BIOTECHNO 2021

The Thirteenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies

ISBN: 978-1-61208-859-4

May 30th – June 3rd, 2021

BIOTECHNO 2021 Editors

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany

BIOTECHNO 2021

Foreword

The Thirteenth International Conference on Bioinformatics, Biocomputational Systems and Biotechnologies (BIOTECHNO 2021), held between May 30 – June 3rd, 2021, covered these three main areas: bioinformatics, biomedical technologies, and biocomputing.

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

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

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

We take here the opportunity to warmly thank all the members of the BIOTECHNO 2021 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 2021.

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 2021 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that BIOTECHNO 2021 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.

BIOTECHNO 2021 Chairs:

BIOTECHNO 2021 Steering Committee
Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA

BIOTECHNO Industry/Research Advisory Committee
Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania
Gilles Bernot, University Nice Sophia Antipolis, France
Erliang Zeng,University of Iowa, USA
Y-h. Taguchi, Chuo University, Japan

BIOTECHNO 2021 Publicity Chairs
Daniel Basterretxea, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

BIOTECHNO 2021

Committee

BIOTECHNO 2021 Steering Committee

Birgit Gersbeck-Schierholz, Leibniz Universität Hannover, Germany
Hesham H. Ali, University of Nebraska at Omaha, USA

BIOTECHNO 2021 Industry/Research Advisory Committee

Steffen Heber, North Carolina State University, USA
Alexandru Floares, SAIA Institute, Romania
Gilles Bernot, University Nice Sophia Antipolis, France
Erliang Zeng,University of Iowa, USA
Y-h. Taguchi, Chuo University, Japan

BIOTECHNO 2021 Publicity Chairs

Daniel Basterretxea, Universitat Politecnica de Valencia, Spain
Marta Botella-Campos, Universitat Politecnica de Valencia, Spain

BIOTECHNO 2021 Technical Program Committee

Behrooz Abbaszadeh, University of Ottawa, Canada
Antonino Abbruzzo, University of Palermo, Italy
A M Abirami, Thiagarajar College of Engineering, Madurai, India
Don Adjeroh, West Virginia University, USA
Jens Allmer, Hochschule Ruhr West - University of Applied Sciences, Germany
Yoseph Bar-Cohen, Electroactive Technologies / NDEAA Lab - Jet Propulsion Laboratory (JPL), USA
Kais Belwafi, King Saud University, Saudi Arabia
Boubaker Ben Ali, University of Bordeaux, France / University of Manouba, Tunisia
Razvan Bocu, Transilvania University of Brasov, Romania
Vincenzo Bonnici, Universita' degli Studi di Verona, Italy
Matthias Chung, Virginia Tech, USA
Peter Clote, Boston College, USA
Giovanni Cugliari, University of Turin, Italy
Santa Di Cataldo, Politecnico di Torino, Italy
Eleftheria Polychronidou, Information Technologies Institute - Centre for Research and Technology Hellas (ITI/CERTH), Greece
Maria Evelina Fantacci, University of Pisa, Italy
Rosalba Giugno, University of Verona, Italy
Asier Ibeas, Universitat Autònoma de Barcelona, Spain
Jan Kubicek, VSB - Technical University of Ostrava, Czech Republic
Antonio LaTorre, Universidad Politécnica de Madrid, Spain
Cedric Lhoussaine, University Lille, France

Copyright Information

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

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

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

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

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

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

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

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

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

Table of Contents

Genetic Markers Associated with Anemia in Individuals with Sickle Cell Disease in Tanzania

Liberata Alexander Mwita[1], Raphael Zozimus Sangeda[1], Upendo Masamu[1], David Dynerman[2]

Liberata Alexander Mwita, Raphael Zozimus Sangeda,
Upendo Masamu
Department of Pharmaceutical Microbiology
Muhimbili University of Health and Allied Sciences
Dar-es-salaam, Tanzania
lmwita@blood.ac.tz, sangeda@gmail.com,
umasamu@blood.ac.tz

David Dynerman
The Public Health Company
5662 Calle Real #222
Goleta, California 93117
California, United States of America
emperordali@block-party.net

***Abstract*—Sickle cell disease is a global health problem, a genetic disease which affects many people, particularly common among those whose ancestors came from sub-Saharan Africa. All individuals with sickle cell disease experience anemia which increases the morbidity and mortality. This research aims to identify genetic variants associated with anemia in individuals with sickle cell disease. In the long-term this will contribute towards efforts to improve the life expectancy of individuals by quickly identifying single nucleotide polymorphisms related to anemia in sickle cell disease and enabling better prediction of the severity of anemia that the individual will experience which enable better preventive treatment. Quality control of the Genome Wide Association Studies data and association between anemia and the genotype data were performed using PLINK software and will be presented. Designing of imputation and replication study of the Genome Wide Association Studies data is in progress. The analysis will identify single nucleotide polymorphisms and genes linked to anemia in individuals with sickle cell disease. The results can also be compared with single nucleotide polymorphisms candidates from other studies.**

***Keywords*-Sickle cell disease; Anemia; Genome Wide Association Studies.**

I. INTRODUCTION

Sickle Cell Disease (SCD) is inherited genetic disorder caused by mutation in the hemoglobin (HBB) gene. SCD is a major public health concern [1]. Worldwide, it is estimated that majority of the 275,000 babies born with SCD annually are in sub-Saharan Africa [2]. The burden of SCD in Tanzania is high where it is estimated that 11,000 children are born with SCD annually [3]. SCD causes shortage of healthy Red Blood Cells (RBC) due to the polymerization of the RBCs into a sickle shaped red blood cells. These aggregate in small blood vessels and slow or block blood flow and oxygen initiating vaso-occlusion. Individuals with SCD become anemic because the sickle shaped cells have a short life (10-20 days) unlike normal RBCs which live for 120 days.

Despite the similarity in the origin of the disease, individuals demonstrate varying symptoms and severity. Our previous studies confirmed known and identified new genetic variants associated with fetal hemoglobin [4] and liver function (manuscript write-up ongoing). Other studies [5, 6] have also identified genetic variants associated with different phenotypes observed in individuals with SCD, however much of the variation in phenotype is yet to be explained.

Anemia in SCD increases the morbidity and mortality of individuals. Considering the amount of hemoglobin (Hb) as one variable, non-SCD individuals have a normal range of 13.5-17.5 grams per deciliter (adult men) and 12-15.5 grams per deciliter (adult women). In our database SCD individuals have an average of 8 grams per deciliter. Genome Wide Association Studies (GWAS) involve studying a set of genetic variants in different individuals to see if any variant is associated with a trait by investigating the entire genome of each individual. This study aims to identify genetic variants associated with anemia in individuals with sickle cell disease using a database of GWAS data for 1952 individuals with SCD in Tanzania.

The methodology used to identify the markers will be presented in Section II, followed by the results of the analysis in Section III. Discussion and conclusion of the research will be presented in Sections IV and V, respectively.

II. METHODOLOGY

A. Sampling of subjects and data collection

The phenotype data contains clinical, laboratory and demographic information. Some of these parameters were used in this analysis. Data of 1952 individuals diagnosed with SCD from a cohort have been genotyped. Samples were collected, DNA extracted and genotyped. These individuals are part of the Muhimbili Sickle Cohort recruited at Muhimbili National Hospital, Dar es Salaam, Tanzania. Full details are provided in [4]. Samples were typed on the Illumina Human Omnichip 2.5 platform.

B. Quality control of the genotype data and Association

Standard technical Quality Control (QC) of the data was performed using PLINK software to remove possible sources of technical and genetic bias [7]. This includes removing missing data, duplicates and individuals and Single Nucleotide Polymorphisms (SNPs) failing QC.

 ISBN: 978-1-61208-859-4

Principal Components Analysis (PCA) and the association of the phenotype (Hb) to the QC genotype were done by using PLINK software.

C. Genotype imputation and replication study

Genotype imputation is a statistical inference of unobserved genotypes which is performed on SNPs using known haplotypes in a population such as 1000 Genomes Project in humans. Genotype imputation is underway. Replication in GWAS studies is performed to confirm the phenotype-genotype association results by providing statistical evidence and rule out associations due to biases. Designing of replication and imputation study of GWAS data is in progress.

III. RESULTS

Fig. 1 shows the relationship of the quality-controlled genotype data from our study to other populations.

Our study population (blue dots) is admixture, most of individuals cluster with individuals of African ancestry while few individuals deviate from the cluster. The individuals deviating from the cluster are of Arabic and Indian origin.

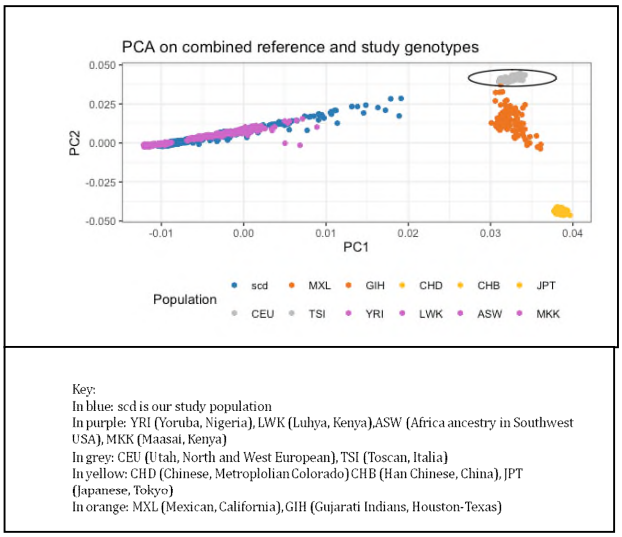


Figure 1. Study population relatedness to other population based on Principal Component Analysis (PCA)

IV. DISCUSSION

The analysis showed that the SNPs associated with anemia are present in the genes that are co-expressed. Individuals with sickle cell anemia experience anemia and frequent infections, this activates the immune response in individuals in order to fight the infections. The SNPs that significantly associated with anemia are found in the genes (Table.1) which function in cell-cell adhesion, antigen receptor-mediated signaling pathway, immune response-activating cell surface receptor signaling pathway and T cell receptor signaling pathway These functions are associated with immune response in humans; it is common for immune system to respond when the human body gets infected.

Other SNPs that significantly associate with anemia are found in the genes (Table.1) that function in hindbrain development and central nervous system neuron differentiation. This is expected in individuals with sickle cell anemia because they experience episodes of pain as well as developmental delays.

Unfortunately, the SNP found to be mostly significant associated with anemia (Fig. 2) at chromosome 3 and 7 have not been annotated hence the functions are not known.

Fig. 2 shows the SNPs (red and blue dots, the p-values on the y-axis) and chromosome in which the SNPs belong on the x-axis.

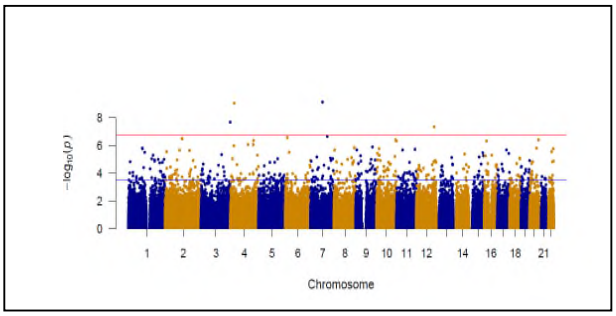


Figure 2. Manhattan plot showing SNPs that associate with anemia in SCD, significant SNPs located at Chromosome 3, 7 and 12.

Some of the SNPs that associate with anemia in individuals with SCD are in Table 1.

TABLE I. FEW SELECTED SNPS AND THEIR LOCATIONS.

