur in the coding regions, as well as random signals for baseline comparison. We focus on identifying k-mers (motifs) that would produce the best results using this approach. The obtained results showed that specific k-mers with biological significance such as restriction enzymes produce excellent results. They also make it possible to obtain good comparisons while using shorter or incomplete sequences, which is a critical property for comparing genomes obtained from next generation sequencers.**

Keywords–sequence comparison; alignment; biological motifs; alrignment-free; k-mers; restriction enzymes; coding sequences; phylogenetic trees

I. INTRODUCTION

The second generation of sequencing provided the bioinformatics domain with more genomes for comparison and analysis, which in turn motivated more researchers to compare these genomes and identify their similar structures and functionalities. The need for more research in the comparative genomic came from the fact that sequence alignment methods have limitations, such in quality and speed. The focus for this research is to find better results for the comparison process.

Although the default sequence comparison methods in the literature are based on alignment, other methods are alignment-free as discussed by S. Vinga et al. [1]. These alignment-free methods introduced alternative solutions to overcome the limitations of alignment-based methods, which led to the question; “how much have these methods achieved to overcome the addressed limitations?” These limitations could be briefed in two major issues, the speed issue and the quality issue. The speed issue was addressed before in the literature, and several accomplishments were reached based on alignment, such as the work of BLAST by S. F. Altschul et al. [2] for pair-wise comparison. Other work focused on multiple sequence alignment with heuristic speeds like the work done in MUSCLE by R. Edgar [3] and DIALIGN by A. Subramanian [4]. In addition to alignment-based methods, other techniques are alignment-free [1] and had a focus on addressing the speed issue as well the quality issue.

Alignment-free methods are not new subject and they are in the literature for a while as discussed by K. Song et al. [5]. Alignment-free methods are categorized mainly into two categories. The first category is based on compression techniques, which improved the speed problem in comparing the biological sequences; the improvement came from the fact that many of the compression algorithms could be implemented in a linear time complexity. Compression-based techniques also showed very good quality with the results, especially those techniques that are dictionary-based. The two major techniques for compression are Lempel-Ziv complexity and Kolomogrov complexity [1].

The second category of alignment-free methods is based mainly on considering all possible k-mers [1, 6] to identify the relatedness between species, and it is specific for each k value. The core of the k-mer method is accomplished by generating vectors that represent the probability of each k-mer within each sequence. The distance is then measured between these vectors. Several proposed techniques were applied to the second approach, either using different formulas for measuring the distance between the vectors, or integrating several vectors of different k values within the same distance measure’s formula.

Several approaches were introduced to construct the measuring vectors, and several formulas were provided and/or designed to calculate the distance between these vectors. Bonham-Carter et al. [7] surveyed the methods that were conducted in this domain/area in [7], and we are summarizing some of these methods.

Liu et al. [8] explained the development of base-base correlation, which is based on generating frequency vectors for all the possible combination of DNA nucleotides of length two (AA, AC…., GT, GG), and each vector is normalized, then a mathematical distance measure is applied to find how closely are the pair sequences. Another approach was discussed by V. Arnau et al. [9] which is called Feature Frequency Approach, and it is also based on generating vectors of specific k-mer, these vectors could also be normalized, then a mathematical measure is applied; which would result in a numerical value for the distance measure. Also application of block-FFP method was necessary, a method similar to the one described by T. J. Wu et al [10]. In another work Sims et al. [11] applied different distance measures that are based on Jensen-Shannon Divergence and Kullback-Leibler Divergence which were discussed in J. Lin [12]. G. Lu et al [6] discussed the same concept of generating the vectors of the k-mers with more in-depth. In their models; they applied several values for k which led to several groups of vectors, each with a different k value, these vectors are called compositions vectors, then applied basic mathematical distance measure, and then tuned up better distance measure that would produce better results.

Application of other distance measures to the composition vectors were borrowed from Z. G. Yu et al. [13] as in the work of R. H. Chan et al. [14] and G Lu et al. [6]. A different approach to generate the vectors was based on suffix trees, a data structure that searches for words of length k, and generates the vectors based on its reading; as of the work of Soares et al. [15].

In general Bonham-Carter et al. [7] discussed in depth more statistical (frequency) measures of different k-mers values, with different distance measures, and these methods would address the frequency and also the occurrence of the k-mers, but they never addressed the order of these k-mers.

Our proposed algorithm is primarily based on exploring information embedded in the k-mers of given sequences, it also considers the order of these k-mers as well as the ability of assigning weights for specific signals (k-mers).

The paper is organized as follows: section II is the motivation for this work; section III is the experimental design and the needed algorithms and the utilized methods for this work; section IV is the provided experiments; section V is the results and analysis; and finally section VI is the conclusion followed by references.

II. MOTIVATION

Sequence comparison has been addressed in the literature for several decades, especially with the birth of the very first alignment-based method, Needleman and Wunsch method for sequence alignment was introduced in 1970. This method dominated the domain for a long time, though its limitations showed up with other advances in the bioinformatics sub-domains, especially with the new sequencing machines and the generation of longer genomes, as well as genomes that have sequencing errors or evolutionary history.

Other problems with sequence alignment were addressed by either biologists or computer scientists. Problems like poor quality of results with longer sequences; misinterpretation of results that include biological assumptions (such as the gap filling part of the alignment algorithm; as there is no proof exists that these filled gaps are results of possible evolutionary mutations). Other errors resulted from the genomic translocations; reverse subsequence; mutations; or any other errors that would result from non-biological assumptions. Other errors that are difficult to address with the alignment algorithm; are errors resulting from the sequencing machines; these errors come from mutations and/or assembly errors. Another limitation with sequence alignment is the speed issue, but this work addresses and focuses mainly on the quality issue.

To address the quality issue, integration of biological features and computational theories, and understanding the nature of the DNA sequences are the major motivations for the work, with a hypothesis that considering these major factors would enhance the quality of the comparison results.

DNA sequences are not random in their structures, and it is believed that each fragment/subsequence of the DNA sequence carries a message or a signal. The hypothesis used in this research is that closely related or similar genomes would carry similar signals/fragments.

For example, sequences that carry the same restriction enzymes’ cut positions [16] might be related and would have similar functions. It would be the same with sequences that carry transcription factor binding sites; other signals would be motifs of specific nature, unique shortest substrings [17] within the sequences, or just motifs with biological relevance that are not known to the literature.

Another feature that DNA sequences has; is that they carry tandem repeats in their structures. These tandem repeats could also be significant signals, and all of these features need to be addressed when comparing the sequences.

A motivation of the comparison problems is based on the fact that similar genomes have similar structures and functions; although subsequences with similar functions do not necessarily have similar exact structures, they carry similar signals within these structures. By identifying these signals, we would be able to classify these genomes and address better measurement for their relatedness.

Notice that these signals might be hidden and/or overlapping with other signals. They might also be of different lengths. To identify these signals or at least take advantage of using them, we need to consider all of the available features. For the previous reasons, we designed an approach that would consider all or a group of prospective signals of specific length k, which could help in addressing the unknown hidden signals. Our approach is variable and would consider different lengths of k, also would consider the overlapping signals.

The challenge of identifying such hidden and unknown signals is not easy. Trying to identify these signals and their functions, taking advantage of their existence and their relevant order within the sequences, and using them for clustering purposes are collectively the focus of this work. The hypothesis of this work is that we can have an approach that takes advantage of these hidden signals within the sequences. Identifying the relatedness among species would be done by considering all the possible chances for the existence of these signals within the sequences and using them to identify the biological distance between the sequences.

Investigating whether or not addressing such signals would improve the clustering process. and reveal a better measurement for the relatedness among species is the focus and challenge of this work. As we consider different signals of different lengths to compare the sequences, we also consider random groups of these signals in order to measure the quality of the results in each case, and to measure whether randomly selected signals would have better results than those that contained all k-mers or signals with biological nature. In addition, this work also considers the use of signals that have biological relevance like restriction enzymes, as well as signals that occur within specific regions that have biological functionality in the DNA sequence, such as those in CDs regions. Finally and as a conclusion of the strength of this approach, applications to datasets with errors were conducted.

III. EXPERIMENTAL DESIGN

The design of the experiment should meet the needed requirements to test the hypothesis. Recall that comparing DNA sequences results in numerical values that represent biological distances between species. These values are subjective with each dataset and would be meaningless if they are not used to address the relationship for the entire group of species.

Verification of the correctness of these distances is not an easy task and simply looking at these numerical values will not reveal the correctness of the results. Hence, we propose another way to measure the correctness.

Clustering the species based on the resulting distances would provide a way to evaluate the correctness of these results. The clustering would be done using bi-clustering algorithms for phylogeny. Using the resulting trees of the phylogeny would be a good way to evaluate the quality of the results. Evaluating the correctness of these trees can then be done by comparing them to known gold standard trees; those are trees that have been verified biologically, hence would be a good proof of the quality of the approach and sequence alignment is used as a baseline for comparison.

The steps necessary to accomplish the proposed experiment in this work are listed as follows:

1. Generate the list of the k-mers. For example, k = 3 for all the possible 3-mers, would result in 64 words (4^3). Alternatively, the list could be a random selection of about 20 percent of all possible words, which would be 13 random 4-mers. It could also be a list of biological signals of different lengths.
2. Convert the DNA sequences according to the compiled list of k-mers (refer to Figure 1).
3. Generate the scoring matrix based on pair-wise comparison, using longest common subsequence (LCS) and Lempel-Ziv complexity of distance measure 2 (LZC).
4. Build the phylogenetic trees using UPGMA and Neighbor-Joining (NJ) phylogenetic algorithms.
5. Repeat Step 4 using the scoring matrix generated by multiple sequence alignment (MSA) [3].
6. Measure the distance between the generated trees and the gold standard tree; the method used to measure this distance is the path-length-difference.

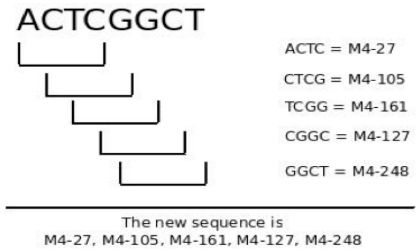


Figure 1. The list on the right side includes the preferred signals to be used for the approach. The sequence on the left is parsed to subsequences each of the same length as the signal lengths on the right, If there is a match, it will be reported with the proper code and in the correct order.

A. *Conversion of DNA sequence to sequence of signals*

Consider all possible signals of specific length k (all possible k-mers); a production of all the possible combination of the length k is generated. This would result in a word's list of size 4^k, where 4 is the number of the used nucleotides in a DNA sequence (A, C, T, and G).

The content of the generated list is used as the main seeds for the signals needed to be identified within the sequences. We substitute any existence of a signal in the DNA sequence with a unique code, which would save conservation of the order for the signals within the sequences. This design would also save computational time when the list is small and the sequences are longer.

Figure 1 shows how to identify the existence of these signals in the sequences, as well as how to convert the DNA sequence to a sequence of signals/words in the proper order of these signals. The used motifs/signals list in Figure 1 is on the right side of the figure. In this list, each motif/signal has a name (code). The left side of the figure has the original sequence, parsed as words of length k (k = 4 in this case).

The used motifs/signals list in Figure 1 is on the right of the figure. Each motif/signal has a name (code). On the left side is the original DNA sequence. We identify the signals from this list that exist in the sequence. If they occur, their codes would be assigned with the proper order to the new sequence of signals. Thus we convert a DNA sequence to a sequence of signals. Also notice that this approach considers all the overlapped signals.

We also need to mention that occasionally some of these signals do not exist in the sequence, or they occur more frequent. Either way would impact the results of the relatedness between the sequences; this would be a major difference between the converted sequences and would address similarity or dissimilarity among species.

B. *The experimental design steps and discussion on the remaining steps*

The conversion step is the heart of this work. In this step, we address the signals in preference, but the work remains incomplete as long as there is no way to compare the converted sequences.

The nature of the converted sequences is that they carry two main features. The first feature is a new alphabet of preferred signals; this motivates us to use similar comparison algorithms/approaches as in regular DNA sequences. Simple and efficient algorithms such as the longest common subsequence (LCS) [18] would address the distances between the converted sequences.

The second feature of converted sequences is the conservation of the signals' order within the sequences. This was missed by other research that was also based on k-mers [6, 7]. The order of the signals would have a great impact on the results, with some possibility of having few or more mobile subsequences. We would still be able to use an algorithm that would address the order feature, like Lempel-Ziv Complexity (LZC) [19]. LZC is based on compression complexity and has a great success in identifying the relatedness of different strings. Please notice that LCS addresses the order factor as well.

The comparison method would result in numerical values that represent the distances between species. As previously discussed, it is necessary to cluster the species to measure the correctness of the resulting distances using hierarchical clustering algorithms such as UPGMA and NJ. Therefore the results of these algorithms are in the form of trees. Although most researchers use visual inspection to evaluate phylogenetic trees, we don't recommend it for the following reasons:

- Visual inspection uses personal judgment, and personal judgment is not usually accurate. It can mislead the evaluation process, especially if it is not compared against some reference.
- Visual inspection cannot identify the correctness of trees with large numbers of species. In fact with some trees that have 1,000 species or more, it would be impossible to find out the relationships between each species.
- Visual inspection does not provide numerical value for the comparison. Therefore, no clear decision could be achieved based on its results. Using a computational method to measure the distance of the resulting tree to a reference tree would yield a decision for the entire experiment.

For these reasons, a computational approach to measuring the distance between resulting trees to a gold standard tree was used. This approach is called path-length-difference, and it was modified to give normalized values. Finally, it is important to compare the trees from our approach to the resulting values from MSA and evaluate whether our approach would have better results.

C. *Different algorithms of the experiments*

This subsection describes some of the methods used to verify the hypothesis of this work, specifically methods that are new to the reader or those that have been modified to fit the work.

- *Normalizing Longest Common Subsequences*

LCS is based on dynamic programming and has a well-established reputation and implementations. However, the generated scores are not normalized, and these scores cannot be used to build a phylogenetic tree. To understand this problem, consider these sequences:

S1 : GTTAATGCCACCAAAAAAAAA (length 21)
S2 : GTTAATGCCACCGA (length 14)
S3 : TCCCTAGCT (length 9)

The LCS for all the pair-wise comparisons is as follows:
S1 : GTTAATGCCACCAAAAAAAAA
S2 : GTTAATGCCACCGA
LCS is GTTAATGCCACCA and the score is 13

S1 : GTTAATGCCACCAAAAAAAAA
S3 : TCCCTAGCT
LCS is TTAGT and the score is 5
S2 : GTTAATGCCACCGA
S3 : TCCCTAGCT
LCS is TTAGC and the score is 5

TABLE I. THE SCORES OF USING LCS ON THE EXAMPLE SEQUENCES

	S3	S2	S1
S3	9	5	5
S2	5	14	13
S1	5	13	21

The resulting scores of using LCS for these sequences are shown in Table 1. Note that the scores in the table are not normalized. To address this issue, we divide the resulting score by the length of the shortest sequences of the measured pair. In addition, to simplify the clustering step we reverse the meaning (average) of the score, by subtracting the normalized score from one, so smaller values imply closer species or higher degree of similarity. The resulting values are shown in Table 2 below.

TABLE II. THE RESULTS AFTER USING THE NORMALIZING FUNCTION AS SUGGESTED

	S3	S2	S1
S3	0	1-5/9	1-5/9
S2	1-5/9	0	1-13/14
S1	1-5/9	1-13/14	0

- *Lempel-Ziv complexity:*

Lempel-Ziv complexity of distance measure 2 was used. Please refer to [19, 20] for more details.

- *Path-Length-Difference:*

The comparison between trees was done by estimating the path-length-difference metric [20]. The main concept here is to give penalties for changing relative positions of species in the generated tree, relative to the reference tree (gold standard tree). Each change would make a species closer to a group of species and further from the rest of species; that should cause a penalty value.

The method begins by generating two matrices; one for the generated tree (resulting tree of our approach), and the other matrix is for the gold standard tree. The dimensions of the matrices are *m* x *m*, where *m* represents the number of species for the dataset (or the tree). Each cell has a value

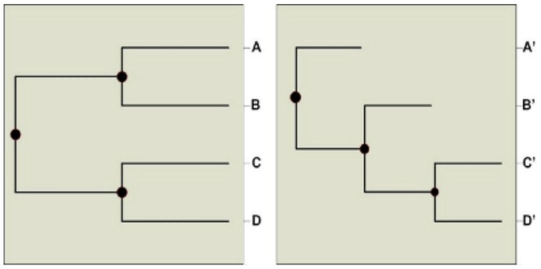


Figure 2. Two trees for the comparison method. The one on the left is the gold standard tree, while the one on the right is the algorithmic tree.

IV. Experiments

The *experiments* are designed and carried out to answer the following questions:

1. Would some motifs/words/signals provide good results for sequence comparison? Would these signals have better comparison results over traditional sequence comparison methods such as those that are alignment based?
2. If the answer for question 1 is positive, is it possible to change the selection of the k-mers for the experiment? Would that enhance the results? In other words, are there certain words/k-mer(s) that would improve the clustering pattern?
3. If the answer for question 2 is positive, would we be able to use signals with biological relevance to improve the results? Like restriction enzymes..etc.
4. If the answer for question 3 is positive, would it be possible to find hidden signals within the sequence with biological relevance and then use them to have valid results?
5. Finally, if the first four questions have been answered positively, is it possible to use the approach on datasets that have errors and still get valid results?

V. Results and Analysis

To answer the above questions, an experiment for each question was conducted. All the experiments follow the same process, as discussed in the Methodology section, with each experiment using a different list of used k-mers.

Datasets:

- The first dataset used was the mycobacterium dataset. We used it for the first three experiments.
- The second dataset was a mitochondrial genomic dataset. We used it for experiments 3, 4, and 5.

