BIOTECHNO 2016

The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-488-6

BIOCOMPUTATION 2016

The International Symposium on Big Data and BioComputation

June 26 - 30, 2016

Lisbon, Portugal

BIOTECHNO 2016 Editors

Pascal Lorenz, Université de Haute Alsace, France

Steffen G. Scholz, Karlsruhe Institute of Technology (KIT), Germany

BIOTECHNO 2016

Foreword

The Eighth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2016), held between June 26 - 30, 2016 - Lisbon, Portugal, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

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

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

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

The BIOTECHNO 2016 conference also featured the following symposium:

- BIOCOMPUTATION 2016 - The International Symposium on Big Data and BioComputation
-

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2016 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 2016. 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 2016 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2016 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 also hope that Lisbon provided a pleasant environment during the conference and everyone saved some time for exploring this beautiful city.

BIOTECHNO 2016 Chairs:

Stephen Anthony, The University of New South Wales, Australia
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada

BIOTECHNO 2016

Committee

BIOTECHNO Advisory Committee

Stephen Anthony, The University of New South Wales, Australia
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada

BIOTECHNO Industrial/Research Chairs

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

BIOTECHNO 2016 Technical Program Committee

Basim Alhadidi, Albalqa' Applied University - Salt, Jordan
Hesham H. Ali, University of Nebraska at Omaha, USA
Jens Allmer, Izmir Institute of Technology, Turkey
Stephen Anthony, The University of New South Wales, Australia
Sansanee Auephanwiriyakul, Chiang Mai University, Thailand
Ganesharam Balagopal, Ontario Ministry of the Environment - Toronto, Canada
Siegfried Benkner, University of Vienna, Austria
Gilles Bernot, University of Nice Sophia Antipolis, France
Tom Bersano, University of Michigan, USA
Christian Blum, IKERBASQUE, Basque Foundation for Science - Bilbao, Spain
Razvan Bocu, University of Brasov, Romania
Magnus Bordewich, Durham University, UK
Sabin-Corneliu Buraga, "A. I. Cuza" University - Iasi, Romania
Eduardo Campos dos Santos, Universidade Federal de Minas Gerais (UFMG), Brazil
Yang Cao, Virginia Tech – Blacksburg, USA
Cesar German Castellanos Dominguez, Universidad Nacional de Colombia - Manizales,Colombia
Yili Chen, Monsanto Company - St. Louis, USA
Rolf Drechsler, DFKI Bremen || University of Bremen, Germany
Esmaeil Ebrahimie, University of Adelaide, Australia
Lingke Fan, University Hospitals of Leicester NHS Trust, UK
Jerome Feret, INRIA, France
Alexandru Floares, SAIA Institute, Romania
Xin Gao, KAUST (King Abdullah University of Science and Technology), Saudi Arabia
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

Alejandro Giorgetti, University of Verona, Italy
Paul Gordon, University of Calgary, Canada
Radu Grosu, Vienna University of Technology, Austria
Ivo Grosse, Martin Luther University of Halle-Wittenberg, Germany
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
Sohei Ito, National Fisheries University, Japan
Attila Kertesz-Farkas, University of Washington, USA
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
Panayiotis Kyriacou, City University London, UK
Christina Rose Kyrtsos, Pennsylvania State University - College of Medicine / University of Maryland - Institute for Systems Research, USA
Cedric Lhoussaine, University 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
Allan Orozco Solano, University of Costa Rica, Costa Rica
Qin Ma, South Dakota State University, USA
Roger Mailler, The University of Tulsa, USA
Igor V. Maslov, EvoCo Inc. - Tokyo, Japan
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
Sebastian Munck, VIB Bio Imaging Core & LiMoNe | Center for the Biology of Disease | KU Leuven, Belgium
Victor Palamodov, Tel Aviv University, Israel
Sever Pasca, Politehnica University of Bucharest, Romania
Maria Manuela Pereira de Sousa, University of Beira Interior, Portugal
Horacio Pérez-Sánchez, Universidad Católica San Antonio de Murcia (UCAM), Spain
Leif E. Peterson, Houston Methodist Research Institute, USA
Clara Pizzuti, ICAR-CNR - Rende (Cosenza), Italy
Enrico Pontelli, New Mexico State University, USA
Ravi Radhakrishnan, University of Pennsylvania, USA
Erin (Cricket) Reichenberger, USDA, USA
Robert Reynolds, Wayne State University, USA
Vincent Rodin, University of Brest (UBO), France
J. Cristian Salgado, University of Chile, Chile
Luciano Sanchez, Universidad de Oviedo, Spain
Thomas Schmid, Universität Leipzig, Germany
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
Christine Sinoquet, University of Nantes, France
Zdenek Smékal, Brno University of Technology, Czech Republic
Bin Song, Oracle - Redwood shores, USA
Andrzej Swierniak, Silesian University of Technology, Poland
Sing-Hoi Sze, Texas A&M University, USA
Yoshihiro Taguchi, Chuo University, Japan
Sophia Tsoka, King's College London, UK
Marcel Turcotte, University of Ottawa, Canada
Ugo Vaccaro, Universita` di Salerno, Italy
Chun Wu, Mount Marty College - Yankton, USA
Boting Yang, University of Regina, Canada
Wang Yu-Ping, Tulane University, USA
Nazar Zaki, United Arab Emirate University (UAEU), United Arab Emirate
Alexander Zelikovsky, Georgia State University, USA
Erliang Zeng, University of South Dakota, USA

BIOCOMPUTATION 2016 Advisory Committee

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

BIOCOMPUTATION 2016 Program Committee Members

Hesham H. Ali, University of Nebraska at Omaha, USA
Khalid Belhajjame, Paris Dauphine University, France
Zhiwei Cao, Tongji University, China
John Carlis, University of Minnesota, USA
Juan Miguel Cejuela, Technische Universität München, Germany
Alexey Cheptsov, High Performance Computing Center Stuttgart, Germany
Matthias Chung, Virginia Tech, USA
Raffaele A. Calogero, University of Torino, Italy
Angelo Facchiano, Istituto di Scienze dell'Alimentazione - CNR, Italy
Fabio Fumarola, University of Bari "Aldo Moro", Italy
Mohamed Ghalwash, Temple University, USA
Saman K. Halgamuge, University of Melbourne, Australia
Steffen Heber, North Carolina State University, USA
Uri Hershberg, Drexel University, USA
Sheng-Jun Huang, Nanjing University of Aeronautics and Astronautics, China
Philippe Hupé, Institut Curie, CNRS UMR 144, INSERM U900, Mines ParisTech, France
Sumit Kumar Jha, University of Central Florida, USA
John Karro, Miami University, USA
Daniel Lorenz, Technical University of Darmstadt, Germany

Copyright Information

Table of Contents

Matroska Feature Selection Method for Microarray Data

Shuichi Shinmura
Faculty of Economics, Seikei Univ.
Tokyo, Japan
e-mail: shinmura@econ.seikei.ac.jp

Abstract— **We propose a Matroska feature selection method (Method 2) for microarray datasets (the datasets). We had already established a new theory of the discriminant analysis (Theory) and developed an optimal Linear Discriminant Function (OLDF) named Revised IP-OLDF. This LDF can naturally select features for the datasets. The dataset consists of several small genes subspaces that we call small Matroskas (SMs) and are linearly separable. We confirmed this feature selection of Revised IP-OLDF by Swiss banknote data and Japanese automobile data, also. Therefore, we need not struggle with high-dimension genes space. In this paper, we develop a LINGO program to find all SMs and confirm that the dataset consists of disjoint union of SMs and high-dimension subspace that is not linearly separable. Because it is very easy for us to analyze these SMs that are small samples, we may be able to find new facts of gene analysis. Lasso researchers will have better results compared with our results.**

Keywords- Minimum Number of Misclassifications (MNM); Revised IP-OLDF; SVM; Fisher's LDF; Gene Analysis; Small Matroska (SM); Basic Gene Subspase (BGS); Lasso.

I. INTRODCTION

Fisher [6] [7] developed a Linear Discriminant Function (Fisher's LDF) under Fisher's assumption and established the theory of discriminant analysis. Because Fisher's assumption was too strict for the real data, a Quadratic Discriminant Function (QDF) was developed. In addition to two discriminant functions, logistic regression [4] and a Regulalized Discriminant Analysis (RDA) [9] were proposed as the statistical discriminant functions. These statistical discriminant functions apply for many applications, and statistical software packages became essential tools for the science and industries. On the other hand, it is well known that Mathematical Programming (MP) can define the discriminant models [16]. Linear Programming (LP) sets out Least Absolute Deviation (LAD) discriminant function. Quadratic Programming (QP) defines an L2-norm discriminant function (Least square method). Nonlinear Programming (NLP) defines Lp-norm discriminant functions. Before 1997, there were many papers of MP-based discriminant functions summarized by Stam [57]. We think the first generation research ended in 1997 because these researches lacked examination of real data and comparison with statistical discriminant functions. Vapnik [61] proposed three Support Vector Machines (SVMs) such as a Hard-margin SVM (H-SVM), Soft-margin SVM (S-SVM) and kernel SVM in 1995. H-SVM clearly defined a Linearly Separable Data (LSD) and generalization ability. However, because most real data are not LSD, and H-SVM can be used only for LSD, we use S-SVM for actual data. QP defines these SVMs. Although kernel SVM is one of nonlinear discriminant function and provides an attractive idea, we do not discuss it in this research because our concern is a comparison of LDFs. Many researchers use SVMs because there are many examinations of real data compared with the first generation research of MP-based discriminant theory. From 1971 to 1974, we became a member of the project to develop a computer system for an Electrocardiogram (ECG) data. Project leader, Doc. Nomura gave us a theme to develop a diagnostic logic using Fisher's LDF. Our research was inferior to Nomura's experimental decision tree algorithm. At first, we thought this failure was caused by our poor experience and knowledge of statistics. However, we considered the discriminant functions based on the variance-covariance matrices were not suitable for the medical diagnosis discussed in Section Ⅱ. Moreover, we found all LDFs cannot correctly discriminate the cases on the discriminant hyperplane (Problem 1).

In Section Ⅱ, although Fisher established discriminant analysis based on variance-covariance matrices, we explain a new theory of MP-based discriminant analysis (Theory) [53]. At first, we developed an Optimal LDF based on a Minimum Number of Misclassifications (minimum NM, MNM) criterion (IP-OLDF) in (1) [19] - [21]. It reveals two important facts of discriminant analysis. Those are 1) the relation of NM and LDF in the discriminant coefficient space, 2) monotonic decrease of MNM that is very crucial for gene analysis. It shows the good result by comparison with Fisher's LDF and QDF using Fisher's iris data [2] and Cephalo Pelvic Disproportion (CPD) data [14]. It finds Swiss banknote data is LSD [8]. All LDFs except for H-SVM and Revised IP-OLDF in (2) cannot discriminate LSD theoretically (Problem 2). Experimentally, Revised LP-OLDF in (2), one of L1-norm LDF using LP, can discriminate LSD. Nevertheless, it tends to gather cases on the discriminant hyperplane (Problem 1). Student data [24] reveals the defect of IP-OLDF caused by Problem 1. Therefore, Revised IP-OLDF is developed. It is only LDF to solve Problem 1. The pass/fail determination using exam scores [28] shows the defect of QDF and RDA caused by the defect of generalized inverse of variance-covariance matrices (Problem 3). If we add random noise to constant values of some particular variable, we can solve Problem 3. Japanese automobile data [35] explain Problem 3, also. Because Fisher never formulate the equation of Standard Error (SE) of error rate and discriminant coefficient, discriminant analysis is not traditional inferential statistics based on normal distribution

(Problem 4). A 100-fold cross-validation for small sample method (Method 1) offers the 95% Confidence Interval (CI) of error rate and discriminant coefficient [23] [25] - [27]. Moreover, because the best model with minimum mean of error rate in the validation samples is powerful model selection method and we can explain the meaning of discriminant coefficient [51][52], we understand to establish Theory. However, we know many researchers have been struggling in the gene analysis for more than ten years (Problem 5) [12].

In Section Ⅲ, we propose a Matroska feature selection method for gene analysis (Method 2). When we discriminate six microarray datasets (the datasets) [12], our three OLDFs can naturally select features [37] - [44]. However, three SVMs cannot select features. Moreover, Fisher's LDF cannot discriminate six datasets correctly because six NMs are not zero. In [42] [43] we explained in detail the results of Fisher's LDF. Revised IP-OLDF by Method 2 reveals the dataset consists of disjoint union of small linearly separable subspaces (SMs) and high-dimensional subspace that is not linearly separable (MNM >=1). This perception is essential for gene analysis.

In Section Ⅳ, we explain how to analyze each SM and find a Basic Gene Subspace (BGS) in each SM by ordinary statistical methods. We can analyze each SM very easy because all SMs are small samples. Moreover, we can understand the structure of dataset by BGSs because of monotonic decrease of MNM.

II. NEW THEORY OF DISCRIMINANT ANALYSIS

We develop four OLDF including IP-OLDF that find two new facts and solve four problems. Moreover, we confirm the best models of Revised IP-OLDF are better than other seven LDFs by six ordinary data introduced in Section Ⅰ.

A. Four Problems of Discriminant Analysis

There are four problems with the discriminant analysis [31][35] [36].

Problem 1: The discriminant rule is very simple. Let f($\mathbf{x}$) be LDF and y_i*f($\mathbf{x}_i$) be a discriminant score for $\mathbf{x}_i$. If y_i*f ($\mathbf{x}_i$) > 0, $\mathbf{x}_i$ is classified to class1/class2 correctly. If y_i*f($\mathbf{x}_i$) < 0, $\mathbf{x}_i$ is misclassified. We cannot properly discriminate $\mathbf{x}_i$ on the discriminant hyperplane (f($\mathbf{x}_i$) = 0). Many researchers ignore this unresolved problem until now. They consider a discriminant rule as follows: If f ($\mathbf{x}_i$) >= 0, $\mathbf{x}_i$ is classified to class 1 correctly. Otherwise, if f ($\mathbf{x}_i$) < 0, $\mathbf{x}_i$ is classified to class 2 correctly. Their discriminant rule is not logical. Only Revised IP-OLDF can treat this problem appropriately. Indeed, except for Revised IP-OLDF, no LDFs can count the NMs correctly. These LDFs should count the number of cases where f($\mathbf{x}_i$) = 0, and display this figure alongside the NM in the output. Student data tells us the defect of IP-OLDF. Therefore, we develop Revised IP-OLDF.

Problem 2: Only H-SVM and Revised IP-OLDF can recognize LSD theoretically. Experimentally, Revised LP-OLDF discriminates LSD correctly. Nevertheless, it tends to collect cases on the discriminant hyperplane (Problem 1). If we discriminate exam scores by four testlets score, and the pass mark is 50 point, we can obtain a trivial LDF such as f = T1 + T2+ T3+ T4 -50 [36]. We can judge the student pass the exam if f($\mathbf{x}_i$) >= 0 and fail the exam if f($\mathbf{x}_i$) < 0. However, error rates of Fisher's LDF and QDF are very high [35] because exam scores do not satisfy Fisher's assumption. Therefore, these LDFs should not be used in important applications such as medical diagnosis, pattern recognition, and rating.

Problem 3: Problem 3 is the defect of generalized inverse. When we discriminated math exam scores by QDF and RDA, all pass students were misclassified in the failed class because all pass students answered some item scores correctly, and scores of failed student vary. In this case, if we add random noise to the constant values, we can solve this problem.

Problem 4: Fisher never formulated the equation of SE of discriminant coefficients and error rates based on the normal distribution. Because there is no model selection procedure instead of a leave-one-out (LOO) procedure [13], we propose Method 1. It offers the 95% CI of error rates and discriminant coefficients. Moreover, it offers simple and powerful model selection procedure such as the best model with a minimum mean of error rate in the validation samples. We confirmed the best models of Revised IP-OLDF were better than Fisher's LDF, logistic regression and five MP-based LDFs using six ordinary data [29] [30] [33] [34]. Fisher's LDF and logistic regression discriminate these data by JMP script [15]. JMP division of SAS Institute Inc. Japan supports us to develop it. Six MP-based LDFs are Revised IP-OLDF, Revised LP-OLDF, Revised IPLP-OLDF, H-SVM and two S-SVMs such as SVM4 (penalty c = 10000) and SVM1 (penalty c = 1) by LINGO program that is supported by LINDO Systems Inc [17]. We can establish Theory by JMP and LINGO.

B. MP-based LDFs

Although we developed a diagnostic logic of ECG data by Fisher's LDF, our research was inferior to the decision tree logic developed by the medical doctor. After this experience, we concluded it is not adequate for the discrimination of the normal and abnormal diseases because of two main reasons [18].

1) There are many cases nearby the discriminant hyperplane. Medical doctors are striving to discriminate the cases nearby the discriminant hyperplane.

2) If the value of some variable increases or decreases, the probability belonging to abnormal disease increases from 0 to 1. Fisher's LDF assumes the typical abnormal patients are the average of the abnormal classes. However, the typical patients are far from the normal patients. Taguchi et al. [58] method was one of multi-class discrimination by Mahalanobis-distance based on the variance-covariance matrices. The authors claim that the cases belonging to abnormal states are far from the normal state. Their claim is the same perception as our claim. If some independent variable of logistic regression increases or decreases, the *probability 'p' belonging to class1 (abnormal symptom) increases from 0 (class2) to 1 (class1). Therefore, most medical users use logistic regression instead of Fisher's LDF. However, because JMP does not support logistic regression for the datasets, we never discuss logistic regression in Section Ⅲ.*

After many experiences of the discriminant analysis [14] [22], we developed IP-OLDF in (1). Because we fix the intercept of IP-OLDF to one, it is in the p-dimensional coefficient space. Although $y_i*({}^t\mathbf{x}_i\mathbf{b}+1)$ is discriminant scores, $y_i*({}^t\mathbf{x}_i\mathbf{b}+ 1) = 0$ is a linear hyperplane and divides discriminant space to two half planes such as plus half plane ($y_i*({}^t\mathbf{x}_i\mathbf{b}+1) > 0$) and minus half plane ($y_i*({}^t\mathbf{x}_i\mathbf{b}+1) < 0$). If we choose $\mathbf{b}_k$ in plus hyperplane as LDF, LDF such as $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1)$ discriminate $\mathbf{x}_i$ correctly because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) > 0$. On the other hand, if we choose $\mathbf{b}_k$ in minus hyperplane, LDF misclassify $\mathbf{x}_i$ because of $y_i*({}^t\mathbf{b}_k\mathbf{x}_i+1) = y_i*({}^t\mathbf{x}_i\mathbf{b}_k+1) < 0$. However, we must solve other two models such as the intercept = -1 and 0. It looks for the right vertex of an Optimal Convex Polyhedron (optimal CP, OCP) if data is a general position. There are only p-cases on the discriminant hyperplane, and it becomes the vertex of OCP. On the other hand, if data is not general position, it may not look for the correct vertex of OCP because there are over (p+1) cases on the discriminant hyperplane, and we cannot correctly discriminate these cases. Therefore, we developed Revised IP-OLDF that looks for the interior point of true OCP in (2) directly. Because b_0 is free variable, it is defined in (p+1)-dimensional coefficient space. If it discriminates $\mathbf{x}_i$ correctly, $e_i = 0$ and $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) >= 1$. If it cannot discriminate $\mathbf{x}_i$ correctly, $e_i = 1$ and $y_i* ({}^t\mathbf{x}_i\mathbf{b}+b_0) >= -9999$. Although support vector (SV) for classified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = 1$, SV for misclassified cases are $y_i*({}^t\mathbf{x}_i\mathbf{b}+b_0) = -9999$. Therefore, we expect a discriminant score of misclassified cases are less than -1, and there are no cases within two SVs. Therefore, if M is small constant, it does not work correctly [27]. Because there are no cases on the discriminant hyperplane, we can understand the optimal solution is an interior point of OCP defined by IP-OLDF. All LDFs except for Revised IP-OLDF cannot solve Problem 1 theoretically. Therefore, these LDFs must check the number of cases (h) on the discriminant hyperplane. Correct NM may increase (NM + h).

$$\text{MIN} = \Sigma\, e_i;\ y_i*({}^t\mathbf{x}_i\mathbf{b} + 1) >= - e_i\ ; \qquad (1)$$

e_i: 0/1 integer variable corresponding to classified/misclassified cases.
y_i: 1/-1 for class1/class2 or object variable.
$\mathbf{x}_i$: p-independent variables.
$\mathbf{b}$: discriminant coefficients.

Because we can consider IP-OLDF in (1) on the data and discriminant coefficients spaces, we find two relevant facts as follows.

1) We explain the notation of IP-OLDF by the Golub et al. dataset [10]. It consists of two classes such as "All (47 cases)" and "AML (25 cases)" with 7,129 genes. Our primary concern is to discriminate two classes by 7,129 variables (genes). The 72 linear hyperplane, the 7,129 coefficients of those are values of each case, divide the discriminant coefficient space into finite CP. The interior points of each CP correspond to the discriminant coefficient of LDF that discriminates the same cases correctly and misclassifies another same case. Therefore, because the interior points of each CP have unique NM, we can define the OCP with MNM. Many examinations show the best models of Revised IP-OLDF are better than other seven LDFs.

2) Let MNM_k be MNM in the k-dimensional subspace. MNM decreases monotonously ($MNM_k >= MNM_{(k+1)}$). If $MNM_k = 0$, all MNMs including these k-variables (genes) are zero. This fact tells us the smallest Matroska (Basic Gene Subspace, BGS) can completely describe the structure of gene space by monotonic decreases of MNM.

When we discriminate Swiss banknote data with six variables, IP-OLDF finds two-variables models, such as (X4, X6), is linearly separable. By the monotonic decrease of MNN, 16 MNMs including these two variables are zero among 63 models (= 2^6-1 = 63). Other 47 MNMs are greater than one. Revised IP-OLDF in (2) can naturally select features for ordinary data and six datasets. However, we develop more powerful model selection procedure such as the best model by Method 1. Therefore, we had ignored the natural feature selection for ordinary data before Method 2.

$$\text{MIN} = \Sigma e_i\ ;\quad y_i* (\ {}^t\mathbf{x}_i\mathbf{b} + b_0) >= 1 - M* e_i\ ; \qquad (2)$$

b_0: free decision variables.
M: 10,000 (Big M constant).

If e_i is non-negative real variable, equation (2) changes Revised LP-OLDF. Revised IPLP-OLDF [32] is a mixture model of Revised LP-OLDF in the first phase and Revised IP-OLDF in the second phase. The equation (3) is S-SVM. If we set c=10^4 or c=1, it becomes SVM4 or SVM1. If we omit "c* Σe_i" and "- e_i", it becomes H-SVM. QP solves both SVMs.

$$\text{MIN} = ||\mathbf{b}||^2/2 + c*\Sigma e_i\ ;\quad y_i* (\ {}^t\mathbf{x}_i\mathbf{b} + b_0) >= 1 - e_i\ ; \qquad (3)$$

c: penalty c to combine two objectives.
e_i: non-negative real value.

C. *New Theory of Discriminant Analysis (Theory)*

We explain the outlook of Theory. There are four serious problems with the discriminant analysis. We developed four MP-based OLDFs. IP-OLDF finds two new facts of discriminant analysis. Revised IP-OLDF solves Problem 1 and Problem 2 related to this paper. Because Method 1 solves Problem 4 and four problems are solved completely, we misunderstand to establish Theory. In 2015, when we discriminated six datasets by MP-based LDFs and Fisher's LDF, only Revised IP-OLDF could naturally select features because coefficients less than 173 are not zero and other coefficients become zeroes [37]. After we recognize this fifth problem, we completely solve Problem 5 by Method 2 in Dec.