SNP	Chromosome	Gene
rs2269688	8	MTMR7
rs11259403	10	PRKCQ
rs13389996	2	CTNNA2
rs10778462	12	CKAP4
rs7136826	12	CLEC1A
rs11632584	15	MEGF11
rs7163369	15	SLCO3A1
rs732523	12	PCED1B
rs17276467	7	CREB3L2
rs10209276	2	KCNH7
rs4578863	2	ZC3H6

 ISBN: 978-1-61208-859-4

It is our hope that the completion of replication and imputation analysis will reveal more and significant associations.

V. CONCLUSION AND FUTURE WORK

This study indicated genetic markers (SNPs) that associate with anemia in individuals with SCD. This is the first step towards developing a tool that will quickly identify the markers linked to anemia in SCD individuals which is an important step in improving preventive treatment of these individuals. Similar analysis has to be extended in same and different sickle cell disease cohorts in order to identify new and confirm the variants linked to anemia in individuals with SCD.

ACKNOWLEDGMENT

This research is sponsored by University of California Global Health Institute (UCGHI) GloCal Health Fellowship.

REFERENCES

[1] Un.org, “Secretary-General's message on sickle-cell anaemia | United Nations Secretary-General” [Online] Available from: https://www.un.org/sg/en/content/sg/statement/2009-0619/secretary-generals-message-sickle-cell-anaemia, 2009, [Accessed 15 Nov. 2018].

[2] E. Ambrose et al., “High birth prevalence of sickle cell disease in Northwestern Tanzania,” Pediatric Blood & Cancer, vol. 65(1), pbc.26735, Jan. 2018, doi: 10.1002/pbc.26735.

[3] F. Tluway and J. Makani, “Sickle cell disease in Africa: an overview of the integrated approach to health, research, education and advocacy in Tanzania, 2004-2016, “British Journal of Haematology,” vol. 177(6), pp.919-929, June. 2017, doi: 10.1111/bjh.14594.

[4] S. Mtatiro et al., “Genome wide association study of fetal hemoglobin in sickle cell anemia in Tanzania,” PLoS ONE, vol. 9(11), pp.e111464, Nov. 2014, doi: https://doi.org/10.1371/journal.pone.0111464

[5] P. Sebastiani et al., “Genetic modifiers of the severity of sickle cell anemia identified through a genome-wide association study,” American Journal of Hematology, vol. 85(1), pp.29-35, Jan. 2010, doi: 10.1002/ajh.21572.

[6] L. Liu et al., “Original Research: A case-control genome-wide association study identifies genetic modifiers of fetal hemoglobin in sickle cell disease,” Experimental Biology and Medicine, vol. 241(7), pp.706-718, April, 2016, doi: 10.1177/1535370216642047

[7] S. Purcell et al., “PLINK: a tool set for whole-genome association and population-based linkage analyses,” The American Journal of Human Genetics, vol. 81(3), pp.559-575, Sep. 2007, doi: 10.1086/519795.

 ISBN: 978-1-61208-859-4

Limits of a Glucose-Insulin Model to Investigate Intestinal Absorption in Type 2 Diabetes

Work in progress

Danilo Dursoniah, Maxime Folschette, Cédric Lhoussaine
Univ. Lille, CNRS, Centrale Lille, UMR 9189 CRIStAL,
F-59000 Lille, France
email: danilo.dursoniah@univ-lille.fr
maxime.folschette@centralelille.fr
cedric.lhoussaine@univ-lille.fr

Rebecca Goutchtat, François Pattou, Violeta Raverdy
Univ. Lille, Inserm, CHU Lille, U1190 - EGID,
F-59000 Lille, France
email: rebecca.goutchtat@vet-alfort.fr
francois.pattou@univ-lille.fr
violeta.raverdy@univ-lille.fr

***Abstract*—Abnormal regulation of glucose absorption in the small intestine is an important cause of Type 2 Diabetes (T2D). Even if this hypothesis is clinically well-known, it has not been fundamentally validated yet, mainly due to a lack of reliable metabolic knowledge on the glucose regulation. The main objective of this paper is to test this hypothesis on a highly referenced model composed of ordinary differential equations. This model is tested on an original dataset featuring the observations of obese diabetic patients. It shows its limits to predict our post-prandial glycemia and insulinemia time series especially with regard to the crucial complexity of gastro-intestinal regulation.**

***Keywords*—Ordinary Differential Equation; Systems Biology; Type 2 Diabetes; Intestinal Glucose Absorption.**

I. Introduction

Diabetes is a chronic metabolic disease characterized by a lack of insulin secretion and a decreased peripheral insulin response. Insulin is a hormone that down-regulates blood sugar concentration. Consequently, the balance of glucose and insulin concentrations in different tissues, called *homeostasis*, is pathologically perturbed: hyperglycemia is observed both during fasting and post-prandial periods. It gradually leads to tissues damages and subsequent diseases, i.e., a high rate of comorbidity [6].

More precisely, *Type 2 Diabetes* (T2D) results from the body's ineffective use of insulin. Most patients ($\sim 90\%$) with diabetes have T2D. Around 400 millions of people are affected worldwide by the disease representing a major public health issue in most developed countries [7]. It is commonly accepted that this type of diabetes is largely caused by physical inactivity combined with an high-carbohydrate diet. However, through bariatric surgery, obese patients with T2D have seen their physiological glycemia immediately restored, independently to their weight loss [1]. This observation leads us to consider *Intestinal Glucose Absorption* (IGA) as a critical cause of T2D, among others. Bariatric surgery, and more precisely *Roux-en-Y Bariatric Surgery* (RYGB), anatomically leads to the decrease of the glucose absorption surface, which would explain, at least partially, this unexpected clinical benefit. Furthermore, the gastro-intestinal tract includes:

- enzymatic and mechanical transformation of starch (amylopectine and amylose) into absorbable glucose,
- incretin secretion and effects on the blood sugar,
- and the small intestine microbiota, which may modulate dietary responses.

This landscape of hypothetical causal factors shows that fundamental research effort on T2D must continue despite precise clinical understanding of the disease. However, all representations of glucose-insulin homeostasis largely underestimate the importance of the gastro-intestinal tract into the blood sugar consequences. Instead, they tend to model with increasing details the interaction of insulin with its related tissues (pancreas, liver and insulino-dependent tissues). We want to investigate the contribution of IGA to glucose homeostasis and its potential role in diabetes. To this aim, and as a preliminary work, we consider a typical and state of the art homeostasis glucose-insulin model [5] formalized as a system of *Ordinary Differential Equations* (ODEs). Our objective is twofold:

- test if this model can predict a significant improvement of glucose homeostasis by simulating RYGB as is observed experimentally,
- test if this model can predict the time-course data of an original dataset of diabetic patients.

In Section 2, we briefly describe the model. In Section 3, we present our parameter fitting results both from the original parameters of [5] and for our own dataset We discuss the partial results in Section 4 and present the on-going and future work in Section 5.

II. Model

Many simulation models of the glucose-insulin system for the postprandial period have been developed [8]. In this work, we consider a highly cited model, proposed in [5], to simulate the postprandial physiological events of their own cohorts of normal subjects and T2D patients. This model is made of 12 ODEs and 36 parameters describing fluxes of glucose and insulin between physiological compartments: gastro-intestinal tract, plasma, liver, pancreas, muscle and adipose tissues (Figure 1). We recall in the following, in informal terms, how the physiological modules interact.

The *Gastro-Intestinal Tract* module describes the digestion process, from the stomach to the gut, and can be considered as the input of the whole system. It includes the complexity of

 ISBN: 978-1-61208-859-4

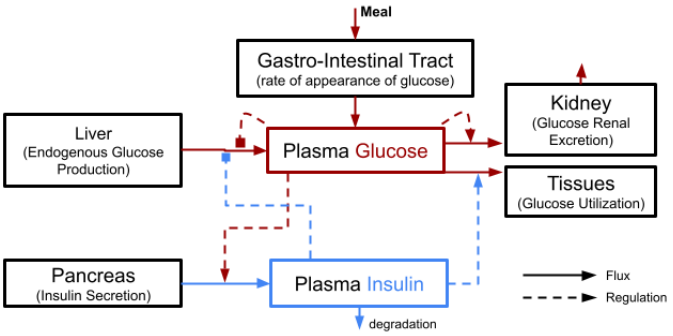


Figure 1: Simplified interaction graph of physiological compartments describing the model of [5].

the gastric emptying depending on the proportion of the solid and liquid phases of the alimentary bolus [4]. Only the liquid phase that ends up in the gut is absorbed by the intestine and discharged to the peripheral blood. Incretins are not modeled. The *Liver* module describes the hepatic activity responsible for the endogenous glucose production down-regulated by the insulin. The glucagon hormone is not modeled. The *Pancreas* secretes insulin, which is up-regulated with the amount of glucose in the blood. The uptake of glucose by the (muscle and adipose) *Tissues* is regulated by the insulin. The *Kidney* is responsible for the glucose excretion-reabsorption. The insulin degradation is due to its lifespan and liver clearance. Even though the insulin independent glucose uptake by the brain is modeled (not in shown in Figure 1) the regulation of the glucose by the brain is neglected in [5].

Using tracer-to-tracee ratio clamp technics [3], the authors of [5] measured the time course of the glucose concentration in various compartments. This was done following a mixed-meal received by several normal and diabetic subjects. The parameters were estimated to fit this experimental data, which resulted in two sets of parameter values modeling respectively normal and T2D behaviors. For practical reasons, we fully reimplemented the model in the *Julia* programming language (version 1.5.3) with the *DifferentialEquations* package (version 6.17.0).

III. Results

A. From T2D to normal model

In the following, we call *normal model*, resp. *T2D model*, the model instantiated with the normal, resp. diabetic, parameter values given in [5] (Table 1). We consider several parameter subsets corresponding to the previous modules: gastro-intestinal tract (also denoted as Ra in [5]), liver (EGP), tissues (GK+U), pancreas (IK+S) and kidney (RE).

Starting from the T2D model, we re-estimated, in turn, each of these subsets of parameters, while leaving the other parameters fixed, in order to fit the plasma glucose dynamics of the normal model. Based on the data of [5], our first objective is to evaluate the capability of the model to predict, for each module alone, its capability to restore a normal glycemia. We estimate the parameters twice: with and without estimation of the basal values for insulin secretion by the pancreas, glucose production by the liver and utilization by the tissues. From the resulting 10 inferred models, we plotted the time course of the observed variables.

In this short paper, we only report (Figures 4a to 4d) the most relevant plots for our purpose: plasma glucose (G), plasma insulin (I) and the rate of intestinal absorption (Ra) for the two models obtained from the estimations of the gastro-intestinal tract and pancreas compartments, with and without basal estimation. In order to compare the models' performance in fitting the normal model, we collect in a bar plot (Figure 3) the residual sum of squares for each model.

B. Parameter estimation of obese diabetics and RYGB

Our second objective is to test whether the model of [5] can predict the time course concentrations of glucose and insulin obtained from our own dataset of diabetic patients who underwent RYGB surgery. For each patient, we use data before (hereafter referred to as *visit A*) and 3 months after (*visit B*) surgery. We first estimate all the parameters in order to fit the time course data of glucose and insulin from the *visit A* dataset. The model that we obtain is called the *visit A* model. Figure 2 shows the glucose and insulin plasma concentrations predicted by this model as well as the fitted data points. Then, as previously, we estimate each subset of parameters in order to fit the *visit B* dataset. Here, we only consider the case where we also estimate the basal concentrations, which indeed changed 3 months after surgery. We report in Figure 4 the time course of G, I and Ra after estimation of the parameters of the gastro-intestinal tract and pancreas.

IV. Discussion

Estimating the basal values for insulin secretion, endogenous glucose production and insulino-dependent glucose utilization, can be interpreted as a prediction of the "long-term" effect of the parameter changes *in the best case* (since the model does not incorporate any long-term recovery mechanisms). Thus, not estimating these basal values can be interpreted as a "short-term" (or worst case) prediction.