A. The First Experiment: Viability of the method

The goal of this experiment is to evaluate whether using generic signals within the sequences would provide good results, as well as whether those results would be better than the results of traditional alignment-based methods (MSA). This experiment deals with all the possible k-mers, as some of them might be hidden signals with strength and within the sequences. The used list of k-mers includes all possible k-mers.

Figure 3 shows the results of using all possible k-mers, and it shows very good results. All distances of any value for k were less than 1.25 percent, while with MSA the results were above 1.8 percent, be aware that smaller values express better distance measures, and are interpreted as closer distance to the gold standard tree.

Figure 3 shows significance in the results using our approach compared to those of MSA. That proves our hypothesis that emphasizing such signals would improve the results and would answer the first question. Please notice that MSA refers to the result of applying Multiple Sequence Alignment, and 7LCS means applying longest common subsequence on k-mer of length 7.

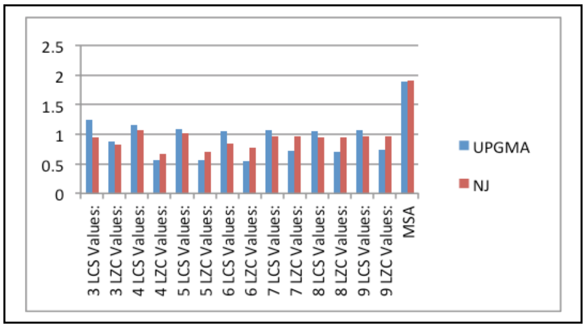


Figure 3. This shows the results of using our algorithm with different parameters. Here, k ranges from 3 to 9. The used methods of comparison are LCS and LZC, and the clustering methods are UPGMA and NJ. The chart shows that in all cases our approach outperformed MSA (multiple sequence comparison) with significant results.

B. The Second Experiment: Using random k-mers

For this experiment, we used lists of random signals selected from all possible k-mers with percentages of 10–90 percent. These selections were applied to k values of 3 to 9, using comparison methods LCS and LZC and clustering NJ and UPGMA. We tested the impact of charging the parameter k on the quality of the results by measuring how similar the obtained phylogenetic tree produced by our approach and MSA from the gold standard tree.

Expectations for the randomly generated lists were that the list carries strong signals, carries weak signals, or carries both. The purpose of carrying on this experiment was to test whether results would be better, worse, or close to those obtained from the first experiment.

Overall, the majority of the experiments we conducted using random k-mers showed better results obtained from our proposed approach as compared to MSA. Some of them were even better than the results obtained in the first experiment. Few cases produced results slightly below the level produced by the first experiment, and some were very close to the results of the first experiment. We tested the approach using the two comparison methods LCS and LZC while using two clustering algorithms NJ and UPGMA.

For example, when using using LCS as a comparison method and NJ as a clustering method. The results of this experiment showed that with just a random selection of k-mers, the approach would still provide better performance than using alignment-based methods for all the values of k that we used. Even with a small list for k-mers (up to 10 percent of all the possible k-mers of specific k), the results would still outperform MSA.

In addition, some runs showed better results than those obtained in the case when for all possible k-mers are used, as in the case with 60 percent random selection of k-mers of length 6. The resulting tree has a distance to the gold standard tree of 0.489 percent, which outperformed any result of all possible motifs, which you can compare by referring to Figure 3. When the random selection of 10 percent for k-mers of length 5, the distance to the standard tree has a value of 1.877005 percent, which is slight worse than any value in Figure 3. That shows that while some signals would do better when they are used alone, others would do slightly worse.

Similar results were produced when the algorithm used LZC as a comparison method and NJ or UPGMA were as the clustering method.

C. *The Third Experiment: Using restriction enzymes' cut positions as the words list*

The second experiment showed that results would be impacted with the selection of the words (k-mers) list, and that some signals would have a higher impact over others. This motivated us to proceed with the third experiment that deals with words that have biological relevance and to see how these words would impact the results. The used signals were obtained from a database of restriction enzymes' cut positions.

Restriction enzymes are special nucleotide signals that cut the DNA double- or single-stranded sequence at specific recognition positions. We believe that DNA sequences that share similar restriction enzymes' cut positions would also have similarities in their functions and structures.

We used restriction enzymes' cut positions that have lengths of 4 to 8 nucleotides. As the number of words for each length was small, we had to use all of them as the words list. Therefore, we used a modified implementation for the conversion algorithm, which would integrate different lengths of the words. The following subsection shows how we modified our conversion approach to take advantage of all restriction enzymes' cut positions.

Since there are a limited number of restriction enzymes, we had to integrate all of them in the converted sequence. To do so, we looked at restriction enzymes of length 4 and identified their locations in the sequences. Then we moved on to restriction enzymes of length 5, 6, 7, and 8. This would give priority to words with shorter lengths first, then move up with longer words.

Again, these words have names/codes in their list, so the generated sequences would have a new alphabet that represents words of different lengths and biological relevance. The rest of the experiment would be the same as in the previous two experiments. The following example shows the new modification for the conversion approach. For example, assume this sequence: ACCGTGC, the restriction enzymes list we have with their codes is:

ACCG	= RE1
CGTG	= RE2
ACCGT	= RE3

Applying the restriction enzymes of length 4 would generate: RE1 (at position 1), RE2 (at position 3), while applying the restriction enzymes of length 5 would generate RE3 (at position 1). The final sequence of restriction enzymes after integrating both lengths would be: RE1, RE3, RE2.

Figure 4 shows the results of using a list of restriction enzymes' cut positions on the mitochondria dataset. The results showed better quality for the application of the restriction enzymes' list than those results from using MSA. Similar results are shown in Figure 9, using mitochondrial genomes. Again the results of the proposed approach outperformed those obtained by MSA.

Figures 4 and 5 show that the results obtained using restriction enzymes are generally better than those obtained using multiple sequence alignment. However in some cases, the random selection (refer to the second experiment) might produce better results, as shown in Figure 4. Using k = 6 and the random selection of 60 percent, we got a 0.489 percent of tree distance difference to the gold standard tree.

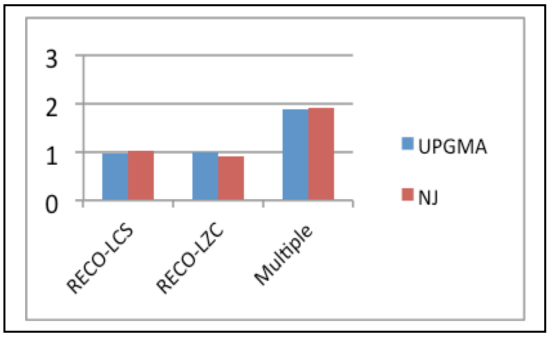


Figure 4. These are the results of using the algorithm with a list of restriction enzymes' cut positions on the mitochondrial dataset and using LZC; MSA results are included for comparison.

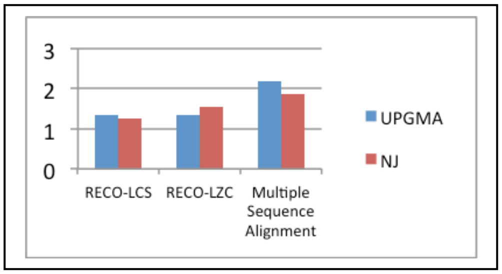


Figure 5. These are the results of using our algorithm with a list of restriction enzymes on the mitochondrial dataset and using LCS. MSA results are included for comparison.

That proves that there are some strong signals known to the literature, and those signals would improve the results of the comparison method. This yields a positive answer for the third question.

D. *The Fourth Experiment: Using k-mers that occur only in CDs regions of the genomes*

As our hypothesis of using words with biological relevance provided promising results, we continued searching for more signals that would also give high quality. One way to find such signals is to use signals/words from the CDs regions of the genomes. As these regions are rich with biological information, we proposed that they would improve the results. In fact, CDs are the main DNA source for functional genes, and a lot of species that are closely related would have similar functions, and in turn genes with similar structures that exist in these CDs regions. Therefore, we generated a list of k-mers that occur in the CDs regions.

We eliminated word lists of lengths 3, 4, and 5, as those lists were all possible k-mers of these lengths and would have same exact results as in the first experiment. The used dataset here was entire genomes. These mitochondrial genomes are rich with CDs regions and were a good fit for this experiment, as they also have a gold standard tree. Figure 6 shows better results when signals from the coding sequences are used in our approach. Figure 6 shows that these signals are rich with information that would improve the quality of the method. Therefore, these signals would be a major source as input lists of the approach. This yields a positive answer for the fourth question.

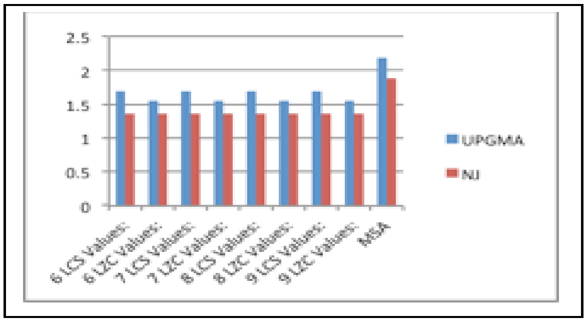


Figure 6. This figure shows the results of using our algorithm with lists that were generated from CDs regions, k ranges from 6 to 9. The used methods of comparison are LCS and LZC, and the clustering methods are UPGMA and NJ.

E. *The fifth experiment: Application of the approach to datasets with different level of gaps errors*

We finally applied the approach to special datasets. These datasets were generated and manufactured from the mitochondrial genome dataset. They are incomplete genomes and/or with errors. The reason for applying the approach to such datasets is to measure if it would be possible to identify the relatedness among species with errors.

These datasets are divided into three categories. The first category is for a dataset where each sequence is a fragment from the original genome, each fragment's content is a percentage of the original genome's content, and was chosen randomly from the genome's content. For this category, we generated two datasets: one with 50 percent content, and the second for 70 percent content.

The second group was for datasets composed of several fragments from the original genomes, and these fragments are in order. So each sequence would be the merging of several fragments from the original sequence, and these fragments would have a content represented as a percentage of the original genome. These fragments were chosen randomly from the genome's content and did not overlap. For this category, we generated two datasets with percentages 30 percent and 90 percent.

The third category is similar to the second one, but the fragments were switched randomly. This means that now a sequence has fragments that are not in order, yet would still have a content that is represented as a percentage amount of the original genome. These datasets were generated with percentages of 40 percent and 80 percent.

We compare the results of using our approach on these datasets to those resulting from MSA on the same datasets. We are evaluating whether our approach would identify the relatedness of the species in these datasets, even if they have errors, and whether these results would be better than those of MSA.

Figure 7 shows the results of applying the approach to these datasets. Each group of columns (blue, red, and green) represents one dataset and the use of one clustering algorithm (NJ or UPGMA). Each column shows the result of using LCS, LZC, or MSA. The results show that in most cases our approach outperforms MSA, except in two cases. As with the dataset of using several fragments, with 30 percent contents of the original sequences, and using all possible 4-mers, and LCS comparison method with UPGMA clustering algorithm, the quality of result was lower than MSA, same with the dataset of (80% contents, several fragments not in order, 6-mers selected from CDs and using LCS and UPGMA), the result again was lower than MSA.

Using the motif-based approach for comparing sequences in datasets that contain errors would be more effective than using MSA, as most of the results of our approach outperformed MSA results. Thirty-four results were better than MSA out of 36 runs (94.44 percent).

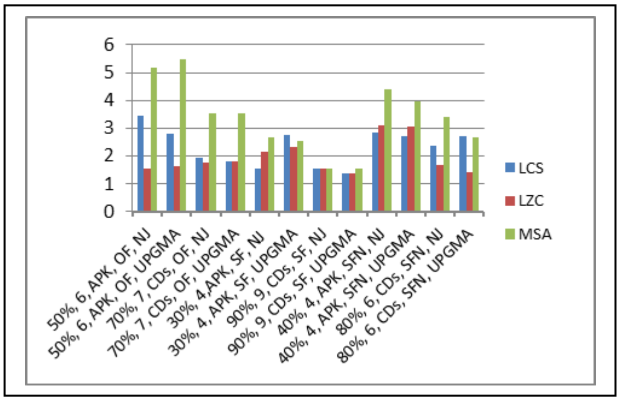


Figure 7. The results of applying proposed approach to datasets with high degree of errors. Abbreviations: APK(All Possible K-mers), CDs(Coding Regions), OF(One Fragment), SF(Several Fragments), SFN(Several Fragments Not in order).

VI. CONCLUSIONS

In this paper, we introduced an alternative approach to compare biological sequences, and that method in several cases outperforms the traditional alignment-based approaches. The proposed method is developed to compare sequences based on their inclusion of short biological signals or motifs. The conducted experiments showed that the effectiveness of the approach depends on which motifs used. In particular, the results showed that with a number of biologically significant signals/words, we should be able to produce results far superior than those obtained using alignment.

The proposed approach produced comparable results, and even better in certain cases, when random or generic motifs are used to compare sequences. Better results were obtained when certain motifs were used. Future work should focus on identifying different types of biological signals that can be utilized for better classification.

When we used short motifs associated with biological significance, such as restriction enzymes, the motif-based comparison produced even better results. Similar results were obtained when motifs are selected from rich regions in the genome such as coding regions. These motifs made it possible for the approach to outperform traditional methods like MSA in identifying the relatedness between genomes.

We also compared the proposed method with MSA using datasets that contain sequencing errors, and again for the majority of the cases, the signals-based comparison produced better results and better identified relationships among species. So for genomic datasets that have errors, it would be better to use this approach instead of using traditional alignment-based methods. The overlapped signals would identity such relationships among species. Overlapped signals are powerful marks for comparing biological sequences and would identify more accurate relationships between species as compared to alignment-based methods.

REFERENCES

[1] S. Vinga, J. Almeida, “Alignment-free sequence comparison – a review,” Bioinformatics, 19(4), pp.513–23, 2003.

[2] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, “Basic local alignment search tool,” J Mol Biol 215 (3), pp. 403–410, 1990.

[3] R. Edgar, “MUSCLE: Multiple sequence alignment with high accuracy and high throughput,” Nucleic Acids Research 32(5), pp. 1792-9, 2004.

[4] A. Subramanian**,** M. Kaufmann, and B. Morgenstern, “DIALIGN-TX: Greedy and progressive approaches for segment-based multiple sequence alignment,” Algorithms for Molecular Biology, 3:6, 27 May 2008

[5] K. Song, J. Ren, G. Reinert, M. Deng, M. S. Waterman, F. Sun: "New developments of alignment-free sequence comparison: measures, statistics and next-generation sequencing", Brief Bioinform (2013) doi: 10.1093/bib/bbt067

[6] G. Lu, Sh. Zhang, and X. Fang: “An improved string composition method for sequence comparison,” Symposium of Computations in Bioinformatics and Bioscience (SCBB07) Iowa City, 28 May 2008.

[7] O. Bonham-Carter, J. Steele and D. Bastola**, “**Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis”, Briefing in Bioinformatics, August 23, 2013

[8] Z. Liu, J. Meng, X. Sun, "A novel feature-based method for whole genome phylogenetic analysis without alignment: application to HEV genotyping and subtyping." Biochem Biophys Res Commun 2008; 223:223–30.

[9] V Arnau, M Gallach, I Marín. “Fast comparison of DNA sequences by oligonucleotide profiling.” BMC Res Notes 2008;1:5.

[10] T. J. Wu, Y. H. Huang, L. A. Li . "Optimal word sizes for dissimilarity measures and estimation of the degree of dissimilarity between DNA sequences." Bioinformatics 2005; 21:4125–32.

[11] G. E. Sims, S. R. Jun, G. A. Wu, SH. Kim. “Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions.” Proc Natl Acad Sci USA 2008;106:2677–82.

[12] J. Lin, "Divergence measures based on the shannon entropy." IEEETrans InfTheory 1991;37:145–51.

[13] Z. G. Yu, L. Q. Zhou, VV Anh, "Phylogeny of prokaryotes and chloroplasts revealed by a simple composition approach on all protein sequences from complete genomes without sequence alignment." J Mol Evol 2005;60:538–45.

[14] R. H. Chan, T. H. Chan, H. M. Yeung and R. W. Wang, “Composition vector method based on maximum entropy principle for sequence comparison.” IEEE/ACM Trans in Comput Biol Bioinf, Mar 3, 2011.

[15] I. Soares, A. Goios, A. Amorim. “Sequence comparison alignment-free approach based on suffix tree and l-words frequency.” SciWorldJ 2012;2012:450124.

[16] R. Sengupta, D. Bastola, H. Ali, “Classification and identifMSAcation of fungal sequences using characteristic restriction endonuclease cut order,” J. Bioinformatics and Comp Biology, pp. 181–198, 2010.

[17] B. Haubold, N. Pierstorff, F. Möller, and T. Wiehe, “Genome comparison without alignment using shortest unique substrings,” BMC Bioinf, 6:123, 23 May 2005.

[18] S. Aluru, Handbook of Computational Molecular Biology, pp. 15–10, 2005.

[19] H. Otu, K. Sayood, “A new sequence measure for phylogenetic tree construction,” Bioinformatics Vol. 19, no. 16, pp. 2122–2130, 2003.

[20] R Mina, D Bastola and H Ali, “Compression- based Algorithms for Comparing Fragmented Genomic Sequences”, BIOTECHNO 2013, The Fifth Int Conf on Bioinformatics, Biocomputational Systems and Biotechnologies, Lisbon, April 2013.