2015. Although we had observed the feature selection of Revised IP-OLDF by Swiss banknote data and Japanese automobile data that are LSD, we ignore this fact because the best model is an excellent model selection procedure for ordinary six data. In gene analysis, if we call all linearly separable models as Matroskas that are linearly separable gene subspaces, Revised IP-OLDF reduces the high-dimension gene space, the big Matroska, to small subspace (SM) drastically. After we remove genes in the first SM1 from the big Matroska, Revised IP-OLDF discriminates the new gene space (the second big Matroska), again. It can find the second different SM2. We repeat this process and locate the dataset that consists of the disjoint union of SMs and high-dimension gene subspace (MNM>=1). Therefore, we develop Method 2. We make a program of Method 2 by LINGO and can list up all SMs of the six datasets very easy. Although many researchers have been struggling to analyze the high-dimension gene datasets by a statistical approach [55] [60], we can analyze each SM very easy because it is a small sample. In Section Ⅳ, we show how to find BGSs by manual operation and analyze one of each SM by the ordinary statistical approach. In Section Ⅴ, we discuss the use and application of our results.

D. *Short Story of Feature Selection*

At the end of October 2015, we presented our Theory at Japanese statistical conference and knew six datasets presented by another researcher presentation. After the conference, we discriminated six datasets by seven LDFs. Because error rates of Fisher's LDF were very high for eighteen exam scores [35], it is self-evident we cannot obtain better results in the gene datasets. Therefore, users never use it for gene analysis. Although NMs of three SVMs are zero, all coefficients are not zero. Therefore, three SVMs are not helpful for the feature selection. Several coefficients of Revised IP-OLDF are not zero, and most of the coefficients are zero. It can naturally select features of the datasets within few seconds and reduce high-dimension genes spaces to the smaller subspace that is one of the Matroska. Next, when we discriminate the Matroska again, we can find smaller Matroska. Therefore, the dataset has the structure of Matroska. When we cannot locate the smaller Matroska again, we call the last subspace as the Small Matroska (SM1). Moreover, after we exclude the first SM1 from the dataset, we find the second different SM2. At last, we can list up all SMs by a LINGO program of Method 2 and conclude the dataset consists of the disjoint union of SMs and another high-dimension gene subspace that is not linearly separable. Six studies [45] - [50] include full genes lists of the SMs about six datasets. If we analyze all SMs, we may be able to obtain new facts of gene analysis. Although some researchers try to discriminate the dataset by LASSO based on variance-covariance matrices, our Theory showed only H-SVM and Revised IP-OLDF can discriminate LSD theoretically, and revealed the structure of datasets. If LASSO researchers compare their results with our results using our two ordinary data and six datasets, it is expected to improve the research of feature selection method more deeply.

III. Matroska Feature Selection Method

In this section, we introduce Method 2.

A. *Outlook of Method 2*

When we discriminate Shipp et al. data [54] on Oct. 28, 2015, only Revised IP-OLDF can select thirty-two genes among 7129 genes [37]. Although we misunderstand the discrimination having 7129 variables requests huge CPU time, Fisher's LDF by JMP ver.12 (JMP12) and other MP-based LDFs coded by LINGO can solve the datasets less than 20 seconds because the datasets are LSD. However, most coefficients of these LDFs except for Revised IP-OLDF are not zero. Therefore, these LDFs are not helpful for feature selection for gene analysis in addition to ordinary data. In this research, we call the smallest Matroska as the BGS with k-variables. The biggest Matroska with 7129 variables includes many smaller Matroskas from 7128 (= 7129 - 1) variables to k variables. LINGO program found the datasets are the disjoint union of SMs with h-variables (p > h >= k) and another high-dimension gene subspace with "MNM >= 1." Now, we must survey the BGSs from SM by manual operation. If Revised LINGO program can find all list of BGSs, we can understand the structure of the dataset by these BGSs completely. Because we can analyze each SM using ordinary statistical methods, we expect to obtain new facts of gene analysis and hope many researchers try to analyze these SMs. By our breakthrough, the feature selection becomes exciting theme.

We guess the reason why Revised IP-OLDF can naturally select features as follows.

1) *MNM criterion works well for the feature selection. This expectation will be right if LASSO cannot list up all SMs or BGSs correctly as same as Revised IP-OLDF because it does not use MNM criterion. We consider the discrimination of LSD requests MNM criterion or maximization of two SVs.*

2) *The algorithm of LINGO IP solve uses the branch and bound. We believe Revised IP-OLDF coded by another IP algorithm cannot naturally select features. On the other hand, we cannot control the flow of the branch and bound. When IP solver finds the model with MNM=0 at first, LINGO program output it and end. This treatment is the reason why LINGO program may not be able to find BGS directly. This research is our future theme.*

B. *Results of Six Microarray Data*

Table Ⅰ shows the summary of six datasets. Rows "Description" show two classes. Rows "Size" are the case number by the gene number. Rows "SM: Gene" are the number of SM [with reference number]: the total number of genes including in all SMs. Six lists of full gene name are in the references. Rows "Min, Mean, Max" are the minimum, mean and maximum values of genes including in all SMs. Rows "JMP12" are 2 by 2 tables of the discrimination by Fisher's LDF. Six NMs are 5, 3, 8, 3, 10 and 29. Rows "% and error rate" are the percentages of (Maximum value/case number) and error rates of JMP12. Maximum percent is 63%

by Alon et al. dataset. Minimum percent is 43% by Golub et al. dataset. Maximum error rate is 17% by Tian et al. dataset and minimum error rate is 1% by Chiaretti et al. dataset

TABLE I. SUMMARY OF SIX MICROARRAY DATASETS

Data	Alone et al. [1]	Chiaretti et al. [2]
Description	Normal (22) vs. tumor cancer (40)	B-cell (95) vs. T-cell (33)
Size	62 *2000	128*12625
SM: Gene	64 [47]:1152	270 [50]:5385
Min/Mean/Max	11/18/39	9/19/62
JMP Ver.12	20:2/3:37	94:1/2:31
% and error rate	63%, 8%	49%, 1%
Data	Golub et al. [10]	Shipp et al. [54]
Description	All (47) vs. AML (25)	Follicular lymphoma (19) vs. DLBCL (58)
Size	72*7129	77 *7130
SM: Gene	69 [46]:1238	213 [45]:3032
Min/Mean/Max	10/18/31	7/14/43
JMP12	20:5/3:44	17:2/1:57
% and error rate	43%, 11%	56%, 4%
Data	Singh et al. [56]	Tian et al. [59]
Description	Normal (50) vs. tumor prostate (50)	False (36) vs. True (137)
Size	102 *12626	173 *12625
SM: Gene	179 [48]:3990	159 [49]:7221
Min/Mean/Max	13/22/47	28/45/104
JMP Ver.12	46:4/6:46	16:20/9:128
% and error rate	46%, 10%	60%, 17%

C. Detail of the Matroska Feature Selection Method

We explain Method 2 briefly. Table Ⅱ is the output of Golub et al. dataset by LINGO program. Two columns "LOOP1 and LOOP2" are the sequence number of big and small loops of Method 2. Revised IP-OLDF discriminate the dataset with 7129 genes in the LOOP1=1 and LOOP2=1, and only 34 coefficients of Revised IP-OLDF are not zero. In general, this number is less than the case number such as 72. In the second small loop (LOOP1=1, LOOP2=2), we discriminate the smaller Matroska with 34 genes again, and only 11 coefficients are not zero. Therefore, we get the Matroska sequence such as Matroska7129 → Matroska34 → Matroska11. We stop at LOOP2=4 because we cannot find the smaller Matroska. We call Matroska11 as the SM1 because Revised IP-OLDF cannot locate the smaller Matroska. We exclude the first SM1 with 11 genes from the big Matroska with 7129 genes and make the second big Matroska with 7118 genes. In the second big loop at LOOP1 = 2, we get the second SM2 with 16 genes.

TABLE II. THE OUTLOOK OF THE THEORY 2

SN	LOOP1	LOOP2	Gene	MNM
1	1	1	7129	0
2	1	2	34	0
3	1	3	11	0
4	1	4	11	0
16	2	1	7118	0
17	2	2	36	0
18	2	3	18	0
19	2	4	16	0
20	2	5	16	0

After LINGO program finds sixty-nine SMs in Table Ⅲ, it stops the big loop when we find MNM is greater than one at LOOP1=70. However, we can continue this loop until it cannot naturally select features and list up all small subspaces with "MNM >= 1." Therefore, Method 2 can discriminate other types of genes datasets that are not LSD. Because Golub et al. dataset consists 69 SMs that are linearly separable models or subspaces, it is very easy for us to analyze all SMs because the 68th and 69th SMs are the biggest samples with 72 cases by 31 genes.

TABLE III. ALL SMALL MATROSKA OF GOLUB ET AL. DATA

LOOP1	LOOP2	Gene	n	MNM
1	11	7129	11	0
2	11	7118	16	0
3	11	7102	11	0
-	-	-	-	-
32	11	6683	19	0
33	11	6664	16	0
34	11	6648	18	0
35	11	6630	17	0
36	11	6613	19	0
37	11	6594	12	0
38	11	6582	16	0
-	-	-	-	-
67	11	5976	23	0
68	11	5953	31	0
69	11	5922	31	0

IV. BGS AND STATISTICAL ANALYSIS

In this section, we introduce how to find BGS and analyze it.

A. How to find BGSs

Because we cannot control the flow of branch and bound algorithm, there may be several BGSs in the SM. We propose how to find BGSs by manual operation as follows:

1) To find the smaller linear separable model in SM

We analyze the first SM1 with 11 genes by the forward stepwise procedure and obtain the five columns from 'Step' to 'BIC' in Table Ⅳ. The last column is NM of logistic regression. Although there is no theoretical guarantee that logistic regression can discriminate LSD correctly [5], we judge it discriminates LSD correctly under the condition of "MNM=0 and NM=0". Therefore, we can judge BGS exists among all combination models in four genes subspace [11]. We know the four-variable model is linearly separable. Cp, AIC and BIC recommend this model. Usually, these three statistics recommend the different models by our many trials.

TABLE IV. FORWARD STEPWISE AND LOGISTIC REGRESSION.

Step	Gene	Cp	AIC	BIC	logistic
1	M11722_at	72.56	137.78	144.26	5
2	X59871_at	38.42	118.62	127.13	2
3	U05259_rna1_at	9.92	96.07	106.54	2
4	D21063_at	3.88	90.15	102.52	0
5	M22919_rna2_at	3.80	90.30	104.49	0
6	M21624_at	4.27	91.09	107.02	0
7	M25280_at	4.63	91.79	109.38	0
8	L13210_at	6.15	93.93	113.09	0
9	X82240_rna1_at	8.02	96.56	117.21	0
10	HG3039-HT3200_at	10.01	99.44	121.47	0
11	L76159_at	12.00	102.41	125.73	0

TABLE V. FIFTEEN MODEL BY FOUR GENES

p	X1	X2	X3	X4	c	MNM	ZERO
4	1	1	1	1	1	0	0
3	1	1	1	0	1	1	0
3	1	1	0	1	1	1	0
3	1	0	1	1	1	3	0
3	0	1	1	1	1	2	0
2	1	1	0	0	1	2	0
2	1	0	1	0	1	4	0
2	0	1	1	0	1	3	0
2	1	0	0	1	1	4	0
2	0	1	0	1	1	13	0
2	0	0	1	1	1	6	0
1	1	0	0	0	1	5	0
1	0	1	0	0	1	25	0
1	0	0	1	0	1	10	0
1	0	0	0	1	1	17	0

2) Search BGSs by all possible combination models

We search BGSs by all possible combination models using Revised IP-OLDF. Table Ⅴ is the 15 models by four genes that are four combinations of 0/1 values from the second column to the fifth column. Column "c" is the intercept of Revised IP-OLDF. The column "p" is the number of independent variables from four-variable model (p=4) to four one-variable models (p=1). The binary values, such as 1/0, mean each model include or not include four variable in the model. Column "MNM" is MNM of 15 models. Column "ZERO" is the number of cases on the discriminant hyperplane. Only full model is linearly separable. Therefore, we find one BGS in the first SM, such as (X1: M11722_at, X2: X59871_at, X3: U05259_rna1_at, X4: D21063_at). All MNMs including these four genes are linearly separable in Golib et al. dataset. Therefore, although there are numerous Matroskas in the dataset, we can understand the structure of Matroska by BGS because of the monotonic decrease of MNM. The big Matroska with 7129 genes includes numerous smaller Matroska from 7128 genes to four genes. Although there are 7129 subspaces with 7128 genes, there are 7125 smaller Matroska with 7128 genes and four subspaces with 7128 genes that are not Matroska. By monotonic decrease of MNM, we can completely understand the structure of Matroska. It is hard for us to analyze the dataset by the ordinary statistical methods without knowledge of this fact.

B. How to analyze each SM

Figure 1 is the output of principal component analysis (PCA). Left figure is the eigenvalues. Two eigenvalues are greater than one and contribution ratio is about 0.75. The middle figure is the scatter plot. The symbol + are "AMLs" that are in the third quadrant. Forty-seven cases of "ALL" are situated in the fourth, first and second quadrant. The right plot is the factor loading plot. "M11722_at" is overlapped on the first component and "X59871_at" is overlap on the second component. It is very important for specialists of gene analysis to consider the reason why two groups are orthogonal. We expect specialists of gene analysis to examine the meaning of statistical outputs of SMs.

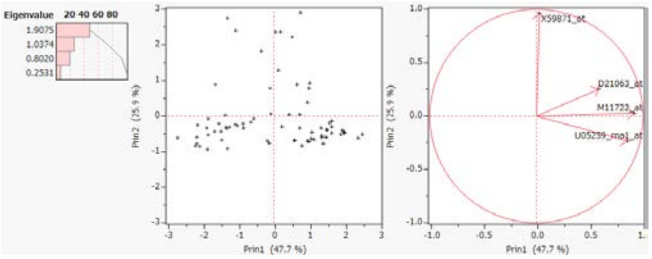


Figure 1. The principal component analysis.

Figure 2 is two score plots. X-axis is the first component. Y-axis of left and right score plots correspond the second component and the third component. Because PCA cannot separate two classes, ordinary statistical analysis such as one-way ANOVA, cluster analysis, and PCA cannot conclude clear results for the datasets directly. Jeffery et al. compared the efficiency of the ten feature selection methods using conventional statistical approaches. It tells us the limitation of conventional statistical methods.

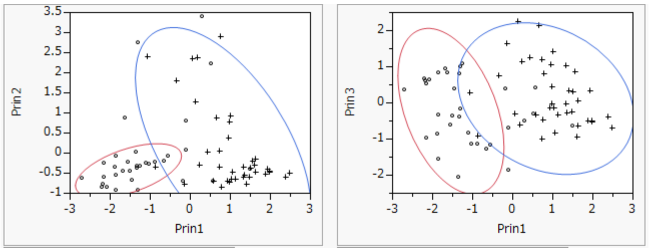


Figure 2. Two score plots.

Table VI is the correlation matrix. The absolute correlations of "X59871_at" with other three genes are less than 0.088 that are the same result as the factor loading plot.

TABLE VI. CORRELATION MATRIX.

Var.	X1	X2	X3	X4
M11722_at	1	0.076	0.713	0.371
X59871_at	0.076	1	-0.088	0.052
U05259_rna1_at	0.713	-0.088	1	0.220
D21063_at	0.371	0.052	0.220	1

V. CONCLUSION

We developed Theory, Method 1 and Method 2. Revised IP-OLDF solves Problem 1, Problem 2 and Problem 5. Moreover, the best models of Revised IP-OLD are better than another seven LDFs. Although H-SVM discriminate LSD correctly, it cannot naturally select features for six datasets. Because Problem 3 is the defect of the generalized inverse and error rates of Fisher's LDF and QDF are very high for LSD, we guess the discriminant analysis and regression analysis based on variance-covariance matrices may not be helpful for gene analysis. Although the discriminant analysis is not the traditionaly inferential statistical method, Method 1 offers the 95% CI of error rate and discriminant coefficient and the validation of Revised IP-OLDF by six ordinary data. In this paper, we do not discuss the validation of six microarray datasets. However, because Method 1 validated already six ordinary data, we will validate the results of six microarray datasets in near future. Because the best model is powerful model selection procedure for ordinary data, we ignore some parameters of Revised IP-OLDF are zeroes in ordinary data. Because other LDFs cannot naturally select features, they may be difficult for gene datasets. If we can develop Revised LINGO program that can find all BGSs, it will be more useful in gene analysis. LINGO program is useful for other gene dataset, such as RNA-Seq., in addition to the six datasets. Although we surveyed to clarify the long-term survivors of the Maruyama vaccine (SSM) administration patients, our trial failed [22]. If we compare two lists of cancer genes, (normal and cancer patient data) vs. (normal and SSM Administration patient data), and find the differences between two gene lists, it may show the proof of the effectiveness of SSM. Now, we plan this new theme and have proposed a joint research with the inspection agency of microarray in Japan.

We would like to propose a joint research with medical doctors in the world.

REFERENCES

[1] A. Alon et al., "Patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays," Proc. Natl. Acad. Sci. USA, 96, pp. 6745-6750, 1999.

[2] E. Anderson, "The irises of the Gaspe Peninsula," Bulletin of the American Iris Society vol. 59, pp. 2-5, 1945.

[3] S. Chiaretti et al., "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," Blood. April 1, 2004, 103/7, pp. 2771-2778, 2004.

[4] D. R. Cox, "The regression analysis of binary sequences (with discussion)," J Roy Stat Soc B 20, pp. 215-242, 1958.

[5] D. Firth, "Bias reduction of maximum likelihood estimates," Biometrika, vol. 80, pp. 27-39, 1993.

[6] R. A. Fisher, "The Use of Multiple Measurements in Taxonomic Problems," Annals of Eugenics, 7, pp. 179–188. 1936.

[7] R. A. Fisher, Statistical methods and statistical inference. Hafner Publishing Co. 1956.

[8] B. Flury and H. Rieduyl, Multivariate Statistics: A Practical Approach. Cambridge University Press. 1988.

[9] J. H. Friedman, "Regularized Discriminant Analysis," Journal of the American Statistical Association, 84/405, pp. 165-175, 1989.

[10] T. R. Golub et al., "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," Science. 1999 Oct 15, 286/5439, pp. 531-537, 1999.

[11] J. H. Goodnight, SAS Technical Report – The Sweep Operator: Its Importance in Statistical Computing – (R-100). SAS Institute Inc. 1978.

[12] I. B. Jeffery, D. G. Higgins, and C. Culhane, "Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data," BMC Bioinformatics. Jul 26 7:359, pp.1-16, Jul. 2006. doi: 10.1186/1471-2105-7-359. (Accessed Oct. 28, 2015).

[13] P. A. Lachenbruch, M. R. Mickey, "Estimation of error rates in discriminant analysis," Technometrics 10, pp. 1-11, 1968.

[14] A. Miyake, S. Shinmura, "An Algorithm for the Optimal Linear Discriminant Function and its Application," Japan Society of Medical Electronics and Biological Engineering, 18/6, pp. 452-454, 1980.

[15] J. P. Sall, L. Creighton, and A. Lehman, JMP Start Statistics, Third Edition. SAS Institute Inc. 2004. (S. Shinmura, supervise Japanese version)

[16] L. Schrage, LINDO ‑ An Optimization Modeling Systems ‑ . The Scientific Press. 1991. (S. Shinmura & H. Takamori, translate Japanese version)

[17] L. Schrage, Optimization Modeling with LINGO. LINDO Systems Inc. 2006.

[18] S. Shinmura, "Medical Data Analysis, Model, and OR," Operations Research, 29/7, pp. 415-421, 1984.

[19] S. Shinmura, "Optimal Linear Discriminant Functions using Mathematical Programming," Journal of JSCS, 11 / 2, pp. 89-101, 1998.

[20] S. Shinmura, “A new algorithm of the linear discriminant function using integer programming,” New Trends in Probability and Statistics, 5, pp. 133-142, 2000.

[21] S. Shinmura, Optimal Linear Discriminant Function using Mathematical Programming. Dissertation, March 200, pp. 1-101, Okayama Univ. 2000.

[22] S. Shinmura, “Analysis of Effect of SSM on 152,989 Cancer Patient,” ISI2001, pp. 1-2. 2001.

[23] S. Shinmura, “New Algorithm of Discriminant Analysis using Integer Programming,” IPSI 2004, Pescara VIP Conference CD-ROM, pp. 1-18, 2004.

[24] S. Shinmura, “Comparison of Revised IP-OLDF and SVM,” ISI2009, pp. 1-4, 2007.

[25] S. Shinmura, “Overviews of Discriminant Function by Mathematical Programming,” Journal of JSCS, 20/1-2, pp. 59-94, 2007.

[26] S. Shinmura, “Practical discriminant analysis by IP-OLDF and IPLP-OLDF,” IPSI 2009, Belgrade VIPSI Conference, CD-ROM, pp. 1-17, 2009.

[27] S. Shinmura, The optimal linear discriminant function. Union of Japanese Scientist and Engineer Publishing. 2010.

[28] S. Shinmura, “Problems of Discriminant Analysis by Mark Sense Test Data,” Japanese Society of Applied Statistics， 40/3, pp. 157-172, 2011.

[29] S. Shinmura, “Beyond Fisher’s Linear Discriminant Analysis - New World of Discriminant Analysis -,” ISI2011 CD-ROM, pp. 1-6, 2011.

[30] S. Shinmura, “Evaluation of Optimal Linear Discriminant Function by 100-fold Cross-validation,” 2013 ISI CD-ROM, pp. 1-6, 2013.

[31] S. Shinmura, “End of Discriminant Function based on Variance-Covariance Matrices,” ICORES, pp. 5-14, 2014.

[32] S. Shinmura, “Improvement of CPU time of Linear Discriminant Functions based on MNM criterion by IP,” Statistics, Optimization and Information Computing, 2, pp. 14-129, 2014.

[33] S. Shinmura, “Comparison of Linear Discriminant Function by K-fold Cross-validation,” Data Analytic 2014, pp. 1-6, 2014.

[34] S. Shinmura, “The 95% confidence intervals of error rates and discriminant coefficients,” Statistics Optimization and Information Computing, 3, pp. 66-78, 2015.

[35] S. Shinmura, “Four Serious Problems and New Facts of the Discriminant Analysis,” In E. Pinson, F. Valente, B. Vitoriano, (Eds.), Operations Research and Enterprise Systems, pp. 15-30, 2015. Springer (ISSN: 1865-0929, ISBN: 978-3-319-17508-9, DOI: 10.1007/978-3-319-17509-6).

[36] S. Shinmura, “A Trivial Linear Discriminant Function,” Statistics, Optimization and Information Computing, 3, Dec. 2015, pp. 322-335, 2015.

[37] S. Shinmura, “The Discrimination of the microarray data (Ver. 1),” Research Gate (1), Oct. 28, 2015: pp. 1-4, 2015.

[38] S. Shinmura, “Feature Selection of three Microarray data,” Research Gate (2), Nov.1, 2015: pp. 1-7, 2015.

[39] S. Shinmura, “Feature Selection of Microarray Data (3) – Shipp et al. Microarray Data,” Research Gate (3), 2015: pp. 1-11, 2015.

[40] S. Shinmura, “Validation of Feature Selection (4) – Alon et al. Microarray Data,” Research Gate (4), 2015: pp. 1-11, 2015.

[41] S. Shinmura, “Repeated Feature Selection Method for Microarray Data (5),” Research Gate (5), Nov. 9, 2015, pp. 1-12, 2015.