A. From T2D to normal model

Our re-estimations of the parameters based on the data of [5] predict (Figure 3) that the best performing compartments to restore a normal glycemia are pancreas and tissues, and then the intestinal tract. As expected, estimating the basal concentrations (i.e., long-term effect) improves the performance especially for the intestinal tract (see also G curve in Figures 4b and 4c), which is consistent with experimental observation. However, the performance of the pancreas should be modulated. Indeed, Figures 4b and 4a show that, in order to improve the glycemia, a very high plasma insulin concentration is necessary if only the pancreas parameters are modified. This seems physiologically unrealistic, meaning that the good performance of this compartment is over-estimated. Similarly, the estimation of the gastro-intestinal tract parameters on the short term (Figure 4d) indicates an unrealistic decrease of

 ISBN: 978-1-61208-859-4

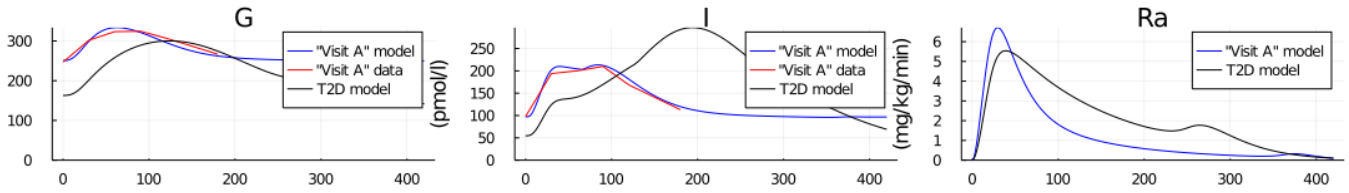


Figure 2: Parameter estimation of all parameters, including basal concentrations, to fit "Visit A" data.

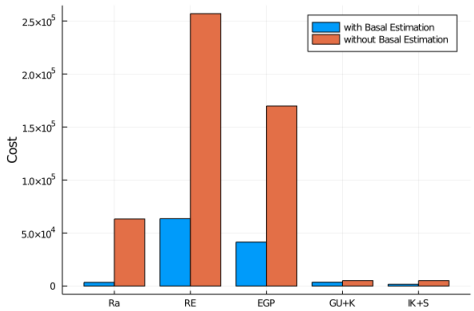


Figure 3: Loss from the parameters estimation applied on the different compartments.

plasma insulin concentration. Finally, the "long-term" estimation of the gastro-intestinal tract parameters (Figure 4c) allows a good improvement of the glycemia with a realistic concentration of insulin and a decrease of intestinal absorption, which is consistent with observation.

B. Parameter estimation of obese diabetics and RYGB

From Figure 2, it is primarily interesting to notice that the T2D data of [5] is significantly different from ours. Indeed, glucose in *visit A* is higher than in *T2D model* whereas insulin in *visit A* is lower than in *T2D model*. Despite this difference, the fitting is satisfying with all parameters set as free. This fitting is sensitive to the parameters estimation methods.

In Figures 4e and 4f, the parameters are set free for the pancreas (IK+S) and the gastro-intestinal tract (Ra), respectively. Such process can be interpreted as surgery simulations targeting respectively the pancreas and the gastro-intestinal tract. No fitting attempt seems satisfying. On the one hand, freeing IK+S parameters seems to be satisfying for fitting the glycemia but clearly overestimates the insulinemia (cf. Figure 4e). On the other hand, by freeing Ra parameters, the rate of appearance is decreased as observed in experimental data (based on our D-xylose data, an alimentary glucose marker). Still, the parameter estimation fails completely to fit *visit B* glycemia and insulinemia (cf. Figure 4f).

V. Conclusion and Future Works

The parameter estimation performed by the authors of [5] is based on training data generated by tracer-to-tracee clamp technique [3], which, despite its efficiency, remains an uncommon and complex method to monitor exogenous solutes. In practice, plasma glycemia and insulinemia are usually the only accessible clinical data. However, and as our model assessment suggests, this may raise parameter identifiability issues. To overcome this problem, we first plan to use the available additional D-xylose data, a marker that can be used to fit the rate of appearance (Ra) [2]. We also plan, by exploiting profile likelihood and sensitivity analysis, to study model reduction in order to eliminate the potential sources of non-identifiability. Other original datasets are currently used for the parameter estimation, generated from experiments on minipigs. Such biological models allow for more experiments and reproducibility, and decreased individual variabilities thus improving the reliability of parameter estimation. In this direction, another possibility, is to use publicly available datasets.

It should be noted that the failure of parameter estimation may be due to structural problems inherent to the model of [5] which sub-model of IGA is largely simplified. For instance, it ignores the spatial none uniform glucose absorption rate along the intestin and the secretion of incretins. We plan to extend the model of [5] with these aspects while simplifying the others to overcome identifiability issues that could emerge from additional parameters related to the gastro-intestinal tract.

References

[1] G. Baud et al., "Sodium glucose transport modulation in type 2 diabetes and gastric bypass surgery," *Surgery for Obesity and Related Diseases*, 12(6), pp. 1206–1212, 2016.

[2] I. T. Bjarnason and R. A. Sherwood, *Clinical biochemistry of the gastrointestinal tract*, pp. 214–230. Elsevier, 2014.

[3] C. Cobelli, G. Toffolo, and D. M. Foster, "Tracer-to-tracee ratio for analysis of stable isotope tracer data: link with radioactive kinetic formalism," *American Journal of Physiology-Endocrinology And Metabolism*, 262(6), pp. 968–E975, 1992.

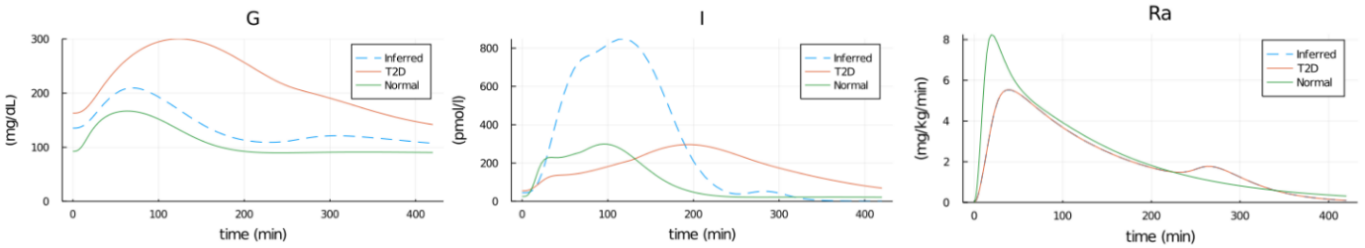
[4] C. Dalla Man, M. Camilleri, and C. Cobelli, "A system model of oral glucose absorption: validation on gold standard data," *IEEE Transactions on Biomedical Engineering*, 53(12), pp. 2472–2478, 2006.

[5] C. Dalla Man, R. A. Rizza, and C. Cobelli, "Meal simulation model of the glucose-insulin system," *IEEE Transactions on Biomedical Engineering*, 54(10), pp. 1740–1749, 2007.

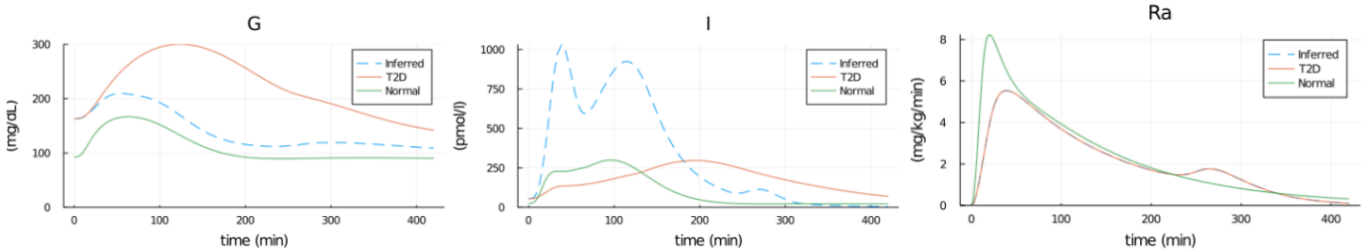
[6] A. Ghasemi and R. Norouzirad, "Type 2 diabetes: An updated overview," *Critical Reviews in Oncogenesis*, 24(3), pp. 213–222, 2019.

[7] L. Goedeke, R. J. Perry, and G. I. Shulman, "Emerging pharmacological targets for the treatment of nonalcoholic fatty liver disease, insulin resistance, and type 2 diabetes," *Annual review of pharmacology and toxicology*, 59, pp. 65–87, 2019.

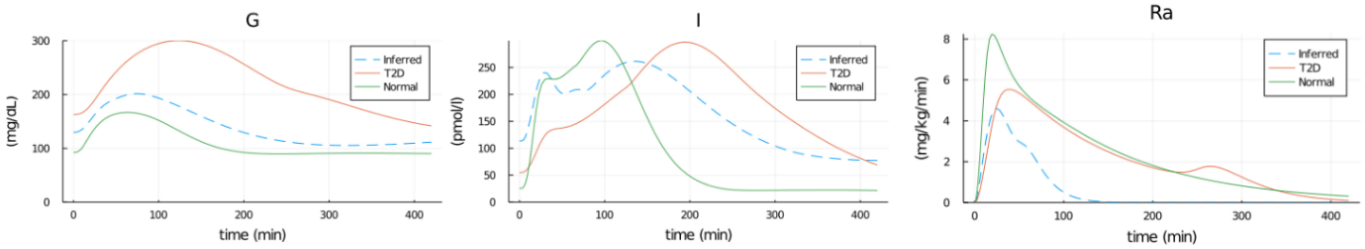
[8] A. Mari, A. Tura, E. Grespan, and R. Bizzotto, "Mathematical modeling for the physiological and clinical investigation of glucose homeostasis and diabetes," *Frontiers in Physiology*, 11, pp. 1548–1565, 2020.



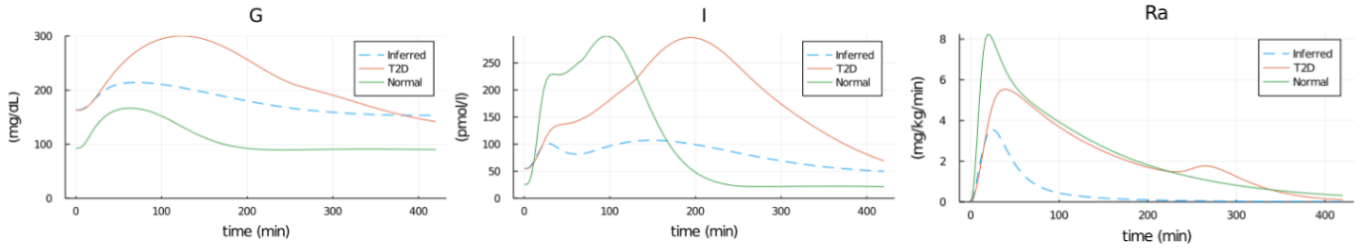
(a) Parameter estimation of IK+S compartment and basal concentration to fit the normal model for the T2D model.



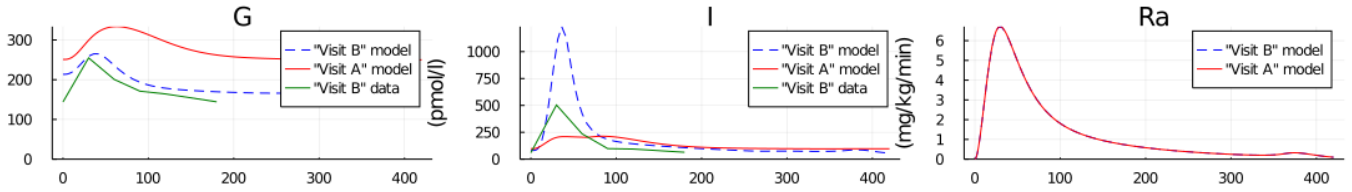
(b) Parameter estimation of IK+S compartment only to fit the normal model for the T2D model.



