THE 12TH INTERNATIONAL SYMPOSIUM
ON HEALTH INFORMATICS AND
BIOINFORMATICS

17 - 19 OCTOBER 2019

HIBIT 2019
ABSTRACT BOOK

THE 12TH INTERNATIONAL SYMPOSIUM ON HEALTH INFORMATICS
AND BIOINFORMATICS

IZMIR BIOMEDICINE
AND GENOME CENTER

TABLE OF CONTENTS

Welcome Message

The International Symposium on Health Informatics and Bioinformatics, (HIBIT), now in its twelfth year HIBIT 2019, aims to bring together academics, researchers and practitioners who work in these popular and fulfilling areas and to create the much-needed synergy among medical, biological and information technology sectors. HIBIT is one of the few conferences emphasizing such synergy. HIBIT provides a forum for discussion, exploration and development of both theoretical and practical aspects of health informatics and bioinformatics and a chance to follow current research in this area by networking with other bioinformatician.

The Organizing Comitee
Izmir Biomedicine and Genome Center
Ezgi KARACA, Gökhan KARAKÜLAH,
Serap ERKEK, Yavuz OKTAY

Organizing commitee

Conference Co-Chairs

Ezgi KARACA
Gökhan KARAKÜLAH
Serap ERKEK
Yavuz OKTAY

Volunteer Students

Aleyna ERAY
Deniz DOĞAN
Esra ESMERAY
Gülden Özden YILMAZ
Hamdiye UZUNER
Işıl TAKAN
Kaan OKAY
Perihan Yağmur GÜNERİ

Administrative Staff

Burcu İNCE
Özlem DALAN
Hakan ÖZLER

Scientific commitee

Abdullah KAHRAMAN, University Hospital Zurich
Alper KÜÇÜKURAL, University of Massachusetts Medical School
Aslı SUNER, Ege University
Athanasia PAVLOPOULOU, Izmir Biomedicine and Genome Center
Attila GÜRSOY, Koc University
Aybar Can ACAR, Middle East Technical University
Bahar DELİBAŞ, Kadir Has University
Barış FİDANER, Izmir Biomedicine and Genome Center
Barış SÜZEK, Mugla Sıtkı Koçman Universty
Bilge KARACALI, Izmir Institute of Technology
Burcu BAKIR-GÜNGÖR, Abdullah Gul University
Emre GÜNEY, Pompeu Fabra University
Ercüment ÇİÇEK, Bilkent University
Evren KOBAN, Ege University
Jens ALLMER, Wageningen University
Joao RODRIGUES, Stanford University
Levent CAVAS, Dokuz Eylül University
Malik YOUSEF, Zafat Academic College
Maria-Jesus MARTIN, European Bioinformatics Institute
Nurcan TUNÇBAĞ, Middle East Techical University
Ogün ADEBALI, Sabanci University
Oğuz DİCLE, Dokuz Eylül University
Özlen KONU, Bilkent University
Öznur TAŞTAN, Sabancı University
Panagiotis KASTRITIS, MLU Halle
Rengül ÇETIN ATALAY, Middle East Techical University
Tolga CAN, Middle East Techical University
Tuğba ÖZAL İLDENİZ, Acibadem University
Tuğba SÜZEK, Mugla Sıtkı Koçman University
Tunca DOĞAN, Hacettepe University, Institute of Informatics
Türkan HALİLOĞLU, Bogazici University
Uğur SEZERMAN, Acibadem University
Volkan ATALAY, Middle East Techical University
Yeşim AYDIN SON, Middle East Techical University
Zerrin IŞIK, Dokuz Eylül University


IZMIR BIYOTIP
IZMIR BIOMEDIC


P VE GENOM MERKEZI
CINE AND GENOME CENTER
STAR WARS
NEW SOUTH

Program

17th October Thursday

09:00–09:15
Mehmet ÖZTÜRK
Welcome &
Opening Speech

09:15–10:00
Keynote Lecture I
Rita CASADIO
Functional and Structural Features of Disease-Related Protein Variants

10:00–10:45
Keynote Lecture II
Hasan DEMİRCİ
Structure-Based Antibiotic Development Driven by Ambient-Temperature Serial Femtosecond X-ray Crystallography

10:45–11:15 Coffee Break

Omics Technologies

11:15 - 11:40 Nurcan TUNÇBAĞ
Personalized Medicine Guided by Integrative Network Modeling

11:40 - 12:05 Tunahan ÇAKIR
Network-based Analysis of Transcriptome Data to Unravel Molecular Mechanisms Behind Cellular Impairments

12:05 - 12:30 Zerrin IŞIK
Network Based Approaches for Compound Target Identification

12:30 - 12:45 Ercüment ÇİÇEK
SPADIS: An Algorithm for Selecting Predictive and Diverse SNPs in GWAS

12:45–13:45 Lunch

13:45 - 14:00 Poster Mounting

14:00 - 14:25 Özlen KONU
Comparative Transcriptomics of Zebrafish and Mammals

14:25 - 14:50 Burcu BAKIR-GÜNGÖR
Network Oriented Multi-Omics Data Analysis Methodologies to Enlighten the Molecular Mechanisms of Human Complex Diseases

14:50 - 15:05 Ogün ADEBALİ
Comparison of Nucleotide Excision Repair Profiles between Gray Mouse Lemur and Human

15:05 - 15:20 Yasin KAYMAZ
A Hierarchical Random Forest Approach for Cell Type Projections Across Single Cell RNAseq Datasets

15:20 - 15:35 İlayda Betül UÇAR
Assessing the Impact of Proteome Redundancy Minimization in UniProtKB

15:35 - 15:50 Arda SÖYLEV
Discovery of Structural Variations in Ancient Genomes

15:50–16:20 Coffee Break

16:20 - 16:30 QiaGen

16:30 - 16:45 Betül BAZ
Genome-Scale Metabolic Network Reconstruction of Klebsiella Pneumoniae HS11286

16:45 - 17:00 Tuğçe BOZKURT & Umut GERLEVİK
Identification of a Novel Missense Variant in the PRKAR1A Gene and Its Pathogenicity Mechanism

17:00–17:45
Keynote Lecture III
Mehmet SOMEL
Ancient Genomics in Anatolia: Challenges and Opportunities

17:45–19:10
Poster Session I

19:15–22:00
Conference Dinner at Bogazici Restraurant

09:00–09:45

Keynote Lecture IV

Matthias SAMWALD

Transformative Artificial Intelligence in Life Science and Health Care: Mapping the Challenges

Computational Structural Biology and Drug Design

09:45 - 10:30 Serdar DURDAĞI

Integration of Machine Learning, Text Mining, Binary QSAR Models and Target-Driven Virtual Screening Approaches for the Identification of Novel Small Molecule Therapeutics

10:30 - 10:45 Nuray SÖĞÜNMEZ

Allosteric Communications in Inactive States of Human β2-Adrenergic Receptor (β2-AR)

10:45 - 10:55 Seyit KALE

The Role of Computational Modeling in Understanding Nucleosome Dynamics

10:55–11:30 Coffee Break

11:30–12:15

Keynote Lecture V

Türkan HALİLOĞLU

Dynamic Order in Allosteric Interactions of Proteins

12:15 - 12:30 Özge KÜRKÇÜOĞLU

Exploring Allosteric Communication on the Ribosomal Tunnel in Human and Bacteria

12:30 - 12:45 Özge DUMAN

Intrinsic Allosteric Dynamics in G-Protein Coupled Receptors

12:45 - 13:00 Ahmet RİFAİOĞLU

Receptor-Ligand Binding Affinity Prediction via Multi-Channel Deep Chemogenomic Modeling

13:00 - 13:15 Mert GÖLCÜK

Molecular Dynamics Simulations of the Dynein Linker Movement

13:15 - 13:30 Merve ARSLAN

Effects of Vernier Zone Residues on Antibody Multi-Specificity

13:30–14:30	14:30-14:45	14:45-16:15
Lunch	Poster Mounting	Poster Session II

Bioengineering and Health Informatics

16:15 - 16:40 Oğuz DİCLE

Medical Informatics Sailing in the Wind of Artificial Intellicence

16:40 - 17:05 Alper SELVER

Combining Clinical and Molecular Data Using Machine Learning

17:05 - 17:30 Serhat TOZBURUN

Decision Support Systems Based on Artificial Intelligence in Colonoscopy

17:30 - 17:45 Heval ATAŞ

CROssBAR: Comprehensive Resource of Biomedical Relations with Network Representations and Deep Learning

17:45 - 18:00 Rafsan AHMED

MEXCOWalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Modules

18:00–18:15

Closing Ceremony & Travel Awards

KEYNOTE LECTURERS

Hasan Demirci
Stanford University & Koç University

Structure-Based Antibiotic Development Driven by Ambient-Temperature Serial Femtosecond X-ray Crystallography

High-resolution ribosome structures determined by cryo X-ray crystallography have provided important insights into the mechanism of translation. Such studies have thus far relied on large ribosome crystals kept at cryogenic temperatures to reduce radiation damage. Here I will describe the application of serial femtosecond X-ray crystallography (SFX) using an X-ray free-electron laser (XFEL) to obtain diffraction data from ribosome microcrystals in liquid suspension at ambient temperature. 30S ribosomal subunit microcrystals programmed with decoding complexes and bound to either antibiotic compounds or their next-generation derivatives diffracted to beyond 3.4 Å resolution. Our results demonstrate the feasibility of using SFX to better understand the structural mechanisms underpinning the interactions between ribosomes and other substrates such as antibiotics and decoding complexes. We have also determined the structure of large (50S) ribosomal subunit in record-short time by using record-low amount of sample during and XFEL beamtime . This structure is the largest one solved to date by any FEL source to near atomic resolution (3 MDa). We expect that these results will enable routine structural studies, at near-physiological temperatures, of the large ribosomal subunit bound to clinically-relevant classes of antibiotics targeting it, e.g. macrolides and ketolides, also with the goal of aiding development of the next generation of these classes of antibiotics. Overall, the ability to collect diffraction data at near-physiological temperatures promises to provide new fundamental insights into the structural dynamics of the ribosome and its functional complexes.

Mehmet Somel
Middle East Techical University

Ancient genomics in Anatolia: challenges and opportunities

The increasing availability of ancient genomes has been accelerating studies in a number of fields, including human history, molecular evolution, and conservation biology. The METU/Hacettepe Ancient DNA Group has been contributing to these endeavors by producing and analyzing ancient DNA data from archaeological human and animal remains of Anatolian origin, from the last 10,000 years. However, the temperate climate conditions across most parts of Anatolia drive rapid DNA degradation, and most biological material examined harbor only trace amounts of endogenous DNA (median <1% of shotgun sequenced molecules). Hence, the genomes produced are typically only partial (median <0.1x coverage). In this talk, I will discuss different computational approaches to overcome the limitations posed by extremely low coverage data. I will further share examples of new genetic-based insights into Anatolian human societies from 10,000 years ago, including their regional movements and social traditions.

Rita Casadio
University of Bologna

Functional and structural features of disease-related protein variants

Modern sequencing technologies provide an unprecedented amount of data about single-nucleotide variations occurring in coding regions and leading to changes in the expressed protein sequences. A significant fraction of these single-residue variations is linked to disease onset and collected in public databases. In recent years, many scientific studies have been focusing on the dissection of salient features of disease-related variations from different perspectives. Here I will give an overview of what the Biocomputing Group at the University of Bologna provides for functional annotation of disease related variations, including recent developments suited to link genotyping to phenotyping. Starting from proteins whose 3D structure is known, we recently described all the functional and structural features that can be of interest for discriminating among neutral and disease-related variations, including functional pathways mostly affected by the presence of disease-related variation and structural/functional domains that contains the highest number of disease-related variations. Moreover, we highlight interesting physico-chemical features related to the type of variations and their relations to the property of being related to disease or not. Our results support previous findings obtained with much larger data sets of protein sequences and it adds to the notion that not all the disease related variations are perturbing the protein stability.

Matthias Samwald
Medical University of Vienna

Transformative Artificial Intelligence in Life Science and Health Care: Mapping the Challenges

The advances of deep learning spark hopes that Artificial Intelligence could radically transform the life sciences and medicine in the 21st century. While we now have more powerful means of recognizing patterns in large datasets we may ask: is this enough to really make new breakthrough discoveries and to offer better therapies to patients? In this talk I will explore the difficult road from data to knowledge to action, and the questions, risks and opportunities we are facing in our quest to bring Artificial Intelligence to its full potential.

Turkan Haliloglu
Bogazici university

Dynamic Order in Allosteric Interactions of Proteins

Allostery is a fundamental dynamic mechanism that underlies functional long distance interactions in proteins. Hybrid methodologies with their combined novel aspects provide means to study and characterize internal dynamics and relate to protein function. Here, we combine Langevin dynamics and Elastic Network Models (ENM) to explore and understand how proteins act for their function. With the disclosed intrinsic dynamic modes of motion from local to global fluctuations, time-delayed correlation patterns with preceding and lagging residue sites in their fluctuations reveal an order in allosteric communication pathways and directionality in function. This posits key events to modulate or control the function as will be demonstrated with the exemplary cases of large dynamic protein complex systems. Understanding and exploiting intrinsic dynamics of proteins are thus likely to open new ways for design and discovery.

INVITED SPEAKERS

ALPER SELVER

Combining Clinical and Molecular Data Using Machine Learning

Deep Learning has changed the course of data-driven medical image analysis and re-enabled the domination of machine learning techniques. Albeit being very complex in terms of interpretability, convergence, architectural formation and parameter adjustment, the performance of black-box deep models constantly outperforms the existing approaches in various clinical problems. As a data-driven science, the genomics field is and will be highly affected by these new approaches of machine learning to produce novel biological hypotheses. This talk will first set the stage for deep learning tools and what is changed/not-changed compared to their older alternatives: artificial neural networks. Then, it will cover the differences and similarities of the application of emerging deep models to imaging and genomics problems. The current trends for data set preparation, deep model application, and evaluation strategies will be discussed through recent examples from multi-centric organizations and large-scale grand-challenges. Finally, the measures of complementarity and diversity will be presented for the integration of clinical and molecular data using ensembles, which corresponds to the machine learning-based fusion of individual models.

NURCAN TUNÇBAĞ

Personalized Medicine Guided by Integrative Network Modeling

Precision medicine aims to find the best treatment strategy based on the information about the patient's tumor. Molecular heterogeneity is the main obstacle in developing treatment strategies. Therefore, transforming patient specific molecular data into clinically interpretable knowledge is fundamental in precision medicine. In this work, we tackle the mutation profiles of patients with Glioblastoma Multiform (GBM), which is the most aggressive type of brain tumors with a poor survival. Our main motivation is that different mutations, that are spatially in close proximity in the same protein, or function in the same pathway, may result in phenotypically similar tumors. 3D spatial clustering of the mutations, that we call "mutation patch", significantly decreases the heterogeneity. We additionally identify the affected patient-specific subnetworks and pathways that are inferred from mutations. For this purpose we use our Omics Integrator software that solve the prize-collecting Steiner forest problem to integrate a variety of 'omic' data as input and identify putative underlying molecular pathways. Indeed, grouping the patients based on the presence of mutations in close proximity together with network-guided grouping is significantly associated with their survival. These results also enable us to suggest several therapeutic hypotheses for each group based on available drug treatment data. We believe that from mutations to networks and eventually to clinical and therapeutic data, this study provides a novel perspective to the analysis of mutation effect towards the network-guided precision medicine.

OĞUZ DICLE

Medical Informatics in the Era of Artificial Intelligence

Medical informatics is one of the new scientific disinclines mainly developed in the second half of the 20th century. The first organization in this area was held in Germany by the affords of Gustave Wager in 1949. In 1960's first training programs started in France, Belgium and Germany. In early 1970's first research centers were established in Poland and USA. Later, in the mid of 1970's IMIA which is still the head of medical informatics society was founded. In these years MUMPS Language and OS were began to use in practice. This brought a new vision to data management in hospital and medical processes. Early PACS applications were began in 1982. DICOM 1.0 was introduced in 1985. In 1993 İnternet came to the scenery as a game changer together with WWW in 1995. Pub med, which is now a great repository of medical literature was introduced in 1996. Google became the choice of internet surfing just after it has been launched in 1998. We have met with a lot of successful hospital information system implementations from the beginning of 2000's. This was also the beginning of big data age.

As the digitalization of every single element in medicine such as medical images, lab data and the patient medical records, the need for a new field of science in data management emerged. Both data collection, storage, data analysis and data distribution had to be done with totally new tools and new methods. Medical informatics emerged to meet this need in this field. Over time, four separate working areas appeared in this area. One of the these areas is Bioinformatics. Practitioners in this specialty are concerned with storing, retrieving, sharing and helping analyze biomedical information for research and patient care. Subspecializes include chemical, nursing and dental informatics. Mostly those who works in genetics involves this area. Public health informatics is the second field where the professionals involve the use of technology to guide how the public learns abouth health and health care while also ensuring access to the latest medical research. Professionals also ensure public health practices have access to the information they need. Web portals serve the main function in this area and epidemiologists, statisticians and public health departments provide the related research and practice. Clinical informatics is the application of informatics and information technologies for clinical research and patient care. Hospital information systems (HIS) re the main interest of clinical informatics. In recent years, decision support systems become one of the most important components of HIS. Computer engineers are in the center of clinical informatics. The fourth subspecialty in medical informatics is called the imaging informatics. Image processing is the main concern of imaging informatics. Health professionals, especially the radiologists collaborate with engineers who have special interest on signal processing.

Alan Turing was the first man in history who succeeded to compute with the machines. Turing is widely considered to be the father of theoretical computer science and artificial intelligence. After his studies many attempts have been done to utilize computers to solve complex problems in every part of human life. We began to use personal computers by 1971. In the following years many algorithms were developed concerning machine learning. Expert systems mainly dominated these algorithms for decades. Meanwhile some of the scientists inspired from the neural networks expecting to model human learning and decision making. Neural networks widely were used in decision support systems.

In recent years a new concept was introduced to technology world which improved machine learning. It is called deep learning. Deep learning evoked an huge hope and by means of big data and processing capacity we have been estimating to be able human-like thinking, learning and decision making machine.

With the use information systems, medicine and health care gained several benefits. Patient data now is totally under control and can be archived without any loss. It can be distributed all over the world and patients got the opportunity to reach more information about health. However, data and information is always need to be interpreted and analyzed in order to be useful as a mature knowledge. In many medical professions such as radiology human power is no longer sufficient to overcome to interpret the data in a reasonable time. It is not possible for a physician to read millions of new scientific articles published every year. It is not easy and feasible for physicians to handle gene data, lab data, medical images and several other data together to achieve a smart benefit for the patients. Hundreds of other examples can be given with the same dilemma.

Deep learning is a strong candidate for the solution. It is very capable to find the features and attributes from big data which helps to verify, classify and identify the problems. It gives very sensitive and precise possibilities for the right choice. It is expected that deep learning will be a great tool for the physicians before their decisions. IBM Watson is one of the successful tools which revealed the power of machine learning. Deep learning algorithms will be a great time saving method. It will enhance the right diagnosis or treatment at the right time with the best rational reasons.

But AI in medicine also raises significant legal and ethical challenges. Several of these are concerns about privacy, discrimination, psychological harm and the physician-patient relationship. This is one of the challenges that medical informatics have to manage in the future. Computers have become smarter; they are now able to anticipate, detect, and offer suggested actions in response to a given set of conditions. Artificial intelligence in medicine offers humans a chance at better healthcare, with more efficiency and precision.

As a general motto "Medicine is a combination of art and science". Now it will be augmented by AI. In this speech I will elaborate the influence of AI on medical informatics.

ÖZLEN KONU

Comparative Transcriptomics of Zebrafish and Mammals

Zebrafish is a freshwater vertebrate species frequently used as a human disease model. Zebrafish genes are highly orthologous and syntenic to those of mammals allowing for comparisons of changes in transcriptomes due to drug treatment or pathologies such as cancer. Herein, I will first present our methodology that is used for comparing transcriptomes of zebrafish and mouse cell lines upon their exposure to rapamycin, an mTOR inhibitor. Next, I will demonstrate Comparizome, a webserver we have developed to statistically analyze selected microarray expression datasets for the zebrafish paralogs along with their human counterpart. Different examples from development, drug exposure and cancer biology studies will be presented using Comparizome demonstrating the feasibility and effectivity of comparative transcriptomics approach between zebrafish and humans for a better understanding human disease and treatment.

SERDAR DURDAĞI

Integration of Machine Learning, Text Mining, Binary QSAR Models and Target-Driven Virtual Screening Approaches for the Identification of Novel Small Molecule Therapeutics

The industry appears to be suffering from innovation costs and would benefit from new, more rational approaches to drug design which exploit the recent advances made in molecular and structural biology and computational chemistry. Computational chemistry provides a range of simulation tools for description of protein-ligand binding, statistical methods for analysis of the binding data that help to predict the optimal ligands, and molecular modeling tools that enable construction of novel ligands. In recent years, a growing number of defined crystal, as well as NMR and cryo EM structures and access to these structures from the protein database (PDB, protein data bank) have enabled these data to be used as template targets in molecular modeling studies. In this talk, examples of structure-based and ligand-based screening of small molecule databases for different targets will be highlighted. Filtered structures based on predicted binding energy results using high throughput virtual screening (HTVS) techniques are used in more sophisticated molecular simulations approaches (i.e., Induced Fit Docking- IFD, and Quantum Mechanics Polarized Ligand Docking- QPLD). Potent high binding affinity compounds that are predicted by molecular simulations are then tested by long molecular dynamics (MD) simulations. The molecular mechanism analysis, free energy calculations (i.e., MM/GBSA) using short/long multiple MD simulations for the identified compounds which show high predicted binding affinity against specific target structures, as well as structure-based pharmacophore development (E-pharmacophore) studies and drug re-positioning studies (Figure 1) will be summarized.1-6

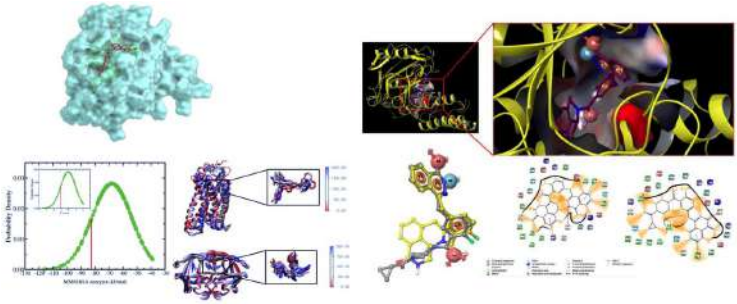


Figure 1. Integrated structure- and ligand-based virtual screening techniques were applied in different projects for the identification of novel therapeutic small molecules.

SERHAT TOZBURUN

Decision Support Systems Based on Artificial Intelligence in Colonoscopy

Machine intelligence provides a platform suitable for applications such as processing, analysis, anomaly detection, and classification from real-time images of medical imaging devices. In particular, it allows the creation of a rapid decision-support system. In highly operator-dependent procedures such as colonoscopy, for example, colorectal polyps can be easily overlooked during the procedure. This medical failure can be attributed to factors such as poor bowel preparation, mucus on the polyp, bowel folds, and blind spots, as well as the inexperience of the endoscopist. At this point, abnormal tissue structures can be automatically detected, localized, and classified using methods such as machine learning and deep learning under artificial intelligence. The most common method in the literature of this particular application is convolutional neural networks, and also autoencoder-decoder architecture for segmentation. Besides, optimization and cross-validation methods are used carefully in the development of the learning model and, therefore, the model can be generalized for other possible medical applications. In addition to contributions such as adaptive contrast enhancement and motion blur removal, the development of a new generation of optical imaging endoscope devices integrated with artificial intelligence can be promising to enable rapid classification of different types of polyps and significantly reduce the rate of biopsy.

TUNAHAN ÇAKIR

Network-Based Analysis of Transcriptome Data to Unravel Molecular Mechanisms Behind Cellular Impairments

Among other functional genomics data types, transcriptomics is still the most accessible one due to its higher genome coverage and lower cost. Availability of public transcriptome databases is another important factor that prompts the use of transcriptomic data to elucidate molecular details of cellular impairment mechanisms. Measured mRNA levels in transcriptome experiments are indicative of corresponding protein activities. Since proteins mostly function by interacting with each other, it is oversimplification to interpret the change in mRNA levels by ignoring interactions between corresponding proteins. There are two major approaches for incorporating interaction information to the analysis of transcriptomic data. The first one is mapping the significance of change of mRNA levels (e.g. p-values) on an organism-specific protein-protein interaction network to identify a highly altered subnetwork. This enables selection of significantly changed genes in response to cellular impairment whose products physically interact with each other. The second approach is to use correlation between mRNA levels of gene pairs to create a co-expression network, which can later be divided into modules to identify different functions affected from cellular impairment. The advantage of the first approach is that it uses experimentally known physical interactions between proteins. The second one, on the other hand, is not limited by the known physical interactions, which is still incomplete, and it does not ignore the possibility of nonphysical interactions between gene products. In this talk, cognitive impairments due to aging and cellular impairments due to neurodegeneration will be analyzed via network-based approaches using corresponding transcriptomic data.

ZERRIN IŞIK

Network Based Approaches for Compound Target Identification

All protein targets of a compound might not be defined in the development stage. Consequently, unexpected side effects of any compound might appear and lead unsuccessful wet-lab experiments due to unknown off-targets. If protein targets of compounds would be identified more exhaustively, the observed side effects after a disease treatment might be also limited. The focus of this talk is explaining of a network-based approach proposed for computational identification of potential targets of a given compound. Our approach computes potential off-targets of a compound by using gene expression data of compound-treated cells and tissue-specific protein-protein interaction (PPI) networks. The method first maps transcriptome level responses of a compound on the PPI network, and computes several network centrality metrics to prioritize the most probable targets of the given compound. Our experiments showed that the functional PPI network achieves a much better performance compared to the physical one. The random walk-based metric provided the best recall values out of six centrality metrics for different cancer tissues. The proposed approach would help to identify potential off-targets in experimental drug developments, eventually reduces development costs.

SELECTED ABSTRACTS FOR ORAL PRESENTATIONS

SPADIS: AN ALGORITHM FOR SELECTING PREDICTIVE AND DIVERSE SNPS IN GWAS

Serhan Yilmaz[1], Oznur Tastan[2], , A. Ercument Cicek[1,3]*

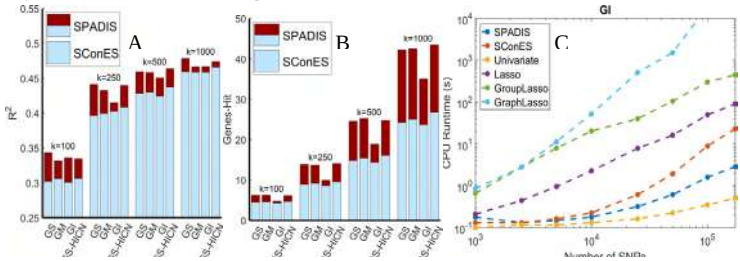
*1. Bilkent University, Computer Engineering Dept. Cankaya, Ankara 06800 2Sabanci University, Faculty of Natural Sciences and Engineering, Tuzla, Istanbul 34906 3Carnegie Mellon University, Computational Biology Dept., Pittsburgh PA 15213 *Correspondance*

Complex traits often cannot be explained by individual variants. Therefore, the efficient selection of multiple loci that explain the phenotype is critical for understanding the genetic basis of these traits. Selecting multiple loci is a computationally challenging problem that grows exponentially with the number of genomic variants. Many methods tackle this problem by focusing on coding regions to reduce the complexity of the problem. However, those approaches ignore the non-coding regions and introduce literature bias. As one alternative, regularized regression methods have been used; however, they do not allow the incorporation of background biological knowledge and suffer from long execution times. Currently, there is only one machine learning method in the literature, which aims to select a large set of loci efficiently by incorporating biological background information - SConES . SConES selects a set of features guided by a SNP-SNP network and favors the selection of SNPs that are connected on the network. We argue that while connectedness assumption is frequently used for functionally related features, it leads to the selection of redundant features when the goal is to explain a complex phenotype. In the current study, we hypothesize that selecting features on an SNP-SNP network that are diverse in term of location would correspond to incorporating complementary terms and thus, would help to explain the phenotype better. We present SPADIS that implements this novel idea by maximizing a submodular set function with a greedy algorithm that ensures a constant factor approximation to the optimal solution. We compare SPADIS to the state-of-the-art method SConES on a dataset of Arabidopsis Thaliana genotype and continuous flowering time phenotypes and show that (i) SPADIS has better average phenotype prediction performance in 15 out of 17 phenotypes when same number of SNPs are selected and provides consistent and statistically significant improvements in regression performance on average across multiple networks and settings (Figure 1A), (ii) it identifies more candidate genes (Figure 1B), and (iii) runs faster much faster (Figure 1C). We also perform rigorous simulation experiments and compare SPADIS with off the shelf regression-based feature selection methods and show that SPADIS outperforms its counterparts. SPADIS is on bioRxiv (https://www.biorxiv.org/content/early/2018/05/17/256677) and the accompanying software tool is readily available at http://ciceklab.cs.bilkent.edu.tr/spadis.

Keywords: SNP Selection, Submodular Optimization

Figure 1. Figure shows the improvement of SPADIS over SCONES in terms of Pearson's squared correlation coefficient (Panel A) and the number of candidate genes identified (Panel B) for different

Keywords: SNP Selection, Submodular Optimization



number of SNPs selected, k. All values shown are averaged over 17 phenotypes. Blue bar indicates the best performance of SCONES for the corresponding SNP-SNP network (GS, GM, GI, and GS-HICN) and k value. The red bar indicates the improvement of SPADIS over SCONES. Panel C shows CPU time measurements of SPADIS, SCONES, Univariate (baseline method), Lasso, GroupLasso and GraphLasso when the number of considered SNPs is varied from 1.000 to 173.219 SNPs.

References: [1] Azencott et. al. (2013) Efficient network-guided multi-locus association mapping with graph cuts. Bioinformatics 29 (13): i171-i179. Corresponding author's address: 1. Sabanci University, Tuzla, Istanbul 34906, otastan@sabanciuniv.edu, http://people.sabanciuniv.edu/otastan/; 2. Bilkent University, Cankaya, Ankara 06800 – cicek@cs.bilkent.edu.tr - http://ercumentcicek.com

ALLOSTERIC COMMUNICATIONS IN INACTIVE STATES OF HUMAN B_2-ADRENERGIC RECEPTOR (B_2-AR)

Nuray Söğünmez[1], E. Demet Akten[2]

1. Graduate Program of Bioinformatics and Genetics, Graduate School of Science and Engineering, Kadir Has University, Istanbul, TURKEY
2. Department of Bioinformatics and Genetics, Faculty of Engineering and Natural Sciences, Kadir Has University, Istanbul, TURKEY

β2-adrenergic receptor (β2-AR) is a member of G-protein coupled receptor (GCPR) superfamily, which is known to be one of the largest and most diverse superfamilies in the mammalian genome, playing a critical role in multiple essential physiological processes. They have become the most important drug targets in the pharmaceutical industry; yet, GPCR-targeted drug discovery still lacks a comprehensive understanding of the target-specific dynamic network and allosteric communication [1]. Allostery, defined as the communication between two distant parts, is a universal characteristic of most proteins [2]. For this study, a detailed investigation of β2-AR dynamics and putative allosteric sites will be conducted for two inactive states, which were previously obtained from a 1.5 μs long molecular dynamics (MD) run [3] using a time-delayed cross-correlation analysis of atomic fluctuations [4].

Prior to this study, a 1.5 μs long MD simulation was conducted on an inactive X-ray structure of human β2-AR (PDB ID: 2RH1) where a distinct inactive state of the receptor was revealed and designated as "Phase II"; along with the known inactive state designated as "Phase I". In this "highly" inactive state, G-protein binding cavity was completely inaccessible by the closure of third intracellular loop (ICL3) of the receptor. In this study, mutual information (MI), which represents the shared information between residue-pairs, was analyzed based on both residue-types and residue-location. Additionally, entropy transfer (TE) was calculated based on changes in the atomic fluctuations of Cα atoms with respect to their mean positions. Number of bins and delay times were optimized for each residue, followed by the correction of finite sampling effect.

According to MI analysis, polar residues contribute more to information sharing than hydrophobic ones. Overall, the average information shared by proximal residues was the highest between loop regions and the lowest between transmembrane and loop regions in both phases. For distal residues, although the average MI was smaller than proximal ones, the highest degree of MI was still shared between loop regions. The amount of MI and TE notably intensified in Phase II in the overall structure. TE analysis identified residues driving the fluctuations of other residues, thereby unraveling the causality. The orthosteric-binding site in β2-AR was distinctly revealed as an information acceptor in Phase I, and donor in Phase II. On the other hand, the donor characteristics of the allosteric inhibitor-binding site switched to an acceptor in Phase II. Interestingly, the intracellular portion of the receptor incorporating the

ionic lock, which majorly contributes to receptor inactivation, turned into a major acceptor site in Phase II. These findings clearly showed that the information transfer was directed from the extra- to the intracellular region in the receptor in this "highly" inactive state, which represented the extreme opposite of an active state where the intracellular region is known to drive the extracellular orthosteric ligand-binding initiated by G protein binding.

Keywords: GPCRs; Human β2-AR; Protein Allostery; Mutual Information, Entropy Transfer

References: [1]. Weis WI, Kobilka BK. The Molecular Basis of G Protein–Coupled Receptor Activation. Annu. Rev. of Biochem. 2018 Jun 20;87:897-919. Pubmed PMID: 29925258
[2]. Kamberaj H, van der Vaart A. Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations. Biophys. J. 2009 Sep 16;97(6):1747-55. Pubmed PMID: 19751680
[3]. Ozcan O, Uyar A, Doruker P, Akten ED. Effect of Intracellular Loop 3 on Intrinsic Dynamics of Human B2-Adrenergic Receptor. BMC Struct. Biol. 2013 Dec;13(1):29. Pubmed PMID: 24206668
[4]. Sogunmez N, Akten ED. Intrinsic Dynamics and Causality in Correlated Motions Unraveled in Two Distinct Inactive States of Human β2-Adrenergic Receptor. J. Phys. Chem. B. 2019 Apr 4;123(17):3630-42. Pubmed PMID: 30946584

Corresponding Author: demet.akten@khas.edu.tr
+90 212 533 65 32 (x1350)

THE ROLE OF COMPUTATIONAL MODELING IN UNDERSTANDING NUCLEOSOME DYNAMICS

Seyit Kale[1,2]

1. Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda MD 20894, USA,
2. Izmir Biomedicine and Genome Center, Balcova, Izmir, TURKEY

Protection of our genomic material and regulation of access to it are inherently conflicting tasks. At the heart of this conflict lies the nucleosome, the smallest repeating architectural unit of chromatin. This nucleo-protein is a disk-shaped, 200 kDa complex that is composed of approximately 150 base pairs of DNA wrapped around eight highly basic, well-conserved proteins, known as histones. Recent advances in structural and complementary experimental techniques have permitted a detailed look into how the histone octamer provides a protective core for the nucleic acid around it, yet how this protection can be tuned by chromatin interactors and epigenetic modifications remains to be elucidated at the molecular level. We use tools of computational modeling and biophysics to complement experimental efforts with a molecular picture of the dynamics of the nucleosome. Guided by experimental input, molecular simulations can elicit unique insights into how intricate epigenetic variations, such as substitutions between histone variants, can lead to DNA-sequence dependent behavior, or how the action of a chromatin remodeler can induce conformational changes in the core octamer [1], both with implications for the nucleosomal DNA.

Keywords: Nucleosome, Chromatin Remodeling, Molecular Dynamics

References:[1] Hada A, Hota SK, Luo J, Lin Y, Kale S, Shaytan AK, Bhardwaj SK, Persinger J, Ranish J, Panchanko AR, Bartholomew B. Histone octamer structure is altered early in ISW2 ATP-dependent nucleosome remodeling. Cell Rep. 2019 Jul 2, 282-294. Pubmed PMID: 31269447.

Corresponding Author: Izmir Biomedicine and Genome Center, seyitkale@gmail.com

EFFECTS OF VERNIER ZONE RESIDUES ON ANTIBODY MULTI-SPECIFICITY

Merve Arslan[1,2], *Dilara Karadag*[1], *Sibel Kalyoncu*[1]

1. Izmir Biomedicine and Genome Center, Izmir, Turkey
2. Izmir Biomedicine and Genome Institute, Dokuz Eylul University, Izmir, Turkey

Monoclonal antibodies are one of the most important biological drugs being developed for targeted therapy. Antibody-antigen interaction specificity is one of main properties to consider for successful of antibody drug development. Although antibody specificity has not been completely understood yet, certain structural and physicochemical characteristics are known to play important roles [1]. In this study, we hypothesize that vernier zone residues are key for modulation of antibody specificity. Vernier zone regions underlie complementarity determining regions (CDRs) of antibodies which is responsible for antigen binding. Humanization approaches mainly engineer vernier zone residues to provide correct canonical structures of CDRs in order to restore/improve binding affinity. Unlike binding affinity, specificity cannot be quantified and it is usually hard to modulate antibody specificity by rational design. By using homology modelling and molecular dynamics simulations for antibody sequences from Bostrom et.al study, we found that certain interactions between vernier zone and CDR residues convert a mono-specific antibody into a bispecific antibody [2]. Surprisingly, bispecific antibodies also have comparable binding affinities with those of mono-specific ones. Because vernier zone regions are usually underestimated in the literature, we tried to represent them in a schematic of antibody secondary structure (Figure 1). Certain locations of vernier zone, especially the one facing antigen, were documented to be important for providing antibody multi-specificity. This study might lead to a novel antibody engineering strategy for antibody specificity modulation.

Keywords: Antibody; Specificity; Vernier Zone

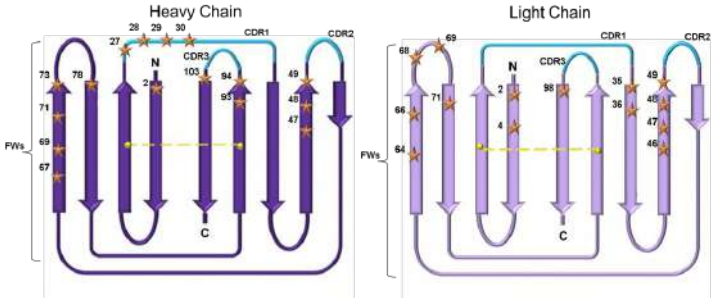


Figure 1. Schematic representation of vernier zone residues on variable regions of heavy and light chains. CDR regions are colored in cyan, framework (FW) regions are colored in dark blue (heavy chain) and light blue (light chain), vernier zone residues are labelled as stars in orange, disulfide bonds are represented in yellow.

References: [1] Peng, H. P., Lee, K. H., Jian, J. W., & Yang, A. S. Origins of specificity and affinity in antibody-protein interactions. Proc Natl Acad Sci. USA 2014 111(26), E2656-2665. PubMed PMID: 24938786.

[2] Bostrom J, Yu S-F, Kan D, et al. Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. Science. 2009;323(5921):1610. PubMed PMID: 19299620.

Corresponding Author's Address: Izmir Biomedicine and Genome Center, Dokuz Eylul University Health Campus, Mithatpasa Cad. No:58/5 Balcova Izmir, Turkey

DISCOVERY OF STRUCTURAL VARIATIONS IN ANCIENT GENOMES

Arda SÖYLEV[1], Can ALKAN[2], Mehmet SOMEL[3]

1. Department of Computer Engineering, Konya Food and Agriculture University, Konya, Turkey, arda.soylev@gidatarim.edu.tr
2. Department of Computer Engineering, Bilkent University, Ankara, Turkey, calkan@cs.bilkent.edu.tr
3. Department of Biological Sciences, Middle East Technical University, Ankara, Turkey, msomel@metu.edu.tr

Structural variations are genomic rearrangements that affect more than 50 bps in the genome. These variations are commonly linked with various genomic disorders and have profound significance in population and evolutionary analysis. With the emergence of high-throughput sequencing (HTS) technologies, studies on the discovery of these variants have become more common and several algorithms have been developed for this purpose. In parallel, sequencing the genomes of ancient populations has become possible with short-read sequencing technologies and thousands of ancient genomes have been sequenced both from Anatolia and the other parts of the World. However, since ancient DNA is generally not well preserved, there are many technical difficulties in analysing such genome data, including the shorter read lengths, higher error rates and lower coverage compared to modern genomes. Indeed, no known algorithms that detect structural variations in ancient genomes have been developed so far. In this paper we describe the ancientSV algorithm that discovers different types of structural variations including deletions, inversions, duplications and mobile element insertions in ancient genomes. Our approach uses various sequencing signatures such as split-read, read-pair and read-depth with HTS short-read technology. We evaluated the performance of our algorithm using real data and simulations, which revealed surprisingly high overlap with variants identified in modern genomes at low depth of coverage, indicating the accuracy of our approach.

Keywords: Structural Variations; High Throughput Sequencing; Ancient Dna

A HIERARCHICAL RANDOM FOREST APPROACH FOR CELL TYPE PROJECTIONS ACROSS SINGLE CELL RNASEQ DATASETS

Yasin Kaymaz[1], *Nathan Lawless*[2], *Timothy Sackton*[1]

1. Informatics Group, Harvard University, Cambridge, MA, USA.
2. Computational Biology, Boehringer Ingelheim Pharma GmbH & Co KG, Biberach an der Riss, DE.

The emergence of single-cell RNA sequencing (scRNAseq) has led to an explosion in novel methods to study biological variation among individual cells, and to classify cells into functional and biologically meaningful categories. These approaches have revealed novel cell types and previously unknown relationships with phenotype and disease, motivating an exponential increase in the scope of scRNAseq studies, including large scale cell atlas projects. Being able to utilize the rich information from such projects provides new dimensions to smaller-scale individual studies. Therefore, integration and accurate information transfer between existing cell atlases and newly generated, targeted data is a critical step for proper interpretation of biological features. While methods exist for cell type assignment across experiments, current approaches have limitations. Existing algorithms work well when the reference training data is composed of a few well-represented cell types, and when the query data contains few or no novel types and a good representation of known cell types. However, an ideal classification should be able to handle many candidate cell classes and not rely on a minimum input threshold of query data as some single-cell protocols produce low-throughput data in which rare cell types are represented with only a few cells.

Here, we present a new cell type projection tool based on hierarchical random forests that overcomes these limitations by using a priori information about cell type relationships for improved classification accuracy. We named our tool as '**HieRFIT**', which stands for "Hierarchical Random Forest for Information Transfer". This novel classification algorithm takes as input a hierarchical tree structure representing the class relationships, along with the reference data. We use an ensemble approach combining multiple random forest models, organized in a hierarchical decision tree structure. We show that our hierarchical classification approach improves accuracy and reduces incorrect predictions. We use a scoring scheme that adjusts probability distributions for candidate class labels and resolves uncertainties while avoiding the assignment of cells to incorrect types by labeling cells at internal nodes of the hierarchy when necessary. Using HieRFIT, we re-analyzed publicly available scRNAseq datasets showing its effectiveness in cell type cross-projections with inter/intra-species examples. HieRFIT is implemented as an R package, freely available through GitHub (https://github.com/yasinkaymaz/HieRFIT).

Key Phrases: single-cell RNAseq, Cell types, Machine learning, Hierarchical classification, HieRFIT

Corresponding Author: Timothy Sackton Informatics Group, Harvard University, Cambridge, MA, USA. Phone: +1 617-495-9492 Email: tsackton@g.harvard.edu

EXPLORING ALLOSTERIC COMMUNICATION ON THE RIBOSOMAL TUNNEL IN HUMAN AND BACTERIA

Pelin Güzel[1,2], Ozge Kurkcuoglu[1]

1. Department of Chemical Engineering, Istanbul Technical University, Istanbul, Turkey
2. Faculty of Engineering and Life Sciences, Istanbul Medeniyet University, Istanbul, Turkey

Background: Ribosome complexes synthesize proteins across all kingdoms of life. These large ribonucleoproteins (rnp) found in both prokaryotes and eukaryotes seem to be evolved from a common structural rnp core with only additional parts at the solvent surface of the complex. On the complex, a conserved rRNA cavity named peptidyl transferase center (ptc) catalyzes peptide bond synthesis to form polypeptides that are emerged from the ribosomal exit tunnel to solvent side. There is accumulating evidence for allostery between the tunnel and ptc in order to control the crucial task. Exploring potential allosteric communication pathways in ribosomal complexes of different organisms is highly important to reveal allosteric regions that can be tested for species-specific drugs with high selectivity to pathogens.

Method: Graph-based k-shortest paths method [1] and suboptimal path calculation using coarse-grained molecular dynamics (CGMD) trajectories are used to predict allosteric paths between ptc and ribosomal exit tunnel in both T. thermophilus and human ribosomal complexes [2,3]. In k-shortest path method; a network is generated using local interaction strengths between nucleotides [1]. In suboptimal path calculation from CGMD; cross-correlations are calculated based on 500 ns long trajectories and used to generate the network. Both networks are then subjected to shortest path algorithms with k=20 to find optimal and suboptimal paths between ptc and tunnel.

Results: In T. thermophilus, both methods point to the same optimal path involving A2503, G2061 and U2504, which link A2062 (tunnel) and A2451(ptc), agreeing with a previous study [4]. In addition, suboptimal paths accomodate the universally conserved non-Watson-Crick base pair A2450-C2063 [5]. Similar optimal and suboptimal pathways are observed for the human ribosome, including the conserved non-Watson-Crick base pair at A4396-C3909.

Conclusion: Results from our graph-based approach [1] and MD simulations point to same network of nucleotides, highlighting the success of this computationally efficient method. Allosteric pathways predicted for T. thermophilus agree with previous data supporting the plausibility of the findings, while there is no experimental study for the tunnel of human complex yet. At this point, calculations suggest that both organisms use similar allosteric communication mechanisms between tunnel and ptc, where relatively low conserved nucleoides are involved. Here, all findings provide useful insights for new strategies to design drugs with high selectivity for bacterial ribosome.

Keywords: Ribosomal Complexes, Allostery, Shortest Paths

References: [1] Guzel, P; Kurkcuoglu O., BBA-General Subjects, 2017, 1861(12), 3131-3141.
[2] Górecki, A., et.al. Journal of Computational Chemistry 2009, 30-2364.
[3] Grant, B. J., et. al., Bioinformatics, 2006, 22(21), 2695-2696.
[4] Vazquez-Laslop, N., Thum, C., Mankin, A.S., Molecular Cell, 2008, 30, 190–202.
[5] Bayfield, M. A., Thompson, J., Dahlberg, A. E., Nucleic Acids Research, 2006, 32, 5512–5518.
Corresponding author's address: olevitas@itu.edu.tr

ASSESSING THE IMPACT OF PROTEOME REDUNDANCY MINIMIZATION IN UNIPROTKB

Ilayda B. Ucar[2], Ramona Britto[1], Borisas Bursteinas[1], Baris E. Suzek[2,3] and Maria Jesus Martin[1]

1. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK
2. Department of Computer Engineering, Muğla Sıtkı Koçman University, Muğla, Turkey
3. Bioinformatics Graduate Program, Muğla Sıtkı Koçman University, Muğla, Turkey

The UniProt Knowledgebase (UniProtKB) is the central hub for the collection of functional information on proteins. It consists of SwissProt, the manually annotated section and TrEMBL, which offers translations of nucleotide sequences supplemented with automatic annotation. Historically, UniProtKB was importing all coding sequences from International Nucleotide Sequence Database Collaboration (INSDC) which operates between EMBL, GenBank and DDBJ. As a result of the vast increase in genome sequencing projects, UniProtKB doubled in size reaching nearly 90 million entries in 2014. Furthermore, the genome sequencing projects', especially when sequencing of similar organisms involved, were increasing the redundancy in UniProtKB as their unique sequence contributions were limited. The size and redundancy started to hinder the user's ability to effectively navigate and find meaningful biological results in UniProtKB. To deal with this issue, UniProtKB team developed a procedure to identify highly redundant proteomes within species groups using a combination of manual and automatic methods. This procedure was first implemented for bacterial species in mid-2015 and the sequences corresponding to redundant proteomes (approximately 47 million entries) were deprecated [1]. These redundant proteomes were made available for download only from the UniProt Archive (UniParc). The procedure is still applied as part of UniProt releases and there is potentially some level of information lost through this effort. Hence, there is a need to develop methods to assess novel information lost as the sequences corresponding to redundant proteomes removed.

In this work, to assess potential information loss, we first created a set of three representative bacterial species each with a large number of redundant proteomes already in UniParc, namely Rhizobium meliloti, Bacillus toyonensis and Mycobacteroides abscessus. The information loss is assessed based on two aspects; the loss of novel sequences and the loss of protein domain annotations. To assess the loss of novel sequences, we first compared the sequences in our set against the non-redundant set of sequences for the same species available in UniProtKB using BLAST. The sequences showing <50% identity to any other sequence in this non-redundant set are further compared to the remaining UniProtKB sequences using BLAST and the sequences showing <50% identity to any other sequence in whole UniProtKB are categorized as "novel" sequences. Similarly, to assess the loss of novel protein domain annotations, the InterPro domains available for the proteins in our set is compared against the domains in the non-redundant set of sequences for the same species available in UniProtKB.

The InterPro domains present in only our set is categorized as "novel" domain annotations. These "novel" sequences or "novel" domain annotations (e.g. corresponding molecular functions, structurally important regions) wouldn't be represented in UniProtKB (or lost) if the corresponding proteome is categorized as redundant based on the current procedure.

Based on our initial results, the number of "novel" sequences in R.meliloti, B. toyonensis, and M. abscessus account for 0.3%(3071 proteins), 0.08%(898 proteins) and 0.17%(2431 proteins) of their proteomes, respectively. Similarly, "novel" domain annotations in R. meliloti, B. toyonensis, and M. abscessus account for 12.2%(800 domains), 5.42%(358 domains) and 7.9%(412 domains) of all the unique InterPro domains represented in their proteomes, respectively. Based on Interpro2GO mappings, on average, ~2% of these "novel" protein domains represent "novel" GO molecular functions.

In conclusion, we identified the redundancy minimization procedure leads to some level of information loss. The significance of the loss in terms needs to be further evaluated by UniProtKB curation team, however, our initial findings suggest the decision on redundancy should also factor in protein domain annotations. We anticipate our work will help in improving the proteome redundancy minimization procedure that is becoming exceedingly important due to the increasing number of genome sequencing projects.

We gratefully acknowledge the EMBL-EBI internship program and Muğla Sıtkı Koçman University BAP Project 19/079/10/2/2 for their support to this work.

Keywords: Protein Databases, Protein Sequence Analysis, Protein Sequence Annotation.

References:

[1] Bursteinas B, Britto R, Bely B, Auchincloss A, Rivoire C, Redaschi N, O'Donovan C, Martin MJ. Minimizing proteome redundancy in the UniProt Knowledgebase. Database (Oxford). 2016 Dec 26;2016:baw139. doi: 10.1093/database/baw139. PubMed PMID: 28025334.

Corresponding Author's Address:

E-mail: ilaydaucar66@gmail.com

COMPARISON OF NUCLEOTIDE EXCISION REPAIR PROFILES BETWEEN GRAY MOUSE LEMUR AND HUMAN

Veysel Kaya[1], Ümit Akköse[1], Zeynep Karagöz[1], Laura Lindsey-Boltz[2], Aziz Sancar[2], Ogün Adebali[1]

1. Molecular Biology, Genetics and Bioengineering Program, Sabanci University, Istanbul, 34956, Turkey
2. Department of Biochemistry and Biophysics, School of Medicine, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Nucleotide excision repair is the primary mechanism to remove bulky DNA adducts in animals.[1] Correspondingly, genome-wide mapping of nucleotide excision repair with XR-seq [2], provides comprehensive profiling of this type of repair. By the assembly of the gray mouse lemur (Microcebus murinus) genome at chromosome-scale [3], we were able to achieve a direct comparison of DNA damage repair profiles between mouse lemur and human. UV-irradiated human and lemur fibroblasts were used to generate excised oligomer pool representing the maps of genome-wide excision repair. To establish overlapping regions between genomes, Nucmer nucleotide aligner from Mummer software package[4] has been used to align the mouse lemur genome (Mmur_3.0) to the human genome (hg19). Chromosomal positions of homologous regions in both genomes are listed. Further details such as the alignment score of each homologous region, strand information, and the possibility of each homologous region overlapping with an annotated gene provided. This additional information on the defined homologous regions was further used to better assess the association between repair profiles of lemur and human genomes. To assess and compare the window-based relative excision repair levels in human and mouse lemur genomes in an unbiased way, we have defined the homologous regions reciprocally. XR-Seq reads for all homologous regions is then normalized, RPKM values are calculated and compared. This work will provide an understanding of the evolutionary conservation of DNA damage repair preferences across mouse lemur and human, which might eventually contribute mouse lemur to become a model organism.

Keywords: Damage Recognition; Repair Mapping; Cross-Species, Genome Alignment, Mouse Lemur

References:[1] Hu J., Selby C.P., Adar S., Adebali O., Sancar A.* (2017) Molecular mechanism of DNA excision repair and excision repair maps of the human and E. coli genomes. The Journal of Biological Chemistry, 292(38), 15588-15597.
[2]Hu J., Li W., Adebali O., Yang Y., Oztas O., Selby C.P., Sancar A.* (2019) Genome-wide mapping of nucleotide excision repair with XR-seq. Nature Protocols, 14(1), 248-282.
[3] Larsen, P. A., Harris, R. A., Liu, Y., Murali, S. C., Campbell, C. R., Brown, A. D., ... Worley, K. C. (2017). Hybrid de novo genome assembly and centromere characterization of the gray mouse lemur (Microcebus murinus). BMC biology, 15(1), 110. doi:10.1186/s12915-

017-0439-6
[4] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.

Corresponding Author's Address:
Ogün Adebali oadebali@sabanciuniv.edu
http://adebalilab.org/Sabanci University, Istanbul, 34956, Turkey

GENOME-SCALE METABOLIC NETWORK RECONSTRUCTION OF KLEBSIELLA PNEUMONIAE HS11286

Betül Baz[1], *Tunahan Çakır*[1]

1. Gebze Technical University, Department of Bioengineering, Gebze, Kocaeli

Global rising in the levels of multidrug-resistant bacteria has been an emerging problem of public health. Carbapenem-resistant or Carbapenemase-producing Enterobacteriaceae (CR-CPE) strain Klebsiella pneumoniae HS11286 is an important gram-negative bacterial pathogen, and has three multidrug-resistance plasmids including carbapenemase production. K.pneumoniae strain HS11286 causes opportunistic and severe hospital-associated infections [1]. It is important to demonstrate metabolic interactions within bacteria to elucidate molecular mechanisms and consequently design novel treatments for infections. Computational systems biology approaches can give further insights by constructing data-driven networks. In silico genome-scale metabolic network reconstruction is a mathematical representation of biochemical reactions catalyzed by gene products [2]. At present, there are several bioinformatics tools for the automated reconstruction of organism-specific draft metabolic network models at the genome-scale. However, these draft models later need laborious manual efforts with extensive literature review for curation and validation [2]. In this study, MATLAB-based RAVEN Toolbox 2.0 was used for semi-automated draft metabolic model reconstruction based on protein similarity via Blastp between the target organism and genetically similar species of it [3]. Within the scope of reconstruction, proteome information of K.pneumoniae strain HS11286 and K.pneumoniae KPPR1, and its recently reconstructed high-quality genome-scale metabolic network model called iKp1289 [4] were provided as input. Hence, the draft metabolic network was generated by retrieving a possible set of reactions from the template model via sequence homology, with specified cut-offs. The generated draft model was improved by a manual curation process to have a complete functional model. The addition of biomass formation reaction, exchange reactions and some reactions that produce missing biomass components were needed to turn the draft model into a model that can simulate phenotypic behavior of the organism via constraint-based analysis. The reconstructed genome-scale metabolic model for K.pneumoniae strain HS11286 consists of 2332 reactions, 1757 metabolites, and 1262 genes. The model could simulate the growth and respiration of the bacteria. In a further study, RAVEN Toolbox was used to reconstruct a KEGG-based metabolic network of the organism automatically. The KEGG-based metabolic network was used to obtain reactions that are specific to our target organism but do not exist in the reference strain used as a template. After that, the inconsistency within the model was checked to have a more accurate and reliable metabolic network. As a validation step of our model, the simulation of growth rate on different carbon sources was performed. The new metabolic network model will clarify the functional capacity of the

bacteria and the mechanisms of infections. This study was financially supported through a grant by TUBITAK (Project Code: 316S005).