[42] S. Shinmura, “Comparison Fisher’s LDF by JMP and Revised IP-OLDF by LINGO for Microarray Data (6),” Research Gate (6), Nov. 11, 2015, pp. 1-10, 2015.

[43] S. Shinmura, “Matroska Trap of Feature Selection Method (7) –Golub et al. Microarray Data,” Research Gate (7), Nov. 18, 2015, pp. 1-14, 2015.

[44] S. Shinmura, “Minimum Sets of Genes of Golub et al. Microarray Data (8),” Research Gate (8), Nov. 22, 2015, pp. 1-12, 2015.

[45] S. Shinmura, “Complete Lists of Small Matroska in Shipp et al. Microarray Data (9),” Research Gate (9), Dec. 4, 2015, pp. 1-81, 2015.

[46] S. Shinmura, “Sixty-nine Small Matroska in Golub et al. Microarray Data (10),” Research Gate, Dec. 4, pp. 1-58, 2015.

[47] S. Shinmura, “Simple Structure of Alon et al. et al. Microarray Data (11),” Research Gate (11), Dec. 4, 2015, pp. 1-34, 2015.

[48] S. Shinmura, “Feature Selection of Singh et al. Microarray Data (12),” Research Gate (12), Dec. 6, 2015, pp. 1-89, 2015.

[49] S. Shinmura, “Final List of Small Matroska in Tian et al. Microarray Data,” Research Gate (13), Dec. 7, pp. 1-160, 2015.

[50] S. Shinmura, “Final List of Small Matroska in Chiaretti et al. Microarray Data,” Research Gate (14), Dec. 20, 2015, pp. 1-16, 2015.

[51] S. Shinmura, “The best model of the Swiss bank note data,” Statistics, Optimization and Information Computing, 3, Spring. 2016, pp. 0-13, 2016. (unpublished).

[52] S. Shinmura, “Discriminant Analysis of the Linearly Separable Data,” Journal of Statistical Science and Application, 2016. (unpublished).

[53] S. Shinmura, New Theory of Discriminant Analysis after R. Fisher, Springer, Dec. 2016. (unpublished).

[54] M. A. Shipp et al., “Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning,” Nature Medicine 8, pp. 68-74, 2002.

[55] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “A sparse-group lasso,” J. Comput. Graph. Statist., 22. pp. 231-245, 2013.

[56] D. Singh et al., “Gene expression correlates of clinical prostate cancer behavior,” Cancer Cell: March 2002, Vol.1, pp. 203-209, 2002.

[57] A. Stam, “Non-traditional approaches to statistical classification: Some perspectives on Lp-norm methods,” Annals of Operations Research, 74, pp. 1-36, 1997.

[58] G. Taguchi and R. Jugulum, The Mahalanobis-Taguchi Strategy-A Pattern Technology System. John Wiley & Sons. 2002.

[59] E. Tian et al., “The Role of the Wnt-Signaling Antagonist DKK1 in the Development of Osteolytic Lesions in Multiple Myeloma,” The new England Journal of Medicine, Vol. 349, 26, pp. 2483-2494, 2003.

[60] R. Tibshirani, “Regression Shrinkage and Selection via the Lasso,” J. R. Statist. Soc. B 58/1, pp. 267-288, 1996.

[61] V. Vapnik, The Nature of Statistical Learning Theory. Springer. 1995.

In Vivo Single-Molecule Dynamics of Transcription of the Viral T7 Phi 10 Promoter in *Escherichia coli*

Nadia S.M. Goncalves[1], Huy Tran[1], Samuel M.D. Oliveira[1], Ramakanth Neeli-Venkata[1], Andre S. Ribeiro[1]
[1] Laboratory of Biosystem Dynamics, Tampere University of Technology, Finland.
e-mail: andre.ribeiro@tut.fi

Leonardo Martins[2], José M. Fonseca[2]
[2]Computational Intelligence Group, CTS/UNINOVA, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, Portugal.
e-mail: jmf@uninova.pt

***Abstract* – We study the dynamics of transcription initiation of the T7 Phi 10 promoter as a function of temperature, using quantitative polymerase chain reaction (qPCR) and *in vivo* single-cell, single-ribonucleic acid (RNA) time-lapse microscopy. First, from the mean and squared coefficient of variation of the empirical distribution of intervals between consecutive RNA appearances in individual cells, we find that both the mean rate and noise in RNA production increase with temperature (from 20°C to 43°C). Next, the process is shown to be sub-Poissonian in all conditions, suggesting the existence of more than one rate-limiting step and absence of a significant ON-OFF mechanism. Next, from the kinetics of RNA production for varying amounts of T7 RNA polymerases, we find that as temperature increases, the fraction of time that the T7 RNA polymerase spends in open complex formation increases relative to the time to commit to closed complex formation, due to changes in the kinetics of open complex, closed complex, and reversibility of the closed complex formation. We conclude that the initiation kinetics of the T7 Phi 10 promoter changes with temperature due to changes in the kinetics of its rate-limiting steps.**

Keywords* – *Transcription; Open and closed complex formation; T7 Phi 10 promoter

I. Introduction

The bacteriophage T7 is an obligate lytic phage that infects *Escherichia coli*, using the host system to produce up to 100 progeny phages in less than 25 min, in optimal conditions [1]. One of the major gene products of T7 bacteriophage is the T7 RNA polymerase (T7 RNAP) [2]. This is a single subunit enzyme, with a high specificity towards T7 promoters via the recognition of a highly conserved 23bp consensus sequence [3]. Early studies have shown that the T7 RNAP transcription rate is sequence dependent and depends on environmental conditions [4][5][6]. Given that the infection process of T7 bacteriophage is not only fast but it also requires a balance between the number of phages and the amount of capsid proteins produced [7], the phage needs to coordinate the dynamics of transcription of the viral genes, as this is likely critical for its success.

It is known that the dynamics of gene expression, as well as of many other cellular processes, depends on environmental factors, particularly temperature [8]. Consequently, microorganisms have evolved mechanisms that allow them to cope with both sudden as well as slow temperature changes [9][10]. *E. coli*, for example, can survive in a wide range of temperatures. Similarly, it has also been shown that the T7 bacteriophage is capable of coping with these fluctuations and wide ranges [5].

Even though robustness to sudden temperature changes and wide temperature ranges is crucial for the survival of microorganisms, so far, little is known about what are the consequences of these environmental changes on the *in vivo* transcription kinetics of the T7 promoter. In addition, most studies characterizing the transcription initiation kinetics of T7 promoters have mostly used *in vitro* measurement techniques [5][11].

To address this issue, here we use recently developed measurement strategies that use single-cell, single-RNA *in vivo* detection techniques [12] and use them to study in detail the kinetics of transcription initiation of the T7 *Phi10* (*Φ10*) promoter as a function of temperature.

The remaining of this article is organized as follows: Section II describes the methods used and measurements conducted. Section III presents the results from these experiments. In Section IV, we conclude by presenting our interpretation of the results and our assessment of their relevance, as well as additional considerations for future work.

II. Methods

In this section, we describe the measurements conducted in this study. Each subsection presents a detailed explanation of the experiments performed.

A. *Strain and plasmids*

The strain *E. coli* BL21(DE3) (New England Biolabs, USA) was used to express the target and reporter genes. This strain has a copy of the T7 bacteriophage gene 1 coding for T7 RNAP controlled by the P_{lacUV5} promoter and integrated in the chromosome [13] (Figure 1A).

The single copy F-plasmid pBELOBAC11, carrying the *Φ10-mCherry-48bs* sequence (constructed for this work) was inserted in the host strain. It produces the target RNA, with an array of 48 MS2 binding sites (*48bs*) under the control of a T7 *Φ10* promoter, cloned from the plasmid pRSET/EmGFP (ThermoScientific, USA).

A second plasmid, pZA25-GFP (Green Fluorescent Protein) [14] (a gift from Orna Amster-Choder, Hebrew University of Jerusalem, Israel), was also inserted in the host strain. It contains the reporter gene *ms2-gfp*, placed under the control of P_{BAD} promoter. This reporter gene encodes for the fusion protein MS2-GFP, which binds the target RNAs and renders them visible as bright spots under the confocal microscope [15] (Figure 1B). From here onwards we refer to the T7 *Φ10* promoter as T7 promoter.

B. *Microscopy*

For live cell microscopy, BL21(DE3) cells were incubated in M63 medium supplemented with Glucose (0.4%) and the appropriate concentration of Chloramphenicol and Kanamycin (Sigma Aldrich, USA) and was grown overnight at 30^{o}C, with shaking (250 rpm). Cells from the overnight culture were then diluted in fresh M63 medium, with an initial OD600 ~ 0.05, and incubated at 37^{o}C, for 90 minutes with shaking (250 rpm). Then, cells were pelleted and re-suspended in ~100 µl of M63 medium. Four microliters of cells were placed between a 3% agarose gel pad, made with M63 medium, and a glass coverslip before assembling the imaging chamber (CFCS2, Bioptechs, USA). Two hours before the microscopy measurements, a flow of fresh M63 medium at 37^{o}C containing the reporter inducer (0.8% L-arabinose) was initiated with a peristaltic pump at a rate of 1 ml/min to produce sufficient MS2-GFP molecules in the cells to detect the target RNA in all experiments. Note that we shifted the temperature in the chamber from 37^{o}C to 20^{o}C or to 43^{o}C (depending on the condition studied), 20 minutes prior to inducing the target system.

To activate the target system, we induced the production of T7 RNAP, controlled by P_{lacUV5}, by introducing a new flow (1 ml/min) of M63 medium containing 0.8% L-arabinose and Isopropyl β-D-1-thiogalactopyranoside (IPTG) at various concentrations (see below). Once synthesized, T7 RNAPs will bind the T7 promoter and transcribe *48bs* RNAs, which are quickly bound by MS2-GPF molecules and appear under the confocal microscope as bright spots (Figure 1B).

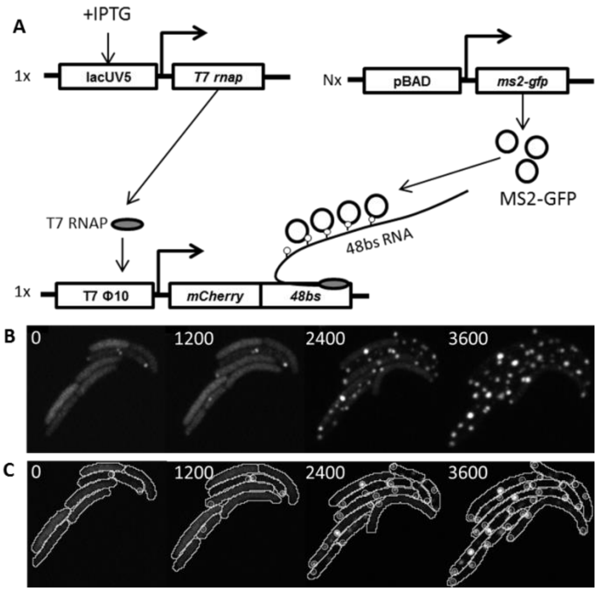


Figure 1. (A) Diagram of the measurement system, depicting the target and reporter genes along with the MS2-GFP tagging process. (B) Confocal microscope images at subsequent time points showing the cells and the MS2-GFP tagged RNA molecules inside. (C) Segmented cells and RNA spots within.

Cells imaging was started at the same time as the introduction of the flow containing IPTG. Images were captured every minute for 2 hours using an inverted microscope Nikon Eclipse (Ti-E, Nikon, Japan). Both confocal images (confocal C2+ scanner connected to LU3 laser system, Nikon) and phase contrast images (DS-Fi2 CCD-camera) were collected.

Examples of confocal images of cells are shown in Figure 1B. Note that, at the end of the time series, the fluorescent background in some cells becomes dimmed due to the produced RNAs having bound most MS2-GFP molecules in the cytoplasm.

C. *Image analysis*

The segmentation of cells and detection of RNA spots were performed by the software “iCellFusion” [16]. It first applies the cell segmentation on phase contrast images using a Gradient Path Labelling Algorithm [17]. Then, it performs the inter-modal image registration between phase-contrast images and the corresponding fluorescence images and exports the segmentation results on fluorescence images. The spot detection was performed as in [18]. Results from the segmentation and spot detection algorithms are shown in Figure 1C.

D. *Data analysis*

The cell-to-cell variability in the kinetics of intake of IPTG, which affects the activation of P_{lacUV5} [19][20], creates extrinsic variability regarding when the first RNA appears in each cell. Since we are only interested in the intrinsic noise of the transcription process, to correct for this, we fit the total spot intensity in each cell over time with an activation function:

$$x(t_{activation}, c, t) = c \times H(t - t_{activation}) \times (t - t_{actiovation}) \quad (1)$$

where t is time, $t_{activation}$ is activation time of T7 when the *48bs* RNA production reaches steady state, c is the mean increment rate of total spot intensity and H is a unit step function. With the function in (1) fitted using least mean squared, we find $t_{activation}$ for each cell. The total spot intensities are then aligned using the inferred $t_{activation}$, so as to compare the kinetics of active T7 promoters in individual cells.

We found by inspection that, at 37^{o}C, in the first ~18 minutes, the mean curve of the aligned total spot intensities can be well fitted with a linear function, indicating that RNA production in most cells reached a steady state after their corresponding $t_{activation}$. After the 18th minute, the mean spot intensity increases with decreasing speed, visibly due to increasing shortage of free MS2-GFP. Therefore, for this condition, we select the data in the first 18 minutes for RNA quantification as in [18][21]. Note also that for different temperatures and IPTG concentrations, the window for RNA quantification differs (data not shown).

E. *qPCR*

Cells grown to OD600 ~0.4 were induced with the appropriate IPTG concentration (5-250 µM) for 1 hour, at the specific temperature (20^{o}C, 37^{o}C and 43^{o}C). Afterwards, cells were fixed with RNAprotect bacteria reagent (Qiagen, Germay), followed by total RNA isolation, DNase I treatment (ThermoScientific, USA) and cDNA synthesis (BioRad, USA). The qPCR master mix contained iQ SYBR Green supermix (Biorad, USA) with primers for the target gene, the T7 RNAP and the reference gene at a final concentration of 200 nM. The primers for the target gene were (Forward: 5' CACCTACAAGGCCAAGAAGC 3' and Reverse: 5' TGGTGTAGTCCTCGTTGTGG 3') for the mCherry region. To quantify the T7 RNAP, the primers used were (Forward: 5' TCCTGAGGCTCTCACCGC 3' and Reverse: 5' GATACGGCGAGACTTGCGA 3'). For the reference gene 16SrRNA, the primers were (Forward: 5' GCTACAATGGCGCATACAAA 3' and Reverse: 5' TT CATGGAGTCGAGTTGCAG 3'). The data from CFX Manager TM Software was used to obtain the relative gene expression and standard error [22].

F. Model of T7 promoter transcription kinetics

To study how the kinetics of the T7 promoter changes with temperature, we assume the modelling strategy of transcription proposed in [23][24][25], derived from both *in vitro* and *in vivo* studies on viral [11][26] and *E. coli* promoters [8][25][27][28][29]. The model of transcription kinetics of T7 promoter is as follows:

$$R + Pr \underset{k'_{cc}}{\overset{k_{cc}}{\rightleftarrows}} Pr_{cc} \xrightarrow{k_{oc}} Pr_{oc} \xrightarrow{\infty} Pr + R + RNA \qquad (2)$$

where R is an active T7 RNAP, Pr is a free promoter, Pr_{cc} is a fully formed closed complex, and Pr_{oc} is a fully formed open complex. The closed complex formation occurs at the rate k_{cc}. Once the closed complex is formed, the promoter can either be unbound by the R at the rate k'_{cc} or undergo open complex formation at the rate k_{oc}. Due to fast promoter escape [30], the low frequency of abortive initiation [6] and the fast rate of elongation of T7 RNAP [5][11][31], we assume that the RNAP and target RNA are released soon after completion of the open complex. Note that this model does not include an ON-OFF mechanism since T7 is a constitutive promoter.

From (2), the mean of the interval distribution (Δt) between consecutive transcription events is:

$$\Delta t(R) = \frac{(k'_{cc} + k_{oc})}{R k_{cc} k_{oc}} + \frac{1}{k_{oc}} = \frac{1+K}{R k_{cc}} + \frac{1}{k_{oc}} = \tau(R) + \tau_{oc} \qquad (3)$$

where R is the abundance of T7 RNAP in the cell, K is ratio between k'_{cc} and k_{oc} indicating the reversibility of the closed complex, $\tau(R)$ is the time for an RNAP to commit to the open complex formation, and τ_{oc} is time for open complex formation. From (3), the production interval $\Delta t(R)$ is a linear function of the inverse of T7 RNAP level ($1/R$), and thus:

$$\tau_{oc} = \Delta t(R = \infty) \qquad (4)$$

With each set of values of $R.k_{cc}$, K, and k_{oc}, we use the Chemical Master Equation (CME) to find the distribution of intervals between consecutive RNA production events, from which the mean rate and noise in transcription are extracted.

III. Results

This section comprises the results, obtained from the measurements, which are presented into three separate subsections.

A. Validation of the construct with the T7 promoter

First, to validate that the T7 promoter inserted in the F-plasmid (Methods) is active, we measured the RNA levels of the T7 RNAP and of the target gene by qPCR for varying IPTG concentrations (which control the expression of T7 RNAP). Results are shown in Figure 2.

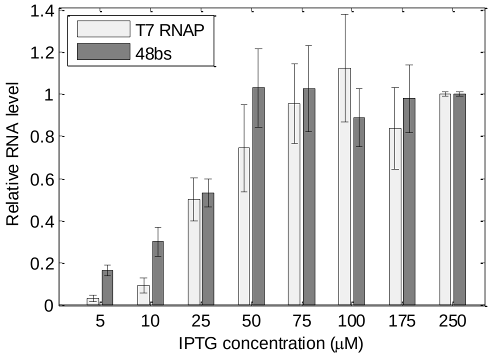


Figure 2. Relative RNA levels of T7 RNAP (light grey) and target gene (*48bs*) (dark grey) at 37^{o}C with varying IPTG concentrations as measured by qPCR. Also shown for each condition are the standard errors from 3 technical replicates.

From Figure 2, first, both the T7 RNAP's and target gene's levels do not increase significantly with increasing IPTG concentrations beyond 100 µM, suggesting that the lacUV5 promoter is fully induced at this concentration. In Figure 2, the data is normalized by the RNA levels at 250 µM IPTG. We validated these measurements, in the case of the target RNA, by observing its production dynamic at 175 µM, 250 µM and 1000 µM IPTG at 37^{o}C under the microscope (via MS2-GFP tagging, Methods).

While we observed changes in the mean activation time of the T7 promoter with changing IPTG concentration (data not shown), we did not observe a significant change in mean transcription rate ($\mu_{\Delta t}$ ~350 s).

Finally, we find an increase in both the T7 RNAP's and target gene's RNA expression with increasing IPTG concentration, demonstrating that both genes are active. Note the close correlation between the activities of the two genes, indicating that the T7 promoter is, as expected, under the control of the T7 RNAP.

B. T7 promoter dynamics at various temperatures

We next observed the transcription dynamics of T7 promoter at different temperatures (within sub-optimal intervals). The IPTG concentration used was 250 µM, in order to ensure that lacUV5 is fully induced in all conditions. Under the microscope, all cells appeared to grow normally, with reduced division rates at lower temperatures. In particular, cells' mean doubling times were 50 min, 60 min and 100 min at 43^{o}C, 37^{o}C and 20^{o}C respectively.

From the RNA numbers over time in individual cells as observed by microscopy at different temperatures, we extracted the mean duration (μ) and coefficient of variation squared (CV^2) of the intervals between consecutive RNA appearances in individual cells as in [18][32]. Results are shown in Table I.

For each temperature, the number of cells observed, the number of samples collected (intervals between consecutive RNAs in individual cells), and the mean and CV^2 of the intervals between consecutive RNA appearances in individual cells are shown. The final column shows the relative RNA levels of T7 RNAP measured by qPCR (normalized by RNA levels at 37°C).

TABLE I. *IN VIVO* TRANSCRIPTION INITIATION DYNAMICS OF THE *T7* PROMOTER AT DIFFERENT TEMPERATURES MEASURED BY MS2-GFP TAGGING OF RNA.

T (°C)	No. Cells	No. Samples	μ (s)	CV^2 (σ^2/μ^2)	Relative T7 RNAp no.
43	150	508	320	0.95	0.86
37	111	311	352	0.85	1
20	68	105	518	0.62	0.46

From Table I, somewhat surprisingly but in agreement with a previous observation by *in vitro* methods [5], the mean length of the RNA production intervals, μ, increases with decreasing temperature. Overall, this indicates that the *in vivo* kinetics of transcription initiation of the T7 promoter is temperature dependent.

Notably, the mean transcription rates *in vivo* are approximately one order of magnitude smaller than those reported from *in vitro* tests [5][11]. This weaker activity in live cells is likely due to the more limited amount of T7 RNAP (bound by the limits in lacUV5's activity) and limited resources (ATP, ribonucleotides, etc.) in the host cells to support the viral transcription process.

Also in Table I, the noise in transcription (as measured by CV^2) decreases with decreasing temperature. A previous work reported a similar result for P_{tetA}, a native promoter of *E. coli* [8].

In addition, in all conditions, the RNA production appears to be a sub-Poissonian process (CV^2<1). This suggests that it consists of multiple rate-limiting steps rather than being dominated by an ON-OFF process [11]. Similar *in vivo* sub-Poissonian dynamics of transcription has been observed in several *E. coli* promoters, native and synthetic, when under full induction [8][28][33].

Overall, the results suggest that the process of transcription initiation of the T7 promoter by the T7 RNAP is similar to that of *E. coli* native promoters.

Meanwhile, from the relative numbers of T7 RNAP as measured by qPCR, we find that unlike when controlling with IPTG concentrations, the kinetics of RNA production of the target promoter T7 no longer follows solely the T7 RNAP numbers, as its production rate is not maximized at 37°C while T7 RNAP numbers are. Therefore, we conclude that the observed changes in the T7 promoter dynamics are due to changes in both the kinetic rates of T7 transcription and in T7 RNAP numbers.

C. *Estimation of kinetic rates of the T7 promoter*

We searched for changes in the underlying kinetics of transcription initiation of the T7 promoter (i.e. in the duration of the closed and open complex formation) with temperature that can explain the changes in the target RNA production with changing temperature.

To quantify how the kinetic rates of T7 promoter evolve with temperature, we followed the strategy proposed in [12] by investigating, for each temperature, how the transcription activity on T7 promoter is affected by the T7 RNAP abundance. This abundance should affect the kinetics of the closed complex formation, but not that of the steps following the closed complex [12].

Here, the T7 RNAP levels, varied by employing different IPTG concentrations (5 μM, 10 μM, 25 μM, 50 μM and 250 μM), and the T7 promoter's activity are measured relatively by qPCR. From these, we infer what would be the relative rate of RNA production given an infinite amount of T7 RNAP in cells (Methods). This rate should correspond to the fraction of time of the transcription initiation process that corresponds to the open complex formation alone [12]. Results for each temperature condition are shown in 'τ plots' in Figure 3.