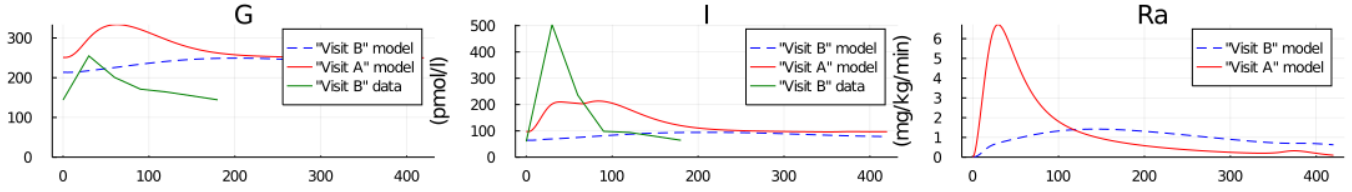
(c) Parameter estimation of Ra compartment and basal concentration to fit the normal model for the T2D model.



(d) Parameter estimation of Ra compartment only to fit the normal model for the T2D model.



(e) Parameter estimation of IK+S compartment and basal concentrations, to fit *visit B* model from *visit A* model.



(f) Parameter estimation of Ra compartment and basal concentrations, to fit *visit B* model from *visit A* model.

Figure 4: Glucose (*G*), insulin (*I*), and rate of appearance (*Ra*) after parameter estimation, with and without basal concentrations, of pancreas (a, b), gastro-intestinal tract (c, d) compartment for fitting normal model from TD2 model and for fitting *visit B* from *visit A* model (e, f).

 ISBN: 978-1-61208-859-4

A Word Recurrence Based Algorithm to Extract Genomic Dictionaries

Vincenzo Bonnici
Department of Computer Science
University of Verona
Verona, Italy
email: vincenzo.bonnici@univr.it

Giuditta Franco
Deptartment of Computer Science
University of Verona
Verona, Italy
email: giuditta.franco@univr.it

Vincenzo Manca
Deptartment of Computer Science
University of Verona
Verona, Italy
email: vincenzo.manca@univr.it

***Abstract*—Genomes may be analyzed from an information viewpoint as very long strings, containing functional elements of variable length, which have been assembled by evolution. In this work, an innovative information theory based algorithm is proposed, to extract significant (relatively small) dictionaries of genomic words. Namely, conceptual analyses are here combined with empirical studies, to open up a methodology for the extraction of variable length dictionaries from genomic sequences, based on the information content of some factors. Its application to human chromosomes highlights an original inter-chromosomal similarity in terms of factor distributions.**

***Keywords*—Genome languages, information content, Kullback-Leibler, word extraction.**

I. Introduction

Human genome computational analysis is one of the most important and intriguing research challenges we are currently facing. Genomes carry the main information underlying life of organisms and their evolution, including a system of molecular rules which orchestrate all cell functions [1]. Our work here follows and outlines some trends of research which analyze and interpret genomic information, by assuming the genome to be a book encrypted in a language to decipher [2–7], in order to convert the genomic information into a comprehensible mathematical form, such as a dictionary of variable-length factors that collects words of the unknown genomic language.

According to a common approach in computational genomics [8–12], a genome is represented by a string over the nucleotidic alphabet. This representation easily leads to affinities with a text, written in a natural language, which is comprehensible by means of its vocabulary, giving both syntax and semantic of *words*.

Several studies define properties for words which result to be salient features in analysing genomic sequences [13]. Minimal absent words, maximal or palindromic repeated words are some examples [14–16]. These approaches are focused on finding specific words to be used as key features of a string for analysing its property or for comparing it to another sequence [17]. The extracted words are often sparsely located in the analysed sequence [18], thus they do not constitute a real linguistic analysis of genomic strings.

According to recent advancements, the concept of *functional element* is central, defined as a genomic segment that codes for a defined biochemical product or displays a reproducible biochemical signature [6, 19]. An information theory based analysis clearly plays an important role in deciphering such elements as the genomic language [20], and it allows us to confirm the linkage between DNA fragments and their information content [4, 8, 19, 21–23].

In [24, 25], the authors applied a methodology developed for literary text to extract fixed length genomic dictionaries. Examples of fixed length dictionary extraction procedures could be provided by applying notions such as word multiplicity or word length distributions. On the other hand, graphical investigative analyses, based on expected frequency gaps, show the unpredictable behaviour of genomic sequences and help to detect peculiar words [26].

If we think of a book, semantically significant words have a fairly medium number of occurrences and they are clustered according to the topic described in specific part of the book. Several works are focused on finding genomic words exhibiting some special kind of (somehow clustered) repetitiveness, with a global frequency quite different than the expected frequency in purely random sequences having the same length of an investigated genome [8, 21, 22, 27–29]. A very relevant and peculiar word periodicity is revealed by the *Recurrence Distance Distribution* (RDD), which measures the frequency at which a given word occurs at given distances [30]. Its application to coding regions shows the informational evidence of the codon language, and in [31–33] some characterizations of recurrence behaviours were pointed out for very short k-mers. However, only fixed length dictionaries were extracted from real genomes by means of such a distribution [25].

In this paper, we start from a modified version of an algorithm introduced in [24], in order to apply it to real genomes. We call it V-algorithm, from the first name of the authors who designed it. Both these original and modified algorithms are aimed at finding words forming local clusters (the approach is explained in Section II-A). Then, we propose a new RDD-based algorithm, we call it W-algorithm, which extracts variable length dictionaries of interests from several real genomic sequences and collects words having a recurrence distribution maximally different than their random distribution. Such a selection is developed by computing the (locally) maximum divergence, from random sequences, of the RDD of each string obtained by elongating an initial *seed word*

 ISBN: 978-1-61208-859-4

over the genome. The divergence from random sequences is a crucial issue in information analysis of strings [34, 35] and in analyzing mathematical properties of dictionaries. The methodology in [24] to find dictionaries is therefore here improved by the V-algorithm, and a more general approach is proposed (Section II-B) by means of the RDD based W-algorithm, that works with the global word recurrence distance distribution rather than with only a first slice of it.

II. Material and Methods

This section summarizes the genomic word extraction methodology reported in [24], which was our starting point to develop a variant of it, the V-algorithm, and then introduces a novel RDD-based extraction algorithm, called W-algorithm. We also propose some criteria to evaluate extracted genomic dictionaries. Following the terminology from our previous work [12], a genome is a string over the genomic alphabet $\Gamma = \{A, C, G, T\}$. Given a genome G, we call $D_k(G) \subseteq \Gamma^k$ the *k-dictionary* of all k-mers occurring in the genome G. Given a word $\alpha \in D_k(G)$), a recurrence distance distribution (RDD) informs how many times α occurs at a given distance d. Thus, a recurrence is a pair of positions (p_1, p_2) (with $p_1, p_2 < |G|$ and $p_1 < p_2$) such that α occurs in p_1 and p_2 and no other occurrences of α are in the middle. The recurrence distance is given by $p_2 - p_1$.

A. A clustering coefficient based approach

RDD has been used to identify keywords by applying a methodology that associates a clustering coefficient C to k-mers [24]. The main idea is based on the fact that keywords are not uniformly distributed among a literary text, instead they are clustered. The approach combines the information provided by the spatial distribution of a word along the text (via the clustering coefficient) and its frequency, since the statistical fluctuation depends on the frequency. This basic approach has been used in [25] to assign a relevance to 6-mers and 8-mers in *Homo sapiens* and *Mus musculus*. The 8-mers were sorted by their normalized clustering coefficient (called σ_{nor}), and it has been shown that part of the top-200 clustered words (about 70%) appears in known functional biological elements, like coding regions and transcription factor binding sites.

The whole recurrence distribution is synthesised with a single parameter σ, to quantify the clustering level, previously presented in [9] for studying the energy levels of quantum disorder systems [36], and a clustering degree σ_{nor} assigned to words, for the identification of keywords in literary texts, obtained by means of the relation between the σ of a real word and the theoretical expected one (coming from a theoretical hypothesized distribution), as in the following.

For a given word, the parameter σ is the standard deviation of its normalized set of recurrence distances, $\sigma = s/\bar{d}$, where s is the standard deviation of the recurrence distance distribution, and $\bar{d}$ is the average recurrence distance. When the RDD is a geometric distribution, the parameter is denoted by σ_{geo} and it is equal to $\sqrt{1-p}$, since $s = \sqrt{1-p}/p$ and $\bar{d} = 1/p$, where p is the word frequency. Thus, the resultant normalized clustering measuring σ_{nor} of the given word is given by $\frac{\sigma}{\sigma_{geo}} = \frac{s/\bar{d}}{\sqrt{1-p}}$. For values of σ_{nor} near to 1, the recurrence distribution of the word is close to the geometric one, thus it indicates a randomness of the word. In fact, a random sequence is generated by a Bernoullian process, then different occurrences of a given word are independent events, and the event of having k occurrences of a word (in a segmentation unit) follows a Poisson distribution. Therefore, according to probability theory [37] its waiting time, that is the distance at which a word recurs, is an exponential distribution (having a geometric distribution as a discrete counterpart).

For words with low multiplicity, the statistical fluctuation is much larger, and it is possible to obtain an higher σ_{nor} for rare words placed at random, and they would be misidentified as keywords. Thus, the authors applied a correction by a Z-score measure that combines the clustering of a word and its multiplicity n. The resultant clustering measure C is given by the following equation: $C(\sigma_{nor}, n) = \frac{\sigma_{nor} - \langle\sigma_{nor}\rangle(n)}{sd(\sigma_{nor})(n)}$, where $\langle\sigma_{nor}\rangle(n) = \frac{2n-1}{2n+2}$ and $sd(\sigma_{nor})(n) = \frac{1}{\sqrt{n}(1+2.8n^{-0.865})}$. Parameter values were obtained via extensive simulations, by taking into account the distribution of σ_{nor} in random texts. They represent the mean value and the standard deviation of such empirical distribution. The C coefficient measures the deviation of σ_{nor} with respect to the expected value in a random text, in units of the expected standard deviation. In this case, $C = 0$ indicates randomness, $C > 0$ that the word is clustered and $C < 0$ that the word *repels* itself.

In [24] also an approach to explore the lineage of a word (from a short word to one of its possible elongations), without any knowledge about the effective word length, was provided. Given an initial word length k_0, some of the words in $D_{k_0}(G)$ are selected, according to their C measure, that must be greater then a C_0 measure corresponding to a fixed percentile (usually 0.05). Successively, for each of these initial words, their lineage is explored by selecting only the *elongations having a C measure greater than C_0, and up to a fixed maximal word length*: these are properly the two points we changed in the V-algorithm presented in the next section. The longest visited lineage is selected as a word with semantic meaning, and the process is repeated for different values of k_0 (ranging from 2 to 35), until a dictionary is obtained by discarding repeating words.

B. The RDD-based W-algorithm

We use RDD to calculate the divergence of the real distribution of a word within the genome from its frequency over a random string with the same genome length [29, 38]. Such a divergence is used as a measure of the information content of a word. Low expressive words are elongated by an expansion procedure, until they reach a reasonable level of *significance* according to which they are classified as genomic words of the extracted dictionary.

We assume that the higher the entropic divergence from the above exponential distribution, the more specialized and evolutionary selected is the genomic element. In this sense, low multiplicity words already represent elements owning high

 ISBN: 978-1-61208-859-4

level of significance. Instead, for what concerns repeats, we associate their *meaning* with their repetitiveness-profile, as it is revealed by their RDD. A word has to occur along the genomic sequence several times and at different distances. See an example in Figure 1, where the exponential distribution represents the random recurrence behaviour of the word. RDD of words along real genomes is often sparse, meaning that several distances (of recurrence) actually do not appear in the genome. This is why we evaluate the sound (i.e., more fitting) exponential distribution after removing peaks, that are absent in exponential functions, and by imposing a normalization ensuring the overall unitary probability.