Codifying Primary Protein Structure as Peptides Frequencies Vector

An Efficient Alternative Method to Investigate Relationships among Genes and Organisms

Braulio R. G. M. Couto[1], Bruna A. Coimbra[1], Gabriel B. Tofani[2], Gustavo P. Irffi[2] and Cinthia T. V. Rocha[1]
[1]Instituto de Engenharia e Tecnologia- IET; [2]Instituto de Ciências Biológicas e da Saúde - ICBS
Centro Universitário de Belo Horizonte - UniBH
Belo Horizonte, MG, Brazil
braulio.couto@unibh.br, brunatecquimica@hotmail.com, bandeiragt@yahoo.com, palmer.gustavo@gmail.com,
cinthia.rocha@prof.unibh.br

Marcos Augusto dos Santos
Departamento de Ciência da Computação - DCC
Universidade Federal de Minas Gerais - UFMG
Belo Horizonte, MG, Brazil
marcos@dcc.ufmg.br

***Abstract*—There are no uniquely correct methods for inferring phylogenies. In this scenario, Linear Algebra methods and visualization techniques are essential to better analyze complex systems and can be very helpful to categorize species. Thus, if proteins are represented as vectors, it will be possible to apply Linear Algebra methods in order to investigate relationships among genes and organisms. To investigate whether or not primary protein sequences and genomes can be represented as vectors and processed by Linear Algebra methods to generate accurate phylogenetic relationships, four sets of sequences were analyzed by classical phylogenetics techniques, based on pairwise multiple alignments, and by Linear Algebra and optimization methods. The results showed that primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space the relationships among species are consistent with classic phylogenetic trees.**

Keywords-protein sequences; phylogenetics; linear algebra methods.

I. Introduction

Phylogenetics or cladistics aims to reconstruct the evolutionary relationship among genes and organisms and to establish classifications that reflect those genealogies. Its fundamental axiom is that, as a product of evolution, nature is hierarchically ordered [1]. Unfortunately, there are no uniquely correct methods for inferring phylogenies, and many methods have been used [2]. Given an alignment and a tree scoring function, there are no efficient methodologies to calculate an optimal phylogeny. In this scenario, Linear Algebra methods and visualization techniques are essential to better analyze complex systems and can be very helpful to categorize species. So, if proteins are represented as vectors, it will be possible to apply Linear Algebra methods to investigate relationships among genes and organisms.

Here, we investigated whether or not primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space it generates relationships among species consistent with classic phylogenetic trees. The objective of our study is to answer three questions: Is it possible to represent proteins and genomes as tripeptide frequency vectors? Are phylogenetic trees constructed by using Euclidean distance between protein vectors consistent with phylogenetic trees constructed with alignments (classical phylogenetic trees)? Do images of genomes represented by multidimensional vectors and visualized in reduced tridimensional space generate relationships among species consistent with those described by classical phylogenetic trees?

In Section II we present the material and methods to codify proteins and genomes as tripeptide frequency vectors, and how to build phylogenetic trees by using Euclidean distance between protein vectors. In the same section, we show the visualization technique of genome vectors in reduced space. In Section III, we present and discuss the results. Section IV concludes the paper.

II. Material and Methods

Four sets of sequences were analyzed by classical phylogenetics techniques, based on pairwise alignments, and by Linear Algebra and optimization methods. Firstly, the origin of the Human Immunodeficiency Virus (HIV) was analyzed, retrieving from GenBank the three longest coding regions from seventeen different isolated strains of the Human and Simian immunodeficiency virus (SIV): the gag protein, the pol polyprotein and the envelope polyprotein precursor [3][4][5]. The second database was composed by the complete genome of five strains of *Chlamydophila pneumoniae* that were retrieved from the NCBI (National Center for Biotechnology Information) website [14].

ISBN: 978-1-61208-335-3

The third dataset was similar to that used by reference [6] being composed by 59 whole mitochondrial genomes from the NCBI genome database, each one with 13 genes, totaling 767 proteins. Mitochondrial DNA (mtDNA) is one of the most fundamental evolutionary markers for phylogenetics due to the absence of recombination, high mutation rate and ease of sequencing. The following genes were analyzed: ATP Synthase F0 subunit 6 (ATP6), ATP Synthase F0 subunit 8 (ATP8), Cytochrome C Oxidase subunit 1 (COX1), Cytochrome C Oxidase subunit 2 (COX2), Cytochrome C Oxidase subunit 3 (COX3), Cytochrome B (CYTB), NADH Dehydrogenase subunit 1 (ND1), NADH Dehydrogenase subunit 2 (ND2), NADH Dehydrogenase subunit 3 (ND3), NADH Dehydrogenase subunit 4 (ND4), NADH Dehydrogenase subunit 4L (ND4L), NADH Dehydrogenase subunit 5 (ND5), and NADH Dehydrogenase subunit 6 (ND6). The 59 species were: Goldfish, Tasseled-mouth loach, Commom Carp, Zebrafish, Atlantic Cod, Olive flounder, Ornate Bichir, inbow trout, Atlantic salmon, Salvelinus, Brook trout, Redhead (Bird), Oriental Stork, White Stork, Peredrin Falcon, Red Junglefowl, Rook (Bird), Grey-headed Broadbill, Village Indigobird, Rhea, Ostrich, Cattle (Cows), Hippopotamus, Sheep, Wild boar, Dog, Cat, Grey Seal, Harbor seal, Blue Whale, Fin Whale, Jamaican Fruit Bat, Nine-Banded Armadillo, Virginia Opossum, Wallaroo, European Headhog, European rabbit, Platypus, White Rhinoceros, Asinus, Horse, Indian rhinoceros, Western Gorilla, Human, Orangutan, Chimp Vellerosus, African Bush Elephant, Guinea Pig, House Mouse, Edible Dormouse, Brown Rat, European mole, Aardvark, American Alligator, Ryukyu odd-tooth snake, Mole Skink, Green Sea Turtle, Painted Turtle, and African helmeted turtle.

The last database analyzed was composed by mitochondrial D-loop sequences for the Hominidae taxa (pongidae) [7]-[10]. We used this dataset because mitochondrial D-loop is very useful for comparing closely related organisms provided that it is one of the fastest mutating sequence regions in animal DNA. The origin of modern man is a highly debated issue that has recently been tackled by using mtDNA sequences. The limited genetic variability of human mtDNA has been explained in terms of a recent common genetic ancestry: all modern-population mtDNAs originated from a single woman who lived in Africa less than 200,000 years ago [7][8][9]. This family embraces the gorillas, chimpanzees, orangutans and the humans. Species description (and GenBank access code): German Neanderthal (AF011222), Russian Neanderthal (AF254446), European Human (X90314), Puti Orangutan (AF451972), Jari Orangutan (AF451964), Mountain Gorilla Rwanda (AF089820), Western Lowland Gorilla (AY079510), Eastern Lowland Gorilla (AF050738), Chimp Schweinfurthii (AF176722), Chimp Vellerosus (AF315498), Chimp Verus (AF1767310), and Chimp Troglodytes (AF176766). Both HIV and primates analysis (the first and fourth datasets) were based on demos of the bioinformatics toolbox from MATLAB (The MathWorks, Inc).

In order to visualize the genomes, we must represent each one as a point in space. The distance between the points should represent the differences in the genomes as a whole. Therefore, we might expect similar species to be close together in space. The genome proteins were represented as vectors of frequencies of groups of amino-acids. In this study, a sliding window of size 3 was used to measure the frequency (tripeptide frequency vectors). To represent the genome we used the vector sum of all its proteins. Therefore, we can obtain a database of genomes, S, as a rectangular matrix, X, where each line corresponds to one of the n genomes:

$$X = (X_1, X_2, ..., X_n)^T = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \dots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{pmatrix}$$

With 20 amino-acids, each primary protein structure is codified as a vector with m = 20^3 = 8,000 dimensions. In the above matrix, X_{ij} represents frequency o tripeptide i in the genome j (i = 1, 2, …n; j = 1, 2, …m; where n = number of genomes and m = 8,000 possible tripeptides). To codify a complete genome, the vector representations of genes from individual species are summed [6][11][12]. Therefore, both genes and genomes are represented as tripeptide frequency vectors with 8,000 dimensions. To generate a suitable visualization of these vectors, it is necessary to reduce the dimensionality of the space, with the minimum loss of information. Reference [13] presented a visualization technique to analyze chemical databases. We used that technique as a basis to develop a method for using genomes to visualize relationships among species in reduced space (3D). Actually, the high-dimensional visualization problem in $\Re^m$ can be formulated as a distance-geometry problem: to find n points in low space (2D or 3D) so that their interpoint distances match the corresponding values from $\Re^m$ as closely as possible. When a representation in reduced space, Y, is generated for the database matrix X, we can calculate an error function **E** as the following:

$$E = \sum_{i=1}^{n} \sum_{j=1}^{n} \left(\delta_{ij} - \gamma_{ij}\right)^2$$

where δ_{ij} is the Euclidean distance between protein vector (or genome) i and j in the original space, represented in the matrix X, and γ_{ij} is the Euclidean distance between protein vector (or genome) i and j in the reduced space, represented in the matrix Y. The best representation of X in the reduced space will be the Y with the minimal associated error function. Therefore, we must solve an unconstrained optimization problem. Many methods can be used to solve this problem [13]. In this study, we used a technique based on the interior-reflective Newton method, actually the conjugated gradient Newton's method [11]. Classical phylogenetic tree reconstructions were made by using

ISBN: 978-1-61208-335-3

pairwise alignments algorithms. Alternative phylogenetic trees were obtained by UPGMA or Neighbor-Joining methods using, as pairwise distances, the Euclidean distance between each protein or genome vector of the dataset.

III. RESULTS

Phylogenetic tree reconstruction of 17 strains of the Human (HIV) and Simian immunodeficiency virus (SIV) in the first dataset were made for each coding region. For the GAG coding region we used the Tajima-Nei method to measure the distance between the sequences and the unweighted pair group method using arithmetic averages (UPGMA) for the hierarchical clustering. Phylogenetic tree for the POL polyproteins was made by using the Jukes-Cantor method to measure distance between sequences and the weighted pair group method using arithmetic averages (WPGMA) for the hierarchical clustering. For the ENV polyproteins we used the normalized pairwise alignment scores as distances between sequences and the UPGMA method. Given that the three trees showed slightly different results, a consensus tree using all three regions was built using a weighted average of the three trees (Figure 1). All three sequences from the GAG, POL, and ENV regions of the 17 HIV and SIV strains were codified as tripeptide frequency vectors that can be visualized in tridimensional space (Figure 2). Phylogenetic tree of HIV and SIV viruses reconstructed by Neighbor-joining method using as pairwise distances the Euclidean distance between each tripeptide frequency vectors (Figure 3). Both phylogenetic trees (Figure 1 and 3) and the vectors in 3D space (Figure 2) illustrate the presence of two clusters and some other isolated strains. The most compact cluster includes all the HIV2 samples, and the second cluster contains the HIV1 strain (not as compact as the HIV2 cluster). From the trees it appears that the Chimpanzee is the source of HIV1.

Nucleotide sequences of five isolated strains of *Chlamydophila pneumoniae* were converted into amino-acid sequences, codified as tripeptide frequency vectors and projected in reduced space (3D): strains J138 and TW-182 occupy the same region of space, near CWL029 strain and far from AR39 and LPCoLN. Before applying the method proposed herein, we tried to generate multiples alignments using the complete genomes of *Chlamydophila pneumonia* for reconstructing the classical dendogram (Figure 4) in MATLAB. An “out of memory error” occurred in a computer with an Intel(R) Core(TM) i5 processor and 6GB RAM memory. It was impossible to reconstruct the classical phylogenetic tree in that computer because the genome was more than 1,200,000 nucleotides long and there were more than 400,000 amino-acids. The computer time and memory consumption grew quickly due to the size of each genome, generating computational fatal error. This was not a problem for the Linear Algebra method proposed here: all five genomes were codified as tripeptide frequency vectors and projected in 3D space using the same computer. Phylogenetic tree reconstructed with the Euclidean distance between each tripeptide vectors (Figure 5) was equal to the classical dendogram in Figure 4. For the third dataset, amino-acid sequences of thirteen mitochondrial genes from 59 different species were codified as tripeptide frequency vectors and projected in reduced space (3D): we observed exactly 13 compact clusters related to each gene analyzed (Figure 6). When these genes vectors were summed and the resultant genome vectors projected in tridimensional space, the four classes analyzed formed compact and distinct clusters (Figure 7). Species from the same class (Actinopterygii, Aves, Mammalia, and Reptilia) were placed in a monophyletic branch in the phylogenetic tree of the 59 whole mitochondrial genomes reconstructed by Neighbor-joining method using as pairwise distances between species the Euclidean distance between each tripeptide frequency vectors. When the phylogenetic tree for Mammalia species was reconstructed by using Euclidean distance between vector species, we observed a pattern compatible with classical trees (Figure 8). The last database evaluated was composed by mitochondrial D-loop sequences for the Hominidae taxa. Pairwise distances using the Jukes-Cantor formula and the phylogenetic tree with the UPGMA distance method were reconstructed (Figure 9). In this classical phylogenetic tree, Human resembles Neanderthal and Chimp species. However, when amino-acid from these mitochondrial D-loop sequences for the Pongidae taxa were codified as tripeptide frequency vectors and projected in reduced space (3D), we observed that European Human was more closely related to Gorilla than to Chimp species (Figure 10). In the phylogenetic tree built using Euclidean distance between vector species (Figure 11) Neanderthal, Chimp and Orangutan form monophyletic groups. However, differently from the classical tree, Human and Gorilla species were in the same branch.

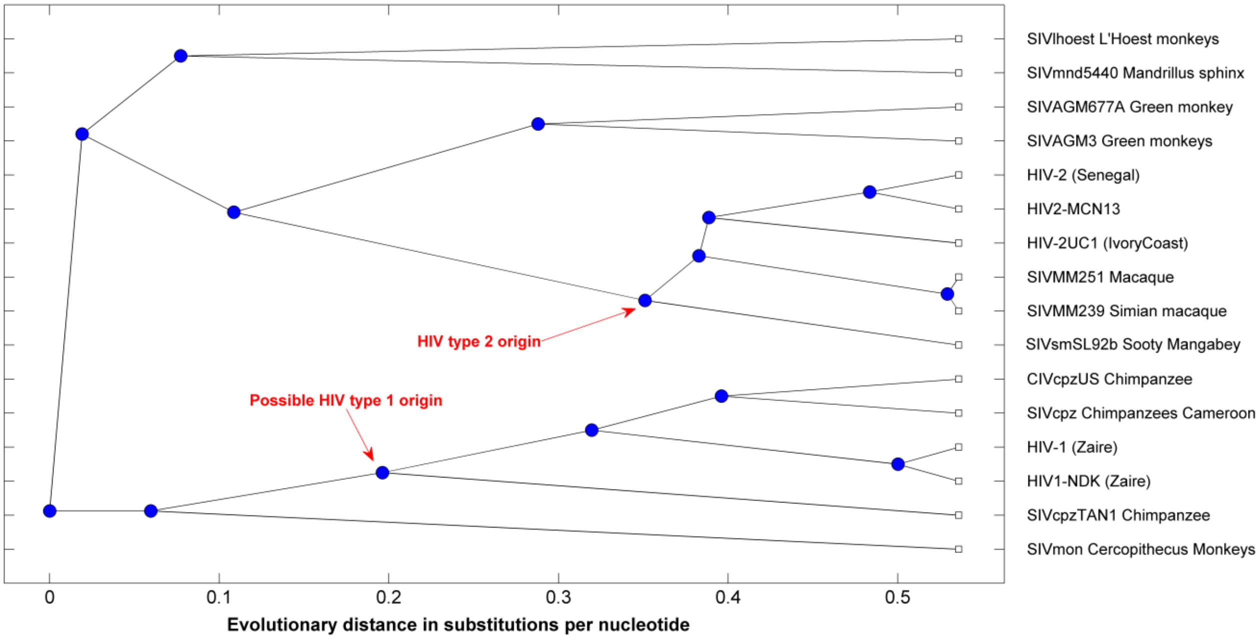


Figure 1. Phylogenetic consensus tree from multiple strains of the HIV and SIV viruses: the tree illustrates the presence of two clusters and some other isolated strains, showing possible origins for two characterized strains of human AIDS viruses, type 1 (HIV-1) and type 2 (HIV-2).

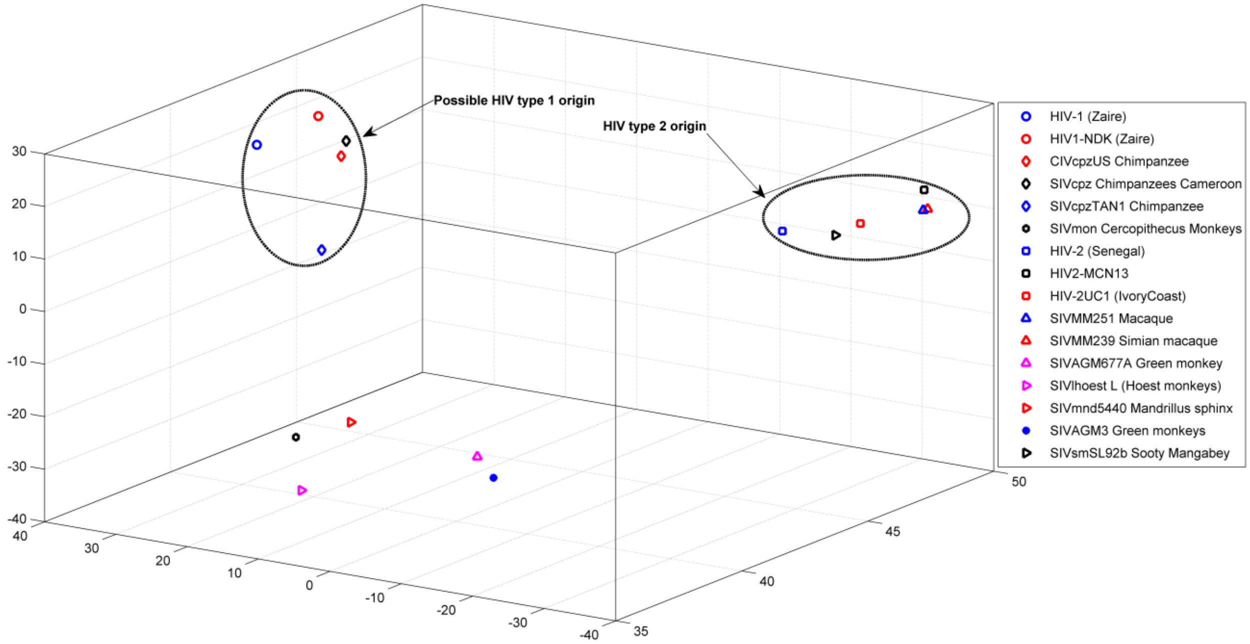


Figure 2. Nucleotide sequences from the GAG, POL, and ENV coding regions of seventeen different isolated strains of the Human and Simian immunodeficiency viruses were converted into amino-acid sequences, concatenated, codified as tripeptide frequency vectors and projected in reduced space (3D): the two clusters containing HIV1 and HIV2 are identified, as in the phylogenetic tree (Figure 1).