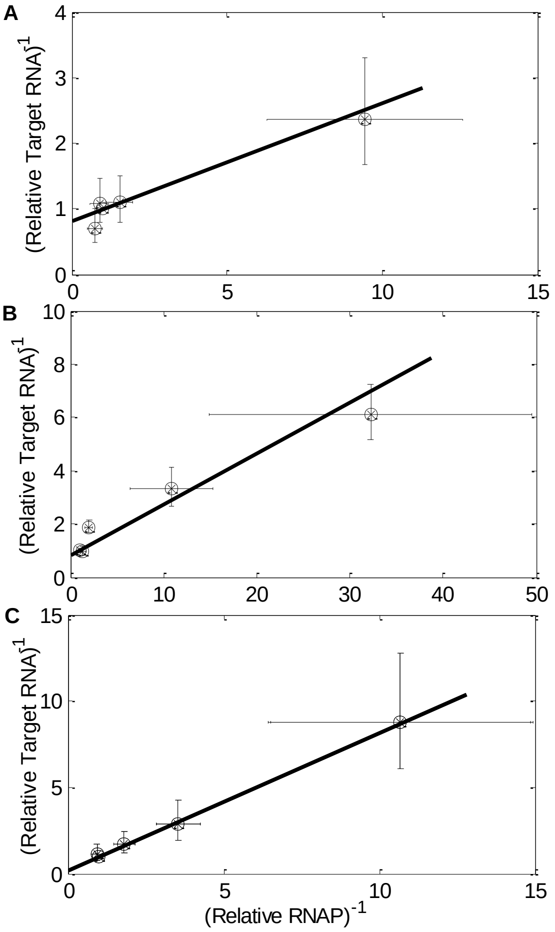


Figure 3. τ plots for T7 promoter activity at different temperatures: (A) 43°C (B) 37°C and (C) 20°C.

In Figure 3, the data is shown relative to the RNA and RNAP levels at 250 μM IPTG. Error bars represent the standard error of the mean (SEM) of the estimate of the inverse of the relative rates of transcription for the target RNA and T7 RNAP in each condition. The lines are Weighted Total Least Squares fits [34]. Errors are calculated including the uncertainty in the 250 µM IPTG condition in the plot (thus removing the error from that point). From Figure 3, the ratio between the inverse of the T7 RNA production rate for infinite T7 RNAP numbers in the cells (R^{-1}=0) equals 0.82 at 43^{o}C, 0.81 at 37^{o}C, and 0.21 at 20^{o}C. These numbers correspond also to the ratio between open complex formation (τ_{oc}) and mean transcription interval (Δt), described in Table I (Methods).

Next, from the ratio (τ_{oc} /Δt), we calculated the rate of open complex formation (k_{oc}). Given the value of k_{oc}, we can find the values of k_{cc} and K to achieve the same mean and noise (with 95% accuracy) of the transcription intervals shown in Table I (Methods). Results are shown in Table II. Shown are the rate of open complex formation (k_{oc}), the reversibility of the closed complex formation (K) and the rate of closed complex formation (R.k_{cc}), given the empirical values of the ratio (τ_{oc}/Δt) extracted from Figure 3.

TABLE II. ESTIMATION OF THE KINETIC RATES OF THE T7 PROMOTER INITIATION PROCESS VERSUS TEMPERATURE.

T (°C)	τ_{oc}/Δt	k_{oc} (s-^{1})	K	R.k_{cc} (s^{-1})
43	0.822	263^{-1}	> 2.00	> 20^{-1}
37	0.808	284^{-1}	1.2±0.5	$(32\pm8)^{-1}$
20	0.206	107^{-1}	<0.11	$(351\pm77)^{-1}$

From Table II, the formation of the open complex, following the T7 RNAP commitment to the closed complex, is faster at 20^{o}C and slower at 37^{o}C and 43^{o}C. This seemingly counterintuitive response suggests that, at higher temperatures, the open complex may be less stable and that, has a consequence, it becomes more reversible to the previous state rather than to committing to the elongation complex.

Namely, the reversibility of the closed complex (K) increases with increasing temperature. At 43^{o}C, the closed complex appears to be highly unstable and T7 RNAP likely binds and unbinds from the T7 promoter several times before being able to form a stable open complex, thus reducing the rate of RNA production. At 37^{o}C, the closed complex appears to be more stable, with a ~50% chance of the RNAP unbinding. At 20^{o}C, the chance of this RNAP unbinding appears to become negligible, likely due to both more stable closed complex formation and faster rate of open complex formation.

Finally, the rate of closed complex formation (R.k_{cc}) becomes slower with decreasing temperature. It should be noted that this rate is highly dependent on lacUV5's strength (which determines R) and therefore is not a property of the natural system. In the future, direct measurements of the relative T7 RNAP protein levels should help revealing the temperature dependence of the closed complex (k_{cc}) of this system.

IV. Conclusion and Future Work

The T7 bacteriophage has only the lytic cycle. Once infecting an *E. coli* cell, its genes transcription is activated and proceeds uninterruptedly until the replication of the viral DNA it achieved [2]. The dynamics of transcription (mean and noise), should therefore play a key role in the success rate of this process. Consequently, for this process to be successful in temperature-fluctuating environments, the transcription process ought itself to be robust to a wide range of temperature conditions.

To assess this robustness, we observed for the first time the *in vivo* transcription initiation kinetics of the T7 promoter at the single RNA level as a function of temperature. Our results suggest that, as temperature decreases, both the mean rate of RNA production and the noise in this process decrease. This somewhat surprising result appears to be made possible by the stabilization of the closed complex formation at lower temperatures.

Our results are, to some extent, similar to those reported for a natural promoter of *E. coli*, P_{tetA}. Namely, its initiation kinetics is also sub-Poissonian, with two rate-limiting steps, the closed and the open complex, whose duration is temperature dependent [8]. However, in P_{tetA}, the noise increases for decreasing temperature.

At the moment, it is unknown what specificities the configuration or composition of the T7 promoter allow this opposite behavior, but this knowledge should be of value to the future engineering of synthetic genes and circuits with robust behaviors at low temperature conditions. From the evolutionary point of view, such noise reduction with lowering temperatures could be associated with the need of the virus for balancing the numbers of phages and capsid proteins more accurately as their total numbers are reduced due to the lowering of the mean production rate [2][7].

In this regard, note from Table I that the relative increase in the interval between RNA productions as temperature decreases from 37^{o}C to 20^{o}C is smaller than the decrease in T7 RNAP numbers (which here are artificially controlled by the LacUV5 promoter). This suggests that, provided a constant number of T7 RNAP for changing temperature, the mean rate of transcription from the T7 promoter will not decrease heavily for decreasing temperature in this range.

In the future, we will employ the system used here and, among other, make use of different promoters controlling the expression of the T7 RNAP so as to, by comparing the various results, isolate the effects of temperature on the T7 promoter alone. Also, we observed that this system is capable of quickly depleting cells from MS2-GFP. This may allow studying the kinetics of binding and unbinding of MS2-GFP to the target RNA as a function of temperature, which might give insights, e.g., on the process by which viral RNAs are protected from the host degradation mechanisms.

Acknowledgment

Academy of Finland [257603, A.R.]; Vilho, Yrjö and Kalle Väisälä fund of Finnish Academy of Science and Letters [S.O.]; TUT President's Graduate Programme

[R.N.], Portuguese Foundation for Science and Technology (FCT/MCTES) [SFRH/BD/88987/2012, LM] and PTDC/BBB-MET/1084/2012, JMF and ASR] and FCT Strategic Program UID/EEA/00066/203 of UNINOVA, CTS [LM and JMF].

REFERENCES

[1] I. Molineux, “The T7 Group,” in *The Bacteriophages*, Second Edi., Oxford: Oxford University Press, pp. 277–301, 2005.

[2] J. Dunn, F. Studier, and M. Gottesman, “Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements,” *J. Mol. Biol.*, vol. 166, 1983, pp. 477-535.

[3] J. L. Oakley, R. E. Strothkamp, a H. Sarris, and J. E. Coleman, “T7 RNA polymerase: promoter structure and polymerase binding,” *Biochemistry*, vol. 18, no. 3, 1979, pp. 528–37.

[4] W. T. McAllister and A. D. Carter, “Regulation of promoter selection by the bacteriophage T7 RNA polymerase in vitro,” *Nucleic Acids Res.*, vol. 8, no. 20, 1980, pp. 4821–4837.

[5] R. Ikeda, A. Lin, and J. Clarke, “Initiation of transcription by T7 RNA polymerase as its natural promoters,” *J. Biol. Chem.*,vol. 267, no. 4, 1992, pp. 2640–2649.

[6] R. A. Ikeda, “The efficiency of promoter clearance distinguishes T7 class II and class III promoters,” *J. Biol. Chem.*, vol. 267, no. 16, 1992, pp. 11322–11328.

[7] M. De Paepe and F. Taddei, “Viruses’ life history: Towards a mechanistic basis of a trade-off between survival and reproduction among phages,” *PLoS Biol.*, vol. 4, no. 7, 2006, pp. 1248–1256.

[8] A.-B. Muthukrishnan, M. Kandhavelu, J. Lloyd-Price, F. Kudasov, S. Chowdhury, O. Yli-Harja, and A. S. Ribeiro, “Dynamics of transcription driven by the tetA promoter, one event at a time, in live Escherichia coli cells,” *Nucleic Acids Res.*, vol. 40, no. 17, 2012, pp. 8472–8483.

[9] F. Arsène, T. Tomoyasu, and B. Bukau, “The heat shock response of Escherichia coli,” *Int. J. Food Microbiol.*, vol. 55, 2000, pp. 3–9.

[10] K. Yamanaka, “Cold shock response in Escherichia coli,” *J. Mol. Microbiol.*, vol. 1, 1999, pp. 193–202.

[11] G. M. Skinner, C. G. Baumann, D. M. Quinn, J. E. Molloy, and J. G. Hoggett, “Promoter binding, initiation, and elongation by bacteriophage T7 RNA polymerase. A single-molecule view of the transcription cycle,” *J. Biol. Chem.*, vol. 279, no. 5, 2004, pp. 3239–3244.

[12] J. Lloyd-Price, S. Startceva, J. G. Chandraseelan, V. Kandavalli, N. Goncalves, A. Häkkinen, and A. S. Ribeiro, “Dissecting the stochastic transcription initiation process in live Escherichia coli,” *DNA Res.*, 2016, in press.

[13] F. W. Studier and B. A. Moffatt, “Use of bacteriophage T7 RNA polymerase to direct selective high-level expression of cloned genes,” *J. Mol. Biol.*, vol. 189, no. 1, 1986, pp. 113–130.

[14] K. Nevo-Dinur, A. Nussbaum-Shochat, S. Ben-Yehuda, and O. Amster-Choder, “Translation-independent localization of mRNA in E. coli,” *Science*, vol. 331, no. 6020, 2011, pp. 1081–1084.

[15] I. Golding and E. Cox, “RNA dynamics in live Escherichia coli cells,” *Proc. Natl. Acad. Sci,*. vol. 101, no. 31, 2004, pp. 11310-11315.

[16] J. Santinha, L. Martins, A. Häkkinen, J. Lloyd-Price, S. M. D. Oliveira, A. Gupta, T. Annila, A. Mora, A. S. Ribeiro, and J. R. Fonseca, “iCellFusion: Tool for fusion and analysis of live-cell images from time-lapse multimodal microscopy,” in *Biomedical image analysis and mining techniques for improved health outcomes, IGI Global*, 2015, pp. 71-99.

[17] A. D. Mora, P. M. Vieira, A. Manivannan, and J. M. Fonseca, “Automated drusen detection in retinal images using analytical modelling algorithms,” *Biomed. Eng. Online*, vol. 10, no. 1, 2011, p. 59.

[18] A. Häkkinen, M. Kandhavelu, S. Garasto, and A. S. Ribeiro, “Estimation of fluorescence-tagged RNA numbers from spot intensities.,” *Bioinformatics*, 2014, pp. 1–8.

[19] J. Mäkelä, M. Kandhavelu, S. M. D. Oliveira, J. G. Chandraseelan, J. Lloyd-Price, J. Peltonen, O. Yli-Harja, and A. S. Ribeiro, “In vivo single-molecule kinetics of activation and subsequent activity of the arabinose promoter,” *Nucleic Acids Res.*, vol. 41, no. 13, 2013, pp. 6544–6552.

[20] H. Tran, S. M. D. Oliveira, N. Goncalves, and A. S. Ribeiro, “Kinetics of the cellular intake of a gene expression inducer at high concentrations,” *Mol. Biosyst.*, vol. 11, no. 9, 2015, pp. 2579–2587.

[21] A. Häkkinen and A. S. Ribeiro, “Estimation of GFP-tagged RNA numbers from temporal fluorescence intensity data,” *Bioinformatics*, vol. 31, no. 1, 2015, pp. 69–75.

[22] K. J. Livak and T. D. Schmittgen, “Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method,” *Methods*, vol. 25, no. 4, 2001, pp. 402–408.

[23] W. McClure, “Rate-limiting steps in RNA chain initiation,” *Proc. Natl. Acad. Sci.*, vol. 77, no. 10, 1980, pp. 5634–5638.

[24] H. Buc and W. R. McClure, “Kinetics of open complex formation between Escherichia coli RNA polymerase and the lac UV5 promoter. Evidence for a sequential mechanism involving three steps,” *Biochemistry*, vol. 24, no. 11, 1985, pp. 2712–2723.

[25] W. R. Mcclure, “Mechanism and control of transcription initiation in prokaryotes,” *Annu. Rev. Biochem.*, vol. 54, 1985, pp. 171–204.

[26] D. K. Hawley and W. R. Mcclure, “Nucleic compilation and analysis of Escherichia coli promoter DNA sequences,” *Nucleic Acids Res.*, vol. 11, no. 8, 1983, pp. 2237–2255.

[27] R. Lutz and H. Bujard, “Independent and tight regulation of transcriptional units in Escherichia coli via the LacR/O, the TetR/O and AraC/I1-I2 regulatory elements,” *Nucleic Acids Res.*, vol. 25, no. 6, 1997, pp. 1203–1210.

[28] M. Kandhavelu, J. Lloyd-Price, A. Gupta, A.-B. Muthukrishnan, O. Yli-Harja, and A. S. Ribeiro, “Regulation of mean and noise of the in vivo kinetics of transcription under the control of the lac/ara-1 promoter,” *FEBS Lett.*, vol. 586, no. 21, 2012, pp. 3870–3875.

[29] E. Bertrand-Burggraf, J. F. Lefèvre, and M. Daune, “A new experimental approach for studying the association between RNA polymerase and the tet promoter of pBR322,” *Nucleic Acids Res.*, vol. 12, no. 3, 1984, pp. 1697–1706.

[30] L. M. Hsu, “Promoter clearance and escape in prokaryotes,” *Biochim. Biophys. Acta - Gene Struct. Expr.*, vol. 1577, no. 2, 2002, pp. 191–207.

[31] S. L. Heilman-Miller and S. A. Woodson, “Effect of transcription on folding of the Tetrahymena ribozyme,” *RNA*, vol. 9, no. 6, 2003, pp. 722–733.

[32] C. Zimmer, A. Häkkinen, and A. S. Ribeiro, “Estimation of kinetic parameters of transcription from temporal single-RNA measurements,” *Math. Biosci.*, vol. 271, 2015, pp. 146–153.

[33] M. Kandhavelu, H. Mannerström, A. Gupta, A. Häkkinen,

J. Lloyd-Price, O. Yli-Harja, and A. S. Ribeiro, "In vivo kinetics of transcription initiation of the lar promoter in Escherichia coli. Evidence for a sequential mechanism with two rate-limiting steps," *BMC Syst. Biol.*, vol. 5, no. 1, 2011, p. 149.

[34] M. Krystek and M. Anton, "A weighted total least-squares algorithm for fitting a straight line," *Meas. Sci. Technol.*, vol. 19, no. 7, 2008, p. 079801.

Effect of Nasal High Flow Therapy on CO_2 Tension - Physico-mathematical Modelling

[1] Cletus F. Adams, [2] Mark Jermy ,[3] P. H. Geoghegan

Centre for Bioengineering
Dept. of Mechanical Engineering
University of Canterbury
New Zealand
email: [1] cleadams.23@gmail.com
[2] mark.jermy@canterbury.ac.nz
[3] patrick.geoghegan@canterbury.ac.nz

C. J. T. Spence

Fisher & Paykel Healthcare
Auckland, New Zealand
email: Callum.Spence@fphcare.co.nz

***Abstract*—The respiratory system of a human embodies complex assembly of tissues and organs (typically internal and external intercostal muscles, diaphragm, lung and rib cage), which are coordinated in a fashion that allows the influx and efflux of air into the airways and lungs. Like all other biological systems, the respiratory system is susceptible to injuries and diseases. Where ventilation has been severely impaired leading to poor gaseous exchange across the lung tissue, biomechanical therapeutic modalities such as continuous positive airway pressure (CPAP) and mechanical ventilators have been prescribed for such patients. Currently, Nasal High Flow therapy (NHFT), a novel ventilation technique has been reported to improve gaseous exchange in both neonates and adults by supplying a constant flow of humidified and warmed air into the lungs. NHFT is presently applied in the management of apnoea of prematurity, respiratory distress syndrome, bronchiolitis, and acute lung injury. In spite of reported success, its mechanisms of action (MOA) are not wholly understood. This work, in terms of relevance, provides some insights into the MOA of NHFT by underscoring the mathematical basis for reported improved gaseous exchange during the administration of NHFT. The mathematical model predictions appreciably agreed with bench-top measurements - indicating 17 % and 24% reduction in end tidal CO_2 concentration upon the respective administration of 30 l/min and 60 l/min NHFT.**

Keywords–nasal high flow therapy; capnography; dead space; alveolar CO_2 tension.

I. Introduction

Capnography is the process of analysing the partial pressure of CO_2 in respiratory gases [1]. Owing to the importance of capnography in current medical practice, medical bodies including the American Association for Respiratory Care [1] and American Society of Anesthesiologists [2] [3] have endorsed it as a method for verifying the correct placement of endotracheal tubes for the provision of respiratory support. Where blood acidosis is clinically diagnosed, capnography may be used to identify the cause by checking for hypercarbia ($PaCO_2 > 45$ mmHg) [4]. Additionally, since CO_2 is transported from the cells (sites of metabolism) to the lungs via the circulatory system, events such as pulmonary and vascular embolism can be detected using capnography [5].

The CO_2 profile recorded during a breathing cycle is unique in terms of morphology for healthy individuals [6][2]. A CO_2 tension profile for a healthy state is shown in Figure 1. Phase I denotes the baseline where inspiration is about to end. The transition stage, Phase II, physically represents a blending of alveolar CO_2-rich air and dead space air. Phase III (alveolar plateau) indicates an almost complete saturation of the airway with alveolar air and peaks at point E_tCO_2, known as end-tidal concentration of CO_2 . Inspiration begins immediately after E_tCO_2 - marking the commencement of phase IV, where influx of fresh atmospheric air speedily dilutes airway air until the baseline value is reached [6][5][7]. Essentially, deviations from healthy state morphology may be suggestive of a pathological condition of the respiratory system [7][3]. It has been mentioned that changes in baseline level, steadiness of the alveolar plateau and slope of transitional portion may be cardinal to the clinical diagnosis of CO_2 rebreathing and pneumothorax [3][5].

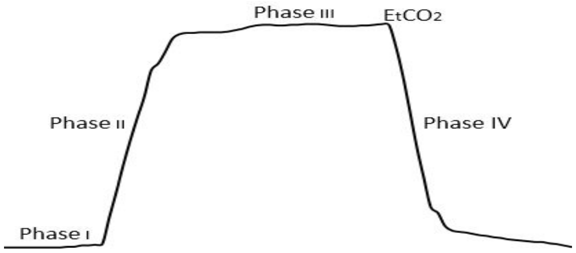


Figure 1. Morphology of a healthy adult capnogram.

Figure 2 shows two polymer models of the upper airway geometry fabricated for experimental work. To fabricate this, a set of computer tomography (CT) images obtained from an adult (age = 44 years and gender = male) was used to reconstruct an *in silico* anatomically representative 3-D model of the upper airway. A detailed description of model making methods has been outlined by Geoghegan et al. [8]. A physical upper airway model, patterned according to the *in silico* model, was built by the use a 3-D printer (fused deposition modelling type), which utilized acrylonitrile butadiene styrene (ABS) as print material.

Nasal high flow therapy (NHFT) involves the administration of humidified and heated air (up to normothermia) at a constant flow rate. Figure 3 is a pictorial representation of the administration of NHFT via a nasal cannula using Fisher & Paykel Healthcare Airvo2 device. It has been reported that NHFT washes the nasopharygeal dead space resulting in an increased proportion of inhaled oxygen content and

Copyright (c) IARIA, 2016.   ISBN: 978-1-61208-488-6

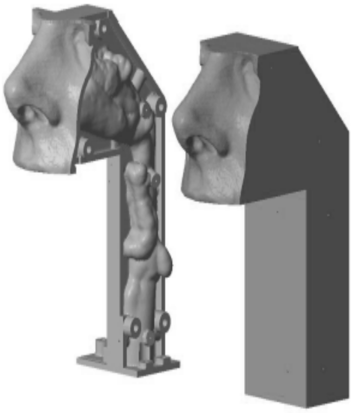


Figure 2. Polymer model of upper airway.

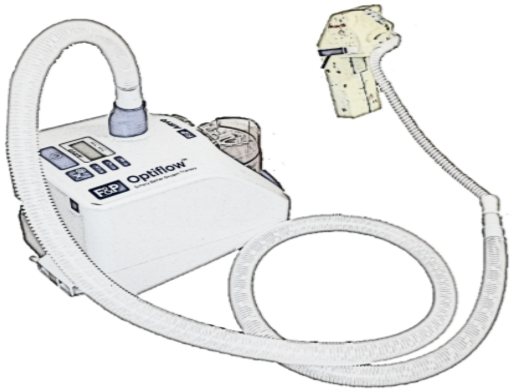


Figure 3. Setup for NHFT administration.

a subsequent improvement in gaseous exchange across the blood-gas barrier [9][10][11]. Arthur et al. [12] pioneered the use of physico-mathematical models to investigate CO_2 fluctuations in the pulmonary system. Their analysis explained the effect of rebreathing on pulmonary CO_2 tension profile. Authors such as Finchman et al. [13] and Milhorn et al [14] used sophisticated compartmental models which included homeostatic response mechanisms, to provide insight into arterial and alveolar reactions to step changes in inhaled CO_2 amounts. Quite recently, Swanson et al. [15] and Benallal et al. [16] applied a two compartment model to shed light on alveolar gas changes during exercise. In the present work, an identical two compartment model has been modified to investigate the amount of CO_2 flushing from the dead space under the influence of NHFT.

In Section 2, formulation of the mathematical model along with the experimental setup is presented. Section 3 comprises of results obtained from both mathematical simulation and bench-top experiment. A discussion of results and concluding observations are presented in Sections 4 and 5 respectively.

II. Method

A. Modelling setup

The model presented in this work is identical to that used by Benallal et al. [16] however the inclusion of a constant volumetric flow term to cater for NHFT distinguishes the present model (Figure 4) from it. In this model, the volume of the respiratory system has been thought of as compartmentalized into two main units, namely dead space unit and alveolar unit. The dead space unit represents the volume of all airway regions where gaseous exchange does not occur whilst the alveolar unit denotes the combined volume of lung and airway regions that exchange gas with pulmonary capillaries. Assumptions made during the formulation of model equations include the following (a) administration of NHFT does not change the dead space volume (time invariant) (b) alveolar volume changes due to NHFT is insignificant (c) there is negligible gas loss across the walls of the dead space.