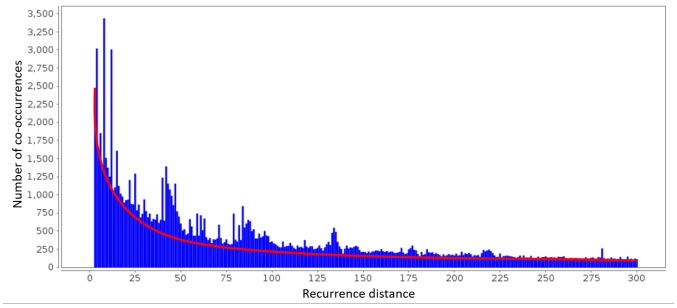


Fig. 1. RDD of word CGC (the jagged curve) in human chromosome 22

The degree of significance of a word is its *random deviation*, measured by the function in 1, based on the the entropic divergence (Kullback-Leibler divergence [27]), between the real RDD of a word (over the analysed genome) and its expected exponential distribution.

More technically, given a word α, which occurs in a genome G, we calculate its random deviation as the entropic divergence between its RDD and a suitable exponential distribution. To this aim, we first extract the real RDD of α over G, which we refer as R_α. Then, we estimate a two parameters exponential distribution E_α, by making use of the Nelder and Mead Simplex algorithm [39]. A denoised distribution is used as input for the estimation procedure: it is obtained by applying a low-pass filter (over R_α) in order to attenuate peaks. Afterwards, we remove from E_α the domain values which are not present in R_α, namely the gaps of R_α. Successively, both R_α and E_α are normalized in order to become probability distributions. Finally, the random deviation of α is chosen as:

$$r(\alpha) = \max(KL(R_\alpha, E_\alpha)\ ,\ KL(E_\alpha, R_\alpha)), \qquad (1)$$

where KL is the asymmetric Kullback-Leibler entropic divergence.

In our algorithm (reported in Listing 1) estimation of the information content of a word α is computed by the function $r(\alpha)$. Word elongation is realized until the random deviation does not start to decrease. As it may be seen in Figure 2, smaller seeds allow the algorithm to generate words α corresponding to the first peak (local maximum) of $r(\alpha)$. To produce a longer significant word α, corresponding to the second peak of $r(\alpha)$, a longer seed has to be taken as a starting string. In all our computational experiments, $r(\alpha)$ showed

```
W:=∅;
ForEach α ∈ D0:
    Elongate(α, W)
W := W \ D0;
Return W
```

Listing 1. Extraction Algorithm

```
if r(αx) ≤ r(α), ∀x ∈ Γ then W := W ∪ {α}
else ForEach x ∈ Γ
        if r(αx) > r(α) then Elongate(αx, W)
```

Listing 2. Elongation procedure: Elongate (α, W)

to have only two peaks, whose localization depends on the genome length.

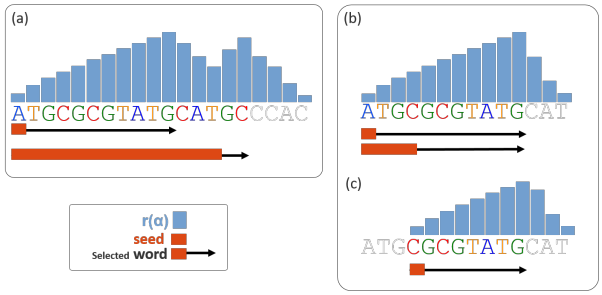


Fig. 2. Expansion procedure

We would like to extract all the words α such that both $\alpha[1, |\alpha|-1]$ and αx (where αx is any elongation of α occurring in G at least once) own a lower level of significance, namely a lower random deviation, with respect to α. The goal can be reached by examining all the words within G from monomers up to a word length equal to the maximum repeat length of G, and by discarding hapaxes. However, such an approach turns out highly expensive, and it cannot be applied efficiently for long genomes. Thus, we developed an expansion procedure with the aim of elongating seed words, let say monomers, up to more meaningful words. The (variable length dictionary) extraction algorithm, combining word elongation and random deviance test (in the expansion procedure) is given by two recursive functions in Listings 1 and 2, where D_0 denotes the set of seeds $D_{k_0}(G)$.

The main idea is to compare the random deviation of a word with those of its elongations. If an elongation results in a word more significant than its root, then the root word is discarded and the elongated word is selected. The process is applied recursively over the word branching of the selected elements (see Listing 2). Seeds are discarded from the output dictionary. Three steps are implemented to compute random deviations. For all factors α of the genome i) RDD of the current word α is computed, by also removing distribution noise (peaks) and transforming R_α into a probability distribution; ii) an exponential distribution E_α is computed from R_α and normalized to be a probability distribution; iii) random deviation r_α is computed by means of the Kullback-Leibler (entropic) divergence.

 ISBN: 978-1-61208-859-4

We employ two elongating functions (along both directions of the genome double string) and the resulting dictionary is the union of the dictionaries obtained with the two elongations. We refer with W_{L2R} and W_{R2L} as the dictionaries extracted by following the $5' - 3'$ and $3' - 5'$ verses, respectively, and with $W = W_{L2R} \cup W_{R2L}$ as the resulting dictionary.

C. Dictionary evaluation

Extracted dictionaries are evaluated by means of information measurements, such as *the word length distribution* of their elements. Two other parameters are the *sequence coverage*, which is the percentage of positions i in the genome such that $G[j,k]$ is a word of the extracted dictionary D for $j < i < k$, and the *average positional coverage*, which is the average over positions i of number of words $G[j,k]$ for $j < i < k$ of the dictionary D. They are denoted by $cov(G,D)$ and $avg(covp(G,D))$, respectively. A *good* dictionary must have a high sequence coverage, but a low overlapping degree among its elements. In fact, if we consider $D_k(G)$ as a language, for a certain value of word length k, then it has the maximum sequence coverage (all positions of the genome would be involved by at least one k-mer) but also the maximum positional coverage, since each position of the sequence is involved by up to k different words of the dictionary. On an ideally good dictionary, both parameters are close to one, meaning that its words cover almost the entire genome and tend to not overlap.

III. Results

Both algorithms described in previous section were run over all human chromosomes belonging to the reference assembly hg19.

A. Dictionaries extracted by the V-algorithm

Table I shows the number of extracted words (that is, dictionary sizes), for each single human chromosome, and their union at the bottom, for both the algorithm in [24] and the V-algorithm, by starting from different seed lengths, and by implementing two filters as redundancy strategies: one discarding duplicates (same words coming from different seed lengths) and the other discarding prefixes (in order to estimate the relative amount of prefixes).

The result is that the V-algorithm is able to select a smaller set of words, with a lower gap between the two redundancy discarding strategies. This is essentially due to the fact that the higher is k the lower are the C measures of k-mers. Therefore, comparing the C measure of a word, relatively longer than k_0, with the measure of its proper prefix is more restrictive than a comparison with the measure of the initial word of length k_0. From this behaviour, we can speculate that the V-algorithm selects words with an higher semantic meaning.

In Table I, it is evident that the V-algorithm extracts a smaller amount of duplicates and prefixes than the algorithm in [24] (even when starting from seeds with different length). Indeed, *smaller variable length dictionaries were extracted* by the V-algorithm, with fewer duplicate discarding steps, and a smaller amount of prefixes (which needed to be discarded in the original algorithm).

TABLE I
NUMBER OF EXTRACTED WORDS BY THE ORIGINAL AND MODIFIED ALGORITHMS

Chr	Orig. no dup.	Orig. no pref.	ratio	V-algo no dup.	V-algo. no pref.	ratio
1	276,178	210,728	0.763	57,064	57,055	1.000
2	281,698	227,544	0.808	119,582	118,368	0.990
3	259,805	203,888	0.785	102,640	101,142	0.985
4	251,067	201,760	0.804	108,229	106,879	0.988
5	259,167	207,300	0.800	112,846	111,581	0.989
6	255,025	198,487	0.778	106,193	104,510	0.984
7	269,392	208,465	0.774	113,139	111,840	0.989
8	259,586	206,241	0.794	118,551	117,295	0.989
9	212,362	152,523	0.718	33,886	33,878	1.000
10	234,663	186,844	0.796	100,616	99,595	0.990
11	249,374	188,012	0.754	94,484	93,417	0.989
12	247,842	187,931	0.758	99,147	97,579	0.984
13	176,546	149,563	0.847	81,634	78,868	0.966
14	209,881	162,515	0.774	94,312	90,313	0.958
15	207,173	177,125	0.855	107,114	103,917	0.970
16	229,208	166,653	0.727	62,732	62,673	0.999
17	204,905	160,475	0.783	85,091	84,303	0.991
18	161,710	131,900	0.816	65,985	65,558	0.994
19	258,781	197,822	0.764	123,913	122,541	0.989
20	171,474	131,434	0.766	66,320	65,597	0.989
21	130,763	100,427	0.768	50,698	50,233	0.991
22	147,002	120,259	0.818	77,797	74,511	0.958
X	279,938	213,093	0.761	124,793	123,006	0.986
Y	194,014	137,284	0.708	66,088	65,986	0.998
union	4,281,701	3,737,766	0.873	1,813,776	1,798,241	0.991

B. Dictionaries extracted by the W-algorithm

The RDD-based W-algorithm was applied (with values for seed length from the range $1 - 12$) to extract genomic dictionaries from each human chromosome, and some analysis was performed also on the union of such 24 dictionaries. However, here we show data only for some (more explicable) chromosomes, for (more significant) seed lengths up to 8.

TABLE II
WORD LENGTH DISTRIBUTION OF HUMAN CHROMOSOME 1

	k_0							
k	1	2	3	4	5	6	7	8
4	2	13	20					
5	31	134	202	272				
6	**63**	**349**	**517**	**995**	**1,261**			
7	57	180	232	350	475	1,343		
8	57	193	277	430	679	3,001	10,668	
9	10	144	241	529	1,073	7,602	29,521	53,314
10	5	**201**	**326**	**794**	**1,391**	**9,126**	59,951	129,872
11	2	151	233	569	923	4,302	**63,089**	**184,296**
12		64	91	198	323	973	24,275	97,646
13		21	30	51	81	225	4,592	20,670
14		2	3	10	18	40	875	3,525
15		2	2	5	6	11	190	724
16		4	5	5	5	9	54	165
17		1	1	2	2	3	17	54
18							5	19
19								5
20								6
21								3
22								6
23								1

The Word Length Distribution (WLD) related to human chromosomes 1 is shown in Table II by reporting the cardinality of words having a given length and being generated by starting from a given seed length. A common feature is to have two modes in the k-dictionary sizes, that is, two local maximum values (indicated in bold) for some lengths k. In

 ISBN: 978-1-61208-859-4

Table II, such values are 6 (for seeds long from 1 to 5) and 10-11 (for seeds long from 2 to 8). Although they do not have fixed values (for tests performed on the other human chromosomes and not shown here), they are not very variable.

Another empirical result, confirmed on all the other chromosomes, is that the dictionary generated by starting from seeds $k-1$ long is a proper subset of that generated by starting from seeds k long, apart of the words long k. In fact, words with the same length of the seed are eliminated by the algorithm and do not appear in the WLD tables.

Extracted dictionaries are evaluated according to both their sequence and their (average) positional coverage: these data related to chromosome 1 are reported in Table III and Table IV respectively, where it is clear that parameter goodness does not increase with the word or seed length k_0.