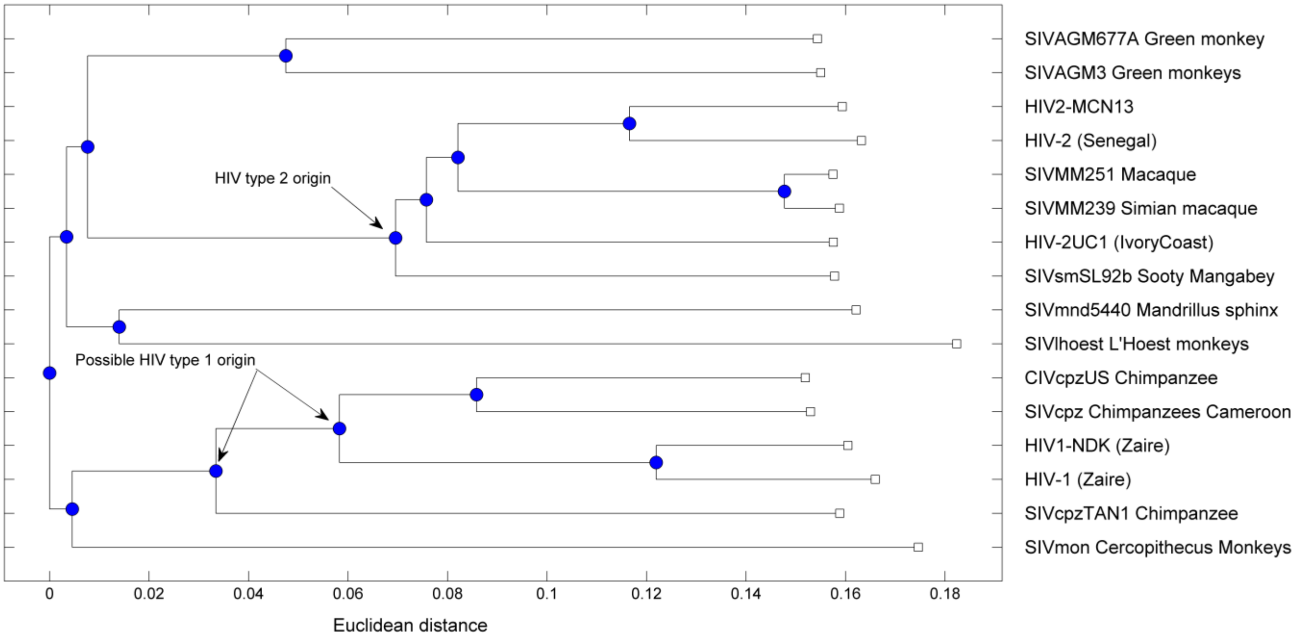


Figure 3. Phylogenetic tree of HIV and SIV viruses reconstructed by Neighbor-joining method using the Euclidean distance between each tripeptide frequency vectors: the pattern obtained is identical as that showed in classical phylogenetic tree from Figure

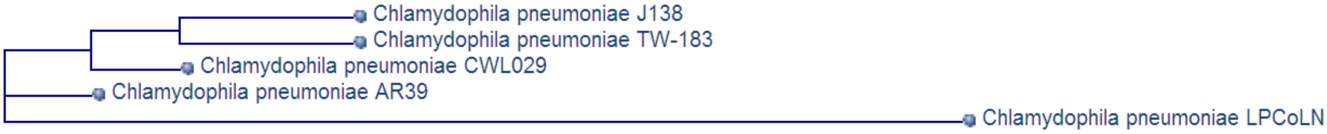

Chlamydophila pneumoniae J138
Chlamydophila pneumoniae TW-183
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae LPCoLN

Figure 4. Classical dendogram, constructed based on genomic BLAST, for five strains of *Chlamydophila pneumoniae* (available at reference [14]).

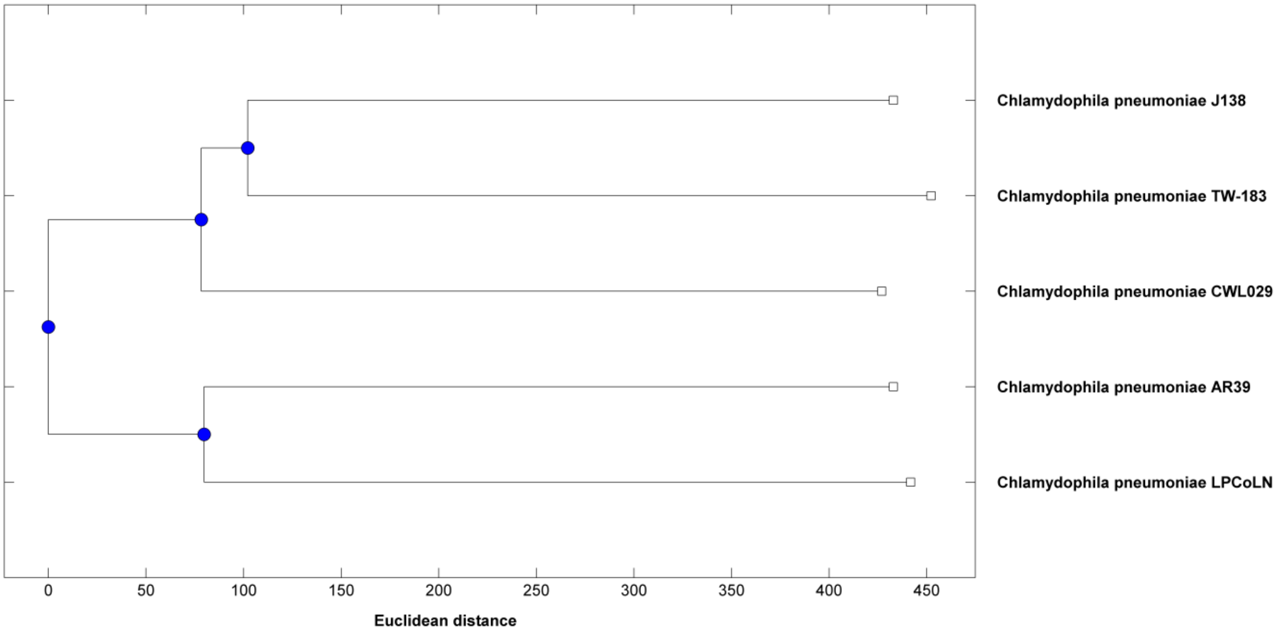

Chlamydophila pneumoniae J138
Chlamydophila pneumoniae TW-183
Chlamydophila pneumoniae CWL029
Chlamydophila pneumoniae AR39
Chlamydophila pneumoniae LPCoLN
0
50
100
150
200
250
300
350
400
450
Euclidean distance

Figure 5. Phylogenetic tree of *Chlamydophila pneumoniae* strains reconstructed by Neighbor-joining method using as pairwise distances between strains the Euclidean distance between each tripeptide frequency vectors: there in no difference between this tree and the classical dendogram from Figure 4.

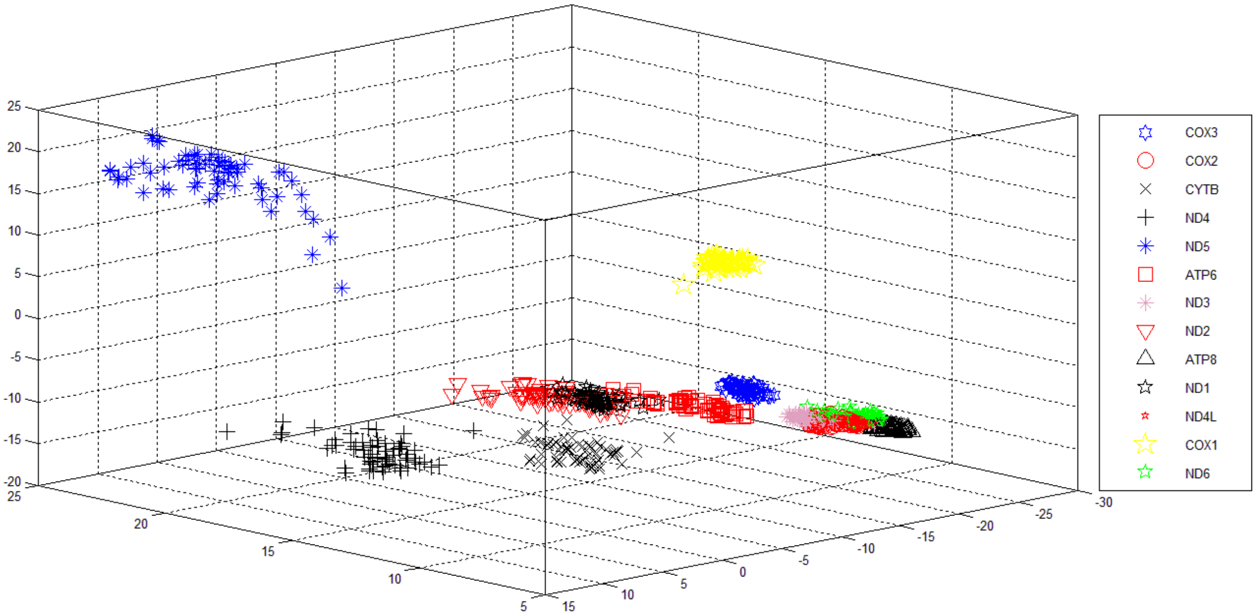


Figure 6. Visualization in reduced space (3D) of thirteen mitochondrial genes from 59 different species, codified as tripeptide frequency vectors: there are exactly thirteen compact and well defined clusters of genes.

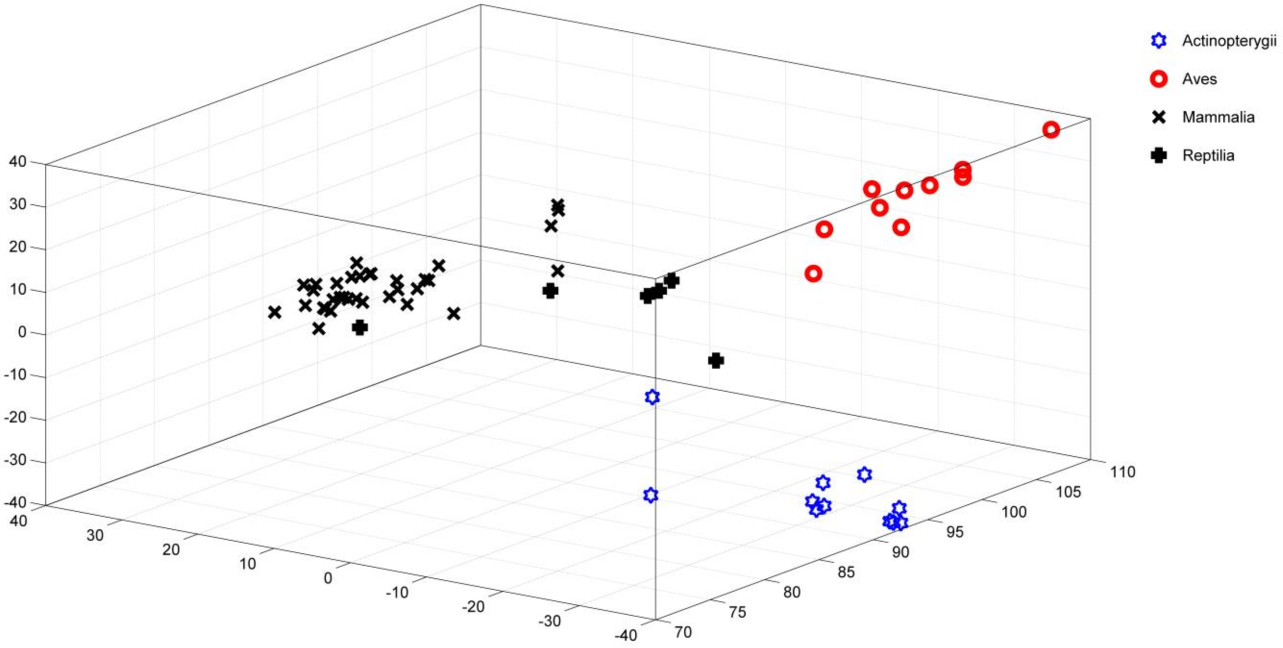


Figure 7. Amino-acid sequences of thirteen mitochondrial genes from 59 different species were codified as tripeptide frequency vectors, summed to represent each specie vector and projected in reduced space (3D): there are exactly four compact clusters related to the four class analyzed (Actinopterygii, Aves, Mammalia, and Reptilia).

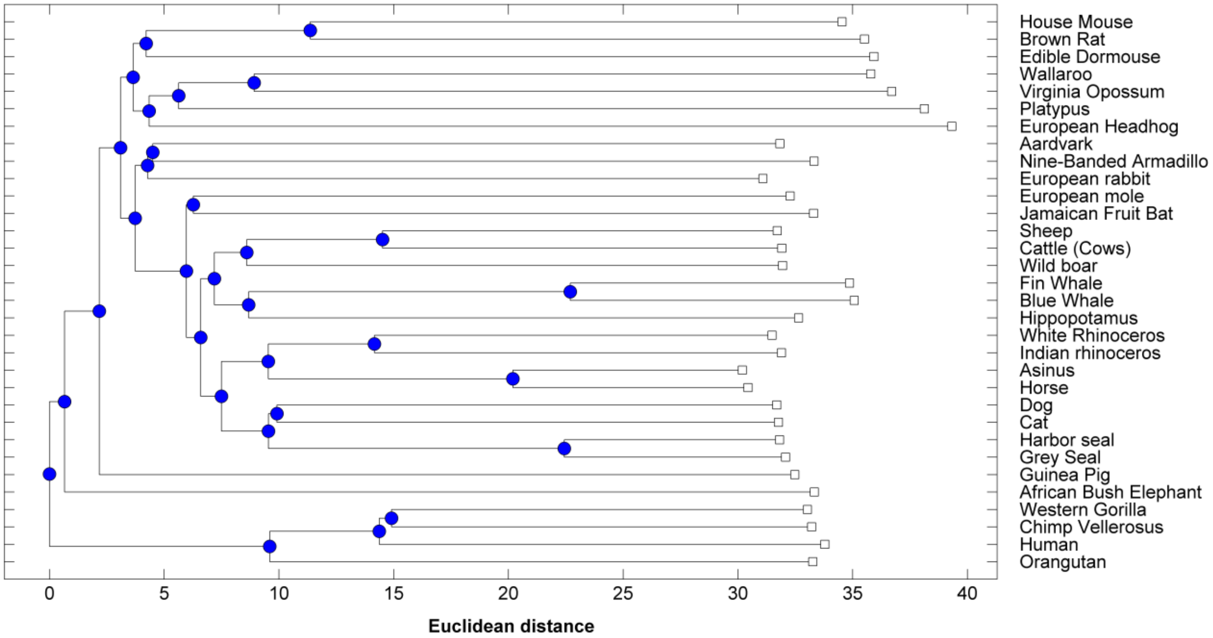


Figure 8. Phylogenetic tree of whole mitochondrial genomes from 32 Mammalia species reconstructed by Neighbor-joining method using Euclidean distance between each sum vector.