Keywords: Metabolic Network; Genome-Scale Reconstruction; Klebsiella Pneumoniae

References: [1] Liu, Lu, et al. "Identification and characterization of an antibacterial type VI secretion system in the carbapenem-resistant strain Klebsiella
pneumoniae HS11286." Frontiers in cellular and infection microbiology 7 (2017): 442.
[2] Thiele, Ines, and Bernhard Ø. Palsson. "A protocol for generating a high-quality genome-scale metabolic reconstruction." Nature protocols 5.1 (2010): 93.
[3] Wang, Hao, et al. "RAVEN 2.0: A versatile toolbox for metabolic network reconstruction and a case study on Streptomyces coelicolor." PLoS computational biology 14.10 (2018): e1006541.
[4] Henry, Christopher S., et al. "Generation and validation of the iKp1289 metabolic model for Klebsiella pneumoniae KPPR1." The Journal of infectious diseases 215.suppl_1 (2017): S37-S43.

E-mail: bbaz@gtu.edu.tr

Intrinsic Allosteric Dynamics in G-protein Coupled Receptors

Özge Duman[1], Burcu Özden Yücel[1], Burcin Acar[1], Burcu Aykac Fas[1], Türkan Haliloğlu

1. Polymer Research Center and Chemical Engineering Department, Bogazici University

G-Protein Coupled Receptors (GPCRs) are the largest family of signaling proteins. GPCRs are divided into six classes according to the classical A-F classification system. Class A, B and C form the main classes and Class A (Rhodopsin-like receptors) is the largest class [1]. More than 800 GPCRs exist in humans [2]. The understanding of how they act for their function is essential in drug design. Although many studies have been performed to date, conformational dynamics in-between inactive/active states remain elusive. We plan to study Rhodopsin-like family (Class A) that includes five subtypes from M1 to M5 responsible from regulating many vital functions of the central and peripheral nervous systems and regulating a variety of physiological functions [3]. Using ANM-LD [4], the Anisotropic Network Model (ANM) combined with all-atom Langevin Dynamics (LD) simulations, we explore the spectrum of dynamic modes accessible and hidden in conformational transition pathways and allosteric interactions for each member of this family. As an initial case study, dissecting the motion underlying the activation of Muscarinic acetylcholine M2 receptor, the inactive and active crystal structures of which are shown in Figure 1, is seen to disclose certain modes of motion that profile the allosteric connection from extracellular to intracellular side. Known functional molecular switches such as DRY and NPxxY motifs align with the hinges of the activation motion, which suggest these sites as means to plausibly modulate the activation pathway and thus G-protein binding and allosteric signaling. The comparison of the dynamic spectrums across the GPCRs subtypes will provide a means towards the conformational dynamic landscape of GPCRs family and their evolutionary link.

Keywords: Protein Dynamics; Allostery; Internal Dynamics; GPCRs; Biased Signaling; Functional Selectivity; ANM-LD; Muscarinic acetylcholine M2 Receptor

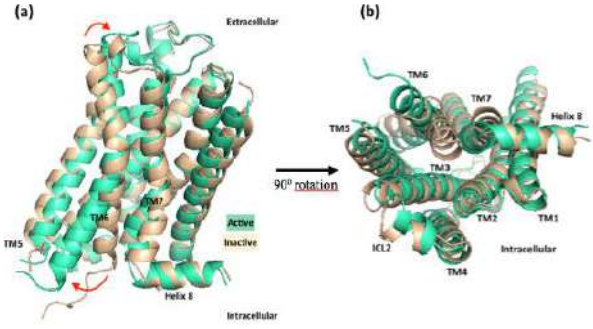


Figure 1. Inactive and active crystal structures of M2 receptor. (a) Red arrows show displacement of TM6 during the activation. (b) Intracellular side view of the structures.

References: [1] Latorraca NR, AJ Venkatakrishnan, RO Dror. GPCR Dynamics: Structures in Motion. Chemical
Reviews. 2017 Jan 11;117(1):139-155. PubMed PMID: 27622975.
[2] Pándy-Szekeres G, C Munk, TM Tsonkov, S Mordalski, K Harpsøe, AS Hauser, AJ Bojarski, DE
Gloriam. GPCRdb in 2018: adding GPCR structure models and ligands. Nucleic Acids Res. 2018 Jan 4;
46(D1): D440–D446. PubMed PMID: 29155946.
[3] Thal DM, B Sun, D Feng, V Nawaratne, K Leach, CC Felder, MG Bures, DA Evans, WI Weis, P
Bachhawat, TS Kobilka, PM Sexton, BK Kobilka, A Christopoulos. Crystal structures of the M1 and M4
muscarinic acetylcholine receptors. Nature. 2016 Mar 17; 531(7594):335-40. PubMed PMID: 26958838.
[4] Yang M, N Livnat Levanon, B Acar, B Aykac Fas, G Masrati, J Rose, T Haliloglu, N Ben-Tal, Y Zhao, O
Lewinson. Single-molecule probing of the conformational homogeneity of the ABC transporter BtuCD. Nat
Chem Biol. 2018 Jul;14(7):715-722. PubMed PMID: 29915236.

Corresponding Author's Address: Polymer Research Center, Bogazici University, Bebek 34342, Istanbul, Turkey
halilogt@boun.edu.tr

CROssBAR: Comprehensive Resource of Biomedical Relations with Network Representations and Deep Learning

Tunca Doğan[1,2,3], Ahmet Sureyya Rifaioglu[1,4], Heval Atas[1], Esra Sinoplu[1], Vishal Joshi[3], Andrew Nightingale[3], Rabie Saidi[3], Vladimir Volynkin[3], Hermann Zellner[3], Rengul Cetin-Atalay[1], Maria Jesus Martin[3], Volkan Atalay[1,4]

1. Cancer Systems Biology Laboratory (Kansil), Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey
2. Institute of Informatics / Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey
3. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), CB10 1SD Hinxton, Cambridge, UK
4.Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey

There are several computational tools and services in the literature that assist the experimental biomedical research. However, they also have shortcomings especially in terms of data connectivity, which limits their application to real-world problems. Here we aim to develop a comprehensive resource, CROssBAR, to address these shortcomings by connecting various biomedical resources, focusing on the fields of drug discovery and precision medicine. The proposed computational system contains 5 modules (Figure 1):

1. Construction of the CROssBAR database by integrating biological data from various resources;

2. Large-scale prediction of unknown drug-target interactions: a novel computational method for the comprehensive prediction of unknown drug/compound - target protein interactions to reveal novel on and off-target effects. For this, we developed a method called DEEPScreen using deep convolutional neural networks which predicts drug-target interactions based on 2D structural compound representations (https://github.com/cansyl/DEEPScreen);

3. Generation of the biomedical networks of integrated information where different types of nodes represent compounds/drugs, genes/proteins, pathways and diseases, and the edges represent the known and predicted pairwise relations in-between (https://github.com/cansyl/CROssBAR-Networks);

4. Biological evaluation (experimental validation) of selected computational results from the deep learning based predictor and from the biological networks of integrated information, on PI3K/AKT/mTOR pathway, in terms of liver cancer mechanisms;

5. Construction of the open access CROssBAR web-service, where it will be possible to browse with an entity ofinterest (i.e., a gene/protein, disease, drug or a pathway) to visualize its related biological network generated on-the-fly with its components.CROssBAR data pipelines, which does the heavy lifting of data from varied data sources like UniProt, InterPro, ChEMBL, PubChem, DrugBank, Reactome, KEGG, OMIM, EFO, OpenTargets and HPO has been developed.

The data is downloaded and processed by applying different rules in implementation logic to filter out noise. The database is hosted in self-sufficient collections in Mongo DB. CROssBAR database, which provide training data for the deep learning based drug discovery method DEEPScreen, and source data for the construction of biological networks, is maintained by EMBL-EBI and will be publicly available to researchers through a REST API service. It is expected that, the CROssBAR system will bridge the biological data resources which provide highly related biomedical information, but fairly disconnected from each other currently. The new system displays a continuous data flow from drugs/compounds to diseases with network representations and will be utilized to aid experimental and computational work in biomedical research.

Keywords: Biological Data Integration, Computational Drug Discovery And Repurposing, Drug-Target Interaction Prediction, Deep Learning, Biomedical Networks, Network Analysis, Biological Databases/Services, Biomedical Entity Relationships.

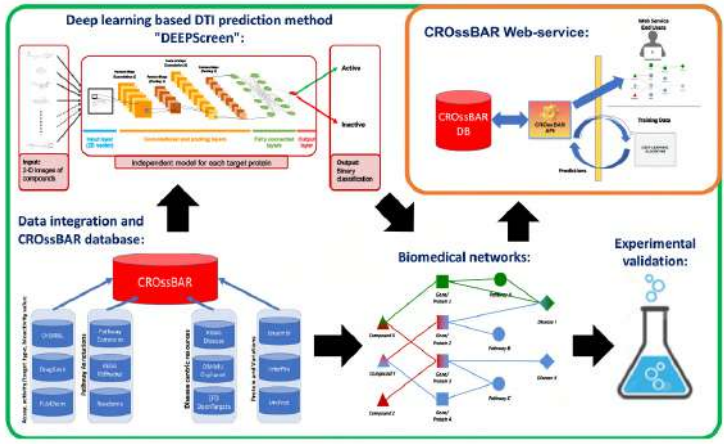


Figure 1. Overall schematic representation of the CROssBAR project.

Corresponding Authors' Addresses:
tuncadogan@hacettepe.edu.tr, rengul@metu.edu.tr,
vatalay@metu.edu.tr,

RECEPTOR-LIGAND BINDING AFFINITY PREDICTION VIA MULTI-CHANNEL DEEP CHEMOGENOMIC MODELING

Ahmet Sureyya Rifaioglu[1], Tunca Doğan[2,3], Maria Martin[4], Rengül Çetin-Atalay[2], Volkan Atalay[1]

1. Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey
2. Cancer Systems Biology Laboratory (Kansil), Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey
3. Institute of Informatics, Hacettepe University, 06800 Ankara, Turkey
4. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), CB10 1SD Hinxton, Cambridge, UK

In the field of drug discovery and development, conducting high-throughput screening experiments is not feasible for the massive compound and protein space. Therefore, several computational methods have been developed to aid experimental drug discovery processes by providing binding affinity predictions. The bioactive small molecules (i.e., ligands of proteins) with the desired predicted binding affinities (i.e. strong binders) can be used as drug candidates for further experimental validation against their targets (i.e., receptors). Here, we propose a multichannel receptor-ligand binding affinity prediction method. Our system employs a chemogenomic modeling approach where the aim is to use both receptor (target) and ligand features as inputs. One advantage of incorporating both ligand and receptor features is that the system can predict binding affinities of any given receptor-ligand pair, even if the corresponding ligand and/or receptor does not have any data point in the training set.

We used three datasets to train our system and to compare our results with the state-of-the-art methods: (1) Davis: Davis et al. performed screening experiments to determine the binding affinity values of 72 kinase inhibitors against 442 kinase proteins in human catalytic protein kinome. (2) Filtered Davis: In the original Davis dataset, Kd value of 10μM assigned to ligand-target pairs, if no bioactivity were observed in primary screen. These data points constitute more than two third of whole data points. They do not represent the actual binding affinities of corresponding ligand-target pairs therefore it is not suitable to use these data points in a regression problem. We filtered-out these data points having ambiguous bioactivity values and created a more reliable dataset. (3) PDBBind: is a comprehensive resource of experimentally measured binding affinities data for protein-ligand complexes, derived from PDB. It contains binding affinity values and 3D structures of protein-ligand complexes.

We describe a new protein encoding method where each protein is represented as a matrix which constitute the base channel of the convolutional part of the system. We first defined a surjective mapping function such that each unordered amino acid pair was mapped to a unique integer. Subsequently, we created an encoding matrix for each protein sequence where rows and columns represent amino acids in

the protein sequence, the elements of the matrices were filled based on the matching amino acid pairs using the encoding function. We also created additional input channels based on pre-defined amino acid matrices which represent various properties of amino acids and proteins. The four protein channels are: (1) the encoding matrix (defined above); (2) evolutionary properties/conservation: BLOSUM62 matrix; (3) physicochemical properties: the Grantham matrix (polarity, composition and molecular volume); (4) information reflecting the 3-D structure/fold: SIMK990101 matrix (i.e., distance-dependent statistical potential form the AAindex database). The diagonal elements in the encoding matrix represent the sequence itself. The remaining elements represent the amino acid matches in different positions of the protein sequence. Protein channels are fed to convolutional neural networks for processing. In the ligand side, we generated ECFP4 fingerprints using the SMILES strings of ligands, which are fed to a feed-forward neural network. Overall, we created a hybrid pairwise input neural network architecture which starts with separate ligand and protein branches, fully connected layers which processes the concatenated protein+ligand vector, and a regressor to predict the actual binding affinity value at the output layer (Figure 1).

We compared our system with three different methods: (1) DeepDTA: a binding affinity prediction method based on convolutional neuralönetworks and 1-D protein and compound encoding [1]. (2) SimBoost: another binding affinity prediction method based on gradient boosting machines and similarity networks [2]. (3) MoleculeNet: a benchmarking platform designed for evaluating and testing computational methods for molecular property predictions, which include prediction models that employ popular graph convolutional networks [3]. We trained our model and the other methods using the same training/validation/test settings. The performance results are given in Table 1. Results show that our method performs significantly better than the other methods in majority of the cases. Pursuing this approach, new models can be constructed by incorporating additional types of input protein channels.

Keywords: Binding Affinity Prediction, Chemogenomics, Receptor, Ligand, Deep Learning, Convolutional Neural Networks, Pairwise Input Neural Network, Protein Encoding.

References: [1] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: Deep drug-target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.
[2] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, "SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines," J. Cheminform., vol. 9, no. 1, pp. 1–14, 2017.
[3] Z. Wu et al., "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.

Corresponding Authors' Addresses: vatalay@metu.edu.tr & arifaioglu@ceng.metu.edu.tr

Method	CI	MSE	Pear-son	Spear-man	AUC	Prec-ision	F1-Score	MCC
PINN (Davis)	0.875	**0.28**	**0.813**	0.674	**0.945**	**0.838**	**0.732**	**0.674**
DeepDTA (Davis)	0.863	0.315	0.795	0.661	0.937	0.749	0.703	0.645
SimBoost (Davis)	**0.876**	0.284	0.804	**0.677**	0.939	0.781	0.699	0.645
PINN (filt.Davis)	**0.722**	**0.597**	**0.964**	**0.681**	**0.712**	**0.842**	**0.98**	**0.709**
DeepDTA (filt.Davis)	0.655	0.873	0.934	0.463	0.649	0.754	0.862	0.64
PINN (PDBBind)	**0.74**	**2.65**	**0.668**	**0.661**	**0.83**	**0.757**	0.781	0.446
Grid Featurizer - RF (PDBBind)	0.729	3.4	0.632	0.634	0.807	0.762	**0.822**	**0.529**
Grid Featurizer - DNN (PDBBind)	0.67	3.616	0.532	0.505	0.735	0.692	0.79	0.406
ECFP4 - RF (PDBBind)	0.657	3.207	0.478	0.483	0.736	0.675	0.76	0.334
ECFP - RF (PDBBind)	0.608	5.255	0.329	0.344	0.664	0.648	0.736	0.25

Table 1: vatalay@metu.edu.tr & arifaioglu@ceng.metu.edu.tr

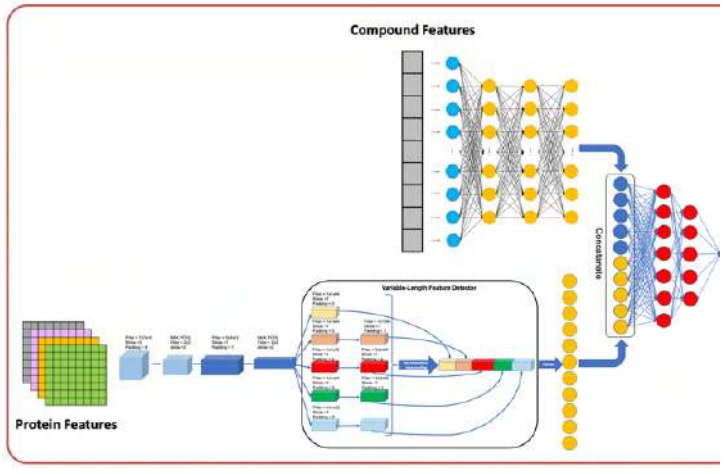


Figure 1. Schematic representation of the architecture.

IDENTIFICATION OF A NOVEL MISSENSE VARIANT IN THE PRKAR1A GENE AND ITS PATHOGENICITY MECHANISM

Tugce Bozkurt[1], *Umut Gerlevik*[1], *Ugur Sezerman*[1]

1. Department of Biostatics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey

Acrodysostosis refers to a rare type of skeletal dysplasias accompanied by common phenotypic features such as facial dysostosis, nasal hypoplasia, brachydactyly, short stature, and mental retardation [1,2]. Heterozygous mutations in PRKAR1A gene (OMIM *188830) have been found as the genetic cause of the subset of acrodysostosis which is characterized with resistance to several hormones, acrodysostosis 1 (OMIM #101800) [3,4]. PRKAR1A gene encodes cAMP-dependent regulatory subunits of protein kinase A (PKA). PKA holoenzyme has critical roles in several signaling pathways by phosphorylating the downstream substrates by catalytic subunits after leaving regulatory subunits upon cAMP binding to these regulatory subunits [5]. Here, we report a 6-years-old girl who has the clinical symptoms: distinctive facial features, skeletal system abnormalities, global developmental delay, hypothyroidism. Whole exome sequencing (WES) was performed on genomic DNA of the patient. After WES analysis, homozygous variants with minor allele frequency <0.1% in the databases, i.e. ExAC [6] and 1KGP [7] were kept. Heterozygous variants which are found in at least one of these databases were excluded. The intronic variants far away from ±10 bases at exon-intron boundaries were eliminated. Then, we prioritized variants based on the symptoms after collecting evidence from OMIM, MGI and literature search. Next, several mutation impact predictors such as SIFT [8], PolyPhen2 [9], MutationTaster [10], and REVEL [11] were considered for pathogenicity estimations. With the combination of whole computational assessments, de novo heterozygous mutation, c.G512A, in PRKAR1A gene was determined as the most prominent variant. To investigate the impact of G171E, anisotropic network modeling (ANM) and molecular dynamics (MD) simulations were performed. Full coverage homology models were built, and one from Robetta [12] was chosen after quality check. Then, cAMP molecules were docked to the known positions in UniProtKB [13] with the binding poses of cAMPs in PRKAR1A structures in RCSB PDB [14]. Wild-type and mutant systems were simulated for 40 ns at 310 K under NPT ensemble with 3 repeats. Loss of cAMP molecules from the binding pockets were observed upon the G171E mutation whereas they were stable in the wild-type. The equilibrated structures were used for ANM analyses. An increase in the mobility of binding pocket surrounding residues were monitored in the mutant structure as supporting MD results. In conclusion, the patient who has a novel missense variant in the PRKAR1A gene was diagnosed with acrodysostosis 1 via our prioritization strategy. We enlightened the pathogenicity mechanism of G171E mutation by ANM and MD studies as disrupted binding of cAMP to the regulatory subunit that prevents the catalytic subunit of PKA from phosphorylating many downstream substrates in numerous

signaling pathways.

Keywords: PRKAR1A; Whole Exome Sequencing; Molecular Dynamics

References: [1] Maroteaux P, Malamut G. [Acrodysostosis]. Presse Med. 1968 Nov 27;76(46):2189–2192. PMID: 5305130
[2] Robinow M, Pfeiffer RA, Gorlin RJ, McKusick VA, Renuart AW, Johnson GF, et al. Acrodysostosis. A syndrome of peripheral dysostosis, nasal hypoplasia, and mental retardation. Am J Dis Child. 1971 Mar;121(3):195–203. PMID: 5551869
[3] Linglart A, Menguy C, Couvineau A, Auzan C, Gunes Y, Cancel M, et al. Recurrent PRKAR1A mutation in acrodysostosis with hormone resistance. N Engl J Med. 2011 Jun 9;364(23):2218–2226. PMID: 21651393
[4] Linglart A, Fryssira H, Hiort O, Holterhus P-M, Perez de Nanclares G, Argente J, et al. PRKAR1A and PDE4D mutations cause acrodysostosis but two distinct syndromes with or without GPCR-signaling hormone resistance. J Clin Endocrinol Metab. 2012 Dec;97(12):E2328-2338. PMID: 23043190
[5] Bossis I, Stratakis CA. Minireview: PRKAR1A: normal and abnormal functions. Endocrinology. 2004 Dec;145(12):5452–5458. PMID: 15331577
[6] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016 18;536(7616):285–291. PMCID: PMC5018207
[7] 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74. PMCID: PMC4750478
[8] Kumar P, Henikoff S, Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc. 2009;4(7):1073–1081. PMID: 19561590
[9] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010 Apr;7(4):248–249. PMCID: PMC2855889 [10]
Schwarz JM, Rödelsperger C, Schuelke M, Seelow D. MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods. 2010 Aug;7(8):575–576. PMID: 20676075
[11] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am J Hum Genet. 2016 Oct 6;99(4):877–885. PMCID: PMC5065685
[12] Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W526–W531. PMCID: PMC441606
[13] UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019 Jan 8;47(D1):D506–D515. PMCID: PMC6323992
[14] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–242. PMCID: PMC102472

MEXCOWalk: Mutual Exclusion and Coverage Based Random Walk to Identify Cancer Driver Modules

Rafsan Ahmed[1], *Ilyes Baali*[1], *Cesim Erten*[2], *Evis Hoxha*[2], *Hilal Kazan*[2]

1. Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya, Turkey
2. Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey

Motivation Genomic analyses from large cancer cohorts have revealed the mutational heterogeneity problem which hinders the identification of driver genes based only on mutation profiles. One way to tackle this problem is to incorporate the fact that genes act together in functional modules. The connectivity knowledge present in existing protein-protein interaction networks together with mutation frequencies of genes and the mutual exclusivity of cancer mutations can be utilized to increase the accuracy of identifying cancer driver modules.

Results: We present a novel edge-weighted random walk-based approach that incorporates connectivity information in the form of protein-protein interactions, mutual exclusivity, and coverage to identify cancer driver modules. MEXCOwalk outperforms several state-of-the-art computational methods on TCGA pancancer data in terms of recovering known cancer genes, providing modules that are capable of classifying normal and tumor samples, and that are enriched for mutations in specific cancer types. Furthermore, the risk scores determined with output modules can stratify patients into low-risk and high-risk groups in multiple cancer types. MEXCOwalk identifies modules containing both well-known cancer genes and putative cancer genes that are rarely mutated in the pan-cancer data. The data, the source code, and useful scripts are available at: https://github.com/abu-compbio/MEXCOwalk [1].

Keywords: Cancer Driver Modules, Mutual Exclusivity, Random Walk

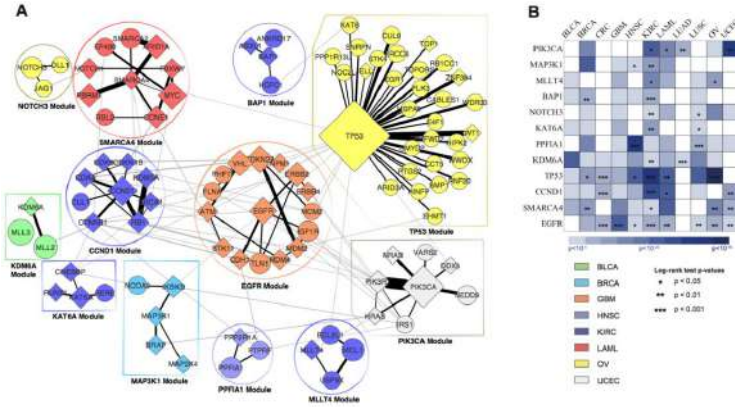


Figure 1. A) MEXCOwalk output modules when total_genes = 100. Diamond shaped nodes correspond to CGC genes. Sizes of the nodes

are proportional with mutation frequencies of corresponding genes. Edges within the module are colored black, whereas the edges between the modules are colored. Edge weights are reflected in the thicknesses of the line segments. Color of a module denotes the cancer type with the strongest enrichment for mutations in genes of that module. The legend for the color codes are shown on the right. Each module is named with the largest degree gene in the module. B) Results of cancer type specificity and survival analyses. Rows correspond to modules and columns correspond to cancer types. Colors of the matrix entries indicate the significance of enrichment for cancer types in terms of Fisher's exact test p-values. Stars indicate the significance of log-rank test p-values in survival analyses.

References: [1] Biorxiv link: https://www.biorxiv.org/content/10.1101/547653v2 (currently in press in Bioinformatics)

Corresponding Author's Address
cesim.erten@antalya.edu.tr, hilal.kazan@antalya.edu.tr

MOLECULAR DYNAMICS SIMULATIONS OF THE DYNEIN LINKER MOVEMENT

Mert Gölcük[1], Elhan Taka[1], Sema Zeynep Yılmaz[1], Mert Gür

1. Department of Mechanical Engineering, Istanbul Technical University (ITU), Istanbul, Turkey

Dyneins are AAA+ (AAA: ATPase associated with various cellular activities) motors that move progressively straight to the minus end of the microtubule (MT) (1). The dysfunctionality of dyneins has been associated with numerous neurodegenerative diseases and disorders. Also, dyneins play important roles in cell division and motility of the cell (1, 2). Understanding how dynein generates motility and force on MTs is essential to develop novel chemical inhibitors/modifiers of dynein function for the treatment. Cytoplasmic dynein is composed of two identical dynein heavy chains (DHCs) and several smaller associated polypeptides (Perrone et al. 2003a). DHC contains a C-terminal motor domain (head) and an N-terminal tail domain. The head contains six AAA modules (AAA1-AAA6) organized into a hexameric ring. ATP hydrolysis occurs in AAA1- AAA4 domains, where AAA1 is the main ATP site and necessary for motility (3). In addition to the AAA+ domain, dynein also contains the linker, the stalk, the buttress, and the C-terminal domain. ATP binding and hydrolysis results in a conformational change of the linker from a straight conformation to a bend conformation. This conformational change is referred to as the priming stroke. Upon microtubule binding, the power stroke of dynein takes place during which the linker returns to a straight conformation. To the best of our knowledge, all-atom Molecular Dynamics (MD) simulations of the linker movement during the priming/power stroke have not been performed in the literature. Moreover, the energetics and structural mechanism remains to be elucidated. For the full length dynein motor domain only a very limited number of all-atom MD simulations studies exists in the literature. We have recently (4), performed MD simulations of human dynein-2 in its pre- power stroke state in the presence of explicit water and ions. Furthermore, MD simulations of engineered dynein constructs were performed. Using the full length motor domain system of our earlier study (~800,000 atoms), an extensive set of new simulations totalling 500ns in length were performed. In these simulations the linker domain is guided towards its post-power stroke structure and free energy surface along the transition is constructed.

Keywords: Dynein, Molecular Dynamics, Advanced Molecular Dynamics, Mechanochemical, Thermodynamics

References: [1] Roberts AJ, Kon T, Knight PJ, Sutoh K, Burgess SA. Functions and mechanics of dynein motor proteins. Nature reviews Molecular cell biology. 2013;14(11):713.
[2] Reck-Peterson SL, Redwine WB, Vale RD, Carter AP. The cytoplasmic dynein transport machinery and its many cargoes. Nature Reviews Molecular Cell Biology. 2018;19(6):382.
[3] Schmidt H, Gleave ES, Carter AP. Insights into dynein motor domain function from a 3.3-Å crystal
structure. Nature structural & molecular biology. 2012;19(5):492.

[4] Can S, Lacey S, Gur M, Carter AP, Yildiz A. Directionality of dynein is controlled by the angle and
length of its stalk. Nature. 2019;566(7744):407.

Corresponding Author's Address:
Gümüşsuyu, İnönü Cd. No:65, 34437 Beyoğlu/İstanbul, Turkey / gurme@itu.edu.tr website: gurlab.itu.edu.tr

POSTER PRESENTATIONS

DIFFERENTIAL CO-EXPRESSION ANALYSIS OF HUMAN CANCERS TO REVEAL SYSTEMS BIOMARKERS FOR DIAGNOSTIC AND THERAPEUTIC TRIALS

Meltem Nur Erdöl[1], *Kazım Yalçın Arğa*[1]

1. Marmara University, Faculty of Engineering, Bioengineering Department, Istanbul

Differential co-expression analysis, a new approach in systems biology, has emerged as a complementary method to traditional differentialgene expression analysis. In order to understand the genetic background of diseases, it is important to reveal the interactions of genes in diseased and healthy conditions.[1] In this study, some types of cancer (gastric cancer, lung cancer, pancreatic cancer) were examined according to their histological differences and it was discussed how these results could be used for research of diagnostic and therapeutic applications in cancer. The relevant gene expression data were obtained from public databases, including NCBI-GEO and TCGA databases. To characterize the differentially expressed genes, each data set was normalized by means of the Robust Multi-Array Average (RMA) expression measure as implemented in the "affy" package of "R / Bioconductor" and statistical analyses were performed by LIMMA. Co-expression profiles were constructed from gene clusters that differ in their expression in each type of cancer examined and genetic networks were established to indicate whether they were expressed differently in tumor and healthy individuals. Highly interacting sub-clusters of these networks were assigned as modules and these modules were investigated whether they are significant for diagnostic and therapeutic applications of cancer.

Keywords: Differential Co-Expression Analysis, Biomarker, Cancer

References: [1] Siska C, Bowler R, Kechris K. The discordant method: a novel approach for differential correlation. Bioinformatics. 2016;32(5):690–6

Corresponding Author's Address: kazim.arga@marmara.edu.tr

Putative drug target identification in Trichophyton rubrum using sequential elimination method and evaluating the phytocompounds from Balanites aegyptiaca against the identified targets – A Computational Approach

Syed Abuthakir Mohamed Husain[1], *Sharmila Velusamy*[2] *and Jeyam Muthusamy*

*1&2. Research scholar, Biochematics lab, Department of Bioinformatics, Bharathiar University *Assistant Professor, Biochematics lab, Department of Bioinformatics, Bharathiar University, Coimbatore,Tamil Nadu, India.*

Trichophyton, Microsporum and Epidermophyton are group of fungi called Dermatophytes causing superficial infections in skin, nail and hair of all living organisms. Trichophyton rubrum is mainly causing tinea pedis, tinea cruris, and tinea corporis in human and animals. Currently, various drugs are targeting the fungal proteins to inhibit the growth of dermatophytes. But, these drugs are producing some side effects in human because of drug interactions with human proteome and human gut microbiome. The present study used the computational methods to identify the proteins unique to T.rubrum by sequential elimination of proteins similar to that of human as well as gut microbiota and to find the phytocompounds from the edible fruit pulp of Balanites aegyptiaca to target these proteins which can lead to better treatment without side effects in human. By this way, the whole proteome of T.rubrum was analysed to identify the essential proteins of T.rubrum which are non-homologous to proteins of human and gut flora, non-homologous against human domain architecture and to find sub-cellular localization of the selected protein, functional classification of hypothetical protein, protein network analysis and druggability of the targets. Finally 7 novel drug targets were identified from T.rubrum, 3D structures of these targets were modelled using I-TASSER and docked with LC-MS derived compounds of fractioned methanol extract of fruit pulp of B.aegyptiaca using GLIDE module of Schrodinger. The compound Cyanidin-3-O-rhamnoside gave better result with all the targets and may have a good multi-targeting potential against Trichophyton rubrum which has to be further confirmed by in vitro and in vivo experiments.

Keywords: Trichophyton Rubrum, Putative Targets, Balanites Aegyptiaca, Gu Flora, Docking.

References: [1] Barry. L, and Hainer.MD. Dermatophyte Infections. American Family Physician. 2003. 67(1): 101-108.

[2] Saboo, SS, Chawan.RW, Tapadiya.GG and Khadabadi.SS. An important ethnomedicinal plant Balanite aegyptiaca Del. International journal of Phytopharmacy. 2014. 4(3):75-78.

[3] Hasan. MA, Rahman.MA, Noore.MS, Ullah.MR, Rahman.MH, Hossain.MA, Ali.Y and Islam.MS. Identification of potential drug targets by subtractive genome analysis of Bacillus anthracis A0248: An in silico

approach. Computational Biology and Chemistry. 2014. 52:66-72.

Corresponding Author's Address:

Dr.M.JEYAM M.Sc.,M.Phil.,Ph.D.,PGDBI

Assistant Professor,

Department of Bioinformatics,

Bharathiar University,

Coimbatore-641 046.

E.mail: jeyam@buc.edu.in

URL: http://cdn.b-u.ac.in/faculty_data/bioinfo_dr_jeyam.pdf

PREDICTING CARBON SPECTRUM IN HSQC FOR ONLINE FEEDBACK DURING SURGERY

E. Onur Karakaslar[1], Baris Coskun[1], Hassiba Outilaft[2], Izzie Jacques Namer[2], A. Ercument Cicek[1,3]

1.Bilkent University, Computer Engineering Dept. Cankaya, Ankara 06800 2. University of Strasbourg, Strasbourg 67081, France.
3. Carnegie Mellon University, Computational Biology Dept., Pittsburgh PA 15213

High Resolution Magic Angle Spinning (HRMAS) Nuclear Magnetic Resonance (NMR) spectroscopy is a technology that can efficiently detect and quantify metabolites in solid tissues. HRMAS-NMR does not need any chemical extraction procedure, which is a must for MS technologies and liquid state NMR. Thus, it is frequently used in biopsy analyses and the results can be obtained in < 20 minutes. Rapid response enables giving feedback to surgeons during an ongoing surgery. Recently, Battini et al. proposed using HRMAS-NMR for pancreatic adenocarcinoma surgeries [1]. Even if it might seem like the tumor is completely removed, it is possible that residual tumor cells are left over in the excision cavity. Then there is the trade-off between removing healthy tissue, which risks the well being of the patient and leaving tumor cells in the body, which risks recurrence and further surgeries. In this system, the surgeon gets samples from the excision cavity for identifying possible left-over tumor cells. After HRMAS analysis, parts of the cavity that have tumor-like spectrum are reported for further resection. This pipeline is possible because the feedback is available within 20 minutes. Even though 1H is commonly used due to high sensitivity and natural abundance in samples, identification of biomarker metabolites can be impossible due to overlapping peaks in 1H-NMR spectra. In that case, a second experiment called Heteronuclear Single Quantum Coherence Spectroscopy (HSQC)-NMR is performed. This analysis generates a 2D correlation plot for 1H and 13C spectra. However, this analysis requires around 15 hours to complete and is outside of the time frame of surgery. In this study, we propose two methods to predict 13C spectra in the HSQC experiment, without performing the HSQC experiment at all. These methods are (i) performing multivariate multiple regression and (ii) repurposing STOCSY for a blind prediction of a single sample. Using a set of 1H-13C HSQC experiments, methods learn how each peak in 1H dimension affects each peak in 13C dimension. With only 1H HRMAS NMR for 14 human brain tissue samples and predicting their corresponding 13C spectra, we show that we can successfully identify presence and absence of 104 groups belonging 39 metabolites. Both methods achieve 97.1% accuracy in less than a second. We also show on one of these samples that regression model can be used to reconstruct the 2D HSQC experiment as well. Figure 1 shows that we are able to predict the presence of creatine even though its peak is overlapping with lysine in 1H dimension. Panel A shows the actual HSQC experiment. The zoomed square (Panel C) shows that HSQC was able distinguish lysine and creatine. Note that their 1H peaks (signal on x axis, on top) is overlapping; thus, one cannot distinguish those by just looking at the

1H dimension. Panel B and Panel D show the prediction of NLSPR for the same sample. Clearly, without performing the HSQC experiment we would be able distinguish creatine and lysine by our prediction. The reconstructed 1D 13C signal is also provided in Panel E, which also shows that the peaks of creatine and lysine are distinguished in 13C dimension and both NLSPR and STOCSY can predict those peaks accurately. Creatine is an indicator of hypoxia and possibly drug resistant tumor tissue [2,3]. Thus, our approach can make it possible to provide accurate feedback to the surgeon during the surgery even if 1H HRMAS NMR results are inconclusive. Even though we experiment on 1H - 13C HSQC NMR dimensions in this study, all methods can be used with any other 2D spectra as well. This work has recently been accepted for publication in IEEE TCBB and is available at https://ieeexplore.ieee.org/document/8730423

Keywords: Multiple Multivariate Regression; Stocsy; Peak Prediction; Nmr; Hsqc; Surgery Feedback

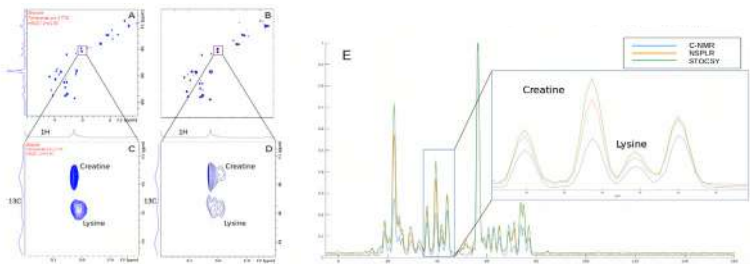


Figure 1. This figure shows the 1H-13C HSQC NMR of Sample 3 and its reconstructed version. (A) Original spectra captured with Bruker TopSpin 3.5. (B) Reconstructed version of same spectra in (A) obtained from only 1H-NMR sample. (C) Zoomed version of sample in (A), this figure shows two metabolites Creatine and Lysine overlapping on 1H dimension of HSQC NMR, yet differentiable on 13C dimension. (D) Zoomed version of (B), same peaks (Creatine and Lysine) are shown in the reconstructed version. (E) Reconstructed 13C NMR spectra of the same sample using NSPLR and STOCSY, both metabolites are shown as well.

References: [1] S. Battini, F. Faitot, A. Imperiale, A. Cicek, C. Heimburger, G. Averous, P. Bachellier, and I. Namer, "Metabolomics approaches in pancreatic adenocarcinoma: tumor metabolism profiling predicts clinical outcome of patients," BMC Med., vol. 15, no. 1, 2017, Art. no. 56.

[2] Wyss M, Kaddurah-Daouk R. Creatine and creatinine metabolism. Physiological reviews. 2000 Jul 1;80(3):1107-213.

[3] Z. Luo, H. Tian, L. Liu, Z. Chen, R. Liang, Z. Chen, Z. Wu, A. Ma, 498 M. Zheng, and L. Cai, "Tumor-targeted hybrid protein oxygen carrier to 499 simultaneously enhance hypoxia-dampened chemotherapy and 500 photodynamic therapy at a single dose," Theranostics, vol. 8, no. 13, 2018, 501 Art. no. 3584.

Corresponding Author's Address: Bilkent University, Cankaya, Ankara 06800 – cicek@cs.bilkent.edu.tr -http://ercumentcicek.com

CONTROLLING FDR IN EPISTASIS TEST PRIORITIZATION

Gizem Caylak[1], *A. Ercument Cicek*[1,2]

1. Bilkent University, Computer Engineering Dept. Cankaya, Ankara 06800
2. Carnegie Mellon University, Computational Biology Dept., Pittsburgh PA 15213

Analyzing single loci associations in GWAS have provided many valuable insights but they alone do not account for the full heritability. Statistically significant interactions between two or more loci is called epistasis and it has been shown to contribute to complex genetic traits. Given a million variants in a genome, a trillion tests are required to process all SNP pairs. Thus, this procedure is not only computationally prohibitive, but also lacks statistical power due to multiple hypothesis testing. A popular approach is to prioritize the tests to be performed rather than discarding pairs from the search space and control for type I error. In this approach, the user can keep performing tests, in the order specified by the algorithm, until a desired number of significant pairs are found. While false negatives may arise, the idea is to provide the user a tractable number of true positives with minimum number of tests performed to ensure statistical power. Cowman and Koyutürk (2017), introduced the state-of-the-art LINDEN algorithm. The method first generates trees that represent genomic regions (LD forest). Then, by comparing the roots of these trees it decides if this pair of genomic regions is a promising candidate for epistasis test. Trees are parsed in depth first manner and leaf pairs are tested only if the estimation at higher levels provides a value larger than a threshold. All three methods aim at avoiding testing pairs that are in LD. The fundamental problem in all of the above-mentioned algorithms is the high number of false positives (FP) (i.e., Linden's false discovery rate (FDR) is ~0.99, 10 TP, ~1800 FP). FPs are SNPs that are being tested not crossing the Bonferroni-corrected significance threshold. While these algorithms aim at minimizing the number of predictions and the number of tests, none has aimed at controlling for the false positive rate, which is an important indication for life scientists using these algorithms. In an orthogonal study, Yilmaz et al. (2018) avoids LD in a different manner and for phenotype prediction problem. They show that while looking for a small set of loci (i.e., 100) that is the most predictive of a continuous phenotype in GWA study, selecting SNPs further away from each other results in better predictive power. This method (SPADIS) is designed for feature selection for multiple regression and as the SNP set it generates contains diverse and complementary SNPs it results in better R2 values. In this study, we conjecture that FDR can be controlled by guiding the prioritization algorithms using SPADIS. The hypothesis is that the set of SNPs selected by SPADIS are likely to be epistatic, since the algorithm is designed to diversify the set and select complementary SNPs. We created a pipeline that first uses SPADIS to generate its candidate set for epistasis test. Instead of using this set for all-pairs epistasis testing which would still return a large number of FPs, we use it to guide the state of the art example of these algorithms: Linden. We let it only form LD trees over SPADIS-selected regions (selected SNPs and a small

number of neighbors) to pick likely epistatic pairs from this set. Thus, LINDEN does not have perform still a sheer number of tests that cover the genome but a pruned search space of likely epistatic pairs. We measure the improvement in precision using this approach on the Wellcome Trust Case Control Consortium (WTCCC) Type 2 Diabetes (T2D) GWAS data, which was also used in (Cowman and Koyutürk, 2017). First, we run Linden on this dataset and, it returned 1792 reciprocally significant epistatic pairs, only 10 passing the significance threshold 0.1 (Bonferonni corrected). This resulted in a precision of 0.0055. Complete results are shown in Figure 1A. The guidance of SPADIS improves the precision substantially, up to ~55%. Figure 1B shows the significance of pairs detected, where green line denotes the significance level to be passed for each approach (k=1000). This figure shows that we get only a few FPs, given more TPs, when our pipeline is used. Moreover, the total runtime of the pipeline is only one forth of the Linden-only run (~15min vs 1+ hour). This project is being supported by TUBITAK 3501 Career Grant 116E148 to A. Ercument Cicek.

Keywords: Epistasis Test Prioritization, Snp Selection, Submodular Optimization, Fdr

METHOD		# Tested loci	# Pairs Reported	Precision A
LINDEN		378016	1792 (10)	0,0055
SPADIS + LINDEN	k=500	9250	20 (11)	0,550
	k=1000	18943	55 (28)	0,509
	k=5000	94939	389 (26)	0,067
	k=10000	189877	773 (29)	0,038

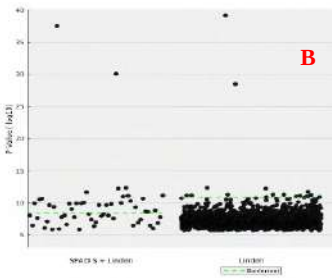


Figure 1. Panel A: Number of pairs reported is the total number of reciprocally significant pairs returned by LINDEN with and without the guidance of SPADIS for varying number of selected SNPs. For each SPADIS-selected SNP 18 closest neighbors are also input to Linden. The number in parentheses denotes the significant pairs passing significance threshold (0.1) with Bonferroni correction based on the number of tests performed by each method. Table shows that the guidance of SPADIS increases the precision substantially as compared to LINDEN only. Panel B: Each approach, SPADIS+LINDEN and LINDEN-only, we show the significance levels (y-axis) of each reported pair (dots) given the Bonferonni corrected significance threshold (0.1, green line). X – axis is just randomly assigned values to pairs for visualization. k=1000. SPADIS clearly minimizes FPs by guiding Linden.

References:[1] Cowman, T. and Koyutürk, M. (2017). Nucleic acids research, 45(14), e131–e131
[2] Yilmaz, S. et al. (2018). bioRxiv https://doi.org/10.1101/256677

Corresponding Author's Address:

Bilkent University, Cankaya, Ankara 06800 – cicek@cs.bilkent.edu.tr - http://ercumentcicek.com

DORMAN: DATABASE OF RECONSTRUCTED METABOLIC NETWORKS

Furkan Ozden[1], Metin Can Siper[1],, Necmi Acarsoy[1], Tugrulcan Elmas[1], Bryan Marty[2], Xinjian Qi[2], A. Ercument Cicek[1,3]

1. Bilkent University, Computer Engineering Dept. Cankaya, Ankara 06800,
2. Case Western Reserve University, EECS Dept., Cleveland OH 44106 3 Carnegie Mellon University, Computational Biology Dept., Pittsburgh PA 15213

With the advancements in the omics platforms and the availability of affordable high throughput data, researchers have been able to capture the genome-scale chemical composition of the cell and integrate this knowledge into genome-scale reconstructed metabolic networks of organisms. Reconstructed models have proven to be indispensable tools for understanding the metabolism in a diverse spectrum of applications such as: (i) metabolic engineering, (ii) model-directed discovery, (iii) interpretations of phenotypic screens, (iv) analysis of network properties, and (v) studies of evolutionary processes. Also many studies have made use of the topology of the networks for understanding disease mechanisms. Ongoing interest in network reconstruction and analyses comes with the need for computational tools to work on the resulting models. This everlasting research interest lead to the development of online systems/repositories that store existing reconstructions and enable new model generation, integration and constraint-based analyses. While features that support model reconstruction are widely available, current systems lack the means to help users who are interested in analyzing the topology of the reconstructed networks. Here, we present the Database of Reconstructed Metabolic Networks (DORMAN); a central database that collects available genome-scale reconstructed metabolic networks from the literature and provides a user friendly and efficient platform for accessing, visualizing and querying the models with multiple interfaces. While above mentioned systems are mostly focused on the integration process as explained above, DORMAN provides novel complementary features, which lets users to analyze the topology of the networks. First of all, current systems provide either no or only pathway-level static visualization (i.e., Escher maps of BiGG) while DORMAN's visualization tool can render the full network (not just pathways) with compartments and lets users to interact with the model (e.g., zoom in/out; move, add or remove nodes in the network) – See Figure 1 panels A and B. Unlike its counterparts, hierarchical navigation feature of DORMAN allows users to efficiently browse through connected entities in the model, level-by-level. For instance, users can start browsing the reactions of a model, and then can list and select the metabolites of a selected reaction, and finally, can list the compartments associated with the selected metabolites. The browser interface is integrated with KEGG data and can be used to navigate whenever KEGG pathway or KEGG molecule data is made available by the model. External databases such as UniProt, ChEBI and KEGG are linked when the information is provided in the reconstruction. DORMAN's built-in graph queries can be used to search for topologically related entities in the graph

such as searching for entities in k-hop neighborhoods of reactions or metabolites. DORMAN also provides an interface to compare models that do not follow the same nomenclature, using approximate name matching of the entities. See Figure 1, panels C, D and E for the performance of the tool with respect to a ground truth matching. Currently, the database contains 199 SBML-based models obtained from external repositories and is continuously updated by screening the literature and available repositories. System is online and available at http://ciceklab.cs.bilkent.edu.tr/dorman

Keywords: Genome-Scale Reconstructed Metabolic Network, Online Workbench, Online Repository

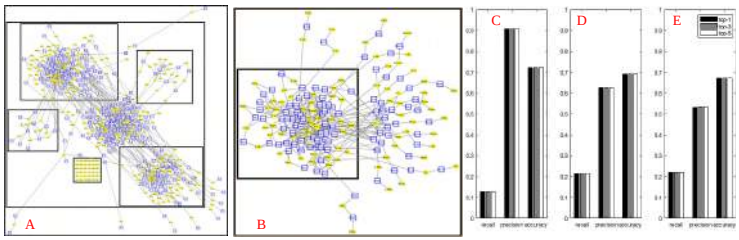


Figure 1. Visualizations of the iAM303 model (A) and E.Coli textbook model (B). Parts C, D and E show the performance of the Compare Networks tool on the following 4 models: IRC1080, E.coli iAF1260, E.coli iJR904 and H. sapiens Recon 1. The ground truth ma matches are obtained from the MetaNetX system. Three configurations are used to tune the strictness of the tool in terms of returning exact matches versus less similar matches: Panel C: q = 7, k = 1 (most strict); Panel D: q = 5, k = 3 and Panel E: q = 3, k = 5 (most relaxed). Results for top-1, top-3 and top-5 matches are shown.

References:[1] Schellenberger, J. et al., "BiGG: a Biochemical Genetic and Genomic knowledgebase of large scale metabolic reconstructions", BMC Bioinformatics, 11:213, 2010

[2] King, Z.A. et al., "BiGG Models: A platform for integrating, standardizing, and sharing genome-scale models", Nucleic Acids

Research, 44(D1):D515-D522, 2016

[3] Henry, C.S. et al., "High-throughput generation, optimization and analysis of genome-scale metabolic models", Nat Biotechnol, 28(9):977-82, 2010

[4] Pabinger, S. et al., "MEMOSys: Bioinformatics platform for genome-scale metabolic models", Syst Biol, 5:20, 2011

[5] Kumar, A. et al., "MetRxn: a knowledgebase of metabo-lites and reactions spanning metabolic models and da-tabases", BMC Bioinformatics, 13:6, 2012

[6] Ganter, M. et al., "MetaNetX. org: a website and reposi-tory for accessing, analysing and manipulating meta-bolic networks", Bioinformatics 29(6): 815-816, 2013

Corresponding Author's Address:

Bilkent University, Cankaya, Ankara 06800 – cicek@cs.bilkent.edu.tr - http://ercumentcicek.com

ABILITY OF BDELLOVIBRIO BACTERIOVORUS TO REMOVE BACTERIAL BIOFILMS

Nedal Said[1], Antonis Chatzinotas[2] and Matthias Schmidt[1]

1. Department of Isotope Biogeochemistry, ProVIS - Centre for Chemical Microscopy;
2. Department of Environmental Microbiology Helmholtz Centre for Environmental Research – UFZ, Leipzig/Germany

Bdellovibrio bacteriovorus is a bacterial predators that attacks Gram-negative bacteria and that has therefore been considered as promising biological agent for the control of unwanted bacteria in an environmental, biotechnological or medical context it inters the periplasmic space of the prey, which is transform into a spherical-shaped bdelloplast. The Bdellovibrio elongates and divides within the host by consuming prey cell components. Finally the Bdelloplast lyses and the offspring are released. This life-cycle is known to last for about three to four hours. In our previous study we showed that B.bacteriovorus HD100 attached to Pseudomonas .putida prey cells within 10 min of co-incubation; first bdelloplasts in early stage were seen after 20 min and bdelloplast lysis occurred after approximately 2 h[1].The formation of biofilms by pathogenic bacteria can increase their resistance to therapeutic substances [2]. The 2017 WHO list of the most critical bacteria with respect to antibiotic-resistance comprises Acinetobacter baumannii, P.aeruginosa and Enterobacteriaceae, all of which are Gram-negative and strong biofilm formers[3].The capability of B.bacteriovorus to remove bacterial biofilms[4] makes it an interesting candidate for "living antibiotics"[5].In this study, we examined the ability of B. bacteriovorus HD100 to prey on P.putida biofilms. Scanning electron microscopy was employed to reveal if P. putida biofilms are removed by B. bacteriovorus HD100. To see the effect of B.bacteriovorus HD100 on the P.putida biofilms we analyzed the predator-prey system during different-time points after inoculation, i.e. 2, 4, 6, 8 and 24 hours. P.putida biofilms were 2 days old, grown on PVC coverslip in LB media at 28°C without shaking. From the microscopic data we conclude that the extracellular polymeric substances in the P.putida biofilm delay the predator's life-cycle but does not protect the prey from being attacked. Thus B.bacteriovorus might hold promise to be a suitable biological agent capable to remove biofilms of pathogenic.

Keywords: Bdellovibrio Bacteriovorus; Biofilm; Pseudomonas Putida; Pathogens; Biological Control

References:

[1] Said N., Chatzinotas A., Schmidt M. (2019). Have an Ion on It: The Life Cycle of Bdellovibrio bacteriovorus Viewed by Helium-Ion Microscopy; Adv. Biosys. 3, 1800250. [2] Gupta P., Sarkar S., Das B., Bhattacharjee S., Tribedi P. (2016). Biofilm, pathogenesis and prevention – a journey to break the wall: a review. Arch. Microbiol.198(1):1–15. [3]https://www.who.int/news-room/detail/27-02-2017-who publishes-list-of-bacteria-for-which-new-antibiotics-are-urgently-needed. [4] Dashiff, A., et al., (2011). Predation of human pathogens by the predatory bacteria Micavibrio aeruginosavorus and Bdellovibrio bacteriovorus. J. Appl Microbiol. 110(2): 431-44. [5] Strauch E., Sebastian Beck S.,

Appel B. (2007). Bdellovibrio and Like Organisms: Potential Sources for New Biochemicals and Therapeutic Agents?, In Predatory Prokaryotes (pp. 131-152)..

Corresponding Author's Address: E-mail: nedal.said@ufz.de

INVESTIGATION OF INTERACTIONS BETWEEN SUBUNITS OF POLYPEPTIDES AND NUCLEIC ACIDS (MONOMERIC-TETRAMERIC) BY MOLECULAR MODELLING

Fulya Çağlar[1], *Cenk Selçuki*

1. Ege University, Institute of Health Sciences, Health Bioinformatics Program
2. Ege University, Faculty of Science, Department of Biochemistry

Transcription factors (TF) are proteins which are consisted of combinations of many different amino acids. The amino acid sequences in the structure provide three-dimensional structure due to the entalpy value of the external environment and the non-covalent interactions. A protein must be in its optimal 3D conformational structure to perform its function: effectively binding to DNA [1].

In this study, the non-covalent interactions between AG and CC nucleotide dimer sequences in the single- stranded DNA, known to be effectively bound to REST, and the triple serine structure of the DNA binding region of RE-1 (Suppressor Element-1) silencer transcription factor (REST) were investigated by computational methods. For this purpose, Spartan 14 program was used to determine initial structures for the serine trimers and nucleotide dimers by conformational analysis. The MOPAC2016 software [2,3] with PM6- D3H4 method and Gaussian09 software [4] with B3LYP-6-311++G(d,p) level were used for optimizations and frequency calculations for each investigated system. Using the calculated results, the most stable structures were determined based on the relative energies. Molecular structures were presented by Discovery Studio Visualizer 2019 software.

The results revealed that serine residues in the active DNA binding region of REST protein can form hydrogen (–H) bonds with DNA. Other non-covalent interactions such as van der Waals forces were not observed. Most of the calculations were performed on TUBITAK-ULAKBIM Truba resources.