TABLE III
HUMAN CHROMOSOME 1: SEQUENCE COVERAGE VALUES

	k_0							
k	1	2	3	4	5	6	7	8
4		0.0291	0.0291					
5	0.0309	0.0790	0.1362	0.1681				
6	0.0269	**0.3149**	**0.5504**	**0.7767**	**0.8426**			
7	**0.0742**	0.2479	0.3878	0.6430	0.7691	**0.8141**		
8	0.0285	0.0616	0.0899	0.1187	0.1384	0.1643	**0.2634**	
9	0.0115	0.0209	0.0303	0.0499	0.0615	0.0714	0.1593	**0.6315**
10	0.0008	0.0054	0.0071	0.0128	0.0206	0.0329	0.0974	0.5388
11	**0.0025**	**0.0077**	**0.0088**	**0.0108**	**0.0127**	0.0174	0.0602	0.3509
12		0.0028	0.0031	0.0081	0.0089	0.0101	0.0342	0.2858
13	0.0000	0.0006	0.0013	0.0054	0.0065	0.0070	0.0155	0.1209
14	0.0035	0.0048	0.0049	0.0056	0.0065	0.0066	0.0101	0.0451
15	0.0026	0.0036	0.0036	0.0050	0.0052	0.0052	0.0065	0.2140
16		0.0016	0.0017	0.0017	0.0071	0.0028	0.0032	0.0090
17		0.0011	0.0011	0.0012	0.0013	0.0013	0.0014	0.0031
18		0.0006	0.0006	0.0006	0.0006	0.0012	0.0012	0.0020
19							0.0000	0.0003
20							0.0000	0.0002
21								0.0001
22								0.0000
23								
24								0.0000

By observing the data in Table III, the best coverage of the chromosome (corresponding value 0.84) is obtained by the examers obtained starting from 5-mers as seeds, while the average positional coverage of such a dictionary is 2.7715 (see Table IV), which is far from one. However, this dictionary was our choice for the chromosome clustering analysis described below, because we gave a priority of importance to sequence coverage. Relatively to only positional coverage values, in Table IV we may notice that words of length 10 (or longer, for instance 15) exhibit good (i.e., less than 2) values for any seed length up to 7, while examers have good positional coverage with shorter seeds (long up to 3).

Finally, we extracted dictionaries of examers on each single human chromosome, and from their pairwise intersections, in absolute and relative terms, we found interesting results, reported in Figure 3, where four groups of chromosomes may be identified at the second level of the dendrogram, having cardinalities of dictionary intersection of the same order of that of the extracted dictionary from each single chromosomes (see leaves of the dendogram). Our dictionary based method was then capable to discriminate by structure similarity the following clusters of human chromosomes.

TABLE IV
HUMAN CHROMOSOME 1: AVERAGE POSITIONAL COVERAGE

	k_0							
k	1	2	3	4	5	6	7	8
4		1.0078	1.0078					
5	1.0807	1.1690	1.2411	1.4198				
6	**1.1539**	**1.3022**	**1.6590**	**2.3201**	**2.7715**			
7	1.0934	1.2876	1.4587	1.9817	2.5877	2.9160		
8	1.1569	1.2590	1.3125	1.4228	1.5184	1.5836	1.5572	
9	**1.4480**	**1.5411**	**1.5211**	**1.7039**	**1.8791**	**1.8661**	1.5470	1.7484
10	1.0006	1.1090	1.1033	1.1697	1.1926	1.2632	1.2580	1.5457
11	**4.0810**	**2.1729**	**2.0809**	**1.9100**	**1.7829**	**1.6131**	**1.3009**	1.3658
12		1.0654	1.0624	1.1926	1.1809	1.1716	1.1507	1.3455
13	1.0000	1.0000	1.0000	1.1355	1.3769	1.3530	1.2340	1.3709
14	1.0000	1.0000	1.0000	1.0551	1.2244	1.2235	1.1687	1.3807
15	1.000	1.1446	1.1445	1.1065	1.1739	1.1725	1.1444	1.2559
16		1.2684	1.2636	1.2588	1.2539	1.1544	1.1447	1.1148
17		1.0000	1.0000	**1.3982**	**1.3957**	**1.3948**	**1.3608**	**1.3440**
18		1.0000	1.0000	1.0000	1.0000	1.0000	1.0015	1.0187
19							1.0000	1.0000
20							1.0000	1.0000
21								1.0000
22								1.0000
23								
24								1.0000

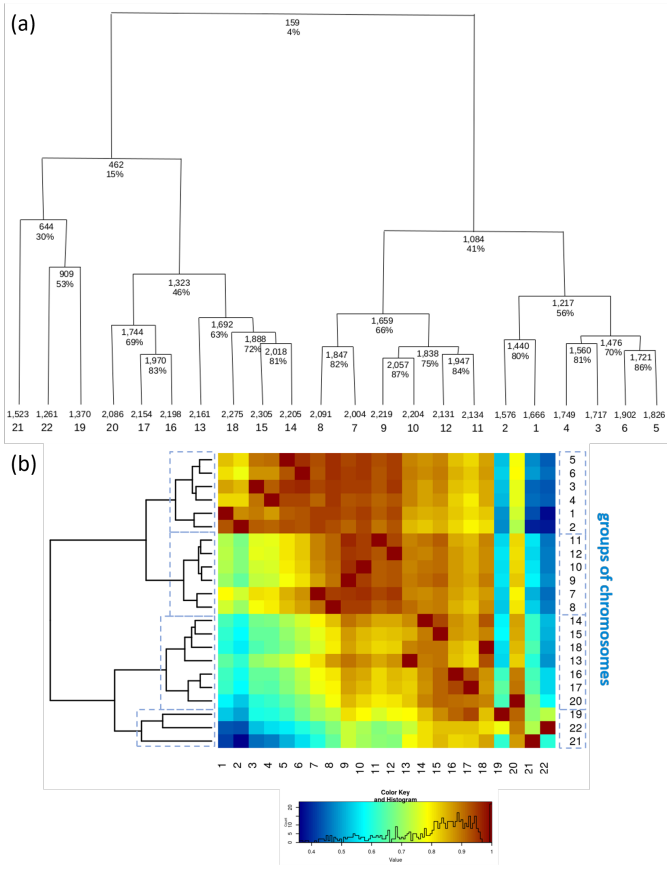


Fig. 3. **(a)** Human chromosome similarities percentages. **(b)** Heatmap of human cromosome similarity.

The dictionary of examers obtained by the algorithm from seeds long 5 was here employed to cluster all human chromosomes (see Figure 3). All chromosomes share very few examers (159 are common to all, over the 1,666 extracted words) which we exhibit as informative conserved sequences, a sort of product by evolution selection, to be further analyzed for their biological characterization.

 ISBN: 978-1-61208-859-4

IV. Conclusions and Discussion

Given a genome, we extract a specific set of its factors which represent the building blocks, or semantic units, of a dictionary significant for the genome language. In this work, we have described an information theoretical methodology to extract relatively small genomic dictionaries, which have good properties in terms of genome coverage.

Three methods were presented. One from the literature, introduced in [24], which was our starting point in terms of basic ideas, the second method is a variant of this, called V-algorithm, more efficient and appropriate to extract genomic dictionaries, and finally, our RDD based W-algorithm, which originally combines a criterion of anti-randomness with a criterion of elongation of seeds to select variable length factors. The application of the state of the art methodology and the V-algorithm to human chromosomes show that both algorithms often fail in extending seeds, and when they success, they more-likely extract very long words, which sparsely cover the investigated sequences. The point of our approach is to produce relatively small dictionaries with both sequence and average positional coverage as close as possible to one. The goal is reached thanks to the proposed W-algorithm. We have shown that preferred seed lengths emerge, from an observation of sequence and positional genome coverage that provide a better coverage. Moreover, dictionaries of examers were identified to reveal a clear similarity pattern for human chromosomes.

References

[1] G. S. Ginsburg and H. F. Willard, Eds., *Genomic and Precision Medicine – Foundations, Translation, and Implementation*. (Third Edition): Elsevier, 2017.

[2] R. Mantegna et al., "linguistic features of noncoding dna sequences," *Physical Review Letters*, vol. 73, no. 23, pp. 3169–72, 1994.

[3] D. B. Searls, "The language of genes," *Nature*, vol. 420, pp. 211–217, 2002.

[4] M. Sadovsky, J. Putintseva, and A. S. Shchepanovsky, "Genes, information and sense: Complexity and knowledge retrieval," *Theory in Biosciences*, vol. 127, no. 2, pp. 69–78, 2008.

[5] S. Neph et al., "An expansive human regulatory lexicon encoded in transcription factor footprints," *Nature*, vol. 489, pp. 83–90, 2012.

[6] G. Franco and V. Manca, "Decoding genomic information," in *Computational Matter*, S. Stepney, S. Rasmussen, and M. Amos, Eds. Springer, Cham, 2018, ch. 9, pp. 129–149.

[7] U. Ferraro Petrillo, G. Roscigno, G. Cattaneo, and R. Giancarlo, "Informational and linguistic analysis of large genomic sequence collections via efficient hadoop cluster algorithms," *Bioinformatics*, vol. 34, no. 11, pp. 1826–1833, 2018.

[8] Z. Zhang et al., "Statistical analysis of the genomic distribution and correlation of regulatory elements in the encode regions," *Genome Res.*, vol. 17, no. 6, pp. 787–97, 2007.

[9] M. Ortuno, P. Carpena, P. Bernaola-Galván, E. Munoz, and A. Somoza, "Keyword detection in natural languages and DNA," *EPL (Europhysics Letters)*, vol. 57, no. 5, p. 759, 2007.

[10] T. E. P. Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–72, 2012.

[11] F. Zambelli, G. Pesole, and G. Pavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era," *Briefings in bioinformatics*, p. bbs016, 2012.

[12] A. Castellini, G. Franco, and V. Manca, "A dictionary based informational genome analysis," *BMC Genomics*, vol. 13, no. 1, p. 485, 2012.

[13] V. Mäkinen, D. Belazzougui, F. Cunial, and A. Tomescu, *Genome-Scale Algorithm Design: Biological Sequence Analysis in the Era of High-Throughput Sequencing*. Cambridge: Cambridge University Press, 2015.

[14] S. P. Garcia, A. J. Pinho, J. M. Rodrigues, C. A. Bastos, and P. J. Ferreira, "Minimal absent words in prokaryotic and eukaryotic genomes," *PLoS One*, vol. 6, no. 1, 2011.

[15] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21, no. suppl_1, pp. i351–i358, 2005.

[16] I. Grissa, G. Vergnaud, and C. Pourcel, "Crisprfinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W52–W57, 2007.

[17] J. Qian and M. Comin, "Metacon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage," *BMC Bioinformatics*, vol. 20, no. 367, 2019.

[18] M. Aharon, M. Elad, and A. Bruckstein, "K-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions*, vol. 54, p. 4311–4322, 2006.

[19] F. Zhou, V. Olman, and Y. Xu, "Barcodes for genomes and applications," *BMC Bioinformatics*, vol. 9, p. 546, 2008.

[20] S. Vinga, "Information theory applications for biological sequence analysis," *Briefings in bioinformatics*, vol. 15, no. 3, pp. 376–389, 2013.

[21] G. E. Sims, S. Jun, G. A. Wu, and S. Kim, "Alignment-free genome comparison with feature frequency profiles (ffp) andoptimal resolutions," *PNAS*, vol. 106, no. 8, pp. 2677–82, 2009.

[22] B. Chor et al., "Genomic dna *k*-mer spectra: models and modalities," *Genome Biology*, vol. 10, p. R108, 2009.

[23] Y. Zheng et al., "Evolutionary mechanism and biological functions of 8-mers containing cg dinucleotide in yeast," *Chromosome Research*, vol. E-pub ahead of print], pp. 1–17, 2017.

[24] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. Coronado, and J. Oliver, "Level statistics of words: Finding keywords in literary texts and symbolic sequences," *Physical Review E*, vol. 79, no. 3, p. 035102, 2009.