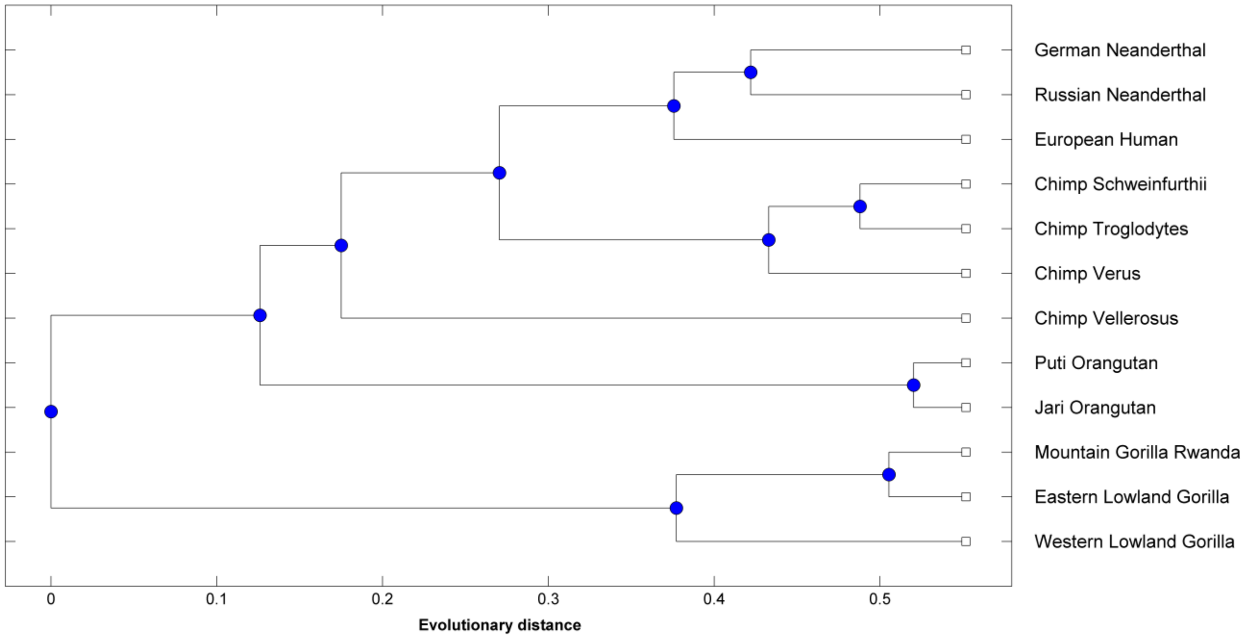


Figure 9. Classical phylogenetic tree from mitochondrial D-loop sequences for the Hominidae taxa reconstructed by pairwise distances using the Jukes-Cantor formula and the UPGMA distance method.

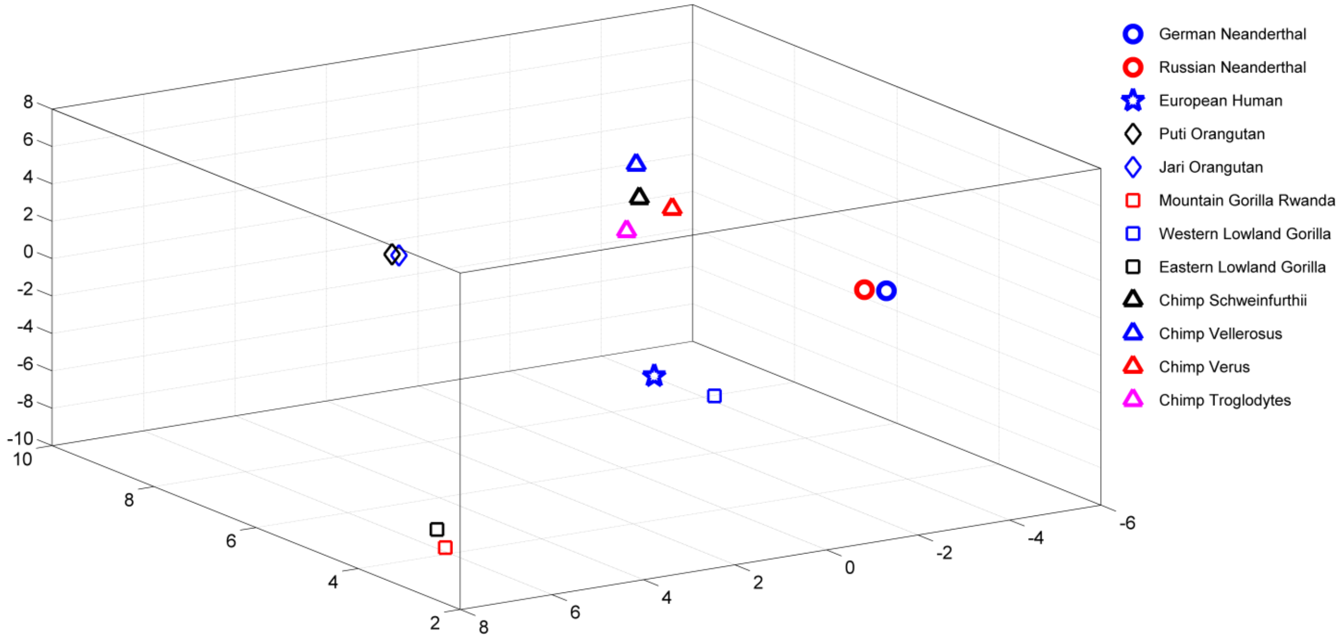


Figure 10. Mitochondrial D-loop sequences for the Hominidae taxa were codified as tripeptide frequency vectors and projected in tridimentional reduced space: groups of Neanderthal, Orangutan, Gorilla, and Chimp species occupy specific regions in the space. European Human is closer to Gorilla than Chimp.

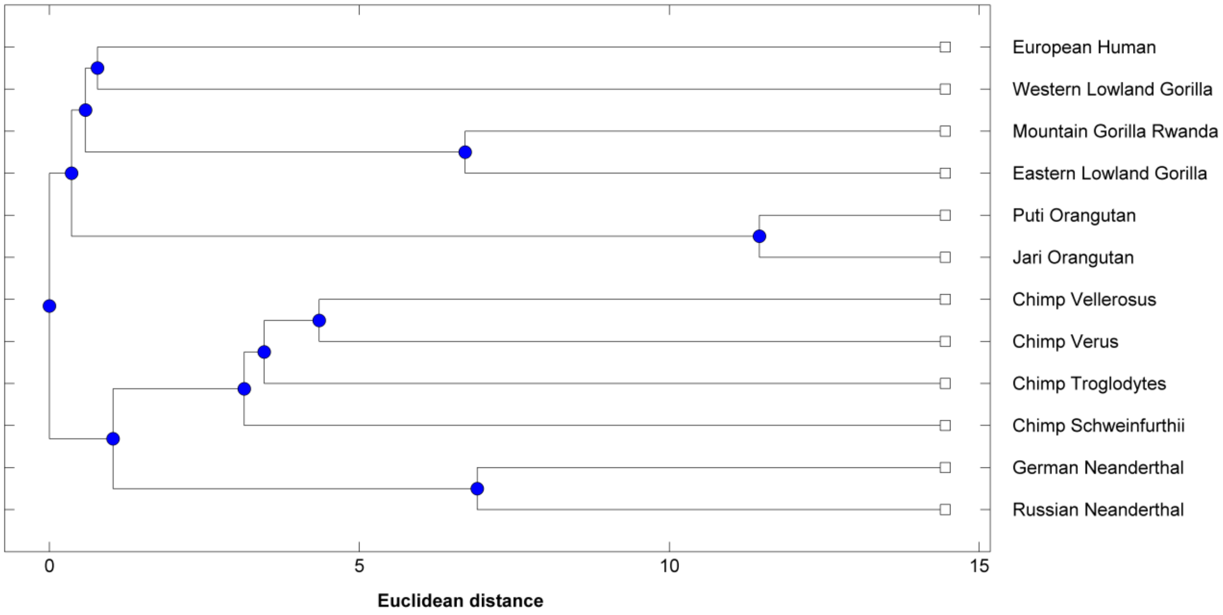


Figure 11. Phylogenetic tree from mitochondrial D-loop sequences for the Hominidae taxa reconstructed by using the Euclidean distance between specie vectors: Human and Gorilla are at the same branch.

IV. CONCLUSION

Primary protein structure from four different datasets were codified as peptides frequencies vector, visualized in reduced space (3D) and analyzed by using Euclidean distance as pairwise distances between genes and species. According to our results, the Linear Algebra method proposed seems to work very well, i.e., primary protein sequences and genomes can be represented as vectors in multidimensional space in such way that when they are mapped into 3D space they generate relationships among species consistent with classic phylogenetic trees. We illustrated by the previous results biological and computational benefits of such a methodology.

Therefore, it is possible to represent proteins and genomes as tripeptide frequency vectors and the analysis done with these vectors are consistent with classical analysis using multiple alignments. Besides, phylogenetic trees constructed by using Euclidean distance between vectors of proteins are compatible with trees constructed with alignments (classical phylogenetic trees). Images of genomes represented by multidimensional vectors and visualized in reduced tridimensional space generate relationships among species consistent with those described by classical phylogenetic trees.

Computationally, and mathematically the proposed method simplifies the study of the evolutionary chain of genes and genomes. It is a richer model to investigate relationships among sets of organisms than classical phylogenetics trees, because it takes into account not only pairwise distances but also the geometric position of each genome in a multidimensional space. As a result, computational load is substantially lowered and complete genome, as shown in *Chlamydophila pneumonia* analysis, can be easily done in a modest computer.

REFERENCES

[1] J. V. Crisci, L. Katinas, and P. Posadas, Historical Biogeography: an Introduction, Harvard University Press, Cambridge, 2003.

[2] M. Salemi and A. M. Vandamme, The Phylogenetic Handbook: a Practical Approach to DNA and Protein Phylogeny, 1st ed., Cambridge University Press, Cambridge, 2003.

[3] F. Gao, *et al*., "Origin of HIV-1 in the chimpanzee Pan troglodytes troglodytes," Nature 397(6718), 1999, pp. 436-41.

[4] H. W. Kestler 3rd, *et al*. "Comparison of simian immunodeficiency virus isolates," Nature 331(6157), 1998, pp. 619-622.

[5] M. Alizon, S. Wain-Hobson, L. Montagnier, and P. Sonigo, "Genetic variability of the AIDS virus: nucleotide sequence analysis of two isolates from African patients," Cell 46(1), 1986, pp. 63-74.

[6] G. W. Stuart, K. Moffett, and J. J. Leader, "A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes," Molecular Biology and Evolution 19(4), 2002, pp. 554–562.

[7] I. V. Ovchinnikov, A. Götherström, G. P. Romanova, V. M. Kharitonov, K. Lidén, and W. Goodwin, "Molecular analysis of Neanderthal DNA from the northern Caucasus," Nature 404(6777), 2000, pp. 490-493.

[8] A. Sajantila, *et al*., "Genes and languages in Europe: an analysis of mitochondrial lineages," Genome Research 5(1), 1995, pp. 42-52.

[9] M. Krings, A. Stone, R. W. Schmitz, H. Krainitzki, M. Stoneking, and S. Pääbo, "Neandertal DNA sequences and the origin of modern humans," Cell 90(1), 1997, pp. 19-30.

[10] M. I. Jensen-Seaman, and K. K. Kidd, "Mitochondrial DNA variation and biogeography of eastern gorilas," Molecular Ecology 10(9), 2001, pp. 2241-2247

[11] L. S. Marcolino, B. R. G. M. Couto, and M. A. Santos, "Genome Visualization in Space," Advances in Soft Computing, 2010, pp. 225-232.

[12] A. Seetharam and G. W. Stuart, "Whole genome phylogenies for multiple Drosophila species," BMC Research Notes 2012, vol. 5, p. 670.

[13] D. Xie and T. Schlick, "Visualization of chemical databases using the singular value decomposition and truncated-Newton minimization," In: C. A. Floudas, P.M. Pardalos (eds), "Optimization in Computational Chemistry and Molecular Biology: Local and Global Approaches," vol. 40, Kluwer Academic Publishers, Dororecht/Boston/London, 2000, pp. 267-286.

[14] National Center for Biotechnology Information [Online] Available from: <http://www.ncbi.nlm.nih.gov/genome/171> 2014.02.28

Geometrical Detection of Pathways in Protein Structures Leading Among More Binding Sites

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*—In this paper, we present a novel algorithm for the detection of pathways connecting two or more specific user defined binding sites, which are deeply buried in a protein macromolecule. These pathways can play an important role in the protein reactivity and overall behavior. However, our new algorithm can be generalized and used for computation of pathways inside an arbitrary set of spheres in three-dimensional space, leading through an ordered set of user-defined sites. Our approach is based on the localized Voronoi diagram approach and the Delaunay triangulation. The greatest benefit of our approach is its independence on the size of the input data set. This is achieved by using only a subset of all atoms in the macromolecule in each phase. This substantially reduces the size of the processed space. The method can also be utilized for determination whether pathways wide and straight enough exist among determined binding sites. This information then serves as the guideline for assessing the migration of products of chemical reaction between these binding sites.**

Keywords–protein; ribosome; tunnel; channel; active site; binding site; Voronoi diagram; Delaunay triangulation

I. INTRODUCTION

Proteins are organic molecules, which are irreplaceable for every life form on our planet. They perform specific tasks and are involved in many vital processes in living organism's cells. Analysis of proteins helps to understand their complex structure and chemical principles. In the quickly evolving area of protein engineering, where new substances like drugs and inhibitors are being designed, there are strong needs to explore possibilities of protein modification via transporting ligands to the desirable spots (active sites) inside the protein structure, causing a chemical reaction. However, verification of each potential reaction by an experiment in a laboratory is extremely time-consuming. Based on this fact, heuristic methods helping to exclude impossible cases and to suggest highly probable variants are utilized for significant decrease of in vitro experiment time. Methods that have been developed so far are mostly based on the computational geometry approach. Most of them are stochastic and present an approximation to the real situation, but as proven by numerous experiments and comparisons performed by the biochemists, they still provide an essential information to the biochemical community.

These methods concentrate on analysis and detection of specific inner structures inside molecules, such as cavities, pockets, tunnels, channels, pores, etc. Long-term research of biochemical properties of proteins revealed that the protein reactivity strongly depends on the presence of specific voids (cavities) inside the molecule, called active sites [1]. These active sites subsequently represent the destination of small ligand molecule that is transported from the outside solvent via pathways called tunnels. The ligand can then react with atoms surrounding the active site and the product of such reaction can serve as a basis of new medications or other important chemical compounds. The presence of tunnels on proteins along with particular examples is thoroughly described in Chapter 17 [2].

Until now, we have concentrated on the detection of tunnels from one active site. But, some enzymes contain two or even more active sites located on separate domains or subunits and all of them can participate on the final product structure. The product of the chemical reaction between the ligand and the protein in the first active site can travel to the second active site. Here, another reaction can modify this product. The direct transfer of ligands between active sites prevents the release of labile substrates into the outer solvent or the entrance of other intermediates, which can compete with already present ligands. Moreover, using the intramolecular tunnels, biochemists are able to regulate a set of consecutive chemical reactions [3].

The first tunnel connecting distinct active sites was discovered for tryptophan synthase from Salmonella typhimurium [4], which contains two active sites. It is obvious that in this case we have to detect not only tunnels connecting the active site with the outside solvent but also intramolecular tunnels between two and more active sites. This situation demands introducing a new solution to tunnel computation, which is the main goal of this article. Intramolecular tunnels are commonly present in ammonia-transferring enzymes. Many such enzymes have been already studied in some detail [5], [6], including the following structures – carbamoyl phosphate synthetase (PDB ID of wild type (WT) 1BXR), glucosamine 6-phosphate synthase (WT 2J6H), glutamate synthase (WT 1OFD), imidazole glycerol phosphate synthase (WT 1KA9) or cytidine triphosphate synthetase (WT 1VCM).

This algorithm can also be generalized and utilized for user-driven tunnel computation, when the biochemists can, according to their prior knowledge, determine some important spots inside the protein that should the tunnel follow.

When dealing with molecules consisting of thousands of atoms, the analysis of their structure at once is possible. However, for large macromolecular structures containing hundreds of thousands of atoms, such as ribosomes, the original approach is inapplicable. But, when dividing the area of interest into smaller regions and computing the tunnel piecewise, we can obtain acceptable and relevant results.

II. Related work

In this section, we concentrate on existing approaches to the computation of various structures (some of them were mentioned in the first chapter). The first group of algorithms aims to calculate the molecular surface. This apparently unrelated structure has two practical exploitations for our purpose. First of all, it can contain so called pockets. The difference between a pocket and a tunnel is that the binding site of a pocket is directly accessible from the surface, whereas the binding site of a tunnel is deeply buried inside the protein. The basic idea of the algorithm is rolling the probe of user-specified radius over the protein. The probe touches the protein boundary atoms and from all probe positions the resulting molecular surface can be constructed. This idea describes the Alpha shape theory [7] and the Reduced surfaces theory [8]. As mentioned above, a suitable probe size is able to reveal pockets on the protein surface. However, this approach is not applicable to the binding sites, which are deeply buried in the protein structures. Such binding sites cannot be accessed by the rolling probe. Some extended approaches [9] or other techniques that compute other pathways for accessing the deeply buried binding sites, e.g., tunnels or channels have to be involved.

In general, detection of tunnels in proteins can be considered as a specialized path detection algorithm. The environment is formed by a set of obstacles (atoms), which are scattered in the 3D space and may overlap. For this problem, many approaches of path planning have been introduced so far. When selecting an appropriate solution, we have to take into account its extendability by involving time changes into the input set. It corresponds to molecular dynamics, when the inner forces and outer environment causes permanent movement of atoms. Our solution of the static case, which is described in this article, was designed with respect to this knowledge.

Most of the existing path planning algorithms are mainly designed for environments, where the obstacles are static. Examples of such approaches can be the Minkowski sum [10], visibility graphs [11], or cell decomposition [12]. In 3D space, the complexity increases significantly; but, several useful algorithms have been also introduced. Again, solutions for static environments are more common, such as the spatial indexing [13] or the velocity obstacles method [14]. Algorithms for 3D dynamic environment, which would satisfy our problem with molecular dynamics, are quite rare. General methods, such as [15], can be mentioned as an example. However, these solutions contain some input constraints or were not designed for large datasets. Also in the field of protein analysis, there is an input set of initial requirements, which have to be fulfilled by the algorithms. Two sophisticated techniques for the detection of tunnels in dynamics were designed [16][17].

The first algorithm designed primarily for the detection of tunnels in static environment was based on the grid method [18]. The key idea is enclosing the whole protein into a bounding box, which is then regularly sampled into a set of cubes (grid). Thanks to the dependence on the grid resolution, the results of this algorithm may vary substantially. Also its time complexity ($\mathcal{O}(n^3)$ with respect to the number of samples) is inapplicable to larger molecular systems. So, this method was successively replaced by Voronoi diagram-based approach [19][20].