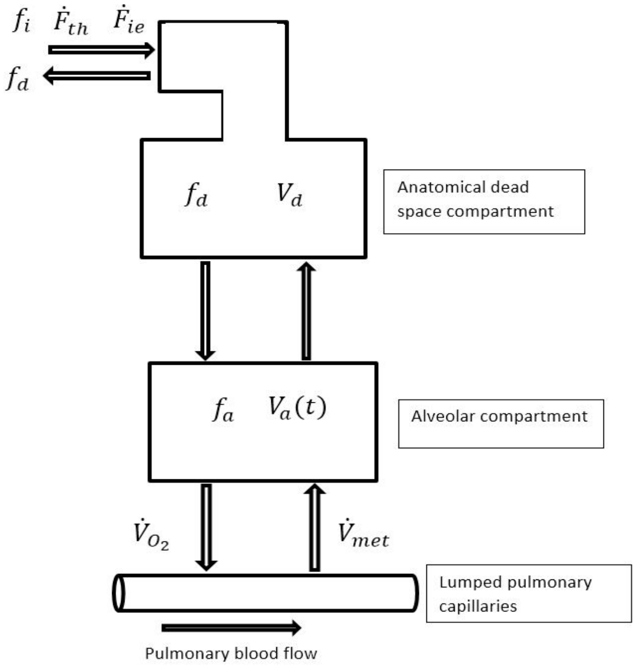


Figure 4. Model representation of respiratory system.

A description of the symbols used in Figure 4 and model equations (1-5) is presented in Table I. Tidal flow data of a healthy male adult (age 23, BMI = 24.6), acquired using a spirometer setup, was used as an input signal to the model. The flow signal was scaled to produce a tidal volume of 500 ml, which is the estimated average for healthy adults [17]. For a resting phase spontaneous breathing of an adult, the functional residual capacity, average metabolic CO_2 production rate ($\dot{V}_{met}$) and O_2 exchange rate ($\dot{V}O_2$) across the blood gas barrier were specified as 2500 ml, 240 ml/min and 300 ml/min respectively as used by several authors [16][15][14][5]. Equations (1) and (2) are descriptive of the CO_2 balance in the dead space unit during inspiration and expiration, respectively. In a like manner, the material balance of CO_2 in the alveolar compartment during the breathing cycle is represented by (4) and (5). The alveolar volumetric transience is described by equation 3. In equations 1, 2, 4 and 5, $\dot{V}_{ie} = \dot{F}_{th} + \dot{F}_i e$.

To solve this 5 non-linear system of equations, a Runge-Kutta (4,5) based solver, ode45, custom-packaged in Matlab (Version 2014b) was employed. The simulation was performed using 50 breathing cycles, which is equivalent to 3.5 minutes of breathing.

TABLE I. MODEL AND EQUATION PARAMETERS.

Symbol	*Parameter*	*Value*
f_d	dead space CO_2 fraction	-
f_a	alveolar CO_2 fraction	-
$\dot{F}_{ie}$	ventilatory flow	-
$\dot{F}_{th}$	NHFT flow	30 and 60 l/min
f_i	inspiratory CO_2 fraction	-
V_a	alveolar volume (time variant)	-
V_d	dead space volume (constant)	150 ml
$\dot{V}_{met}$	metabolic CO_2 influx	240 ml/min
$\dot{V}O_2$	O_2 exchange rate	300 ml/min

$$\frac{df_dCO_2}{dt} = \frac{(\dot{V}_{ie})(f_i - f_d)}{V_d} \tag{1}$$

$$\frac{df_dCO_2}{dt} = \frac{\dot{V}_i e(f_a - f_d)}{V_d} \tag{2}$$

$$\frac{dV_a}{dt} = \dot{V}_{ie} + \dot{V}_{met} - \dot{V}0_2 \tag{3}$$

$$\frac{df_aCO_2}{dt} = \frac{\dot{V}_{ie} f_d + \dot{V}_{met} - f_a(\dot{V}_{ie} + \dot{V}_{met} - \dot{V}O_2)}{V_a} \tag{4}$$

$$\frac{df_aCO_2}{dt} = \frac{\dot{V}_{met} - f_a\dot{V}_{ie} + f_a(\dot{V}_{ie} + \dot{V}_{met} - \dot{V}O_2)}{V_a} \tag{5}$$

A sensitivity study was performed to evaluate the independent contribution of the model parameters to E_tCO_2. The resolution of the sensitivity scale was limited to a 1 % change in E_tCO_2 because a 1 % change significantly affects amount of CO_2 flushed when NHFT is applied. Findings for a 10 % increment on each parameter value is presented in Table II. E_tCO_2 was found to be most responsive to initial value of alveolar CO_2 partial pressure ($PACO_2$) and metabolic CO_2 production rate. Variation in initial dead space CO_2 fraction produced the least change in E_tCO_2.

TABLE II. SENSITIVITY STUDY.

Parameter	*10% value increment*	Change in E_tCO_2	*Comment*
initial $PACO_2$	5.86 %	7.71%	very sensitive
$\dot{V}_{met}$	275 ml	1.06 %	sensitive
initial f_i	0.048 %	0.05 %	less sensitive
FRC	2750 ml	0.12 %	less sensitive
V_d	165 ml	0.43 %	less sensitive
$\dot{V}O_2$	330 ml/min	0.15%	less sensitive

B. Experimental work

The same physiological flow signal as specified for the mathematical model was programmed into a LabVIEW (Version 8.6) application that operates a pulsatile pump (Figure 5). See component labelling of Figure 5 for setup description. The pulsatile pump (2) connects to a 3-D printed upper airway model via a tubing. NHFT is administered through a nasal cannula (4) by means of a Fisher & Paykel Airvo2 device (3). In performing CO_2 experiments, CO_2 was metered at a bleed rate of 250 ml/min into the pump chamber (piston barrel) , allowing a back pressure of 101.3 KPa in the CO_2 source (1). Measurement of CO_2 concentration at the trachea opening of the airway model is performed using a capnograph (5). After a minute, a steady CO_2 profile peaking at E_tCO_2 of 5.2% was observed on the computer (6) connected t the capnograph.

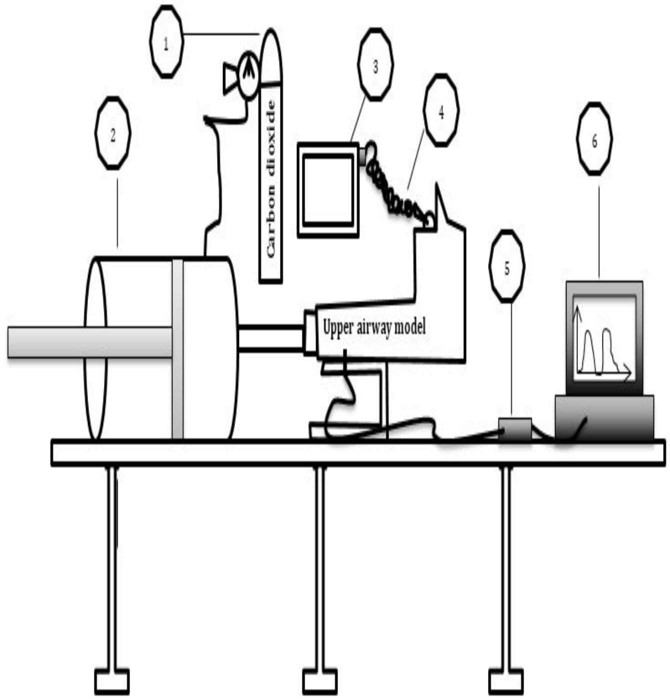


Figure 5. Complete experimental setup

At this point, recording of CO_2 data was performed for 3 minutes followed by 30 l/min NHFT - administered using Fisher & Paykel Healthcare Airvo2 device. Two minutes was allowed for equilibration after which CO_2 recording proceeded for 3 more minutes. An identical procedure was repeated for 60 l/min NHFT, after residual effects from the previous experiment had been eliminated by shutting off CO_2 bleed valve whilst the pump was in operation - allowing CO_2 levels to plummet to near atmospheric values.

III. RESULTS

On Figure 6, plot A shows fractional CO_2 profile in the alveolar unit during a 50 breathing cycle simulation. Transience in breath-by-breath CO_2 fraction is observable up to 60 seconds and then a steady state is reached. The flow wave used for both simulation and experiment had a period of 4.3 seconds - inspiratory time being 2 seconds and expiratory time being 2.3 seconds. Plot B (Figure 6) represents two cycles of zero-therapy (ZT) ventilatory flow patterns superimposed on corresponding alveolar CO_2 fraction profile at steady state. The distortions on the flow wave at the interface between inspiration and expiration are artefacts introduced by the limited resolution of the spirometer flow device, which is directly linked to the observable local perturbations in fractional alveolar CO_2 profile at the flow transition points. The alveolar CO_2 fraction is seen to rise up to 0.3 seconds into inspiration. Furthermore, over the respiratory cycle, alveolar CO_2 fraction fluctuates between 4.9 % and 5.3 % with a mean value corresponding to 5.1 % .

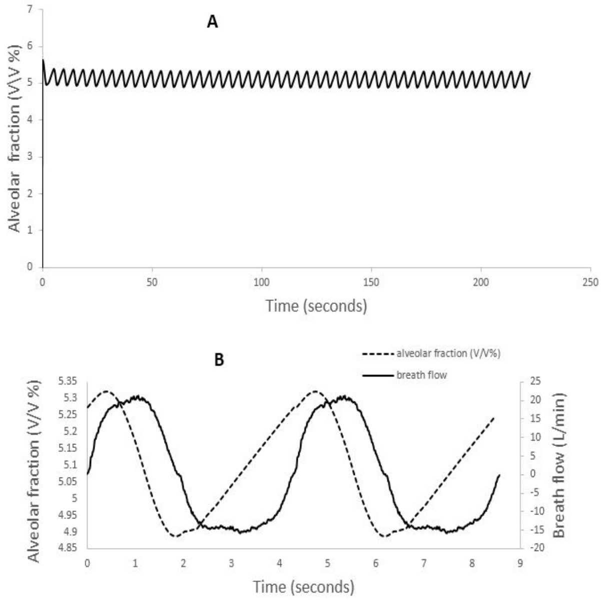


Figure 6. Variation of alveolar CO_2 tension with time.

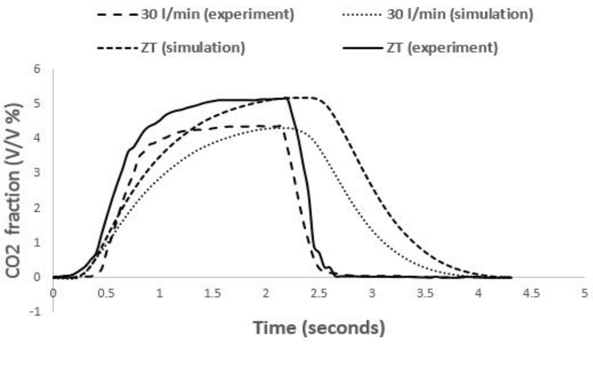


Figure 7. Variation of deadspace CO_2 tension with time.

Figure 7 shows a single respiratory cycle plot of experimental and simulated dead space CO_2 fraction profiles for both ZT and 30 l/min NHFT conditions. When 30 l/min NHFT was applied , dead space E_tCO_2 fell from 5.2 % to about 4.3 % (17% decrement). It is observable from Figure 8 that for 60 l/min NHFT, model simulation under predicts experimental E_tCO_2 by a margin of 8%. In morphology, the simulated dead space CO_2 fraction does not match very well with the experimental measurements. This may be due to the inherent perfect mixing of CO_2 in the dead space compartment (not realistic) which can be made physiological by sub-dividing the dead space volume into several control volumes. The area under the plots (Figures 7 and 8) contain information about dead space CO_2 volume. In quantitative terms, mathematical simulation over-predicts the combined inspiratory and expiratory CO_2 volume for both ZT and 30 l/min NHFT by 8 %. On the same note, CO_2 volume over the respiratory cycle for 60 l/min NHFT is over-predicted by 4%.

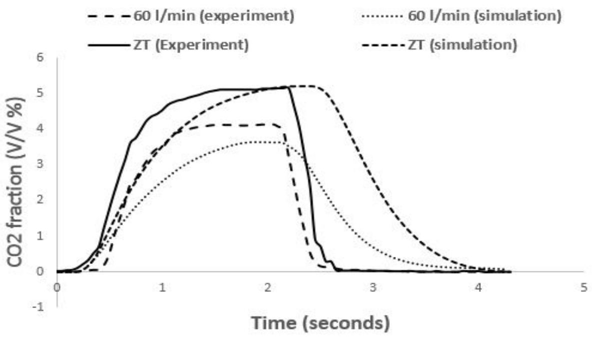


Figure 8. Variation of deadspace CO_2 tension with time.

IV. DISCUSSION

Generally, the pulsatile pump setup differs from *in vivo* conditions in two aspects, i.e., compliance mismatch (high rigidity of piston barrel) and absence of O_2 exchange mechanism. Since lung elasticity does not enter into the model equations (1-5) and the pulsatile pump is able to deliver the expected tidal volume, the influence of tissue elasticity is eliminated. Given that the ratio of exchanged O_2 volume to tidal volume is quite small (1:25), the present experimental results may therefore considerably approximate the results for the case where an O_2 absorption unit is included in the pulsatile pump assembly. Simulation of no-oxygen condition, however, has indicated a 5 % reduction in expired CO_2 volume.

E_tCO_2 obtained via simulation is comparable to corresponding experimental values for ZT and 30 l/min NHFT. For all cases, slopes of capnogram phases II and IV as predicted by simulation are lesser compared to experimental observations (Figures 7 and 8). This may be due to the model compartmental configuration, which allows for full CO_2 mixing at all times whilst in the experimental setup (same as *in vivo*) there is an established CO_2 front which takes a finite time to travel along the airway. For 60 l/min NHFT, the model under-predicts experimental E_tCO_2 by a change of 8 %. This may be considered as a pronounced effect of the afore-mentioned full CO_2 mixing (not physiological), being heightened by high levels of NHFT.

ZT E_tCO_2 of 5.2 % obtained from simulation corresponds to a tension of 38 mmHg . This value is within physiologic range (37 - 44 mmHg) for healthy adults [18]. The mean alveolar CO_2 fraction of 5.1 % is equivalent to alveolar CO_2 tension ($PACO_2$) of 39 mmHg, which agrees with the reported physiologic range of 35 mmHg to 45 mmHg [19][20][21]. The rise in alveolar CO_2 fraction for about 0.3 seconds into inspiration is suggestive of rebreathing of CO_2 from the dead space unit [12].

Over a spontaneous breathing cycle, mean arterial CO_2 pressure ($PaCO_2$) is approximately the mean of $PACO_2$. It has been reported that the lungs can sufficiently engage in oxygen exchange when it is in the state of apnoea, however in this state, CO_2 diffusing across the blood-gas barrier accumulates in the lung and can potentially cause blood acidosis ($PaCO_2 >$ 45 mmHg). Several studies have mentioned an increase in ventilatory rate in severe COPD patients experiencing hypercapnic events ($PACO_2 > 45$ mmHg), which has been interpreted as a homeostatic reaction to maintain $PACO_2$ in physiological range [14]. The mean $PACO_2$ output from the present model may potentially provide some insights into expected $PaCO_2$ for COPD related hypercapnic events under NHFT conditions. On Figure 7, there is a 17 % fall in $E_t\mathrm{CO}_2$ when 30 liters/min NHFT is applied. The washout volume of CO_2 during expiration, as predicted by the model is 2 ml - representing 15 % of the total expired CO_2 volume for spontaneous breathing. Experimentally, administration of 60 l/min NHFT yielded 24% reduction of ZT $E_t\mathrm{CO}_2$ though the simulated results over-predicts this change. It is however noticeable that the amount of reduction in $E_t\mathrm{CO}_2$ is dependent on flow rate at which NHFT is administered.

Spence et al. [11] used particle imaging velocimetry (PIV) techniques to investigate flow distribution in a silicone upper airway model under NHFT conditions. Their conclusion was that recirculation currents observed in the nasopharynx resulted in CO_2 flushing. Chatila et al. investigated exercise tolerance of severe COPD patients and concluded that NHFT leads to a gain in exercise endurance attributable to increased oxygenation. Flushing of CO_2 may increase the proportion of alveolar ventilation in reference to minute ventilation, thereby boosting oxygen exchange across the blood-gas interface in the lungs [10]. In the light of these reports, the presented two-compartment model , in spite of the outlined limitations, appreciably predicts changes in $E_t\mathrm{CO}_2$ of the bench-top model capnogram for both zero-therapy and NHFT conditions.

V. Conclusion

The results from this work show that it is possible to make appreciably satisfactory predictions of the end tidal CO_2 fraction, alveolar CO_2 tension and flushed CO_2 volume for nasal high flow therapy conditions though in terms of morphology, results from the presented two-compartment model show a width-wise disparity from physiologic CO_2 profiles.

Acknowledgment

The authors would like to thank UC Doctoral Scholarship and Fisher & Paykel Healthcare for their support.

References

[1] B. K. Walsh, D. N. Crotwell, and R. D. Restrepo, “Capnography/Capnometry During Mechanical Ventilation: 2011,” *Respiratory Care*, vol. 56, pp. 503–509, Apr. 2011.

[2] M. B. Jaffe, “Volumetric Capnography, The Next Advance in CO2 Monitoring,” *Respironics Inc (Critical Care)*, 2006.

[3] J. E. Thompson and M. B. Jaffe, “Capnographic waveforms in the mechanically ventilated patient,” *Respiratory care*, vol. 50, no. 1, pp. 100–109, 2005.

[4] H. Soleimanpour, “Capnography in the Emergency Department,” *Emergency Medicine: Open Access*, vol. 02, no. 09, 2012.

[5] J. O. Den Buijs, L. Warner, N. W. Chbat, and T. K. Roy, “Bayesian tracking of a nonlinear model of the capnogram,” in *Engineering in Medicine and Biology Society, 2006. EMBS’06. 28th Annual International Conference of the IEEE*, pp. 2871–2874, IEEE, 2006.

[6] G. Tusman, A. Scandurra, S. H. Bhm, F. Suarez-Sipmann, and F. Clara, “Model fitting of volumetric capnograms improves calculations of airway dead space and slope of phase III,” *Journal of Clinical Monitoring and Computing*, vol. 23, pp. 197–206, Aug. 2009.

[7] B. You, R. Peslin, C. Duvivier, V. D. Vu, and J. P. Grilliat, “Expiratory capnography in asthma: evaluation of various shape indices,” *European Respiratory Journal*, vol. 7, no. 2, pp. 318–323, 1994.

[8] P. H. Geoghegan, N. A. Buchmann, C. J. T. Spence, S. Moore, and M. Jermy, “Fabrication of rigid and flexible refractive-index-matched flow phantoms for flow visualisation and optical flow measurements,” *Experiments in fluids*, vol. 52, no. 5, pp. 1331–1347, 2012.

[9] S. et al., “Beneficial effects of humidified high flow nasal oxygen in critical care patients: a prospective pilot study,” *Intensive care medicine*, vol. 37, no. 11, pp. 1780–1786, 2011.

[10] K. Dysart, T. L. Miller, M. R. Wolfson, and T. H. Shaffer, “Research in high flow therapy: Mechanisms of action,” *Respiratory Medicine*, vol. 103, pp. 1400–1405, Oct. 2009.

[11] C. J. T. Spence, N. A. Buchmann, and M. C. Jermy, “Unsteady flow in the nasal cavity with high flow therapy measured by stereoscopic PIV,” *Experiments in fluids*, vol. 52, no. 3, pp. 569–579, 2012.

[12] A. B. Chilton and R. W. Stacy, “A mathematical analysis of carbon dioxide respiration in man,” *The bulletin of mathematical biophysics*, vol. 14, no. 1, pp. 1–18, 1952.

[13] W. F. Fincham and F. T. Tehrani, “A mathematical model of the human respiratory system,” *Journal of biomedical engineering*, vol. 5, no. 2, pp. 125–133, 1983.

[14] H. T. Milhorn, R. Benton, R. Ross, and A. C. Guyton, “A mathematical model of the human respiratory control system,” *Biophysical Journal*, vol. 5, no. 1, pp. 27–46, 1965.

[15] G. D. Swanson and D. L. Sherrill, “A model evaluation of estimates of breath-to-breath alveolar gas exchange,” *Journal of Applied Physiology*, vol. 55, no. 6, pp. 1936–1941, 1983.

[16] H. Benallal and T. Busso, “Analysis of end-tidal and arterial PCO2 gradients using a breathing model,” *European journal of applied physiology*, vol. 83, no. 4-5, pp. 402–408, 2000.

[17] M. P. Hlastala, “A model of fluctuating alveolar gas exchange during the respiratory cycle,” *Respiration physiology*, vol. 15, no. 2, pp. 214–232, 1972.

[18] J. Ritchie, A. Williams, C. Gerard, and H. Hockey, “Evaluation of a humidified nasal high-flow oxygen system, using oxygraphy, capnography and measurement of upper airway pressures,” *Anesth Intensive Care*, vol. 39, no. 6, pp. 1103–1110, 2011.

[19] E. D. Robin, R. D. Whaley, C. H. Crump, and D. M. Travis, “Alveolar Gas Tensions, Pulmonary Ventilation and Blood pH During Physiologic Sleep in Normal Subjects1,” *Journal of Clinical Investigation*, vol. 37, pp. 981–989, July 1958.

[20] J. N. Mills, “Changes in alveolar carbon dioxide tension by night and during sleep,” *The Journal of Physiology*, vol. 122, pp. 66–80, Oct. 1953.

[21] C. J. Allen, N. L. Jones, and K. J. Killian, “Alveolar gas exchange during exercise: a single-breath analysis,” *Journal of Applied Physiology*, vol. 57, no. 6, pp. 1704–1709, 1984.

Using a Synthetic Probe to Study the Robustness of the Segregation Process of Protein Aggregates in *Escherichia coli*

Andre S. Ribeiro, Samuel M.D. Oliveira
Laboratory of Biosystem Dynamics, Department of Signal Processing, Tampere University of Technology
Tampere, Finland.
e-mail: andre.ribeiro@tut.fi, samuel.oliveira@tut.fi

***Abstract*—Even though the processes of protein production and folding are not immune to errors, *Escherichia coli* lineages are capable to maintain a stable cell lineage, provided viable environmental conditions. One of the internal processes that makes this possible consists of segregating unwanted protein aggregates to the cell poles by nucleoid exclusion, which, combined with cell divisions, generates asymmetries in the aging process of the population, with some individuals aging faster while others exhibit rejuvenation. A recent study showed that this process is not immune to sub-optimal temperature conditions due to increased cytoplasm viscosity, which weakens the anisotropy in aggregate displacements at the nucleoid borders. This was made possible by the usage of a synthetic fluorescent probe, consisting of a RNA sequence with multiple binding sites for the MS2-GFP synthetic protein, which can be tracked in time-lapse microscopy images. Here we provide a description of the findings from these measurements and investigate with an *In Silico* model the consequences in the context of cell lineages.**

Keywords-segregation; polar retention; protein aggregates; cold temperature conditions; synthetic probes; cell lineages.

I. INTRODUCTION

Escherichia coli are able to segregate unwanted protein aggregates to the cell poles by nucleoid exclusion. This process is essential for cell lineages to generate cells that are free from aggregates. Such ‘rejuvenated’ cells have been shown to exhibit faster division time than ‘older’ cells, where aggregates accumulate at, and are thus essential for the maintenance of vitality of the lineages [1].

The exclusion of aggregates from midcell is made possible by the presence of the nucleoid at midcell, which causes anisotropy in the dynamics of the aggregates that generates the preference for polar localization [2].

Recent studies, making use of a synthetic fluorescent probe that allows observing the processes of segregation and retention with single aggregate sensitivity, showed that at lower temperatures, the degree of viscosity of the cytoplasm increases, which hampers the anisotropy [3]. These synthetic probes are ideal in that they behave similarly to natural aggregates, have long life-times with highly stable fluorescence levels, and are robust to photobleaching [2]. In addition, and contrary to natural aggregates, the synthetic aggregates have all the same fluorescence level and do not interact with one another or with other cellular components, facilitating their counting from the images.