[25] M. Hackenberg, A. Rueda, P. Carpena, P. Bernaola-Galván, G. Barturen, and J. L. Oliver, "Clustering of DNA words and biological function: A proof of principle," *Journal of theoretical biology*, vol. 297, pp. 127–136, 2012.

[26] G. Franco and A. Milanese, "An investigation on genomic repeats," in *Conference on Computability in Europe – CiE*, ser. Lecture Notes in Computer Science, vol. 7921. Springer, 2013, pp. 149–160.

[27] A. Thomas and T. M. Cover, *Elements of Information Theory*. John Wiley, 1991.

[28] J. H. Holland, *Emergence: from chaos to order*. Perseus books: Cambridge, Massachusetts, 1998.

[29] S. G. Kong et al., "Quantitative measure of randomness and order for complete genomes," *Phys Rev E*, vol. 79, no. 6, p. 061911, 2009.

[30] P. Kolekar, M. Kale, and U. Kulkarni-Kale, "Alignment-free distance measure based on return time distribution for sequence analysis: Applications to clustering, molecular phylogeny and subtyping," *Molecular phylogenetics and evolution*, vol. 65, pp. 510–22, 2012.

[31] A. S. Nair and T. Mahalakshmi, "Visualization of genomic data using inter-nucleotide distance signals," *Proceedings of IEEE Genomic Signal Processing*, vol. 408, 2005.

[32] V. Afreixo, C. A. Bastos, A. J. Pinho, S. P. Garcia, and P. J. Ferreira, "Genome analysis with inter-nucleotide distances," *Bioinformatics*, vol. 25, no. 23, pp. 3064–3070, 2009.

[33] C. A. Bastos, V. Afreixo, A. J. Pinho, S. P. Garcia, J. Rodrigues, and P. J. Ferreira, "Inter-dinucleotide distances in the human genome: an analysis of the whole-genome and protein-coding distributions," *Journal of Integrative Bioinformatics*, vol. 8, no. 3, p. 172, 2011.

[34] L. Gatlin et al., *Information theory and the living system*. Columbia University Press, 1972.

[35] S. P. Harter, "A probabilistic approach to automatic keyword indexing," Ph.D. dissertation, University of Chicago, 1974.

[36] P. Carpena, P. Bernaola-Galván, and P. C. Ivanov, "New class of level statistics in correlated disordered chains," *Physical review letters*, vol. 93, no. 17, p. 176804, 2004.

[37] W. Feller, *An introduction to probability theory and its applications*. John Wiley & Sons, 1968, vol. 1.

[38] A. Kolmogorov, "On tables of random numbers," *Theoretical Computer Science*, vol. 207, no. 2, pp. 387–395, 1998.

[39] J. A. Nelder and R. Mead, "A simplex method for function minimization," *The computer journal*, vol. 7, no. 4, pp. 308–313, 1965.

 ISBN: 978-1-61208-859-4

A Bioinformatics Pipeline for Evaluating Protein Misfolding Impact on the Tertiary Structure in Alzheimer's Disease

Antigoni Avramouli, Eleftheria Polychronidou, Panayiotis Vlamos
BiHELab – Bioinformatics and Human Electrophysiology Lab
Department of Informatics of Ionian University
Corfu, Greece
e-mail: c15avra@ionio.gr, c13poly@ionio.gr, vlamos@ionio.gr

***Abstract*— Alzheimer disease (AD) is the most common cause of neurodegenerative disorder in the elderly individuals. To support the biomarker research on Alzheimer's Disease progression, this study describes a bioinformatics pipeline for the evaluation of the mutations impact on the tertiary structure of AD causative genes.**

Keywords: protein structure; protein misfolding; machine learning; Alzheimer's Disease.

I. INTRODUCTION

Proteins are large, complex biomolecules made up of amino acids. Proteins play a significant role in almost all biological processes. The functional properties of proteins rely upon their three-dimensional structures. The three-dimensional structure arises because the polypeptide chains fold to produce (starting from linear sequences) compact and independent structural regions with specific structures. Predicting the three-dimensional structure of proteins by their amino acid sequence contributes to understanding their biological function. Prediction is not always possible: despite the remarkable efforts of recent years, the problem of folding remains one of the major problems in molecular biology. In addition, proteins that do not get the right configuration can bind abnormally to other biomolecules, as well as form aggregates that are highly toxic to the body [1]. Aggregates are organized into fibrillar structures, a common feature of many neurodegenerative diseases [2].

Alzheimer's Disease (AD), characterised as a protein misfolding disease, is the most common progressive form of dementia [3]. Typical pathological findings are misfolded and aggregated amyloid-β (Aβ) peptides and intracellular neurofibrillary tangles of tau protein. The most well-known predisposing genetic factor for the disease is the presence of the e4 allele of apolipoprotein E (ApoE) [4]. In the e4 allele (frequency 13.7%), the codon 112 has been replaced by arginine. However, the frequency of the e4 allele increases dramatically to ~ 40% in patients with AD. This mutation is associated with a change in the tertiary structure of the protein and the accumulation of β-amyloid in neurons, as well as with the induction of inflammatory responses, while it is the most prone isoform to proteolysis. In this context, changes in the tertiary structure of proteins, which are components of major signaling pathways of AD, could justify the genetic background of this heterogeneous disorder.

In recent years, the correlation of the different tertiary structures of the isoforms of the ApoE gene with the pathogenesis of AD has been studied worldwide [5, 6]. In particular, a study published by the Paralvrez-Marin group in Sweden proposed a computational model of the abnormal interaction of the β-amyloid peptide with the e4 isoform of ApoE, due to the incorrect tertiary structure of the second [7]. However, apart from ApoE-related studies, to date, changes in the tertiary form of proteins due to gene mutations have not yet been investigated in AD. Prior to the discovery of mutations in genes associated with disease onset, no molecular signaling pathways were implicated. Recent genetic studies have identified many candidate genes that are associated with an inherited form of AD. Even if mutations in these genes account for a small proportion of Familial AD (FAD), knowledge of these genes and correlated biochemical cascades will provide several potential targets for treatment of AD and aging-related disorders. Also, the different pathogenetic mechanisms of the disease involve a combination of genetic factors (with different severity for the disease from person to person), indicating that it is essentially a set of disorders with common characteristics rather than a distinct disease.

The present research paper aims to contribute to the reduction of the research gap created by the study of the tertiary structure, to understand the pathogenesis of the disease. In recent years, research interest has focused on identifying all the genetic sites associated with the disease and the different alleles of these genes using high-resolution technologies. In contrast, there is the tertiary form of these mutant proteins, which has not yet been studied in depth. In addition, some of the AD-related proteins have not yet had their crystal structure determined.

Approaches that allow the prediction of three-dimensional structures of proteins through computers are relatively new in the medical sciences [8], but their contribution is increasingly recognized as a tool for characterizing changes in the structure of proteins and detecting rare molecular events. These principles make it easier for us to understand how the protein structure is created, to identify common structural issues, to relate structure and function, but also to see the fundamental relationships between different proteins. Deciphering the mechanisms of the loss of the tertiary structure of a protein is essential for understanding the pathogenesis of diseases, such as AD and essential for explaining neuronal damage during aging.

 ISBN: 978-1-61208-859-4

This pipeline is described by four steps: (a) the evaluation of the online prediction tools and the selection of the most suitable for AD protein structures, (b) the prediction of the mutated structures, (c) the AI/ML classification of the tertiary structures into discrete groups and (d) the evaluation of the pathogenicity of each group to gain evidence for the impact of the mutations and to suggest a characterization for the mutations with unclear etiology. This is an on-going research and thus preliminary results on Presenilin one will be presented here.

II. METHODS

The first step towards the implementation of the pipeline is to collect data from biological databases, to evaluate the existing data and finally to apply machine learning approaches and classify proteins into groups with similar characteristics.

A. Data Consolidation

Here some of the most AD pathogenic mutated alleles will be studied. As many of these mutations affect protein stability, modeled protein structures for the mutant proteins will be compared with the native protein to evaluate stability changes. The genetic loci that will be analysed further through protein 3D structure include APP (Amyloid precursor protein), PSEN1 (Presenilin one), PSEN2 (Presenilin two), CLU (Clusterin), CR1 (Complement receptor 1), PICALM (Phosphatidylinositol binding clathrin assembly protein), BIN1 (Myc box- dependent- interacting protein 1), ABCA7 (ATP binding cassette transporter 7), MS4A (Membrane- spanning 4- domains, subfamily A), EPHA1 (Ephrin type-A receptor 1), CD33 (CD33 antigen), CD2AP (CD2 associated protein), SORL1 (Sortilin-related receptor 1), TPEM2 (Triggering receptor expressed on myeloid cells 2) [9]. These genes are linked to inflammation, oxidative stress, vascular regulation, immune system function, and the function of specific proteases.

Successful mapping of these genes and their association with the onset of the disease has led to the formulation of the amyloid hypothesis [10]. This hypothesis sets as the main pathogenetic mechanism the increased production of β amyloid peptide fragments. Nevertheless, there are cases where the onset of symptoms occurs at a much younger age. In a unique clinical case so far, the onset of the disease occurred in the mid-forties and in some people from the age of thirty. Members of this family had a mutation in the PSEN1 gene (Presenilin 1 E280A) [11]. The mutations related to the proteins were identified through literature and used for the next steps of this pipeline. More particular, so far 69 mutations were identified for APP, 112 for MART, 326 for PSEN1, 68 for PSEN2, and 68 for TREM2.

B. Evaluation of Protein Structures

Since the three-dimensional shape of most of the related proteins is not determined through experimental methodologies, the most established servers were evaluated for predicting the mutated structures and estimate the impact of the mutations to the 3-dimensional structure. A list of the selected methodologies is presented on the Table I below:

TABLE I. LIST OF SELECTED METHODOLOGIES

Methodology	Description	How was used
Uniprot [12]	A comprehensive resource for protein sequence and annotation data	To understand the protein function, and the most related protein structures
PolyPhen-2 [13]	A tool which predicts possible impact of an amino acid substitution on the structure and function	To understand how mutations affect the structure and function of the protein
iTASSER [14]	A hierarchical approach to protein structure prediction and structure-based function annotation	To predict the mutated and unmutated 3D protein structures
PDBeFold [15]	An interactive service that allows you to identify structures that are similar to that of your reference protein	To compare the mutated and unmutated structures on residues level
CATH / Gene3D [16]	A protein family classification methodology	To identify if there is any relationship between mutations impact and protein families

The methodologies are currently used based on the order of the table, to determine the protein structures and understand in detail the impact of the mutations to the proteins. Furthermore, STRING [17] server is used to analyse protein-protein association networks and assess any change that might occur on the mutated protein networks (Figures 1&2).

C. Clustering of protein structures

To analyze further the mutated structures, an established methodology from the field of 3D object recognition was applied [18]. The combination of the above local descriptors was applied to the 3D structures to extract the appropriate features for the comparison.

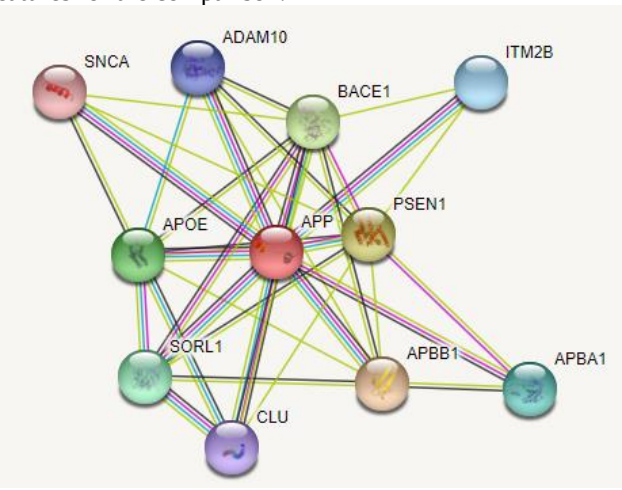


Figure 1. Example of APP network in STRING network analysis.