Nowadays, the leading approaches in tunnel and channel detection are based on construction of the Voronoi diagram (VD) and its dual structure, the Delaunay triangulation (DT). The construction of DT is independent on any initial settings – the complexity of the QuickHull algorithm for DT computation is expected to be $\mathcal{O}(n \cdot log(n)$ [21]. According to the duality principle, the VD is obtained from the DT in $\mathcal{O}(n)$). Thanks to this time complexity, it is applicable to molecular dynamics in reasonable computational time.

The basic idea is as follows. For the whole protein structure, the DT for all centers of atoms is computed (for example using 4D Quick-Hull algorithm [21]). Afterwards, it is converted to its dual VD. However, atoms in the protein may be of different radii. In most cases, this fact is not taken into account [20], [22] and the resulting tunnels are still biochemically relevant. Of course, involving atom radii to the computation increases the preciseness. Few recent approaches take into account the different radii of atoms – by introducing some approximation [23] or implementing the additively weighted Voronoi diagram (AWVD) (see [24]). AWVD splits the space more naturally, thus the results are more accurate. But, it also leads to higher time and memory complexity.

After the VD or AWVD is constructed, it is converted into a graph. In this graph, each Voronoi vertex is represented by a node and connections of Voronoi vertices form edges. Every edge is then evaluated by a number representing the radius of maximal sphere with the center located on this edge and non-colliding with any atom of the protein (Figure 1). In the last phase, a search algorithm (such as Dijkstra or A*) is initiated. As a result, a set of collision-free paths is obtained. Note that the additional search criteria resulting in the detection of the shortest or the best-evaluated path can be applied (in our case the widest tunnels).

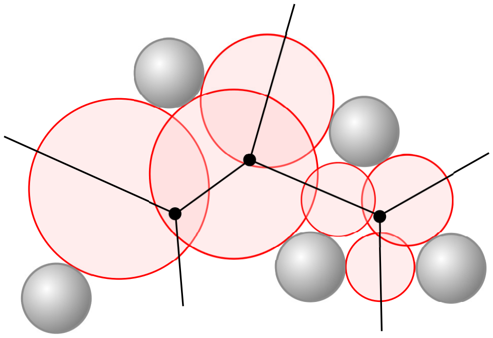


Figure 1: Description of the empty space inside the protein expressed by spheres. Spheres with maximized radius not intersecting any of atoms are centered on the edges of Voronoi diagram.

The main advantage of this VD-based method lies mainly in its precision. But still, the running time and memory consumption increase rapidly when being applied on large structures, such as ribosomes. In these cases, our newly designed approach is able to divide the structure into smaller parts which are then processed individually.

To our best knowledge, the above described methods were by now applied to the computation of tunnels leading from a binding site to the outer solvent. However, our solution aims to detect intramolecular tunnels connecting several deeply buried binding sites.

III. Algorithm

Our new method presented in this paper focuses, generally, on the detection of pathways of guaranteed minimal width connecting sites of interest inside a set of points. To demonstrate the applicability of this general approach, we customized it for the detection of tunnels among more active sites in large macromolecular structures, e.g., ribosomes.

The input set consists of points in 3D space (representing the centers of atoms). Users have to define points of interest – starting and ending points of each tunnel local segment (e.g., chemically relevant binding sites inside the ribosome). Previously described algorithms were designed for searching for arbitrary paths from the active site to the molecular surface. Users had no control over the trace of the computing tunnels. They could only choose the most suitable and relevant tunnels from the computed set. Using this new approach, users can control the trace of the tunnel by determining the desired spots, through which the tunnel should traverse.

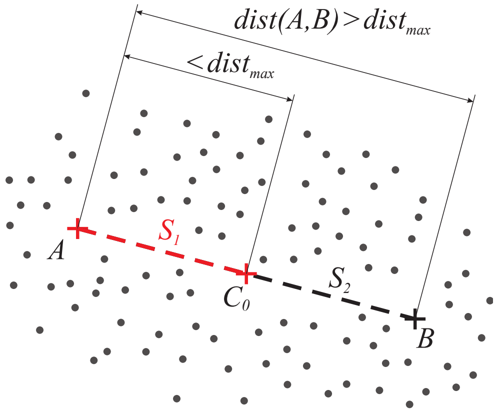


Figure 2: A plan for start-to-end connection, pivot segments S_1, S_2

By default, the criteria of our algorithm are set for searching ”straight” tunnels (more straight tunnels are often also more biochemically relevant). However, this requirement can be easily customized by users and easily adapted to compute the widest tunnels or tunnels fulfilling other properties or their combination.

```
Require: points $A$, $B$, distance $dist_{max}$
  if $dist(A,B) > dist_{max}$ then {split long segment}
    $S \leftarrow \{[A,C_0],...,[C_n,B]\}$
  else
    $S \leftarrow \{[A,B]\}$
  end if
  return $S$
```

Figure 3: Preprocessing phase: construction of the set of pivot segments

In the preprocessing phase (Figure 3), the algorithm evaluates the assignment and prepares a set of individual tasks for searching paths along shorter segments. Afterwards, this set is iteratively processed in four phases and pathways along the determined pivot segments are computed. Finally, individual paths are connected to form the resulting pathway – tunnel.

As the input, the algorithm requires a set of points in 3D P, starting A and ending B points. All those points have to be located inside the convex hull of the input set P. To overcome memory limitations and to avoid the calculation of large Voronoi diagram, large pivot segments are split into shorter segments. The distance $dist(A,B)$ between points A and B is compared to the user defined value $dist_{max}$. If $dist(A,B)$ is larger, it is split to the set of shorter *pivot segments* $S_i, i = 1,...,m$ of equal lengths, see Figure 2. Each *pivot segment* S_i is represented by its starting point $A(S_i)$ and ending point $B(S_i)$, i.e., for the case when $|C_j| = 0$ there is only one *pivot segment* $S_1 = [A,B]$, when $|C_j| = 1$ there are two segments $S_1 = [A,C_0]$ and $S_2 = [C_0,B]$ and so on. The algorithm finds a path connecting $A, C_0, ..., C_n, B$. Because the connection between every two subsequent points is computed similarly, in the following paragraphs we concentrate only on the computation of the path between A and B ($|C_j| = 0$).

Phase 1: Selection of relevant points

```
Require: set of points $P$, pivot segment $S_i$, user-defined
  distance $r$
  $P_i \leftarrow$ empty
  for $p \in P$ do
    if dist$(p,S_i) \leq r$ or dist$(p,A(S_i)) \leq 2r$ or dist$(p,B(S_i))$
    $\leq 2r$ then
      $P_i \leftarrow p$
    end if
  end for
  return $P_i$
```

Figure 4: Phase 1: Selection of relevant points

The iteration process of the algorithm starts with selecting of relevant atoms (Figure 4). During the computation of a partial path along pivot segment S_i, only the set of relevant atoms P_i is taken into account. Relevant atoms are those having the distance to the segment S_i lower than r or the distance from $A(S_i)$ or $B(S_i)$ lower than $2r$, where $r > 0$ is a user-defined distance (Figure 5). Such a construction assures that the starting or ending points of each segment are always enclosed inside the set P_i. At the same time, according to the current value of parameter r, the massive reduction of input data is achieved. On the other hand, the lower the ratio r, the narrower the resulting pathway could be or it may not even be detected.

Phase 2: Computation and refinement of the Delaunay triangulation

From the set P_i using the Quick Hull algorithm [21], the Delaunay triangulation $T(P_i)$ is computed (Figure 6). In the three dimensional space, the triangulation is formed by tetrahedra. Since we compute the triangulation only for a limited set of atoms P_i, there may be differences between the triangulation computed for this subset and the triangulation

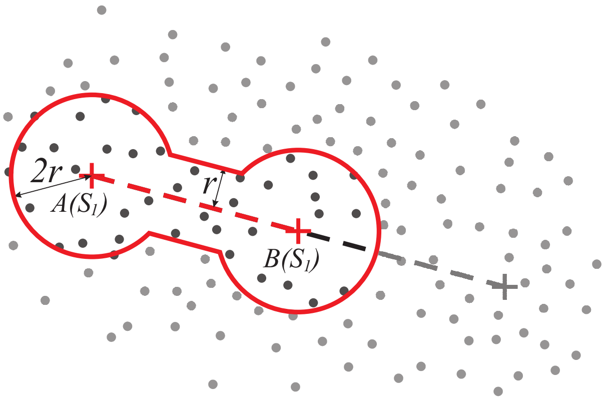


Figure 5: Selection of relevant points P_i.

T computed for the whole protein. The differences occur mainly on the boundaries, where narrow tetrahedra forming the hull appear. These narrow tetrahedra would misrepresent the resulting tunnel. Therefore, the triangulation $T(P_i)$ must be refined to include only tetrahedra that are also contained in $T(P)$. All boundary tetrahedra from $T(P_i)$ must be checked whether they satisfy the Delaunay condition in the context of the whole protein. I.e., for a tangent sphere constructed for every triple from P_i no other point from P is allowed to lie inside such a sphere (Figure 7). If a point lies inside the tangent sphere, the old tetrahedron has to be replaced by two newly constructed tetrahedra. The refinement is processed incrementally and to prevent the worst case – traversing all points from P – we stop it when there is no unchecked atom within the distance lower than $3r$ from the segment S_i. The rest of tetrahedra that do not fulfill the Delaunay condition are simply removed from the triangulation.

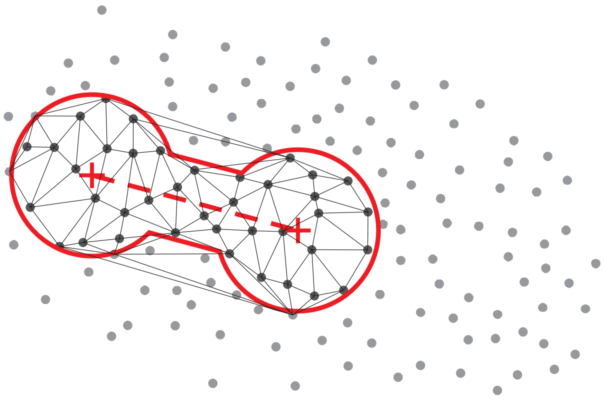


Figure 6: Local triangulation of $T(P_i)$.

Phase 3: Conversion to graph and evaluation

When the triangulation $T(P_i)$ is computed, it is converted into a graph $G(P_i)$ in the same way as described in [20]. The duality between the Delaunay triangulation (DT) and the Voronoi diagram (VD) is employed, i.e., every tetrahedron represents a Voronoi vertex and every shared face of two tetrahedra represents a Voronoi edge. The graph $G(P_i)$ is then constructed in the following steps. Every node N in graph $G(P_i)$ represents a Voronoi vertex from triangulation $T(P_i)$ and stores a reference to its dual tetrahedron $tetra(N)$. For each two nodes whose referenced tetrahedra share a face, a new edge e_k is added into the graph $G(P_i)$. In the last step, the weight $w(e_k)$ of every edge e_k is computed as

$$w(e_k) = \frac{length(e_k)}{dist(e_k)^3} \tag{1}$$

where $length(e_k)$ is the geometrical length of the edge e_k and $dist(e_k)$ is the distance from e_k to the surface of the nearest atom from P_i. The power of 3 was determined experimentally and it means that shorter and wider edges in graph are preferred to longer and narrower ones.

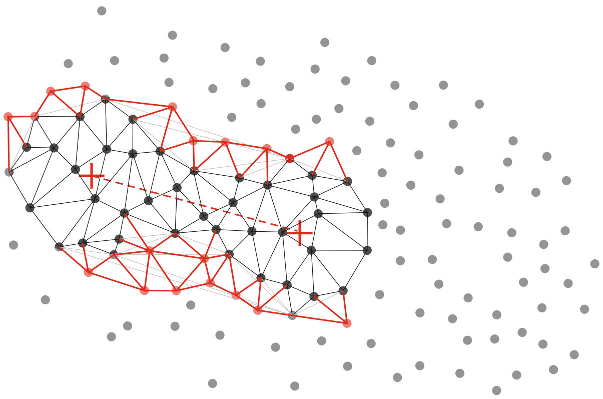


Figure 7: Local triangulation of P_i refinement – every tetrahedron has to satisfy the Delaunay condition.

Phase 4: Traversing the graph

Before the start of the searching algorithm, the starting nodes $A(N)$ and ending nodes $B(N)$ from $G(P_i)$ have to be determined. Firstly the nodes $A(N)$ or $B(N)$ whose referenced tetrahedra are enclosing the starting point $A(S_i)$ or $B(S_i)$ are selected. However, the weight $w(e)$ of all edges originating from the node $A(N)$ or $B(N)$ may be low, i.e., causing the starting or ending site inaccessible for the probe of a selected size. To prevent these cases, in places where the path is completely blocked at the starting or ending node, nodes $A(N)$ and $B(N)$ are selected (optimized) in the following way. Let Y be a set of nodes whose referenced tetrahedra $tetra(N)$ are accessible ($tetra(N)$ and $tetra(A(N))$ share a face) from $tetra(A(N))$ at most in 3 hops. From the set Y a node, from which the edge e_k with the greatest $w(e_k)$ originates, is selected as $A(N)$. The ending node $B(N)$ is determined similarly.

The optimization is not performed, when $A(S_i) \in \{C_0, ..., C_n\}$ or $B(S_i) \in \{C_0, ..., C_n\}$. In other words, $A(S_i)$ or $B(S_i)$ can be optimized only if it was created when splitting the long pivot segments by $createSegments$ method and not set by the user. This ensures that if the path is found it always leads through the user specified points of interest. However, this restriction may be loosen to allow more flexible path searching.

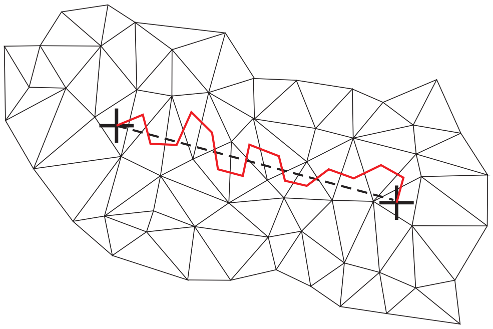


Figure 8: Resulting pathway in segment S_1.

After the nodes $A(N)$ and $B(N)$ are known, the A* search algorithm [25] is launched. As the result, the best weighted path $path(S_i)$ of segment S_i is obtained (Figure 8).

Construction of the resulting path

After all the segments S_i are processed, a set of paths is obtained. As the last step, all paths $path(S_i)$ have to be concatenated in order to form one continuous path connecting A with B. Since the ending point $B(S_i)$ of S_i is equal to the starting point $A(S_{i+1})$ of S_{i+1}, the connection process is straightforward. The pseudocode (Figure 9) presents the overview of the whole algorithm.

```
Require: set of 3D points $P$, points $A$ and $B$
  $S \leftarrow createSegments(A, B)$
  $T \leftarrow$ empty list of tunnels
  for $S_i \in S$ do
    $P_i \leftarrow selectRelevantPoints(P, S_i)$
    $T(P_i) \leftarrow$ Delaunay triangulation of $P_i$
    adjust $A(S_i)$ and $B(S_i)$
    refine $T(P_i)$
    $G(P_i) \leftarrow$ convert $T(P_i)$
    evaluate all edges in $G(P_i)$
    $A(N) \leftarrow$ select starting node from $G(P_i)$
    $B(N) \leftarrow$ select ending node from $G(P_i)$
    $path(S_i) \leftarrow$ search algorithm($G(P_i)$, $A(N)$, $B(N)$)
    $T \leftarrow path(S_i)$
  end for
  $path(A, B) \leftarrow concat$(T)
  return $path(A, B)$
```

Figure 9: Tunnel connecting two points of interest

IV. Results and Discussion

In the testing and verification phase, this algorithm was launched on protein structures of various sizes and inner arrangement. We will mention only several examples: epoxide hydrolase with PDB ID 1CQZ, acetylcholinesterase complexed with huperzine A (PDB ID 1VOT), haloalkane dehalogenase (PDB ID 2HAD), GroEL-GroES-(ADP)7 chaperonin complex (PDB ID 1AON) and ribosome indexed as 70S_RF2. The binding sites for the input were taken from the CSA database [1] or, when not present, selected by biochemists according to their prior knowledge or by analyzing inner cavities (coming out from the fact that each binding site is enclosed in an inner cavity).

In each case, the algorithm was able to compute a suitable pathway between selected binding sites. Computation of pathways connecting three active sites of 1CQZ (selected from the CSA database) can be seen in Figure 10 (left). Figure 10 (right) shows the capability of our solution to detect intramolecular tunnels in large ribosomal structure 70S_RF2. This particular case presents four manually selected points as an input.

The algorithm was implemented in Java as a single-thread process. On a common 2.5GHz computer with 32-bit operating system the running time for 1CQZ (8218 atoms) was approximately 8 seconds whereas for 70S_RF2 (149412 atoms) was 45 seconds. This measurement was done with experimental settings of parameters ($dist_{max} = 50$Å, $r = 8$Å).

In the worst case, finding a path among two points on opposite sides of the structure and with the specific shape of the structure, the algorithm has to process all atoms in one step. But, for the computation of paths in Figure 10 (left) only approx. 32% and for 70S_RF2 only approx. 18% of all atoms were taken into account. The algorithm starts with the selection of relevant atoms for the current segment which is done in $\mathcal{O}(n)$. Then, the Delaunay triangulation using the QuickHull algorithm is obtained in worst in $\mathcal{O}(n^2)$ ($\mathcal{O}(n \cdot log(n)$ expected [21]). Converting DT to a graph and evaluation of all edges is done in $\mathcal{O}(n)$. The last step is the A* search algorithm which finds the path in worst in $\mathcal{O}(n^2)$ (better complexity expected since it is heavily dependent on the heuristic function used).