Keywords: REST; DNA; Molecular Modelling

References: [1] Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015;16(3):129–143. PubMed PMID: 25693126
[2] J.D.C. Maia, G.A. Urquiza Carvalho, C.P. Mangueira, S.R. Santana, L.A.F. Cabral, G.B. Rocha, GPU Linear Algebra Libraries and GPGPU Programming for Accelerating MOPAC Semiempirical Quantum Chemistry Calculations, J. Chem. Theory Comput. 8 (2012) 3072–3081. doi:10.1021/ct3004645.
[3] S.C.C. James Stewart, MOPAC2016, (2016). http://openmopac.net/.
[4] Rhodes R. and Vargas V. Gaussian multiplicative chaos and applications: a review. Probability Surveys, 2014; (11): 315–392. ISBN: 978-1-935522-03-4

Corresponding Author's Address: Ege University, Faculty of Science, Department of Biochemistry cenk.selcuki@gmail.com

MAKING MOST FROM FERMENTATION DATA TO LEARN ABOUT PHYSIOLOGY FOR PRODUCTION OF COMMERCIALLY SIGNIFICANT BIOMOLECULES

Ezgi Tanıl[1], *Emrah Nikerel*[1]

1. Yeditepe University, Department of Genetics and Bioengineering

Since 1980's, microbial production of biopharmaceuticals has gained increasing importance due to its lower production cost and allowing production of tailor-made, specific drugs that are impossible or impractical to produce using chemical methods such as hormones, proteins, and even DNA-RNA based products for using in various therapy applications [1]. Such bioprocesses have their own challenges in designing to produce these biomolecules with high and specific activity and productivity in the most efficient manner. This requires careful investigation carbon, electron and other elemental as well as ATP balances, not only for overall production efficiency but also to monitor and control quality during production, which calls a number of physiological parameters specific to the production host of interest, e.g. ATP maintenance parameters to determine the energetic needs of organisms for growth, cellular maintenance and product formation [2]. However, obtaining these parameters is challenging, despite several datasets available in literature. Efficient use of these fermentation data for estimation of physiological parameters is crucial to better understand cellular energetics, improve substrate use efficiency thereby decrease production costs.

The aim of this study is two fold: (i) present a framework to estimate the parameters related to energetics of growth and production as growth associated (Kx) and non-growth associated (mATP) maintenance constants (ii) set up and analysis of unstructured black-box kinetic model to describe the kinetics of citric acid production by Candida oleophila ATCC20177 from published fermentation data of diverse sources resulting from different fermentation modes [3-5]. Interestingly, the ATP balance can be reformulated as a linear regression problem ($\alpha = Y \cdot \beta$) to estimate these parameters from data, allowing not only the application of regression diagnostic methods such as detection of collinearities and partial least squares but also integration of data from various sources in a single problem. Kx and mATP were found to be 2.3 ± 1.7 and 5.25 ± 2.75 respectively for published P/O ratio of 1.45. It was also found that effects of P/O change on Kx was much lower than on mATP. It was also shown that, parameters obtained using data from relatively less laborious batch fermentation can be used to predict exometabolome levels of chemostat experiments.

Keywords: Energetic Parameters, Black-Box Kinetic Model, Atp Balance

References: [1] Jozala AF, Geraldes DC, Tundisi LL, Feitosa VA, Breyer CA, Cardoso SL, Mazzola PG, Oliveira- Nascimento L, Rangel-Yagui CO, Magalhães PO, Oliveira MA, Pessoa A Jr. Biopharmaceuticals from microorganisms: from production to purification. Braz J Microbiol.

2016 Dec;47 Suppl 1(Suppl 1): 51-63. PubMed PMID: 27838289
[2] vanGulik WM, Antoniewicz MR, deLaat WT, Vinke JL, Heijnen JJ. Energetics of growth and penicillin production in a high-producing strain of Penicillium chrysogenum. Biotechnol Bioeng. 2001 Jan;72(2):185-93. PubMed PMID: 11114656
[3] Anastassiadis S, Wandrey C, Rehm HJ. Continuous citric acid fermentation by Candida oleophila under nitrogen limitation at constant C/N ratio. World J. Microbiol. Biotechnol. 2005 Jul;21: 695–705.
[4] Anastassiadis S, Rehm HJ. Continuous citric acid secretion by a high specific pH dependent active transport system in yeast Candida oleophila ATCC 20177. Electron. J. Biotechnol. 2005 Aug;8: 146–161.
[5] Anastassiadis S, Rehm HJ. Citric acid production from glucose by yeast Candida oleophila ATCC 20177 under batch, continuous and repeated batch cultivation. Electron. J. Biotechnol. 2006 Jan;9: 26-39.

Corresponding Author's Address: Yeditepe University, Department of Genetics and Bioengineering, 26 Ağustos Yerleşimi, Kayışdağı Cad, 34755 Ataşehir, İstanbul e-mail: emrah.nikerel@yeditepe.edu.tr

COMPUTATIONAL ANALYSIS OF THE STRUCTURE, GLYCOSYLATION AND CMP BINDING OF HUMAN ST3GAL SIALYLTRANSFERASES

Muhammet Uslupehlivan[1], *Ecem Şener*[1], *Savaş İzzetoğlu*[1]

1. Ege University, Faculty of Science, Department of Biology, Molecular Biology Section, İzmir/ Turkey

Sialyltransferases (STs) are the fundamental enzymes which are related to many biological processes such as cell signalling, cellular recognition, cell-cell and host-pathogen interactions and metastasis of cancer [1-3]. All STs catalyze the terminal sialic acid addition from CMP donor to the glycan units[4,5]. ST3GAL family is one of the most important STs and divided into the six subfamily in mouse and humans which are ST3Gal I, ST3Gal II, ST3Gal III, ST3Gal IV, ST3Gal V, and ST3Gal VI. The members of the ST3GAL family transfer sialic acid to the terminal galactose residues of glycochains through an α2,3-linkage [6-9]. There are many reports on the ST3GAL function in mammals [10,11] but, there is a paucity of information about the three-dimensional structure, glycosylation, and CMP interaction of human ST3GAL family. Herein, we investigated the structure, glycosylation and CMP binding site of human ST3GAL family using computational methods. We found for the first time N-glycosylation positions in ST3Gal IV and VI, mucin type glycosylation in ST3Gal III and O-GlcNAcylation in ST3Gal V and their relation with sialylmotifs. In addition, we predicted CMP binding positions of human ST3GAL enzyme family on three-dimensional structure using molecular docking and first demonstrated the sialylmotifs relation with the CMP binding positions in ST3Gal III-VI subfamilies. In conclusion, we suggest that predicted positions may contribute to design of selective inhibitors and generating position specific therapeutic targets for metastasis since the control of ST activity in prevention of metastasis in different types of cancer may have been an effective approach.

Keywords: Human St3gal Family, Sialylmotif, Glycosylation, Cmp Ligand Binding, Molecular Docking, 3D Molecular Modelling

References:[1] Varki A. Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins, Nature. 2007 446(7139), 1023.
[2] Traving C, Schauer R. Structure, function and metabolism of sialic acids. Cell. Mol. Life Sci. 1998 54, 1330e1349
[3] Schultz MJ, Swindall AF, Bellis SL. Regulation of the metastatic cell phenotype by sialylated glycans. Cancer Metastasis Rev. 2002 31: 501.
[4] Harduin-Lepers A, Vallejo-Ruiz V, Krzewinski-Recchi MA, Samyn-Petit B, Julien S, Delannoy P. The human sialyltransferase family, Biochimie. 2001 83(8), 727-737.
[5] Harduin-Lepers A, Mollicone R, Delannoy P, Oriol R. The animal sialyltransferases and sialyltransferase-related genes: a phylogenetic approach. Glycobiology. 2005 15: 805-817.
[6] Kitagawa H, Paulson, JC. Cloning of a novel alpha 2, 3-sialyltransferase that sialylates glycoprotein and glycolipid carbohydrate groups. Journal of Biological Chemistry. 1994 269(2), 1394-1401.

[7] Kim YJ, Kim KS, Kim SH, Kim CH, Ko JH, Choe IS, ... & Lee YC. Molecular cloning and expression of human Galβ1, 3GalNAc α2, 3-sialyltransferase (hST3Gal II). Biochemical and biophysical research communications. 1996 228(2), 324-327.
[8] Okajima T, Fukumoto S, Miyazaki H, Ishida H, Kiso M, Furukawa K., ... & Furukawa K. Molecular cloning of a novel α2, 3-sialyltransferase (ST3Gal VI) that sialylates type II lactosamine structures on glycoproteins and glycolipids. Journal of Biological Chemistry. 1999 274(17), 11479-11486.
[9] Priatel JJ, Chui D, Hiraoka N, Simmons CJ, Richardson KB, Page DM, ... & Marth JD. The ST3Gal-I sialyltransferase controls CD8+ T lymphocyte homeostasis by modulating O-glycan biosynthesis. Immunity. 2000 12(3), 273-283.
[10] Rao FV, Rich JR, Rakić B, Buddai S, Schwartz MF, Johnson K, ... & Strynadka NC. Structural insight into mammalian sialyltransferases. Nature Structural and Molecular Biology. 2009 16(11), 1186.
[11] Rakić B, Rao FV, Freimann K, Wakarchuk W, Strynadka NC, & Withers SG. Structure-based mutagenic analysis of mechanism and substrate specificity in mammalian glycosyltransferases: porcine ST3Gal-I. Glycobiology. 2013 23(5), 536-545.

Corresponding Author's Address: Ege University, Faculty of Science, Department of Biology, Molecular Biology Section, 35100, Bornova/İzmir
E-mail: savas.izzetoglu@ege.edu.tr

A Molecular Characterization and genetic structure of Domestic Cavies (Cavia porcellus) Population in Cameroon

Youchahou Poutougnigni Matenchi[1] *and Evren Koban Baştanlar*[1]

1. Ege University, Dept. of Biology, İzmir, Türkiye

To study the variability of domestic cavies (Cavia porcellus) populations in Cameroon, a panel of 16 microsatellite markers was used. Genomic DNA was extracted from blood samples of 110 unrelated animals from accessible farms in the bimodal agroecological zone, taking into account two groups according to their origin - either the Centre or the East.

Among the 16 selected microsatellites markers, 3 (Cavy 1, 13 and 14) were non-informative while the rest was highly polymorphic. A total of 87 alleles were observed at 13 loci; varying from 4 (cavy 3) to 13 (cavy 6), with a mean of 6.69 alleles per locus. The estimated inbreeding coefficient (FIS) varied from -0.14 at locus cavy 4 to 0.77 at locus cavy 16 with a mean of 0.37. The estimated PIC value for each locus ranged between 0.05 and 0.86 with a mean of 0.45 such that; markers with a high PIC estimation can further be used for the assessment of the genetic relationship between different cavy populations at molecular level. A deviation from Hardy-Weinberg's equilibrium was observed, except for Cavy 4 marker. AMOVA revealed that only 3% of the observed genetic diversity were due to differences among populations, while 35% of the diversity were due to differences within populations, and 62% was due to differences between individuals within the total population. The genetic distance between the two populations was small (0.041) and suggested a relatively recent separation or still ongoing genetic mixing through interchanging breeding individuals. The estimated inbreeding level was significant, i.e. 28.5% and 33.3% for the East and Centre populations, respectively. The expected heterozygosity was higher (0.504 ± 0.061 and 0.442 ± 0.070) than the observed heterozygosity (0.317 ± 0.047 and 0.295 ± 0.052) in both the Center and the East region respectively, suggesting a deficit in heterozygotes in the population. Moreover, the F indices per locus were estimated: the FIT varied from 0.004 (locus cavy 5) to 0.043 (locus cavy 6); the FST varied from -0.150 (locus cavy 4) to 0.742 (locus cavy 5); and the FIS varied from -0.154 (locus cavy 4) to 0.739 (locus cavy 16). All these estimations suggest a high inbreeding and low differentiation leading to an excess in homozygotes at the locus level. In addition, the gene flow, which is found high (20.866 ± 5.309), between the populations reinforces the lower differentiation and the tendency of homogeneity in the total population.

The phylogenetic tree of individuals and the population structure analysis revealed 3 genetically distinct groups: an eastern group which is more closed a center widespread group, and a mixture group made up of animals from both region of individual in the bimodal agro-ecological zone in Cameroon.These preliminary results pave the way to genetic improvement scheme for domestic cavy in Cameroon.

Keywords: Genetic Diversity, Guinea Pigs, Microsatellites, Cameroon

References: Numbela ER, Valencia CR, 2003. Guinea pig management manual. Benson Agriculture and Food Institute. Provo, UT, USA.: 54p. Salganik MJ, Heckathorn DD, 2004. Sampling and estimation in hidden populations using respondent driven sampling. Sociological methodology 34(1): 193-239.

Corresponding Author's Address: poutougnigni2005@yahoo.fr

META-ANALYSIS OF TRANSCRIPTOMIC DATA REVEALS POTENTIAL BIOMARKERS AND THERAPEUTIC TARGETS IN CERVICAL CANCER

Medi Kori, Kazim Yalcin Arga

1. Department of Bioengineering, Marmara University, Istanbul, Turkey

Cervical cancer is the second most common cancer and one of the leading causes of cancer death among women worldwide [1]. Although

infection by "highly oncogenic" Human Papillomavirus (HPV) is essential for cervical cancer development, it alone is not sufficient ; therefore, other cancer related risk factors are required for this disease to develop [2, 3]. Also, despite the presently available screening tests, approximately cervical cancer causes 570.000 new cases and 311.000 deaths and described as a fourth leading cause of death at 2018 [4]. In this context, considering the unclear etiology of cervical cancer and the inaccuracy of present screening methods, systems-level approaches are needed to elucidation of potential biomarkers for the screening, diagnosis and treatment for the disease. Accordingly, in this study, a meta-analysis of the cervical cancer associated transcriptomic datasets was performed by taking into consideration five independent studies and the core differentially expressed genes (DEGs) was obtained by statistical analyses. DEGs were further integrated with genome-scale human biomolecular networks i.e. protein-protein interaction network, genome-scale human metabolic model and transcriptional and post-transcriptional regulatory network. As a result of that, biomolecules were identified at RNA (mRNA, miRNA), protein (receptor, transcription factor, etc.), and metabolite levels. Moreover, survival analyses were performed and the prognostic power of selected reporter biomolecules was identified via cross-validation using independent gene expression datasets. This approach revealed already-known biomarkers, tumor suppressors and oncogenes in cervical cancer as well as various receptors (e.g. ephrin receptors EPHA4, EPHA5, and EPHB2; endothelin receptors EDNRA and EDNRB; nuclear receptors NCOA3, NR2C1, and NR2C2), miRNAs (e.g., miR-192-5p, miR-193b-3p, and miR-215-5p), transcription factors (particularly E2F4, ETS1, and CUTL1), other proteins (e.g., KAT2B, PARP1, CDK1, GSK3B, WNK1, and CRYAB), and metabolites (particularly, arachidonic acids) as novel biomarker candidates and potential therapeutic targets. Consequently, this systematic study reports candidate biomolecules that can be considered as diagnostic/prognostic biomarkers or potential therapeutic targets for further experimental and clinical trials for cervical cancer.

Keywords: Cervical Cancer; Meta-Analysis; Biomarker

References: [1] Tota JE, Chevarie-Davis M, Richardson LA, Devries M, Franco EL. Epidemiology and burden of HPV infection and related diseases: implications for prevention strategies. Prev Med. 2011; 53(1): 12–21.
[2] Haedicke J, Iftner T. Human papillomaviruses and cancer. RadiotherOncol. 2013; 108(3): 397-402.
[3] Agarwal SM, Raghav D, Singh H, Raghava GP. CCDB: a curated

database of genes involved in cervix cancer. Nucl Acids Res. 2011; 39: 975-9.
[4] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov; 68(6):394-424.

Corresponding Author's Address: Medi Kori Department of Bioengineering, Faculty of Engineering, Marmara University, Building D, Kadıkoy, Istanbul, Turkey. E-mail: medi_kori@hotmail.com

COMPUTATIONAL ANALYSIS OF SEQUENCES AND STRUCTURES OF PROTEINS AND NUCLEIC ACIDS INVOLVED IN RESPIRATORY DISEASES

Seda Yüzeren[1], *Cenk Selçuki*[2]

1.Ege University Institute of Health Sciences Health Bioinformatics Program
2.Ege University Faculty of Science Biochemistry Department

Biofilm is an ecosystem of exopolysaccharide-producing bacteria that cause many infectious diseases. In this study, S-ribosylhomocysteine lyase enzymes synthesized from the luxS gene in Quorum Sensing, an important mechanism in biofilm formation of 7 selected microorganisms that play a role in respiratory diseases, and nucleic acids encoding this protein were investigated by molecular modeling and sequence comparison.

The nucleic acid and amino acid sequences of proteins of Bacillus anthracis, Haemophilus influenzae, Klebsiella pneumoniae, Staphylococcus aureus, Streptococcus pneumoniae, Streptococcus pyogenes and Neisseria meningitidis microorganisms were obtained from the NCBI and Uniprot databases.

Alignment was performed to find common nucleic acid and amino acid sequences by using MEGA 7.0.26 free software.

Three dimensional structures of proteins were created in the Swiss-Model/ Expasy database, these models were compared using the free software UCSF Chimera and common structures were found. In binary comparisons, the similarities of proteins were calculated by Root Mean Square Deviation (RMSD) by using UCSF Chimera.

As a result, common sequences and structures were obtained. Common sequences and structures was observed in double and triple comparisons of *H. influenzae- K. pneumoniae, H. influenzae-N. meningitidis, N. meningitidis- K. pneumoniae, S. pyogenes- S. pneumoniae, B. anthracis- S. pyogenes- S. pneumoniae and H. influenzae-K. pneumoniae-N. meningitidis* comparisons.

It is thought that these results can contribute to the production of solutions to prevent biofilm formation.

I would like to thank TÜBİTAK-BİDEB for providing scholarships under the 2210-C program in this study.

Keywords: Biofilm; Quorum Sensing; Respiratory Diseases; S-Ribosylhomocysteine Lyase; Three Dimensional Structures of Proteins.

Corresponding Author's Address: Ege University, Faculty of Science, Biochemistry Department cenk.selcuki@gmail.com

PREDICTION OF TRANSCRIPT PROFILE OF CELLS VIA COMPUTATIONAL METHODS

Gülben AVŞAR[1], Pınar PİR[1]

1. Gebze Technical University, Department of Bioengineering, Çayırova/Kocaeli

Connecting epigenetic modifications to cellular functions can be performed by utilizing the gene expression levels [1]. It is well known that aberrant epigenetic marks are related to several diseases such as cancer, autoimmune diseases, cardiovascular diseases and schizophrenia [2-4]. Several studies showed that epigenetic marks are reversible in contrast to DNA mutations, hence drugs targetting epigenetic profile of cells may have big potential on treatments of the diseases [1]. In addition, use of stem cells for treatment of diseases, which requires investigation of cellular mechanisms with respect to epigenomics, transcriptomics and also proteomics level of stem cells may also have huge impact on current treatment methods [5]. The effects of epigenetic modifications (epigenomic level) such as histone marks on gene regulation can be investigated by evaluating the gene expression levels (transcriptomic level). The computational methods such as machine learning techniques may lead construction of mathematical models based on epigenetic factors and capabale of predicting the gene expression levels in the target cell. Several computational models have been released to predict gene expression from histone modification profiles using support vector machine, random forest, and deep learning methods [6]. However, the aforementioned studies focus on differential gene expression levels that need two conditions for comparison instead of absolute gene expression values. Here we perform analysis of next generation sequenced data on RNA-seq and ChIP-seq analysis results of undifferentiated cells and/or iPSCs. The pipeline includes trimming of low-quality reads, mapping on the reference genome, finally identification of gene expression levels and peak heights. The obtained gene expression levels and peak heights will be used for training the deep neural networks. Convolutional neural networks and recurrent neural networks with several hyperparameters will be created. In convolutional neural networks, the parameters such as filter size, number of filters, pooling size, window size, and repeating time will be optimized. Number of hidden nodes and layers, and addition of different activation functions will be determined for recurrent neural networks. The model will be validates using test from various cell types.

Keywords: Epigenomics, Transcriptomics, Ngs Data Analysis, Deep Neural Networks

References: [1] Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from
histone modifications. Bioinformatics. 2018 Sep 1;34(17):i891-900.
[2] Gluckman PD, Hanson MA, Buklijas T, Low FM, Beedle AS. Epigenetic mechanisms that underpin
metabolic and cardiovascular diseases. Nature Reviews Endocrinology. 2009 Jul;5(7):401.

[3] Urdinguio RG, Sanchez-Mut JV, Esteller M. Epigenetic mechanisms in neurological diseases: genes,
syndromes, and therapies. The Lancet Neurology. 2009 Nov 1;8(11):1056-72.
[4] Ballestar E. Epigenetic alterations in autoimmune rheumatic diseases. Nature Reviews Rheumatology.
2011 May;7(5):263.
[5] Avior Y, Sagi I, Benvenisty N. Pluripotent stem cells in disease modelling and drug discovery. Nature
reviews Molecular cell biology. 2016 Mar;17(3):170.
[6] Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from
histone modifications. Bioinformatics. 2016 Aug 29;32(17):i639-48.

Corresponding Author's Address: Gebze Technical University, Department of Bioengineering, Gebze/KOCAELİ/TURKEY
(pinarpir@gtu.edu.tr)

SINGLE-CELL TRANSCRIPTOME ANALYSIS OF NEURONAL CELL DIFFERENTIATION PROCESSES

Batuhan Çakır[1], *Pınar Pir*[1]

1. Department of Bioengineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkey

Induced pluripotent stem cells (iPSCs) resemble embryonic stem cells and has the potency to differentiate into all types of mature cells. Mature somatic cells can be reprogrammed into iPSCs by inducing high expression of the genes important for maintaining the essential properties of embryonic stem cells (ESCs) iPSC can be used in medical research areas such as disease modelling, toxicity testing or transplantation applications [1]. Their ability to transform into any cell type can lead to development of new treatment therapies and discovery of new drug targets. To understand the reprogramming processes of iPSCs in detail, analysis of high-resolution data is crucial. Gene expression profiles are measured at the cell level instead of bulk samples with single-cell transcriptomics (SCT). Measuring gene expression in single-cell level is critical to better understand the heterogeneity, cellular behaviour and molecular composition in developing, adult and pathological tissues. With time-course single-cell RNAseq data, dynamics of biological systems can be predicted more accurately than ever before. One useful way to gain biological insights from single-cell RNA-seq data is to computationally sort the cells according to the gradual transition of their transcriptomes. Using single-cell RNA-seq data, one may construct an ordered sequence of cells to describe the gradual transition of the single-cell transcriptome [2]. Ordering cells can be done by using trajectory inference (TI) methods (also known as pseudotemporal ordering methods).As shown in Figure 1, trajectory inference plots show the branching of a time-dependent process by scRNAseq data. Analysis of scRNAseq data changes the understanding of diseases or biological processes, and it reveals intracellular heterogeneity within a tissue at very high resolution [3,4]. The aim of this study is the analysis of single cell omics data to investigate and model the transformation of fibroblasts and stem cells into neuron cells, and expected to generate new theories on the progress of these transformation processes.

Keywords: Induced Pluripotent Stem Cell; Rod Photoreceptor Cell; Differentiation; Trajectory Inference; Single-Cell Rnaseq; Single-Cell Transcriptomics; Scrnaseq

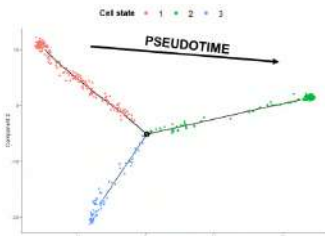


Figure 1. Example of trajectory inference result (produced by Monocle)

References: [1] National Institute of Health. Regenerative Medicine 2006 [Internet]. National Institutes of Health. 2006.
[2] Ji Z, Ji H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. Nucleic Acids Res.
2016;44(13):e117.
[3] Poirion OB, Zhu X, Ching T, Garmire L. Single-cell transcriptomics bioinformatics and computational challenges. Front Genet.
2016;7(SEP).
[4] Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. 2018; Available from: https://www.biorxiv.org/content/biorxiv/early/2018/03/05/276907.full.pdf

Corresponding Author's Address: Pınar Pir Gebze Technical University, Department of Bioengineering,41400, Çayırova/Kocaeli, Turkey pinarpir@gtu.edu.tr

IDENTIFICATION CELL-CELL COMMUNICATION BY RECEPTOR-LIGAND INTERACTIONS IN CANCER

Cemal Yıldız [1], *Pınar Pir* [1], *Devrim Gözüaçık* [2]

1. Gebze Technical University, Department of Bioengineering, Çayırova/Kocaeli
2. Sabancı University, Department of Genetics and Bioengineering, Tuzla/İstanbul

Cancer is mainly caused by mutations in single somatic cells. The mutations cause abnormality in cell growth, proliferation, migration that leads to tumor formation and eventually secondary tumors (metastasis) which are composed of cells with mutation burden different from their origin. There are several cancer types and all of them have common progression profiles such as single somatic mutations, complex interaction between cells and extracellular matrix (ECM) components in the host tissue. Cancerous cells communicate with their microenvironment and their growth, their survival and metastasis can be supported via proteins that are present in their microenvironment. [1]

Tumors are also heterogeneous because of their environment, a sample that taken from patient includes more than one cell type such as B cells, T cells, Endothelial cells, Macrophages, Fibroblast cells. Single-cell RNA sequencing, can reveal complex cell populations and uncover regulatory relationships between cells. [2] Cells communicate each other by receptor-ligand interactions in tumor microenvironment that include extracellular matrix components and other cells. Identifying interactions between cell types could lead to right treatment an individual patient. [3]

In this study, cell-cell communication is identified in cancer tissues and results are associated with signaling pathways. Transcriptome data is taken from GEO and analyzed with bioinformatics software packages developed in MATLAB. After identifying receptor-ligand interactions between detected cell types, active signaling pathways are found and tumor characteristics are predicted from related receptor-ligand interactions.

Keywords: Single-Cell Rna Sequencing, Cell-Cell Communication, Receptor-Ligand Interactions

References: [1] A. Masoudi-Nejad, G. Bidkhori, S. Hosseini Ashtiani, A. Najafi, J. H. Bozorgmehr, and E. Wang, "Cancer systems biology and modeling: Microscopic scale and multiscale approaches," Semin. Cancer Biol., 2015, Feb. vol. 30, pp. 60–69. Pubmed PMID: 24657638
[2] B. Hwang, J. H. Lee, and D. Bang, "Single-cell RNA sequencing technologies and bioinformatics pipelines," Exp. Mol. Med., 2018, Aug. vol. 50, no. 8, p. 96. Pubmed PMID: 30089861
[3] M. P. Kumar et al., "Analysis of Single-Cell RNA-Seq Identifies Cell-Cell Communication Associated with Tumor Characteristics.," Cell Rep., 2018, Nov. vol. 25, no. 6, p. 1458–1468.e4. Pubmed PMID: 30404002

Corresponding Author's Address: Gebze Technical University, Department of Bioengineering, Room 210. Email: cyildiz@gtu.edu.tr

RECONSTRUCTION OF A GENOME-SCALE MODEL OF EUBACTERIUM LIMOSUM (ATCC 8486) BY INTEGRATION OF TRANSCRIPTOME DATA AND A KIST612 MODEL

Simge Sengul[1], *Emre Taylan Duman*[2], *Pınar Pir*[2]

1. Gebze Technical University, Department of Molecular Biology and Genetics
2. Gebze Technical University, Department of Bioengineering

Eubacterium limosum is an acetogenic bacteria which is a species found in gut microbiata. E. limosum was shown to have theraputic probiotic activity such as reducing inflammation in colitis[1]. Acetogens also have industrial importance as they can convert syngas to produce biomass acetic acid by utilizing in-air carbon sources and Hydrogen [2]. These carbon sources are taken to the metabolism through Wood-Ljungdahl pathway which is the main metabolical steps of this conversion process [3]. Systems Biology approach is used in generating genome-scale metabolic models to simulate whole metabolism or some sub-metabolic processes in certain culture conditions [4]. Draft models are constructed based on genomic information, then they nee to be tuned by using experimental data and manual curation prior to simulations. In this study, Genome Scale Metabolic Model (GSMM) of E. limosum KIST612 generated automatically from genome data and KEGG pathways by BioModels Database is used as a template for E.limosum ATCC 8486 model reconstruction. Integration of ATCC 8486 transcriptome data (GSE97613) from GEO Database and a draft metabolic model of KIST612 by Thiele et.al [5] will allow us to find missing reactions and metabolites in the model to simulate potential syngas fixation and acetic acid production in addition to its probiotic activity.

Keywords: Genome Scale Metabolic Model; Eubacterium Limosum; Acetogen; Gut Microbiata; Flux Balance Analysis.

References: [1] Kanauchi, Osamu, et al. Eubacterium limosum ameliorates experimental colitis and metabolite of microbe attenuates colonic inflammatory action with increase of mucosal integrity. World journal of gastroenterology: WJG, 2006, 12.7: 1071.
[2] Roh, Hanseong, et al. Complete genome sequence of a carbon monoxide-utilizing acetogen, Eubacterium limosum KIST612. Journal of bacteriology, 2011, 193.1: 307-308.
[3] Ragsdale, Stephen W.; PIERCE, Elizabeth. Acetogenesis and the Wood–Ljungdahl pathway of CO2 fixation. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 2008, 1784.12: 1873-1898.
[4] Kitano, Hiroaki. Systems biology: a brief overview. science, 2002, 295.5560: 1662-1664.
[5] Magnúsdóttir, Stefanía, et al. Generation of genome-scale metabolic reconstructions for 773 members of the human gut microbiota. Nature biotechnology, 2017, 35.1: 81.

Corresponding Author's Address: Gebze Technical University Bioengineering 41070 Gebze/Kocaeli Turkey pinarpir@gtu.edu.tr

IDENTIFICATION OF TF-MEDIATED DYNAMICS OF SIGNALLING PATHWAYS IN STEM CELL REPROGRAMMING AND DIFFERENTIATION USING RNA-SEQ DATA

Enes Ak[1], Batuhan Çakır[1], Hamza Umut Karakurt[1], Furkan Kurtoğlu[2], Şükrü Öztürk[3], Gülben Avşar[1], Pınar Pir[1]

1. Department of Bioengineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkey
2. School of Informatics, Computing and Engineering, Indiana University, Bloomington, IN, USA
3. Department of Pharmaceutical Basic Sciences, Faculty of Pharmacy, Hacettepe University, Ankara, Turkey

Use of induced pluripotent stem cells (iPSCs) holds great promise in medicine, especially regenerative and personalized medicine [1]. iPSCs originate from somatic cells which are reprogrammed into pluripotency via directed methods such as expression of specific transcription factors [2]. iPSC-associated technologies provide tools to understand disease mechanisms, discover drug contents and develop new therapeutic interventions [3]. Reprogramming of iPSC from somatic cells and differentiation of somatic cells from iPSC takes place by activation of reprogramming mechanisms and their signaling pathways [4]. Expression or repression of transcription factors (TFs) affect these regulatory networks. The identification of transcription factors affecting signaling pathways is an important step in revealing cell differentiation mechanisms, and further insights about cell differentiation mechanisms can be obtained by studying the time-dependent activities of these pathways. Time dependent change in the epigenetic marks and expression of the TFs can be profiled using omics technologies such as RNAseq and ChipSeq. In this study two transcriptome (RNAseq) datasets from GEO database was used, dataset GSE86790 (Chen et al.) [5] was used for profiling the differentiation Mus musculus retinal organoids from iPSCs and dataset GSE87064 (Aldiri et al.) [6] was used for profiling to the reprogramming of iPSCs from Mus musculus retinal organoids. Main target of this study is comparison of the differentiation and reprogramming processes of cells and determine TFs and their expression trends which affect the regulatory networks during differentiation process using RNAseq data. We aim to propose new strategies for efficient cell differentiation and reprogramming processes, in addition to shed light on dysregulation of these processes in health and disease states.

Keywords: Induced Pluripotent Stem Cell,Rod Photoreceptor Cell,Transcription Factor,Signalling Pathway, RNAseq

References:[1] Attwood, S., & Edel, M. (2019). iPS-Cell Technology and the Problem of Genetic Instability—Can It Ever Be Safe for Clinical Use? Journal of Clinical Medicine, 8(3), 288. https://doi.org/10.3390/jcm8030288
[2] Gomes, K. M. S., Costa, I. C., Santos, J. F. dos, Dourado, P. M. M., Forni, M. F., & Ferreira, J. C. B. (2017). Induced pluripotent stem cells

reprogramming: Epigenetics and applications in the regenerative medicine. Revista Da Associação Médica Brasileira, 63(2), 180–189. https://doi.org/10.1590/1806-9282.63.02.180
[3] Mertens, J., Marchetto, M. C., Bardy, C., & Gage, F. H. (2016). Evaluating cell reprogramming, differentiation and conversion technologies in neuroscience. Nature Reviews Neuroscience, 17(7), 424–437. https://doi.org/10.1038/nrn.2016.46
[4] Lu, X., Chen, X., Xing, J., Lian, M., Huang, D., Lu, Y., ... Feng, X. (2019). miR-140-5p regulates the odontoblastic differentiation of dental pulp stem cells via the Wnt1 / β -catenin signaling pathway, 4–11.
[5] Chen, H. Y., Kaya, K. D., Dong, L., & Swaroop, A. (2016). Three-dimensional retinal organoids from mouse pluripotent stem cells mimic in vivo development with enhanced stratification and rod photoreceptor differentiation. Molecular Vision, 22(May), 1077–1094. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5017542/
[6] Aldiri, I., Xu, B., Wang, L., Chen, X., Hiler, D., Griffiths, L., ... Dyer, M. A. (2017). The DynamicEpigenetic Landscape of the Retina During Development, Reprogramming, and Tumorigenesis. Neuron, 94(3), 550-568.e10. https://doi.org/10.1016/j.neuron.2017.04.022

Corresponding Author's Address: Pınar PİR Gebze Technical University, Department of Bioengineering,41400, Çayırova/Kocaeli, Turkey E-Mail: pinarpir@gtu.edu.tr

METABOLIC EFFECTS OF BIPOLAR DISORDER ON DORSOLATERAL PREFRONTAL CORTEX: A GENOME-SCALE METABOLIC MODEL APPROACH

Hamza Umut Karakurt[1], *Pınar Pir*[1]

1. Gebze Technical University, Department of Bioengineering, Kocaeli, Turkey

Bipolar Disorder, formerly known as manic depression, is a common, disabling and recurrent major psychiatric condition. Bipolar disorder causes mental periods of depression and elevated mood which is also called as mania or hypomania [1]. Frequent change of moods causes disturbances on daily lives. Attemp to suicide is one of the frequent results in bipolar disorder results. At least 25% to 60% of the patients attemp suicide at least once in their lifetime. 4% to 19% or patients complete suicide [2]. Key to diagnose bipolar disorder is the history or presence of mania or hypomania. Some psychotic symptoms can be mistaken for schizophrenia. The overlap between other disorders and examination without previous health record of the patient may lead to missed diagnosis of the condition.

The exact mechanism of bipolar disorder is still unclear. Previous studies suggested that specific polymorphisms [3], alterations in Wnt and Notch signalling [4,5], glucose and phosphorus metabolisms [6,7] and certain brain-related metabolites such as dopamin [8] and glutamate [9] are associated with bipolar disorder.

In this study, transcriptome data from bipolar disorder patients and healthy samples combined with brain specific genome-scale metabolic model to identify the reporter metabolites, reporter pathways and highly altered biochemical fluxes in the dorsolateral prefrontal cortex. Due to fact that brain functions via neurotransmitters, which are essentially metabolites produced by the metabolism, metabolic modelling approaches in mood disorders such as bipolar disorder are promising.

Keywords: Bipolar Disorder; Transcriptomics; Genome-Scale Metabolic Models

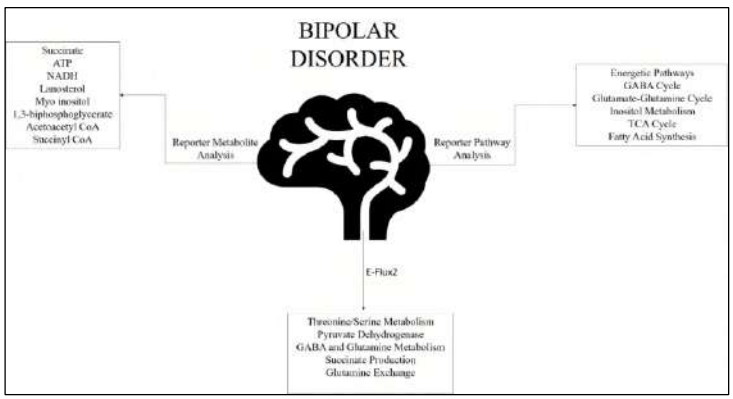


Figure 1. Foreseen Metabolic Alterations in Bipolar Disorder in Genome-Scale Metabolic Model Approach

References: [1] Anderson, I. M., Haddad, P. M. & Scott, J. Bipolar disorder. Bmj 345, e8508–e8508 (2012).
[2] Novick, D. M., Swartz, H. A. & Frank, E. Suicide attempts in bipolar I and bipolar II disorder: a review and meta-analysis of the evidence. 12, 1–9 (2015).
[3] Gao, J. et al. TPH2 gene polymorphisms and bipolar disorder: A meta-analysis. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. 171, 145–152 (2016).
[4] Pedroso, I. et al. Common genetic variants and gene-expression changes associated with bipolar disorder are over-represented in brain signaling pathway genes. Biol. Psychiatry 72, 311–317 (2012).
[5] Van Meter, A. R., Moreira, A. L. R. & Youngstrom, E. A. Meta-Analysis of Epidemiologic Studies of Pediatric Bipolar Disorder. J. Clin. Psychiatry 72, 1250–1256 (2011).
[6] Hosokawa, T., Momose, T. & Kasai, K. Brain glucose metabolism difference between bipolar and unipolar mood disorders in depressed and euthymic states. Prog. Neuropsychopharmacol. Biol. Psychiatry 33, 243–250 (2009).
[7] Kato, T., Takahashi, S., Shioiri, T. & Inubushi, T. Alterations in brain phosphorous metabolism in bipolar disorder detected by in vivo 31P and 7Li magnetic resonance spectroscopy. J. Affect. Disord. 27, 53–59 (1993).
[8] Salvadore, G. et al. The neurobiology of the switch process in bipolar disorder: A review. Journal of Clinical Psychiatry 71, 1488–1501 (2010).
[9] Michael, N. et al. Acute mania is accompanied by elevated glutamate/glutamine levels within the left dorsolateral prefrontal cortex. Psychopharmacology (Berl). 168, 344–346 (2003).

Corresponding Author's Address: Gebze Technical University, Department of Bioengineering, Kocaeli/Turkey hkarakurt@gtu.edu.tr

IDENTIFICATION OF TRANSCRIPTIONAL REGULATORY NETWORKS CHARACTERIZING INVASIVE AND NON-INVASIVE BLADDER CANCER CELL LINES

Perihan Yağmur Güneri1, Gülden Özden Yilmaz2, Aleyna Eray1, Neşe Atabey2, Şerif Şentürk[1,2], Serap Erkek[2]

1.Izmir Biomedicine and Genome Institute, Izmir, Turkey
2.Izmir Biomedicine and Genome Center, Izmir, Turkey

It is known that epigenetic regulations lead to changes in gene expression programs thus affecting cell fate. Mutations in chromatin modifier genes can readjust chromatin states[1]. Mutation rate in chromatin modifier genes in bladder cancer is especially high (~20-25%) compared to many other cancer types. Therefore, bladder cancer is one of the major cancer types for which chromatin regulatory mechanisms should be studied. In this study, we investigate the transcriptional regulatory networks differentially regulated between muscle invasive bladder cancer (MIBC) and non-muscle invasive bladder cancer (NMIBC) cell lines .via identifying the active chromatin regions and transcription factors and their interacting partners specific to the respective cell lines. We use H3K27ac chromatin immunoprecipitation followed by sequencing (ChIP-seq) to define active regulatory regions in 3 non-muscle invasive bladder cancer cell lines (RT4, RT112 and 5637) and 3 muscle invasive bladder cancer cell lines (J82, T24 and HT1376). Until now, we have completed ChIP-seq experiments of two MIBC cell lines (J82 and T24) and one NMIBC cell line (RT4) and we called H3K27ac peaks on these cell lines. Our preliminary findings suggest differential H3K27ac occupancy in MIBC and NMIBC around promoter regions of several genes such as CDH1 and VIM, which might be involved in invasive characteristics. After completing the ChIP-seq experiments for other three cell lines and extending our computational analysis, we will identify which transcription factors have roles in respective regulatory regions and which chromatin modifiers are associated with these transcription factors by performing the assay "ChIP Selective Isolation of Chromatin Associated Proteins (ChIP-SICAP)". Our findings will provide important insights about transcriptional regulatory pathways implicated in invasive characteristics of bladder cancer.

Keywords: Epigenetics, BLCA, ChIP-Seq, Invasion

References: [1] Flavahan WA, Gaskell E, Bernstein BE. Epigenetic plasticity and the hallmarks of cancer. Science. 2017 Jul 21;357(6348). doi: 10.1126/science.aal2380.

Corresponding Author's Address: serap.erkek@ibg.edu.tr

PHYLOGENETIC ANALYSIS OF KERATINS: FUNCTIONAL IMPLICATIONS

Işil Takan[1,2], Hani Alotaibi[1,2], Athanasia Pavlopoulou[1,2]

1. Izmir Biomedicine and Genome Center, 35340 Balcova, Izmir, Turkey,
2. Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, 35340 Balcova, Izmir,Turkey

Keratins (KRTs) constitute the intermediate filament-forming proteins of epithelial cells. They are also grouped as "soft" and "hard" keratins, on the basis of their physicochemical properties. Hard keratins make up morphological structures such as scales, claws and feathers in birds and hair and nails in mammals. Soft bepithelial keratins are highly involved in epithelial cell protection from mechanical and non-mechanical stressors, and also play a regulatory role in apical–basal plasma membrane polarity, cell size, cell motility, protein synthesis, membrane trafficking and signaling pathways that regulate cell response to wound healing, cell growth and cell death [1, 2]. Because of these properties soft keratins are known to have a prominent role in several aspects of cancer pathophysiology, including cancer cell invasion and metastasis, and several members of the KRT family serve as diagnostic or prognostic markers [3]. However, there is a lack of a comprehensive phylogenetic analysis of keratins which would enhance our understanding regarding the functional implications of keratins in cancer biology among tumor-bearing species. The human genome contains both functional KRT genes and nonfunctional KRT pseudogenes which are arranged in two uninterrupted clusters on chromosomes 12 and 17 [4]. This exceptional characteristic of KRTs makes them ideal for evolutionary studies aiming to infer the direction and timing of gene duplication events. In order to reconstruct the evolutionary history of the type I KRT gene family, we have performed comprehensive phylogenetic analyses of KRT homologous proteins in multiple genomes. Toward this end, the publicly available genomes, including those that have been completed recently, were searched extensively for keratin homologs. Phylogenetic trees based on the entire length keratin homologous proteins were reconstructed. Furthermore, the chromosomal arrangement of KRT genes was examined, and it was found that their position, transcriptional orientation, and number is preserved in the species under investigation.

Keywords: Keratins; Phylogeny; Gene Duplication

References: [1] Coulombe PA, Omary MB. 'Hard' and 'soft' principles defining the structure, function and regulation of keratin intermediate filaments. Curr Opin Cell Biol. 2002 Feb;14(1):110-22. PubMed PMID: 11792552.
[2] Moll R, Divo M, Langbein L. The human keratins: biology and pathology. Histochem Cell Biol. 2008 Jun;129(6):705-33. PubMed PMID: 18461349.
[3] Karantza V. Keratins in health and cancer: more than mere epithelial cell markers. Oncogene. 2011 Jan 13;30(2):127-38. PubMed PMID: 20890307.
[4] Hesse M, Zimek A, Weber K, Magin TM. Comprehensive analysis

of keratin gene clusters in humans and rodents. Eur J Cell Biol. 2004 Feb;83(1):19-26. PubMed PMID: 15085952.

Corresponding Author's Address: athanasia.pavlopoulou@ibg.edu.tr

THE USE OF MACHINE LEARNING AND DEEP LEARNING ON SNP DATA TO PREDICT THE CASE-CONTROL STATUS OF INDIVIDUALS IN THE CHILDHOOD LEUKEMIA DISEASE

IR. Onur ÖZTORNACI[1], Bahar TAŞDELEN[1], T. Mehmet DORAK[2], Erdal COŞGUN[3]

1. Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Mersin
2. School of Life Sciences, Pharmacy and Chemistry, Kingston University London, UK,
3.Microsoft Research, 14820 NE 36th Street, Building 99, Redmond, WA, 98052, USA

Using machine learning (ML) and deep learning (DL) approaches may have as alternatives for the genome wide association studies (GWAS) that detecting to SNPs with associated to the diseases. GWAS methodology is used with a large amount of data, this case may have a problem for researchers in terms of using for high thresholds. On the other hand, ML and DL don't require any assumption such as normality and also false discovery rate corrections. Besides, a large amount of data may take an advantage for using ML and DL because of heterogeneous patterns as well as this situation allows to comparison for the best model between ML and DL on SNP data. Nevertheless, computation time is one of the problems for ML algorithms. Basically, two different methods can be used on the preparing the data for using ML methods and DL as well. i) Using clumping on PLINK, ii) Using Chi-square for feature selection. We would like to propose a feature selection approach for handling this problem. The method depends on minor allele frequency; through this method, it can combine two methods and it may allow to keep the SNPs that may be associated with the disease when using ML methods and DL. Minor alleles can be given clue about the complex diseases [1]. We re-analysed an already published GWAS data that study designed to identify genetic markers of gender-specific differences in childhood [2]. In this study, analysis was performed by the data which was passed QC procedure. In case-control studies in the GWAS, Logistic regression is used frequently to identify to SNPs with associated to diseases in GWAS. The results are shown that SVM is the best method with a high accuracy rate as well as high sensitivity and specificity results. RF has reached %73 accurate classifying. However, RF shown a low specificity %43 which means using RF may not giving a good solution for choosing non-disease individuals on our GWAS data. The accuracy value of MLP is very low, thence; it may not appropriate for using GWAS data. DL was able to reach %65 accuracy rate and also it shown that quite similar PPV to RF. Besides, it can be said that DL has reported the results in a very short time (7 Min). SVM appeared to be the most accurate method (PPV=1 and NPV=0.96, sensitivity=0.97, specificity=1, Accuracy=0.98) (Table1). Our results concordant with the consensus view in the field that SVM is probably best suited for GWAS analysis as an alternative to conventional methods. We used 10-fold cross-validation method to avoid for bias effect and over-fitting [3].

Keywords: Genome-Wide Association Studies, Machine Learning, Deep Learning.

Methods	*PPV*	*NPV*	*Sensivity*	*Specificity*	*Accuracy [Conf. Interval]*	*Cpu Time (Average)*
Support Vector Machine RK*	**1**	**0.968**	**0.973**	**1**	**0.985 [0.92-0.99]**	**14 Min.**
Random Forest	**0.678**	**1**	**1**	**0.419**	**0.739 [0.61-0.83]**	**42 Min.**
Multi-Layer Perceptron	**0.550**	**--**	**1**	**0.000**	**0.550 [0.42-0.67]**	**34 Min.**
Deep Learning-CNTK	**0.616**	**0.888**	**0.973**	**0.258**	**0.652 [0.527-0.762]**	**7 Min.**

Table1: The performances of Machine Learning Methods with a-priori feature selection

References:[1] Kido, T., Sikora-Wohlfeld, W., Kawashima, M., Kikuchi, S., Kamatani, N., Patwardhan, A., ... & Butte, A. J. (2018). Are minor alleles more likely to be risk alleles?. BMC medical genomics, 11(1),3.
[2] Singh, S. K., Lupo, P. J., Scheurer, M. E., Saxena, A., Kennedy, A. E., Ibrahimou, B., ... & Dorak, M. T. (2016)
[3] Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

Corresponding Author's Address:

R. Onur Öztornacı, onur.oztornaci@gmail.com

GENE ENRICHMENT ANALYSIS TO ELUCIDATE THE ROLE OF MITOCHONDRIAL GENES IN PROSTATE CANCER

*Mete Emir Özgürses[1], Ahmet Melih Öten[2], Öykü İrigül Sönmez[3], Aslı Yenenler[2]**

1. Department of Molecular Biology and Genetics, Faculty of Engineering and Natural Sciences, Biruni University, 2. Department of Biomedical Engineering, Faculty of Engineering and Natural Sciences, Biruni University, 3. AYA R&D Biotechnology Inc.

Abstract

Mitochondria is a double-membrane organelle and plays an important role in cell metabolism, calcium balance, oxidation-reduction reactions and cell fate determination [1]. The metabolism of each cell differs from each other and this difference is quite significant between normal and tumor cells [2]. Cancer cells, metabolic activity differs from normal cells because most cancer cells lose their mitochondrial function due to oncogenic signals and mitochondrial DNA damage increases glycolysis dependent energy production and increase lactate formation in cells, which is referred as Warburg Effect, as a final product [3]. Due to the lactate production, generated hypoxic condition increases the reactive oxygen species (ROS) production. According to the literature, ROS is involved in stabilizing the Hypoxia Inducing Factor (HIF1-alpha) in adapting tumors to extreme hypoxic conditions. Prostate cancer a highly progressive cancer with a 1 million new cases every year. Even though it effects vast majority of the population and has a high mortality rate, the initiation of prostate cancer is not clearly known but recent studies have shown that proliferative inflammatory atrophy (PIA) and prostatic intraepithelial neoplasia (PIN) may be the initiators of prostate cancer [4]. To diagnose prostate cancer, prostate specific antigen level is generally considered. However, it has been found that the PSA can also be increased by body weight and daily carbohydrate intake. These properties limit the early diagnosis of prostate cancer and increases the mortality (5]. When the molecular background of prostate cancer is studied, it has been found that the receptors for steroid hormones and hormonal growth factors is highly similar with breast cancer and the tumors generated from these organs are hormone dependent and have biological similarities [6]. These similarities may be used to generate a prognostic test which applies both for prostate and breast cancer. There have been different types of prognosis kits in the world and one of them is a PCR based test for 50 gene (PAM50 signature) which helps to identify 4 different subtypes of breast cancers. Since the similarities of two different cancers are highly common, enriching the PAM50 breast cancer prognose kit can be valid to use in both cancers. In our project, due to the ROS ability of both effecting the mitochondrial dynamics and hypoxic conditions in a cell and the similarities between breast and prostate cancer, understanding the relationship with mitochondrial events with cancer metabolism and stress conditions will enable us to better understand how high energy is sustained in cancer cell and the enriched genes will enable us to enhance the power of diagnosis of

PAM50 prognose kit for prostate cancer patients. Before developing a prognose kit, as a further work, we are going to check the expression level of critical genes with nanostring nCounter technology from dissected prostate cancer and reciprocal healthy tissue as in-patient control. According to the PPI network analysis from STRING database, hub genes between the mitochondrial genes and PAM50 genes and interaction scores has been identified.

Keywords: Mitochondrial Dynamics; Prostate Cancer; Gene Enrichment

References: 1. Wallace DC. Mitochondria and cancer. Nat Rev Cancer. 2012 Oct;12(10):685-98. doi: 10.1038/nrc3365. PubMed PMID: 23001348
2. Annibaldi, A., Widmann, C. Glucose metabolism in cancer cells Curr Opin Clin Nutr Metab Care. 2010. 13 (4), 466-470. PMID:
3. Wen S, Zhu D, Huang P. Targeting cancer cell mitochondria as a therapeutic approach. Future Med Chem. 2013 Jan;5(1):53-67. doi: 10.4155/fmc.12.190. PubMed PMID: 23256813
4. Vincent AE, Turnbull DM, Eisner V, Hajnóczky G, Picard M. Mitochondrial Nanotunnels. Trends Cell Biol. 2017 Nov;27(11):787-799. doi: 10.1016/j.tcb.2017.08.009. Epub 2017 Sep 19. PubMed PMID: 28935166
5. Balacescu O, Petrut B, Tudoran O, et al. Urinary microRNAs for prostate cancer diagnosis, prognosis, and treatment response: are we there yet? Wiley Interdiscip Rev RNA. 2017. doi:10.1002/wrna.1438
6. Risbridger, G. P., Davis, I. D., Birrell, S. N., & Tilley, W. D. (2010). Breast and prostate cancer: more similar than different. Nature Reviews Cancer, 10(3), 205–212. doi:10.1038/nrc2795

Corresponding Author's Address:
Aslı Yenenler ayenenler@biruni.edu.tr Biomedical Engineering, Faculty of Engineering and Natural Sciences, Biruni University

COMPARISON OF ANTIGLYCATING PROPERTIES OF PHYTOCHEMICALS ON HEMOGLOBIN USING IN-SILICO APPROACH

Saba Shaikh Amir[1], Aparna Patil[1], Ahmad Ali[2]

1. Dept. of Bioinformatics, GNIRD, G.N. Khalsa College, Matunga, Mumbai- 400019, India
2. Department of Life Sciennces, University of Mumbai, Vidyanagari, Santacruz (East), Mumbai 400098, India

Glycation is a noneznymatic process of interaction between the carbonyl group of sugars and amino groups of proteins. This interaction leads to generation of a group of harmful compounds collectively called as advanced glycation end products (AGEs). The glycating potential of sugars vary according to their interaction, binding affinity and concentrations. Lysine and arginine are two the most important amino acids in the protein structure to which carbonyl groups of the sugars interact. The amount of glycation products generated depends on the number of these basic amino acids and their accessibility. In the present study the binding sites of different glycating and antiglycating agents on Hemoglobin were analyzed with the help of molecular docking approach. Growmacs was used to optimize the protein structure and Autodock was used for docking studies. Phytochemicals and artificial sweeteners showed higher binding affinity than the sugars. These results indicate that among phytochemicals Ferulic acid had the highest binding affinity and most common interacting residues with Hemoglobin. Other phenolic acids were also had high binding affinity. It can be concluded that molecular docking tools can be used for identifying antiglycating agents.

Keywords: Glycation, Glycated Hemoglobin, Molecular Docking, Phytochemicals,

References: [1] Ali, A., More, T., Hoonjan, A. K., Sivakami, S. 2017. Antiglycating potential of Acesulfame potassium: An artificial sweetener. Applied Physiology, Nutrition and Metabolism. 42(10):1054-1063. http://dx.doi.org/10.1139/apnm-2017-0119.
[2] Morris, G. M., & Lim-Wilby, M. (2008). Molecular Docking. Molecular Modeling of Proteins, 365–382. doi:10.1007/978-1-59745-177-2-19

Corresponding Author's Address: Department of Life Sciennces, University of Mumbai, Vidyanagari, Santacruz (East), Mumbai 400 098, India

DRUG REPOSITIONING AND DOCKING STUDIES ON MELANOMA

Nevin Betul Imat[1], *Beste Turanli*[2], *Saliha Ece Acuner-Ozbabacan*[2]

1. Institute of Graduate Studies, Nanosicence and Nanoengineering Program, Istanbul Medeniyet University, Istanbul 34700, Turkey,
2. Department of Bioengineering, Istanbul Medeniyet University, Istanbul 34700, Turkey

There is a rapid increase in the melanoma incidences globally with high mortality rates and the challenges in the treatment makes development of efficient drug discovery methods crucial [1]. Drug repositioning is an innovative approach providing new usage areas for existing drugs. We focused on melanoma and used a gene expression-guided search tool, geneXpharma [2], which provides gene expressions and their drug interactions. After determining the significant gene-drug pairs in melanoma, we categorized them into four subgroups based on their relevance to cancer; namely, already being used in cancer treatment; being used as complementary in cancer treatment; not related to cancer treatment yet; and random gene-drug pairs. Lastly, we performed docking studies for the proteins encoded by the listed significant genes and their ligand drugs and analyzed the structural details for important representative case studies in detail. This study reveals the importance of using a combinatorial approach including drug repositioning and docking and the ability of computational tools to help in the optimization of drug development efforts.