Here, based on the empirical data that was obtained by observing cells containing these probes and placed in environments at different temperatures while under microscope observation, we investigate the long-term consequences to future cell generations of the temperature-dependence of the aggregate segregation and subsequent polar retention processes, following the occurrence of sub-optimal conditions.

II. PREVIOUS FINDINGS

In our previous study [3], we compared the efficiency with which aggregates are segregated to and retained at the cell poles by the nucleoids in optimal and in sub-optimal temperature conditions.

Observing cells with one nucleoid, and by probing the positioning of both nucleoids and aggregates, we found that at lower temperatures the aggregates are not preferentially located at the poles. Results are shown in Table I.

From the table, note how the relative concentration of aggregates at the poles is close to 1 (corresponding to uniform distribution along the major cell axis) for low temperatures. Meanwhile, at the higher temperatures, it is much larger than 1. Note also how, according to the Kolmogorov-Smirnov test, the behavioral change is statistically significant between 24 and 37 degrees.

TABLE I. AGGREGATES AT THE POLES

T (°C)	**Concentration of Aggregates at the poles**		
	Relative nucleoid length	***Relative concentration of Aggregates at poles***	***P value of KS test***
10	0.63	1.32	
24	0.56	1.09	0.11
37	0.53	1.86	< 0.01
43	0.47	1.79	0.05

Next, in cells with two nucleoids, the concentration of aggregates in between the nucleoids was measured. From Table II, the relative concentration of aggregates in between nucleoids in cells close to division decreases significantly as the temperature increases [3]. Thus, one can conclude that the relative concentration of the aggregates at the poles is increasing with increasing temperature.

TABLE II. AGGREGATES IN BETWEEN NUCLEOIDS

T (°C)	**Concentration of Aggregates in between nucleoids in cells close to dividing**		
	Relative nucleoids length	***Relative concentration of Aggregates in between nucleoids***	***P value of the permutation test***
10	0.75	0.85	
24	0.68	0.78	< 0.01
37	0.72	0.69	< 0.01
43	0.70	0.68	< 0.01

III. RESULTS AND DISCUSSION

Note that, in division, while the aggregates at the poles will remain at the old pole of the cells of the new generation, those at midcell will be at the new poles. As such, changes in the fractions at midcell prior to division should affect the distributions of aggregates in individual cells of future generations. In particular, we hypothesized that at lower temperatures, as the ability of cells to exclude the aggregates to the poles is significantly reduced, future cell generations will have more homogenous distributions of unwanted protein aggregates, which is expected to hamper the rejuvenation process of the lineage. To validate our hypothesis, we developed a simple stochastic model.

In this model, we start with a cell near division (generation 0), with 200 aggregates whose location (in between nucleoids or at the poles) is defined by the empirical values in Table II. Then, the cell divides and the aggregates are placed in the 'old' and 'new' pole of the two daughter cells, in accordance with their location in the mother cell prior division (i.e., in the pole or in between nucleoids, respectively). In this regard, the aggregates that were at midcell were placed randomly in either daughter cell. Note that, at this stage (i.e., generation 1) all aggregates are at the poles in all cells. Finally, these daughter cells also divide, producing four cells (generation 2). Two of these cells will inherit the original poles of the mother cell, while the remaining ones will inherit only poles generated during the two division processes. Meanwhile the partitioning processes of the aggregates follow the same rules as before.

Using this model, we compared the outcomes at different temperatures, by setting different concentrations of aggregates in between nucleoids of the original mother cell in accordance with the empirical values in Table II. Namely, for each condition, we obtained the mean and standard deviation of the numbers of aggregates in individual cells in the last generation from 10.000 independent simulations. Results are shown in Table III.

TABLE III. AGGREGATES IN THE LAST GENERATION

T (°C)	**Distributions of aggregates in cells of the last generation**	
	Mean number of Aggregates per cell	***Standard deviation of the number of aggregates per cell***
10	50	10.5
24	50	12.2
37	50	16.7
43	50	17.4

From Table III, first, as expected, temperature does not affect the mean number of aggregates in each cell (50 as we started with 200 and 2 rounds of division took place). Also, we find that as temperature increases (and thus, the relative concentration of aggregates at the poles decreases), as expected, the variability in the aggregates numbers in cells of future generations increases. The decrease at lower temperatures, most likely, will result in the hampering of the rejuvenation process of the lineage in these conditions.

We conclude that the effects of lower temperatures at the single cell level have long term consequences in the functioning of cell lineages aging and rejuvenation processes.

REFERENCES

[1] A.B. Lindner, R. Madden, A. Demarez, E.J. Stewart, F., and Taddei (2008) "Asymmetric segregation of protein aggregates is associated with cellular aging and rejuvenation". Proc Natl Acad Sci USA 105: 3076–3081.

[2] A. Gupta, J. Lloyd-Price, R. Neeli-Venkata, S.M.D. Oliveira, and A.S. Ribeiro (2014) In vivo kinetics of segregation and polar retention of MS2-GFP-RNA complexes in Escherichia coli. Biophys J 106: 1928–1937.

[3] S.M.D. Oliveira, R. Neeli-Venkata, N. Goncalves, J.A. Santinha, L. Martins, H. Tran, J. Mäkelä, A. Gupta, M. Barandas, A. Häkkinen, J. Lloyd-Price, J.M. Fonseca, and A.S. Ribeiro (2016) "Increased cytoplasm viscosity hampers aggregate polar segregation in Escherichia coli." Mol. Microbiol. 99(4), 686–699.

Computational Analysis of the Linear Motif Mediated Subversion of the Human Protein Synthesis Machinery

Andrés Becerra, Victor A. Bucheli, Pedro A. Moreno,
School of Systems Engineering and Computer Science
Faculty of Engineering
Universidad del Valle
Email: {andres.becerra.sandoval, victor.bucheli, pedro.moreno}@correounivalle.edu.co

***Abstract*—We argue that virus-host interactions mediated by short linear motifs can be used to analyze common viral attack strategies. In this direction we develop a method for predicting interactions between human protein-synthesis machinery and viral proteins mediated by linear motifs in order to study common protein-synthesis subversion strategies. The method consists in finding viral instances of host linear motifs. We filter these instances by conservation in viral sequences, location in protein disordered regions and scarcity in randomized protein sets. With the filtered motifs we deduce virus-host interactions using the motif-domain associations in the Eukaryotic Linear Motifs (ELM) database. We validate the results against the Linear Motif mediated Protein Interaction Database (LMPID) and obtain a network of interactions between the human protein-synthesis machinery proteins and viruses influenza AH1N1, Dengue1, Ebola, MERS, Rotavirus, WestNile, and Zika.**

Index Terms*—*virus; host; protein; interaction; short; linear; motif; prediction; eukarya; protein-synthesis; subversion

I. Introduction

The objective of this paper is to present a work in progress for predicting virus-host protein-protein interactions (VHPPIs) between several viruses and the human protein-synthesis machinery (HPSM) mediated by short linear motifs (SLiMs). Our motivation to conduct this study is to unveil common viral strategies to subvert protein translation.

There is no known virus that encodes a complete protein-synthesis system. This implies that viruses are forced to use the HPSM to translate their messenger RNA (mRNA) into products: microRNA (miRNA), peptides and proteins. Viruses must control the HPSM and disrupt innate host defense systems capable of disabling protein synthesis [1].

The control and disruption of host signaling pathways is conducted through VHPPIs like the ones DNA viruses engage with the PI3K–Akt–mTOR pathway (phosphatidylinositol 3-kinase-Akt-mammalian target of rapamycin) [2]. The consequences of VHPPIs can be as significant as the shutdown of host protein synthesis done by Rotavirus protein NSP3 [3].

There are open questions about the viral control of the HPSM like the role of phosphorylation in activity of protein eIF4E and how viral mRNA is preferentially translated [4]. These questions could be investigated with a systems biology approach.

Systems biology uses VHPPIs for the discovery of infection mechanisms [5]. However, the scarcity of virus-host PPIs with experimental evidence is an obstacle to system approaches [6]. This lack of data has encouraged the development of VHPPI prediction methods.

VHPPI prediction methods have been mostly based on machine learning classifiers like random forests [7] and support vector machines [8]–[10]. Most of these classifiers use protein sequences and other features like gene ontology (GO) function and gene expression as inputs to infer the interactions because structural data for viral proteins is scarce [11].

There are other prediction methods like information integration [12], asking experts [13], literature mining [14] and focusing on PPIs mediated by SLiMs [15].

We focus our study on SLiM-mediated interactions. The inference of this kind of interactions is guided by biological hypotheses like the conservation of motifs and localization of motifs in protein disordered regions.

Recently, the role of SLiMs has been studied in a wide set of viruses. These pathogens use SLiMs extensively as means to interact with host proteins [16]. Human proteins targeted by viruses have a high number of SLiMs [17].

If virus-host PPIs are divided in domain-motif interactions (DMI) and domain-domain interactions (DDI), DMI are the predominant ones. Furthermore, DMI are used by several viruses while DDI are virus-specific [17]. This supports our use of SLiMs as a way to find common viral subversion strategies.

Eukaryotic organisms use SLiM instances as as mechanisms to tune the regulation of multi-protein complexes. These instances are short, allowing viruses to evolve them de novo and retain them if they are useful to disrupt o subvert a host protein complex [18]. If the SLiM instances are encoded in different host genomic locations, the viral evolution of SLiM instances is robust in a virus-host coevolutionary arms race [19].

SLiMs are represented computationally as regular expressions like **PxIxIT** for the PCNA-binding PIP box motif of Flap endonuclease 1 (FEN1), where the **x** stands for any amino acid. A SLiM instance is a subsequence in a protein that

ISBN: 978-1-61208-488-6

matches the regular expression, like **PRIEIT** in the human protein NFATC1 [18].

Viral instances of regular expressions representing host SLiMs can be found by chance. For this reason, filtering methods of viral instances must be implemented.

Evans et al. find that HIV-1 instances of human SLiMs are significantly conserved in HIV-1 proteins [15]. They propose a criterion to filter SLiMs if they are conserved above a 70% in the available viral sequences.

Hagai et al. propose two criteria to filter SLiMs: the first is based on SLiM location in protein disordered regions and the second in SLiM rarity in a big set of randomized (chimeric) proteins [16]. A SLiM is judged as rare, or hard to form by pure chance, if it is counted in less than a fraction of the sequences in the set of randomized proteins, e.g. 1% of the sequences.

We implement a combinaion of filtering criteria: 1) conservation, 2) location in disordered region and 3) difficulty to find the SLiM by chance. Our contribution is computational, the development of a platform to predict SLiM-mediated interactions that can be generalized to other subsystems and hosts. The clear limitation of our platform is our reliance on the ELM motifs database that makes the method appropriate for eukaryotic hosts only.

The organization of this paper is as follows. In Section III we present the results or our work. In Section II we describe the computational methods used and the Section IV contains the conclusion and directions for further research.

II. METHODS

Algorithmically, the prediction of SLiM-mediated VHPPI we propose is divided into: 1) collecting regular expressions representing SLiMs in the HPSM proteins, 2) finding instances of the collected SLiMs in viral proteins, 3) filtering the instances, 4) infer VHPPIs using SLiM instances in viral proteins and counter domains (CDs) in host proteins.

In order to complete the phases enumerated above we: 1) use the ELM database as a catalog of SLiMs [22], 2) implement software to find SLiM instances in protein sequences, 3) develop three filtering criteria described below, and 4) use the SLiM-domain associations in the ELM database together with Pfam protein-domain associations to infer protein-protein interactions [23].

A. Sequences and disorder prediction

HPSM proteins are taken from reference [1] and the Ribosomal Protein Gene database (RPG) [24]. All proteins are mapped to Uniprot identifiers in order to match protein entries in the ELM database [25].

Viruses are selected for their availability of protein sequences in the National Center for Biotechnology Information (NCBI) viral genomes resource: Dengue virus, West Nile virus, Middle East Respiratory Syndrome coronavirus (MERS), Ebolavirus, Rotavirus and Zika virus [26]. For influenza we choose type A, subtype H1N1, for Dengue we choose type 1, for Ebola the Zaire species.

We download every viral protein for each virus. For all viruses, we set the parameter region as any, the parameter “Full-length sequences only” to true and the parameter host as human. For Influenza AH1N1 proteins we set the parameter collapsed sequences, with the exception of proteins M1,M2 and NS2 for which the collapsed sequences option was deactivated.

For viruses Dengue type 1, West Nile and Zika the NCBI viral genomes resource gives the complete polyprotein sequence that must be manually cleaved. The viral reference genomes stored in Genbank files are computationally translated to protein sequences that are used as reference for cleaving the polyprotein into viral proteins.

Disorder prediction is computed with IUPred [27]. We develop a wrapper to call IUPred on each protein sequence to compute the disordered regions with a sliding-window algorithm proposed by Hagai et al. [16].

B. SLiMs

We download all the SLiMs, instances and interactions from the ELM database and create a SLiM dictionary indexed by the ELM unique identifiers containing the SLiM name, class and its full regular expression [28]. We develop scripts to compute for a set of sequences: the number of sequences with a given SLiM, the number of SLiM instances per protein, the number of SLiMs conserved above a percentage of sequences (set C) and the number of SLiMs in disordered regions (set D).

We write a script to randomize viral sequences. For each sequence in a protein file, we create 1000 shuffled versions randomizing the residues located in disordered regions of the sequence, as computed with IUPred. Then, we counted the rare (scarce) SLiMs in these shuffled data sets, i.e. the SLiMs that are found in 1% of the randomized sequences or less (set R).

Finally, we use the scripts to generate the sets C, D and R for every viral protein using all the SLiMs in the ELM database.

C. Interactions

We compute the SLiM instances in viral proteins for all the human SLiM regular expressions in the set $C \cup D \cup R$. With the SLiM instances we infer PPIs between humans and the corresponding virus using the SLiM-domain associations in the ELM database and the protein-domain associations in the Pfam database [23]. We validate the interactions obtained with the LMPID database [29].

D. Analysis of the interactions

The PPIs inferred are analyzed statistically. The proteins in the HPSM are sorted by the number of interactions predicted with viral proteins. The viral proteins are classified by the number of interactions with different human proteins.

We classify the interactions as tentatively disrupting or bridging the human protein-protein interaction network. A viral protein that interacts with only one protein in the HPSM probably disrupts a pathway, while a viral protein that interacts with two or more HPSM proteins probably wires a new path.

ISBN: 978-1-61208-488-6

TABLE I
NUMBER OF INTERACTIONS PREDICTED WITH VIRAL PROTEINS

Human HPSM protein	Interactions with viral proteins
EIF4A1	30
EIF4A2	30
EIF4A3	30
EIF3B	44
EIF3G	44
PABPC5	44
PABPC1	50
PABPC3	50
PABPC4	50
EIF4E	55
EIF4E1B	55
EIF4E2	55
EIF4E3	55
EIF3I	78

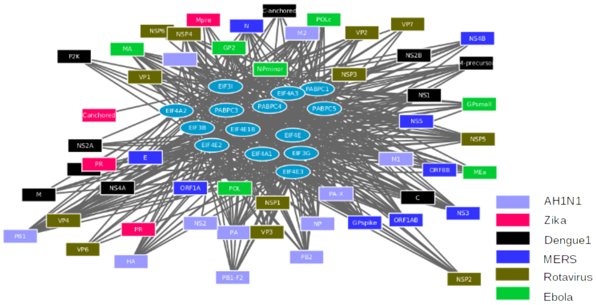


Fig. 1. Protein-protein interaction network predicted for protein-synthesis and viral proteins. Human protein-synthesis proteins are represented as ellipses and viral proteins as boxes. Boxes are colored differently for each virus.

The disrupting or wiring interactions are contrasted with the information in the KEGG pathway database [21] and gene ontology [20].

III. RESULTS

There are only two kinds of human proteins in the HPSM targeted by the selected viruses: 1) eukaryotic Initiation Factors (EIF*), 2) polyadenilate-binding proteins (PABPC*). No cytoplasmic ribosomal proteins or components of the ribosomal units are predicted to interact with the viral proteins. The number of interactions with viral proteins for the targeted proteins is reported in Table I.

Targeted proteins EIF3B, EIF3G and EIF3I belong to the module A of the EIF3 complex involved in the recruitment of the 43S ribosomal complex at the translation initiation phase.

Proteins EIF4A1, EIF4A2, EIF4A3, EIF4E, EIF4E1B, EIF4E2 and EIF4E3 are part of the EIF4 complex that binds to capped mRNAs in the translation initiation phase.

Finally, proteins PAPBPC1, PAPBPC3, PAPBPC4 and PAPBPC5 bind to the tail (end) of mRNAs recognizing poly(A) regions. This helps to mRNA circularization.

We obtain a network of interactions between human proteins in the HPSM subsystem and the proteins of the selected viruses represented in Figure 1.

We present two degree distributions for the network, one for the human proteins with respect to the number of interactions with viral proteins in Table II, and other for viral proteins with respect to the number of interactions with human proteins in Table III. For human proteins there is a clear hub, the protein EIF3I, predicted to interact with 78 viral proteins through SLiMs, but the other proteins have a large degree, Table II. On the other hand, there are 27 viral hub proteins that have 14 interactions with human proteins, Table III.

TABLE II
DEGREE DISTRIBUTION FOR HUMAN PROTEINS

Human protein degree	Number of proteins
30	3
44	3
50	3
55	4
78	1

TABLE III
DEGREE DISTRIBUTION FOR VIRAL PROTEINS

Viral Degree	Number of proteins
1	16
4	1
5	11
7	5
8	5
10	1
11	11
14	27

We classify the viral proteins in two groups: 1) the ones that have only one interaction with human proteins, potentially disrupting the protein-synthesis process and 2) the ones that have two or more interactions with human proteins, potentially bridging unexpected interactions between human protein-synthesis proteins or proteins in other pathways. These first group of potentially disrupting proteins is presented in Table IV and the viral hubs are presented in table V.

We find that the protein EIF3I, the Eukaryotic translation initiation factor 3 subunit I is the hub of the HPSM system, with 78 interactions. EIF3I is involved in the formation of translation preinitiation complex, regulation of translational

TABLE IV
VIRAL PROTEINS WITH ONE INTERACTION (POTENTIALLY DISRUPTING)

Virus	Viral protein
Zika	PR
MERS	NS5
Rotavirus	NSP6
WestNile	C
Dengue1	NS4B
Dengue1	NS4A
WestNile	C-anchored
Dengue1	M
WestNile	M
Dengue1	NS2A
MERS	E
WestNile	NS4A
WestNile	NS4B
Zika	C
AH1N1	NS2
Zika	Canchored

Copyright (c) IARIA, 2016. ISBN: 978-1-61208-488-6

TABLE V
VIRAL HUB PROTEINS (POTENTIALLY BRIDGING)

Virus	Protein
AH1N1	NP
AH1N1	M1
AH1N1	NS1
AH1N1	PA
AH1N1	PB1
AH1N1	PB2
Dengue1	NS3
Dengue1	NS5
Ebola	GPspike
Ebola	MA
Ebola	NP
Ebola	NPminor
Ebola	POL
Ebola	POLc
MERS	N
MERS	ORF1AB
Rotavirus	NSP4
Rotavirus	NSP5
Rotavirus	VP2
Rotavirus	VP4
WestNile	E
WestNile	NS1
WestNile	NS3
WestNile	NS5
Zika	NS1
Zika	NS3

initiation and assembly of the eukaryotic 48S preinitiation complex [20]. The EIF3I protein is in the hsa03013 RNA transport KEGG pathway, in which it is part of a multifactor complex with EIF1, EIF2 and EIF5 [21].

We tried to validate the interactions found against the LMPID database but did not found any candidate interaction there. Perhaps the coverage of SLiM-mediated VHPPIs is too limited at the moment.

IV. CONCLUSION AND FUTURE WORK

We propose the prediction of SLiM-mediated host-virus PPIs between the human HPSM and some selected viruses. Further analysis of the interactions obtained might yield clues about common viral strategies for subverting protein translation.

Our main contribution is the combination of SLiM filtering methods. Having a general implementation of SLiM finding and filtering allows that the methods can be extended to other subsystems like the interferon [30] and apoptosis proteins [31] to investigate viral infection mechanisms at different stages. The methods can even be used with non-human eukaryotic hosts.

ACKNOWLEDGEMENTS

This study was supported by a Colciencias scholarship granted to AB. The authors would like to thank Dr. Aydin Tozeren for receiving AB in the Biomedical Engineering Department at Drexel university and suggesting the protein translation machinery as a subsystem to study.

REFERENCES

[1] D. Walsh and I. Mohr, “Viral subversion of the host protein synthesis machinery,” *Nat. Rev. Microbiol.*, vol. 9, pp. 860–875, Dec 2011.

[2] N. J. Buchkovich, Y. Yu, C. A. Zampieri, and J. C. Alwine, “The TORrid affairs of viruses: effects of mammalian DNA viruses on the PI3K-Akt-mTOR signalling pathway,” *Nat. Rev. Microbiol.*, vol. 6, pp. 266–275, Apr 2008.

[3] L. Padilla-Noriega, O. Paniagua, and S. Guzman-Leon, “Rotavirus protein NSP3 shuts off host cell protein synthesis,” *Virology*, vol. 298, pp. 1–7, Jun 2002.

[4] S. Flint, V. Racaniello, G. Rall, and A. M. Skalka *Principles of Virology*. Third edition. American Society for Microbiology, 2009.

[5] S. Durmuş, T. Çakir, A. Özgür, and R. Guthke, “A review on computational systems biology of pathogen-host interactions,” *Front Microbiol*, vol. 6, p. 235, 2015.

[6] S. D. Durmuş Tekir and K. O. Ülgen, “Systems biology of pathogen-host interaction: networks of protein-protein interaction within pathogens and pathogen-human interactions in the post-genomic era,” *Biotechnol J*, vol. 8, pp. 85–96, Jan 2013.

[7] S. Wuchty, “Computational prediction of host-parasite protein interactions between P. falciparum and H. sapiens,” *PLoS ONE*, vol. 6, no. 11, p. e26960, 2011.

[8] M. D. Dyer, T. M. Murali, and B. W. Sobral, “Supervised learning and prediction of physical interactions between human and HIV proteins,” *Infect. Genet. Evol.*, vol. 11, pp. 917–923, Jul 2011.

[9] G. Cui, C. Fang, and K. Han, “Prediction of protein-protein interactions between viruses and human by an SVM model,” *BMC Bioinformatics*, vol. 13 Suppl 7, p. S5, 2012.

[10] R. K. Barman, S. Saha, and S. Das, “Prediction of interactions between viral and host proteins using supervised machine learning methods,” *PLoS ONE*, vol. 9, no. 11, p. e112034, 2014.

[11] J. M. Doolittle and S. M. Gomez, “Structural similarity-based predictions of protein interactions between HIV-1 and Homo sapiens,” *Virol. J.*, vol. 7, p. 82, 2010.

[12] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, “Prediction of interactions between HIV-1 and human proteins by information integration,” *Pac Symp Biocomput*, pp. 516–527, 2009.

[13] O. Tastan, Y. Qi, J. G. Carbonell, and J. Klein-Seetharaman, “Refining literature curated protein interactions using expert opinions,” *Pac Symp Biocomput*, pp. 318–329, 2015.

[14] T. Thieu, S. Joshi, S. Warren, and D. Korkin, “Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches,” *Bioinformatics*, vol. 28, pp. 867–875, Mar 2012.