V. Conclusion

In this paper, we presented a novel algorithm, which is able to determine non-colliding pathways among a set of spheres randomly scattered in the 3D space. This algorithm can be applied for solving a general path planning problem and it is based on the computational geometry. To demonstrate its application, we adopted this solution to detect intramolecular tunnels in protein macromolecules. The curvature of the detected tunnels and the preciseness of the algorithm is dependent on initial settings. By changing the input parameters biochemists are able to reach the desired pathways. Our solution is equipped with simple and user friendly interface where users can change the input parameters and thus influence the size, shape and curvature of resulting tunnel. The algorithm is not limited to the size of the input data set because only a subset of necessary atoms is required in every phase of computation. Paths computed by our algorithm can also reveal potentially interesting places of protein structures that lie on the path between two or more active sites.

VI. Future work

Our future research in this area will concentrate on an extension of this algorithm to process molecular dynamics, computation of multiple tunnels between more binding sites and more automatized settings of initial parameters. Moreover, our aim is also to decrease the computational time of the algorithm. This can be reached by parallelization of the algorithm. In fact, because all path segments are identified once before the main computation starts, we can process each segment in different processing unit independently.

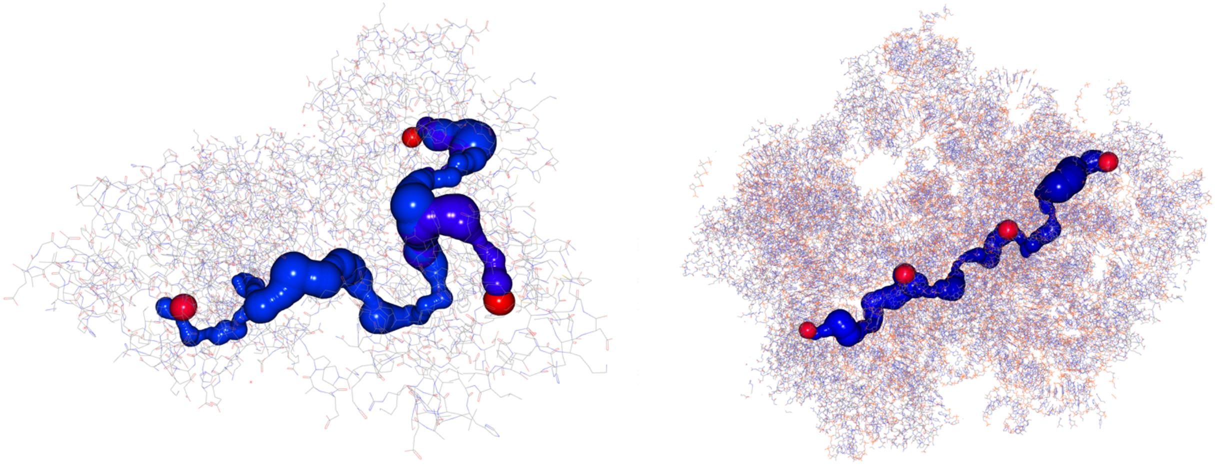


Figure 10: Left - path among 3 known binding sites in the 1CQZ (8218 atoms) protein structure. Right - path among 4 custom sites in 70S_RF2 structure (149412 atoms).

References

[1] N. Furnham, G. L. Holliday, T. A. P. de Beer, J. O. B. Jacobsen, W. R. Pearson, and J. M. Thornton, "The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes," *Nucleic Acids Research*, vol. 42, no. D1, pp. D485–D489, 2014. [Online]. Available: http://nar.oxfordjournals.org/content/42/D1/D485.abstract [retrieved: 03, 2014]

[2] S. Lutz and U. Bornscheuer, *Protein Engineering Handbook*, ser. Protein Engineering Handbook. Wiley, 2012, no. sv. 3.

[3] E. W. Miles, S. Rhee, and D. R. Davies, "The molecular basis of substrate channeling," *Journal of Biological Chemistry*, vol. 274, no. 18, pp. 12 193–12 196, 1999.

[4] C. C. Hyde, S. A. Ahmed, E. A. Padlan, E. W. Miles, and D. R. Davies, "Three-dimensional structure of the tryptophan synthase alpha 2 beta 2 multienzyme complex from Salmonella typhimurium," *J. Biol. Chem.*, vol. 263, no. 33, pp. 17 857–17 871, Nov 1988.

[5] A. Gora, J. Brezovsky, and J. Damborsky, "Gates of enzymes," *Chemical Reviews*, vol. 113, no. 8, pp. 5871–5923, 2013. [Online]. Available: http://pubs.acs.org/doi/abs/10.1021/cr300384w

[6] F. M. Raushel, J. B. Thoden, and H. M. Holden, "Enzymes with molecular tunnels," *Accounts of Chemical Research*, vol. 36, no. 7, pp. 539–548, 2003.

[7] H. Edelsbrunner and E. P. Mucke, "Three-dimensional alpha shapes," *ACM Trans. Graph.*, vol. 13, no. 1, pp. 43–72, 1994.

[8] M. F. Sanner, A. J. Olson, and J. C. Spehner, "Reduced surface: an efficient way to compute molecular surfaces," *Biopolymers*, vol. 38, no. 3, pp. 305–320, Mar 1996.

[9] M. Krone, M. Falk, S. Rehm, J. Pleiss, and T. Ertl, "Interactive exploration of protein cavities," in *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, ser. EuroVis'11. Aire-la-Ville, Switzerland, Switzerland: Eurographics Association, 2011, pp. 673–682. [Online]. Available: http://dx.doi.org/10.1111/j.1467-8659.2011.01916.x [retrieved: 03, 2014]

[10] K. Ghosh, "A solution of polygon containment, spatial planning, and other related problems using minkowski operations," *Comput. Vision Graph. Image Process.*, vol. 49, no. 1, pp. 1–35, 1990.

[11] J. Janet, R. Luo, and M. Kay, "The essential visibility graph: an approach to global motion planning for autonomous mobile robots," *Robotics and Automation, 1995. Proceedings., 1995 IEEE International Conference on*, vol. 2, pp. 1958 –1963 vol.2, may. 1995.

[12] F. Lingelbach, "Path planning using probabilistic cell decomposition," *Robotics and Automation, 2004. Proceedings. ICRA '04. 2004 IEEE International Conference on*, vol. 1, pp. 467 – 472 Vol.1, apr. 2004.

[13] K. Fujimura and H. Samet, "A hierarchical strategy for path planning among moving obstacles [mobile robot]," *Robotics and Automation, IEEE Transactions on*, vol. 5, no. 1, pp. 61 –69, feb. 1989.

[14] P. Fiorini and Z. Shillert, "Motion planning in dynamic environments using velocity obstacles," *International Journal of Robotics Research*, vol. 17, pp. 760–772, 1998.

[15] Y. Kwon, J. Cho, S. Kwon, and J. Joh, "Collision avoidance of moving obstacles for underwater robots," *Journal of Systemics, Cybernetics and Informatics*, vol. 4, no. 5, pp. 86–91, 2006.

[16] P. Beneš, P. Medek, O. Strnad, and J. Sochor, "Computation of dynamic channels in proteins," in *Proceedings of Biotechno*, U. o. S. Pei-Yuan Qian, KAUST and H. K. Technology, Eds. Venice/Mestre, Italy: Neuveden, 2011, pp. 78–83.

[17] P. Beneš, O. Strnad, and J. Sochor, "New path planning method for computation of constrained dynamic channels in proteins," *WSCG Full papers proceedings*, pp. 81–88, 2011.

[18] M. Petřek et al., "Caver: A new tool to explore routes from protein clefts, pockets and cavities," *BMC Bioinformatics*, vol. 7, p. 316, 2006.

[19] G. Voronoi, "Nouvelles applications des parametres continus a la theorie des formes quadratiques. duesieme memoire: Recherches sur les parallelodères primitifs," *J. Reine Angew. Math*, vol. 134, p. 198287, 1908.

[20] 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.

[21] C. Barber, C. Bradford, D. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, 1996.

[22] M. Petřek, P. Košinová, J. Koča, and M. Otyepka, "Mole: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels," *Structure*, vol. 15, pp. 1357–1363, 2007.

[23] E. Chovancová et al., "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 [retrieved: 03, 2014]

[24] N. Lindow, D. Baum, and H. Hege, "Voronoi-based extraction and visualization of molecular paths," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 12, pp. 2025–2034, Dec. 2011. [Online]. Available: http://dx.doi.org/10.1109/TVCG.2011.259 [retrieved: 03, 2014]

[25] P. E. Hart, N. J. Nilsson, and B. Raphael, "A formal basis for the heuristic determination of minimum cost paths," *IEEE Transactions on Systems, Science, and Cybernetics*, vol. SSC-4, no. 2, pp. 100–107, 1968.

Improving Protein Sub-cellular Localization Prediction Through Semi-supervised Learning

Jorge Alberto Jaramillo-Garzón
Instituto Tecnológico Metropolitano
Medellín, Colombia
Email: jorgejaramillo@itm.edu.co

César Germán Castellanos-Domínguez
Universidad Nacional de Colombia
Manizales, Colombia
Email: cgcastellanosd@unal.edu.co

***Abstract*—Prediction of sub-cellular localization of proteins is a fundamental task in bioinformatics, since it can provide useful information to determine its function. Several prediction techniques have been proposed in the recent years and methods based on machine learning techniques have achieved state of the art classification, usually employing support vector machines and neural networks. However, those methods need high amounts of labeled samples (proteins with known function) in order to train accurate classifiers, and such information is not easily available for this task. In this paper, an alternative methodology that uses semi-supervised learning is proposed. This type of machine learning allows to use unlabeled samples (which are easily available) in order to improve the estimation of the classifiers. All the needed steps for using semi-supervised learning in the problem of predicting protein sub-cellular localizations are described in detail and the methodology is compared with the standard supervised alternative. The results show that using semi-supervised learning significantly improves the prediction performance of the classifier in several cases, proving to be a valuable tool in bioinformatics.**

Keywords-Sub-cellular localization, Gene Ontology, Semi-supervised, Support Vector Machines.

I. INTRODUCTION

One of the most important tasks in modern bioinformatics is to provide reliable functional annotations for gene products. Predicting protein sub-cellular localizations allows researchers to obtain useful information for revealing protein functions and helping to understand the pathways that regulate biological processes [1]. The localization of specific proteins can be experimentally determined by assays of expression of green fluorescent proteins in order to monitor its intrinsic fluorescence and subsequently locate it in the cell [2]. However, such procedures become expensive and highly time consuming when they have to be applied in high-throughput projects, which yields to the need of developing computational predictors able to identify the sub-cellular location of novel proteins based on its sequence information alone [3].

Several predictors have been proposed in the recent years (for full surveys, see [4, 5, 6]). In particular, most recent methods have used machine learning techniques trained over feature spaces of physical-chemical, statistical or locally-based attributes. Those methods employ techniques such as neural networks (ProtFun [7]), Bayesian multi-label classifiers 8]) and support vector machines (SVM-Prot [9], GOKey [10], PoGO 11]), obtaining high performance results in their own respective databases, mostly composed by model organisms such as bacteria and a few high order species [12].

One of the main limitations of machine learning methods, however, is that they need relatively high amounts of training data in order to learn reliable classification models. Such training data refers to “labeled instances”, that is, enough protein sequences which function must be already known. It is a known fact, however, that only a small number of proteins have actually been annotated for certain functions [4]. Under such circumstances, semi-supervised learning methods provide an alternative approach to protein annotation. In semi-supervised learning methods, additionally to labeled data, the algorithm is provided with an amount of unlabeled data that can be used to improve the estimations of the classifier.

This work presents an implementation of semi-supervised learning using semi-supervised support vector machines (S^3VM) for predicting protein sub-cellular localizations. The results obtained with this approach show that using semi-supervised learning significantly improves the prediction performance of the standard support vector machine (SVM) in several cases, proving to be a valuable tool in bioinformatics.

The following section describes the theoretical background about SVM and S^3VM. Next, the “Experimental setup” section describes the database and all the components of the proposed methodology. The final two sections present the results and conclusions, respectively.

II. THEORETICAL BACKGROUND

A. Support vector machines

Support vector machines (SVM) are powerful tools for solving classification problems, designed over a strong theoretical background based on the idea of minimizing the structural risk [13]. For a non-linear SVM, the objective is to find a classification function of the form:

$$f_{(\boldsymbol{w},b)}(\boldsymbol{x})=\langle\boldsymbol{w},\boldsymbol{x}\rangle+b \qquad (1)$$

ISBN: 978-1-61208-335-3

where $\langle\cdot,\cdot\rangle$ represents the dot product. A vector of parameters can be defined as $\boldsymbol{\theta} = [\boldsymbol{w},b]$, and the optimization problem can be stated as follows:

$$\boldsymbol{\theta}^{*}=\arg\min_{\boldsymbol{\theta}\in\mathcal{T}}\left\{\frac{1}{2}\|\boldsymbol{\theta}\|^{2}+C\sum_{i=1}^{L}\ell\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right)\right\} \qquad (2)$$

where $\ell(t)=\max(0,1-t)$ is the hinge loss function and C is a trade-off parameter regulating the complexity of the model. For the non-linear case, the data are first mapped in a high dimensional Hilbert space $\mathcal{H}$ through a mapping $\Phi:\mathcal{X}\mapsto\mathcal{H}$, and then a linear decision boundary is constructed in that space. The mapping Φ can be explicitly computed or only implicitly through the use of a kernel function K such that $K(\boldsymbol{x}_1,\boldsymbol{x}_2)=\langle\Phi(\boldsymbol{x}_1),\Phi(\boldsymbol{x}_2)\rangle$. The Representer Theorem can be used to show that the solution function has the form:

$$f_{\boldsymbol{\theta}^{*}}(\boldsymbol{x})=\sum_{i=1}^{L}\alpha_i K(\boldsymbol{x},\boldsymbol{x}_i) \qquad (3)$$

where the coefficients α_i can be found with a conventional quadratic optimization algorithm. The Gaussian kernel is the most commonly used because of its attractive features such as structure preservation [14]. This kernel is computed by:

$$K(\boldsymbol{x}_1,\boldsymbol{x}_2)=e^{-\frac{\|\boldsymbol{x}_1-\boldsymbol{x}_2\|_2}{2\sigma}} \qquad (4)$$

where σ is the dispersion parameter that must be properly chosen by the user. In this work, the SVM is trained with the '*kernlab*' package, available in R-CRAN [15].

B. Semi-supervised support vector machines

Semi-Supervised SVMs (S^3VMs) emerged as an extension to standard SVMs for semi-supervised learning. S^3VMs find a labeling for all the unlabeled data, and a separating hyperplane, such that maximum margin is achieved on both the labeled data and the (now labeled) unlabeled data. As a result, unlabeled data guides the decision boundary away from dense regions. The assumption of S^3VMs is that the classes are well-separated, such that the decision boundary falls into a low density region in the feature space, and does not cut through dense unlabeled data [16, chapter 6].

In a similar way than the conventional SVMs, the optimization problem for an S^3VMs can be stated as follows:

$$\boldsymbol{\theta}^{*}=\arg\min_{\boldsymbol{\theta}\in\mathcal{T}}\left\{\frac{1}{2}\|\boldsymbol{\theta}\|^{2}+C\sum_{i=1}^{L}\ell\left(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)y_i\right)+\ldots\right.$$
$$\left.+\ \lambda\sum_{i=L+1}^{L+U}\ell\left(|f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)|\right)\right\} \ (5)$$

where $\ell(t)=\max(0,1-t)$ is the hinge loss function, C is the trade-off parameter and λ is a new regularization parameter. The first two terms in the above equation correspond to the traditional solution for the standard supervised SVM shown in equation (2), while the last term puts $f_{\boldsymbol{\theta}}(\boldsymbol{x}_i)$ of the unlabeled points $\boldsymbol{x}_i$ away from 0 (thereby implementing the low density assumption) [17].

Again, as in the supervised case, the kernel trick can be used for constructing non-linear S^3VMs. While the optimization in SVM is convex and can be solved with QP-hard complexity, optimization in S^3VM is a non-convex combinatorial task with NP-Hard complexity. Most of the recent work in S^3VM has been focused on the optimization procedure (a full survey in this matter can be found in [18]). Among the proposed methods for solving the non-convex optimization problem associated with S^3VMs, one of the first implementations is the S^3VM^{light} by Joachims [19], which is based on local combinatorial search guided by a label switching procedure. Chapelle et. al. [20] presented a method based on gradient descent on the primal, that performs significantly better than the optimization strategy pursued in S^3VM^{light}; the work by Chapelle et. al. [17] proposes the use of a global optimization technique known as "continuation", often leading to lower test errors than other optimization algorithms; Collobert et. al. [21] uses the Concave-Convex procedure, providing a highly scalable algorithm in the nonlinear case.

III. Experimental Setup

A. Database

This work uses the database designed in [12]. Such database comprises all the available *Embryophyta* proteins at UniProtKB/Swiss-Prot database [22], with at least one annotation in the Gene Ontology Annotation (GOA) project [23]. In order to avoid the presence of protein families that could bias the results, the dataset was filtered at an identity cutoff of 30%.

The main set comprises a total of 2210 sequences associated to 20 different sub-cellular localizations. Those localizations correspond to the Cellular component ontology defined by the plants GO slim [24]. Categories with less than 30 proteins were discarded because they did not have enough samples to train a statistically reliable classifier. All the available *Embryophyta* proteins at UniProtKB/Swiss-Prot database that has no entries in the GOA project were added as the core set of unlabeled instances. Proteins associated to the nodes in the functional path of each GO term were also left as unlabeled instances regarding that classifier. Finally, 22000 unlabeled instances were randomly chosen in order to accomplish an approximate relation of ten unlabeled instances per each labeled one.