Keywords: Drug Repositioning, Docking, Melanoma

References: [1] Khosravi A, Jayaram B, Goliaei B et al. Active repurposing of drug candidates for melanoma based on GWAS, PheWAS and a wide range of omics data. Molecular Medicine. 2019; 25:30. doi: 10.1186/s10020-019-0098-x
[2] Turanli B, Gulfidan G, Arga KY. Transcriptomic-Guided Drug Repositioning Supported by a New Bioinformatics Search Tool: geneXpharma. Omics : a journal of integrative biology. 2017;21(10):584-91. doi: 10.1089/omi.2017.0127. PubMed PMID: 29049014

Corresponding Author's Address: Department of Bioengineering, Istanbul Medeniyet University, Istanbul 34700, Turkey, E-mail:ece.ozbabacan@medeniyet.edu.tr

AN IN SILICO DETECTION FOR ANTIGENIC POTENTIAL OF TOXOPLASMA GONDII APICOPLAST PROTEINS

Hüseyin Can[1], Sedef Erkunt Alak[1], Ahmet Efe Köseoğlu[1], Mert Döşkaya[2], Cemal Ün[1]

1. Ege University Faculty of Science Department of Biology, Molecular Biology Section, 35040-İzmir
2. Ege University Faculty of Medicine Department of Parasitology, 35100-İzmir

Toxoplasma gondii, causative agent of the disease known as toxoplasmosis, is an obligate intracellular apicomplexan parasite. Toxoplasmosis is acquired by ingesting the food or water contaminated by sporulated oocysts or by consuming raw or undercooked meat containing bradyzoites. During prenatal period, toxoplasmosis can also be acquired by transplacental infection. In healthy individuals, toxoplasmosis generally does not lead to serious clinical symptoms but, in immuno-compromised patients such as those with acquired immuno-deficiency syndrome, immunosuppressed cancer patients and transplant recipients, it leads to life-threatening clinical manifestations. In terms of veterinary medicine, toxoplasmosis is a common cause of abortion in livestock such as sheep and goat, and leads to considerable economic losses worldwide. To date, wide range of nuclear genome proteins such as dense granule protein GRA1, rhoptry protein ROP2, heat shock protein BAG1 and microneme protein MIC1, have been used in development of vaccine against T. gondii infection but, an effective vaccine antigen has not been discovered yet. Apart from its nuclear genome, T. gondii contains an apicoplast genome in 35 kb length which is originated from a secondary endosymbiotic event and the apicoplast genome includes both large and small-subunit rRNAs genes, RNA polymerase genes, the elongation factor tufA gene, tRNA (transfer RNA) genes, a putative ribosomal protein gene operon, five ORFs of unknown function, sufB gene (formerly named ycf24), and ClpC gene. In this study, we aimed to investigate the antigenic potential of apicoplast genome encoded proteins (n:28) of T. gondii using in silico analysis. For this purpose, proteins were primarily predicted to reveal antigenic probability and then, several bioinformatics analyses were applied for all predicted antigenic apicoplast proteins to analyze physico-chemical parameters, subcellular localization, and transmembrane domain. Antigenic proteins that have a signal peptide or a high antigenicity value, were further analyzed for structural conformation, B cell and T cell (MHC-I/II) epitope sites as well as post-translational modifications motifs. Among the 28 apicoplast proteins, 19 were predicted as probable antigen. Among antigenic proteins, ribosomal protein S5, L11 and S2 were predicted to have signal peptide whereas ribosomal protein L36 and S17 were predicted to have a significantly high antigenicity value ($P<0.05$). In addition, ribosomal protein S5, L11, S2, L36, and S17 were predicted to have a lot of epitopes which have low IC50 and percentile rank value indicating a strong binding among epitopes and MHC-I/II alleles, and post-translational modifications such as N-linked glycosylation,

acetylation and phosphorylation. To the best of authors' knowledge, this is the first study to show the antigenic potential and other properties of apicoplast-derived proteins of T. gondii.

Keywords : T. Gondii; Antigen; Apicoplast; Epitope; Ribosomal Protein

Corresponding Author's Address: erkuntsedef@gmail.com

INVESTIGATION OF POLYMORPHISMS ON OMPB GENE OF RICKETTSIA AESCHLIMANNII STRAINS ISOLATED FROM TICK SAMPLES IN TURKEY AND THEIR EFFECTS ON OMPB PROTEIN

Ahmet Efe Köseoğlu[1], Hüseyin Can[1], Sedef Erkunt Alak[1], Samiye Demir[2], Cemal Ün[1]

1. Ege University Faculty of Science Department of Biology, Molecular Biology Section, 35040-İzmir
2. Ege University Faculty of Science Department of Biology, Zoology Section, 35040-İzmir

Rickettsia causing Rickettsiosis is an obligate intracellular bacterium and transmitted by mainly arthropod vectors such as tick. In Rickettsia strains, OmpB (outer-membrane protein B) is a highly conserved gene and plays remarkable roles in bacterial pathogenesis. Also, the gene is used in diagnosis and strain identification. In this study, it was aimed to investigate the polymorphisms on OmpB gene of Rickettsia aeschlimannii strains isolated from tick samples in Turkey and the effects of detected polymorphisms on OmpB protein. For this aims, OmpB gene was amplified from three different Rickettsia strains detected in tick samples in Turkey using PCR methods and sequenced by ABI3730XL. Later, generated sequences were aligned by MEGA7 software and analyzed for nucleotide sequence identity by comparing them with reference strains in National Center for Biotechnology Information (NCBI) and translated into protein sequence. Thereafter, effects of detected polymorphisms on antigenicity, stability, solubility, surface accessibility, allergenicity, B and T cell epitope regions, post-translational modifications, secondary and tertiary structure of OmpB protein were predicted by several bioinformatics tools. As a result, total of 10 single nucleotide polymorphisms in ompB gene of three Ri. aeschlimannii strains were found when compared to reference strain with accession number, AF123705.1. Among 10 nucleotide variations, codon alteration caused amino acid changes at 7 positions. For remaining three variations, though codon alteration, no amino acid change revealed. Obtained NCBI data showed that, 2 of the SNPs were detected for the first time in this study when compared to 42 reference strains of Ri. aeschlimannii. According to in silico analysis, it was found that polymorphisms on OmpB gene increased to antigenicity, changed the physico-chemical parameters associated with stability and solubility and provide the increase in number of random coil and the decrease in number of alpha helices. It was detected that B cell and T cell epitope regions were altered by polymorphisms. There were several alterations in post translational modification regions such as N/O- glycosylation and phosphorylation sites, except methylation and acetylation sites. Overall, the presence of polymorphisms on OmpB gene causes a lot of changes in characteristics of protein.

Keywords : R. Aeschlimannii; Ompb; Polymorphism; Antigen; Epitope; Post-Translational Modification

Corresponding Author's Address: ahmetefekoseoglu@gmail.com

PHYLOMAF: NEXT GENERATION PHYLOGENETIC MICROBIOME ANALYSIS FRAMEWORK

Farid MUSA[1], *Efe SEZGIN* [1,2]

1. Biotechnology Program, 2. Department of Food Engineering, Izmir Institute of Technology, Izmir, Turkey

Recent technological leaps in genome sequencing technologies have transformed the research potential in metagenomics, microbiome, and environmental DNA fields. [1] Microbiome research community swiftly grows and responds to the new advancements in microbiome research. Nevertheless, microbiome research methodologies and approaches evolve rapidly with frequent introduction of novel concepts such as Amplicon Sequence Variants (ASV) that are expected to replace traditional Operational Taxonomic Units (OTU). [2] Furthermore, 16S/18S rRNA or ITS based reference databases for phylogenetic and taxonomic classification of microorganisms have a considerable level of disparity between each other. [3] Result of this inconsistency poses serious trouble for scholars that perform research that involves merging, comparison or validation of independent studies, each of which ideally could have been processed using different reference databases such as Greengenes, SILVA, RDP, etc. Due to such disparity issues, currently, there is no effective tool that can merge two or more OTU tables with different taxonomies and phylogenetic trees. Traditional and probably all of the currently available tools make practically impossible to properly merge two or more OTU tables along with taxonomies and phylogenetic trees. Lack of such tools become particularly bothersome if a researcher has OTU abundance and taxonomy data without any raw sequence data. Given the lack of useful community tools, we aimed to develop a new framework that would address the discussed shortcomings. PhyloMAF is a next-generation phylogenetic microbiome analysis framework written in Python. The framework is under active development and will be published on GitHub (https://git.io/fjME1) at the time of article publication. PhyloMAF support importing both traditional TSV formatted OTU tables, BIOM file formats and popular phylogenetic tree file formats such as Newick or Nexus. Currently supported reference taxonomy databases are Greengenes, SILVA and Open Tree of life Taxonomy (OTT), while support of NCBI and Ribosomal Database Project (RDP) are under active development. Overall, PhyloMAF can work with heterogeneous microbiome data and its features include but are not limited to maintaining, manipulating, comparing, contrasting, merging, analyzing, plotting, exporting and report generation. For example, it is possible to aggregate taxonomy data based on the higher taxonomic ranks and then rematch taxonomy to the reference database to produce phylogenetic tree topology with tips that represent merged rank nodes. Likewise, OTU tables with or without similar taxonomies and/or samples can be merged by user-defined rules, and produced taxonomy can then be used to reconstruct phylogenetic tree topology from the reference database. Furthermore, if data have a low-quality taxonomic resolution, it is possible to improve resolution via any reference database or online services such as Taxonomic Name Resolution Service (TNRS). Finally, in order to adjust the phylogenetic tree, it is possible to use tools such as

RAxML [4] and OTU sequence data, which can be added manually or obtained via internal framework features. PhyloMAF allows the user to add sequence data for every taxon in tree manually by importing FASTA files or by fetching NCBI database. Other options are the possibility to directly import and apply multiple sequence alignment with sequence names associated with taxa IDs, or apply pre-evaluated distance matrix that can be used for tree construction/correction. The primary motivation for the development of PhyloMAF is to introduce a comprehensive taxonomy based microbiome data analysis Python framework to the microbiome research community that is designed by the S.O.L.I.D principles. [5] In conclusion, PhyloMAF will be useful for small and large scale studies as either independent framework or by integrating it into popular frameworks such as scikit-bio or QIIME2.

Keywords : Microbiome; Taxonomy; Phylogenetics

References: [1] Abdul-Aziz MA, Cooper A, Weyrich LS. Exploring Relationships between Host Genome and Microbiome: New Insights from Genome-Wide Association Studies. Front Microbiol. 2016;7:1611. Epub 2016/10/28. doi: 10.3389/fmicb.2016.01611. PubMed PMID: 27785127; PubMed Central PMCID: PMCPMC5061000.
[2] Callahan BJ, McMurdie PJ, Holmes SP. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J. 2017;11(12):2639-43. Epub 2017/07/22. doi: 10.1038/ismej.2017.119. PubMed PMID: 28731476; PubMed Central PMCID: PMCPMC5702726.
[3] Balvociute M, Huson DH. SILVA, RDP, Greengenes, NCBI and OTT - how do these taxonomies compare? BMC Genomics. 2017;18(Suppl 2):114. Epub 2017/04/01. doi: 10.1186/s12864-017-3501-4. PubMed PMID: 28361695; PubMed Central PMCID: PMCPMC5374703.
[4] Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22(21):2688-90.
[5] Martin RC, Martin M. Agile principles, patterns, and practices in C# (Robert C. Martin): Prentice Hall PTR; 2006

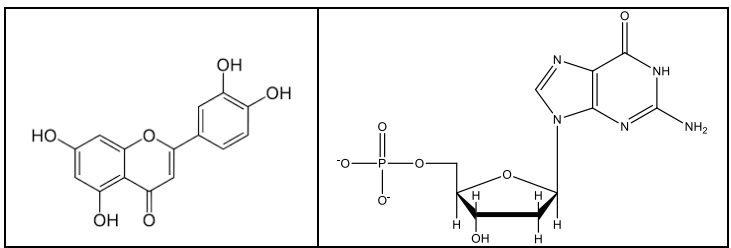
Corresponding Author's Address: Assoc. Prof. Efe SEZGIN: E-mail: efesezgin@iyte.edu.tr,
URL: https://food.iyte.edu.tr/staff/doc-dr-efe-sezgin

INVESTIGATION OF INTERACTIONS OF LUTEOLIN WITH DEOXYCYTIDINE MONOPHOSPHATE USING COMPUTATIONAL METHODS

Tuğçe Şener Raman, Nursel Açar Selçuki
Department of Chemistry, Faculty of Science, Ege University, İzmir, Turkey

Deoxyribonucleic acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. The two DNA strands are also known as polynucleotides as they are composed of simpler monomeric units called nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases (Cytosine [C], Guanine [G], Adenine [A] or Thymine [T]), a sugar called deoxyribose, and a phosphate group [1]. DNA is an important drug target facilitating expression and gene transcription, which also form the background of many life-threatening diseases. Drug molecules great importance today as many clinical anticancer drugs their derivatives. Many natural drug molecules are more potent than synthetic ones [2]. Luteolin (3,4,5,7-tetrahydroxyflavone is a crucial member of the flavones class and has been identified in many edible plants such as carrots, peppers, celery, olive oil, peppermint, thyme, and green tea. Luteolin has multiple biological effects such as antioxidant, antiinflammatory, antimicrobial, anticancer, and antiallergic. The anti-inflammatory, anticarcinogenic, and chemopreventive effects exhibited by luteolin is in part attributed to its antioxidant and free radical scavenging capacities [3]. In this study, the basic interactions between deoxycytidine monophosphate(dCMP) that is one of the DNA nucleotides and Luteolin will be investigated by computational tools. Conformal analyzes were performed to determine the initial structures for dCMP and Luteolin using Spartan 08 [4]. Ground state geometry optimizations are first performed with Gaussian 09 [5] at the ωB97XD/6-31+G(d,p) level without symmetry constraint in water, solvation calculations were performed by Tomasi's Polarizable Continuum Model (PCM) [6,7]. Interactions (van der Waals interactions, dispersive interactions and weak interactions of π-π stacking) between molecules and distances were determined. Also, molecular orbitals, energy differences of frontier orbitals and electrostatic potentials for studied molecules were investigated.

Keywords: Density Functional Theory, DNA, Luteolin and dCMP



Luteolin	d(CMP) (Deoxycytidine monophosphate)

Figure 1. Molecular structures of Luteolin and d(CMP)

References: [1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell (6th ed.). Garland. p. Chapter 4: DNA, Chromosomes and Genomes 2014 July. ISBN 978-0-8153-4432-2.
[2] Harborne, J. B., Williams, C. A. Advances in Flavonoid Research since 1992. Phytochemistry 2000; (55): 481–504.
[3] Lopez-Lazaro, M. Distribution and Biological Activities of the Flavonoid Luteolin. Mini-Rev. Med. Chem. 2009; (9), 31–59.
[4] Spartan08 for Windows, Wavefunction, Inc., Irvine, CA 92612 USA.
[5] Frisch M. J., et al. Gaussian09 Version C.01, 2009, Gaussian, Inc., Wallingford CT 8985.
[6] Tomasi J., Mennucci B., Cancès E.J. Journal of Molecular Structures (Theochem). 1999; (464):211-226.
[7] Tomasi J., Mennucci B., Cammi R., Chemical Reviews. 2005; (105):2999-3093.

MOLECULAR DYNAMICS SIMULATIONS OF PEPTIDE ANTIBIOTICS DEVELOPED BY MIMICKING NATURAL ANTIMICROBIAL PEPTIDES

Aslihan Ozcan[1], Tugba Arzu Ozal Ildeniz[1,2]

1. Institute of Science, Department of Medical Engineering, Acibadem University, Istanbul, Turkey
2. Department of Medical Engineering, Faculty of Engineering, Acibadem University, Istanbul, Turkey

In recent years, the rapid increase of infectious microorganisms' resistance against antibiotics has led to insufficiencies in the treatment of many fatal infections. As microorganisms rapidly gained resistance against each type of newly developed antibiotic, the number of studies on the development of antibiotics which microorganisms cannot develop resistance to have increased[1]. There are already naturally occurring antibiotics with a peptide structure that bacteria cannot develop resistance to[2]. The primary aim of this project is to develop antimicrobial peptides that mimic naturally occurring Cathelicidin through the application of molecular dynamicstechniques. Molecular dynamicstechniques are used not only to quickly observe how newly designed peptides act on the bacterial membrane to save time but also to observe phenomena that cannot be observed through experimental methods[3]. The method is quite interactive and involves the following steps; the three-dimensional structures of a newly developed peptide and a pre-equilibrated lipid bilayer are obtained. The protein is aligned on the lipid bilayer and also inside the water. Then, the water molecules that overlap with the proteins are removed. Equilibration and minimization of the newly developed system are carried out as final steps[4].

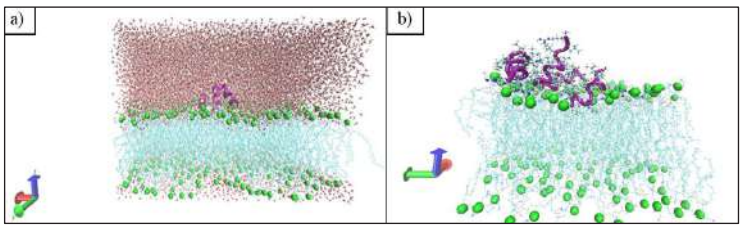


Figure 1. a) Result of 250ns(nanosecond) simulation. Phosphates were visualized with green balls, membrane lipids light blue, peptides alpha helix purple colour, water molecules red and white and CPK representation was chosen in VMD (Visual molecular dynamics)[5]. b) The water molecules were made invisible to be able to observe peptides' movements much more easily.

Through molecular dynamics procedures, at the end of the 250ns simulation, as seen figure 1, it is observed that the newly developed peptide headed towards the bacterial membrane and can penetrate into the membrane from its end parts.

Keywords: Antimicrobial Peptides; Peptide Design; Molecular Dynamics; Peptide Antibiotics.

Acknowledgements: This work was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) (Project no. **217S060** and Project no. **118Z859**).

References: [1] N. Unubol et al., "Peptide Antibiotics Developed by Mimicking Natural Antimicrobial Peptides," Clin. Microbiol. Open Access, 2017.
[2] D. Andreu and L. Rivas, "Animal Antimicrobial Peptides: An Overview," pp. 415–433.
[3] J. A. McCammon, B. R. Gelin, and M. Karplus, "McCammon, Gelin & Karplus Nature 1977 Dynamics of folded proteins," Nature, 1977.
[4] Theoretical Biophysics Group, "NAMD-Tutorial (Win)," Univ. Illinois, Urbana, USA, no. April, 2017.
[5] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics.," J. Mol. Graph., vol. 14, no. 1, pp. 33–8, 27–8, Feb. 1996

Corresponding Author's Address:
Aslihan Ozcan: aslihan.ozcan@live.acibadem.edu.tr

ALLOSTERIC REGULATION IN PROTEINS THROUGH RESIDUE-RESIDUE CONTACT NETWORKS

Melike Çağlayan, Nurcan Tunçbağ[1]

1. Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, Ankara, Turkey

Protein-protein interaction networks are helpful to understand functional organization of the proteome. They bind each other through non-covalent interactions. The native structure of a protein is also determined by the non-bonded forces such hydrophobic or van der Waals forces. Proteins are flexible and a perturbation in one region may affect a distant site which is called allosteric regulation. Allosteric regions play a very important role in regulating protein activity by binding a ligand to a non-active site. In this study, we examine the protein allosteric regions from through a structure-based network view. We represent each protein and protein complex structure as a residue-residue contact network and extract potential allosteric paths connecting two distant sites. For each structure we construct a residue-residue contact graph R (v, e) where v is the set of residues in the protein or protein complex and e is the set of contacts between these residues where if the distance between the Cα atoms of any two residues are less than 7Å then these residues are labelled as contacting. In this work, we assume that the minimum number of residues connecting allosteric regions will be located in allosteric paths. Therefore, we calculate all pair shortest paths between the set of residues in allosteric regions. We additionally analyze these paths if they overlap with any disease causing mutations, phosphorylation sites and the drug binding sites in proteins and protein-protein interactions. In this study, we also predict the presence and location of allosteric paths in protein structures whether the allosteric effect is transmitted through the core or the surface (Figure 1). Furthermore, we will discuss whether these regions are between different regions of proteins or between different regions in the same binding regions. For this purpose, we retrieved protein structures which have allosteric region from Protein Data Bank (PDB) and AlloSteric Database (ASD) database. Then the coverage of the structure (the number of residues present in the structure) were checked and the best coverage were considered. Additionally, we eliminated proteins having 90% sequence similarity to have one represent for each protein. To enrich the structural dataset, we include the homology models in ModBase. Protein-protein interactions and structural data are obtained from PDB and enriched with the predictive approaches such as Interactome3D. Drug binding site information are obtained from Drugport which lists the protein-drug binding regions and their corresponding PDB identifiers. To associate the mutations to allosteric paths, we collect the disease associated mutations from COSMIC database. If a mutation is in an allosteric site or path, it may have a significant effect in the binding of the protein with its partners or with drug molecules.

Keywords: Allostery, Protein Interaction, Residue-Residue Contact Network, Allosteric Paths

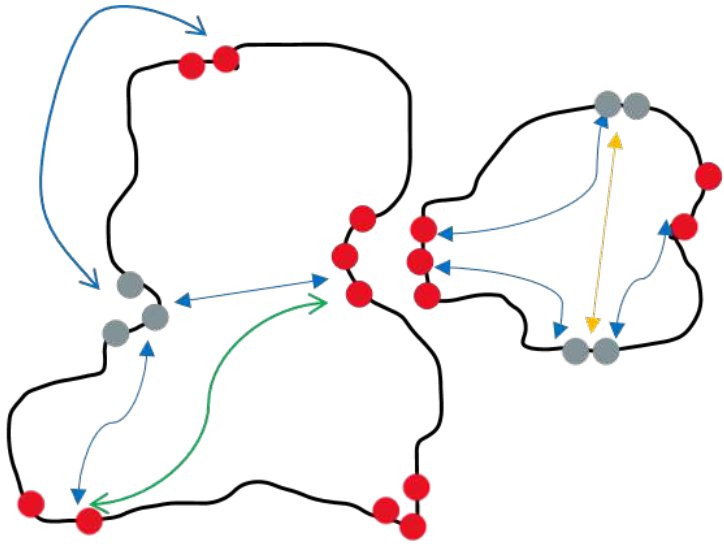


Figure 1. A cartoon representation of the slice mappings. Red points are active sites; gray points are allosteric sites. All lines represent interactions and sector (sectors connect some surface positions to the active site. The surface areas connected to the sector are hot spots for allosteric arrangement.). Blue lines are interaction between allosteric sites and active sites of protein complex (core or surface). Green line represents interaction between two active sites. Yellow line shows interaction between two allosteric sites.

Corresponding Author's Address: Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, Ankara, Turkey ntuncbag@gmail.com http://users.metu.edu.tr/ntuncbag/

RECONSTRUCTION AND INTEGRATION OF MATHEMATICAL MODELS OF SIGNALLING PATHWAYS IN STEM CELL DIFFERENTIATION AND REPROGRAMMING PROCESSES

Gülçin Selçuk, Pınar Pir[1]

Gebze Technical University, Department of Bioengineering, Çayırova, KOCAELI

Induced pluripotent stem cells (also known as IPS cells or iPSCs) are pluripotent stem cells that can be produced directly from adult cells. In recent years, induced pluripotent stem (IPS) cells are being increasingly used in development of in vitro models of neurological diseases in humans. During differentiation and reprogramming mechanisms, new cell identities are generated by restructuring the gene regulatory networks. Mostly, these regulatory networks rely on expression or repression of transcription factors. The identification of signaling pathways affecting the transcription factors is an essential step in cell differentiation mechanism studies, and dynamics of the mechanisms can be obtained by studying the time-dependent activities of these pathways. There is limited information in the literature about how gene-regulation networks, signaling pathways, and epigenetic profiles, are activated or inhibited in the processes of reprogramming and differentiation. Some of the pathways which are thought to be effective in neuron differentiation are: Notch, Shh, WNT, RA (Retinoic Acid Signaling) and BMP (Bone Morphogenetic Proteins) signaling pathways [1]. BIOMODELS [2] database is a a repository of mathematical models of biological processes and has mathematical models of Notch, WNT and Shh pathways implemented in SBML. However, a plausible model of RA and BMP pathways is not available. Hence we aimed to construct new models of these pathways using COPASI [3] tool. Ultimate aim of this study is modelling the gene regulatory networks and multiple signaling pathways together in the differentiation and reprogramming processes of induced pluripotent stem cells (IPS), fibroblasts and neuron cells. Better understanding of neural differentiation and reprogramming process may lead to new approaches for identification of new drug targets and treatment of neurodegenerative diseases.

Keywords: Neural Stem Cells, Differentiation, Signalling Pathways, Mathematical Modeling

References: [1] Sivakumar KC1, Dhanesh SB, Shobana S, James J, Mundayoor S. A systems biology approach to model neural stem cell regulation by notch, shh, wnt, and EGF signaling pathways. OMICS. 2011 Oct;15(10):729- 37 PubMed PMID: 21978399

[2] Chelliah V1, Laibe C, Le Novère N. BioModels Database: a repository of mathematical models of biological processes. Methods Mol Biol. 2013;1021:189-99. Pubmed PMID:23715986

[3] Stefan Hoops, Sven Sahle, Ralph Gauges, Christine Lee, Jürgen Pahle, Natalia Simus, Mudita Singhal, Liang Xu, Pedro Mendes COPASI—a COmplex PAthway SImulator, Bioinformatics, Volume 22, Issue 24, 15 December 2006, Pages 3067–3074.

Corresponding Author's Address: Pınar Pir e-mail: pinarpir@gtu.edu.tr

GENETIC CHARACTERISTICS OF AGEING-RELATED DISEASES AND WAYS TO IMPROVE HEALTHSPAN

Handan Melike Dönertaş[1], Matias Fuentealba Valenzuela[1,2], Linda Partridge[2,3], Janet M. Thornton[1]

1. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
2. Department of Genetics, Evolution and Environment, Institute of Healthy Aging, University College London, London, UK
3.Max Planck Institute for Biology of Aging, Cologne, Germany

Ageing is a major risk factor for many diseases. With the rise in life expectancy, overall burden of ageing- related diseases increases. The molecular link between ageing and age-related diseases, however, has not been explored in a systematic manner. In this study, we test whether diseases with similar age-of-onset share a genetic component that is also implicated in ageing. We perform GWAS on UK Biobank data, which includes genomic, medical and lifestyle measures for almost 500k participants. Our analysis comparing 116 diseases based on their age of onset profiles suggest late-life diseases do share a genetic component that is not prevalent in other diseases. Moreover, these results cannot be explained only by disease categories (e.g. cardiovascular, endocrine) or comorbidities. In order to explore the link between ageing and these diseases, we are now combining our results with publicly available datasets for ageing such as gene expression profiles of senescence and lifespan assays using model organisms. Identifying a shared ageing-related mechanism among multiple diseases offer an opportunity to target or even prevent multiple pathologies with a limited number of drugs and decrease the effect of polypharmacy on elderly while retaining the benefits.

Keywords: Ageing; GWAS; UK Biobank

Corresponding Author's Address:
melike@ebi.ac.uk, thornton@ebi.ac.uk

REVEALING PROBIOTIC TIR DOMAINS USING BIOINFORMATICS TOOLS

Bahar Bakar[1], *Burcu Kaplan Türköz*[2]

1. Ege University, Graduate School of Natural and Applied Sciences,Department of Food Engineering, İzmir,TR
2. Ege University, Faculty of Engineering, Department of Food Engineering, İzmir, Turkey

Toll-interleukin-1 receptor (TIR) is structural domain which is found in Toll-like receptors (TLR) and cytoplasmic adaptor proteins. TIR domain dimerization has a key role in TLR signaling pathway in immune system [1,2]. Recently, bacteria have been known to have TIR domain proteins (BTP) and pathogens have been shown to manipulate TLR signaling pathways using their BTPs via structural mimicry to adaptor TIR domains [3,4]. Research points out evidence that probiotics can also manipulate TLR signaling, but the mechanism has not been elucidated. In this context, our hypothesis is that probiotic bacteria can also produce TIR domain proteins. Following this, a gene region encoding TIR domain in the genome of probiotic Lactobacillus casei was found using BLAST search and this hypothetical protein was named as LcTIR. Results of multiple sequence alignment between LcTIR and other TIR domain proteins showed the presence of conserved regions on LcTIR [5]. Pairwise alignments showed that, LcTIR has high sequence identity with bacterial TIR domains (Table 1). Furthermore, LcTIR was modeled with RaptorX and I-TASSER and high quality models were obtained based on the quality criteria. Structural alignment was performed using PyMOL and RMSD values were evaluated between the obtained LcTIR model and other known TIR domains structures. LcTIR structural model was found to be similar to bacterial and human TIR domains (Table 2). The results of bioinformatics analysis supported our hypothesis that probiotics possess TIR domain proteins which is structurally similar to other TIR domains and thus has the potential to modulate TLR signaling. Future studies will focus on investigating the interactions of LcTIR with human TLR signaling components.

Keywords: Bioinformatics; Probiotic, TIR Domain; TLR Signaling

	***Brucella*-TIR (4lzp)**	***Salmonella*-TIR (WP_000028416)**	***Paracoccus*-TIR (3h16)**	**Human-TLR6-TIR (4om7)**	**Human-TLR1-TIR (1fyv)**
LcTIR	40.14%	39.10%	35.56%	25.23%	24.21%

Table 1: Sequence identity of LcTIR with other TIR domains

	***Brucella*-TIR (4lzp)**	***Paracoccus*-TIR (3h16)**	**Human-MyD88-TIR (2z5v)**	***Hydra vulgaris* – TIR (4w8h)**	**Human-TLR6-TIR (4om7)**
LcTIR	0.394	0.580	1.405	1.434	2.859

Table 2: RMSD from structural alignment between LcTIR model and other TIR domains

References:[1] O'Neill LA, Bowie AG. The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nat Rev Immunol. 2007 May;7(5):353-64. PubMed PMID: 17457343.
[2] Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. Nat Immunol. 2010 May;11(5):373-84. PubMed PMID: 20404851.

[3] Chan SL, Low LY, Hsu S, Li S, Liu T, Santelli E, Le Negrate G, Reed, JC, Woods VL, Pascual J. Molecular mimicry in innate immunity: crystal structure of a bacterial TIR domain. J Biol Chem. 2009 Aug 7; 284(32): 21386–21392. PubMed PMID: 19535337.
[4] Kaplan-Türköz B, Koelblen T, Felix C, Candusso MP, O'Callaghan D, Vergunst AC, Terradot L. Structure of the Toll/interleukin 1 receptor (TIR) domain of the immunosuppressive Brucella effector BtpA/Btp1/TcpB. FEBS Lett. 2013 Nov 1;587(21):3412-6. PubMed PMID: 24076024.
[5] Salcedo SP, Marchesini MI, Lelouard H, Fugier E, Jolly G, Balor S, Muller A, Lapaque N, Demaria O, Alexopoulou L, Comerci DJ, Ugalde RA, Pierre P, Gorvel JP. Brucella control of dendritic cell maturation is dependent on the TIR-containing protein Btp1. PLoS Pathog. 2008 Feb 8;4(2):e21. PubMed PMID: 18266466.

Acknowledgements: This work is supported by TÜBİTAK- 3501 (No:116Z299).

Corresponding Author's Address: Ege University, Graduate School of Natural and Applied Sciences,Department of Food Engineering, İzmir,TURKEY E-mail: baharbakarege@gmail.com

Prediction of the Effects of Single Amino Acid Variations on Protein Functionality with Annotation Centric Modeling

Fatma Cankara[1], *Tunca Doğan*[2]

1 Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey
2. Institute of Informatics / Department of Computer Engineering, Hacettepe University, 06800 Ankara, Turkey

Whole-genome and exome sequencing studies have indicated that genomic variations may cause deleterious effects on protein functionality via various mechanisms. It has been reported that single nucleotide variations that alter the protein sequence (and thus, the structure and the function), namely nsSNPs, are highly associated with genetic diseases in human. These variations also affect the receptor-ligand interactions, which is one of the key reasons that many of the early drug candidates fail in clinical trials. In this study, we propose a new methodology to collect and organize the information related to the effects of sequence variations from various biological databases and to utilize this information in a machine-learning based system to predict the function changing capabilities of mutations with unknown consequences.

The studies aiming to predict the effect of sequence alterations in proteins often exploit sequence information (mostly using alignment) and 3-D structural information. In this study, we took a different perspective and evaluate these variations' function altering capabilities using an annotation-centric focus (i.e., domains, motifs and other sequence features), with the aim of complementing conventional approaches in the literature. Functional regions/sites of proteins are the parts, where a sequence variation may have a more significant consequence and should be evaluated accordingly. In the proposed methodology, we make use of a variety of descriptive features including:

i.

a. Correspondence between the mutated site and different protein sequence feature annotations (obtained from the UniProt database) such as residues taking part in disulfide bonding, nucleotide binding, zinc fingers, glycosylation, helix, repeats etc. (30-D,binary).

b. For the cases where the mutation does not perfectly correspond to the annotated sites on the sequence, the atomic distances (on the 3-D structure) between the mutated residue and the annotated residue (based on alpha-carbons) are incorporated in an additional 30-D vector with real values (angstrom).

c. The information about InterPro/Pfam domain entry, where the mutation resides in (1-D, categorical).

ii. Mutated residue'sstructural information that we deduce from PDB structures by aligning the protein sequence with the corresponding PDB structure's sequence and by categorizing the mutated residues based on their accessible surface area as core, surface or interface

(1-D,categorical). iii. Physicochemical properties of the mutation obtained from the widely accepted Grantham Matrix's 3 distance scores: the change in polarity, composition and molecular volume, upon the occurrence of the variation (3-D, real-values). We employed decision tree (DT) and random forest (RF) classifiers, first, to train our models (using the mutations with known consequences from UniProt, Clinvar and PMD, a total of 108,072 mutations), and then, to predict the effect of unknown variations by categorizing them either as deleterious or neutral, by querying the finalized 65-D feature vector of each variation, on our prediction models. Currently, we are in the process of running our models on widely accepted benchmark datasets from previous studies to calculate our performance, and to compare them with the state-of-the-art. Finally, we plan to combine our method with the state-of-the-art methods, which employ different featurization approaches, to maximize the prediction performance in an ensemble-based tool.

Keywords: Single Amino Acid Variations, Machine Learning, Decision Trees, Random Forest, Predicting Disease Capacity Of Mutations, UniProt, Clinvar, Annotation Centric Modeling, Personalized Medicine.

Corresponding Author's Email Address:

tuncadogan@hacettepe.edu.tr

INVESTIGATION OF STRUCTURAL AND ELECTRONIC PROPERTIES OF BIOLOGICALLY IMPORTANT SMALL NUCLEOTIDES BY MOLECULAR MODELING

Hasip Cirkin[1], *Cenk Selcuki*[2]

1.Ege University, Institute for Natural and Applied Science, Biochemistry Program
2.Ege University, Faculty of Science, Biochemistry Department

In eukaryotic organisms, the flow of genetic information is regulated by the transfer of information from DNA to mRNA ending with a functional protein. In addition, some proteins implicated in the transcription of a gene by binding to short DNA sequences in promoter region in order to initiate RNA synthesis [1].

In this study, we aimed to generate 3D molecular modelling of the nucleotide (primarily mono-tetramer) structures whose specific biological functions are known and to reveal electronic properties of these structures. The Spartan'14 program was used to analysis of the conformer structures of the mono, di, tri and tetramer structures of the TATA-box sequence. The MOPAC2016 program [2,3] with PM6-D3H4 method and Gaussian09 program [4] at ωB97XD/6-311++G(d,p) level were used for optimizations and frequency calculations for each investigated conformer. Using the calculated results, the most stable structures were determined using the relative energies. Molecular structures were drawn by Discovery Studio Visualizer 2019. As a result, we found that the –H bond was formed between the –OH group of the ribose ring and -PO4 group of the nucleotide structure. In addition, we observed nucleobase conjugates (A=T) of tetranucleotide structures and CH$-\pi$ bond, as non-covalent interaction. Most of the calculations were performed on TUBITAK- ULAKBIM Truba resources.

Keywords: Nucleotide; Molecular Modelling; Density Functional Theory

References: [1] Sainsbury S, Bernecky C, Cramer P. Structural basis of transcription initiation by RNA polymerase II. Nat Rev Mol Cell Biol. 2015;16(3):129–143. PubMed PMID: 25693126
[2] Stewart J. J. P. 1989,Optimization of Parameters for Semiempirical Methods II. Applications. J. Comput.
Chem. 10, 221–26410.1002/jcc.540100209.
[3] Stewart J. J. P., MOPAC2016, (2016). http://openmopac.net/.
[4] Frisch, M.J. et al., (2009). https://gaussian.com/g09citation/

Corresponding Author's Email Address: Ege University, Faculty of Science, Biochemistry Department cenk.selcuki@gmail.com

CHARACTERIZING IMPACT OF SOMATIC MUTATIONS IN BREAST CANCER: SF3B1 CASE STUDY

Asmaa Samy[1], Baris Suzek[2], Mehmet Kemal Ozdemir[1] and Ozge Sensoy [1]

1. School of Engineering and Natural Science, Istanbul Medipol University, Istanbul, Turkey
2.Department of Computer Engineering, Muğla Sıtkı Koçman University, Muğla, Turkey

Breast cancer is the most common and the second lethal type of cancer among women. The etiology of the disease has remained elusive since it is caused by a combination of various factors such as genetic, epigenetic as well as the environmental. The somatic mutations are one of such factors that affect breast cancer development and prognosis. Many somatic mutations have been identified in breast cancer and made publicly available through bioinformatics resources. For instance, as of March 2019, the Catalogue of Somatic Mutations in Cancer (COSMIC) database reports over 73,000 somatic mutations in breast cancer. Unfortunately, in many cases, our knowledge on these mutations is limited to their allele frequencies and their relations to cancer deserve further investigation.

In this work, we defined an in silico process to investigate the impact of somatic mutations in breast cancer. The process involves (1) identification of genes participating in a biological process potentially involved in cancer (e.g. harboring highly mutant genes), (2) building a gene network, (3) superimposing cancer somatic mutation frequencies to the network, (4) identifying a gene with high network centrality and showing high mutation rates, and finally, (5) performing molecular dynamics simulations to assess the impact of the frequent somatic mutations for the protein encoded by this gene. To test the process, we use the biological process of "pre-mRNA splicing" and build a network of 80 genes using STRING database. The somatic mutation frequencies for these genes are collected from COSMIC (March 2019 release) [1,2] and superimposed onto the network. Considering both of centrality metrics and mutation frequencies, we selected Splicing factor 3B subunit 1 (SF3B1) and its hotspot K700E mutation for further study. SF3B1 is the core component of SF3B spliceosome complex which is responsible for branch point recognition in mRNA splicing. We performed atomistic molecular dynamics simulations using the cryo-EM structure of the native SF3B1 protein (PDB:5z56) [3] as well as the mutant. Our results showed that K700E mutation decreases the stability of SF3B complex's components, particularly those in direct interaction with SF3B1 and mRNA. Furthermore, the SF3B1's interaction with mRNA is adversely affected at a scale that may lead to aberrant splicing of mRNA and translation of non-functional proteins.

Our work, though need experimental validations, sheds light on SF3B1 K700E mutation's impact on mRNA splicing in breast cancer. We anticipate our in silico process can be improved and employed in the characterization of impact for other cancer somatic mutations.

Keywords: Breast cancer; Somatic Mutations; mRNA Splicing; SF3B1

References: [1] Tate, J. G., et al. (2019). "COSMIC: the Catalogue Of Somatic Mutations In Cancer." Nucleic Acids Res 47(D1): D941-d947. PubMed PMID: 30371878.
[2] COSMIC - Catalogue of Somatic Mutations in Cancer. Retrieved from https://cancer.sanger.ac.uk/cosmic
[3] Zhang, X., et al. (2018). "Structure of the human activated spliceosome in three conformational states." Cell Res 28(3): 307-322. PubMed PMID: 29360106.

Corresponding Author's Address: amahmoud@st.medipol.edu.tr

GRPCLASSIFIEREC : A NOVEL CLASSIFICATION APPROACH BASED ON THE ENSEMBLE CLUSTERING SPACE

Loai Abdallah[1] *and Malik Yousef*[2]

1. The Department of Information Systems, The Max Stern Yezreel Valley Academic College, Israel, E-mail: Loai1984@gmail.com).
2. The Department of Community Information Systems, Zefat Academic College, Zefat, 13206, Israel, E-mail: malik.yousef@gmail.com

Abstract

Background

Advances in molecular biology have resulted in big and complicated data sets, therefore a clustering approach that able to capture the actual structure and the hidden pattern of the data is required. Moreover, a new classification approach that based on space that capture the structure is also essential. Additionally, the performance of many supervised or unsupervised machine learning algorithms depends considerably on distance metrics to determine similarity between data points. A suitable distance metric could improve the classification performance and clustering process significantly.

Distance metrics over a given data space should reflects the actual similarity between objects. One of the obvious weaknesses of the Euclidean distance is dealing with data that is represented by a large number of attributes, where the Euclidean distance does not capture the actual relationship between those points. However, objects belonging to the same cluster usually share some common traits even though their Euclidean distance might be relatively large.

Results

In this study, we propose a new classification method named GrpClassifierEC that replacesthe given data space with categorical space based on ensemble clustering (EC). The similarity between two objects is defined as the number of times that these objects were not belong to the same cluster. The EC space is defined by tracking the membership of the points over multiple runs of clustering algorithms. Different points that were included in the same clusters will be represented as a single point. Our algorithm classifies all these points as a single class. In order to evaluate our suggested method, we compare its results to the k nearest neighbors, Decision tree and Random forest classification algorithms on several benchmark datasets. The results confirm that the suggested new algorithm GrpClassifierEC outperforms the other algorithms.

Conclusions

Our algorithm can be integrated with many other algorithms. In this research, we use only the k-means clustering algorithm with different k values. In future research, we propose several directions: (1) checking

the effect of the clustering algorithm to build an ensemble clustering space. (2) finding poor clustering results based on the training data, (3) reducing the volume of the data by combining similar points based on the EC.

Availability and implementation

The KNIME workflow, implementing GrpClassifierEC, is available at http://malikyousef.com

Keywords: Decision Trees, Ensemble Clustering, Classification.

Corresponding author's address : malik.yousef@gmail.com

Corresponding Author's Address:

Loai1984@gmail.com

TOWARDS AN INTERNET OF SCIENCE

Jens Allmer[1]

1. Medical Informatics and Bioinformatics, Hochschule Ruhr West, University of Applied Sciences, Mülheim an der Ruhr, Germany

Bioinformatics is heavily invested in big data and complex data analysis workflows (pipelines). Such workflows may consist of many computational tools. Many computational tools for bioinformatics data analyses have been developed in the past decades. The same is true for workflow management systems, which allow the combining of computational tools into data analysis pipelines. A few examples of workflow management systems used in bioinformatics include Taverna [1], Galaxy [2], and KNIME [3]. A difficulty in creating reproducible and resilient workflows stems from the availability of many options for each data analysis step. For example, more than 50 computational tools for read mapping are available and it is not easily discernable which ones may work best in a given situation. Apart from this uncertainty, the availability of more than 50 tools for a given purpose represents a large duplication of effort. Apart from these points, it is not feasible for a single researcher to ensure that all tools work correctly to pick the most efficient one for the intended use case. This problem is aggravated since only a limited number of the tools are used frequently and are actively developed that potential errors have already been reported and addressed. Thus, the current state of workflow management in bioinformatics is combining largely untested tools into computational pipelines and using them for data analysis even in critical fields such as medical decision support. Data analyses are presently performed on personal computers/workstations/clusters. However, in the future, a shift of development and analysis to the cloud will likely occur. It appears that no current workflow management system is ready for this transition to the cloud. A community effort can fill the gap and produce many positive side effects for the field of bioinformatics [4]. The envisioned system, loosely based on the Internet of Things, will overcome current duplications of effort, introduce proper testing, allow for development and analysis in public and private clouds, and include reporting features leading to interactive documents. Thereby, the collaborative development of versioned computational workflows can build on correct tools. Adding proper integration testing to the workflow development ensures that workflows can be confidently used in data analysis, publication, and medical decision support (Fig. 1).

Keywords: Workflow Management; Computational Pipelines; Internet Of Things; Code Smells; Scientific Computing

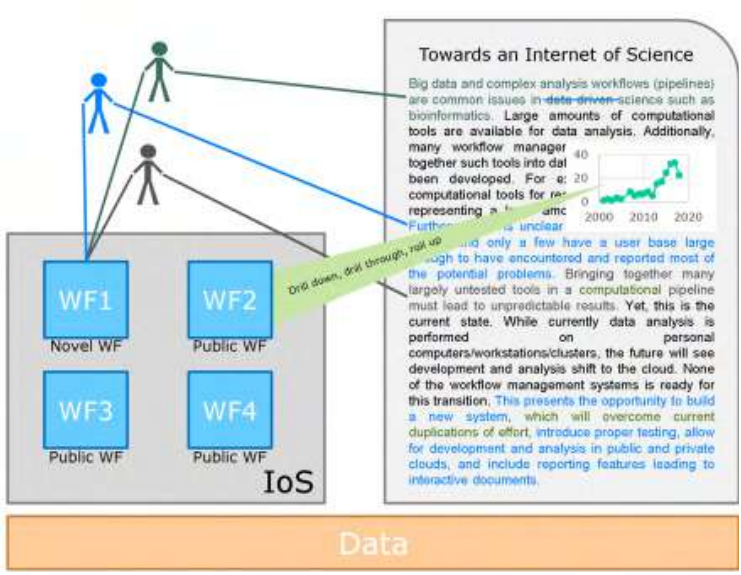


Figure 1. The IoS shall provide end to end collaborative data analysis and publication. All parts fit together such that figures and tables within a document can be traced to their supporting raw data.

References: [1] Hull D, Wolstencroft K, Stevens R, Goble C, Pocock MR, Li P, et al. Taverna: a tool for building and running workflows of services. Nucleic Acids Res. 2006 Jul 1;34(Web Server issue):W729-32. Pubmed PMID: 16845108.
[2] Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018 Jul 2;46(W1):W537–44. PMID: 29790989
[3] Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz Information Miner. In: Preisach C, et al., editors. Data Analysis, Machine Learning and Applications. Springer; 2008.
[4] 1. Allmer J. Towards an Internet of Science. J Integr Bioinform. 2019 May 30. PubMed PMID: 31145694.

Corresponding Author's Address:

Prof. Dr. Jens Allmer, Hochschule Ruhr West, University of Applied Sciences, Medical Informatics and Bioinformatics, 45479 Mülheim an der Ruhr, Germany, jens@allmer.de, https://www.hochschule-ruhr-west.de/forschung/forschung-in-den-instituten/institut naturwissenschaften/beschaeftigte/prof-dr-jens-allmer/.

Predicting Potential Allosteric Communication Pathways in Pyruvate Kinase Using Residue Network Model

Zehra Sarıca[1], Ozge Kurkcuoglu[1]

1.Department of Chemical Engineering, Istanbul Technical University, Istanbul, Turkey

Background: Pyruvate kinase catalyzes the last step in glycolysis and is therefore a key enzyme in cellular metabolism. The homo-tetrameric structure accommodates four active sites and uses allostery to modulate its activity. The latter property makes this protein an attractive target for the treatment of infectious diseases, where the design of species-specific drugs with high selectivity is possible, especially for the lowly conserved allosteric sites.

Method: In this study, twenty crystal structures belonging to H. sapiens and three structures of S. aureus are modeled as residue-networks composed of inter-connected nodes, [1] which trace the skeleton of the protein topology. Centrality measures – degree, closeness and betweenness – are used to reveal critical residues that mark differences between two species, which can be evaluated as new drug target sites.

Results: The model captures functionally important residues, located at the active sites, at known allosteric sites as well as at the monomer interfaces. Residues with high betweenness measures point to short and long pathways of residues that can transmit a perturbation among functional sites of the protein. In addition, webs of potential allosteric pathways have different patterns in tense and relaxed conformations of the enzyme.

Conclusion: Results revealed the similarities and differences between allosteric mechanisms employed by two species. Here, two different sites on pyruvate kinase of S. aureus are proposed as new drug target sites; at the monomer interface and between the active and allosteric sites. The low sequence conservation and critical location of these regions make these sites attractive for novel species-specific drugs with high selectivity for pyruvate kinase of S. aureus.

Keywords: Allostery; Pyruvate Kinase; Centrality Measures; Residue Network Model

References: [1] Kürkçüoğlu, Ö. Exploring allosteric communication in multiple states of the bacterial ribosome using residue network analysis. Turkish Journal of Biology, 42(5): 392–404 (2018). doi:10.3906/biy-1802-77

Corresponding Author's Address: olevitas@itu.edu.t

FAST DETECTION OF MAXIMAL EXACT MATCHES WITH UNIVERSAL K-MER SAMPLING

Baris Ekim[1], *Zeynep Harcanoglu*[2]

1. Massachusetts Institute of Technology (MIT), Cambridge, MA, USA
2. TED Atakent College, Istanbul, Turkey

As next-generation sequencing (NGS) becomes much cheaper and faster, an urgent need for faster and memory-efficient methods, algorithms, and data structures arises. In order to efficiently manage and analyze these data, newer computational approaches and data structures are essential. One widely-studied concept in sequence analysis is maximal exact matches (MEMs), which are exact matches between two strings that cannot be extended in either direction without a mismatch. MEMs are used in whole-genome alignment [1–3], short-and long-read alignment [4, 5], and as anchor points in comparison of closely related genomes [6, 7].

Despite recent advancements in computing MEMs, the original problem remains challenging, both in resources and runtime, and cannot be effectively handled for large genomes. We present a novel k-mer indexing method for detecting MEMs based on the sampling of k-mers from a universal hitting set [8], which is a set of k-mers such that every sequence of length L contains at least one k-mer from the set. Using existing tools [9], we compute a small UHS for a specific value of k and L and build a universal k-mer index of the reference genome. Consequently, we show that universal k-mer sampling of the query genome to find and extend matches to detect MEMs uses less space and processes queries faster than existing sampling methods. We conclude that filtering and indexing only the universal k-mer in the reference and query genomes reduces memory footprint and decreases running time, mitigating the MEM enumeration problem.

Keywords: Maximal Exact Matches; Universal Hitting Sets; K-Mer Sampling

References: [1] Bray N, Dubchak I, and Pachter L. Avid: A global alignment program. Genome Research,13(1):97–102, 2003.
[2] Choi JH, Cho HG, and Kim S. Game: a simple and efficient whole genome alignment method using maximal exact match filtering. Computational Biology and Chemistry, 29(3):244–253,2005.
[3] Hohl M, Kurtz S, and Ohlebusch E. Efficient multiple genome alignment. Bioinformatics,18: S312–S320, 2002.
[4] Liu Y and Schmidt B. Long read alignment based on maximal exact match seeds. Bioinformatics, 28(18): i318–i324, 2012.
[5] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997, 2013.
[6] Bray N and Pachter L. Mavid: Constrained ancestral alignment of multiple sequences. Genome Research, 14(4):693–699, 2004.
[7] Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, and Salzberg SL. Versatile and open software for comparing large

genomes. Genome Biology, 5(2): R12, 2004.
[8] Orenstein Y, Pellow D, Marcais G, Shamir R, and Kingsford C. Designing small universal k-mer hitting sets for improved analysis of high-throughput sequencing. PLoS Computational Biology, 13(10): e1005777, 2017.
[9] Ekim B, Berger B, and Orenstein Y. Memory-efficient parallel algorithms for approximating compact universal hitting sets. bioRxiv, 2019.

Corresponding Author's Address: 3 Ames Street, Cambridge, MA, USA baris@mit.edu http://people.csail.mit.edu/ekim

EXPLORING ALTERED REGULATORY TRANSCRIPTION FACTOR-GENE INTERACTIONS INDUCED BY BORIC ACID AT HALF MAXIMAL INHIBITORY CONCENTRATION THROUGH COMBINED USE OF NETWORK MODELING ALGORITHMS

Ayegül Tombuloğlu[1], *Yeşim Aydın Son*[2]

1. Middle East Technical University, Graduate School of Informatics, Health Informatics Department
2. Middle East Technical University, Graduate School of Informatics, Health Informatics Department

Boric acid is well known for its concentration dependent effects on biological processes both at the cellular level and organism level. Treating cell lines with boric acid may help reduce toxic effects which might reinforce cell survival and proliferation. On the other hand, higher concentrations of boric acid have been reported to inhibit proliferation for many cell lines. In a gene expression profiling study performed with HepG2 hepatoma cell line, we have observed substantial decrease in the expression of genes involved in cell-cycle progression and many metabolic processes at half-maximal inhibitory concentration of boric acid. To discover regulatory mechanisms lying beneath the expression changes, a network made up of regulatory interactions differentiating in boric acid exposed and unexposed HepG2 cells was obtained by using PANDA algorithm [1]. Consequently, forest algorithm of Omics-Integrator software[2] was used to find optimal subnetworks which facilitates understanding the essential boric acid induced regulatory changes in the network and their visualization in a brief and concise manner. In the future, we plan to validate the most prominent hubs of regulatory network and perform similar network modeling approach at lower concentrations of boric acid.