In order to evaluate the accuracy of clustering using the 3D descriptors, a first round of experiments was conducted, using an annotated dataset. This dataset included every mutated structure while the label of each structure was aligned with the pathogenetic impact of each mutation according to literature data.

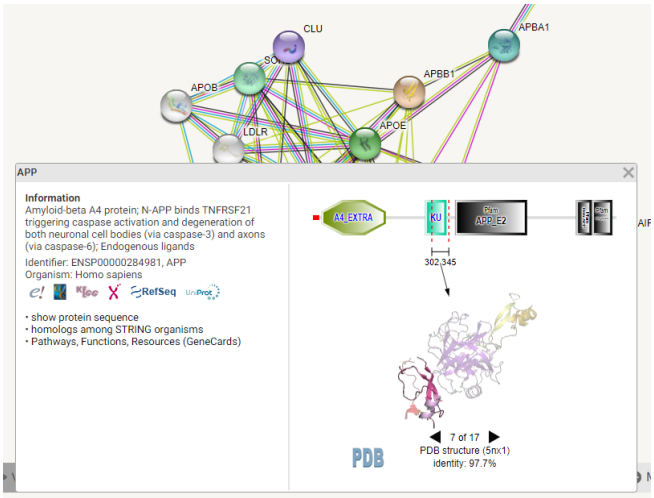


Figure 2. Structural Analysis of the protein. Each protein is mapped to the experimental determined structures (one or more) included in PDB(e).

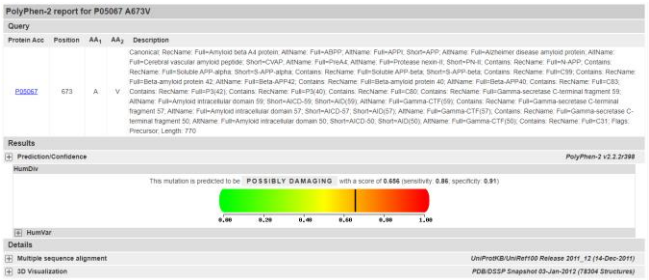


Figure 3. Example of PolyPhen-2 output for the A673V mutation of the APP protein. Percentage of prediction is taken into consideration for the annotation of the clustering output.

The k-medoids, Agglomerative Hierarchical clustering and Density-based spatial clustering of applications with noise (DBSCAN) methods were used to cluster the data using the extracted descriptors [19]. After extracting descriptors from each pair of aligned proteins, the root-mean-square distances (RMSD) between each pair of descriptors is computed, forming a square distance matrix.

In this computational analysis work, preliminary results of our study on PSEN1 mutations are presented and are compared to available clinical data for PSEN1 variants known to cause AD (Figures 3&4). To the best of our knowledge, this is the first study of its kind investigating performing comparative and ab initio prediction of protein structure for mutated forms of PSEN1. The experimental results verify that the use of 3D descriptors can be effectively applied to distinguish structural differences of proteins based on the pathogenic categories of the mutations.

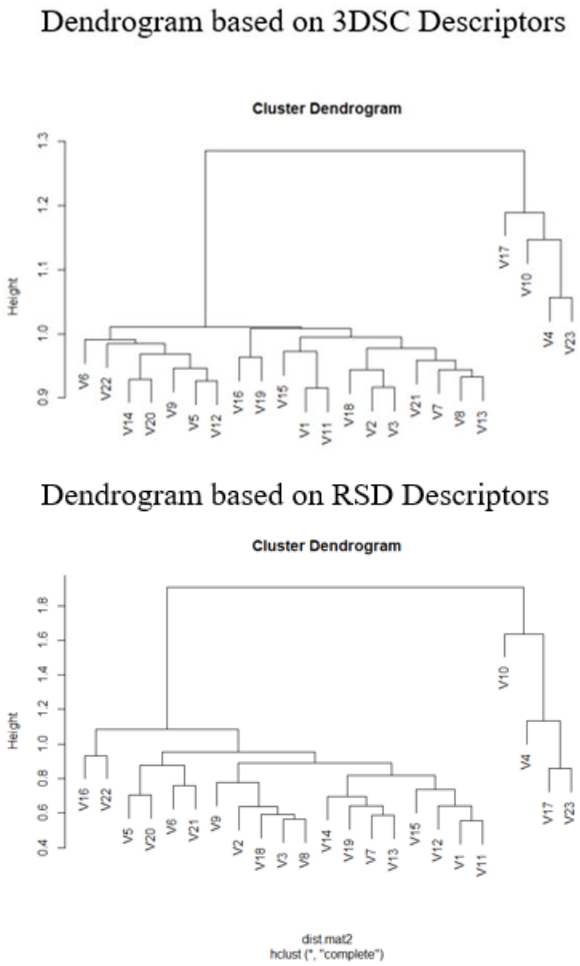


Figure 4. Results of hierarchical clustering are presented for the PSEN1 protein based on the 3DCS and the RSD descriptors type that was applied to each case.

The same process will be repeated for all the other proteins related to the AD progression described in the Data Consolidation section. However, due to the limitations in the prediction time of the online servers, the proof of concept of the PSEN1 is presented here.

III. CONCLUSION

It is known that there is no cure for AD to date. The collective failure of recent clinical trials in the treatment of AD suggests the need for a fuller understanding of the complex biological processes underlying this disease to develop effective, targeted therapeutic approaches. To date, several genetic sites have been identified that are involved in the onset or evolution of AD. Also, AD, like other neurodegenerative diseases, seems to be a biological phenomenon distinct from the phenomenon of normal aging and not an accelerated and pathological version of it. These

 ISBN: 978-1-61208-859-4

data indicate the multiplicity of etiological factors that contribute to the occurrence of AD.

The therapeutic targeting of protein folding has created unique challenges for the discovery and development of new drugs. To achieve this, we must first understand the dynamic nature of the protein species involved and discover the structure and folding of each protein (formation of monomers, oligomers or insoluble aggregates) as well as whether this leads to cell toxicity. To date, our lack of understanding of how proteins interact with other cell proteins and the lack of well-characterized biomarkers that can be used in clinical trials is another bet for the research community.

In the present study, a comprehensive methodology for the analysis of the impact of the AD related proteins is presented. Based on the approach, a combination of well-established online tools can support the prediction of 3D protein structures that have not been determined experimentally yet. Furthermore, the use of Poly-phen2 and CATH can support the identification of evidence of the impact of mutations to the protein structure. Finally, a combination of bioinformatic and object recognition clustering methodology is applied to group the tertiary structures. The annotation of the groups based on the pathogenic characterization of the mutations along with the networks produced by STRING server can reveal evidence on how each mutation affects the protein network.

As mentioned in Section II, the prediction process through online servers consumes significant time and thus a proof of concept is presented here. Since this is an on-going work, the complete analysis will be available as soon as the models are obtained.

ACKNOWLEDGMENT

This research is co-financed by Greece and the European Union (European Social Fund- ESF) through the Operational Programme «Human Resources Development, Education and Lifelong Learning 2014-2020» in the context of the project “Analysis of the tertiary protein structure and correlation of mutations with the clinical characteristics of Alzheimer's disease”, Project no. 5067210.

REFERENCES

[1] C. M. Dobson. “Protein folding and misfolding”. *Nature* 426, pp. 884–890, 2003.

[2] C. Soto and S. Pritzkow. “Protein misfolding, aggregation, and conformational strains in neurodegenerative diseases”. *Nature neuroscience*, 21(10), pp. 1332–1340, 2018.

[3] V. Vingtdeux, N. Sergeant and L. Buee, “Potential contribution of exosomes to the prion-like propagation of lesions in Alzheimer's disease”. *Front Physiol.* 3 pp. 229, 2012.

[4] B. V. Zlokovic, “ Cerebrovascular effects of apolipoprotein E: implications for Alzheimer disease”. *JAMA Neurol.* 70 pp. 440–444, 2013.

[5] V. V. Giau, E. Bagyinszky, S. S. An and S. Y. Kim, “Role of apolipoprotein E in neurodegenerative diseases”. *Neuropsychiatric disease and treatmen*t, 11, pp. 1723–1737, 2015.

[6] P. Huebbe and G. Rimbach, “Evolution of human apolipoprotein E (APOE) isoforms: Gene structure, protein function and interaction with dietary factors”. *Ageing Research Reviews*, 37, pp. 146–161, 2017.

[7] J. Luo, J. D. Maréchal, S. Wärmländer, A. Gräslund and A. Perálvarez-Marín, “In silico analysis of the apolipoprotein E and the amyloid beta peptide interaction: misfolding induced by frustration of the salt bridge network”. PLoS Comput Biol. 5;6(2) pp. e1000663, 2010.

[8] E. Polychronidou, I. Kalamaras, A. Agathangelidis, L. A., Sutton, X. J. Yan, V. Bikos, et. al, “Automated shape-based clustering of 3D immunoglobulin protein structures in chronic lymphocytic leukemia”. *BMC bioinformatic*s 19.14: 414, 2018.

[9] M. Calabrò, C. Rinaldi, G. Santoro, and C. Crisafulli, “ The biological pathways of Alzheimer disease: a review”. *AIMS neuroscience*, 8(1), pp. 86–132, 2020.

[10] D. J. Selkoe and J. Hardy, “The amyloid hypothesis of Alzheimer’s disease at 25 years”. *EMBO Mol Med.* 8(6), pp. 595–608, 2016.

[11] D. Sepulveda-Falla, L. Chavez-Gutierrez, E. Portelius, J. I. Vélez, S. Dujardin, A. Barrera-Ocampo, F. Dinkel, et al, “A multifactorial model of pathology for age of onset heterogeneity in familial Alzheimer's disease”. *Acta neuropathologica*, 141(2), pp. 217–233, 2021.

[12] UniProt Consortium “UniProt: a worldwide hub of protein knowledge”. Nucleic acids research, 47 (D1), pp. D506-D515, 2019.

[13] I. A. Adzhubei, S. Schmidt, L. Peshkin, V. E. Ramensky, A. Gerasimova, P. Bork and S. R. Sunyaev, “ A method and server for predicting damaging missense mutations”. *Nature* methods, 7(4), pp. 248-249, 2010.

[14] J. Yang and Y. Zhang, “I-TASSER server: new development for protein structure and function predictions”. *Nucleic Acids* Research, 43: W174-W181, 2015.

[15] E. Krissinel and K. Henrick, “Protein structure comparison in 3D based on secondary structure matching (PDBeFold) followed by Ca alignment, scored by a new structural similarity function. In: Andreas J. Kungl & Penelope J. Kungl (Eds.)”, Proceedings of the 5th International Conference on Molecular Structural Biology, Vienna, September 3-7, p.88, 2003.

[16] I. Sillitoe, N. Bordin, N. Dawson, V. P. Waman, P. Ashford, H. M. Scholes, et.al., “CATH: increased structural coverage of functional space”. *Nucleic Acids Res.* 49(D1) pp. D266-D273,2021.

[17] D. Szklarczyk, A. L. Gable, D. Lyon, A. Junge, S. Wyder, J. Huerta-Cepas, M. Simonovic, et. al, “STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets”. *Nucleic Acids Research*, 47(D1), pp. D607-D613, 2019.

[18] E. Polychronidou, A. Avramouli and P. Vlamos, “Alzheimer’s Disease: The Role of Mutations in Protein Folding”. Adv Exp Med Biol., 1195, pp. 227-236, 2020.

[19] M. Ester, H.P. Kriegel, J. Sander, X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”. *KDD Proceedings* pp. 226-231, 1996.

 ISBN: 978-1-61208-859-4