[15] P. Evans, W. Dampier, L. Ungar, and A. Tozeren, “Prediction of HIV-1 virus-host protein interactions using virus and host sequence motifs,” *BMC Med Genomics*, vol. 2, p. 27, 2009.

[16] T. Hagai, A. Azia, M. M. Babu, and R. Andino, “Use of host-like peptide motifs in viral proteins is a prevalent strategy in host-virus interactions,” *Cell Rep*, vol. 7, pp. 1729–1739, Jun 2014.

[17] R. R. Halehalli and H. A. Nagarajaram, “Molecular principles of human virus protein-protein interactions,” *Bioinformatics*, vol. 31, pp. 1025–1033, Apr 2015.

[18] N. E. Davey, M. S. Cyert, and A. M. Moses, “Short linear motifs - ex nihilo evolution of protein regulation,” *Cell Commun. Signal*, vol. 13, no. 1, p. 43, 2015.

[19] T. J. Gibson, H. Dinkel, K. Van Roey, and F. Diella, “Experimental detection of short regulatory motifs in eukaryotic proteins: tips for good practice as well as for bad,” *Cell Commun. Signal*, vol. 13, p. 42, 2015.

[20] M. Ashburner et al., “Gene ontology: tool for the unification of biology. The Gene Ontology Consortium,” *Nat. Genet.*, vol. 25, pp. 25–29, May 2000.

[21] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets,” *Nucleic Acids Res.*, vol. 40, pp. D109–114, Jan 2012.

[22] H. Dinkel et al., “ELM–the database of eukaryotic linear motifs,” *Nucleic Acids Res.*, vol. 40, pp. D242–251, Jan 2012.

[23] R. D. Finn et al., “The Pfam protein families database: towards a more sustainable future,” *Nucleic Acids Res.*, vol. 44, pp. D279–285, Jan 2016.

[24] A. Nakao, M. Yoshihama, and N. Kenmochi, “RPG: the Ribosomal Protein Gene database,” *Nucleic Acids Res.*, vol. 32, pp. D168–170, Jan 2004.

[25] A. Bateman et al, “UniProt: a hub for protein information,” *Nucleic Acids Res.*, vol. 43, pp. D204–212, Jan 2015.

[26] J. R. Brister, D. Ako-Adjei, Y. Bao, and O. Blinkova, “NCBI viral genomes resource,” *Nucleic Acids Res.*, vol. 43, pp. D571–577, Jan 2015.

[27] Z. Dosztanyi, V. Csizmok, P. Tompa, and I. Simon, “The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins,” *J. Mol. Biol.*, vol. 347, pp. 827–839, Apr 2005.

[28] H. Dinkel et al., “The eukaryotic linear motif resource ELM: 10 years and counting,” *Nucleic Acids Res.*, vol. 42, pp. D259–266, Jan 2014.

[29] D. Sarkar, T. Jana, and S. Saha, “LMPID: a manually curated database of linear motifs mediating protein-protein interactions,” *Database (Oxford)*, vol. 2015, 2015.

[30] V. Navratil, B. de Chassey, L. Meyniel, F. Pradezynski, P. Andre, C. Rabourdin-Combe, and V. Lotteau, “System-level comparison of protein-protein interactions between viruses and the human type I interferon system network,” *J. Proteome Res.*, vol. 9, pp. 3527–3536, Jul 2010.

[31] S. E. Hasnain et al., “Host-pathogen interactions during apoptosis,” *J. Biosci.*, vol. 28, pp. 349–358, Apr 2003.

BioGraphDB: a New GraphDB Collecting Heterogeneous Data for Bioinformatics Analysis

Antonino Fiannaca, Massimo La Rosa, Laura La Paglia, Antonio Messina, Alfonso Urso
ICAR-CNR, National Research Council of Italy
Palermo, Italy
Email: {fiannaca, larosa, lapaglia, messina, urso}@pa.icar.cnr.it

Abstract—**Current bioinformatics databases provide huge amounts of different biological entities such as genes, proteins, diseases, microRNA, annotations, literature references. In many case studies, a bioinformatician often needs more than one type of resource in order to fully analyse his data. In this paper, we introduce BioGraphDB, a bioinformatics database that allows the integration of different types of data sources, so that it is possible to perform bioinformatics analysis using only a comprehensive system. Our integrated database is structured as a NoSQL graph database, based on the OrientDB platform. This way we exploit the advantages of that technology in terms of scalability and efficiency with regards to traditional SQL database. At the moment, we integrated ten different resources, storing and linking data about genes, proteins, microRNAs, molecular pathways, functional annotations, literature references and associations between microRNA and cancer diseases. Moreover, we illustrate some typical bioinformatics scenarios for which the user just needs to query the BioGraphDB to solve them.**

Keywords–Integrated database; Graph database; GraphDB; OrientDB; Bioinformatics database.

I. Introduction

In the last years, the use of computational approaches allowed researchers in bioinformatics and systems biology to produce, store and share a lot of data, such as genes, proteins, metabolic pathways, and so on. In most cases, data are collected in different databases, each of which has a proper framework and storage technology. For this reason, although the scientific community makes available to biologists and bioinformaticians a large amount of data, it is a big challenge to interconnect results from heterogeneous data sources, where each database can identify the same biological entity on one's own account. For all those reasons, it is important to provide an integrated database offering, in a modular framework, all the information contained in different available databases.

In this work, we propose BioGraphDB, an efficient bioinformatics NoSQL graph database, collecting data related to genes, microRNA (miRNA), proteins, pathways and diseases from 10 online public resources. Since we aim at integrating heterogeneous resources modelling pathways, interactions and relations among a lot of biological entities, we chose to implement a graph database; it has been highlighted by [1] that graph databases both allow for efficient queries and give advantages in scalability with respect to any relational database. The proposed database is built on the OrientDB platform [2], because previous works [3], [4] demonstrated that it outperforms the other NoSQL databases in terms of flexibility and performances.

Moreover, in this work we propose some cases of study in the field of biological and clinical research that can be resolved using the proposed database. The paper has the following structure: in Section II similar integrated DBs are presented; Section III presents the main components of the proposed DB; in Section IV it is described how the different resources have been imported and linked each other; in Section V we present four application scenarios and finally in Section VI some conclusions, as well as future developments, are drawn.

II. Related works

Due to the overwhelming size and type of biological data, the need of biological databases that integrate many different resources has risen. The National Center for Biotechnology Information (NCBI) [5] perhaps offers the most popular platform of integrated biological databases. It includes, among the others, the Entrez database [6] consisting of 37 different databases containing data related to genes, proteins, taxonomy, gene expression and so on; the PubMed system [7] for the scientific literature, the RefSeq [8] database that hosts non-redundant sets of curated genomic, proteomic and transcriptomic sequences; and the BioSystems [9] database that integrates and cross-links information about molecular pathways. The molecular pathways are at the basis of the KEGG integrated databases project [10]. In addition to information about pathways, KEGG has also information about genes, compounds, reactions, diseases and drugs.

In recent years, with the focus on the study of non-coding RNA and especially miRNA, many integrated resources have been developed considering miRNA as their core. The miRò knowledge base [11] is a system that integrates data about miRNAs, their validated and predicted gene targets, functional annotations provided by Gene Ontology (GO) [12] and gene-disease relations taken from the Genetic Association Database (GAD). Another miRNA-centric integrated database is miRWalk 2.0 [13], [14]. Besides data about miRNAs, GO annotations, miRNA-mRNA interactions and gene-disease associations, miRWalk stores and integrates data about pathways and gene and protein classes. Moreover miRWalk web service implements several pre-defined search methods that allow the user to query the database in order to find, for example, gene-miRNA-pathway relations, gene-miRNA-GO annotations, disease-miRNA relations. Even if miRWalk integrates several type of biological data, it however only allows to query them using the above described pre-defined search tools. The proposed BioGraphDB, in turn, lets the user access all of the data in order to assemble his own set of queries, thanks to its graph structure and a specialized query language (see Section IV and Section V).

Since in many cases it is needed only a limited set of bioinformatics resources, it would be useful to build a cus-

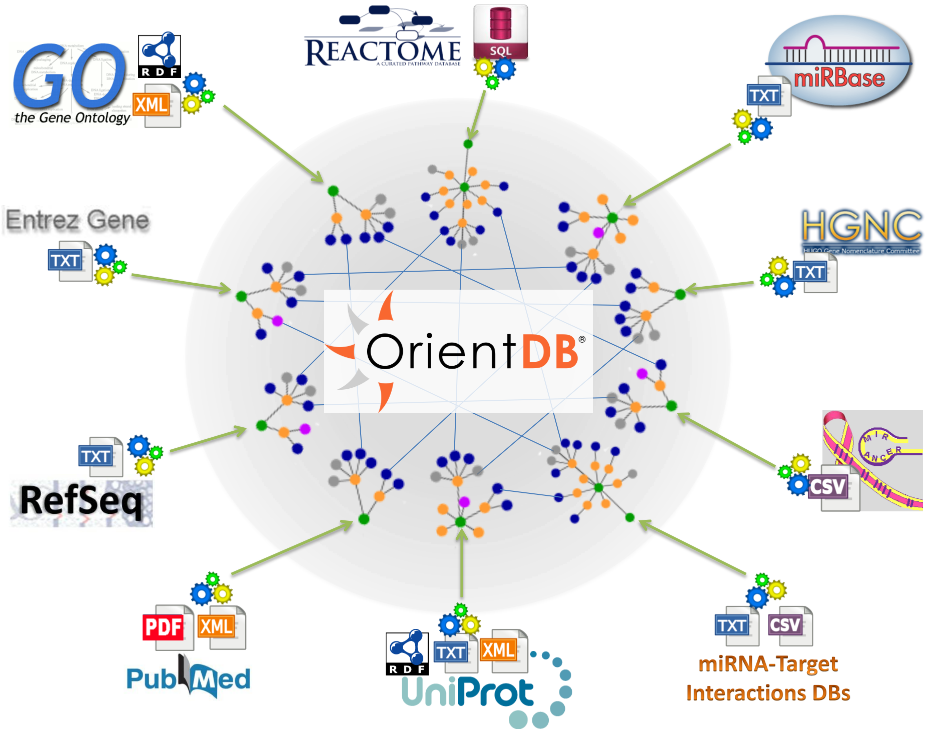


Figure 1. A graphical representation of the proposed integrated database based on OrientDB framework.

tom integrated database. The Java BioWareHouse (JBioWH) platform [15] offers a Java library that allows the importation and integration of different data sources into a SQL-based framework which defines a set of data types related to bio-entities, such as genes, proteins, pathways and drugs.

All the above described integrated databases, as well as the JBioWH library, are built upon a standard SQL architecture. With advances in the developing of NoSQL databases, which provide a more flexible and performing environment, new ways for integrating different resources have been studied. For example authors in [16] presented ncRNA-DB, a NoSQL database based on the OrientDB platform that put together many biological resources that deal with several classes of non-coding RNA (ncRNA) such as miRNA, long-non-coding RNA (lncRNA), circular RNA (circRNA) and their interactions with genes and diseases. More recently a graph-based database, called Bio4j, has been developed by [17]. Bio4j is based on a Java library that allows to build an integrated cloud-based data platform upon a graph structure. Bio4j is protein-centric, in fact it only includes data about proteins, GO and enzymes.

Since Bio4j has fewer resources rather than our proposed BioGraphDB (see Section III-B), it is difficult do directly compare them, especially because the number and type of resolvable scenarios are quite different (see [17] and Section V).

III. BioGraphDB components

All the components used in this work are discussed in the following. In details, in the next subsection we introduce the OrientDB framework, that represents the platform used to build the proposed work. In the subsection III-B we define all the databases we used in this work. Figure 1 shows a graphical representation of public databases integration.

A. *OrientDB*

OrientDB is an open source NoSQL database management system (DMBS) developed in Java by Orient Technologies LTD. It collects features of document databases and graph databases, including object orientation. In graph mode, referenced relationships are like edges, accessible as first-class objects with a start vertex, end vertex, and properties. This interesting feature let us represent a relational model as a document-graph model, maintaining the relationships.

OrientDB supports an extended version of SQL, to allow all sort of Create, Read, Update and Delete (CRUD) and query operations, and Atomicity, Consistency, Isolation, Durability (ACID) transactions, helpful to recover pending document at the time of crash. It is easily embeddable and customizable and it handles HTTP Requests, RESTful protocols and JSON without any 3rd party libraries or components. Finally, it is fully compliant with TinkerPop Blueprints [18], the standard of graph databases. It is distributed under the open source Apache 2 license [19], therefore it is totally free for any kind of use and its enterprise features are not limited.

B. *Data Source*

In order to build a database containing the most updated resources related to genes, proteins, miRNAs, metabolic pathways and their references in literature, it is useful to integrate the last versions of different publicly available data sources. For this aim, we take into account those on-line databases that represent the state-of-art in bioinformatics. In the following the list of databases we have considered for populating the proposed graph database, as showed in Figure 1.

1) miRBase [20]: The microRNA database (miRBase) is a searchable database of published miRNA sequences and annotation. It contains both hairpin and mature sequences of

223 species, and for each of them, it provides name, keywords, genomic location, references and annotations.

2) UniProtKB [21]: The UniProt Knowledgebase (UniProtKB) is the largest public collection of annotated functional information on proteins and it is updated every four weeks. It stores both computationally analysed and manually annotated records, including classifications, cross-references and quality indications available to scientific researchers.

3) Gene Ontology [12]: The Gene Ontology (GO) is the most complete and daily updated public resource for genes and proteins annotation. It provides annotations for gene products in biological processes, cellular components and molecular functions.

4) Reactome [22]: Reactome is a database containing validated metabolic pathways in human biology and computationally inferred pathways for 20 non-human species. Each pathway is annotated as a set of biological events, dealing with genes and proteins.

5) Entrez Gene [6]: The NCBI Entrez Gene database contains a wide set of details related to all the genes that have been studied in literature. For each gene, there is a record containing a lot of information, such as the genomic context, a list of ortholog/homolog genes, annotated pathways, interactions with other genes and so on.

6) Refseq [8]: The Reference Sequence (RefSeq) database is a collection of computationally and manually curated annotations for identification and characterization of genomes, transcripts and proteins.

7) Pubmed: Pubmed is a structured information resource on scientific publications in the field of biomedical literature. It allows to perform clinical queries for specific studies, categories and scopes. Due to copyright restrictions, only an open-access subset of this database is available for download.

8) mirCancer [23]: The microRNA Cancer association database (mirCancer) provides associations between miRNAs and related human cancers Pubmed entries. These associations are first extracted from Pubmed database by means of text mining algorithms and then manually revised. In addition, mirCancer gives, for each association, the miRNA expression profile.

9) HGNC [24]: The HUGO Gene Nomenclature Committee (HGNC) is the authority responsible for the gene nomenclatures (also known as gene symbols) for the human species. The HGNC database contains, for each gene symbol, a list of synonyms and a list of corresponding entries in the most popular gene databases (e.g. Refseq, Entrez gene). HGNC is the main source for synonyms disambiguation for genes and proteins.

10) miRNA-Target Interactions: This resource is a collection of publicly available miRNA-target interactions databases. It contains both validated and predicted interactions. The published experimentally validated interactions, including their experimental conditions, are provided by mirTarBase database [25]. A list of putative interactions are obtained by combining results of five different databases: miRNATIP [26], TargetScan [27], Diana micro-T [28], Pita [29] and miRanda [30].

IV. Data Integration

The publicly databases listed in the previous section give us a huge amount of data, that we have to integrate in an harmonious and consistent way. It is relatively easy to read and parse the various source files, but they often contain redundancies and useless data for our purpose, because, for example, at the moment we are only interested in the human species. Loading and linking the actual useful data is the goal.

Moreover, the databases are available for download in several different formats, such as tab-delimited plain-text, structured XMLs, SQL database dumps. The latest available release of OrientDB has a powerful tool to move data from and to a database by executing an Extract-Transformer-Loader (ETL) process, described by a JSON configuration file. However, its Extractor supports almost all data source types but XML. Therefore, in order to avoid mixed solutions, we decided to develop an ad hoc set of Java based ETLs.

As general rule, each biological entity and its properties have been mapped respectively into a vertex and its attributes, and each relationship between two biological entities has been mapped into an edge. If a relationship has some properties, they are also saved as edge's attributes. Vertices and edges are grouped into classes, according to the nature of the entities. For example, all the genes imported from NCBI Gene become instances of the *gene* vertex class, and all the proteins from UniProtKB become instances of the *protein* vertex class. Moreover, all the relationships between genes and proteins extracted from HGNC, in the form of “gene *G* codes for protein *P*”, become instances of the *coding* edge class.

The ETLs can be grouped in the following five categories:

- *Pubmed ETL*: It is not a real ETL, because actually we do not import any Pubmed publication. It is just used to create a vertex class used to store those *Pubmed IDs* found in the other databases.
- *Tab-delimited ETLs*: They were used to import NCBI Gene, miRNA-target interactions, HGNC, and mirCancer. Because all interactions have several virtually-searchable attributes, they have been mapped to vertices and then linked to the related gene and miRNA. By using the *protein-coding gene* field from HGNC, we were able to link each gene to its encoded proteins.
- *XML ETLs*: Starting from the related XML Schema Definition (XSD) [31] file and thanks to the unmarshalling capabilities of the standard JAXB library [32], they were used to import UniprotKB and GO.
- *miRBase ETL*: miRBase is available in a EMBL format text file, hence we used the *BioJava* library [33], in order to process the data in a simple and efficient way.
- *Reactome ETL*: The Reactome database import was not so easy. It is available for download only as SQL database dumps and its schema is not documented, hence we have installed the relational DBMS MySQL [34] and followed the available installation guide [35] in order to properly load the database from the dumps. After studying the database structure and tables definitions, we have created some ad hoc SQL views to extract the useful data, afterwards exported as a set of tab-delimited text files. Finally, we were able to import pathways data and to link the proteins to their pathways.

HGNC and UniprotKB databases provide conversion tables storing the synonyms for respectively gene and protein names,

TABLE I. OVERALL SIZE OF BOTH IMPORTED AND PROPOSED DATABASES.

External DBs		**BioGraphDB**	
Public DBs size	*Overall input lines*	*Vertices*	*Edges*
> 10 GB	> 185 millions	∼ 7.4 millions	∼ 15 millions

as well as their accession IDs to the most common biobanks. In our BioGraphDB we inserted those data into two different vertexes and linked them to the corresponding gene and protein vertexes. The same strategy can be applied for managing synonyms for other kinds of data.

The imported DBs have not overlapping information, that eventually could be contradictory, because we selected one source for each considered biological entity. In any case, when a new database will be imported, its data will be labelled (as attribute) with details about the source. For instance, if we import more than a miRNA target prediction database, then each prediction will contain an attribute declaring its original source. The advantage of this representation is that a user can define specific queries implementing consensus among different predictors or apply proper filters.

In order to guarantee data consistency and proper relationships, ETLs were executed in a precise order. Since each imported DB has dependencies with the other ones, it is of course important that all the depending resources are already present into the graphDB when a new resource is loaded. The following importation order assures that the dependencies among the integrated resources are correctly satisfied:

1) Pubmed *(schema creation)*
2) NCBI gene *(import)*
3) miRBase *(import, links to Pubmed)*
4) mirCancer *(import, links to miRBase and to Pubmed)*
5) miRNA-target interactions *(import, links to gene and to miRBase)*
6) UniprotKB *(import, links to Pubmed)*
7) HGNC *(links from genes to proteins, gene synonyms import, links from synonyms to gene)*
8) Reactome *(import, links from pathways to proteins)*
9) GO *(import, links to genes and to pathways)*

The import process lasted several hours and most of the time was spent in the creation of the vertices and links related to the miRNA-gene interactions. The size of both imported DBs, in terms of data size and number of input lines, and BioGraphDB, in terms of number of vertices and edges, is reported in Table I. The whole graph assembled by means of the intgration of all the DBs can be traversed using proper query languages, such as Gremlin [18]. Each graph traversal represents a set of queries that are enough in order to solve several bioinformatics scenarios, and some of them will be described in Section V.

V. Results

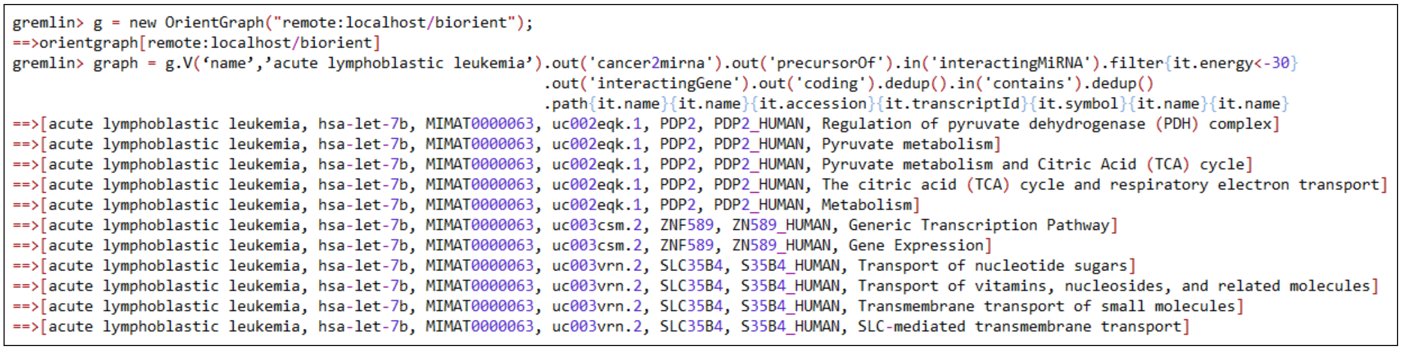
BioGraphDB can be used for the analysis in clinical research of different real life problems. Here we briefly introduce four scenarios representing typical bioinformatics problems that can be faced by means of suitable queries over the proposed DB. As an example, for the last scenario we provide a more detailed explanation and a query in the graph traversal language Gremlin [18] that resolves it.

```
g.V('name',cancer_name)
   .out('cancer2mirna')
   .out('precursorOf')
   .in('interactingMiRNA')
   .filter{it.energy<-energy_score}
   .out('interactingGene')
   .out('coding').dedup()
   .in('contains')
   .path{it.name}{it.accession}{it.accession}
        {it.transcriptId}{it.symbol}
        {it.name}{it.name}
```

Figure 2. A gremlin query for the proposed "*Target analysis of differentially expressed miRNAs in cancer*" scenario.

- *Analysis of gene functions and pathways*. Starting from the gene ID or gene sequence it is possible to investigate its role in the cellular context by exploring its functional annotations and location in pathways. Moreover it can be investigated the enrichment of that gene. This scenario requires the use of different databases: Entrez gene, RefSeq, GO, Reactome.
- *Analysis of protein motifs linked to cellular pathway*. The aim is searching the most representative protein motifs related to a specific cellular pathway. In this context, the study can be implemented by means of functional annotations related to these proteins. This scenario can be resolved using 3 databases: UniProt, Reactome, Gene Ontology.
- *Analysis of tumour-suppressor/oncogenic miRNA*. Starting from group of genes involved in a specific cellular pathway or cellular condition it is possible to identify potential miRNA targets that could have oncogenic or tumour-suppressor functions. This implies the use of 4 resources: Reactome, miRNA-target interactions, mirBase, mirCancer.
- *Target analysis of differentially expressed miRNAs in cancer*. Starting from a list of differentially expressed miRNAs linked to a specific disease, we would verify what are the major target proteins of these miRNAs belonging to particular cellular pathways. This analysis needs the use of 4 resources: mirCancer, mirBase, miRNA-target interactions, Reactome.