Keywords: Boric Acid, HepG2, Biological Network Modelling, Regulatory Network, Microarray

References: [1] Glass K, Huttenhower C, Quackenbush J, Yuan GC. Passing messages between biological networks to refine predicted interactions. PLoS One. 2013 May 31;8(5):e64832. doi: 10.1371/journal.pone.0064832. PubMed PMID: 23741402
[2] Tuncbag N, Gosline SJ, Kedaigle A, Soltis AR, Gitter A, Fraenkel E. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. PLoS Comput Biol. 2016 Apr 20;12(4):e1004879. doi: 10.1371/journal.pcbi.1004879. PubMed PMID: 27096930
Corresponding Author's Address: Middle East Technical University, Graduate School of Informatics, Health Informatics Department

RAPID PATHOGEN DETECTION OF USING CONVOLUTIONAL NEURAL NETWORKS ON METAGENOMICS DATA

Meryem Altın KARAGÖZ[1], *Özkan Ufuk NALBANTOĞLU*[1,2]

1. Computer Engineering Department, Erciyes University, Kayseri, Turkey

2. Genome and Stem Cell Center (GenKök), Kayseri, Turkey

The targeted treatment of infectious diseases is crucial in several scenarios such as sepsis in which broad spectrum antibiotics are usually ineffective, yet the action has to be taken in a short time window as the disease progresses rapidly in hours towards terminal outcome. Rapid identification of pathogens that are the causative agents of an infection is therefore in need, in order to guide diagnosis and antimicrobial treatment correctly. Around 72 hours, which leads to a late time of intervention in some cases, is required for conventional detection methods. Novel methods that can perform fast detection such as immunological tests, PCR panels, and biosensors target a small biodiversity range and cannot correspond to generic use. On the other hand, it is possible to design detection systems that can search all pathogenic species and diagnose with high accuracy using next generation sequencing technologies. In this new paradigm, pathogen detection task can be stated as the taxonomic classification of each sequencing read to species of origin. To date, several machine learning-based methods for the classification of DNA sequencing reads to taxa have been proposed, and many of them are adopted for clinical use. With the success of deep learning algorithms in a broad spectrum of data classification problems, the corresponding taxonomic detection problem has been also addressed by the new architectures of neural- network classifiers, and successful results have been reported recently [1]. The main approach in the proposed methodologies in the current literature is basically feeding the oligonucleotide count vectors to Convolutional Neural Networks (CNN) feature extractors and cascading the taxonomic classifiers on top of these feature extractors. However, as CNNs are nonlinear feature extractors for short term correlated sequences (or as nonlinear counterparts of finite impulse response filters), the correlation exploitation capacity of CNNs are restricted to k-mer size, which saturates quickly. We claim that instead of using k-mer profiles, directly embedding the DNA sequences via one-hot-encoding (OHE) and running the CNN on the encoded sequences would be able to exploit further correlations and yielding better classification performances consequently. From these perspectives, pathogen data that was generated by different NGS platforms, respectively Illumina whole genome sequencing (WGS) and MinION WGS pathogen database have been used for classification in our study. In order to compare the performances of two CNN approaches, spectral representation method that has been generated with k-mer (the length of the oligonucleotide=k) co-occurrence of counts and OHE have been used. CNN models were trained on represented data with several read lengths ranging from 500 to 10000, L= {500, 1000, 10000} also

with distinct k-mer size ($3 \leq k \leq 7$) at genus level. The resulting detection accuracies can be found in Table 1. CNN model has achieved better accuracy rates with OHE representation in several cases. As a result, CNN is capable of learning directly from the nucleotide sequences without intermediate feature extraction steps enabling-end-to-end detection systems.

Keywords: Pathogen Detection, Dna Sequence Analysis, Metagenomics, Machine Learning

Dataset	Read Length	Algorithms	k=1	k=3	k=4	k=5	k=6	k=7
MinION	500 bp	CNN-ohe	79,81%					
		CNN-cor		70,50%	75,58%	78,58%	77,23%	**80,21%**
	1000 bp	CNN-ohe	**86,71%**					
		CNN-cor		80,17%	86,03%	85,98%	86,68%	86,12%
	10000 bp	CNN-ohe	**85,93%**					
		CNN-cor		77,29%	85,09%	90,55%	84,86%	63,69%
Illumina	500 bp	CNN-ohe	**88,56%**					
		CNN-cor		81,06%	84,47%	86,29%	87,95%	87,23%
	1000 bp	CNN-ohe	92,47%					
		CNN-cor		89,66%	92,80%	94,16%	94,67%	**95,24%**
	10000 bp	CNN-ohe	93,52%					
		CNN-cor		96,28%	97,82%	98,34%	**98,44%**	98,24%

Table 1: Comparison of pathogen detection accuracy for OHE (CNN-ohe) and k-mer based (CNN-cor) methods. The algorithms for each dataset have been tested with 10k-fold validation.

References: [1] Fiannaca, A., La Paglia, L., La Rosa, M., Renda, G., Rizzo, R., Gaglio, S., & Urso, A. (2018). Deeplearning models for bacteria taxonomic classification of metagenomic data. BMC bioinformatics, 19(7), 198.

Corresponding Author's Address: Meryem Altın KARAGÖZ (maltinkaragoz@gmail.com)

HEAD-AND-NECK CANCER: PERFORMING FUNCTIONAL GENE ENRICHMENT STUDY TO DISCOVER THE NEW POTENTIALS AS BIOMARKER

Evren Atak[1], Ahmet Melih Öten[2], Şeyma Elbeyoğlu[1], Mete Emir Özgürses[1],Öykü İrigül Sönmez[3], Aslı Yenenler[2]

1. Department of Molecular Biology&Genetics, Biruni University, İstanbul
2. Department of Biomedical Engineering, Biruni University, İstanbul
3.AYA R&D Biotechnology Inc, İstanbul

Head and neck cancer (HNC) is the one of the most widespread cancers with high mortality in the world ..It is defined as a complex disease, existing in the different locations of head and neck including; hypopharynx, oropharynx, lip, oral cavity, nasopharynx, or larynx [1]. There are many risk factors affecting head and neck cancer. These risk factors are; smoking, alcohol, occupational toxics, HPV-infection, chemoprevention and genetics factor. Many different treatment methods can be followed depending the combination of the risk factors and the sub-type of head and neck cancer. Followed by the surgical treatment, radiation therapy, chemotherapy, targeted therapeutics and novel agents are commonly used for the effective treatment of HNC. Even CT and PET are used to detect and to stage HNC; the novel diagnostic approaches are required for the early diagnosis of HNC with highest accuracy and specificity, especially clarifying the subtypes of HNC [2]. Among many, the discovery and further development of biomarker is seen as the most popular approach with highest specificity in the field of genomics and proteomics.

In this study, we basically aim to perform functional gene enrichment in HNC to identify the significantly expressed genes that play a crucial role for the development and staging of HNC in terms of function, mechanism and metabolic processes. As already demonstrated in the literature, we have revealed 6 genes having played a role in HNC that are FRMD5, PCMT1, PDGFA, TMC8, YIPF4 and ZNF324B [3]. Through the miRwalk, we firstly listed the names of miRNA that are regulated by FRMD5, PCMT1, PDGFA, TMC8, YIPF4 and ZNF324B genes. Then, we moved on the second part such that the corresponding genes regulated by the sets of miRNAs are listed and our gene pool was expanded. The miRNAs and regulated genes are displayed. With the finalized list of genes, gene-gene interaction network has been created via STRING. Pathways playing a role in HNC are highlighted with the indication of gene names.

Lastly, HNC RNAseq data was taken from GEO dataset with number GSE83519 to perform the path analysis with R package of pathfindR that is run by significantly expressed genes. As a result, 3 different pathways regulated by ZNF324B gene are revealed as cell cycle, p53 signaling pathway and protein processing in endoplasmic reticulum. Experimental and computational evidences are found in the literature about the strong association between these listed pathways and HNC [4]. This demonstrates us that ZNF324B gene could be a candidate of development of biomarkers for HNC that basically follows cell cycle,

p53 signaling and protein processing pathways. As a further, we want to elucidate the role of ZNF324B gene for subtyping of HNC, of course with the association of other genes significantly expressed in HNC with high specificity. We believe that our further demonstration will be used for the introduction of new biomarker kit for early diagnosis and subtyping of HNC.

Keywords: Gene-Gene Interaction, Molecular Diagnostic, Head And Neck Cancer, Discovery of Biomarkers

References: [1]: Lo Nigro C, Denaro N, Merlotti A, Merlano M. Head and neck cancer: improving outcomes with a multidisciplinary approach. Cancer Manag Res. 2017 Aug 18;9:363-371. doi: 10.2147/CMAR.S115761. PubMed PMID: 28860859; PubMed Central PMCID: PMC5571817.
[2]: Palka KT, Slebos RJ, Chung CH. Update on molecular diagnostic tests in head and neck cancer. Semin Oncol. 2008 Jun;35(3):198-210. doi: 10.1053/j.seminoncol.2008.03.002. PubMed PMID: 18544435; PubMed Central PMCID: PMC2490629.
[3]: Guo W, Chen X, Zhu L, Wang Q. A six-mRNA signature model for the prognosis of head and neck squamous cell carcinoma. Oncotarget. 2017 Oct 10;8(55):94528-94538. doi: 10.18632/oncotarget.21786. PubMed PMID: 29212247; PubMed Central PMCID: PMC5706893.
[4]: Moon, H., Han, H. and Jeon, Y. (2018). Protein Quality Control in the Endoplasmic Reticulum and Cancer. International Journal of Molecular Sciences, 19(10), p.3020. PubMed PMID: 30282948 ; PubMed Central PMCID: PMC6213883.

Corresponding Author's Address: Biomedical Engineering, Faculty of Engineering and Natural Sciences, Biruni University, İstanbul - ayenenler@biruni.edu.tr

AN ATTEMPT TO IDENTIFY THE POTENTIAL BIOMARKERS IN THE BLADDER CANCER VIA FUNCTIONAL GENE-ENRICHMENT APPROACH

Şeyma ELBEYOĞLU[1], Ahmet Melih ÖTEN2, Evren ATAK[1], Mete EmirÖZGÜRSES[1], Öykü İRİGÜL SÖNMEZ[3], Aslı YENENLER[2]

1. Department of Molecular Biology & Genetics, Biruni University, İstanbul
2. Department of Biomedical Engineering, Biruni University, İstanbul
3. AYA R&D Biotechnology Inc, İstanbul

Bladder cancer (BC) is one of the most common cancer types in the world. BC considered as complex disease and it develops along two "tracks" that have very different implications for prognosis: muscle invasive bladder cancer (MIBC) and non-muscle invasive bladder cancer (NMIBC). Up to now, several treatment methods are developed for BC depended on whether being muscle invasive or not. For example, the resection of tumour followed by induction and conservation immunotherapy with intravesical BCG vaccine or intravesical chemotherapy is preferred in treatment of NMIBC; radical cystectomy, tri-modality therapy with neoadjuvant chemotherapy offers the best chance for treatment of MIBC. Despite the new treatments and improvements on health field, molecular basis of bladder cancer remains quite unknown. To elucidate the molecular basis of BC, several aspects have to be combined well such as somatic mutations, mRNA and miRNA expression data, DNA methylation, PTMs levels, clinical correlations and etc.

As similar to other researches in the field of cancer, our aim is to identify the biomarkers for early diagnosis of BC, and also to reveal the specific agent for the effective determination of BC. For this purpose, we perform functional gene enrichment to expand the gene pool for BC as identification of 22 genes that proven to be significant in BC with text mining [1] in the first step, and then identify related miRNAs through miRWalk in the second step. Prior to the finalization of our gene pool, we have also mined the associated genes with the list of miRNAs coming from miRWalk. Among many genes in our extended pool, we select the ones that counted at least 4 times. Then, gene-gene interaction networks could be created in the STRING to reveal their relationship with others, in medium confidence level. There are only 9 genes out of 33 in connections within 3 different clusters. For remaining ones, we couldn't identify any connections. We constructed a scheme to show these 9 gene's shared miRNAs to be used for further studies.

In general, mRNA expression data was used to subtype the BC [2] and helps us to elucidate the mechanism behind BC, especially from the point of gene-gene interactions that has had a potential to provide an insight for development of biomarkers and important pathways have to be targeted. For this purpose, we used functional gene enrichments approach to expand gene pool for bladder cancer. First, we identified 22 genes that proven to be have functional activities in bladder cancer, from literature [1] and found associated miRNAs through miRWalk platform. Then, identified miRNAs again was put on miRWalk to find

their related genes. With the unification of all these, our final gene pool was created, and genes counted at least 4 times in our gene pool were selected for further processes. To identify the related pathways in BC, we have put final gene list in KEGG and Reactome. Among many, Pathways in cancer, Viral carcinogenesis, PI3K-Akt signalling pathway, Human papillomavirus infection pathway, Hepatitis B and MAPK signalling pathways were hit most in both KEGG and Reactome with CDK2, CREB1, MAPK1, TP53, CBL, IGF2, SP1and MAP3K9 genes. Next, we took RNA expression data of BC with GSE13507 code and prioritized them according p-values just before running pathfindR. Among more than 110 pathways, we selected pathways with Fold Enrichment more than 75 and then upregulated and downregulated genes of each pathway displayed. We believe that our study is pioneer to find new potential biomarkers for BC and a good template for any other studies that will provide new knowledge about gene-miRNA interaction and their role in BC related to their expression levels.

Keywords: Bladder Cancer, Gene Enrichment, MiRNA-Gene Relationship, Pathfindr

References:[1]: Pichler, R., Fritz, J., Tulchiner, G., Klinglmair, G., Soleiman, A., & Horninger, W. et al. (2017),"Increased accuracy of a novel mRNA-based urine test for bladder cancer surveillance", BJU International,121(1), 29-37. PubMed PMID: 28941000
[2]: Choi W, Ochoa A, McConkey DJ, Aine M, Höglund M, Kim WY, Real FX, Kiltie AE, Milsom I,Dyrskjøt L, Lerner SP. Genetic Alterations in the Molecular Subtypes of Bladder Cancer: Illustration in theCancer Genome Atlas Dataset. Eur Urol. 2017 Sep;72(3):354-365. doi: 10.1016/j.eururo.2017.03.010. Epub2017 Mar 30. PubMed PMID: 28365159; PubMed Central PMCID: PMC5764190.

Corresponding Author's Address: Biomedical Engineering, Faculty of Engineering and Natural Sciences, Biruni University, İstanbul - ayenenler@biruni.edu.tr

DEEP AND SHALLOW CHEMOGENOMIC MODELLING FOR COMPOUND-TARGET BINDING AFFINITY PREDICTION

Heval Ataş[1], Ahmet Rifaioğlu[2], Tunca Doğan[1,3,] 4, Maria Martin[4], Rengül Çetin-Atalay[1], Volkan Atalay[1,2*],*

1. Cancer Systems Biology Laboratory (CanSyL), Graduate School of Informatics, Middle East Technical University , 06800 Ankara, Turkey
2. Department of Computer Engineering, Middle East Technical University , 06800 Ankara, Turkey
3.Institute of Informatics / Hacettepe University , 06800 Ankara, Turkey
4.European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL EBI), CB10 1SD Hinxton, Cambridge, UK

Machine learning techniques are frequently used in the field of drug discovery and repurposing for the prediction of interactions between drug candidate compounds and target proteins. Recently, chemogenomic modelling approaches became popular, which utilize both compound and target space in one model, so that they can be used to predict novel ligands for targets with limited or no training data. In this study, we developed two chemogenomics based computational methods, using deep (pairwise input deep neural networks -PINNs-) and shallow (random forests -RFs-) supervised learning techniques, to predict the binding affinities of a large set of kinases against several drug candidate compounds. We represented compounds with ECFP4 fingerprints and proteins with k-separated-bigram-PSSM feature vectors as a homology-based protein descriptor. For approach1, we used RF algorithm, which takes a concatenated feature vector (compound + target) as input (Figure 1.a). For approach2, we used PINN architecture, which takes a pair of feature vectors for compounds and targets from disjoint input nodes simultaneously, following two hidden processing layers, latent representation of compound and target features are concatenated and further processed on two additional hidden feed-forward layers (Figure 1.b). For both approaches, output is a single node (a regressor), which predicts binding affinity for the input compound-target pair in terms of pChEMBL values. Performance calculation and parameter optimization was carried out via cross-validation (Table 1). We participated the IDG-DREAM Drug-Kinase Binding Prediction Challenge (https://www.synapse.org/#!Synapse:syn15667962/ wiki/583305) with our models. The challenge is based on the prediction of 430 pKd values between 70 compounds and 199 kinases (round1) and 394 pKd values between 25 compounds and 207 kinases (round2). Considering the model performance results on challenge round1, our best performing model has reached an RMSE value of 1.119 (5th best team). In round2/ final round, our RMSE performance was 1.066 (4th best team, overall). The considerably high performance of our models in this challenge demonstrates the usefulness of chemogenomic approach for the computational prediction of compound-protein interactions.

Keywords: Binding Affinity Prediction; Machine Learning; Chemogenomics; Pairwise Input Deep Neural Networks; Random Forests

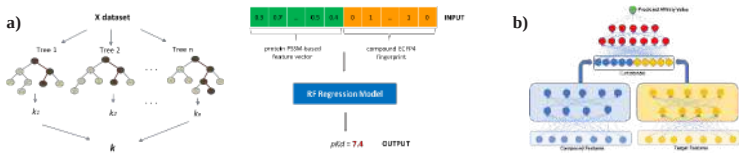


Figure 1. The representation of (a) Random forest, and (b) PINN based model architecture.

Model name	RMSE	Pearson correlation	Spearman correlation	F1-score
Model 1 (RF)	0.64	0.87	0.87	0.85
Model 2 (RF)	0.63	0.87	0.87	0.86
Model 3 (PINN)	0.73	0.72	0.65	0.65
Model 4 (PINN)	0.73	0.72	0.64	0.65

Table 1: Predictive model performance results in 5-fold cross-validation for round 1.

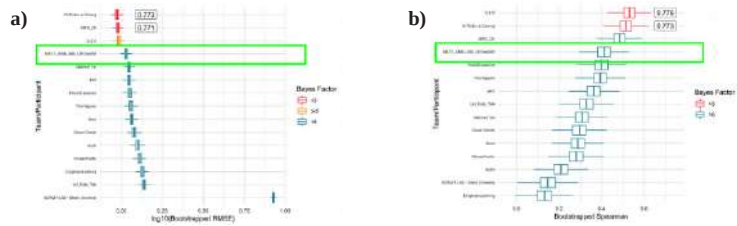


Figure 2. IDG-DREAM Drug-Kinase Binding Prediction Challenge final results. **(a)** Subchallenge1: RMSE scores (our model = 1.066); **(b)** Subchallenge2: Spearman correlation scores (our model = 0.412).

Corresponding Author's Address:

* To whom correspondence should be addressed: tuncadogan@hacettepe.edu.tr, vatalay@metu.edu.tr

BIOINFORMATIC ANALYSIS OF POLYMORPHISMS IN INTRON 7 OF PYHIN1 GENE ASSOCIATED WITH ASTHMA

Cansu Pirim[1], Cemal Ün[2]

1,2. Molecular Biology Section, Department of Biology, Faculty of Science, Ege University, Bornova,

Asthma is a chronic inflammatory disease characterized by airway obstruction due to gene-environment interactions that are believed to be mediated by epigenetic mechanisms triggered by environmental factors as well as genetic factors [1,2]. In case the prevalence of asthma is evaluated, it is known that African individuals have a higher prevalence than other ethnic groups[3]. Genome-wide association studies showed that asthma-related single nucleotide polymorphisms were identified at PYHIN1 gene in individuals of Africandescent only[4].

In this study; sequences containing two different single nucleotide polymorphisms detected in the intron 7 region of the PYHIN1 gene, were examined by bioinformatics tools in 11 different species of the Order Primates and conserved regions containing predicted miRNA sequences were identified. With reference to the conserved sites; precursor miRNA sequences with appropriate hairpin secondary structure and thermodynamic stability have gotten from human miRNA sequence. Potential mature miRNA sequences have been identified from pre-miRNA sequences via the MatureBayes program that using the Naive Bayes Classification algorithm and finally, the genes thought to be associated with asthma, targeted by these miRNAs were estimated.

Keywords: Asthma, PYHIN1, MicroRNA Detection, Target Gene Prediction, Computational Methods, Bioinformatics

References:[1] Global Initiative for Asthma (GINA). Global strategy for asthma management and prevention. Updated 2018. http://www.ginasthma.org. Accessed Aug 16, 2019. [2] Ho SM. Environmental epigenetics of asthma: an update. J Allergy Clin Immunol. 2010 Sep;126(3):453-65. PubMed PMID: 20816181

[3] Centers for Disease Control and Prevention (CDC). Asthma surveillance data. Updated March 2019. https://www.cdc.gov/asthma/asthmadata.htm. Accessed Aug 16, 2019. [4] Torgerson DG, Ampleford EJ, Chiu GY, Gauderman WJ, Gignoux CR, Graves PE, Himes BE, Levin AM, Mathias RA, Hancock DB, Baurley JW, Eng C, Stern DA, Celedón JC, Rafaels N, Capurso D, Conti DV, Roth LA, Soto-Quiros M, Togias A, Li X, Myers RA, Romieu I, Van Den Berg DJ, Hu D, Hansel NN, Hernandez RD, Israel E, Salam MT, Galanter J, Avila PC, Avila L, Rodriquez-Santana JR, Chapela R, Rodriguez-Cintron W, Diette GB, Adkinson NF, Abel RA, Ross KD, Shi M, Faruque MU, Dunston GM, Watson HR, Mantese VJ, Ezurum SC, Liang L, Ruczinski I, Ford JG, Huntsman S, Chung KF, Vora H, Li X, Calhoun WJ, Castro M, Sienra-Monge JJ, del Rio-Navarro B, Deichmann KA, Heinzmann

A, Wenzel SE, Busse WW, Gern JE, Lemanske RF Jr, Beaty TH, Bleecker ER, Raby BA, Meyers DA, London SJ; Mexico City Childhood Asthma Study (MCAAS), Gilliland FD; Children's Health Study (CHS) and HARBORS study, Burchard EG; Genetics of Asthma in Latino Americans (GALA) Study, Study of Genes-Environment and Admixture in Latino Americans (GALA2) and Study of African Americans, Asthma, Genes & Environments (SAGE), Martinez FD; Childhood Asthma Research and Education (CARE) Network, Weiss ST; Childhood Asthma Management Program (CAMP), Williams LK; Study of Asthma Phenotypes and Pharmacogenomic Interactions by Race-Ethnicity (SAPPHIRE), Barnes KC; Genetic Research on Asthma in African Diaspora (GRAAD) Study, Ober C, Nicolae DL. Meta-analysis of genome-wide association studies of asthma in ethnically diverse North American populations. Nat Genet. 2011 Jul 31;43(9):887-92. PubMed PMID: 21804549

Corresponding Author's Address:
Prof. Dr. Cemal ÜN Ege University, Faculty of Science

MOLECULAR MODELLING OF HYDROPHOBIC INTERACTIONS BETWEEN AMINO ACIDS

Seda GEZER[1], *Cenk SELÇUKİ*[2]

1.Ege University Institute for Graduate Studies in Science and Engineering
2.Ege University Faculty of Science Biochemistry Department

Hydrophobic molecules generally consist of long carbon chains and do not interact with water. Hydrophobic interactions are important for the folding of proteins. This is important for the protein to remain stable and biologically active. Because it reduces unwanted interactions of protein with water.In addition to protein folding, there are many biological phenomena due to hydrophobic interaction for the survival of the cells in our bodies. The amino acids selected in this study are; valine, leucine and isoleucine. These amino acids are essential and non-polar amino acids. Dealing with the conformation of these amino acids helps us to understand the basic properties and relative stability of these amino acids. In order to make molecular modeling of the amino acids we selected for this thesis, ωB97xD / 6-311 ++ G (d, p) level was selected. Energy calculating was done to be in vacuum and water.In order to see the interaction of amino acids with each other, the most stable structures of each amino acid we selected were selected. For molecular modeling of these structures, ωB97xD / 6-311 ++ G (d, p) level was selected. After the energy calculations were completed, a graph was created showing the interactions between them and the density of the bonds.

Keywords: Hydrophobic Interaction; Isoleucine; Leucine; Molecular Modeling; Valine.

Corresponding Author's Address:
Ege University, Faculty of Science, Biochemistry Department
cenk.selcuki@gmail.com

MISSENSE MUTATION EVALUATION WITH DEEP NEURAL NETWORK

Oguzhan KALYON[1,2], Nur Sena ULUSKAN[2], Gizem SOMUNCUOGLU[2], EmreTEPELI[2], Ahmet ULUDAG[2]

1. Yildiz Technical University, Faculty of Arts and Sciences, Department of Molecular Biology and Genetics, Istanbul, Turkey
2. Next Genetic Centre, Department of Molecular Genetic, Istanbul, Turkey

It is difficult to interpret the variations have fewer allel frequency. When a mutation, not defined in general mutation databases detected, it is usually evaluated by some features e.g. prevalence, type of mutation,result of in silico algorithm classification, etc.

Researchers at Illumina has developed an algorithm that provide predicting the clinical impact of human mutation with deep neural networks[1]. Algortihm uses lots of features such as six non-human primate species common variations and protein structure to train neural networks. Algorithm enables identify pathogenic mutation with 88% accuracy. We have used program to interpret detected mutation in 30 patient that rare genetic disease preliminary diagnosis.

The program, which we used effectively in filtering the variants during whole exome analysis (WES), provided the diagnosis of 5 in 30 patients easily. For remaining 25 patients, a different approach to determine the variations have fewer allel frequency is needed. It should also be noted that since there are no functional studies of mutations, it is recommended that patients be evaluated together with their clinic

Keywords: Missense Mutation, Deep Neural Network

References:[1] Laksshman Sundaram, Hong Gao, Samskruthi Reddy Padigepati, Jeremy F. McRae, Yanjun Li, Jack A.
Kosmicki, Nondas Fritzilas, Jörg Hakenberg, Anindita Dutta, John Shon, Jinbo Xu, Serafim Batzoglou,
Xiaolin Li, Kyle Kai-How Farh. Predicting the clinical impact of human mutation with deep neural networks.
Nat Genet. 2018 Aug; 50(8): 1161–1170. PubMed PMID: 30038395
Corresponding Author's Address:
ozyyk61@gmail.com

OPPORTUNITIES AND PROBLEMS RELATED TO NOVEL TRAINABLE PROTEIN REPRESENTATIONS

Serbulent Unsal[1,2], *Aybar C. Acar* [2], *Tunca Dogan*[3]

1. Karadeniz Technical University,
2. Middle East Technical University, 3. Hacettepe University

One of the key points for accurately predicting protein features/ properties is generating a holistic representation of proteins. Using these representations, inherent features of proteins can be learnt efficiently by a machine learning classifier. There are two main types of representations: fixed and trainable. Fixed representations are based on pre-defined rules designed by human experts, mostly inspired from the natural properties of these molecules. Trainable representations, on the other hand, are data and/or task specific, and generated based on the patterns found in the data. Lately, trainable representations are getting popularity in the life-sciences domain.

In this study, we aim to investigate the potential of trainable embeddings for protein-function-prediction, especially for the prediction of ontological terms with low number of training instances. For this, we mainly considered novel trainable protein and molecular data representation approaches such as word2vec/doc2vec-based methods inspired by the NLP field and methods from neural network encoding, all of which reported significant improvements over the state-of-the-art in different predictive tasks related to protein science in recent literature. We classified these representation models according to their technical aspects and their objectives, and we discussed our results and propose new directions for protein representation construction.

Key points inferred and exemplified in this study can be summarized as:

• **Importance of representations:** Representing complex structures as numerical entities, such as vectors, is crucial for any computational task.. The tests showed that the representation itself is significantly more important compared to the machine learning technique used, when it comes to predicting the properties of proteins, due to the complexity of biological entities. Hence, exploring trainable, data-driven representations is crucial for any computational protein analysis.

• **Interpretability problem:** One of the key problems in machine learning is interpretability of models. A trainable protein representation should be generated in such a way that it can be interpreted by domain experts to enlighten latent features. Moreover, such a representation should allow the reuse for tasks different than the one it was created for.

• **Representation stability problem:** In stable semantic representations, the vector arithmetic should give consistent results (e.g. king - man + woman ~= queen) and similar entities (i.e. k-mers) should have similar vectors. A problem in protein representations is the dramatic value changes in the resulting vectors when a small perturbation is introduced to the system (e.g., the removal or addition

of a few sequences to/from the whole protein dataset) or with small parameter value changes in the embedding process. Currently, there are a limited number of benchmarks/tests for assessing the stability of available protein/gene representations. We are actively working on potential solutions for this problem.

• **Baseline comparison problem:** In most studies, proposed trainable protein representations are being evaluated via ablation studies and by comparing with very similar representations. However, a few number of studies showed that classical representations that are accepted to be baseline models (e.g., TF-IDF) can be more successful compared to novel trainable representations. Because of this, a standardization, regarding baseline methods, is required.

• **Benchmarking of representations:** Although the number of alternative protein representations is growing rapidly in the literature, there is no independent benchmark for their evaluation. A benchmarking pipeline is proposed in this study.

Keywords: Protein Representation; Deep Learning; Learning Representations

Corresponding Author's Address: Tunca Dogan, tuncadogan@gmail.com, Health Informatics Program, Institute of Informatics, Hacettepe University, 06800 Ankara, Turkey

TAXONOMIC METRIC LEARNING FOR 16S RRNA TAXONOMIC CLASSIFICATION

Samed SAKA[1,2], *Özkan Ufuk NALBANTOĞLU*[2,3]

1. Erciyes University, Graduate School of Health Sciences, Bioinfomatics Systems Biology Deparment, Kayseri
2. Erciyes University, Genome and Stem Cell Center (GenKök), Kayseri
3. Erciyes University, Faculty of Engineering, Computer Engineering Department, Kayseri

In microbiome studies, high-throughput sequencing of the 16S rRNA gene is a commonly used method for identifying the microbial composition [1]. 16S rRNA gene allows classification by its evolutionarily conserved and variable regions. Due to the phylogenetic signals contained in the gene sequence, the sequences can be classified taxonomically, consequently revealing the biodiversity profile of a microbiota sample. Classification methods used in natural language processing (e.g. Naïve Bayesian Classifiers [1] or Random Forest classifiers on bag-of-words, or k-mers for DNA sequences specifically) are usually adopted thanks to the close nature of the problems. Although this approach can result in accurate detection while close relatives of a test sample are contained in the database, remote relatives are often placed to taxonomically unrelated units, distorting the final taxonomic composition significantly. This problem frequently can cause defected microbiota analyses as prokaryotic biodiversity is mostly unexplored with incomplete taxonomic databases. In order to address the problem, and being able to accurately detect remote relatives of taxa of unexplored ranks, we propose an auxiliary learning problem where 16S rRNA sequences are embedded to a feature space, where metric distances are phylogenetically (or at least at taxonomical quanta) meaningful. Once embedding 16S rRNA genes into this space is achieved, the classification to close or remote relatives becomes staightforward, proposing a solution to the underrepresented phylogeny problem.

At first, we created a dataset from SILVA [2] which is database for 16s rRNA sequences.The dataset contained 13880 samples belonging to 694 genera, 199 families, 113 orders, 95 classes, and 35 phyla hierachially. For each hierachial clade level from genus to class, we have created training and test datasets excluding the close taxa from the training in order to simulate the unexplored close relatives problem. Taxonomic similarity matrices on the training data were constructed as follows. Each pair of samples attained a score of similarity based on their taxonomic similarity, incrementing from the ones in the same domain but different phyla, up to the ones sharing the same genus. Therefore, the a gram matrix locating each 16S rRNA sequence implicitly to a taxonomically meaningful metric space was defined. A neural network embedding the pentanucleotide frequency vectors of each gene to this space explicitly via satisfying the similarity metrics of each given pair was trained using Siamese neural network architecture. After the training, one of the twin networks was adopted as the taxonomic embedding network. For a

test sample, the corresponding represantation vector (i.e. the output of the embedding neural net) was assigned to the taxonomic label of the nearest training neighbor in this space, expecting that even if the same taxonomic class has not been seen/trained before, the nearest neighbor should be a taxonomic relative in a higher clade.

Independent experiments for different clade levels from genus to class were performed, and the classification accuracies were compared to the state-of-the-art conventional 16S rRNA classification methodology, which is a Random Forest classifier based algorithm [3] using the same features and conditions (i.e. run on pentanucleotide profiles with the same experimental train/test sets). Table 1 shows the summary of the results suggesting that proposed taxonomic embedding methodology is significantly outperforms the current convention on microbial taxonomic classification.

The results claims that deep learning based methodologies learning biologically informed structures might result in better inference on taxonomic or phylogenetic prediction tasks. Moreover, in practice the performance of microbiota analysis tools can bee boosted following the proposed framework.

	Taxonomic rank			
	Genus	**Family**	**Order**	**Class**
Random Forest	0.894	0.757	0.358	0.480
Proposed	0.955	0.896	0.670	0.724

Table 1. Classification accuracy performances of the proposed and the reference methods. Independent experiments for each clade level is conducted by removing the samples of the same taxonomic class from the training set for a corresponding test sample.

Keywords: Microbiome, 16S rRNA Classification, Artificial Neural Networks

References: [1] Q. Wang, G. M. Garrity, J. M. Tiedje, and J. R. Cole, "Naïve Bayesian classifier for rapid assignment of
rRNA sequences into the new bacterial taxonomy," Appl. Environ. Microbiol., 2007.
[2] C. Quast et al., "The SILVA ribosomal RNA gene database project: Improved data processing and web-
based tools," Nucleic Acids Res., 2013.
[3] Chaudhary, Nikhil, et al. "16S classifier: a tool for fast and accurate taxonomic classification of 16S
rRNA hypervariable regions in metagenomic datasets." PloS one 10.2 (2015): e0116106.

Corresponding Author's Address: Erciyes University, Genome and Stem Cell Center (GENKÖK), Kayseri sakasamed@outlook.com

PHYLOGENETIC ANALYSIS OF CALCIUM-SENSING RECEPTORS

Aylin Bircan[1], *Ogün Adebali*[1]

1. Molecular Biology, Genetics and Bioengineering Program, Faculty of Engineering and Natural Sciences, Sabancı University, Istanbul 34956 Turkey

The calcium-sensing receptor (CaSR) is a member of class C G protein-coupled receptors which is composed of a large bilobed extracellular domain called Venus Flytrap domain with a cystine rich domain, seven transmembrane domain with three intracellular and three extracellular loops [1]. CaSR functions by forming a homodimer and couples with different G proteins to activate different intracellular signaling pathways [2]. CaSR mainly detects even smallest changes in the extracellular calcium level and different germline loss- and gain-of-function CaSR mutations cause different diseases such as familial hypocalciuric hypercalcaemia and autosomal dominant hypocalcemia [3]. Studies on these disease-causing mutations show that specific residues on different domains of CaSR are important for receptor activation and function [3], however, the effect of each amino acid substitution at a certain position on the receptor function and diseases is not fully understood. Here we aim to establish the precise evolution history of CaSR to reveal the evolutionary significance of each amino acid to understand how changes in sequence modulate the receptor function during evolution.

Keywords: Calcium-Sensing Receptor; G Protein-Coupled Receptors; Casr Mutations

References: [1] Mayr B, Glaudo M, Schofl C. Activating Calcium-Sensing Receptor Mutations: Prospects for Future Treatment with Calcilytics. Trends Endocrinol Metab. 2016;27(9):643-52. Epub 2016/06/25. doi: 10.1016/j.tem.2016.05.005. PubMed PMID: 27339034.

[2] Wootten D, Christopoulos A, Marti-Solano M, Babu MM, Sexton PM. Mechanisms of signalling and biased agonism in G protein-coupled receptors. Nat Rev Mol Cell Biol. 2018;19(10):638-53. Epub 2018/08/15. doi: 10.1038/s41580-018-0049-3. PubMed PMID: 30104700.

[3] Gorvin C. Molecular and clinical insights from studies of calcium-sensing receptor mutations. J Mol Endocrinol. 2019. Epub 2019/06/13. doi: 10.1530/JME-19-0104. PubMed PMID: 31189130.

Corresponding Author's Address: To whom correspondence may be addressed: Sabancı University, Faculty of Engineering and Natural Sciences, Tuzla, Istanbul, 34956 Turkey Tel.: 216-568-7043; E-mail: oadebali@sabanciuniv.edu.

REVIEW OF DRUG REPOSITIONING WITH NETWORK ANALYSIS IN CANCER

Ülkü ÜNSAL[1], Ali CÜVİTOĞLU[2], Zerrin IŞIK[2], Kemal TURHAN[1]

1. Karadeniz Technical University, Department of Biostatistics and Health Informatics
2. Dokuz Eylul University, Department of Computer Engineering

Drug developing is a complicated, time-consuming (approximately 13 years per drug), high-priced (average $12 billion per drug) and a risky investment. For these reasons, drugs are designed for binding multiple targets and eventually can be used in treatment of different diseases. In view of these findings, computational drug repositioning approaches were proposed by using bioinformatics analysis tools. Using drug repositioning, drug development can be completed in a shorter time (average 8 years), less costly (mean $1.6 billion) and less risky. There are three main approaches for drug repositioning: computational, biological experiment and mixed approaches. One of the computational methods is the network-based approach which uses physical relationship between two proteins and functional similarities between genes to discover a new usage for a known drug/compound. In network-based approach, network content representing different hypotheses including protein-protein interaction networks, drug-target / drug -drug / drug -disease / drug-side effect relationships or disease-disease relationships have been established In this study, network-based drug repositioning studies in cancer area are reviewed. Key problems and opportunities in this field are summarized to guide researchers for further studies.

Keywords: Drug Repositioning; Network Biology; Computational Approach

Corresponding Author's Address:
Ülkü ÜNSAL Karadeniz Technical University Department of Biostatistics and Health Informatics ulkunsal@gmail.com

GWAS ON VEILLONELLA PARVULA STAINS COLONIZED IN HUMAN GUT REVEALS GENETIC VARIATIONS ASSOCATED WITH LIVER CIRRHOSIS

Kübra YALÇIN [1,2], *Özkan Ufuk NALBANTOĞLU*[2,3]

1.Erciyes University, Graduate School of Health Sciences, Bioinfomatics Systems Biology Deparment, Kayseri
2.Erciyes University, Genome and Stem Cell Center (GenKök), Kayseri
3Erciyes University, Faculty of Engineering, Computer Engineering Department, Kayseri

Human gut microbiome consists of thousands of microbial species, harboring several millions of genes actively involving in crucial host-commensal metabolic and signalling interactions. During the last decade, we have come to the knowledge that the gut microbiome is altered, either as pathogenesis or complication factors, in case of several complex diseases. Recent studies1,2 have shown that these alterations may appear as taxonomic diversity and relative abundance of certain organisms. Yet, accumulating evidence show that only taxonomic analyses are no longer sufficient to reveal the full spectrum of variations[3]. It is anticipated that genomic variations of microorganisms should be studied along with the taxonomic and functional changes. Therefore, in this study, in addition to the existing taxonomic analyses previously performed for liver cirrhosis microbiome, Single Nucleotide Polimorphism (SNP) analysis on a Veillonella parvula was performed. V. parvula is a common human gut commensal that is hypothesized to be a significant biomarker of liver cirrhosis. Firstly, liver cirrhosis metagenome data1 were obtained. There were a total of 314 samples, including 169 liver cirrhosis patients and 145 healthy controls. V. parvula strain UTDB1-3 (NCBI accession number CP019721.1) was used as the reference. The RAST tool was used for anotation of the Veillonella parvula genome. First of all, low quality and low abundance samples were excluded from the analysis and 152 liver liver cirrhosis 136 healthy controls and a total of 288 samples were analyzed. Using BWA (Burrows-Wheeler Aligner) - MEM, these 288 metagenomes were aligned against V. parvula. Then vcf files created with samtools mpileup which utility provides a summary of the coverage of mapped reads on a reference sequence at a single base pair resolution. To keep the SNP quality high, low quality SNPs were filtered and as a result, a total of 1,985,732 SNPs were detected. The SNP distance matrix can be seen in **Figure1.** The visible segmentation on the heatmap implies that the V. parvula strains are clustered by disease phenotype on the cirrhosis-healthy control cohort.

Machine learning was performed by using gradient boosted trees algorithm over SNP profiles. Important SNPs were determined by selecting features on the model created. Approximately 0.82 AUC of ROC curve value was obtained on the test data using Xgboost algorithm. A total of 68 SNPs were selected after feature selection on the trained model. In conclusion, our findings suggest that genetic

variations of V. parvula in human gut is significantly associated with liver cirrhosis raising potential applications of non-invasive diagnosis or novel theuropathic drug targets for the liver disease.

Keywords: Metagenome; Microbiome; Liver cirrhosis; SNP

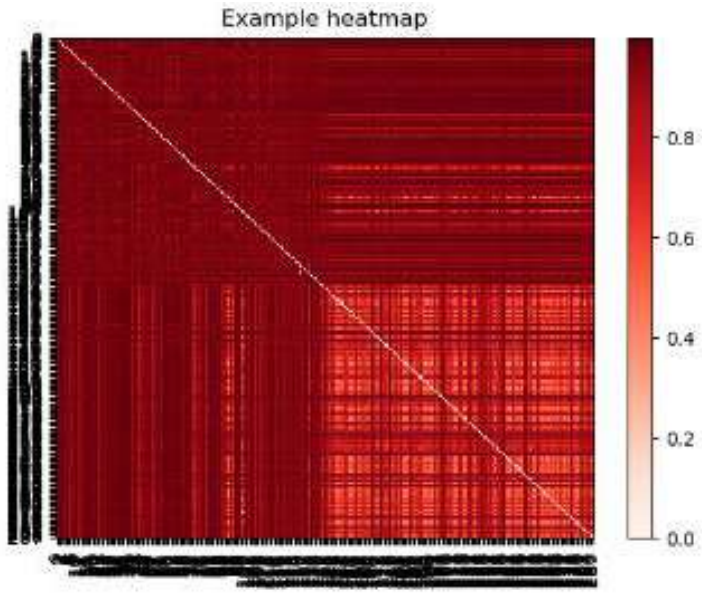


Figure 1. SNP distance heatmap for V. parvula colonized in healthy subjects and subjects with cirrhosis.

References: [1] Qin, N., et al. (2014). Alterations of the human gut microbiome in liver liver cirrhosis. Nature. 10.1038/nature13568. PMID: 25079328
[2] Qin, J., et al. (2012). A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature. 490. 55-60. 10.1038/nature11450. PMID: 23023125
[3] Chen Y., et al. (2017). Gut metagenomes of type 2 diabetic patients have characteristic singlenucleotide polymorphism distribution in Bacteroides Coprocola. Microbiome. 5. 10.1186/s40168-017-0232-3.

Corresponding Author's Address: Erciyes University, Genome and Stem Cell Center (GENKÖK), Kayseri
Kübra Yalçın yalcin.kub@gmail.com

CANSLPRED: A MULTI-VIEW METHOD TO SUBCELLULAR LOCALIZATION PREDICTION OF HUMAN PROTEINS

Gökhan Özsarı[1], Ahmet Süreyya Rifaioğlu[1], Tunca Doğan[2,4], Rengül Çetin-Atalay[3], Volkan Atalay[1]

1. Department of Computer Engineering, Middle East Technical University, Ankara, Turkey, 2. European Bioinformatics Institute (EMBL-EBI), Hinxton, Cambridge, CB10 1SD, UK, 3.CanSyL, Graduate School of Informatics, Middle East Technical University, Ankara, 06800, Turkey,
4. Department of Computer Engineering, Hacettepe University, Ankara, Turkey,

Determining the subcellular localization of proteins is crucial for understanding the functions of proteins, drug targeting, systems biology, and proteomics research. The subcellular localization of proteins can experimentally be identified by purification or imaging methods which are expensive and time-consuming. Therefore, several computational methods for automated prediction of protein subcellular localization are proposed in the last two decades; however, there is still room for better performance. Here, we introduce a multi-view classification method (CanSLPred) that provides subcellular localization predictions for human proteins. In the proposed multi-view approach, we employ seven feature-based probabilistic prediction models that provide seven distinct representations (protein descriptors) and seven probabilistic predictions for each protein sequence. There are three major contributions in this study: First, a new subcellular location hierarchy is introduced by merging Universal Protein Knowledge Base (UniProtKB) Subcellular Location (SL) hierarchy and Gene Ontology (GO) Cellular Component (CC) hierarchy. Second, a dataset of protein sequences is generated by taking the proteins whose subcellular localization is experimentally annotated in UniProtKB/SwissProt and applying the new SL hierarchy to propagate the proteins according to their subcellular localization. This dataset is called Trust dataset. Third, a new classification method is described to predict the subcellular localization of human proteins by employing a weighted mean voting multi-view Support Vector Machine (SVM) approach. The new subcellular location hierarchy is formed to unite the characteristics of both hierarchies: UniProtKB SL hierarchy and GO CC hierarchy. To generate the new SL hierarchy, UniProtKB SL identifiers are first mapped to GO CC terms. GO CC hierarchy is then extracted by considering 'is_a' relations among GO CC terms in GO hierarchy. The mapping of UniProtKB SL identifiers to GO CC terms is applied at the end. CanSLPred consists of nine independently constructed classification models where each model provides predictions for one of nine subcellular locations: cytoplasm (CYT), nucleus (NUC), cell membrane (MEM), mitochondrion (MIT), endoplasmic reticulum (ERE), secreted (EXC), Golgi apparatus (GLG), lysosome (LYS), and peroxisome (PEX). The classification models are developed by considering the subcellular localization problem as

a binary classification problem where each of the models decides if a protein localizes to the corresponding subcellular location or not. The prediction is given in four steps: Feature extraction and normalization, prediction by probabilistic models, weighted-mean voting, and thresholding. In the feature extraction process, seven protein descriptors are selected out of 160 cases (40 descriptors from three tools: iFeature, POSSUM, SPMap and 4 normalization methods: Standard normalization, MinMax normalization, Robust scaler, Power transformation), where these seven protein descriptors contribute the best in the combination of probabilistic prediction models. SVM is used to construct probabilistic prediction models, which produces probabilistic scores indicating the localization probability for a query protein sequence. A weighted score is calculated based on the obtained probabilistic scores from seven feature-based probabilistic prediction models (SVMs) by employing weighted mean voting. Binary prediction is given by applying thresholding on the weighted score. We evaluate CanSLPred by using three datasets: Trust-Test dataset (our in-house dataset), Golden dataset (SubCons' benchmark dataset), and Golden-Trust dataset (a refined version of Golden dataset). Trust dataset is created by applying the new SL hierarchy on the proteins whose subcellular localization is experimentally annotated in UniProtKB/SwissProt. Golden dataset consists of protein sequences whose subcellular localization is experimentally annotated in at least two out of three protein resources: mass spectrometry (Mass-Spec), SLHPA, and UniProtKB. Golden-Trust dataset is a refined version of Golden dataset where the steps we follow to generate Trust dataset are applied for the protein sequences in Golden dataset. We compare the results of CanSLPred with five state-of-the-art methods: SubCons [1], LocTree2 [1], CELLO2.5 [1], MultiLoc2 [1] and DeepLoc. CanSLPred outperforms the others with 59%, 68%, 61% overall Matthews correlation coefficient (MCC) scores on Trust-Test dataset, Golden-Trust dataset, Golden dataset, respectively where SubCons' overall MCC scores are 43%, 53%, and 56%.

Keywords: Subcellular Localization, Prediction, Human Proteins, SVM, Multi-View

References: [1] Salvatore M, Warholm P, Shu N, Basile W, Elofsson A. SubCons: a new ensemble method for improved human subcellular localization predictions. Bioinformatics. 2017;33(16):2464–70.

Corresponding Author's Address: gozsari@ceng.metu.edu.tr

HONESTY IN COMMUNICATION – GAME THEORETICAL ANALYSIS OF BACTERIAL SIGNALING

Özgür Yüksel, Emrah Nikerel

Yeditepe University, Department of Genetics and Bioengineering

Cooperation and competition play crucial role in every scale of social interactions. Cooperative behavior often increases the overall fitness of the population whereas individuals compete for limited resources [1]. In microscopic scale, bacteria form cooperative communities embedded in extracellular polymeric substances (EPS) called biofilms. Trough cell-cell communication biofilm community coordinate behavior and regulate the expression of surface molecules, antibiotic resistance, nutrient utilization and virulence factors [2]. Particularly, bacteria communicate and coordinate metabolism through long-range electrical signals within the biofilm [3]. In microscopic communities, competitive behaviors are often difficult to observe whereas experimental studies readily reveal cooperation. In order to identify competitive behaviors, individual's strategy and its stability in dynamic environments must be taken into consideration. In an explicit manner, Evolutionary Game Theory (EGT) analyzes evolutionary stability of a strategy and specifies dynamics for the population [4].

The aim of this study is to analyze cooperation and competition in biofilm based electrical communication and to calculate frequency dependent fitness of strategies with the EGT approach. The complex communication phenomenon issimplified in Lewis signaling terms where the game is played between a sender and a receiver [5]. The cellular and molecular processes are executed via cellular automata in 2-D lattice space where an individual cell only interacts with immediate neighbors [6]. The model demonstrates spatial propagation of electrical signal and maintenance of metabolic coordination among distant cells. The strategies prevail depending on the initial spatial configurations of each cell-type.

Keywords: Electrical Signals, Evolutionary Game Theory, Lewis Signaling Game, Cellular Automata

References: [1] Liu J, Prindle A, Humphries J, Gabalda-Sagarra M, Asally M, Lee DY, Ly S, Garcia-Ojalvo J, Süel GM. Metabolic co-dependence gives rise to collective oscillations within biofilms. Nature. 2015 Jul 30;523(7562):550-4. doi: 10.1038/nature14660. Epub 2015 Jul 22. PubMed PMID: 26200335; PubMed Central PMCID: PMC4862617.
[2] Hall-Stoodley L, Stoodley P. Evolving concepts in biofilm infections. Cellular Microbiology. 2009;11(7):1034-1043.
[3] Prindle A, Liu J, Asally M, Ly S, Garcia-Ojalvo J, Süel GM. Ion channels enable electrical communication in bacterial communities. Nature. 2015 Nov 5;527(7576):59-63. doi: 10.1038/nature15709. Epub 2015 Oct 21. PubMed PMID: 26503040; PubMed Central PMCID: PMC4890463.
[4] J. McKenzie A. Evolutionary Game Theory (Stanford Encyclopedia

of Philosophy) [Internet].Plato.stanford.edu. 2019 [cited 2019 Aug 18]. Available from: https://plato.stanford.edu/entries/game-evolutionary/
[5] Huttegger S, Skyrms B, Smead R, Zollman K. Evolutionary dynamics of Lewis signaling games: signaling systems vs. partial pooling. Synthese. 2009;172(1):177-191.
[6] Wolfram S. Statistical mechanics of cellular automata. Reviews of Modern Physics. 1983;55(3):601-644.

Corresponding Author's Address:
Yeditepe University, Department of Genetics and Bioengineering, 26 Ağustos Yerleşimi, Kayışdağı Cad, 34755 Ataşehir, İstanbul e-mail: emrah.nikerel@yeditepe.edu.tr

INTEGRATIVE ANALYSIS OF TRANSCRIPTOME DATA AND CELLULAR NETWORKS IDENTIFIES MOLECULAR INTERACTIONS OF METASTASIS MECHANISMS IN CANCER

Dilara Uzuner[1], Pınar Pir[1], Devrim Gözüaçık[2], Tunahan Çakır[1]

1 Department of Bioengineering,Gebze Technical University, Kocaeli.
2 Department of Molecular Biology, Genetics and Bioengineering, Sabancı University, İstanbul.

Cancer is one of the common causes of death in the 21st century. The cause of most of the cancer-related deaths is metastasis, which is the spreading of cancer cells to different tissues. Although significant advances have been made in clinical oncology, metastasis remains a major challenge in cancer therapy. Before metastatic cancer cells get cancerous at a new tissue, they stop proliferation for a while to adapt to that tissue. Traditional chemotherapies kill proliferative cells, therefore non-proliferative metastatic cancer cells can escape from chemotherapy. Because pre-metastatic cancer cells have resistance to treatment, cancer can relapse in patients years after the removal of the primary tumor. Therefore, understanding the underlying mechanisms of metastasis is crucial for efficient cancer treatment. Recent studies show that microenvironmental factors of the target tissues such as hypoxia and content of extracellular matrix are important for the formation of this premetastatic niche. However, the exact mechanism is still unknown [1].

Transcriptomics is a commonly used genome-wide approach to elucidate the mechanism behind the formation of pre-metastatic cells. Although there are multiple computational approaches to process transcriptomic data, network-based analyses are among the most promising ones to discover unknown molecular mechanisms by mapping transcriptome data on molecular interaction networks. Two network-based analysis tools KeyPathwayMiner and BioNet have been reported to have better performance than other tools in the literature [2,4]. Using different algorithms, KeyPathwayMiner and BioNet map transcriptome data on genome-wide interaction networks to identify subnetworks that are enriched with genes that are significantly changed between the compared conditions [2,3]. The aim of this study is using these computational tools to reveal the mechanisms of metastasis by identifying crucial genes and proteins and their interactions that have a role in metastasis. In this study, transcriptome data of cancer samples available in the literature were integrated with protein-protein interaction networks and gene- regulatory networks to reveal functional subnetworks and elucidate the survival mechanisms of pre-metastatic cancer cells in human and mice. This study was financially supported through a grant by TUBITAK (Project Code: 216S489)

Keywords: Cancer, Metastasis, Protein-Protein Interaction Networks, Gene-Regulatory Networks, Transcriptome

References: 1. Sleeman JP. The metastatic niche and stromal progression. Cancer Metastasis Rev. 2012;31(3-4):429-40. Epub 2012/06/16. doi: 10.1007/s10555-012-9373-9. PubMed PMID: 22699312; PMCID: PMC3470821.
2. Alcaraz N, Friedrich T, Kotzing T, Krohmer A, Muller J, Pauling J, Baumbach J. Efficient key pathway mining: combining networks and OMICS data. Integr Biol (Camb). 2012;4(7):756-64. Epub 2012/02/23. doi: 10.1039/c2ib00133k. PubMed PMID: 22353882.
3. Beisser D, Klau GW, Dandekar T, Muller T, Dittrich MT. BioNet: an R-Package for the functional analysis of biological networks. Bioinformatics. 2010;26(8):1129-30. Epub 2010/03/02. doi: 10.1093/bioinformatics/btq089. PubMed PMID: 20189939.
4. Batra R, Alcaraz N, Gitzhofer K, Pauling J, Ditzel HJ, Hellmuth M, Baumbach J, List M. On the performance of de novo pathway enrichment. NPJ Syst Biol Appl. 2017;3:6. Epub 2017/06/27. doi: 10.1038/s41540-017-0007-2. PubMed PMID: 28649433; PMCID: PMC5445589.

Corresponding Author's Address: Tunahan CAKIR, Assoc. Prof. Department of Bioengineering Gebze Technical University Kocaeli, TURKEY tcakir@gtu.edu.tr

IN SILICO DETECTION OF PUTATIVE MIRNA IN HUMAN GUT MICROBIOME

Ayşenur SOYTÜRK[1,2], Aycan GÜNDOĞDU[2,4], Özkan Ufuk NALBANTOĞLU[2,3]

1. Erciyes University, Graduate School of Health Sciences, Bioinfomatics Systems Biology Deparment, Kayseri
2. Erciyes University, Genome and Stem Cell Center (GenKök), Kayseri
3. Erciyes University, Faculty of Engineering, Computer Engineering Department, Kayseri
4. Erciyes University, Faculty of Medicine, Department of Medical Microbiology, Kayseri

MicroRNA (miRNA) are small noncoding RNA fragments of about 19-25 nucleotides in length. miRNAs have important regulatory functions. Therefore it is important to identify miRNAs and the target genes regulated by them in the context of health and disease [1]. Whether bacterial miRNAs harbored in human microbiome affect human cells is also a new and important research interest [2]. The miRNA structure and the multiplicity of potential targets make experimental estimation difficult and economically unfavorable.

The aim of this study is to determine the presence of structures similar to human miRNA in human intestinal microbiota by developing an in silico approach mining the gut metagenomes. In this study, pre-miRNAs, which constitute the first structure of 1917 human miRNAs in Mirbase, all of which have experimentally proven to be miRNAs, were identified and used as a training set. A miRNA detector based on sequence features was developed. Several classification methods (Random forests, Support Vector Machines, Gradient boosted trees, k-Nearest Neighbors) have been experimented for the miRNA classification process and support vector machine (SVM) has given the best accuracy value. The precision recall graph for the SVM classifier method is shown in Figure 1. Based on this characteristics, a threshold of detection (0.98) was determined for a moderately sensitive, yet highly specific detector. Whole genome sequencing (shotgun, assembled after filtering human genome contamination) data were obtained from EBI Metagenomics - EMBL-EBI from human feces associated with cirrhosis, which is common worldwide as a terminal liver disease. As a subset of these data, 10 human metagenomes were screened according to the determined threshold value. In these scans, 5 healthy people and 5 people with cirrhosis were considered. Screening was performed on 70 nucleotide running windows considering the pre-miRNA size of 70-110 nucleotides, which governs the secondary structure of miRNAs. The microbial contigs exceeding the threshold value were located and the species harboring that loci were determined using Kraken2 microbial taxonomic assignment method.