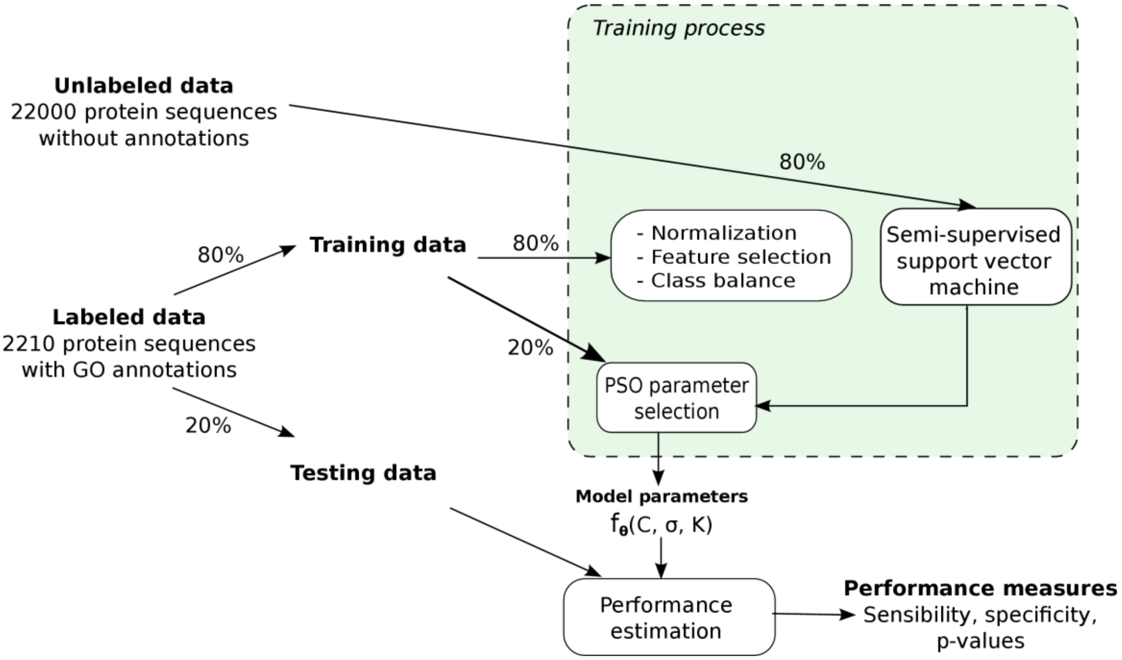


Figure 1: Main methodology

B. *Classification Methodology*

Figure 1 shows the main methodology for classification. The CCP - S^{3}VM [21] was used as base classifier, with the Gaussian kernel. All the parameters of the algorithm, including the dispersion of the kernels, the trade-off parameters of the SVMs, the regularization constants were tuned with a particle swarm optimization meta-heuristic [25].

In order to allow samples to be associated to multiple categories, decision making was implemented following the one-against-all strategy. The method produced a strong class imbalance that was tackled using the Synthetic Minority Over-sampling Technique (SMOTE) [26].

Feature selection was carried out before trying to induce any decision rule (classifier) because, having a limited number of training examples, excessive features would possibly overfit the training data. For this purpose, the *Fast Correlation-Based Filter* presented in [27] was used.

In order to estimate the performance of the predictive model, a 5-fold cross-validation strategy is implemented. In such strategy, the test procedure is repeated five times, and each time an 80% of the data is used for adjusting the SVM parameters and training the model, while the remaining 20% is used as testing samples.

IV. Results

In order to analyze the results obtained with the proposed methodology, sensitivity and specificity for each GO term were computed. The obtained results are compared with the ones obtained in [12] with the commonly used BLASTp method (Figure 2), as well as with a standard SVM (Figure 3). Bars in the left plots show sensitivity and specificity of the Lap-S^{3}VM and lines depict geometric mean for S^{3}VM (orange), BLASTp (blue) and the classical supervised SVM (green). Right plots depict the p-values obtained by paired t-tests at a 95% significance level. Orange bars show the cases when the S^{3}VM significantly outperforms BLASTp and the supervised SVM, in Figures 2 and 3, respectively. On both figures, the best predicted categories are ordered from top to bottom.

While the comparison with BLASTp provides information about the applicability of the methodology compared with the alignment based methods, the main purpose of this comparing the SVM with the S3VM is to verify whether or not the inclusion of the additional cluster-based semi-supervised term in the training of the SVM improves the performance of the system. This can be understood as the accomplishment of the cluster assumption when the unlabeled data is incorporated to the training process.

Figure 2 shows that there are only two cellular components for which there is no statistically significant difference between BLASTp and the S3VM: *Perixosome* and *Endosome*. For all the remaining eighteen cellular components, the semi-supervised method obtained statistically significant superior performance.

Regarding Figure 3, it can be observed that eight cellular components were significantly improved, while another two (*Mitochondria* and *Cytoplasm**) also reached high p-values over 0.9.

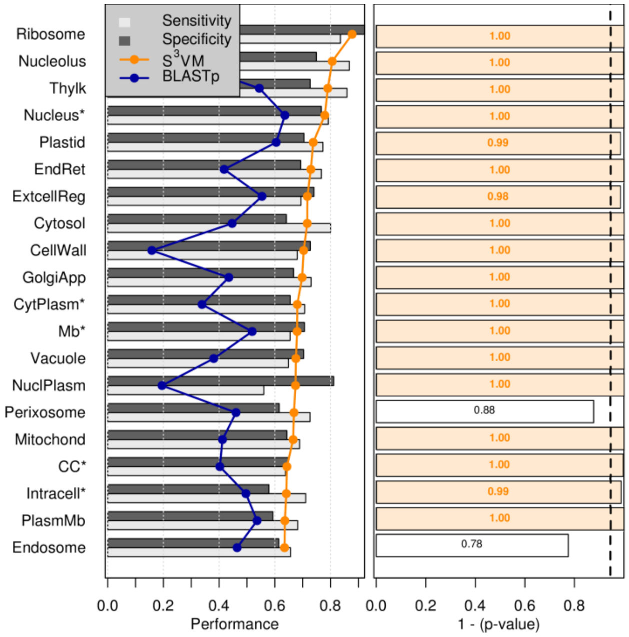


Figure 2: Comparison between the S3VM method and BLASTp

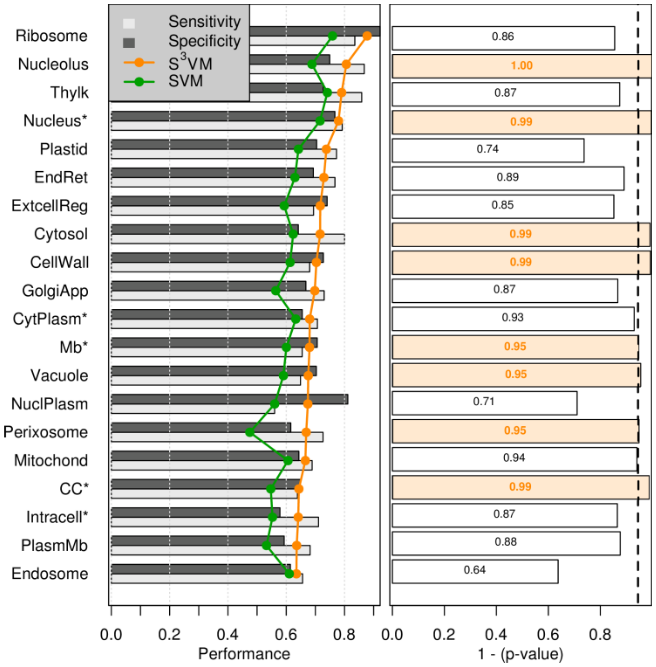


Figure 3: Comparison between the S3VM method and the supervised SVM

This results show that the inclusion of the additional information improves the estimation of the classification models and thus, provides an efficient way for alleviating the lack of labeled data in the field of bioinformatics. Although several localizations were not improved over a statistically significant threshold, no one of them degraded its performance with the inclusion of the additional data, proving to accomplish the underlying assumptions of semi-supervised learning.

Finally, in order to verify the influence of the number of unlabeled instances included in the training, process, several experiments were performed, varying the number of unlabeled instances from 0 to 2200. As exemplary cases, the results for six cellular components (*nucleus*, *cell wall*, *vacuole*, *cytosol*, *membrane* and the root node of the ontology, *cellular component*) are depicted on Figure 4. It is important to point out that these tests where done with the same SVM parameters across all the experiment and, consequently, the predictor is not optimized for each case. However it allows understanding the main influence of the unlabeled data.

From these results, it can be observed that the effect of progressively including unlabeled instances is reducing the specificity of the classifier, while increasing the sensitivity. In general terms, when no unlabeled instances are included, specificity is very high and sensitivity is almost zero. This means that the classifier is rejecting all the samples for that given GO term. The semi-supervised assumption allows the system to recognize the positive samples, thus increasing the overall performance of the predictor.

V. Conclusion

This work presented an experimental analysis of the suitability of semi-supervised methods for the prediction of protein sub-cellular localizations. The results show that semi-supervised learning applied to the prediction of GO terms, significantly outperforms the supervised learning approach in several cases. As future work another semi-supervised strategies must be explored in order to analyze if different assumptions (for example, graph-based methods) can be able to provide better results for the cases where this methodology was not significantly superior.

Acknowledgment

This work is within the framework of the Dirección de Investigaciones de Manizales (DIMA) of the Universidad Nacional de Colombia and the Centro de Investigación of the Instituto Tecnológico Metropolitano. The work has been partially founded by Colciencias grant 111952128388.

References

[1] K. Chou, H. Shen, and E. Newbigin, “Plant-mPLoc: A Top-Down Strategy to Augment the Power for Predicting Plant Protein Subcellular Localization,” PloS one, vol. 5, no. 6, 2010, pp. 259–270.

[2] P. Baldi and S. Brunak, Bioinformatics: the machine learning approach. 1em plus 0.5em minus 0.4emThe MIT Press, 2001.

[3] K. Chou and H. Shen, “Recent progress in protein subcellular location prediction,” Analytical Biochemistry, vol. 370, no. 1, 2007, pp. 1–16.

[4] X. Zhao, L. Chen, and K. Aihara, “Protein function prediction with high-throughput data,” Amino Acids, vol. 35, no. 3, 2008, pp. 517–530.

[5] G. Pandey, V. Kumar, and M. Steinbach, “Computational approaches for protein function prediction: A survey,” Department of Computer

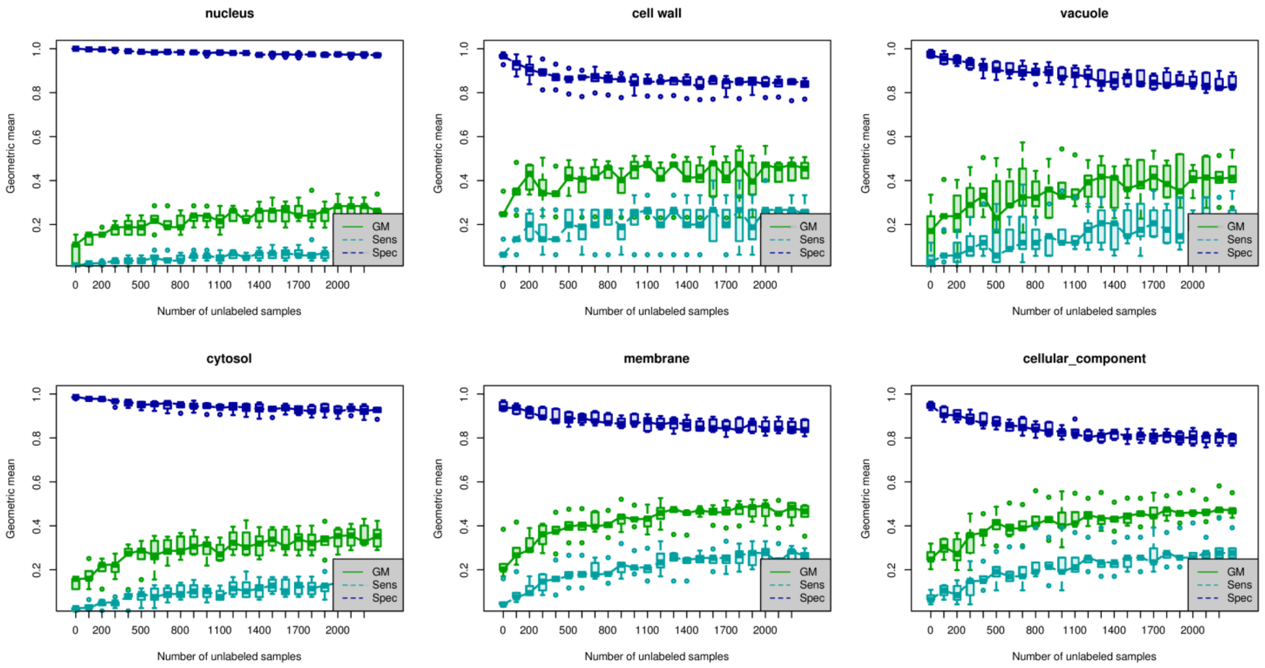


Figure 4: Variation of the number of unlabeled samples included in the training process

Science and Engineering, University of Minnesota, Twin Cities, Tech. Rep. 06-028, 2006.

[6] I. Friedberg, “Automated protein function prediction–the genomic challenge,” Briefings in Bioinformatics, vol. 7, no. 3, 2006, p. 225.

[7] L. Jensen, R. Gupta, H. Staerfeldt, and S. Brunak, “Prediction of human protein function according to Gene Ontology categories,” pp. 635–642, 2003.

[8] J. Jung and M. R. Thon, “Gene function prediction using protein domain probability and hierarchical Gene Ontology information,” 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4.

[9] C. Z. Cai, “SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence,” Nucleic Acids Research, vol. 31, no. 13, 2003, pp. 3692–3697.

[10] R. Bi, Y. Zhou, F. Lu, and W. Wang, “Predicting Gene Ontology functions based on support vector machines and statistical significance estimation,” Neurocomputing, vol. 70, no. 4-6, 2007, pp. 718–725.

[11] J. Jung, G. Yi, S. a. Sukno, and M. R. Thon, “PoGO: Prediction of Gene Ontology terms for fungal proteins.” BMC bioinformatics, vol. 11, 2010, p. 215.

[12] J. A. Jaramillo-Garzón, J. J. Gallardo-Chacón, C. G. Castellanos-Domínguez, and A. Perera-Lluna, “Predictability of gene ontology slim-terms from primary structure information in embryophyta plant proteins,” BMC bioinformatics, vol. 14, no. 1, 2013, p. 68.

[13] V. Vapnik, Statistical learning theory. plus 0.5em minus 0.4emWiley New York, 1998.

[14] Z. Liu, M. J. Zuo, and H. Xu, “Parameter selection for Gaussian radial basis function in support vector machine classification,” 2012 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, Jun. 2012, pp. 576–581.

[15] A. Karatzoglou, A. Smola, K. Hornik, and A. Zeileis, “kernlab – an S4 package for kernel methods in R,” Journal of Statistical Software, vol. 11, no. 9, 2004, pp. 1–20.

[16] X. Zhu and A. B. Goldberg, “Introduction to semi-supervised learning,” Synthesis lectures on artificial intelligence and machine learning, vol. 3, no. 1, 2009, pp. 1–130.

[17] O. Chapelle, M. Chi, and A. Zien, “A continuation method for semi-supervised SVMs,” Proceedings of the 23rd international conference on Machine learning - ICML ’06, 2006, pp. 185–192.

[18] O. Chapelle, V. Sindhwani, and S. S. Keerthi, “Optimization techniques for semi-supervised support vector machines,” The Journal of Machine Learning Research, vol. 9, 2008, pp. 203–233.

[19] T. Joachims, “Transductive inference for text classification using support vector machines,” in ICML, vol. 99, 1999, pp. 200–209.

[20] O. Chapelle and A. Zien, “Semi-supervised classification by low density separation,” Proceedings of the tenth international workshop on, 2005.

[21] R. Collobert, F. Sinz, J. Weston, and L. Bottou, “Large scale transductive SVMs,” The Journal of Machine …, vol. 1, 2006, pp. 1687–1712.

[22] E. Jain, A. Bairoch, S. Duvaud, I. Phan, N. Redaschi, B. Suzek, M. Martin, P. McGarvey, and E. Gasteiger, “Infrastructure for the life sciences: design and implementation of the UniProt website,” BMC bioinformatics, vol. 10, no. 1, 2009, p. 136.

[23] D. Barrell, E. Dimmer, R. Huntley, D. Binns, C. O’Donovan, and R. Apweiler, “The GOA database in 2009–an integrated Gene Ontology Annotation resource,” Nucleic Acids Research, 2008.

[24] T. Berardini, S. Mundodi, L. Reiser, E. Huala, M. Garcia-Hernandez, P. Zhang, L. Mueller, J. Yoon, A. Doyle, G. Lander et al., “Functional annotation of the Arabidopsis genome using controlled vocabularies,” Plant Physiology, vol. 135, no. 2, 2004, p. 745.

[25] J. Kennedy and R. Eberhart, “Particle swarm optimization,” Proceedings of ICNN’95 - International Conference on Neural Networks, vol. 4, 1995, pp. 1942–1948.

[26] N. Chawla, K. Bowyer, L. Hall, and W. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” Journal of Artificial Intelligence Research, vol. 16, no. 3, 2002, pp. 321–357.

[27] L. Yu and H. Liu, “Efficient feature selection via analysis of relevance and redundancy,” The Journal of Machine Learning Research, vol. 5, 2004, pp. 1205–1224.