With regard to the last scenario, using publicly available resources, the following interactions steps are required. First of all, starting from a specific cancer type, a set of differentially expressed (DE) miRNAs can be obtained by the miRCancer database. The obtained miRNAs represent the input for the miRNA target interaction tools. Querying those tools, a list of putative miRNA targets is obtained. Filtering by energy scores, it is possible to evidence those targets that are more strongly linked to the DE miRNAs. The last step of the analysis is to verify if there are specific pathways that the selected targets belong to. This last step can be done through the use of pathways analysis tools such as Reactome. Reactome, in fact, given a list of input genes, provides a set of pathways containing those genes. A typical way in order to solve the described scenario would be to use each different DB (mirCancer, mirBase, miRNA-target predictors, Reactome) at once. In this situation, the user has to collect intermediate results and has to gain enough skill for using all the DBs. Instead of querying each biological resource singularly, all



```
gremlin> g = new OrientGraph("remote:localhost/biorient");
==>orientgraph[remote:localhost/biorient]
gremlin> graph = g.V(‘name’,’acute lymphoblastic leukemia’).out('cancer2mirna').out('precursorOf').in('interactingMiRNA').filter{it.energy<-30}
                                                           .out('interactingGene').out('coding').dedup().in('contains').dedup()
                                                           .path{it.name}{it.name}{it.accession}{it.transcriptId}{it.symbol}{it.name}{it.name}
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Regulation of pyruvate dehydrogenase (PDH) complex]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Pyruvate metabolism]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Pyruvate metabolism and Citric Acid (TCA) cycle]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, The citric acid (TCA) cycle and respiratory electron transport]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc002eqk.1, PDP2, PDP2_HUMAN, Metabolism]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003csm.2, ZNF589, ZN589_HUMAN, Generic Transcription Pathway]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003csm.2, ZNF589, ZN589_HUMAN, Gene Expression]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transport of nucleotide sugars]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transport of vitamins, nucleosides, and related molecules]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, Transmembrane transport of small molecules]
==>[acute lymphoblastic leukemia, hsa-let-7b, MIMAT0000063, uc003vrn.2, SLC35B4, S35B4_HUMAN, SLC-mediated transmembrane transport]
```

Figure 3. BioGraphDB response to the gremlin query depicted in Figure 2 for the proposed "*Target analysis of differentially expressed miRNAs in cancer*" scenario.

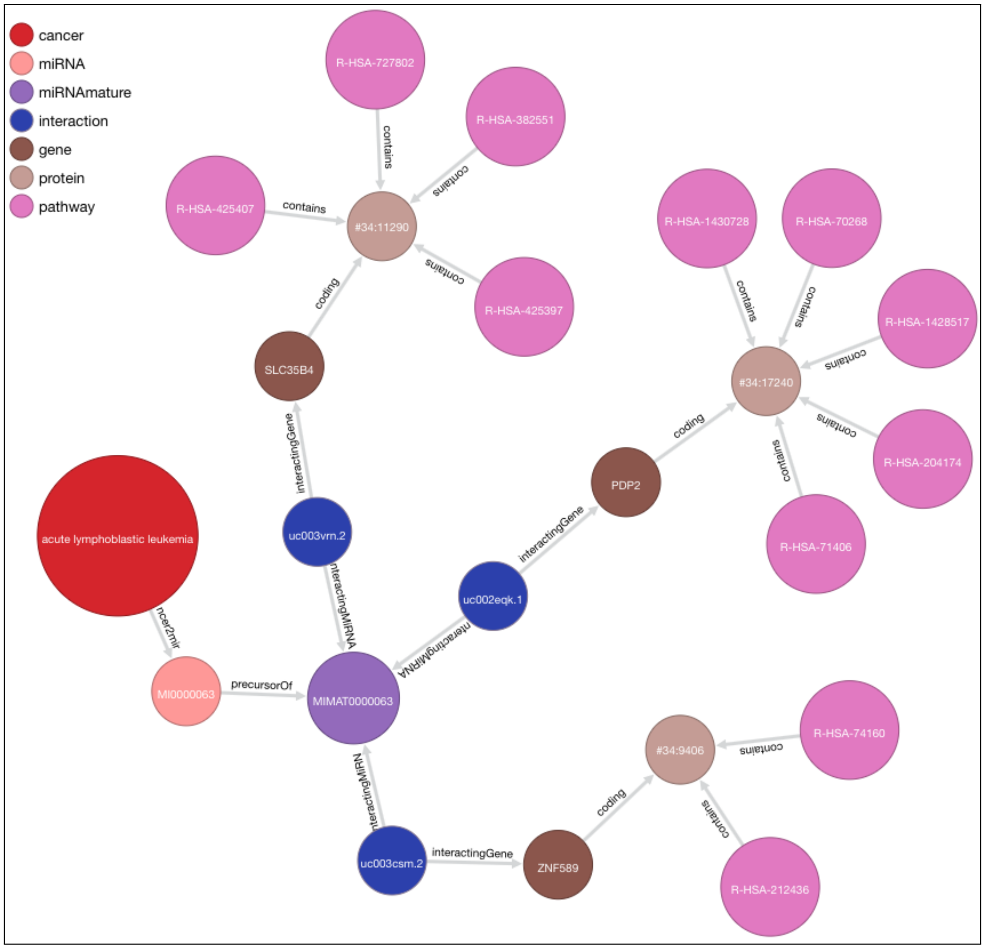


Figure 4. A graphical representation of the response produced by BioGraphDB, as seen in Figure 3. Starting from a specific disease ("*acute lymphoblastic leukaemia*"), we obtain 11 correlated biological pathways, marked with their Reactome ID.

of these steps can be easily performed using our integrated database by means of the Gremlin query shown in Figure 2.

For instance, if we set the cancer_name to "*acute lymphoblastic leukaemia*" and the energy_score threshold to "−30", we obtain as result eleven pathways, as showed in Figure 3. Figure 4 reports a graph representation of this result: starting from the "*acute lymphoblastic leukaemia*" disease, we obtain the "*hsa-let-7b*" DE miRNA, that interact with three genes (*SLC35B4*, *ZNF589*, *PDP2*). Each gene codes for a protein, that, in turns, is contained in at least a biological pathway. In this scenario, the query provides eleven Reactome pathways, marked with their Reactome ID. The complete set of results is summarized in Table II, where we reported the miRNA name and the pathway descriptions lacking in Figure 4.

VI. Conclusions and Future Works

In this work, we proposed BioGraphDB, an integrated graph database for biological data. This database was designed to overcome problems related to the lack of a structural organization and interoperability of publicly available biological

TABLE II. RESULT OF THE GREMLIN QUERY IN FIGURE 2.

Pathology	Mature miRNA	Gene	Reactome Pathway	Pathway Description
Acute lymphoblastic leukemia	hsa-let-7b-5p	SLC35B4	R-HSA-425407	SLC-mediated transmembrane transport
			R-HSA-727802	Transport of nucleotide sugars
			R-HSA-382551	Transmembrane transport of small molecules
			R-HSA-425397	Transport of vitamins, nucleosides, and related molecules
		PDP2	R-HSA-1430728	Metabolism
			R-HSA-70268	Pyruvate metabolism
			R-HSA-1428517	The citric acid cycle and respiratory electron transport
			R-HSA-204174	Regulation of pyruvate dehydrogenase
			R-HSA-71406	Pyruvate metabolism and Citric Acid
		ZNF589	R-HSA-74160	Gene Expression
			R-HSA-212436	Generic Transcription Pathway

resources. Finally we presented some cases of study where the use of the database can give a concrete advantage to the scientific community. Because our BioGraphDB stands at a prototypal stage, we are unable to provide at the moment a full performance evaluation, that will be done in future works.

Further developments will be done in the near future. Of course, thanks to the flexibility of the proposed database, other biological resources will be integrated where necessary. At the same time, we are developing proper automated mechanism in order to update on a regular schedule our BioGraphDB with the latest releases of its integrated DBs. After the data sources integration, we will develop a collection of web services with a common user-friendly web-interface and explicit search methods implementing proper database views. This way, it will be possible to solve some of the most common bioinformatics scenarios, like the ones proposed in this paper. In addition, we are working on a web service in order to provide the users a computer aided methodology to build their own custom views and search methods.

REFERENCES

[1] C. T. Have and L. J. Jensen, “Are graph databases ready for bioinformatics?” *Bioinformatics*, vol. 29, no. 24, pp. 3107–3108, 2013.

[2] Orient Technologies LTD, “OrientDB.” [Online]. Available: http://orientdb.com [accessed: 2016-02-19]

[3] M. Dayarathna and T. Suzumura, “XGDBench: A benchmarking platform for graph stores in exascale clouds,” in *4th IEEE International Conference on Cloud Computing Technology and Science Proceedings*. IEEE, 2012, pp. 363–370.

[4] A. Messina, P. Storniolo, and A. Urso, “Keep it simple, fast and scalable: a Multi-Model NoSQL DBMS as an (eb)XML-over-SOAP service,” in *The 30th IEEE International Conference on Advanced Information Networking and Applications (AINA-2016)*. IEEE, 2016, pp. 220–225.

[5] NCBI Resource Coordinators, “Database resources of the National Center for Biotechnology Information,” *Nucleic Acids Research*, vol. 41, no. D1, pp. D8–D20, 2013.

[6] G. D. Schuler, J. A. Epstein, H. Ohkawa, and J. A. Kans, “Entrez: molecular biology database and retrieval system.” *Methods in enzymology*, vol. 266, pp. 141–62, 1996.

[7] “PubMed” [Online] Available: http://www.ncbi.nlm.nih.gov/pubmed [accessed: 2016-02-19]

[8] K. D. Pruitt, T. Tatusova, and D. R. Maglott, “NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins,” *Nucleic Acids Research*, vol. 35, no. Database, pp. D61–D65, 2007.

[9] L. Y. Geer, et al., “The NCBI BioSystems database.” *Nucleic acids research*, vol. 38, no. Database issue, pp. D492–6, 2010.

[10] M. Kanehisa, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, “KEGG as a reference resource for gene and protein annotation,” *Nucleic Acids Research*, vol. 44, no. D1, pp. D457–D462, 2016.

[11] A. Lagana, et al., “miRo: a miRNA knowledge base,” *Database*, vol. 2009, 2009.

[12] The Gene Ontology Consortium, “Gene Ontology Consortium: going forward,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D1049–D1056, 2015.

[13] H. Dweep, N. Gretz, and C. Sticht, “miRWalk Database for miRNA-Target Interactions.” *Methods in Molecular Biology*, vol. 1182, pp. 289–305, 2014.

[14] H. Dweep and N. Gretz, “miRWalk2.0: a comprehensive atlas of microRNA-target interactions,” *Nature Methods*, vol. 12, no. 8, pp. 697–697, 2015.

[15] R. Vera, Y. Perez-Riverol, S. Perez, B. Ligeti, A. Kertesz-Farkas, and S. Pongor, “JBioWH: an open-source Java framework for bioinformatics data integration,” *Database*, vol. 2013, pp. bat051–bat051, 2013.

[16] V. Bonnici, F. Russo, N. Bombieri, A. Pulvirenti, and R. Giugno, “Comprehensive Reconstruction and Visualization of Non-Coding Regulatory Networks in Human,” *Frontiers in Bioengineering and Biotechnology*, vol. 2, 2014.

[17] P. Pareja-Tobes, R. Tobes, M. Manrique, E. Pareja, and E. Pareja-Tobes, “Bio4j: a high-performance cloud-enabled graph-based data platform,” Era7 bioinformatics, Tech. Rep., 2015.

[18] Apache Software Foundation, “Apache TinkerPop.” [Online]. Available: http://tinkerpop.incubator.apache.org [accessed: 2016-02-19]

[19] Apache Software Foundation, “Apache License Version 2.0.” [Online]. Available: http://www.apache.org/licenses/LICENSE-2.0 [accessed: 2016-02-19]

[20] A. Kozomara and S. Griffiths-Jones, “miRBase: integrating microRNA annotation and deep-sequencing data.” *Nucleic acids research*, vol. 39, no. Database issue, pp. D152–7, 2011.

[21] The UniProt Consortium, “UniProt: a hub for protein information,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D204–D212, 2015.

[22] D. Croft,et al., “The Reactome pathway knowledgebase,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D472–7, 2014.

[23] B. Xie, Q. Ding, H. Han, and D. Wu, “miRCancer: a microRNA-cancer association database constructed by text mining on literature,” *Bioinformatics*, vol. 29, no. 5, pp. 638–644, 2013.

[24] K. A. Gray, B. Yates, R. L. Seal, M. W. Wright, and E. A. Bruford, “Genenames.org: the HGNC resources in 2015,” *Nucleic Acids Research*, vol. 43, no. D1, pp. D1079–D1085, 2015.

[25] S.-D. Hsu, et al., “miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions,” *Nucleic Acids Research*, vol. 42, no. D1, pp. D78–D85, 2014.

[26] A. Fiannaca, M. La Rosa, L. La Paglia, R. Rizzo, and A. Urso, “MiRNATIP: a SOM-based miRNA-target interactions predictor,” *BMC Bioinformatics*, in press.

[27] B. P. Lewis, C. B. Burge, and D. P. Bartel, “Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.” *Cell*, vol. 120, no. 1, pp. 15–20, 2005.

[28] M. D. Paraskevopoulou,et al., “DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows,” *Nucleic Acids Research*, vol. 41, no. W1, pp. W169–W173, 2013.

[29] M. Kertesz, N. Iovino, U. Unnerstall, U. Gaul, and E. Segal, “The role of site accessibility in microRNA target recognition.” *Nature genetics*, vol. 39, no. 10, pp. 1278–1284, 2007.

[30] B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks, “Human microRNA targets,” *PLoS Biology*, vol. 2, no. 11, 2004.

[31] World Wide Web Consortium (W3C), “W3C XML

Schema Definition Language (XSD) 1.1." [Online]. Available: https://www.w3.org/TR/xmlschema11-1/ [accessed: 2016-02-21]

[32] Java Community Process, "JSR 222: Java Architecture for XML Binding (JAXB) 2.0." [Online]. Available: https://jcp.org/en/jsr/detail?id=222 [accessed: 2016-02-21]

[33] A. Prlic, et al., "BioJava: an open-source framework for bioinformatics in 2012," *Bioinformatics*, vol. 28, no. 20, pp. 2693–2695, 2012.

[34] Oracle Corporation, "MySQL." [Online]. Available: http://www.mysql.com [accessed: 2016-02-22]

[35] Website 3 Installing SOP, "Reactome." [Online]. Available: http://wiki.reactome.org/index.php/Website_Installing_SOP [accessed: 2016-02-22]

Automated Identification of Molecular Structures for NMR Based Metabolomics

Arianna Filntisi, George K. Matsopoulos
School of Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
Email: afilntisi@biomig.ntua.gr, gmatso@esd.ece.ntua.gr

Charalambos Fotakis, Panagiotis Zoumpoulakis
Institute of Biology, Medicinal Chemistry and Biotechnology
National Hellenic Research Foundation
Athens, Greece
Email: bfotakis@yahoo.com, pzoump@eie.gr

Dionisis Cavouras, Pantelis Asvestas
Department of Biomedical Engineering
Technological Educational Institute of Athens
Athens, Greece
Email: cavouras@teiath.gr, pasv@teiath.gr

***Abstract*—The Nuclear Magnetic Resonance (NMR) based metabolomics approach is implemented in medicine and pharmacology to assess, identify and quantify metabolites in different biological samples. Metabolite determination, which is a challenging task due to the complexity of the biological matrices, can benefit from bioinformatics tools. In this context, our research has focused on the development of a new computational metabolite identification method from ^{1}H-NMR spectra.**

Keywords- metabolomics; automated metabolite identification; ^{1}H NMR.

I. INTRODUCTION

Metabolomics is the research discipline that is concerned with the qualitative and quantitative assessment of the metabolic response of biological systems to pathophysiological stimuli or genetic modifications. Metabolomics provides information of *in vivo* multi-organ functional integrity in real time [1]. The NMR metabolomics approach has a number of applications, such as the identification of biomarkers of disease and pharmacodynamic response.

In alignment with the increasing applications of the NMR metabolomics, the chemoinformatics field has evolved focusing on data processing and the metabolite identification algorithms. Several methods have targeted the task of metabolite profiling with line fitting and Bayesian modelling, [2-4]. Another strategy has been based on the matching of the input spectrum with a set of reference compounds [5-6].

Herein, we introduce a new computational method for the automatic identification of metabolites from 1D ^{1}H-NMR spectra. In Section II, the main steps of the method are presented. In Section III, preliminary results from the analysis of two mixtures are described.

II. METHOD

The steps A, B presented below are preprocessing steps, while step C constitutes the core of our method.

A. Preprocessing

The input spectrum can be preprocessed by removing low significance regions. Denoising and thresholding can be performed if necessary. The mean, median and Gaussian denoising filters were tested and the last one has been chosen as the best. However, denoising was performed frugally as an auxiliary step, and therefore the filter selection did not affect significantly the results.

B. Data Reduction

The Adaptive Intelligent binning algorithm [7] is applied to the spectrum, resulting in a number of bins corresponding to the local minima across the frequency spectrum. Subsequently, peak picking is performed, selecting for every bin the frequency corresponding to the maximum intensity.

C. Metabolite Search

The input spectrum is screened for metabolites, as defined by a specific database (see Section II.D). For each multiplet peak of every candidate metabolite in the database, a number of peak combinations are being considered as a possible fit. A candidate peak combination $p = \{p_{c1}, p_{c2}, \dots, p_{cn}\}$ for a multiplet is scored differently depending on the order of the spectral lines (first order or higher order multiplets, singlets and multiplets without rules). The scoring of a multiplet is based on its properties, such as the *j* coupling and the height ratios. The optimal peak combination is chosen for each metabolite.

D. Database

A database file containing 850 metabolites was synthesized from the available Human Metabolome Database (HMDB) [8]. For every metabolite, information such as the multiplet type, the expected frequency ranges, the number of hydrogen atoms, the j coupling values, the height ratios, as well as the number of peaks has been stored.

ISBN: 978-1-61208-488-6

III. Results

The method described in Section II has been tested on an amino acid mixture and a human amniotic fluid sample. The former comprised of L-Alanine, L-Valine, L-Methionine, L-Proline, L-Glutamic acid, L-Leucine, L-Isoleucine, L-Arginine, Trigonelline, which were all successfully identified. The latter was screened for the presence of 40 metabolites (experimentally identified), 36 of which were positively recognized.

The performance of our method upon those spectra has been compared to an existing metabolite recognition tool, MetaboHunter (Table 1) [5] and proved to be enhanced despite a higher execution time. Our method seems to be more robust seeking the optimal peaks for a candidate metabolite at each input spectrum with a four digit accuracy, as opposed to assigning predetermined peaks to a metabolite.

In Figure 1, we can see characteristic multiplet peaks assigned to metabolites in an area of the spectrum of the human amniotic fluid mix.

IV. Discussion

This work presents briefly a new chemoinformatics method for metabolite recognition under development. Also, preliminary results from the application of the method on two spectra have been described.

A limitation of our method is the fact that its performance has not been verified on different sample types. Our future goals include the refinement of the method and the default parameter values used. Further validation of the proposed method and comparison with other metabolite identification methods are also necessary.

Acknowledgment

Funding was received through an IKY Fellowship of Excellence for postgraduate studies in Greece – Siemens Program. The authors confirm that the funder had no influence over the study design, content of the paper, or selection of this conference.

References

[1] J. K. Nicholson, J. C. Lindon, and E. Holmes, ““Metabonomics”: understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data”. Xenobiotica, vol. 29(11), pp 1181-1189, 1999.

[2] C. Zheng, S. Zhang, S. Ragg, D. Raftery, and O. Vitek, “Identification and quantification of metabolites in (1)H NMR spectra by Bayesian model selection,” Bioinformatics, vol. 27(12), pp. 1637–44, 2011.

[3] P. Mercier, M. J. Lewis, D. Chang, D. Baker, and D. S. Wishart, “Towards automatic metabolomic profiling of high-resolution one-dimensional proton NMR spectra,” J. Biomol. NMR, vol. 49(3-4), pp. 307–23, 2011.

[4] W. Astle, M. De Iorio, S. Richardson, D. Stephens, and T. Ebbels, “A Bayesian Model of NMR Spectra for the Deconvolution and Quantification of Metabolites in Complex Biological Mixtures,” J. Am. Stat. Assoc., vol. 107(500), pp. 1259–71, 2012.

[5] D. Tulpan, S. Léger, L. Belliveau, A. Culf, and M. Cuperlović-Culf, “MetaboHunter: an automatic approach for identification of metabolites from 1H-NMR spectra of complex mixtures,” BMC Bioinformatics, vol. 12, pp. 400, 2011.

[6] D. Jacob, C. Deborde, and A. Moing, “An efficient spectra processing method for metabolite identification from 1H-NMR metabolomics data,” Anal. Bioanal. Chem., vol. 405(15), pp. 5049-61, 2013.

[7] T. D. De Meyer, et al, “NMR-based characterization of metabolic alterations in hypertension using an adaptive, intelligent binning algorithm,” Anal. Chem., vol. 80(10), pp. 3783–90, 2008.

[8] D. S. Wishart, et al, “HMDB: the Human Metabolome Database,” Nucleic Acids Res., vol. 35(Database issue), D521–6, 2007.

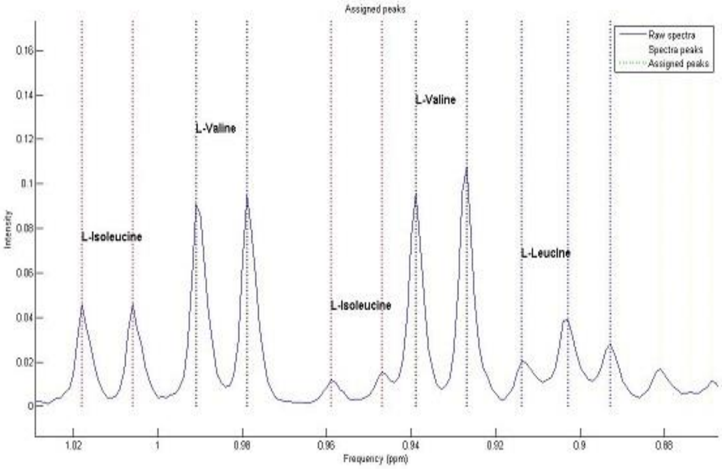


Figure 1. Peak assignment in an area of the human amniotic fluid spectrum.

TABLE I. Comparative peak assignment for three metabolites of the human amniotic fluid spectrum

Metabolites	Metabolite identification methods	
	Our method	***MetaboHunter***
L-Valine	0.9269, 0.9389, 0.9789, 0.9909, 3.5571, 3.5641, 2.1900, 2.2020, 2.2100, 2.2140, 2.2200, 2.2270, 2.2320	0.98, 1.02, 1.04, 2.23, 2.24, 2.25, 2.26, 2.27, 2.28, 2.29
L-Isoleucine	0.9469, 0.9589, 1.0059, 1.0179, 1.9430, 1.9570, 1.9690, 1.9790, 1.9910	0.91, 0.93, 0.94, 0.99, 1, 1.21, 1.22, 1.26, 1.28, 1.42, 1.43, 1.44, 1.45, 1.46, 1.47, 1.48, 1.49, 1.94, 1.96, 1.97, 1.98, 1.99, 2, 3.66
L-Leucine	0.8929, 0.9029, 0.9139, 1.6470, 1.6590, 1.6720, 1.6840, 1.6980, 3.7001, 3.7111, 3.7191	0.94, 0.95, 0.96, 1.65, 1.67, 1.69, 1.7, 1.71, 1.72, 1.73, 1.75, 3.71, 3.72, 3.73, 3.74