As a result, in the group of 5 cirrhosis patients, 1141 different microbial species were found to contain miRNAsequences similar to human miRNA in different numbers and locations. Similarly, 5 healthy human- generated groups showed similar structures to human miRNA

in 1793 different species. Student's t-test was used to determine whether there was a significant difference for the number of putative miRNA for each species between the groups. The most significant species were identified as Streptococcus pluranimalium p=0.043, Bacillus sp p = 0.036, Limnochorda pilosa p = 0.048, Cellulophaga lytica p = 0.024 and the most significant genus were identified as Bacteroidia p = 0.047, Actinomyces p = 0.045, Nocardioidaceae p=0.043.

The study suggests that human gut microbiome contains abundant number of sequences resembling human miRNA. These sequences show up in a broad biodiversity spectrum. Moreover, in the context of disease, certain species exhibit variation in the putative miRNA sequences they harbor. Further experimental validations should be carried out in order to reveal to what extent these sequences are expressed and if they function in regulating human gene expression. In case of such validations, a new dimension in host- microbiome interactions will be in question, as well as a practical opportunity in the detection/treatment of disease with novel potential biomarkers

Keywords: MicroRNA; Bacterial MicroRNA; Metagenome; Host-Microbiome Interaction

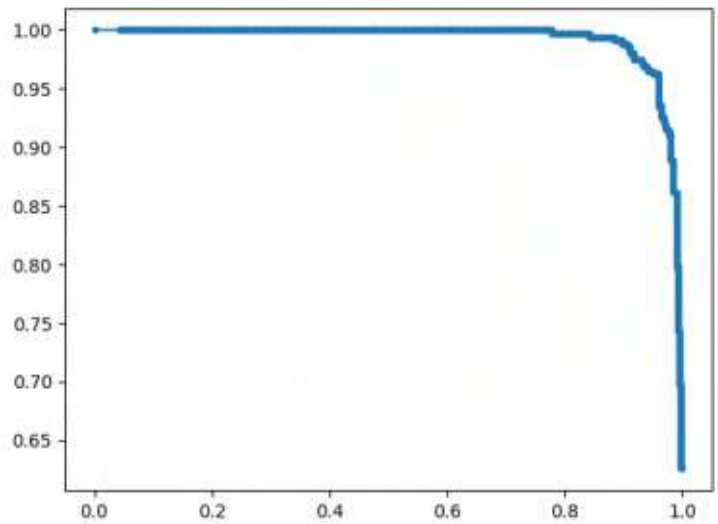


Figure 1. Precision recall graph for support vector machine classifier

References: [1] Dong Yue, Hui Liu and Yufei Huang , "Survey of Computational Algorithms for MicroRNA Target
Prediction", Curr Genomics. 2009 Nov; 10(7): 478–492.
[2] Shmaryahu A, Carrasco M and Valenzuela PD, "Prediction of bacterial microRNAs and possible targets
in human cell transcriptome.", J Microbiol. 2014 Jun;52(6):482-489
Corresponding Author's Address: Erciyes University, Genome and Stem Cell Center (GENKÖK), Kayseri

INVASIVENESS-RELATED NOVEL GENOMIC BIOMARKERS FOR PROGNOSIS OF NON-FUNCTIONING PITUITARY ADENOMAS VIA DIFFERENTIAL CO-EXPRESSION NETWORK ANALYSIS

Busra Aydin[1], *Kazim Yalcin Arga*[1]

1. Marmara University, Department of Bioengineering, Systems Biomedicine Laboratory, ISTANBUL

Non-functioning pituitary adenomas (NFPAs) are the most common intracranial tumors in the central nervous system with clinically challenging features. They are known as benign in general but somehow tumors can exhibit gross invasion into surrounding tissues rarely. Invasion eventuates resistance to conventional treatment methods and leading to early and frequent recurrences. Revealing the multi-layered molecular mechanism lay behind the invasion of PA, there is a great need for omics-level data and their integration into meta-analyses. Differential co-expression network analysis is an outstanding approach for elucidation of groups of genes which show distinct co-expression patterns among phenotypes. In this study, we carried out a differential co- expression network analysis of NFPA-associated transcriptome dataset (n = 40) considering invasive (n = 22) and non-invasive (n = 18) phenotypes. We identified differentially co-expressed and co-regulated mRNA modules, which might be considered as potential systems biomarkers for NFPA prognosis and invasiveness. As a result, we have identified a novel 13-gene module, including CEACAM6, CYP4B1, EIF2S2, HID1, IFFO1, MYO18A, PDCD2, SGIP1, SWSAP1, and four unknown genes (A_24_P127621, A_24_P255786, A_24_P683553, and A_24_P916979), which was able to categorize the patients into two groups as invasive and non-invasive NFPA with distinct prognosis [1] . GATA1, GATA2, ETS1, ESR1, and PRDM14 were top five regulators with highest degree values that co-regulates all module genes. Moreover, identified gene module was in silico validated in association with the indicator of invasiveness and prognosis of PA, plus some related cancer types. Furthermore, these module genes were also expressed in blood, salivary gland, and spinal cord tissues. These results may provide the evidence on featured gene module which might play a prominent role in NFPA prognosis and sub-typing as effective biomarkers and therapeutic targets in the future.

Keywords: Differential Co-Expression Network, Non-Functional Pituitary Adenoma, Invasiveness, Prognosis, Biomarker

References: [1] Aydin, B., & Arga, K. Y. Co-expression Network Analysis Elucidated a Core Module in Association

With Prognosis of Non-functioning Non-invasive Human Pituitary Adenoma. Front Endocrinol. (2019) 10, 361.

Corresponding Author's Address: Kazim Yalcin Arga, E-mail: kazim.arga@marmara.edu.tr

TRANSCRIPTOME ANALYSIS OF AUTISTIC DIZYGOTIC TWINS AND THEIR PARENTS REVEALED ABERRANT SPLICING OF NEURONAL GENES AND ALTERED TRANSCRIPTIONAL NETWORKS IN BLOOD

Kaan Okay [1], Pelin Unal Varis [2], Suha Miral [2], Burcu Ekinci [1], Tutku Yaras[1], Gokhan Karakulah[1,4], Yavuz Oktay[3,4]

1.Department of Genome Sciences and Molecular Biotechnology, Izmir International Biomedicine and Genome Institute, Dokuz Eylul University Health Campus, Balcova, Izmir, Turkey.
2. Deparment of Child and Adolescent Psychiatry, Faculty of Medicine, Dokuz Eylul University Health Campus, Balcova, Izmir, Turkey.
3. Department of Medical Biology, Faculty of Medicine, Dokuz Eylul University Health Campus, Balcova, Izmir, Turkey.
4.Izmir Biomedicine and Genome Center, Dokuz Eylul University Health Campus, Balcova, Izmir, Turkey.

Autism spectrum disorder (ASD) is a neurodevelopmental disorder characterized by deficits in social skills, repetitive behaviors, abnormality of speech and nonverbal communication. As much as 1 in 59 children is diagnosed with ASD. We hypothesized that disruption of some biological phenomenon like alternative splicing (AS) and co-expression networks of genes may be associated with autism. To this end, we performed gene and exon level expression analysis in between non-syndromic autistic dizygotic twins and their parents by using the RNA-seq method to compare co-expression networks and their association with AS events in autism. Such combinatorial analyses revealed that some of the genes in co-expression modules related to synapse maturation and neurotransmitter release displayed aberrant AS events. Specifically, we found that NR4A2 gene, which is a transcription factor involved in neurotransmitter synthesis had aberrant splicing pattern in children compared to their parents and was a member of a co-expression module enriched for p53 downstream, AP1 and alpha linolenic acid (ALA) metabolism pathways. We also observed that PCA analysis based on differential alternative splicing between children and parents led to better correlation with autistic traits, compared to PCA analysis based on differential expression. Overall, our results suggest that transcriptome-wide alternative splicing analysis in peripheral blood leukocytes of patients with ASD provides further insight into disease etiopathogenesis and should be considered for biomarker development purposes.

Keywords: Autism Spectrum Disorders; Alternative RNA Splicing; Co-Expression

References: [1] Parikshak NN, Swarup V, Belgard TG, Irimia M, Ramaswami G, Gandal MJ, Hartl C, Leppa V, de la Torre Ubieta L, Huang J, Lowe JK, Blencowe BJ, Horvath S, Geschwind DH. Genome-wide changes in lncRNA, splicing, and regional gene expression patterns in autism. Nature. 2016 Dec 15;540:423-7. PubMed PMID: 29995847

[2] Voineagu I, Wang X, Johnston P, Lowe JK, Tian Y, Horvath S, Mill J, Cantor RM, Blencowe BJ, Geschwind DH. Transcriptomic analysis of autistic brain reveals convergent molecular pathology. Nature. 2011 May 25;474(7351):380-4. PubMed PMID: 21614001.
[3] Lord C, Bishop SL. Recent advances in autism research as reflected in DSM-5 criteria for autism spectrum disorder. Annu Rev Clin Psychol. 2015 Jan 2;11:53-70. PubMed PMID: 25581244.
[4] Smith RM, Sadee W. Synaptic signaling and aberrant RNA splicing in autism spectrum disorders. Front Synaptic Neurosci. 2011 Jan 26;3:1. PubMed PMID: 21423409.
[5] Ancín I, Cabranes JA, Vázquez-Álvarez B, Santos JL, Sánchez-Morla E, Alaerts M, Del-Favero J, Barabash A. NR4A2: effects of an "orphan" receptor on sustained attention in a schizophrenic population. Schizophr Bull. 2013 May;39(3):555-63. PubMed PMID: 22294735.

Corresponding Author's Address: yavuz.oktay@ibg.edu.tr

CYCLOPAMINE AFFECTS THE GENE EXPRESSION OF ASTROCYTES, GLIOBLASTOMA AND GLIOBLASTOMA STEM CELLS IN MONO AND CO-CULTURE SYSTEMS

Duygu Calık Kocaturk[1], Berrin Ozdil[1, 2], Huseyin Aktug[1], Aysegul Uysal[1]

1. Ege University, Faculty of Medicine, Department of Histology and Embryology, 35100, Izmir Turkey
2.Suleyman Demirel University, Faculty of Medicine, Department of Histology and Embryology, 32260, Isparta Turkey

Among the members of the Hedgehog gene family, Sonic Hedgehog (Shh) has the most prominent effects on the embryonic development at the brain and extremities.[1,2]. The Sonic Hedgehog signaling pathway is one of the important signaling pathways that active in both embryogenesis and carcinogenesis. Many cancer types, such as medulloblastoma, rhabdomyosarcoma, colorectal carcinoma, have increased activity of this pathway. Glioblastoma multiforme (GBM) is the most common malignant primary brain tumor in adult humans and is a rapidly growing, infiltrating tumor, even if it does not metastase to the distant organs[3]. Tumor stem cells, like stem cells, are characterized by self-renewal and differentiation and resistant to treatment regimens such as radiotherapy and chemotherapy besides, they can spread distant organs and are thought to be affected by cancer recurrence also found in tumor mass. Cancer stem cells have also been identified in tumor masses of patients with Glioblastoma multiforme, and these cells may cause worsening prognosis. In recent years, Hedgehog signaling pathway has become an important target for Glioblastoma multiforme treatment. Cyclopamine is an important agent used in the inhibition of the Shh signaling pathway by binding to the Smoothened (Smo) receptor and has shown to inhibit tumor growth in gliomas, pancreatic and colon carcinomas in animal studies[4]. Here, we prospectively investigated the effect of Shh signaling pathway and its inhibition on GBM and GBM cancer stem cells co-cultured with astrocytes. Cell-cell interactions in the co-culture conditions were evaluated to determine whether the Shh signaling pathway active in GBM and GBM cancer stem cells, will undergo a change in the inhibition effect of cyclopamine in the presence of astrocytes. At the same time, the ability of the cells to secrete Shh protein as paracrine signalling, and cell-cell interactions were expected to answer the question whether Shh signaling pathway has a positive or negative effect on cyclopamine inhibition. Although cyclopamine is an effective agent to suppress the Shh pathway, it does not cause sufficient cell death alone. Results of GBM and GBM cancer stem cells control groups and cyclopamine effect on them are different at RNA and protein levels. Different results were often obtained when co-culture with these cells was obtained. As a result of this study, especially when the differences in single cell culture and co-culture results are considered, the importance of the micro-environment effect in diseases and treatments has been revealed.

Keywords: Glioblastoma Multiforme; Astrocyte; Cancer Stem Cell; Co-Culture; Sonic Hedgehog; Cyclopamine

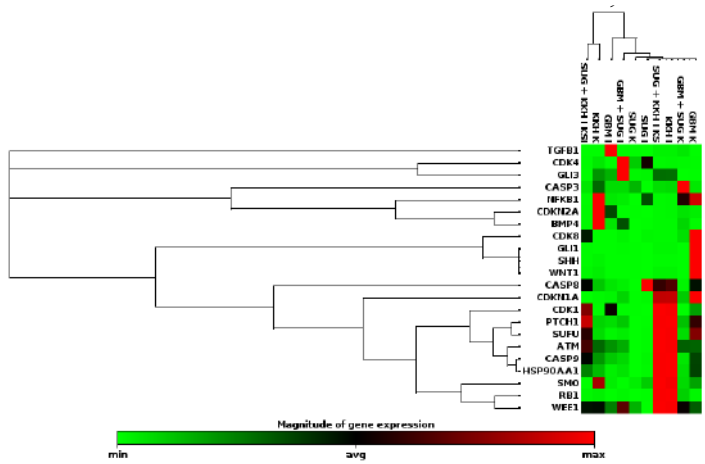


Figure 1. Gene expression analysis of mono and co-culture of astrocytes (SVG), GBM and GBM stem cells (KKH) of selected genes in the absence and presence of cyclopamine.

References:[1]. Hui M, Cazet A, Nair R, Watkins DN, O'Toole SA, Swarbrick A. The Hedgehog signalling pathway in breast development, carcinogenesis and cancer therapy. Breast Cancer Research. 2013. PMID: 23547970

[2] Alvarez-Medina R, Le Dreau G, Ros M, Marti E. Hedgehog activation is required upstream of Wnt signalling to control neural progenitor proliferation. Development. 2009;136(19):3301–3309. PMID: 19736325

[3] McNeill KA. Epidemiology of Brain Tumors. Neurol Clin. United States; 2016 Nov;34(4):981–998. PMID: 27720005

[4] Braun S, Oppermann H, Mueller A, Renner C, Hovhannisyan A, Baran-Schmidt R, Gebhardt R, Hipkiss AR, Thiery J, Meixensberger J, Gaunitz F. Hedgehog signaling in glioblastoma multiforme. Cancer Biol Ther. 2012;13(7):487–495. PMID: 22406999

Corresponding Author's Address: Aysegul Uysal, Department of Histology and Embryology Faculty of Medicine, Ege University 35100, Izmir, Turkey. Email: aysegul.uysal14@gmail.com

EVALUATING CLUSTERING METHODS TO IDENTIFY SUB-NETWORKS FOR BIOMARKER DISCOVERY USING BREAST CANCER TRANSCRIPTOME DATA

Nehir Kızılilsoley[1], *Emrah Nikerel*[1]

1. Department of Genetics and Bioengineering, Yeditepe University

Biomarkers are, typically used in medical context, any measurable molecule or set of molecules, or an (reconstructed) image, pointing to a feature, often the presence of diseases, sometimes even its state. It is utilized in diagnosis, prognosis and monitoring of numerous heterogeneous diseases, evaluating potentially effective therapeutic regime and intrinsically constitutes main gateway to P4 medicine[1].

The discovery of biomarkers underwent a transition from knowledge-driven practice i.e. based on known mechanisms of the target disease, to data-driven discovery especially with the advent of high-throughput technologies and available large scale -omic datasets. As such, typical biomarker discovery workflow consists of using a collection of statistical, clustering, pruning and classification techniques to e.g. assess the candidates and identify networks, on carefully collected datasets, e.g. from case-control studies[2]. Owing to the recently introduced systems biology approach, not only the measurements but also networks (protein-protein interaction, metabolic reaction etc.) are also taken into consideration to link different -omics datasets. Focusing on the use of large scale transcriptome data, main challenges are large number and complex nature of gene-gene interactions, small sample sizes despite recently available large cohorts and validation of the results with external data.

The aim of this work is to evaluate alternative clustering methods (Partitioning methods k-means clustering, Hierarchical, Fuzzy, Density-based and Model-based clustering) for subnetwork identification using transcriptome data within biomarker discovery context. To do so, we compiled a breast ductal carcinoma transcriptome dataset[3] from public databases[4,5], assessed differentially expressed gene sets and clustered these sets using the above mentioned methods using a handful of distance measures. The identified networks are benchmarked against previously known/identified networks.

Keywords: Breast Ductal Carcinoma, Biomarker Discovery, Clustering, Transcriptomics, P4 Medicine

References: [1] Srinivas PR, Kramer BS, Srivastava S. Trends in biomarker research for cancer detection. Lancet Oncol. 2001;2(11):698-704. PMID:11902541
[2]Wang J, Zuo Y, Man YG, Avital I, Stojadinovic A, Liu M, Yang X, Varghese RS, Tadesse MG, Ressom HW. Pathway and network approaches for identification of cancer signature markers from omics data. J Cancer. 2015;6(1):54-65. PMID:25553089
[3]Cancer Genome Atlas Network. Comprehensive molecular

portraits of human breast tumours. Nature. 2012;490(7418):61-70. PMID:23000897
[4]Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015;19(1A):A68-77. PMID:25691825
[5]Athar A, Füllgrabe A, George N, Iqbal H, Huerta L, Ali A, Snow C, Fonseca NA, Petryszak R, Papatheodorou I, Sarkans U, Brazma A. ArrayExpress update - from bulk to single-cell expression data. Nucleic Acids Res. 2019;47(D1):D711-D715. PMID:30357387

Corresponding Author's Address: Nehir Kızılilsoley nehirkizililsoley@gmail.com
Emrah Nikerel emrah.nikerel@yeditepe.edu.tr

COMPARATIVE ANALYSIS OF HUMAN RENAL CELL CARCINOMA TYPES IN RESPECT TO MOLECULAR SIGNATURES AT DIFFERENT OMICS LEVELS

Aysegul CALISKAN[1,2], *Kazim Yalcin ARGA*[1]

1. Department of Bioengineering, Marmara University, Istanbul, Turkey,
2. Department of Pharmacy, Istinye University, Istanbul, Turkey,

Human Renal Cell Carcinoma (RCC) has the highest mortality rate of the genitourinary cancers in adults[1]. Understanding biological mechanism of RCC may improve the current diagnosis, treatment and prognosis of RCC. Renal clear cell carcinoma (KIRC), papillary renal cell carcinoma (KIRP) and chromophobe renal cell carcinoma (KICH) significantly dissociates from each other in many respects including location of a cancer, cell type which they originate and genetic alterations[2]. The aim of this study is to identify molecular signatures at RNA (mRNA, miRNA), protein (receptor, transcription factor, etc.), and metabolite levels by the integration of gene expression profiles with genome-scale biomolecular networks, and to analyze three RCC types (KICH, KIRP, KIRC) in a comparative perspective. A multi-stage analysis method was developed and applied here. Initially, gene expression datasets were statistically analyzed and differentially expressed genes were identified. Hub-proteins were identified by taking account protein-protein interactions. Then, results were integrated with transcriptional regulatory and metabolic networks to identify reporter biomolecules at protein and metabolite levels. The prognostic performance of the biomolecules were also analyzed using survival time statistics. Six TFs, YBX1, ETS1, GATA2, E2F4, AR and GATA1, were determined as mutual transcriptional regulatory components for all three diseases. hsa-miR-335-5p and hsa-miR-887-5p were determined as mutual miRNAs for all three diseases. BUB1 was determined as only mutual receptor for three diseases. Retinoid derivatives, Acetyl-CoA and Proline oxidase were identified as affected metabolites in RCC types. Reporter TFs, receptors, miRNAs and metabolites specific to each disease were determined in this study. This approach revealed already-known biomarkers, tumor suppressors and oncogenes in RCC types as well as various receptors, miRNAs, transcription factors, other proteins, and metabolites as novel biomarker candidates and potential therapeutic targets. In addition, we predicted a significance difference among KIRP and KIRC with KICH, probably originating from the tumor location and tissue origin. Our results support that these three diseases are significantly dissociate from each other although all of them forms in kidney. The whole mechanism of how these diseases are formed needs to be elucidated urgently one by one, so that specific treatment can be produced. In addition, we reported valuable data for further experimental and clinical efforts, because the proposed biomolecules have significant potential as systems biomarkers for screening or therapeutic purposes in RCC types.

Keywords: Kidney; Renal Cell Carcinomas; Biomarkers

References: [1] Yan F, Wang Y, Liu C, Zhao H, Zhang L, Lu X, Chen C, Wang Y, Lu T, Wang F. Identify clear cell renal cell carcinoma related genes by gene network. Oncotarget. 2017, Vol. 8, (No. 66), pp: 110358-110366. PubMed PMID: 29299153.

[2] Cairns P. Renal Cell Carcinoma. Cancer Biomark. 2011 ; 9(1-6): 461–473. PubMed PMID: 22112490.

Corresponding Author's Address: Postal address: Maltepe Mah., Edirne Çırpıcı Yolu Sokak, Mira Rezidans, A 24, Zeytinburnu, İstanbul,Turkey.

e-mail: gulaysecaliskan@hotmail.com

URL: https://eczacilik.istinye.edu.tr/tr/kadro/106

DETERMINATION OF CANDIDATE BIOMARKERS THROUGH DIFFERENTIAL INTERACTOME IN COLORECTAL ADENOCARCINOMA

Hande Beklen[1], Gizem Gulfidan[1], Beste Turanli[2], Pemra Ozbek Sarica[1], Betul Karademir[3], Kazim Yalcin Arga[1]

1. Department of Bioengineering, Marmara University, Istanbul, Turkey
2. Department of Bioengineering, Istanbul Medeniyet University, Istanbul, Turkey
3. Department of Biochemistry, Marmara University, Istanbul, Turkey

Colorectal cancer is one of the most lethal types of cancers common in both men and women. According to the data base published by the International Agency for Research on Cancer, colorectal cancer is the third most common type of cancer found in Turkey and worldwide [1]. The high heterogeneity of colorectal cancer leads to difficulties explaining the biology and behavior of this cancer. The aim of this study is to identify prognostic biomarkers and potential therapeutics for colorectal cancer using the protein interactions differentiated among healthy and tumor groups. Among this purpose at first stage, the differential protein-protein interactions (dPPIs) were identified by using "Differential Interactome" algorithm[2] which is published by our research group. Two independent data sets were obtained from "The Cancer Genome Atlas (TCGA)" containing 644 tumor samples and 51 normal samples and "Gene Expression Omnibus (GEO)" containing 32 tumor samples and 32 normal samples. As a result of differential interactome analysis, significant dPPIs were determined (2434 dPPIs in GEO data set, 1619 dPPIs in TCGA data set) and highly interacting protein modules were identified. Principal Component Analysis for diagnostic purpose and Kaplan-Meier analysis for prognostic purpose were performed for each module. 16 modules were determined significant having diagnostic potential while 6 modules were found having prognostic potential. In addition common significant dPPIs in both data sets were observed in point of drug repositioning and 6 dPPIs and 13 drug targets for these interactions were identified. By using molecular dynamics simulations root mean square deviation (RMSD) and root mean square fluctuation were taken as performance metrics to perform further investigation in vitro cell culture. This study will shed light on the identification of specific biomarkers and drug targets for early detection, disease progression, and accurate treatment, helping to understand some systems of colorectal cancer by preventing high mortality.

This study was supported under FEN-C-1206-0199 project.

Keywords: Cancer, Differential Interactome, Molecular Dynamics

References: [1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin. 2018 Nov 68(6):394-424. PubMed PMID: 30207593

[2] Ayyildiz, D, Gov, E, Sinha R, and Arga K. Y. Ovarian Cancer Differential Interactome and Network Entropy Analysis Reveal New Candidate Biomarkers. OMICS: A Journal of Integrative Biology. 2017 May 21(5): 285-294. PubMed PMID: 28375712

Corresponding Author's Address: Department of Bioengineering, Marmara University, Istanbul, Turkey, kazim.arga@marmara.edu.tr

CLUSTERING-BASED METABOLISM-ORIENTED ANALYSIS OF MOUSE MODELS OF PARKINSON'S DISEASE

Ecehan Abdik[1], *Tunahan Çakır*[1]

1. Department of Bioengineering, Gebze Technical University

Parkinson's disease (PD) is the second most common neurodegenerative disease in the world. The number of people suffering from neurodegenerative diseases has dramatically increased in recent decades. Therefore, it is essential to understand the disease mechanisms in order to improve treatment and diagnostic techniques[1].

Metabolism has a principle role in molecular mechanisms of diseases since it provides nutrients that are necessary to fuel pathways necessary for cellular processes. Most of the metabolic changes are initiated at the transcriptome level. Transcription of genes affects the translation of enzymes catalyzing metabolic reactions, leading to changes in the activities of metabolic pathways. Since genome-scale metabolic networks enables mapping of transcriptome data on metabolic pathways, they have been used for systematic investigation of complex disease mechanisms[2] .

Experimental investigation of diseases including drug tests are mostly carried out on model organisms. Mus musculus (Mouse) is one of the most commonly used model organisms for human diseases. Mouse models of PD are created with two different approaches: genetic-based, and chemical (toxin) based. Numerous transcriptome data of mouse models of PD created with both approaches are available in the public transcriptome database Gene Expression Omnibus (GEO). This study pursues a clustering-based metabolism- oriented comparison of the transcriptome data of PD mouse models with each other and with the transcriptome data of PD patients by mapping the data on genome-scale brain-specific metabolic networks.

To this aim, an updated version[3] of previously reconstructed brain-specific genome-scale metabolic network model of human[4] was used, with improved compartmentalization of reactions and metabolites as cytosolic and mitochondrial. For mapping transcriptome data, the improved network model of human was used as a template to reconstruct the first brain-specific metabolic network model of mouse by a homology based approach. Since many reactions are controlled by multiple genes, it is important to map gene expression values on reactions to obtain reaction scores. If genes control a reaction independently, their expression was summed up to calculate the corresponding reaction score, and if the genes code for different subunits of an enzyme complex, minimum of their expression levels was assigned as reaction score. These scores indicate the potential capacity of reactions based on transcriptional activities of genes for that condition. By using many PD related transcriptome datasets for

both mouse PD models and PD patients, disease related metabolic changes were investigated. Fold changes between reaction scores for control groups and PD groups were calculated for a number of datasets from GEO. Hierarchical clustering was applied to the fold changes from the datasets to identify metabolic similarities and differences between different mouse models of PD. Clustering of human and mouse datasets together also enabled identification of which experimental PD models can reflect PD metabolism more realistically.

Keywords: Parkinson's Disease, Genome-Scale Metabolic Network, Animal Models, Clustering, Transcriptome

References: [1] Dawson TM, Dawson VL. Molecular pathways of neurodegeneration in Parkinson's disease. Science (80-). American Association for the Advancement of Science; 2003;302(5646):819–22.
[2] Bordbar A, Feist AM, Usaite-Black R, Woodcock J, Palsson BO, Famili I. A multi-tissue type genome-scale metabolic network for analysis of whole-body systems physiology. BMC Syst Biol. BioMed Central; 2011;5(1):180.
[3] Ali Kaynar, Integrative Analysis of Transcriptome Data and Genome-Scale Metabolic Networks to Identify Drug Target and Drug Candidates for Parkinson's Disease, (Master's thesis, Gebze Technical University, 2019)
[4] Sertbaş M, Ülgen K, Çakır T. Systematic analysis of transcription-level effects of neurodegenerative diseases on human brain metabolism by a newly reconstructed brain-specific metabolic network. FEBS Open Bio. Elsevier; 2014;4:542–53.

Corresponding Author's Address: Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey tcakir@gtu.edu.tr

BRIDGING ENHANCERS AND TARGET GENES THROUGH CONDITION-SPECIFIC REGULATORY PROTEIN COMPLEXES

Emel Kökrek[1], *Pınar Pir*[1]

1. Gebze Technical University, Department of Bioengineering, Çayırova, KOCAELİ

Transcriptional regulation involves many proteins, along with protein-protein and protein-DNA interactions. When these interactions occur simultaneously to initiate the basic transcriptional machinery, the resulting structure can be defined as a protein complex. Protein complexes providing communication among enhancer– promoter–target gene regions can be studied further with protein interaction networks and omic data integration. Based on dense subnetworks in protein-protein, domain-domain interaction networks and integrated proteome data, both qualitative and quantitative predictions on protein complexes can be made[1,2]. In our attempt to understand transcriptional regulation, we formulate a methodology involving two major prediction steps: 1. Protein complex prediction 2. Enhancer-target gene prediction[3]. Integrating proteome data from two cellular conditions with PPI and DDI, differential protein complexes will be obtained. This differentia can be analyzed in 3 categories: absence/presence of protein complex, missing protein members in the same complex, or differential abundance of the protein complex. Taking the first one into account, differential protein complexes list will be formed. Protein members of complexes will be browsed for their role in transcription: The complex will be kept if it includes at least two DNA-binding transcriptional factors, left out otherwise. Experimentally validated target genes of these transcriptional factors will be collected from relevant databases. Second prediction algorithm will be run for finding out the enhancers of the target genes. By the integration of accessibility data of these enhancer and promoter regions, the mechanism uniting the regulatory regions will be inferred. The association of differential protein complexes to these regulatory sequences and target genes will be made based on the correlation occurring between enhancer accessibility, the occurrence of protein complex and target gene expression. The distinction between the protein complex profiles in healthy and disease conditions will provide more efficient drug targets. Although transcriptional factors were thought to be "untargettable" previously, nowadays, by different modes of action they are inhibited or activated.[4] Hence, for most disease conditions where transcriptional factors are major drivers, our framework can point out the most plausible drug targets.

Keywords: Regulatory Protein Complexes; Protein Complex Prediction; Protein-Protein Interaction Network; Enhancer-Target Gene Prediction

References: [1] Will T, Helms V. Differential analysis of combinatorial protein complexes with CompleXChange. BMC Bioinformatics.

2019;20(1):1–14.
[2] Hernandez C, Mella C, Navarro G, Olivera-Nappa A, Araya J. Protein complex prediction via dense subgraphs and false positive analysis. Ruan J, editor. PLoS One [Internet]. 2017 Sep 22;12(9):e0183460. Available from: https://dx.plos.org/10.1371/journal.pone.0183460
[3] Mora A, Sandve GK, Gabrielsen OS, Eskeland R. In the loop: promoter–enhancer interactions and bioinformatics. Brief Bioinform [Internet]. 2015 Nov 19;17(July 2015):bbv097. Available from: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbv097
[4] Lambert M, Jambon S, Depauw S, David-Cordonnier M-H. Targeting Transcription Factors for Cancer Treatment. Molecules [Internet]. 2018 Jun 19;23(6):1479. Available from: http://www.mdpi.com/1420- 3049/23/6/1479

Corresponding Author's Address: Pınar Pir Gebze Technical University, Department of Bioengineering, 41400, Çayırova, KOCAELİ. pinarpir@gtu.edu.tr

CAN INFORMATION COMMUNICATION THEORY BE USED TO EXPLAIN ALLOMETRIC RELATION BETWEEN GENES AND PROTEINS?

Yekbun Adiguzel[1]

1. Biophysics Department, School of Medicine, Altinbas University, Istanbul, Turkey

Allometric scaling formulates relations between distinct biological parameters. There, the logarithmic relations that are observed or present between the parameters of interest can be formulated as the equation of the fitted power trendline to the plot of the parameters, or as the antilogarithm of the equation of the linear trendline that is fitted to the log-log plot of the same parameters. A distinct concept, information communication theory deals originally with reliable transmission of messages through noisy channels. It has a wide scope of applications. Independent of the allometric scaling concept that is formerly mentioned, we tried in our earlier studies to use a modified form of the information communication theory to relate the encoding DNA's and the encoded protein's residue numbers, to equalize their information contents. This was based on the assumption that there is transformation rather than a loss of information, during protein translation from its encoding gene. The extension in the information amount due to the presence of introns in the eukaryotic DNA was solved by using an exponent of the protein length. This was basically establishing an allometric relation between the information contents of the parameters of interest. Here it is claimed that information theory, and the related concepts, can be a means to explain the source of allometric relations between such molecular entities that are characteristically possessing information, and involving in information communication events of the cells, in the form of information transformation, such as we see during translation of proteins from genes.

Keywords: Shannon's Information Communication Theory; Gene Size; Protein Length; Allometric Scaling Relation

Corresponding Author's Address: Biophysics Department, School of Medicine, Altinbas University, Kartaltepe Mah. Incirli Cad. No11, 34147 Bakirkoy, Istanbul, Turkey.
Email: yekbun.adiguzel@altinbas.edu.tr

IDENTIFICATION OF SYSTEMS BIOMARKERS AND CANDIDATE DRUGS IN PROSTATE ADENOCARCINOMA

Gizem Gulfidan[1], *Beste Turanli*[2], *Hande Beklen*[1], *Kazim Yalcin Arga*[1]

1. Department of Bioengineering, Marmara University, Istanbul, Turkey
2. Department of Bioengineering, Istanbul Medeniyet University, Istabul, Turkey

Deciphering the alterations in the protein interactome is mandatory to reach a systems-level understanding of tumorigenesis, since physical interactions among proteins influence cellular pathways, and mediate various physiological processes in all living organisms [1]. Also, the elucidating of the molecular mechanisms underlying cancer and the identification of efficacious biomarkers is crucial for accurate diagnosis and prognosis of cancers as well as the prevention of tumorigenesis. However, it is a challenging task to develop highly accurate and robust biomarkers considering the complexity of the molecular biology behind these pathologies. On the other hand, the discovery and production of therapeutic agents for cancer treatment need investments in point of money, time and labor and thus it causes waste of money and time. Increasing studies and improvements on omic technologies and computational analysis provide opportunities for drug repostioning which is a useful approach to find out already approved drugs with high confidence and good pharmacokinetic properties on new diseases [2]. The present study highlights the concept of systems biomarkers with special focus on prostate adenocarcinoma using our differential interactome approach and provides the presenting of new drug candidates for prostate adenocarcinoma. To that end, differential interactome algorithm [3] was developed and applied to gene expression profile of prostate adenocarcinoma having 550 samples (52 normal - 498 tumor samples) by using the human protein interactome data in order to find significant protein-protein interactions differentiated at tumor state (dPPI). Our results show that 183 dPPIs among 194 differentially interacting proteins (DIPs) taking roles in dPPIs were significant in prostate adenocarcinoma, which 19 of these dPPIs were repressed, while 164 were activated in tumor phenotype. In addition, the features of DIPs were investigated and it was found that 77 of DIPs were druggable, 49 DIPs were tumor suppressor proteins and 50 DIPs were oncogenes. Also, the gene set enrichment analysis were carried out for DIPs. Moreover, we found various significant modules consisting of DIPs and their diagnostic and prognostic features were examined. Several modules were determined as having diagnostic properties with sensitivity and specificity values > 0.9 and as having prognostic properties with hazard ratio>7, p-value<0.05. Finally, some candidate therapeutic agents were submitted for prostate adenocarcinoma by the aid of drug repositioning approach. This study will pave the way for further studies integrating with systems-level analyses of cancers and provide an insight for clinic studies to design diagnostic kit for early diagnosis

of cancer and to investigate repositioned cancer drugs.

This study was supported under TUBITAK/SBAG/117S489 project.

Keywords: Protein-Protein Interactions; Prostate Adenocarcinoma; Differential Interactome; Drug Repositioning

References: [1] Sevimoglu T, Arga KY. The role of protein interaction networks in systems biomedicine. Comput Struct
Biotechnol J. 2014 Sep 3;11(18):22-7. PubMed PMID: 25379140.
[2] Banno K, Iida M, Yanokura M, Irie H, Masuda K, Kobayashi Y, Tominaga E, Aoki D. Drug
repositioning for gynecologic tumors: a new therapeutic strategy for cancer. ScientificWorldJournal.
2015;2015:341362. PubMed PMID: 25734181.
[3] Ayyildiz D, Gov E, Sinha R, Arga KY. Ovarian Cancer Differential Interactome and Network Entropy
Analysis Reveal New Candidate Biomarkers. Omi A J Integr Biol. 2017 May;21(5):285-294.. PubMed
PMID: 28375712.

Corresponding Author's Address: Department of Bioengineering, Marmara University, Istanbul, Turkey, kazim.arga@marmara.edu.tr

TOWARDS IN SILICO IDENTIFICATION OF RESCUE SITES IN RAC1 ONCOGENIC MUTATIONS

R. Busra Ozguney[1], S. Ece Acuner – Ozbabacan[2], Turkan Haliloglu[1]

1.Polymer Research Center and Chemical Engineering Department, Bogazici University
2.Department of Bioengineering, Istanbul Medeniyet University, Istanbul 34700, Turkey

Ras-related C3 botulinum toxin substrate 1 (Rac1) is a small Rho GTPase and a member of Ras superfamily involved in cell proliferation, adhesion and migration [1]. The abnormal activity of Rac1 due to hot spot mutations yields cancer development, cancer progression and metastasis [2]. In addition to that Rac1 has been recently identified as a central player in cancer therapy resistance [3]. As part of Ras superfamily, Rac1 maintains its function via cycling through two states: inactive GDP bound state and active GTP bound state in which it interacts with its effectors [4]. Regulation of this switch mechanism is tightly controlled and maintained by GTPase-activating proteins (GAPs), guanine nucleotide exchange factors (GEFs) and guanine nucleotide exchange inhibitors (GDIs) [4]. One of the most frequent oncogenic mutations in Rac1 is the gain-of-function P29S; which increases the intrinsic GDP/GTP nucleotide exchange rate leading to a "spontaneously activated" state with a "fast-cycling" property and ultimately increases the effector activation resulting in melanoma [2]. In this study, we propose a novel approach to identify "cancer rescue mutation(s)" so that drugs mimicking suppressor mutations might be designed to rescue the effect of the oncogenic mutations [4]. Accordingly, we aim to explore the dynamic mechanism of the conformational shift due to mutations and mechanistically identify key sites that would control and modify the ensemble of conformations towards the wild type.

The rescue positions at/in close vicinity to the mechanistically informative regions are hypothesized to have the capacity to alter the dynamics related to protein's function. In silico mutations are introduced to the first and second slowest global mode hinge residues[5,6,7] of P29S mutant Rac1 structure as perturbations and their dynamic response effects on the wild type (inactive and active states) and the P29S mutant active Rac1 structures are analyzed by Molecular Dynamics (MD) Simulations. Mutations are observed to generally perturb residues which are either important in the Rho switch mechanism or at the interaction interface of the effector proteins. Further, unlike the wt active, P29S active takes at least two states that adopt different conformations with distinct differences in H-bonding occupancies, residue mobilities and their cooperativity. Intriguingly, particularly for the residues related to Mg2+coordination, while one of these states is similar to the wt inactive, the other state deviates significantly. Some of the P29S mutants indeed restores the impaired Mg2+ coordination and the behavior of GTP and regulatory/ downstream effector binding residues observed in both states of P29S active, however the effects of P29S mutation on residues 29 and 30 are not totally recovered. Lastly, we emphasize the importance of the

regulatory/effector binding residues in gaining fast cycling property. Overall, this study proposes a framework to identify rescue positions for the oncogenic mutations and the results would be complemented with an experimental study to confirm the restored Mg2+ and GTP affinities in mutant cases.

Keywords: Biomolecules, Allostery, Molecular Dynamics Simulation, Drug Design, Protein Dynamics, Internal Dynamics, G proteins, Functional Selectivity, Mutagenesis, Computational Biology, Protein Function, Rho GTPases, Functional Control, Cancer

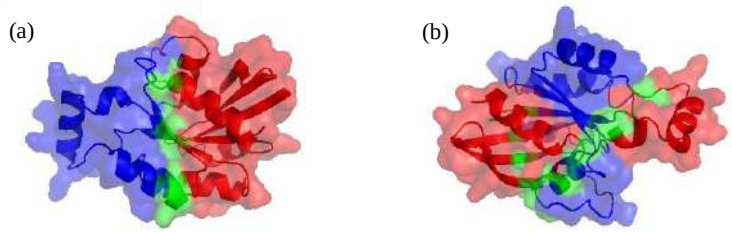


Figure 1. P29S active structure (a) Slowest mode and (b) Second slowest mode hinge residues by Gaussian Network Model are colored as green.

References: [1] Matos P, Skaug J, Marques B, Beck S, Veríssimo F, Gespach C, Boavida MG, Scherer SW, Jordan P. Small GTPase Rac1: structure, localization, and expression of the human gene. Biochem Biophys Res Commun. 2000 Nov 2;277(3):741-51. PubMed PMID: 11062023

[2] De P, Aske JC, Dey N. RAC1 Takes the Lead in Solid Tumors. Cells. 2019 Apr 26;8(5). PubMed PMID: 31027363

[3] Cardama GA, Alonso DF, Gonzalez N, Maggio J, Gomez DE, Rolfo C, Menna PL. Relevance of small GTPase Rac1 pathway in drug and radio-resistance mechanisms: Opportunities in cancer therapeutics. Crit Rev Oncol Hematol. 2018 Apr;124:29-36. PubMed PMID: 29548483

[4] Baronio R, Danziger SA, Hall LV, Salmon K, Hatfield GW, Lathrop RH, Kaiser P. All-codon scanning identifies p53 cancer rescue mutations. Nucleic Acids Res. 2010 Nov;38(20):7079-88. PubMed PMID: 20581117

[5] Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997;2(3):173-81. PubMed PMID: 9218955

[6] Haliloglu T, Ivet B, Erman B. Gaussian Dynamics of Folded Proteins. Phys Rev Lett. 1197; 79(16):3090-93.

[7] Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T. HingeProt: automated prediction of hinges in protein structures. Proteins. 2008 Mar;70(4):1219-27. PubMEd PMID: 17847101

Corresponding Author's Address: Department of Chemical Engineering and Polymer Research Center, Bogazici University, Istanbul, Turkey, E-mail: halilogt@boun.edu.tr

PREFERENTIAL NUCLEOTIDE EXCISION REPAIR IN EXONS

Cem Azgari, Defne Çirci, Zeynep Kılınç, Berk Turhan and Ogün Adebali[1]

1.Faculty of Engineering and Natural Sciences, Molecular Biology, Genetics and Bioengineering, Sabancı University Orhanlı, Tuzla, 34956, Istanbul

Exons are evolutionary conserved regions where mutagenesis is less tolerable compared to introns due to their direct informational contribution to the protein sequence and therefore function. Because of the eliminating effect of hazardous mutations encoded at amino acid level, a lower rate of natural variants are observed in exons compared to introns which are spliced out in the process of transcription. However, there is a similar trend in tumor samples which are less prone to selection.[1]Such a fact in skin and lung cancer implies differential nucleotide excision repair levels within same gene for bulky-adduct inducing damaging agents such as ultraviolet and bezno[a]pyrene. Here, we aim to reveal differences between the intronic and exonic regions by analyzing the genome-wide repair and damage datasets, generated with recent NGS- based techniques XR-seq[2] and Damage-seq[3], respectively. Our research suggests that exons are not damaged differentially but are repaired more efficiently compared to introns. We found that the bias in nucleotide excision repair is independent of transcription and transcription-coupled repair. On the contrary, the exonic preference is more prominent in the poorly repaired genes. The results suggest that the preference is likely to be due to the differential histone markers between introns and exons. Our findings shed light into intronic mutation burden in some cancer types.

Keywords: Nucleotide Excision Repair; Exonic And Intronic Region; Cisplatin; Oxaliplatin; Bezno[A]Pyrene; (6-4)Pps; CPDs; XR-Seq; Damage-Seq;

References: [1] Frigola, J, et al. Reduced mutation rate in exons due to differential mismatch repair. Nat Genet. 2017 Nov 29;49(12):1673-1674. PMID: 29186127

[2] Hu J, Adar S, Selby CP, Lieb JD, Sancar A. Genome-wide analysis of human global and transcription-coupled excision repair of UV damage at single-nucleotide resolution. Genes Dev. 2015 May 1;29(9):948-60. Pubmed PMID: 25934506.

[3] Hu J, Lieb JD, Sancar A, Adar S. Cisplatin DNA damage and repair maps of the human genome at single-nucleotide resolution. Proc Natl Acad Sci U S A. 2016 Oct 11;113(41):11507-11512. Pubmed PMID: 27688757.

Corresponding Author's Address: To whom correspondence may be addressed: Sabancı University, Faculty of Engineering and Natural Sciences, Tuzla, Istanbul, 34956
Turkey Tel.: 216-568-7043; E-mail: oadebali@sabanciuniv.edu.

3D Spatial Organization and Network Guided Comparison of GBM Mutations

Cansu Dincer[1], *Atilla Gursoy*[2,4], *Ozlem Keskin*[3,4], *Nurcan Tuncbag*[1,5]

1.Department of Health Informatics, Graduate School of Informatics, METU, Ankara, Turkey
2. Department of Chemical and Biological Engineering, Koc University, Istanbul, Turkey
3. Department of Computer Engineering, Koc University, Istanbul, Turkey
4. Research Center for Translational Medicine (KUTTAM), Koc University, Istanbul, Turkey
5. Cancer Systems Biology Laboratory (CanSyL-METU), Ankara, Turkey

Molecular alterations on genome accumulate through time and disrupt cellular functions and lead to diseases like cancer. One of the deadliest type of brain tumor, Glioblastoma Multiforme is well known for its molecular heterogeneity, which makes the disease as incurable. Patients still have not been efficiently grouped for precision therapy and the survival has been remained very low. In this study, we aimed to decrease the heterogeneity among GBM patients from The Cancer Genome Atlas (TCGA), classify the patients and propose therapeutic hypothesis for patient groups by using patient mutation profiles. We therefore implemented a systems level approach using three dimensional (3D) spatial organization of the mutations (mutation patches), organization of mutated proteins in patient specific protein interaction networks, and drug responses of the GBM cell lines. In conclusion, we found approximately 10% of the mutations are located in patches. Oncogenes statistically tend to have multiple patches that are relatively small, whereas tumor suppressors prone to have a small number of very large patches. Moreover, different patches in the same protein are often located at different domains that can mediate different functions. Mutation patches decreased the inter-patient heterogeneity while network guided analysis classified patients into five groups based on similarity in their affected pathways. These patient groups have a set of signature mutation patches and each group has a significant association with patient survival. By integrating drug sensitivity data through signature patches, we inferred potential therapeutics for each patient group. According to our results, Pazopanib can be effective for Group 3 yet Group 2 can be resistant to inhibition of ATM which is a mediator of PTEN phosphorylation. We believe that from mutations to networks and eventually to clinical and therapeutic data, this study provides a novel perspective in the network-guided precision medicine.

Keywords: 3D Mutation Patch; Protein-Protein Interactions; Patient-Specific Network Modelling; Precision Medicine

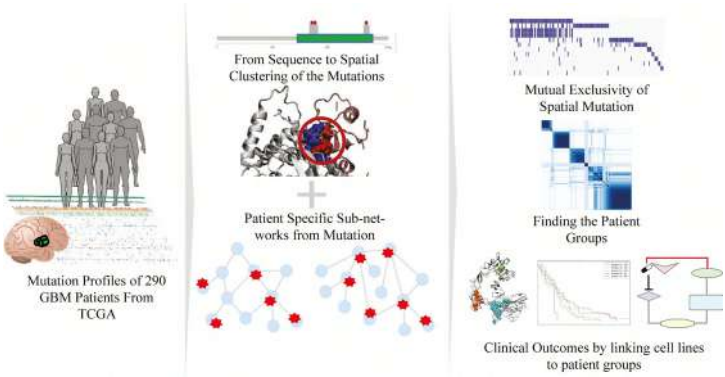


Figure 1. Overview of the Method.

Corresponding Author's Address: Assoc. Dr. Nurcan Tuncbag Office Address: Middle East Technical University Informatics Institute B-204 Cankaya/Ankara E-mail: ntuncbag@metu.edu.tr URL:http://mistral.ii.metu.edu.tr/

MOLECULAR DYNAMICS SIMULATION STUDY ON THE INTERACTIONS BETWEEN DNA& A CONJUGATED POLYELECTROLYTE (CATIONIC OLIGOTHIOPHENE)

Nehir NALINCI BARBAK, Erman KIBRIS, and Nuran ELMACI IRMAK

The absorption spectra of the cationic polythiophenes are shifted to the red when single-stranded DNA (ssDNA) is added, or the color changes in the solution are visible to the naked eye, so that they can be used as a tool for DNA sensor, DNA cleavage reaction and theranostic polyplex applications. The red shift or color change was explained by the fact that the ssDNA molecule leads to a conformational change on the polythiophene, but the form of structural change is not known clearly (i.e. flattening, twisting, stacking, etc.) [1-3].

In this study, molecular dynamics simulations of complexes formed by ssDNA chains with different nucleotides and oligothiophene containing cationic side groups were performed to enlighten the experimental studies. The essential thing for a molecular simulation is the force field definition of the system of interest. For this purpose, the force field parameters of oligothiophene which are not present in the current databases (CHARMM, AMBER, GROMOS), were generated by comparison of quantum chemical calculations and molecular mechanical calculations. The interactions between the oligothiophene and the ssDNA (electrostatic N-O, S-H / O-H bonds, etc.) were analyzed to determine the nature of the conformational change on the oligothiophene when ssDNA is added.

Keywords: Polymer, DNA, Polyelectrolytes, Cationic Polythiophene, Molecular Dynamic Simulation, Computational Chemistry

References: [1] Zheng, Weiming, and Lin He. 2014. "Quantitative Measurements of Thermodynamics and Kinetics of Polythiophene–DNA Complex Formation in DNA Detection." Biomaterials Science 2 (10): 1471.
[2] Rubio-Magnieto, Jenifer, Elias Gebremedhn Azene, Jérémie Knoops, Stefan Knippenberg, Cécile Delcourt, Amandine Thomas, Sébastien Richeter, et al. 2015. "Self-Assembly and Hybridization Mechanisms of DNA with Cationic Polythiophene." Soft Matter 11 (32): 6460–71
[3] Preat, Julien, David Zanuy, Eric A. Perpete, and Carlos Aleman. 2011. "Binding of Cationic Conjugated Polymers to DNA: Atomistic Simulations of Adducts Involving the Dickerson's Dodecamer." Biomacromolecules 12 (4): 1298–1304.

Corresponding Author's Address: Department of Chemistry, Izmir Institute of Technology, 35430 Urla İzmir. nehirnalinci@iyte.edu.tr

DYNAMIC ALTERNATIVE SPLICING EVENTS IN THE FRONTAL CORTEX DURING LATE ADOLESCENCE-EARLY ADULTHOOD PERIOD AND IMPLICATIONS FOR SCHIZOPHRENIA

Kübra Çelikbaş[1], Kerem Mert Şenses[2], Ali Osmay Güre[1,3], Timothea Toulopoulou[1,4]

1. Bilkent University, Department of Neuroscience
2. Zonguldak Bülent Ecevit University, Department of Molecular Biology and Genetics
3. Bilkent University, Department of Molecular Biology and Genetics
4. Bilkent University, Department of Psychology

Alternative splicing (AS) or differential exon usage (DEU) is a regular process after gene expression and it contributes to the diversity of the genome by generating multiple protein isoforms. According to recent studies, the majority (92-94%) of all human multi-exon genes undergoes AS [1] and brain, especially the neocortex, has the highest number of AS events compared to other tissues [2]. While contributing to the complexity of the brain, AS may lead to neuropsychiatric disorders such as schizophrenia or autism if dysregulated [3]. Although there are many studies investigating the possible roles of AS in the function of specific neuron types and during neurogenesis, there is no study investigating AS changes in the brain of an individual during different life periods. In this study, we have analyzed the publically available Affymetrix Human Exon 1.0 ST array data (GSE25219)[4] by AltAnalyze and R, and investigated the changes in DEU of normal brain samples during late adolescence-early adulthood (LAEA) period and compared the DEU data of this period to early infancy, early childhood, and young, middle and late adulthood since many of the neuropsychological disorders first appear during LAEA. We only focused on the frontal cortex region since its importance is implicated in several neuropsychological disorders. Our results revealed that many neuropsychiatric disorder-related genes show dynamic AS changes during LAEA period compared to other periods. Among these genes there are some schizophrenia-related genes (PCDH9, ANKR39) which were previously associated with schizophrenia by genome-wide association studies (GWAS), and our results suggest that DEU of these genes may be an important mechanism in the disease context. We also found novel genes which are not previously implicated to be schizophrenia-related but can be important candidates since they are involved in the disease-related pathways.

Keywords: Alternative Splicing; Differential Exon Usage; Schizophrenia; Frontal Cortex

References: [1] Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, & Burge CB. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008 Nov 27; 456(7221),470–476. PubMed PMID: 18978772
[2] Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-

throughput sequencing. Nat Genet. 2008 Dec; 40(12):1413-5. PubMed PMID: 18978789
[3] Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. Neuron. 2006;52(1):93–101. PubMed PMID: 17015229
[4] Kang HJ, Kawasawa YI, Cheng F, Zhu Y et al. Spatio-temporal transcriptome of the human brain. Nature 2011 Oct 26;478(7370):483-9. PubMed PMID: 22031440

BIOINFORMATICS BASED APPROACH TO DESIGN A THERMOPHILIC P450 FOR INDUSTRIAL BIOCATALYSIS

Ekin Kestevur Doğru[1], *Nur Başak Sürmeli*[1]

1. İzmir Institute of Technology, Bioengineering Department

Enzyme catalyzed biosynthesis of steroidal drugs is an important process for pharmaceutical manufacturing. Cytochrome P450 (P450) monooxygenases are important for hydroxylation of steroid structures because they can catalyze the oxidation of inactive carbon bonds with high selectivity and efficiency. CYP119 is an acidothermophilic P450 from Sulfolobus acidocaldarius, which can be used as biocatalyst for industrial production since it shows activity at high temperature and low pH conditions[1]. In this work we aim to utilize CYP119 for selective hydroxylation of progesterone, which is not the original substrate of CYP119, for production of precursor molecules of important hormones like cortisone and aldosterone. Crystal structure of CYP119 (PDB ID: 1F4T) was used for selecting residues that will be mutated according to structural alignment with other CYPs that can catalyze progesterone hydroxylation naturally. Progesterone-docking performed with CYP119 to identify residues that create clashes with substrate. We also used COACH-D server [2] for ligand-binding site prediction and selected consensus binding residues with progesterone-hydroxylating P450s. Finally, 12 residues were selected (69, 151, 153, 155, 205, 208, 209, 213, 214, 254, 257, 354) and mutated with PyRosetta program[3] to Gly, Glu, Phe, Met, Ala, His, Arg and Ile. Progesterone-docking was performed by using DockMCM Protocol of PyRosetta. We used two different starting coordinates of progesterone for docking since progesterone can bind to two different positions in P450s. Docking results were eliminated according to their energy scores. Best mutants were used for creating double mutants and second round of docking and elimination process was performed with using double mutant enzymes. Triple/quadruple mutants will be designed to obtain efficient substrate binding and selective hydroxylation of progesterone

Keywords: Rational Design, CYP119, Progesterone

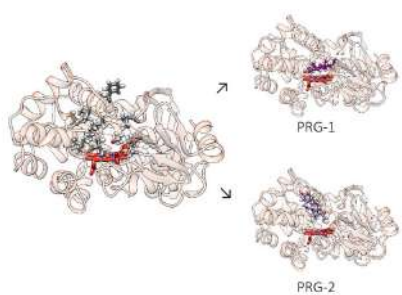


Figure 1. Crystal structure of CYP119 showing mutated residues and two different positions of progesterone in CYP119.

References: [1] Park SY, Yamane K, Adachi S, Shiro Y, Weiss KE, Maves SA, Sligar SG. Thermophilic cytochrome P450 (CYP119) from Sulfolobus solfataricus: high resolution structure and functional properties. J Inorg Biochem. 2002 Sep 20;91(4):491-501. PMID:12237217.
[2] Wu Q, Peng Z, Zhang Y, Yang J. COACH-D: improved protein–ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Res. 2018 Jul 2; 46: W438–W442. PMID: 29846643.
[3] Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics 2010 Mar 1;26(5):689-691. PubMed PMID: 20061306.

Corresponding Author's Address: IZTECH, Bioengineering Department, K003 Gülbahçe/ İzmir
e-mail: ekinkestevur@iyte.edu.tr

PREDICTION OF TRANSMEMBRANE REGIONS OF G PROTEIN-COUPLED RECEPTORS USING MACHINE LEARNING TECHNIQUES

Muazzez Çelebi Çınar[1], *Çağdaş Devrim Son*[1], *Tolga Can*[1]

1. Middle East Technical University

G protein-coupled receptors (GPCRs) are one of the largest and the most significant membrane receptor family in eukaryotes [1]. They transmit extracellular stimuli inside the cell by undergoing conformational changes. GPCRs can recognize a diversity of extracellular ligands including hormones, neurotransmitters, odorants, photons, and ions. These receptors are associated with a variety of diseases in humans such as cancer and central nervous system disorders, and can be proclaimed as one of the most important targets for the pharmaceutical industry[2]. A GPCR spans the cell membrane seven times. Structurally, they have seven transmembrane helices where essential regions such as ligand binding sites, actuator protein (e.g. G protein) binding sites and cholesterol binding sites are located [3]. Data on membrane protein topology are lacking owing to the technical and experimental limitations resulting from unstable environment of the membrane. In UniProt, which is a freely available database of protein sequences and structural and functional information, only 29 GPCRs among the thousands have experimentally solved transmembrane (TM) region data[4]. The topology information of other membrane proteins is provided using the TMHMM prediction tool, which is based on hidden Markov models [5]. However, TMHMM fails to predict the total number of TM regions for 6 of the 29 experimentally determined GPCRs. With this thesis study, we try to develop a GPCR-specific TM prediction algorithm using machine learning techniques. The algorithm is based on hydrophobicity of each amino acid in the protein sequence and the secondary structure information. As hydrophobicity scale, both Moon-Fleming and Kyte-Doolittle hydrophobicity scales are implemented separately [6]. The secondary structures are derived from the JPred server. With this algorithm, we obtain more than 85\% accuracy with higher true positive rate. This study can shed light on other scientific research and facilitate structure-based drug discovery by opening up further potential therapeutic opportunities for many severe diseases.

Keywords: GPCR; Transmembrane Regions; Machine Learning

Measurements	Values
Correctly Classified Instances	86.57 %
Precision	0.98
Recall	0.87
F-measure	0.92
ROC Area	0.88

Table 1: Evaluation results of 10-fold cross validation of SMO performed with the training set, which is generated using mfHydrophobicity and 17-residue sliding windows

References:[1] Liao, Z., Ju, Y., & Zou, Q. (2016). Prediction of G protein-coupled receptors with SVM-prot features and random forest. Scientifica, 2016. PubMed PMID: 27529053.
[2] Zhang, D., Zhao, Q., & Wu, B. (2015). Structural studies of G protein-coupled receptors. Molecules and cells, 38(10), 836. PubMed PMID: 26467290.
[3] Milligan, G. (2001). Oligomerisation of G-protein-coupled receptors. Journal of Cell Science, 114(7), 1265-1271. PubMed PMID: 11256993.
[4] UniProt: the universal protein knowledgebase. Nucleic acids research, 2016, 45.D1: D158-D169. PubMed PMID: 27899622.
[5] https://www.uniprot.org/help/transmem
[6]https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/midas/hydrophob.html

Corresponding Author's Address: e171647@metu.edu.tr

GENE EXPRESSION HETEROGENEITY IN AGING HUMAN BRAIN

Ulaş Işıldak[1], *Mehmet Somel*[1], *Janet M. Thornton*[2], *Handan Melike Dönertaş*[2]

1. Department of Biological Sciences, Middle East Technical University, 06800, Ankara, Turkey.
2. European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.

Aging is characterized by gradual decline in maintenance and repair mechanisms, accompanied by a gradual accumulation of stochastic genetic and epigenetic alterations [1]. This accumulation may cause stochastic deregulation of gene expression over time, leading to increased gene expression heterogeneity between individuals [2]. By analyzing 19 transcriptome datasets across diverse human brain regions covering whole lifespan, we observed a consistent increase in gene expression heterogeneity during aging (20 to 98 years of age), but not in development (0 to 20 years of age). We also found that the genes showing consistent heterogeneity increase during aging are associated with the pathways and biological processes that are related to longevity (e.g. autophagy, mTOR signaling) and neuronal function (e.g. axon guidance, postsynaptic specialization). Furthermore, the number of regulators (miRNAs and transcription factors) is also positively associated with heterogeneity increase, implying that gene regulation might be related to underlying mechanism. Overall, our results showed that human brain aging is associated with increased gene expression heterogeneity, which is associated with multiple lifespan and disease-related pathways. We also showed that increased heterogeneity is not only driven by the general effect of time, but it is a specific effect of the aging process.

Keywords: Aging; Development; Gene Expression; Transcriptome; Heterogeneity; Human; Brain

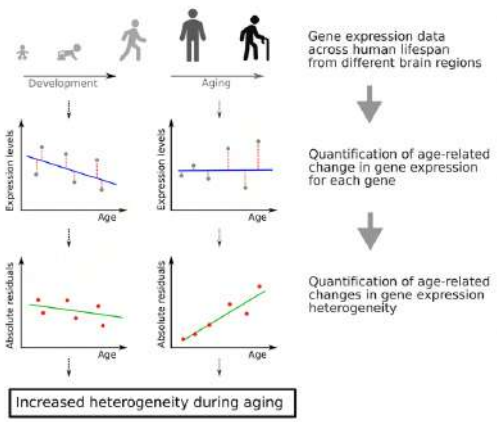


Figure 1. Overview of the study.

References: [1] Lu T, Pan Y, Kao S-Y, Li C, Kohane I, Chan J, et al. Gene regulation and DNA damage in the ageing human brain. Nature. 2004;429(6994):883–91. PubMed PMID: 15190254.
[2] Somel M, Khaitovich P, Bahn S, Pääbo S, Lachmann M. Gene expression becomes heterogeneous with age. Current Biology. 2006;16(10). PubMed PMID: 16713941.

Corresponding Author's Address: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
E-mail: melike@ebi.ac.uk
Tel: + 44 (0) 1223 49 4363

TCGA DRUG SUMMARY INTERFACE

Basak Abak[1], *Tugba Onal-Suzek*[1,2]

1.Department of Computer Engineering, Mugla Sitki Kocman University, Mugla, Turkey, 48000
2. Bioinformatics Graduate Program, Mugla Sitki Kocman University, Mugla, Turkey, 48000

Abstract:
TCGA[1] is a publicly available resource of 11,000 patients with tumor tissue and matched normal tissue. Public patient based genomic data repositories like TCGA are vital for in-house pipeline development and in- silico hypothesis generation towards development of new therapeutics. Although TCGA contains sparse data for survival times and drugs, the information provided is enough to get statistically significant results. However, the drug names in the clinical records are embedded as free texts in TCGA data. Our aim for this study was, for each cancer type in TCGA, to implement a user friendly interface enabling the download of a summary table of 1) the list of drugs 2) and for each drug how many patients survived. For this purpose, when a user clicks on the chosen drug, the list of patients with their drug, the survival information and smoking status were listed. To generate this interface, first, clinical and drug data of each cancer's patients are downloaded from TCGA (The Cancer Genome Atlas) database. We joined these two dimensions of data based on the bcr_patient_barcode and tumor type and filtered the chemotherapy/ immunotherapy/molecular therapy drugs. As the next step, for each drug, the most active protein target name is determined by using the NCBI Eutils R packages and javascript scripts to scrap the results of javascript pop-ups from the PubChem BioAssay[2] database for each drug. Most active protein tarrget from PubChem BioAssay database was incorporated to the data table. At the backend of this user interface, our algorithm provides the user filtering options by drug name and downloading the summary table in csv format. Web User Interface was developed using Rshiny packages.

Keywords: TCGA; Drug Data; PubChem BioAssay; R; Javascript

References: [1]https://bioconductor.org/packages/release/bioc/vignettes/TCGAbiolinks/inst/doc/clinical.html

[2] https://www.ncbi.nlm.nih.gov/pcassay/

Corresponding Author's Address:
basakkabakk@gmail.com
tugbasuzek@mu.edu.tr

A CUSTOMIZED FORCE-DIRECTED LAYOUT ALGORITHM WITH GENETIC ALGORITHM TECHNIQUES FOR BIOLOGICAL GRAPHS

Fırat Aksoydan[1], *Mehmet Volkan Atalay*[2], *Rengül Çetin Atalay*[1,2]

1,2.Department of Computer Engineering, Middle East Technical University, Ankara, TURKEY
3. Graduate School of Informatics, Middle East Technical University, 06800, Ankara, Turkey

Graphs can be used to represent various domain-specific information. By using vertices to symbolize the data items and edges to represent the relation between these items, various type of information can be represented by the graphs. Biological graphs which consist of components such as genes, proteins, and enzymes; have great importance in bioinformatics. Some of the biological graphs contain vertices that represent the enzyme structure. In these type of graphs, there could be some vertices which have clustering information dedicated to Enzyme Commission (EC) number, which is the numerical classification of an enzyme. With the help of these EC numbers, clustering can be conducted. The vertices that belong to the same EC class are members of the same cluster, proportional with the distance in the distance tree.

In our previous study, we described EClerize[1] which is a customized and improved Kamada Kawai[2] force-directed algorithm to visualize pathways that contain nodes with attributes as EC numbers. EClerize creates clusters of nodes with enzymes that belong to the same EC class. Here, in our study EClerize Type GA, our purpose is to avoid the local optima and obtain global optimum solutions for EClerize during graph drawing. By use of a well-known heuristic technique, Genetic Algorithm(GA), we integrate undirected graph layout drawing with GA. We have used the previous study EClerize as a fine-tuner on GA. The genetic algorithm draws strength from the diversity for providing global optima, and the mutation and the crossover are the most important resources of the diversity. In our study, 5 techniques in the mutation phase and 2 techniques in the crossover phase are employed. In mutation, vertices of a selected graph are moved randomly within a limited area or selected edges/vertices are exchanged according to the routines of the selected mutation technique. In the crossover, the operation of exchanging vertices is performed between two selected graphs. In our study, the aesthetic criteria of undirected graphs are served as fitness measurements of GA. In each iteration, to measure how well graphs are drawn, fitness values of graphs are calculated by 6 different fitness measurements ranging from the number of edge crossings to the size of the drawing area. Overall relative fitness values are used to choose parent individuals.

We have applied our study to 3 pathways and the results are better than those of the base study EClerize with respect to certain some measurable fitness criteria with a reasonable longer execution time. From the perspective of global optimum, as genetic algorithm promises, now we can reach better results to draw biological graphs

whose vertices are associated with Enzyme Commission attributes. From the perspective of the graph drawing studies to the best of our knowledge, our work is the first one with the implementation of a GA, a force-directed algorithm and some clustering techniques altogether. There are several studies which combine GA with some force-directed algorithms, however, none of them considers clustering.

Keywords: Graph Visualization, Clustering, Force-directed Graph Layout, Enzyme Commission Numbers, Genetic Algorithm

Table 1: A Comparison Between Results of Our Study and Base Study on above datasets.

Fitness Type	*Signaling by EGFR*		*Signaling by ERBB2*		*Visual phototransduction*	
	Native EClerize	Our Study	Native EClerize	Our Study	Native EClerize	Our Study
Edge Length Deviation	161.65	153.08	190.94	157.21	161.93	158.66
Edge Crossing	322.00	164.00	686.00	370.00	292.00	138.00
Inter-Cluster Distance	3927.00	5228.09	2308.56	3254.40	3795.35	3268.54
Execution Time (ms)	1300	8000	5300	27500	1100	7400

References: [1] H.F. Danaci, R. Cetin Atalay, V. Atalay, "EClerize: A customized force-directed graph drawing algorithm for biological with EC attributes", Int. Journal of Bioinformatics and Computational Biology, Vol. 16, No. 04, 1850007 (2018).
[2] T. Kamada and S. Kawai. An algorithm for drawing general undirected graphs. Inform. Process. Lett., 31:7,15, 1989

Corresponding Author's Address: aksoydan.firat@metu.edu.tr

A WEB BASED DECISION SUPPORT TOOL FOR GASTROINTESTINAL SUBMUCOSAL TUMORS

Asım Leblebici[1], Göksel Bengi[2], Emine Acar[1,3], Serhat Tozburun[4], Müjde Soytürk[2], Ender Ellidokuz[2]

1. Dokuz Eylul University, Institute of Health Sciences, Department of Translational Oncology,
2.Dokuz Eylul University, School of Medicine, Department of Gastroenterology,
3.Izmir Katip Celebi University, Ataturk Training and Research Hospital, Department of Nuclear Medicine,
4.Izmir Biomedicine and Genome Center

In this study, a web based application was realized by using a real data set in order to establish a decision support system for gastrointestinal submucosal tumors (Gastrointestinal stromal tumor & Leiomyoma tumor). In the application, classification algorithm with decision tree, which is one of the data mining methods, is used and it has been found that it provides accurate classification performance in 89%.

The web-based and dynamic operation of this application is provided with the Shiny package of R sofware environment. It is shown that the results can be calculated by taking the input values from the users in the web environment via the interface created with this package. In addition, it is explained that it will enable the users to make the analysis to be more dynamic and visual with R.

The presented study is performed on a real-life data set, which is gathered from texture analysis, allows this analysis to be performed remotely by accessing users over the web instead of performing the analysis on the local computer, and has not been previously performed on the gastrointestinal submucosal tumors with the R program.

The disadvantage of the study is that the number of patients in the gastrointestinal submucosal tumors dataset used is not very large. This can be overcome by increasing the relevant patient data.

Keywords: Texture Analysis, Gastrointestinal Submucosal Tumors, Decision Tree, Shiny Web Ui.

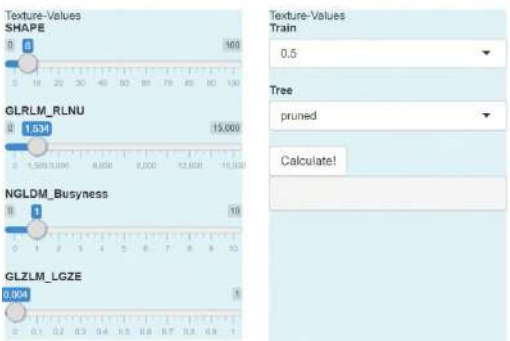


Figure 1. Shiny Web UI for gastrointestinal submucosal tumor

References: [1] Chang, W., Cheng, J., Allaire, J. J., Xie, Y. ve McPherson, J., 2015, Shiny: Web Application Framework for R, http://CRAN.R-project.org/package=shiny.
[2] Kartal E, Balaban, E., "M. E. "R ile Veri Madenciliği Uygulamaları", Çağlayan Kitabevi, 1st edition, 2016

Corresponding Author's Address: asim.leblebici@gmail.com.

THE EFFECTS OF SEQUENCING TECHNOLOGY ON THE AMPLICON BASED MICROBIOME ANALYSIS

Deniz Ece Kaya[1], *Eray Şahin*[1,2], *Orhan Özcan*[1,2], *Osman Uğur Sezerman*[1,2]

1. Acibadem Mehmet Ali Aydinlar University School of Medicine Dept. of Biostatistics and Bioinformatics,Istanbul,
2. Epigenetiks Genetik Biyoinformatik Yazılım A.Ş., İstanbul

The second genome that effect the human health is the microbiome[1]. The importance of the microbial diversity is well understood but we have still could not handle several biases in whole microbiome analysis. In order to get trustworthy microbiome analysis, researchers are developing methods for the various DNA extraction protocols, different kinds of library preparations, different kinds of microbiome analysis pipelines and now different kinds of sequencing techniques. In this research we have presented how sequencing method effects whole microbiome analysis. We have totally 66 samples (50 Illumina and 16 Oxford Nanopore Technology (ONT)) for observing the sequencing technique bias on the microbiome analysis. Naïve Bayes learning and the SVM Learner joined for scoring to observe whether a sequencing technique implementing any bias on the microbiome OTU tables. We have presented that there is a bias on sequencing techniques especially for the Bacteroidetes and Proteobacteria distribution in the OTU tables.

Keywords: Illumina Sequencing; ONT Sequencing; Microbiome Analysis

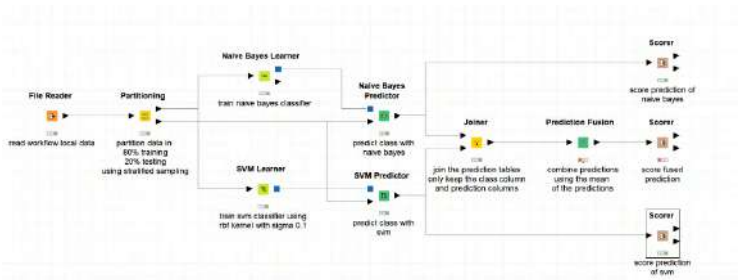


Figure 1. The pipeline used for the learning of sequencing technique effect on microbiome analysis. The microbiome OTU tables converted into clustering table where addition of a clustering column: I (Illumina) or O (ONT).

Correct classified: 7	Wrong classified: 4
Accuracy: 63.636 %	Error: 36.364 %
Cohen's kappa (κ) 0	

Table 1: Prediction table of clustering scoring. 50 Illumina and 16 ONT reads partitioned by 80% versus 20%. The test partition is scored with respect to pipeline showed in Figure1

References: [1] Grice E.A., Segre J.A. The human microbiome: Our second genome. Annu. Rev. Genomics Human Genet. 2012;13:151–170. doi: 10.1146/annurev-genom-090711-163814. [PMC free article] [PubMed] [CrossRef] [Google Scholar]

Corresponding Author's Address: ugur.sezerman@acibadem.edu.tr http://epigenetiks.com.tr/

USING METAGENOMICS AND MACHINE-LEARNING TO UNDERSTAND THE ORAL HEALTH

Emrah Kırdök[1], *Andres Aravena* [2]

1. Mersin University, Department of Biotechnology
2.Istanbul University, Department of Molecular Biology and Genetics

Our bodies host a wide variety of commensal microbes from the oral cavity to the gastrointestinal tract. In a healthy oral cavity, Firmicutes, Proteobacteria, Fusobacteria, Bacteroidetes, and Spirochaetes species constitute 93% of the total oral microbial profile. In a healthy state, the abundance of those microbes stays in specific proportions. However, significant deviations from this balance affect the bacterial diversity in the oral cavity. This process is called dysbiosis and could indicate numerous oral and non-oral diseases[1]. Traditional microbiological methods help to find the bacterial species associated with dysbiosis. However, these methods capture a small portion of the microbial composition. Metagenomics proposes high-throughput methods to study the patterns of dysbiosis by comparing the microbial abundances of thousands of taxa between healthy and disease conditions. In this study we used shotgun libraries of 22 healthy, caries and periodontitis cases to (i) find the bacterial composition in healthy and dysbiotic salivary microbiome, (ii) find a set of bacterial species that are associated with a particular disease state, (iii) use machine learning approaches to predict the oral health of the individual, given the bacterial abundances. To do this, we aligned DNA sequences from the salivary microbiome to a database that contains the full-length genomes of reference bacterial species. Then, we calculated the absolute abundances of each microbe; that is, the number of DNA reads that correspond to a specific genome. We have modeled the abundance shifts between healthy and disease states by fitting a beta- binomial probability distribution[2]. With this method, we have identified a set of bacterial species that are associated with caries and periodontitis disease configurations. Next, we used machine learning methods[3] to train a classifier to predict oral health using bacterial abundances. We have predicted the healthy oral state with 80% success rate.

Keywords: Oral Microbiome; Metagenomics; Machine-Learning

References: [1] Gao L, Xu T, Huang G, Jiang S, Gu Y, Chen F. Oral microbiomes: more and more importance in oral cavity and whole body. Protein and Cell. 2018; 9:488-500 PubMed PMID: 29736705
[2] Martin, B.D., Witten, D., Willis, A.D.. Modeling microbial abundances and dysbiosis with beta-binomial regression. arXiv:1902.02776. 2019
[3] Beck D, Foster JA. Machine learning techniques accurately classify microbial communities by bacterial vaginosis characteristics. PLoS One. 2014 9(2):e87830. Pubmed PMID: 24498380

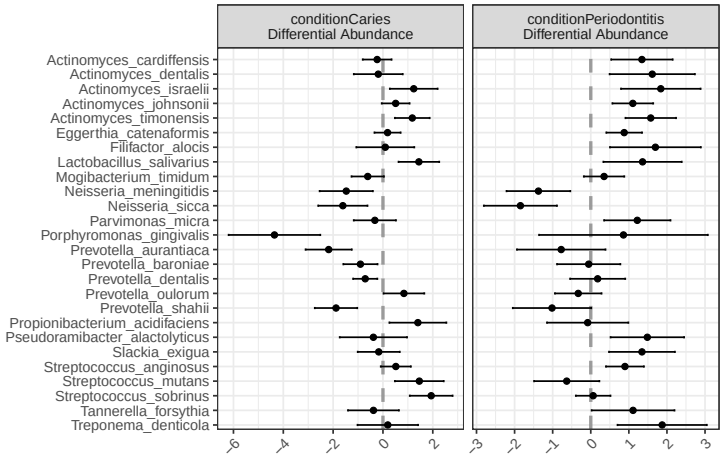


Figure 1. Significant taxa that changed in dysbiotic states. The horizontal axis shows the differential abundance change compared to control group, and the vertical axis shows the bacterial species. Condition label definesthe different dysbiotic states. Bacterial abundances of healthy state are normalized to zero. In this figure we see the bacterial abundance change, relative to healthy sample. Lines represent 95% confidence intervals.

Corresponding Author's Address: emrahkirdok@mersin.edu.tr

NUCLEICACID CONTAMINANTS FROM LIBRARY PREPARATION

Deniz Ece Kaya1, Orhan Ozcan1,2

1Acibadem Mehmet Ali Aydinlar University School of Medicine Dept. of Biostatistics and Bioinformatics, Istanbul,
2. Epigenetiks Genetik Biyoinformatik Yazılım A.Ş., İstanbul

In order to get trustworthy sequencing reads, researchers qualified libraries by a fluorometric method (e.g. Qubit, PicoGreen) or by qPCR. Although the library preparation for the Illumina sequencing is quite efficient and reproducible, the enzymes used in library preparations are not DNA-free. NCBI Sequence Read Archive (SRA) database have been analysed (cancer, metagenomics, microbiome and pure cultures reads). We have recovered two distinct bacteriophages (5463bp) coming in almost all Illumina reads. These two bacteriophages are important as some of the researchers have already misguided by their presence. Although, these two bacteriophages are Enterobacter phage researchers have already showed them in Clostridium (Accession: WP_010791285) and Chlamydia (WP_002934172). DNA and RNA contamination coming from the library preparations will not be problem whenever the researchers mapped out these two sequences from their raw reads.

Keywords: Illumina Sequencing; Library preparation; Nucleic acid contaminations

Figure 1. Two Enterobacter bacteriophages coming from library preparation

	Cancer Reads	Microbiome	Pathogenomics	Viromics
Phage1	+	++	+	+++
Phage2	+	++	+	+++

Table 1. Relative coverages of phages with respect to different library preparations

Corresponding Author's Address:

Cryptomicromicrobiology@gmail.com http://epigenetiks.com.tr/

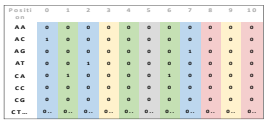
GENOMIC DATA COMPRESSION BY DEEP LEARNING METHODS LSTM AND CNN

Emre Taylan DUMAN[1], Yusuf Sinan AKGÜL[2], Pınar PIR[1]

1. Gebze Technical University, Department of Bioengineering, Çayırova/KOCAELI
2. Gebze Technical University, Department of Computer Science, Çayırova/KOCAELI

Sequencing of the genome is one of the milestones of personalized medicine and healtcare research. One of the main challenge of the genomics is storage and the handling of the huge amount of data. Because of the accelerating increase in data produced by new sequencing technologies, the challange of data storage is growing[1]. Sequence data is usually produced in fastq format, which is a text file that contains sequence reads and quality scores. Sequence reads contain five characters that represent four nucleotides ('A','C','G','T') and 'N' letter for the non-readable nucleotides[2]. Phred quality scores of the fastq file is sequence of ASCII characters assignedby the sequencing device, which are logarithmically calculated base-calling error probabilities. Phred scores represented by single ASCII character can be converted into error rates. These scores are relatively random and can contain 40 different characters. A single run with a commercial sequencing device can produce 1 to 5 Tb of raw data. That makes over 100 Tb in one experiment[3]. Storage of this amount of data requires special compressing techniques capabale of losseless size reduction. Because of the nature of the genomic data, there is no tolerance for losing any single nucleotide reads. Commonly used text compressing techniques allow losseless compression of the text based files such as gzip[4]. But these text compression techniques uses word or letter occurance probabilities for changing their bit encoding. And this technique shows low performance because of the randomness and the wide character scale of the phred scores. Because of this issue, there are variety of different compressing algorithms which use different techniques and approaches for size reduction of genomic data[5]. In this work, deep learning techniques were applied to the sequence data for prediction of the quality scores by using the information on the nucleotide type and location of the read in a fragment. Two different deep learning approaches was chosenfor bench marking. CNN (Convolutional Neural Networks) is a technique that is mainly used for image recognition and prediction[6]. CNN requires two dimensional conversion of the one dimensional genomic data. Second approach, LSTM (Long Short Term Memory), is generally used for one dimensional data as text/voice recognition or translation between languages. LSTM uses previous word relationships to predict new words or letters[7]. Aim of this work is to predict quality scores to be able to for omit storage of quality scores in sequence data files and to use more general compression techniques to reduce size of the files.

Keywords: Compression; Deep Learning; CNN; LSTM; Genomic Data.



References: [1] Kahn, Scott D. On the future of genomic data. science, 2011, 331.6018: 728-729.
[2] Cock, Peter JA, et al. The Sanger FASTQ file format for sequences with quality scores, and the
Solexa/Illumina FASTQ variants. Nucleic acids research, 2009, 38.6: 1767-1771.
[3] Stephens, Zachary D., et al. Big data: astronomical or genomical?. PLoS biology, 2015, 13.7: e1002195.
[4] Shanmugasundaram, Senthil; LOURDUSAMY, Robert. A comparative study of text compression algorithms. International Journal of Wisdom Based Computing, 2011, 1.3: 68-76.
[5] Brandon, Marty C.; WALLACE, Douglas C.; BALDI, Pierre. Data structures and compression algorithms
for genomic sequence data. Bioinformatics, 2009, 25.14: 1731-1738.
[6] Krizhevsky, Alex; SUTSKEVER, Ilya; HINTON, Geoffrey E. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. 2012. p. 1097-1105.
[7] Hochreiter, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. Neural computation, 1997, 9.8: 1735-1780.

Corresponding Author's Address: Gebze Technical University Bioengineering 41400 Gebze/Kocaeli Turkey
pınarpir@gtu.edu.tr

IDENTIFYING COMMON PATHOGENESIS OF DISEASES USING LITERATURE MINED GENE INTERACTIONS

Özge Dinçsoy1, Arzucan Özgür2, Ahmet Okay Çağlayan3

1,2. Bogazici University, Computer Engineering
1. Idea Technology Solutions
3.Department of Medical Genetics, School of Medicine, Istanbul Science University

In the genomic era, the elucidation of the relationship among illnesses is a crucial task for medical research. Among these, there are brain related diseases seen in the same person and whether there is a common pathogenesis has been the subject of interest. In this project, specifically, we focused on the relationship between Autism Spectrum Disorder (ASD) and Epilepsy because up to 20% of people with epilepsy also have ASD[1]. Both of them are lifelong pervasive illnesses that reveals in childhood and may trigger other neurological disorders as well. In these patients, having knowledge about the source of the precipitating genetic defect can be helpful in treatments. Here, by using text mining on articles, our aim was to find connections between diseases in gene base and to propose candidate genes for the two illnesses. To do this, we have aimed to make literature research automatically with the help of PubMed database. We have shown which genes are connected with each other. We have made four different search queries consisting of articles on epilepsy, articles on autism, articles on both epilepsy and autism, and articles on either epilepsy or autism. We have worked on 244159 PubMed articles which represent the union query result and marked the genes by using SciMiner[2], a tool for target identification and functional enrichment analysis. We have followed two different approaches, disease weighting and PageRank scoring. In the former approach, we rank genes according to the number of occurrences per document for these two particular illnesses. Then, we inspected the top 100 genes for both illnesses separately and obtained intersection set of those genes. The intersection set contained 7 genes. In the latter approach, we have generated weighted gene graphs from the results of beforementioned four search queries according to the number of documents that hold both genes in an edge. We sort genes according to PageRank algorithm. We have used SFARI Gene for evaluation of the resulting genes of the two approach. From the result of the first approach, 5 out of 7 genes have been found previously linked with autism and we are recommending the rest 2 genes to be examined in further research. For the second approach, the results are shown in Table 1. In conclusion, by using text mining methods, we have obtained genes that are previously associated with diseases. In addition to these genes, we have come up with new genes which are not found autism related in literature and to be investigated in DNA data of patients

Keywords: Autism; Autism Spectrum Disorder; Epilepsy

	Query #1	Query #2	Query #3	Query #4
Total number of genes	3381	1423	484	3764
Number of genes found in SFARI Gene	464	350	157	539

Table 1. Performance result of the second approach.

References: [1] Besag FM. Epilepsy in patients with autism: links, risks and treatment challenges. Neuropsychiatr Dis Treat. 2017 Dec 18;14:1-10. doi: 10.2147/NDT.S120509. PubMed PMID: 29296085; PubMed Central PMCID: PMC5739118.
[2] Hur J, Schuyler AD, States DJ, Feldman EL. SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinformatics. 2009 Feb 2;25(6):838-40.
[3] Abrahams BS, Arking DE, Campbell DB, Mefford HC, Morrow EM, Weiss LA, Menashe I, Wadkins T, Banerjee-Basu S, Packer A. SFARI Gene 2.0: a community-driven knowledgebase for the autism spectrum disorders (ASDs). Molecular autism. 2013 Dec;4(1):36.

Corresponding Author's Address: 1. ozge.dincsoy@boun.edu.tr 2. arzucan.ozgur@boun.edu.tr 3. okaycaglayan@yahoo.com

GUMMING UP THE WORKS: SOMAN AND SARIN EFFECTED DYNAMICS OF HUMAN ACETYLCHOLINESTERASE

Brian J. Bennion[1], Lau EY[1], Fattebert J-L[2], Emigh A[1], Lightstone FC[1], Sevilay Güleşen[3], Şebnem Eşsiz[4]

1.Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory,
2.Center for Applied Scientific Computing, Lawrence Livermore National Laboratory,
3. Computer Engineering Department, Faculty of Engineering and Natural Sciences, Kadir Has University,
4. Bioinformatics and Genetics Department, Faculty of Engineering and Natural Sciences, Kadir Has University

Intoxication of human acetylcholinesterase (hAChE) by organophosphorous pesticides (OPs) and chemical weapon agents (CWA) leads to cognitive deficiencies, seizures, paralysis, and eventually death[1]. 80 classical short 50 ns length molecular dynamics (MD) simulations and quantum mechanics/molecular mechanics (QM/MM) simulations of the apo and soman-adducted forms of hAChE were analyzed to examine the effects on the dynamics and protein structure when the catalytic Serine 203 is phosphonylated. When the hAChE is adducted by soman, the correlation of gorge entrance and back door motions for substrate entrance is disrupted. These motions support the hypothesis that substrate and product can use two different pathways as entry and exit sites. Recently, the sarin adduction to hAChE has been added to the analysis. 40 classical 50 ns long MD simulations of the apo and 40 classical 50 ns long MD simulations of the sarin-adducted forms of hAChE were analyzed to investigate sarin-dependent changes in backbone and sidechain motions. Principal component analysis (PCA) of apo, sarin and soman adducted simulations are analyzed to understand the differences of apo and different adducts effects on the entrance site dynamics of the substrate to the active site.

Keywords: Human acetylcholinesterase (hAChE); Soman; Sarin; Molecular Dynamics; Principal Component Analysis (PCA)

References: [1] Bennion BJ, Essiz SG, Lau EY, Fattebert J-L, Emigh A, Lightstone FC. A wrench in the works of human acetylcholinesterase: soman induced conformational changes revealed by molecular dynamics simulations. PLoS ONE. 2015;10(4):1-31.

Corresponding Author's Address: Brian J. Bennion: Biosciences and Biotechnology Division, Lawrence Livermore National Laboratory, 7000 East Ave, Livermore CA, United States of America. bennion1@llnl.gov https://bbs.llnl.gov/BrianBennion.html Şebnem Eşsiz: Bioinformatics and Genetics Department, Faculty of Engineering and Natural Sciences, Kadir Has University, 34083 Fatih, Istanbul, Turkey sebnem.gokhan@khas.edu.tr http://sites.khas.edu.tr/bioinformatics/research/sebnem-essiz/

GENOME-WIDE EFFECT OF DNA REPLICATION ON NUCLEOTIDE EXCISION REPAIR OF UV-INDUCED DAMAGES.

Cem Azgari[1], Jinchuan Hu[2], Yanchao Huang[2], Yi-Ying Chiou[3],Aziz Sancar[4] and Ogün Adebali[1]

1. Faculty of Engineering and Natural Sciences, Molecular Biology, Genetics and Bioengineering, Sabancı University Orhanlı, Tuzla, 34956,Istanbul
2.Fifth People's Hospital of Shanghai and Institute of Biomedical Sciences, Fudan University, Shanghai 200032, China
3. Institute of Biochemistry, National Chung Hsing University, Taiwan
4. Department of Biochemistry and Biophysics, University of North Carolina School of Medicine, Chapel Hill, North Carolina

Maintaining genome integrity is crucial for healthy cells to avoid cancer. Considering that DNA damages occur approximately 70,000 times per cell per day, repair of these damages is vital for the maintenance of genome stability[1]. On the other hand, replication is the mechanism that causes unrepaired DNA damages to turn into mutations that might lead to cancer. The role of replication on DNA repair in general is yet to be clarified. Recently developed methods Damage-seq and XR-seq map damage formation and nucleotide excision repair events respectively, in various conditions[2]. In this study, we analyze the Damage-seq and XR-seq results of cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone (6-4) photoproducts) [(6-4)PPs] from UV-irradiated HeLa cells synchronized at two stages of the cell cycle: early S phase, and late S phase. We compare these datasets with the localized replication domains of HeLa cells[3]. We aim to reveal how replication domains are influencing the repair preferences. We found out that in both early and late S phased cells, early replicating domains are more efficiently repaired relative to late replicating domains. We aim to investigate a potential differential repair efficiency between leading and lagging strand.

Keywords: Nucleotide Excision Repair; UV Damage; (6-4)PPs; CPDs; XR-Seq; Damage-Seq; Replication

References: [1]Tubbs A, Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. Cell. 2017 Feb 9;168(4):644-656. Pubmed PMID: 28187286.
[2]Hu J, Adebali O, Adar S, Sancar A. Dynamic maps of UV damage formation and repair for the human genome. Proc Natl Acad Sci. . 2017 Jun 27;114(26):6758-6763. Pubmed PMID: 28607063.
[3]Liu F, Ren C, Li H, Zhou P, Bo X, Shu W. De novo identification of replication-timing domains in the human genome by deep learning. Bioinformatics. 2016 Mar 1;32(5):641-9. Pubmed PMID: 26545821.
Corresponding Author's Address: To whom correspondence may be addressed: Sabancı University, Faculty of Engineering and Natural Sciences, Tuzla, Istanbul, 34956 Turkey Tel.: 216-568-7043; E-mail: oadebali@sabanciuniv.edu.

EVOLUTIONARY PERSPECTIVE TO UNTANGLE DNA SEQUENCE ORGANIZATION OF A MULTI-DRUG RESISTANT PLASMID

Faruk Üstünel[1], Terje M. Steinum[1], Andres Aravena[1]

1.Molecular Biology and Genetics Department, Istanbul University

In a previous study we isolated two specimens of a multi-drug resistant plasmid from E.coli cells found in a human host, and from Citrobactreium freundii found in cow host. To understand which characteristics enable its horizontal transfer, we sequenced the two specimens in two NGS libraries. Since the DNA extraction included some parts of the host bacterial genome, plasmid assembly required an ad hoc approach. We screened our reads to determine which ones were likely to be from the plasmid and not from the chromosome, and which ones were probably part of a repeated region. We used two assembly stages with two strategies (overlay-layout-consensus and De Bruijn graph) to determine a single scaffold revealing all places where the sequence organization is ambiguous. There are multiple instances of transposable elements across the plasmid, that restrict us from achieving a conclusive sequence. To untangle this scaffold, we collected a set of similar plasmids' sequences. By comparing our scaffolds with the set of reference plasmids we are able to propose a likely organization, and a set of PCR primers that can validate our proposal. In this presentation we will discuss the bioinformatic strategies used to achieve this result.

Keywords: Genome Assembly; Scaffold Untangling; Transposable Elements In Plasmid Assembly

References: [1] Oppegaard, H., Steinum, T. M., & Wasteson, Y. (2001). Horizontal transfer of a multi-drug resistance plasmid between coliform bacteria of human and bovine origin in a farm environment. Applied and Environmental Microbiology, 67(8), 3732–3734. https://doi.org/10.1128/AEM.67.8.3732-3734.2001

Corresponding Author's Address:
andres.aravena@istanbul.edu.tr

PRAMP: A PREDICTION PIPELINE COMBINING A META-PREDICTOR AND ACCELERATED MOLECULAR DYNAMICS FOR FUNCTIONAL IMPACT OF MUTATIONS

Umut Gerlevik[1], *Aslı Yenenler*[2], *Ugur Sezerman*[1]

1.Department of Biostatics and Medical Informatics, School of Medicine, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey
2.Department of Biomedical Engineering, Faculty of Engineering and Natural Science, Biruni University,Istanbul, Turkey

One crucial step in variant prioritization is functional impact prediction for mutations, which gives critical directive information about the variants in genome/exome data of patients with genetic diseases or complex genetic diseases such as cystic fibrosis, cancers and Alzheimer's disease. Moreover, these pathogenicity information are helpful when diagnosing the patients and trying to enlighten pathogenicity mechanism of diagnosed disease [1]. There are numerous variant effect predictors in the literature, and each has different basis and algorithm. On the other hand, better predictions can be obtained by creating meta-predictors via machine learning methods such as decision tree modeling. We formerly developed Pathogenicity RIsk Detection Algorithm (PRIDA), which is one of those meta-predictors, and it uses the categorical predictions from MutationTaster, Condel and CanDrA[2]. PRIDA had 78.26% accuracy (with Matthews correlation coefficient: 0.61) on an enormous amount of test cases (n=63321, as 21290 pathogenic and 42031 neutral variants)[2]. However, we observed that there are still deficiencies and considered that these inadequencies can be overcome by using structural and functional feautures on protein level [2]. As far as we observed, there is no predictor/pipeline in the literature, which uses such features from complex methods such as molecular dynamics (MD) simulations while evaluating a vast amount of mutations. Therefore, we developed a novel method called PRIDA-aMD Pipeline (PRaMP) shown in Figure 1 after deriving a reliability score from the scores of MutationTaster, Condel and CanDrA, and offering a short accelerated MD (aMD) study for those unrelaibles to finally classify the mutations. To validate this pipeline, three type of cases were selected randomly from the false predictions of PRIDA on the test cases: (1) one predicted by PRIDA as pathogenic but known as neutral, (2) one predicted by PRIDA as neutral but known as pathogenic, and (3) control/support cases as in Table 1. Moreover, analyses of the resulted MD trajectories provide many structural and functional feautures such as RMSD, RMSF, Rg, changes in destabilization tendency ($\Delta\Delta G$) and ligand distances if exist. Although MD is a time-consuming method, there are tricky methods to overcome this such as aMD which lowers activation energy barriers to accelerate conformation changes. Furthermore, we demonstrated that aMD can be used with shorter trajectories instead of MD, and the resulted structural-functional features provided 100% accurate classification on the cases in Table

1. Thus, we suggest involvement of structural and functional features in variant effect prediction processes as in PRaMP.

Keywords: Functional Impact of Mutations; Accelerated Molecular Dynamics; Meta-predictor; Pipeline

Type	Protein Name	PDB ID	Mutation	PRIDA's Prediction	Actual Effect
(1)	EP300	5lkz	C1275W	Pathogenic	Neutral
(3)	EP300	5lkz	L1276F	Pathogenic	Neutral
(3)	EP300	5lkz	A1629V	(not in the dataset)	Pathogenic
(2)	TNR6	1ddf	T241K	Neutral	Pathogenic
(3)	TNR6	3tje	T122I	Pathogenic	Neutral

Figure 1. Workflow of PRaMP, referring PRIDA-aMD Pipeline. T

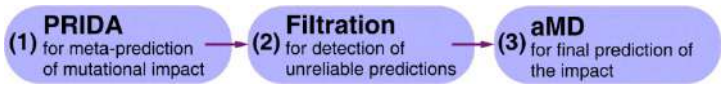


Table 1: List of the cases to validate structural and functional features for the classification of mutations.

References: [1] Saygı C, Alanay Y, Sezerman U, Yenenler A, Özören N. A possible founder mutation in FZD6 gene in a Turkish family with autosomal recessive nail dysplasia. BMC Med Genet. 2019 Jan 14;20:15. PubMed PMID: 30642273.
[2] Gerlevik U, Cingiler O, Ulgen E, Bingol C, Sezerman U. PRIDA: a meta-predictor for functional impact of mutations. In: HIBIT2018 Committees, editors. Student symposium. HIBIT2018: 11th International Symposium on Health Informatics and Bioinformatics; 2018 Oct 25-27; Antalya, Turkey. Antalya: HIBIT2018; 2018. p. 83.

Corresponding Author's Address: umut.gerlevik@gmail.com

ALLOSTERIC TUBULIN BINDING INTERACTIONS IN KINESIN-1

Zeynep Erge Akbaş[1], Burçin Acar[1], Burcu Aykaç Fas[1], Fidan Sümbül[1], Yiğit Kutlu[1], Hamdi Torun[2], Türkan Haliloğlu[1]

1. Polymer Research Center and Chemical Engineering Department, Bogazici University
2. Department of Mathematics, Northumbria University, Newcastle upon Tyne, UK

Molecular motors are key components of the cell responsible for all kinds of movements to realize cellular processes in living organisms. Kinesin-1 is a prototype of molecular motors responsible from carrying cargoes along microtubules [1]. The movement of kinesin on microtubule is mediated by the nucleotide binding (ATP) to the leading motor head of kinesin dimer. Depending on the nucleotide type; ATP or ADP, the state of the neck linker (NL) is respectively either docked (ordered) or undocked (disordered). Switch regions (switch-1 and switch-2), are also responsible for conformational change of NL via ATP hydrolysis. We aim to explore how global dynamics are coupled to the binding interaction of human kinesin-tubulin complex. To this, we integrate Gaussian Network Model (GNM)[2,3] analysis and Molecular Dynamics (MD) simulations with Atomic Force Microscopy (AFM) single-molecule pulling experiments on the wild-type and D72N, S175A and N332A mutants of human kinesin protein in complex with αβ-tubulin (Figure 1). The magnitude of unbinding forces and unbinding rates measured by AFM pulling experiments [4] suggest that N332A/S175A kinesin are effective in lowering the activation energy barrier, whereas the dissociation constant is largest for D72N kinesin in the αβ-tubulin interaction. N332 residue in human resides in between β9-β10 strands of NL; N332 being the latch, docks β10 (V333-T336) on the motor head (G76 (α1), K226 (β7)) and initiates formation of parallel beta-sheet called cover-neck bundle (CNB) through β9-CS (cover strand, β0) interactions[5]. At this event, CS is responsible for generating the force for a walking stroke (power stroke) while N332 and β10 stabilizes the docked state of NL. Forward movement of the trailing head requires the tight binding of NL to the motor head (by N332 latch and β10). D72N mutant forms more hydrogen bonding than wild-type and S175A/N332A between tubulin binding sites (L11; switch-2) and αβ-tubulin (α-tubulin with L11 and switch-2 cluster; β-tubulin with β5L8) as well as between NL and the motor head. This implies a stronger attachment of kinesin to the tubulin as well as of NL to the motor head, which disrupts its dissociation from tubulin during power stroke. In contrast, a decrease in the hydrogen bonding formation between switch-1 and switch-2 residues in D72N, switch regions and also NL being sensitive to the binding and hydrolysis of ATP, implies possibly reduced response to the nucleotide. Moreover, compared to WT and S175A/N332A, the correlation between D72N tubulin binding sites (L11 with α-tubulin and β5L8 with β-tubulin) and tubulin becomes stronger, while the correlations between switch 1 and L11-switch 2 regions and between switch regions and NL are weakened. NL; on the other hand, becomes highly correlated with switch-2 cluster for D72N. Thus, we propose

that the stronger docking of kinesin monomer and decreased ability to hydrolyze ATP as well as higher dynamic cooperativity between kinesin and tubulin result in stronger interactions with tubulin that corroborate the experimental observations. Lower dissociation rate of D72N, which plausibly result in the kinesin motility failure, could be associated with SPG10 disease [6]. The mutation positions being at the hinge sites of the most cooperative mode that integrate kinesin and tubulin dimer rationalize their effects in tubulin binding.

Keywords: Molecular Motors; Kinesin-1; Atomic Force Microscopy; Gaussian Network Model; Molecular Dynamics Simulations; Allostery

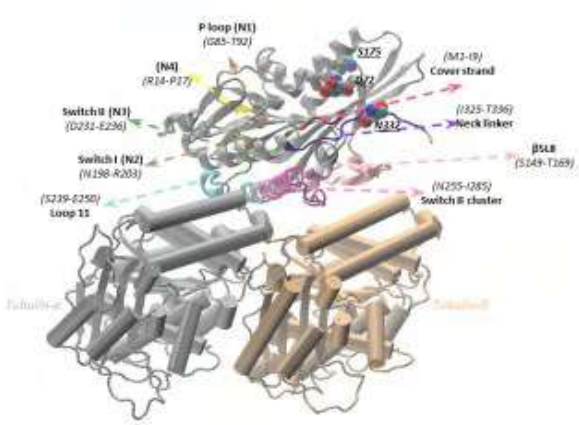


Figure 1. Human kinesin molecule in complex with αβ-tubulin

References: [1] Vale RD and Milligan RA. The way things move: looking under the hood of molecular motor proteins. Science. 2000 Apr 7;288(5463):88-95. PubMed PMID: 10753125.
[2] Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997;2(3):173-81. PubMed PMID: 9218955.
[3] Haliloglu T, Bahar I, Erman B. Gaussian Dynamics of Folded Proteins. Phys Rev Lett. 1997; 79(16):3090-93.
[4] Lee CK, Wang YM, Huang LS, Lin S.Atomic force microscopy: Determination of unbinding force, off rate and energy barrier for protein-ligand interaction.Micron, 2007;38(5):446-61. Pubmed PMID: 17015017.
[5] Hwang, W., M. J. Lang, and M. Karplus. Force generation in kinesin hinges on cover-neck bundle formation. Structure. 2008 Jan;16(1):62-71. PubMed PMID: 18184584.
[6] Karle KN, Möckel D, Reid E, Schöls L. Axonal transport deficit in a KIF5A-/- mouse model. Neurogenetics. 2012 May; 13(2):169-79. PubMed PMID: 22466687
Corresponding Author's Address: Polymer Research Center and Chemical Engineering Department, Bogazici University, Istanbul, Turkey E-mail: halilogt@boun.edu.tr

INVESTIGATING COMMON PATTERNS OF GUT MICROBIOME DYSBIOSIS OVER A RANGE OF COMPLEX DISEASES

Özkan Ufuk NALBANTOĞLU[1,2]

1. Computer Engineering Department, Erciyes University, Kayseri, Turkey
2. Genome and Stem Cell Center (GenKök), Erciyes University, Kayseri, Turkey

The last decade has witnessed the enthusing journey of human microbiome rediscovery, and its relation with human health. The current assumption is that human microbiome, especially the gut microbiome is associated with over 90% of the defined chronic disorders[1]. For healthy individuals, human host and the gut microbiome is in a homeostatic equilibrium, promoting health via interacting metabolic and signalling pathways, over a complex interactome. However, in case of disease, this homeostasis is subject to imbalance either as the cause, a driver or a complication of the disease. The resulting phenomenon is defined as "dysbiosis". Since gut microbiome is a complex ecosystem with thousands of microorganisms, carrying millions of genes, no quantitative traits, biomarkers or measurement indices has been successfully defined to characterize, or even detect dysbiotic states[2]. It is still unclear, and subject of a heated debate, whether dysbiosis is a specific response, forming uniquely for different diseases and individuals, or if it follows certain patterns common to multiple diseases. This study proposes a data driven approach to investigate the latter hypothesis, and the results support that certain diseases exhibit shared dysbiotic components.

The experimental set up is conducted as follows. The 16S rRNA sequencing data which contains the taxonomic profiles of around 16000 individuals' gut microbiomes as well as their metadata were obtained from American Gut Project[3]. For the selected 8 different chronic diseases (Table 1), disease detection was performed using Artificial Neural Network classifier models, taking the gut microbiota taxonomy vectors as input and returning disease/healthy diagnostics. The dysbiosis scores generated for a disease model were later transferred to the remaining disease classifications and decision tree classifiers were trained solely on these scores to classify the other diseases. Therefore, the predictive model trained to evaluate the dysbiosis of one disease is used to predict another disease outcome. The experiments were conducted independently using subsampling to overcome unbalanced subsets, and 20% of the samples were used as test samples. Area under ROC curve was used as the classification performance measure. Table 1 shows that the microbiome features learned to classify a disease can be used to classify other diseases with minimal fine tuning on the classifiers.

Our results suggest that a model trained to learn dysbiosis patterns of a disease can be used to predict another disease accurately to an extent. Therefore, it is plausible to claim that certain chronic disorders share similar pathological mechanisms impacting the homeostasis

of gut microbiome. Further implications of this hypothesis could be searching for the definition of a healthy gut microbiome, as well as the prospective research on the preventive and theuropathic interventions on the microbiome in order to maintain a healthy human-microbiome interactome.

Keywords: Microbiome, Machine Learning, Chronic Diseases, Human Gut Microbiota

Disease		*Diabetes*	*IBD*	*Cancer*	*Thyroid*	*Lung disease*	*Cardiovascular*	*Alzheimer*	*Autism*
Diabetes	0.88	-	0.73	0.69	0.63	0.63	0.69	0.61	0.73
IBD	0.89	0.59	-	0.64	0.53	0.65	0.56	0.60	0.56
Cancer	0.70	0.69	0.56	-	0.69	0.53	0.68	0.65	0.90
Thyroid	0.79	0.73	0.61	0.70	-	0.52	0.59	0.59	0.60
Lung disease	0.80	0.71	0.53	0.56	0.56	-	0.60	0.57	0.63
Cardiovascular	0.69	0.73	0.62	0.66	0.62	0.52	-	0.61	0.73
Alzheimer	0.93	0.57	0.50	0.52	0.54	0.58	0.58	-	0.58
Autism	0.89	0.55	0.61	0.59	0.53	0.51	0.56	0.56	-

Table 1: Detection performance of diseases using predictive models on gut microbiota data. Each row corresponds to a disease and the first column is the AUC value for diagnosis. The remaining columns show the AUC values obtained for the remaining diseases using only the dysbiosis score generated by the disease model of the corresponding row.

References: [1] Durack, Juliana, and Susan V. Lynch. "The gut microbiome: Relationships with disease and opportunities for therapy." Journal of Experimental Medicine 216.1 (2019): 20-40.
[2] Olesen, Scott W., and Eric J. Alm. "Dysbiosis is not an answer." Nature microbiology 1.12 (2016): 16228.
[3] McDonald, Daniel, et al. "American gut: an open platform for citizen science microbiome research."
mSystems 3.3 (2018): e00031-18.

Corresponding Author's Address: Merciyes University, Genome and Stem Cell Center (GenKök), http://genkok.erciyes.edu.tr/en/bioinformatics/, nalbantoglu@erciyes.edu.tr

RECEPTOR-LIGAND BINDING AFFINITY PREDICTION VIA MULTI-CHANNEL DEEP CHEMOGENOMIC MODELING

Ahmet Sureyya Rifaioglu[1], *Tunca Doğan*[2,3], *Maria Martin*[4], *Rengül Çetin-Atalay*[2], *Volkan Atalay*[1]

1.Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey
2. Cancer Systems Biology Laboratory (Kansil), Graduate School of Informatics, Middle East Technical University, 06800 Ankara, Turkey
3. Institute of Informatics, Hacettepe University, 06800 Ankara, Turkey
4. European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), CB10 1SD Hinxton, Cambridge, UK

Here, we propose a multi-channel receptor-ligand binding affinity prediction method. Our system employs a chemogenomic modeling approach where the aim is to use both receptor (target) and ligand features as inputs. One advantage of incorporating both ligand and receptor features is that the system can predict binding affinities of any given receptor-ligand pair, even if the corresponding ligand and/or receptor does not have any data point in the training set. We used three datasets to train our system and to compare our results with the state-of-the-art methods: (1) Davis (2) Filtered Davis (3) PDBBind We describe a new protein encoding method where each protein is represented as a matrix which constitute the base channel of the convolutional part of the system. We also created additional input channels based on pre-defined amino acid matrices which represent various properties of amino acids and proteins. In the ligand side, we generated ECFP4 fingerprints using the SMILES strings of ligands, which are fed to a feed-forward neural network. Overall, we created a hybrid pairwise input neural network architecture which starts with separate ligand and protein branches, fully connected layers which processes the concatenated protein+ligand vector, and a regressor to predict the actual binding affinity value at the output layer. We compared our system with three different methods: (1) DeepDTA: a binding affinity prediction method based on convolutional neural networks and 1-D protein and compound encoding [1]. (2) SimBoost: another binding affinity prediction method based on gradient boosting machines and similarity networks[2]. (3) MoleculeNet: a benchmarking platform designed for evaluating and testing computational methods for molecular property predictions, which include prediction models that employ popular graph convolutional networks [3]. The performance results are given in Table 1. Results show that our method performs significantly better than the other methods in majority of the cases. Pursuing this approach, new models can be constructed by incorporating additional types of input protein channels.

Keywords: Binding Affinity Prediction, Chemogenomics, Receptor, Ligand, Deep Learning, Convolutional Neural Networks, Pairwise Input Neural Network, Protein Encoding.

Method	CI	MSE	Pear-son	Spear-man	AUC	Prec-ision	F1-Score	MCC
PINN (Davis)	0.875	**0.28**	**0.813**	0.674	**0.945**	**0.838**	**0.732**	**0.674**
DeepDTA (Davis)	0.863	0.315	0.795	0.661	0.937	0.749	0.703	0.645
SimBoost (Davis)	**0.876**	0.284	0.804	**0.677**	0.939	0.781	0.699	0.645
PINN (filt.Davis)	**0.722**	**0.597**	**0.964**	**0.681**	**0.712**	**0.842**	**0.98**	**0.709**
DeepDTA (filt.Davis)	0.655	0.873	0.934	0.463	0.649	0.754	0.862	0.64
PINN (PDBBind)	**0.74**	**2.65**	**0.668**	**0.661**	**0.83**	**0.757**	0.781	0.446
Grid Featurizer -RF (PDBBind)	0.729	3.4	0.632	0.634	0.807	0.762	**0.822**	**0.529**
Grid Featurizer - DNN (PDBBind)	0.67	3.616	0.532	0.505	0.735	0.692	0.79	0.406
ECFP4 - RF (PDBBind)	0.657	3.207	0.478	0.483	0.736	0.675	0.76	0.334
ECFP - RF (PDBBind)	0.608	5.255	0.329	0.344	0.664	0.648	0.736	0.25

Table 1: Performance comparison with state-of-the-art on 3 Datasets

References: [1] H. Öztürk, A. Özgür, and E. Ozkirimli, "DeepDTA: Deep drug-target binding affinity prediction," Bioinformatics, vol. 34, no. 17, pp. i821–i829, 2018.
[2] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester, "SimBoost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines," J. Cheminform., vol. 9, no. 1, pp. 1–14, 2017.
[3] Z. Wu et al., "MoleculeNet: A benchmark for molecular machine learning," Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018.

Corresponding Author's Address: vatalay@metu.edu.tr & arifaioglu@ceng.metu.edu.tr

INVESTIGATING GENETIC CONTINUITY OF HUMAN POPULATIONS IN ANATOLIA OVER THE PAST 15,000 YEARS

Dilek Koptekin[1,2], *Füsun Özer*[2,3], *Mehmet Somel*[2]

1. Department of Health Informatics, Middle East Technical University, 06800, Ankara, Turkey
2. Department of Biological Sciencess, Middle East Technical University, 06800, Ankara, Turkey
3. Department of Anthropology, Hacettepe University, Beytepe, 06800 Ankara, Turkey

Past demographic changes and migrations can be traced using modern genomes, as well as ancient genome data. In recent years, ancient genome analyses of human populations from west and central Anatolia have revealed a series of notable results, which suggested continuous gene flow into Anatolia from different geographic sources over time. This observation has led us to hypothesize that the gene pool of Anatolia may be more dynamic compared to those of neighbouring regions, possibly because of Anatolia's intermediate geographical location. From this perspective we investigated temporal changes in a region's gene pool by calculating the genetic dissimilarity among individuals in time using 164 published ancient genomes from Anatolia and also from four neighboring regions: the Levant, Iran, Caucasus, and the Aegean. The individuals studied dated back to the Mesolithic/Epipaleolithic Period and proceeded into the Medieval Period. We find that, as a general trend, gene pools have diverged over time. Our preliminary results further suggest that the populations who lived in Anatolia and Iran during the past 15,000 years may have changed more dramatically than the gene pools of neighbouring regions, most likely as a result of gene flow from external sources. We conclude by describing a timeline of major admixture in Anatolia of the last 15,000 years.

Corresponding Author's Address: dilek.koptekin@metu.edu.tr

EVALUATION OF CANCER SEQUENCING PIPELINES ON PATIENT SAMPLES WITH DIFFERENT HETEROGENEITY LEVELS

Sahin Sarıhan[1], Batuhan Kısakol[1], Mehmet Baysan[1]

1. İstanbul Sehir University, Computer Science and Engineering Department

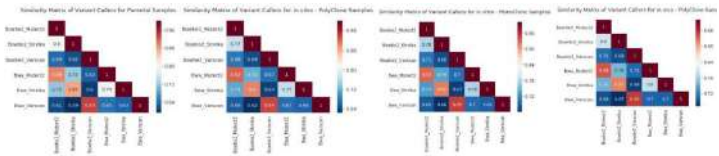
The importance of Next Generation Sequencing (NGS) rises in cancer research as accessing this central technology becomes easier for researchers. Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) are the most common methods for comprehensive variant detection in cancer research. The sequenced raw data by WGS and WES must be processed by various bioinformatics algorithms within a pipeline, in order to convert raw data to meaningful information. Mapping and variant calling are the two main steps of these analyses pipelines and there are many algorithms available for these steps[1]. These algorithms have different approaches and filters, therefore obtained results might vary significantly. Therefore, choosing the appropriate algorithms with optimal filters in a sequencing project is crucial for efficient utilization of sequencing technologies.

Benchmarking these sequencing pipelines is an active field of research and most of these studies rely on simulated data or data from primitive organisms such as bacteria or yeast. Unfortunately constructing realistic human sequencing data is extremely difficult, especially in cancer samples with complex selection dynamics and high level of genetic instability and heterogeneity[2]. Recently, we have published a dataset which has 55 high-resolution homogeneous and heterogeneous samples that belongs to a glioblastoma patient [3]. These samples share a substantial portion of mutations which allows us to declare these mutations as validated mutations; since for an algorithm identifying a non-existing mutation twice in two independent samples is almost impossible. Availability of these samples presented us a unique opportunity to test sequencing pipelines at different heterogeneity levels.

The dataset we used in this study had four different sample types: (i) parental tumor samples were obtained as tissues from different parts of patients tumor, (ii) in vitro polyclone samples are cultured tumor stem cell lines from parental-tumors, (iii) in vivo polyclone samples were obtained from mouse xenografts after in vitro polyclones are injected to mouse brain and formed a tumor, (iv) in vitro monoclone samples were obtaine from in vitro polyclone samples through isolation of single cells and subsequent culturing until there is enough cells for exome sequencing. For 55 samples from these four sample types; we applied two mapping algorithms: Bwa, Bowtie2 and three variant calling algortihms: Mutect2, Varscan, and Strelka. This resulted in six mapping-variant calling combinations labeled as "Bwa_Mutect2, Bwa_Varscan, Bwa_Strelka2, Bowtie2_Mutect2, Bowtie2_Varscan, Bowtie2_Strelka2". For these pipelines, first we compared the mutation lists by pairwise comparisons (Figure 1). Then,

we declared the mutations which are detected in two independent samples as "validated" mutations and evaluated the performance of each pipeline on detecting validated mutations.

In our analyses, we observed that different mapping and variant calling algorithms perform differently for different heterogeneity levels. This suggests attaching to a single pipeline is not optimal for cancer sequencing analyses and sample heterogeneity should be considered in algorithm optimization. Hopefully this will lead to more accurate variant detection and better results in clinical studies.



Keywords: Clinical Bioinformatics, Next Generation Sequencing, Cancer, Mapping Algorithms, Variant Discovery Algorithms

References: [1] Roberts ND, Kortschak RD, Parker WT, Schreiber AW, Branford S, Scott HS, Glonek G, Adelson DL. A comparative analysis of algorithms for somatic SNV detection in cancer. Bioinformatics. 2013 Sep 15;29(18):2223-30. PubMed PMID: 23842810.
[2] Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinformatics. 2019 Jun 17;20(1):342. PubMed PMID: 31208315.
[3] Baysan M. , Woolard K. , Cam M. C., Zhang W. , Song H. , Kotliarova S. , Balamatsias D. , Linkous A. , Ahn S. , Walling J. , Belova G. I. and Fine H. A. Detailed longitudinal sampling of glioma stem cells in situ reveals Chr7 gain and Chr10 loss as repeated events in primary tumor formation and recurrence. Int. J. Cancer 2017 Jul 14;141: 2002-2013. PubMed PMID: 28710771.

Corresponding Author's Address: Asst. Prof. Mehmet Baysan Orhantepe Mahallesi, Turgut Özal Bulvarı, No: 21, Dragos, Kartal – İstanbul, 34865 Office: AB4-4017 Tel: +90 216 559 9000 (ext. 9736) Fax: +90 216 474 53 53 Email: mehmetbaysan@sehir.edu.tr

REDUCTION OF DEPRESSION SYMPTOMS VIA THE METABOLITES OF GUT MICROBIOTA

İsra Mavaldi[1], Alper Yılmaz[2]

1,2. Yildiz Technical University, Department of Bioengineering, Istanbul, Turkey

Depression is a common mental disorder and according to World Health Organization, globally more than 300 million people of all ages suffer from depression[1]. Even though, until now the real reason behind depression is unknown, but in most depressive patients low levels of serotonin are detected and because serotonin is a mood controlling neurotransmitter, depression also called mood disorder. Most of the treatments aimed to elevate the serotonin levels but at the same time, these treatments have many side effects[2].

In this study, we aimed to reveal possible metabolites produced by gut microbiota that can elevate serotonin levels by inhibiting tryptophan degradation which is called kynurenine pathway that is a competitive pathway with serotonin synthesis from tryptophan.

So, we combine many datasets that contain information about the reactions including their products that gut bacteria can catalyze them by expressing the catalyzing enzymes. These datasets are taken from many databases including Brenda, ChEBI, MetanetX, Uniprot, and Microbiome databases. We used R for data analysis steps.

We ended with three distinct reactions catalyzed by two distinct enzymes expressed by eight distinct species of gut bacteria.

Keywords: Depression, Serotonin, Tryptophan, Gut Microbiota, Kynurenine Pathway.

References: [1] 'WHO | Depression'. [Online]. Available: https://www.who.int/mental_health/management/depression/en/.
[2] M. Vaváková, Z. Ďuračková, and J. Trebatická, 'Markers of Oxidative Stress and
Neuroprogression in Depression Disorder', Oxid. Med. Cell. Longev., vol. 2015, pp. 1–12, 2015.

Corresponding Author's Address: Yildiz Technical University, Department of Bioengineering, Davutpasa, Istanbul, Turkey israamaw96@gmail.com

AUTOMATION OF GENETIC TESTING AND REPORTING WITH SNAKEMAKE

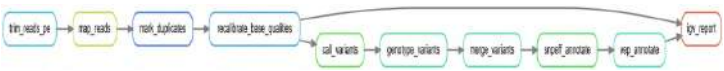
Oguzhan KALYON[1], *M. Hamza MUSLUMANOGLU*[1], *Alper YILMAZ*[2]

1. Yildiz Technical University, Faculty of Arts and Sciences, Department of Molecular Biology and Genetics, Istanbul, Turkey
2.Yildiz Technical University, Department of Bioengineering, Istanbul, Turkey

Next Generation Sequencing (NGS) technology has boosted genetic research. especially allowing fast and accurate diagnosis. However, analysis of multiple samples by passing them through multiple steps in commandline is tedious and error-prone. In this study we aim to facilitate analysis of NGS data with Snakemake[1] which provides management of Python-based workflows. Additionally, Snakemake benefits from conda environments thus installation or configuration of numerous softwares becomes effortless..

We modified an existing Snakemake workflow inspired from GATK best practices[2] and we integrated ENSEMBL Variant Effect Predictor (VEP)[3] into annotation step of the workflow. Additionally, we integrated Integrative Genome Viewer (IGV)[4] in final report via javascipt library[5]. Final report also includes interactive HTML tables generated by R script so that end user dynamically analyze the results.

As a result, any user can clone our code and then with any raw fastq file, initiate the mapping, annotation and report generation steps easily. This approach is reproducible and portable, it can be implemented in a personal laptop, server or even a cluster. Since Snakemake can run parallel jobs, the SNP analysis can be done in parallel fashion if multiple CPUs are available. Such an approach will allow a user or genetic analysis center to save time by running analysis and generating reports in automated way. More importantly, using a workflow approach will prevent errors even though large number of samples are processed.



Keywords: NGS, Snakemake, Genetic Test

Figure 1. Workflow chart

References: [1] Köster J, Rahmann S. "Snakemake - a scalable bioinformatics workflow engine", Bioinformatics, 2012. 28(19):2520-2522, https://doi.org/10.1093/bioinformatics/bts480
[2] https://github.com/snakemake-workflows/dna-seq-gatk-variant-calling [3] McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GR, Thormann A, Flicek P, Cunningham F. "The Ensembl Variant Effect Predictor". Genome Biology, 2016. 17(1):122 doi:10.1186/s13059-016-0974-4 [4] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. "Integrative Genomics Viewer". Nature Biotechnology 29, 24–26 (2011) [5] https://github.com/igvteam/igv-reports

Corresponding Author's Address: ozyyk61@gmail.com

SYMPTOM RANKING ALGORITHM FOR RARE DISEASES USING TF-IDF SCORING

Yasemin Utkueri[1], Hüseyin Okan Soykam[2], Uğur Sezerman[3]

1. Molecular Biology, Genetics and Bioengineering / Computer Science & Engineering , Faculty of Engineering and Natural Sciences, Sabanci University
2. Department of Biostatistics and Bioinformatics, Health Sciences Institute, Acibadem Mehmet Ali Aydinlar University
3. Department of Biostatistics and Bioinformatics, Health Sciences Institute, Acibadem Mehmet Ali Aydinlar University

Rare diseases, which European Union defines as the diseases that affect less than 1 in 2000 people, are mostly life threatening [1]. In countries like Turkey where consanguineous marriages are more common, rare diseases has relatively high prevalence in those populations compared to European countries. In spite of relatively high prevalence, diagnosing rare diseases is hard to accomplish. Identifying casual symptoms particular to a disease and differentiating them from other diseases remains challenging. And, this makes finding cures and offering treatments especially difficult. Because of this, patients can not be diagnosed accurately and had to go through incorrect treatments [2]. This preliminary study aims to help eliminating inaccurate diagnosis by creating a list of symptoms for diseases that is scored and sorted by their importance via information retrieval algorithms. We regard importance as relevance to the disease and expect symptoms particular to a rare disease are to be ranked as most important.

With the TF-IDF (Term Frequency – Inverse Document Frequency) scoring, it is possible to rank the symptoms of a rare disease based on their frequency in literature. The scoring process is made up of two parts: term frequency and inverse document frequency. Term frequency is the scoring of how many times each term, in this case the symptom, appears in a document. To normalize the score, the count is divided by the number of documents in the corpus. Inverse document frequency is the score of the importance of a term. This is calculated by taking the logarithm of the number of documents with the term in it divided by the total number of documents. [3,4] The more frequent the term is, the smaller the IDF score will be. With the two scores, we can calculate the total TF-IDF score of a symptom for the given disease by multiplying the TF and IDF scores. In our study, the corpus is made up of abstracts extracted from pub-med. The TF-IDF scores of the symptoms that are most commonly seen in the corpus are the lowest. These symtpom are ranked the highest because they are more closely associated with the disease.

Table 1 shows the top 10 symptoms with the lowest scores calculated by out algorithm for three diseases that are very similar, parkinson's disease, atypical juvenile parkinsonism and hereditary late-onset parkinson's. Although atypical juvenile parkinsonism and postencephalitic parkinsonism are variations of parkinson's disease, the top symptoms of each disease are different. The TF-IDF scoring

correctly distinguishes diseases from each other by highlighting the symptoms that are specific to that exact disease.

Keywords: Rare Disease; Information Retrieval; Text Mining; Symptom; Disease; Diagnosis

Disease	parkinson's	atypical juvenile parkinsonism	postencephalitic parkinsonism
Rank of symptoms	1. melanoma 4.053956501316812 2. dyskinesia 4.053956501316812 3. dystonia 5.034785754328539 4. carcinoma 5.034785754328539 5. fasciitis 5.034785754328539 6. psychosis 5.034785754328539 7. progressive 5.034785754328539 8. gastroparesis 5.034785754328539 9. periodontitis 5.034785754328539 10. memory impairment 5.034785754328539	1. delirium 3.4760986898352733 2. strabismus 4.23823874188217 3. pneumonia 4.392389421709428 4. dysphonia 4.574710978503383 5. vestibular dysfunction 4.5747109785O3383 6. dyskinesia 4.7978545298175925 7. shock 5.085536602269373 8. dementia 5.085536602269373 9. diplopia 5.085536602269373 10. hypotension 5.085536602269373	1. narcolepsy 2.7642204725692645 2. cataplexy 3.392829131991639 3. ptosis 4.085976312551584 4. camptocormia 4.373658385003365 5. seborrheic dermatitis 4.373658385003365 6. ileus 4.77912349311153 7. dementia 4.77912349311153 8. hypersomnia 4.77912349311153 9. lewy bodies 4.77912349311153 **10.** pharyngitis 4.77912349311153

Table 1: Table showing the top 10 symptoms for parkinson's disease, atypical juvenile parkinsonism and postencephalitic parkinsonism.

References: [1] Rare Disease UK [Internet]. London: Rare Diseases UK. Available from: https://www.rarediseases.org.uk.
[2] Fernandez-Marmiesse A, Gouveia S, Couce ML. NGS Technologies as a Turning Point in Rare Disease Research , Diagnosis and Treatment. Current medicinal chemistry 2018;25(3):404-32.
[3] Ramos,J. Using [3] Ramos, J. (n.d.). Using TF-IDF to Determine Word Relevance in Document Queries. [4] Salton, G. & Buckley, C. (1988). Term-weighing approaches in automatic text retrieval. In Information Processing & Management, 24(5): 513-523.

Corresponding Author's Address: Orta Mahallesi, Sabancı Ünv. No:27, 34956 Tuzla/İstanbul yaseminutkueri@sabanciuniv.edu

PREDICTION OF MUTATION SUSCEPTIBILITY BASED SOLELY ON DNA SEQUENCES WITH WORD2VEC

Busra Nur Darendeli[1], *Alper YILMAZ*[1], *Oguzhan KALYON*[2]

1. Yildiz Technical University, Department of Bioengineering, Istanbul, Turkey
2.Yildiz Technical University, Faculty of Arts and Sciences, Department of Molecular Biology and Genetics, Istanbul, Turkey

With the advent of natural language processing (NLP) techniques empowered with deep learning approaches, more detailed relationships between words have been unraveled. Word2vec[1] is a shallow neural network that generates word embeddings. Word2vec is quite robust in discovering contextual and semantic relationships. Genome being a long text, is subject to similar studies to unravel yet to be discovered relationships between DNA k-mers. Dna2vec[2] applies Word2vec approach to whole genome so that DNA k-mers are represented as vectors. The cosine similarity queries on DNA vectors reveal unusual relationships between DNA k-mers.

In this study, we examined DNA sequence based prediction of mutation susceptibility. Initially,we generated word vectors for human and mouse genome via dna2vec.. On the other hand, we retrieved coordinates of common and all SNPs from dbSNP[3]. For each coordinate, we extracted 8 nucleotide k-mers intersecting SNPs and results are aggregated. such a way that number of SNPs for each 8-mer has been tabulated. These results are incorporated with dna2vec cosine similarity data. Our results showed that for a given k-mer, k-mers with highest cosine similarity coincide with highest SNP count k-mer. In other words, the neighbor with the highest cosine similarity for a k-mer was also seen to be the neighbor overlapping the SNP count. As a result of our studies, human and mouse, dna2vec vs. SNP overlap is 80% and 70%, respectively. In conclusion, dna2vec and other word embedding approaches can be used to reveal mutation or variation characteristics of genomes without sequencing or experimental data, solely using the genome sequence itself. This might pave the way for understanding the underlying mechanism or dynamics of mutations in genomes.

Keywords: SNP, Word2vec, Cosine Similarity, K-Mer

Figure 1.
Workflow of study

References: [1] Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint. 2013. arXiv:1301.3781.
[2] Ng P. (2017). dna2vec: Consistent vector representations of variable-length k-mers. arXiv preprint. 2017. arXiv:1701.06279.
[3] https://www.ncbi.nlm.nih.gov/snp

Corresponding Author's Address: bndarendeli@gmail.com

IDENTIFICATION OF LATENT DRIVER MUTATIONS IN PAN-CANCER DATA

Bengi Ruken Yavuz and Nurcan Tunçbağ

Graduate School of Informatics, Department of Health Informatics, Middle East Technical University, 06800, Ankara, Turkey.

A major challenge in cancer genomics is identifying driver mutations from the background passenger mutations within a given tumor. A central assumption used in discovering driver mutations is that exact positional recurrence is unlikely to occur by chance. There are various databases that already labelled many mutations as drivers or passengers. Recently a new concept, "latent driver mutations" was introduced which is defined as mutations that does not provide a growth advantage to cancer cell but together with a newly evolved mutation, they can express a cancer cell phenotype. These cooperatively acting mutations can change the conformation of the proteins and hence their interactions. Moreover, latent driver mutations can help to explain the rewiring mechanisms of oncogenic networks in developing drug resistance. Differentiating these latent mutations among the passengers is a challenging task.

In this study, we use a computational approach to differentiate the latent driver mutations from the passenger ones using protein structures. We perform a Pan-Cancer Analysis on MC3 Call Set that contains of more than 10,000 tumor-normal exome pairs across 33 different cancer types. First, we identified the number of recurrent mutation pairs on within the same gene across all patients. Afterwards, we obtained the genes on which mutation pairs observed in more than 3 patients and involved in kinase activity or function as a transcription factor. Additionally, we identify significantly mutually exclusive mutations in these Pan-Cancer dataset to find out epistatic relationships. As a result we obtained that mutations on residue pairs in the same protein can rarely co-occur in different patients. Next, we check mutations on different proteins whether any mutation pairs co-occur across different tumors. We mapped these pairwise recurrent mutations to the 3D structure of the protein and determine the effects of these mutations on kinase activity or DNA binding and eventually we label these mutations as driver, latent driver or passenger.

Keywords: Latent Driver Mutations; Pan-Cancer Analysis; Structural Bioinformatics

Corresponding Author's Address: Nurcan Tunçbağ Middle East Technical University Informatics Institute B-204 Cankaya/Ankara ntuncbag@metu.edu.tr

CONNECTIVITY ANALYSIS OF CELL TYPE SPECIFIC PROTEIN INTERACTION NETWORKS

Ertuğrul Dalgıç [1]

1. Department of Medical Biology, Zonguldak Bülent Ecevit University School of Medicine, Zonguldak, Turkey

Systems biology view of different cell types requires analysis of specific molecular interaction networks. Protein-protein interaction (PPI) datasets do not contain cell type specific interactions. Therefore, only a fraction of PPI data could be present in a cell type of interest. Unlike PPI datasets, protein expression datasets for different cell types are available [1]. Here, human PPI data was integrated with human protein expression levels in order to generate potential specific protein interaction networks for various cell types. Specific networks are composed of proteins with low, medium and high expression levels. Compared to randomly selected networks of the same size, specific networks have higher connectivity. For the majority of cell type specific networks, proteins with medium and high expression have significantly higher degree values, based on randomization tests. For the proteins with medium or high levels of expression, a general trend of significantly high number of connections between themselves, and low number of connections with the remaining members of the network (proteins with low expression) was observed; suggesting a modularity effect. Connectivity analyses of cell type specific protein interaction networks suggest a dominant role for proteins with medium and high expression levels.

Keywords: Systems Biology; Protein-Protein Interactions

References: [1] Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson Å, Kampf C, Sjöstedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Pontén F. Proteomics. Tissue-based map of the human proteome. Science. 2015 Jan 23;347(6220):1260419. PMID: 25613900.

Corresponding Author's Address:Tıbbi Biyoloji Anabilim Dalı, Zonguldak Bülent Ecevit Universitesi Tıp Fakültesi Dekanlığı, Kozlu, Zonguldak, 67600, ertugruld@beun.edu.tr

A MACHINE LEARNING APPROACH TO CLASSIFY INTERGENIC SPACER (IGS) REGIONS PREDICTING THE LINEAGE INFERENCES.

Yasin Kaya[1]

1. Hacettepe University, Faculty of Science, Department of Biology, 06800, Ankara - Turkey

Intergenic spacers (IGS) are short, tandemly repeated and rapidly evolving motifs which can be related to length polymorphisms observed between non-coding DNA sequences. They contain sequences that are essential for the initiation of transcription, RNA processing, transcription termination and replication processes of ribosomal DNA as well [1,4]. Many of the repeat motifs in the non-coding DNA have been described as a conservative motif in the promoter region of some crucifers such as Brassica spp.[7], Raphanus spp.[8] and Arabidopsis[9].

Amount of the IGS evolutionary distance rates is used to train machine learning algorithms which gives the final classification model and a list of top decisive features. Frequency of the different IGS distance rates are Poisson-Dirichlet distribution suggested that they could be explained by Random Forest (RF) and Naïve Bayes (NB). The principle behind the working of the RF algorithm is to construct many decision trees during the training and classifies the test case based on the mode of the classes given by individual trees. Naïve Bayes (NB) algorithm is based on the Bayes theorem with the assumption that the features are independent of each other. NB provided a method of calculating the posterior probability of class given the prior probability and learning the likelihood of features given the class.

As a result of the statistical model of intergenic spacer regions demonstrated that could be considered of both hypothesis testing and phylogenetic inferences.

References: [1] Fernandez, M., Polanco, C., Ruiz, M.L. and Perez de la Vega, M. (2000) A comparative study of the structure of the rDNA intergenic spacer of Lens culinaris Medik. and other legume species. Genome 43, 597–602.
[2] Lachance, M.A., Starmer, W.T., Bowles, J.M., Phaff, H.J. and Rosa, C.A. (2000) Ribosomal DNA, species structure, and biogeography of the cactophilic yeast Clavispora opuntiae. Can. J. Microbiol. 46, 195–210.
[3] Irac , abal, B. and Labaré re, J. (1994) Restriction site and length polymorphism of the rDNA unit in the cultivated basidiomycete 189. Pleurotus cornucopiae. Theor. Appl. Genet. 88, 824–830.
[4] VanÕt Hof, J. and Lamm, S.S. (1991) Single stranded replication intermediates of ribosomal DNA replicons of pea. EMBO J. 10, 1949–1953.
[5] Grellet, F., Delcasso-Tremousaygue, D. and Delseney, M. (1989) Isolation and characterization of an unusual repeated sequence from the ribosomal intergenic spacer of the crucifer Sisymbrium irio. Plant Mol. Biol. 12, 695–706.

[6] Barker, R.F., Harberd, N.P., Harvis, M.G. and Flavell, R.B. (1988) Structure and evolution of the intergenic region of a ribosomal DNA repeat unit of wheat. J. Mol. Biol. 201, 1–17.
[7] Bhatia, S., Negi, M.S. and Lakshmikumaran, M. (1996) Structural analysis of the rDNA intergenic spacer of Brassica nigra: Evolutionary divergence of the spacers of the three diploid Brassica species. J. Mol. Evol. 43, 460–468.
[8] Delcasso-Tremousaygue, D., Grellet, F., Panabieres, F., Ananiev, E.D. and Delseney, M. (1988) Structural and transcriptional characterization of the external spacer of a ribosomal RNA nuclear gene from a higher plant. Eur. J. Biochem. 172, 767–776.
[9] Gruendler, P., Unfried, I., Pointner, R. and Schweizer, C. (1991) Nucleotide sequence of the 25S–18S ribosomal gene spacer from Arabidopsis thaliana. Nucleic Acids Res. 17, 6395–6396.

Corresponding Author's Address: yyasinkkaya@gmail.com

MODELING THE TUMOR SPECIFIC NETWORK REWIRING BY INTEGRATING ALTERNATIVE SPLICING EVENTS WITH STRUCTURAL INTERACTOME

Habibe Cansu Demirel, Nurcan Tunçbağ

1. Graduate School of Informatics, Department of Health Informatics, Middle East Technical University,06800, Ankara, Turkey.

Analysis of genetic variations, increasing or decreasing gene expression levels, changes in protein expression those trigger tumor formation and elucidating signaling pathways by using these data are crucial for developing personalized therapeutic strategies. One of the most important mechanisms occurs at transcriptional level which increases the diversity of the proteome is alternative splicing. It has been estimated that in 90% of all genes in human undergo alternative splicing This process is regulated during gene expression and different final mRNA sequences are created via the exclusion of exons, exon parts or even with the inclusion of introns. In this way, one gene can code multiple proteins with potentially different structures. While some protein isoforms may lose their interactions because of a change in their binding region, some others may gain new interactions. Hence, the diversity in isoform proteins has a direct effect on protein interaction networks and signaling pathways. This finding changes our viewpoint on classical protein interaction networks. Because alternative splicing events regulate the activity of almost all genes, abnormalities in this process may be effective in progression of many diseases and cancer.

In this study, we reconstructed patient specific networks with tumor specific protein isoforms by integrating the protein structures and the interaction losses they bring with. For this purpose, we collected 400 breast cancer tumors and 112 normal RNA-seq data from the Cancer Genome Atlas (TCGA). After the detection of expressed transcripts using a transcriptome assembler, we calculated the log fold changes between tumor and an averaged pool of normal samples for each patient to find the transcripts that show increased expression in tumor samples. At the same time, we mapped the transcripts to protein isoforms to detect the lost regions differing from the canonical protein. Additionally, we compiled a structural human interactome from multiple sources and aligned the missing residues on isoforms with the known/predicted protein interfaces to find potential interaction losses. Then, we constructed two interactomes for each sample; one filtered based on the lost interfaces as a result of predominant isoforms called "terminal set" and one filtered based on the expression. Terminal sets included the proteins that lost at least one interaction and show an increased expression favoring the non-canonical isoforms. We used the same terminal set with Omics Integrator to model two sets of networks based on the two patient-specific interactomes. Finally, we compared the resulting network sets to find the proteins that are only found in tumor-specific networks. Their enrichment analysis showed that patient subgroups can be obtained through clustering

based on enriched gene sets and telomer related gene sets were the most distinctive among such clusters. In addition, as a result of the examination of interactions, we found 12125 lost protein-protein interactions for all samples. There were 53 unique lost protein – drug interactions and 43 protein that lost DNA interactions. At the same time, the analysis of the proteins losing interactions resulted in interesting examples including interaction losses between cancer drugs and their targets.

In our ongoing analysis, we focus on breast cancer pathway with pathways in cancer and how these pathways are altered given the spliced isoforms and their lost interactions. We observed that for all samples, the terminal sets include up to 16 pathways in cancer genes. Moreover, we also compare the resulting networks to reveal pathway, protein-protein interaction and protein patterns that can cluster the tumors according to their similarities. The results of our analysis will contribute to the elucidation of tumor mechanisms and will help for target selection and developing therapeutic strategies.

Keywords: Alternative Splicing; Protein Interaction Network; Network Rewiring

Corresponding Author's Address: Nurcan Tunçbağ Middle East Technical University Informatics Institute B-204 Cankaya/Ankara ntuncbag@metu.edu.tr

CONSTRUCTING ACCURATE COMPOUND NETWORKS FOR RELIABLE NETWORK ANALYSIS

Derya Alpaydın1, Ravza Öztürk1, Alper Yılmaz1

1. Yildiz Technical University, Department of Bioengineering, Istanbul, Turkey

Cellular metabolism includes highly complex processes which are essential for the survival of any organism. To enlighten mechanism of many diseases and treatments, it is very important to understand human metabolic system. Analysis of metabolic system as a network is crucial and applied in recent years with the help of holistic approaches over human disease studies. Various methods have been proposed to identify interactions between nodes in the reactions by the help of standard graph theory measurements. Available metabolic networks connect all possible reactant pairs of a reaction to construct metabolite network. This approach introduces inaccuracy due to connections between currency metabolites and actual reactants.

In order to increase accuracy we removed the connections between unrelated molecules. In our project, we applied tanimoto similarity and flexible most common substructure functions available in ChemmineR package[1] to construct accurate compound networks. After constructing more accurate metabolic network, we applied L-value [2], a recently developed node importance measure, to identify critical compounds in human metabolic network.

Keywords: Metabolic Network; Network Analysis; Tanimoto Similarity

References: [1] Cao Y, Charisi A, Cheng L, Jiang T, Girke T. ChemmineR: a compound mining framework for R. Bioinformatics, 2008; 24(15), 1733-1734. https://doi.org/10.1093/bioinformatics/btn307

[2] Liu J, Xiong Q, Shi W, Shi X, Wang K. Evaluating the importance of nodes in complex networks. Physica A: Statistical Mechanics and its Applications. 2016;452:209-219. https://doi.org/10.1016/j.physa.2016.02.049.

Corresponding Author's Address: alyilmaz@yildiz.edu.tr

AN INTEGRATIVE NETWORK-BASED APPROACH FOR PRIORITIZING DRIVER GENES FOR BREAST CANCER

Cesim Erten[2], Aissa Houdjedj[1], Hilal Kazan[2]

1. Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya, Turkey
2. Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey

Recent cancer genomic studies have generated detailed molecular data on a large number of cancer patients. A key remaining problem in cancer genomics is the identification of driver genes. We propose a computational approach that integrates genomic data (gene expression, somatic mutation data) with protein-protein interaction network data. As a first step, a bipartite graph is constructed where one side contains the mutated genes in the cohort and the other side contains patient-specific "outlier" genes that are defined as the set of genes with altered functions. Similar to the existing approach DriverNet [1], the impact of a potential driver gene is determined by its effect on the "outlier " genes that it regulates. Our approach differs from DriverNet and other related methods [e.g., 2] in two ways. First, our definition of outlier gene sets is not simply based on differential gene expression but on betweenness centrality measurements calculated from paired patient-specific normal-tumor networks that are perturbed based on gene expression and somatic mutation data. Second, we apply a network diffusion step on the bipartite graph that results in an edge weighted graph. We applied our approach to TCGA breast cancer data as it contains the largest number of normal / tumor samples. XXX recovers a larger number of known drivers compared to DriverNet and Subdyquency when known drivers are defined using NCG [3] and CancerMine [4] databases. Additionally, we show that the overlap between the GO terms enriched in the set of XXX predicted driver genes and the GO terms enriched in reference gene sets defined using NCG or CancerMine is higher compared to the other methods. We confirm the same observation when we use KEGG pathway enrichment. In conclusion, we Show that utilizing graph theoretical measures together with genomic data and applying a network diffusion step improve the identification of driver genes on breast cancer data.

Keywords: Driver Gene Prioritization, Bipartite Graph, Betweenness Centrality, Network Diffusion

References: [1] Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol, 2012,13(12):R124 1698 [2] Song J, Peng W, Wang F. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. BMC Bioinformatics, 2019, 20:238 [3] Repana D., Nulsen J., Dressler L., Bortolomeazzi M., Venkata S. K., Tourna A., Yakovleva A., Palmieri T., and Ciccarelli F.D.The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing

screens. Genome Biol, 2019, 20:1 https://doi.org/10.1186/s13059-018-1612-0 [4] Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods, 2019, 16:505–507

Corresponding Author's Address:cesim.erten@antalya.edu.tr
hilal.kazan@antalya.edu.tr

ASSESSMENT OF WEBSERVERS IN THE PREDICTION OF POINT MUTATIONS' IMPACT ON KINASE:LIGAND INTERACTIONS

Mehmet Ergüven[1,2], *Tülay Karakulak*[1,2], *Muhammed Kasım Diril*[1,2], *Ezgi Karaca*[1,2]

1. Izmir International Biomedicine and Genome Institute
2. Izmir Biomedicine and Genome Center

Protein kinases are integral players of cellular metabolism. The essentiality of proper kinase functioning is challenged by mutations occurring in kinases, which often result in severe diseases, like cancer. Point mutations occurring within or in the vicinity of catalytically important kinase motifs can switch the kinase conformation to constitutively active or drug resistant state. For many years, it has been of great interest to assess structural/kinetic impact of protein kinase mutations. Thus, it is a requisite to compile accurate binding affinity prediction workflows to estimate the functional impact of kinase mutations in silico. Expanding on this, we compiled the first high-resolution kinase:ligand benchmark to assess the field's capability in predicting the impact of kinase mutations on kinase:ligand binding affinities.

To that end, we collected a set of 12 wild type kinase structures and their 49 mutant states from Protein Data Bank. These numbers represent the cases in which both wild type and mutant states of the kinase of interest are bound to the same ligand. Our benchmark is made of cytosolic and membrane-associated protein kinases, mainly functioning in cell cycle, cell growth, DNA damage, metabolism, and transcriptional regulation. Ligands bound to these kinases are mostly nitrogen-rich poly-heterocyclic compounds (46% are halogenated). Together with the structures, we have also collected experimentally determined kinase:ligand binding data (either IC50, Kd, or Ki) from PDBbind.

As of today, there is no tool specialized to predict the functional impact of kinase mutations. Though, there are web-based approaches poised to estimate protein-ligand binding affinities. Within this context, we chose the most commonly used protein:ligand affinity predictors (HADDOCK2.2[1](refinement interface), PRODIGY-LIG[2], DSXonline[3], and KDEEP[4]) to estimate the binding affinities of our wild type and mutant complexes. When run only on the wild type complexes, PRODIGY-LIG correlated computed andexperimental binding affinities (log(IC50)) the best (Pearson's R2=0.76) (Figure 1). When the mutant forms were included in the data set, none of the webservers could produce a meaningful correlation. In this case, the best performer KDEEP could relate the calculated affinities to experimental Ki values with a Pearson's R2 of 0.32. We are currently analyzing the results to understand if any particular characteristics of the mutation, protein type, or ligand are responsible for this sharp

drop in the prediction accuracy. After this, we plan to probe the predictive capacity of these webservers on a derived benchmark set, made of predicted and experimentally determined binding affinities of the mutant cases normalized according to their wild type values. Our ultimate aim is to present an optimal affinity prediction workflow that can aid experimentalists in designing kinase mutations during their experimental setups.

Keywords: Binding Affinity Prediction; Protein Kinase; Point Mutation; Benchmarking

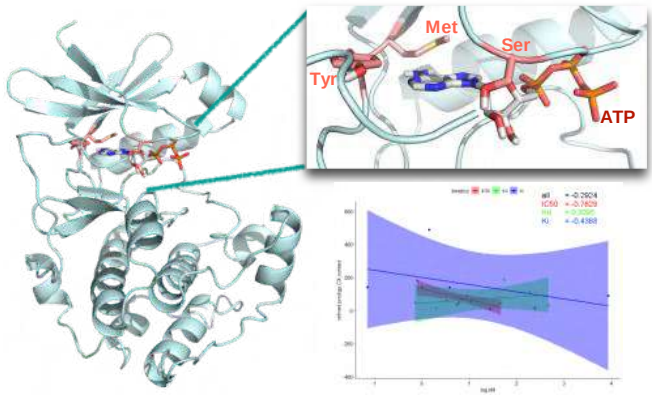


Figure 1. ATP binding pocket of MastI kinase (left). To manipulate substrate specificity, commonly mutated residues are indicated as sticks (top-right). PRODIGY-LIG yielded the highest correlation on wild type complexes (bottom-right).

References: [1] Van Zundert GCP, Rodrigues JPGLM, Trellet M, Schmitz C, Kastritis PL, Karaca E, Melquiond ASJ, Van Dijk M, De Vries SJ, Bonvin AMJJ. The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. Journal of Molecular Biology 2016 Feb 22;428(4):720-725. PubMed PMID: 26410586.

[2] Vangone A, Schraarschmidt J, Koukos P, Geng C, Trellet ME, Xue LC, Bonvin AMJJ. Large-scale prediction of binding affinity in protein-small ligand complexes: the PRODIGY-LIG web server. Bioinformatics 2019 May 1;35(9):1585-1587. PubMed PMID: 31051038

[3] Neudert G, Klebe G. DSX: A knowledge-based scoring function for the assessment of protein-ligand complexes. Journal of Chemical Information and Modeling 2011 Oct 24;51(10):2731-45. PubMed PMID: 21863864.

[4] Jiménez J, Škalič M, Martínez-Rosell G, De Fabritiis G. KDEEP: Protein-Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. Journal of Chemical Information and Modeling 2018 Feb 26;58(2):287-296. PubMed PMID: 29309725.

Corresponding Author's Address: Dr. Ezgi Karaca, ezgi.karaca@ibg.edu.tr, https://www.ibg.edu.tr/research-programs/groups/karaca-lab/

AN INTEGRATIVE NETWORK-BASED APPROACH FOR PRIORITIZING DRIVER GENES FOR BREAST CANCER

Ahmed Amine Taleb Bahmed[1], Cesim Erten[2], Aissa Houdjedj[1],

Hilal Kazan[2]

1. Electrical and Computer Engineering Graduate Program, Antalya Bilim University, Antalya, Turkey
2. Department of Computer Engineering, Antalya Bilim University, Antalya, Turkey

Recent cancer genomic studies have generated detailed molecular data on a large number of cancer patients. A key remaining problem in cancer genomics is the identification of driver genes. We propose BetweenNet, a computational approach that integrates genomic data (gene expression, somatic mutation data) with protein-protein interaction network data. As a first step, a bipartite graph is constructed where one side contains the mutated genes in the cohort and the other side contains patient-specific "outlier" genes that are defined as the set of genes with altered functions. Similar to the existing approach DriverNet [1], the impact of a potential driver gene is determined by its effect on the "outlier" genes that it regulates. Our approach differs from DriverNet and other related methods [e.g., 2] in two ways. First, our definition of outlier gene sets is not simply based on differential gene expression but on betweenness centrality measurements calculated from paired patient-specific normal-tumor networks that are perturbed based on gene expression and somatic mutation data. Second, we apply a network diffusion step on the bipartite graph that results in an edge weighted graph. We applied our approach to TCGA breast cancer data as it contains the largest number of normal / tumor samples. BetweenNet recovers a larger number of known drivers compared to DriverNet and Subdyquency when known drivers are defined using NCG[3] and CancerMine[4] databases. Additionally, we show that the overlap between the GO terms enriched in the set of predicted driver genes and the GO terms enriched in reference gene sets defined using NCG or CancerMine is higher for BetweenNet compared to the other methods. We confirm the same observation when we use KEGG pathway enrichment. In conclusion, we Show that utilizing graph theoretical measures together with genomic data and applying a network diffusion step improve the identification of driver genes on breast cancer data.

Keywords: Driver Gene Prioritization, Bipartite Graph, Betweenness Centrality, Network Diffusion

References: [1] Bashashati A, Haffari G, Ding J, Ha G, Lui K, Rosner J et al. DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome Biol, 2012,13(12):R124 1698
[2] Song J, Peng W, Wang F. A random walk-based method to identify driver genes by integrating the subcellular localization and variation frequency into bipartite graph. BMC Bioinformatics, 2019, 20:238
[3] Repana D., Nulsen J., Dressler L., Bortolomeazzi M., Venkata S. K.,

Tourna A., Yakovleva A., Palmieri T., and Ciccarelli F.D.The Network of Cancer Genes (NCG): a comprehensive catalogue of known and candidate cancer genes from cancer sequencing screens. Genome Biol, 2019, 20:1 https://doi.org/10.1186/s13059-018-1612-0
[4] Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nat Methods, 2019, 16:505–507

Corresponding Author's Address: cesim.erten@antalya.edu.tr
hilal.kazan@antalya.edu.tr

INTEGRATIVE ANALYSIS OF PATHOGEN HOST METABOLIC NETWORK OF SALMONELLA ENTERICA WITH DUAL RNA SEQ DATA

Kadir KOCABAŞ , Tunahan ÇAKIR

1. Gebze Technical University, Department of Bioengineering, Gebze, Kocaeli

Salmonella enterica is an important pathogen for both humans and animals. Being thoroughly studied as an organism, there is a lot more to discover about this pathogen. It has been linked to a variety of infectious diseases along with Typhoid fever and nontyphoidsalmonellosis, which cause public health issues [1]. In this study, it is aimed to determine metabolic characteristics of the pathogen by using constraint based computational methods. Genome scale metabolic network (GMN) models are widely used to understand metabolism of cells. GMN models have shown utility in integrating omic data for systemic analysis of metabolism. Pathogen-host metabolic modeling is a multi-cellular interaction modelling approach, where GMNs of both pathogen and host organisms are integrated in the simulations of metabolic phenotypes. Integration of dual RNA-seq data with a pathogen-host metabolic network model is a convenient way to analyze the effect of infection on the pathogen and host metabolisms together. Dual RNA sequencing (dual RNA-seq) is a rather recent method to study gene expression profiles. It enables simultaneous measurement of the global transcriptome of both host and pathogen, during infection [2]. A pathogen-host metabolic model can be turned into a condition-specific model by mapping dual RNA-seq data, leading to elimination of inactive reactions in both organisms and identification of interactions between a pathogen and its host. We used the genome-scale metabolic network of S. Enterica from literature [3], which included 2545 reactions controlled by 1271 genes. As host GMN, a recent reconstruction of human metabolism, called iHsa, was used, which covered 8336 reactions and 5627 genes [4]. Dual RNA-seq data of S. enterica infection of human cell lines [5] was obtained from the public transcriptome database, Gene Expression Omnibus. Integrative analysis by a contraintbased modeling approach shows promising results in terms of understanding interactions between a pathogen and its host from metabolic perspective, providing a list of affected metabolic pathways in both organisms.

This study was financially supported through a grant by TUBITAK (Project Code: 316S005).

Keywords: Infection, Genome Scale Metabolic Network; Pathogen-Host Metabolic Network; Dual RNA-Seq Data

References: [1] Su LH, Chiu CH. Salmonella: clinical importance and evolution of nomenclature. Chang Gung Med J. 2007 May-Jun;30(3):210-9.
[2] Rienksma RA, Suarez-Diez M, Mollenkopf H-J, Dolganov GM, Dorhoi A, Schoolnik GK, et al. Comprehensive insights into

transcriptional adaptation of intracellular mycobacteria by microbe-enriched dual RNA sequencing. BMC Genomics. 2015;16(1):34.
[3] Thiele I, Hyduke DR, Steeb B, Fankam G, Allen DK, Bazzani S, Charusanti P, Chen FC, Fleming RM, Hsiung CA, De Keersmaecker SC, Liao YC, Marchal K, Mo ML, Özdemir E, Raghunathan A, Reed JL, Shin SI, Sigurbjörnsdóttir S, Steinmann J, Sudarsan S, Swainston N, Thijs IM, Zengler K, Palsson BO, Adkins JN, Bumann D. A community effort towards a knowledge-base and mathematical model of the human pathogen Salmonella Typhimurium LT2. BMC Syst Biol. 2011 Jan 18;5:8.
[4] Blais EM, Rawls KD, Dougherty BV, Li ZI, Kolling GL, Ye P, Wallqvist A, Papin JA. Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nat Commun. 2017 Feb 8;8:14250.
[5] Westermann AJ, Förstner KU, Amman F, Barquist L et al. Dual RNA-seq unveils noncoding RNA functions in host-pathogen interactions. Nature 2016 Jan 28;529(7587):496-501.

Corresponding Author's Address:

Tunahan Çakır, tcakir@gtu.edu.tr

META-ANALYSIS PIPELINE FOR GENOMICS AND TRANSCRIPTOMICS VARIATIONS IN TCGA DATA

Talip Zengin[1], *Tuğba Süzek*[2]

1. Department of Bioinformatics, Graduate School of Natural and Applied Sciences, Mugla Sitki Kocman University, Kotekli Mugla Turkey, talipzengin@mu.edu.tr
2. Department of Bioinformatics, Graduate School of Natural and Applied Sciences,Mugla Sitki Kocman University, Kotekli Mugla Turkey, talipzengin@mu.edu.tr

TCGA (The Cancer Genome Atlas) is a vast and comprehensive database including molecular dataset consisting of genomics, transcriptomics, proteomics, epigenomics and clinical data of more than 11,000 cases across 33 tumor types [1]. In order to identify molecular nature of cancers, integrated molecular analysis must be performed and in accordance with this purpose, we have built a pipeline to carry out an integrated meta-analysis of the mutations including single-nucleotide variations (SNVs), the copy number variations (CNVs); gene expression detected by RNAseq; and clinical data of cancer patients downloaded from The Cancer Genome Atlas (TCGA) and reported their influence on overall survival through Cox Proportional Hazards Model. The workflow of R based designed pipeline is started with downloading of TCGA datasets of cancer of interest by TCGAbiolinks R package. Then, significant Simple Nucleotide Variations are detected by SomInaClust R package, validated by COSMIC mutations. Differentially Expressed Genes (DEGs) are determined by Limma/EdgeR R packages and then active subnetworks of these DEGs are determined by DEsubs R package. Copy Number Variations (CNVs) are determined by Gaia R package. Then all significant alterations are integrated through multi-variate Cox Proportional Hazard Regression Analysis in order to analysis effect of them all together on overall patients survival. Lastly, cancer patients are clustered based on detected impactful alterations and Kaplan-Meier survival analysis is performed for these patient clusters. In this study, we used lung adenocarcinoma (LUAD) dataset from TCGA database. We determined that the significantly mutated genes (RB1, STK11, EGFR and KRAS), with expression of 25 DEGs (WNT3A, VEGFD, GPC3, GNG7, ARHGEF6, PTGER4, PDE3B, ANGPT1, complement C2, CD244, CD74, IL11RA, KLRD1, COL4A3, LEPR, LPL, PLA2G3, KLF2, NR3C2, SCN4B) in active subnetworks, have the most impact on survival of the patients. The heatmap of these DEGs showed apparent dichotomous pattern and when the patients were clustered into two by hierarchical clustering and survival analysis was performed, survival plot showed a significant divergence between survival probability of two clusters of patients. Although RB1, STK11, EGFR, KRAS, WNT3A and VEGFD are known lung cancer related genes; and GPC3, GNG7, ARHGEF6, PTGER4, PDE3B, ANGPT1, complement C2, CD244, CD74, IL11RA, KLRD1, COL4A3, LEPR, LPL, PLA2G3, KLF2, NR3C2, SCN4B are recently identified lung cancer related genes, there is no evidence of direct interaction between PRKCE, RAPGEF3, COL6A5, ROBO2, CACNA1D genes and

lung cancer, yet. These 29 genes are strong candidates as molecular signature for clinical prediction for LUAD.

Keywords: TCGA; LUAD; SNV; CNV; DEA; Active Subnetwork; Cox; Clustering; Survival Analysis

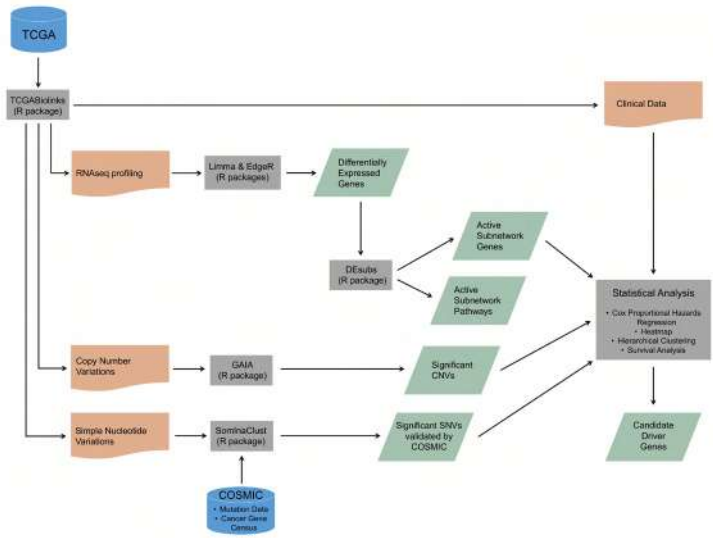


Figure 1. The flowchart of the pipeline

References: [1] Cancer Genome Atlas Research Network,, Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C., and Stuart, J.M. The Cancer Genome Atlas Pan-Cancer analysis project. Nat. Genet. 2013; 45; 1113–1120.

Corresponding Author's Address: Department of Bioinformatics, Graduate School of Natural and Applied Sciences, Mugla Sitki Kocman University, Kotekli Mugla Turkey, talipzengin@mu.edu.tr

TOPOLOGICAL ANALYSIS OF GENOME-SCALE METABOLIC NETWORK OF KLEBSIELLA PNEUMONIAE FOR DRUG TARGET DISCOVERY

Merve Yaşar Semiz, Müberra Fatma Cesur, Saliha Durmuş, Tunahan Çakır

1. Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey

Klebsiella pneumoniae is an opportunistic bacterial pathogen, associated with various life-threatening nosocomial infections (e.g., meningitis, pneumonia, endophthalmitis, cellulitis, and pyogenic liver abscess). Hypervirulent strains of the pathogen counteract the current therapeutic strategies through the development and dissemination of antibiotic resistance [1]. To manage this ever-increasing threat for public health and global economic cost, new treatment efforts must be introduced. Therefore, the aim of this study is to determine candidate drug targets using a topology-based approach, considering the role of the network topology in conservation and essentiality of the genes [2].

Genome-scale metabolic network (GMN) models provide a valuable way for a systems-based understanding of bacterial metabolism, through the mathematical representation of metabolic networks. These platforms provide a significant reduction in the solution space of the putative drug targets via in silico analyses. Thus, they are especially promising to reduce the need for expensive, time-consuming, and labor-intensive laboratory approaches. In this study, a GMN model of K. pneumoniae strain MGH 78578, called iYL1228 [3], consisting of 1,229 genes and 1,658 metabolites involved in 2,262 reactions was used. It includes a set of reactions dedicated to the metabolism of amino acids, fatty acids, and nucleic acids.

iYL1228 was converted into three different metabolic graphs for topological analysis (gene, metabolite, and reaction graphs). Metabolites involved in the same reaction were connected to construct the metabolite graph. For the gene graph, genes controlling the same or succeeding reactions were connected. The reaction graph was obtained by linking reactions with shared metabolites. Currency metabolites involved in many reactions were ignored in constructing the graphs. Each graph representing the same genome-scale metabolic network of the K. pneumoniae MGH 78578 was separately analyzed topologically in terms of degree distribution and betweenness centrality. The degree distribution gives information about the number of connections for each node while the betweenness centrality refers to the number of short paths passing over a node. They are commonly used metrics to identify the key nodes in a network, which have high potential to be drug targets to inactivate networks[4]. Using these metrics, essential nodes (enzymes) were identified through the gene, reaction, and metabolite graphs. They were comparatively analyzed based on the literature. The comparative analysis enabled identification of the most suitable graph structure to identify efficient drug targets. Comparison of the results with the candidate drug targets

from our recent work on constraint-based computational analysis for the same microorganism[5] provided a more comprehensive picture of metabolic mechanisms.

Keywords: Klebsiella pneumoniae; Infection; Genome-scale Metabolic Networks; Network Topology; Drug Target

References: [1] Paczosa MK, Mecsas J. Klebsiella pneumoniae: Going on the Offense with a Strong Defense. Microbiol Mol Biol Rev. 2016;80(3):629–61.
[2] Hwang YC, Lin CC, Chang JY, Mori H, Juan HF, Huang HC. Predicting essential genes based on network and sequence analysis. Mol Biosyst. 2009;5(12):1672–8.
[3] Liao Y, Huang T, Chen F, Charusanti P, Hong JSJ, Chang H, Tsai S, Palsson BO, Hsiung CA. An experimentally validated genome-scale metabolic reconstruction of Klebsiella pneumoniae MGH 78578, iYL1228. J Bacteriol. 2011;193(7):1710–7.
[4] Dickerson JE, Pinney JW, Robertson DL. The biological context of HIV-1 host interactions reveals subtle insights into a system hijack. BMC Syst Biol. 2010;4.
[5] Cesur MF, Siraj B, Uddin R, Durmuş S, Çakır T. Network-based metabolism-centered screening of potential drug targets in Klebsiella pneumoniae at genome scale. Front Cell Infect Microbiol (In Rev.)

Corresponding Author's Address:tcakir@gtu.edu.tr

EMBO
excellence in life sciences

Gold Sponsors

labgen QIAGEN LABORATUVAR SİSTEMLERİ SAN. VE TİC. A. Ş.

Eduline

Silver Sponsors

MedSanTek®
LABORATUAR MALZEMELERİ SAN. VE TİC. LTD. ŞTİ.

In-Kind Sponsors

MERCK

ŞAHİNLER KİMYA

TÜBİTAK 2223-B Organizing National Scientific Meetings Grant Programme

IZMIR BIOMEDICINE AND GENOME CENTER
Contact: hibit19@gmail.com
Dokuz Eylul University Health Campus Mithatpasa Ave. 58/5 35330 Balcova, Izmir/TURKEY