THE 14th INTERNATIONAL SYMPOSIUM ON HEALTH INFORMATICS AND BIOINFORMATICS

HIBIT 2021

ANKARA

10 - 11 SEPTEMBER 2021

The 14th International Symposium on Health Informatics and Bioinformatics

10-11 September 2021
Virtual

Contents

Committees

General Chair

A. Ercüment Çiçek Bilkent University

Organizing Committee

Can Alkan Bilkent University
A. Ercüment Çiçek Bilkent University
Özlen Konu Bilkent University

Program Committee

Aybar Acar Middle East Technical University
Ogün Adebali Sabancı University
Can Alkan Bilkent University
Jens Allmer Izmir Institute of Technology
Volkan Atalay Middle East Technical University
Ferhat Ay La Jolla Institute
Yeşim Aydın Son Middle East Technical University
Burcu Bakır-Güngör Abdullah Gül University
Ali Çakmak Istanbul Technical University
Tolga Can Middle East Technical University
A. Ercüment Çiçek Bilkent University
Tunca Doğan Hacettepe University
Serap Erkek Izmir Biomedicine and Genome Center
Cesim Erten Kadir Has University
Emre Güney Pompeu Fabra University
Gamze Gürsoy Yale University
Zerrin Işık Dokuz Eylül University
Ezgi Karaca Izmir Biomedicine and Genome Center
Gökhan Karakülah Izmir Biomedicine and Genome Center
Hilal Kazan Antalya International University
Özlen Konu Bilkent University
Mehmet Koyutürk Case Western Reserve University
Alper Küçükural University of Massachusetts

Maria Martin EMBL-EBI
Yavuz Oktay Dokuz Eylül University
Burçak Otlu University of California San Diego
Arzucan Özgür Boğaziçi University
Elif Özkırımlı Boğaziçi University
Athanasia Pavlopoulou University of Thessaly
João Rodrigues Stanford University
Osman Uğur Sezerman Acıbadem University
Aslı Süner Karakülah Ege University
Tuğba Süzek Muğla Sıtkı Koçman University
Barış Süzek Muğla Sıtkı Koçman University
Öznur Taştan Sabancı University
Nurcan Tunçbağ Middle East Technical University
Tuğba Arzu Özal İldeniz Acıbadem University

Poster Award Committee

Ogün Adebali Sabancı University
Tolga Can Middle East Technical University
Ali Çakmak Istanbul Technical University
Tunca Doğan Hacettepe University & EMBL-EBI
Burcu Güngör Abdullah Gul University
Hilal Kazan Antalya Bilim University
Özlen Konu Bilkent University
Yavuz Oktay Izmir Biomedicine & Genome Center
Ceren Sucularlı Hacettepe University

Student Organization Committee

Doruk Çakmakçı Bilkent University
Farid Ahadli Bilkent University
Fatma Kahveci Bilkent University
Furkan Özden Bilkent University
Halil İbrahim Kuru Bilkent University
İlayda Beyreli Bilkent University
Kerem Ayöz Bilkent University
Kübra Çalışır Bilkent University
Melike Tombaz Bilkent University
Merve Vural Bilkent University
Mohamad Fakhouri Bilkent University
Yaman Yağız Taşbağ Bilkent University

Speakers

Keynote Speakers

Ivet Bahar University of Pittsburg, USA
Ziv Bar-Joseph Carnegie Mellon University, USA
Igor Jouline Case Western Reserve University, USA

Invited Speakers

Rayan Chikhi Institut Pasteur, France
Tunca Doğan Hacettepe University, Turkey & EMBL-EBI, United Kingdom
Martin Kircher Berlin Institute of Health, Germany
Gioele La Manno EPFL, Switzerland
Niranjan Nagarajan Genome Institute of Singapore, A*STAR & National University of Singapore, Singapore
Elif Özkırımlı Roche, Switzerland
Irene Papatheodorou EMBL-EBI, United Kingdom
Maria Secrier University College London, United Kingdom

Welcome Address

The International Symposium on Health Informatics and Bioinformatics (HIBIT) is in its fourteenth year. It aims to bring together academics, researchers, and practitioners from medical, biological, and information technology sectors to create a synergy. It is one of the few conferences emphasizing such a synergy. It provides a forum for discussion, exploration, and development of theoretical and practical aspects of health informatics and bioinformatics. Also, it gives researchers a chance to follow current research in their field by constructing networks.

This year, we will host 3 keynote speakers, 8 invited speakers, 16 selected talks and 81 posters.

Welcome to HIBIT 2021!

The Organizing Committee
Can Alkan
A. Ercüment Çiçek
Özlen Konu

Abstracts

DiMA: Protein sequence diversity dynamics analyser for viruses

Yongli Hu[1], Shan Tharanga[2], Olivo Miotto[3,4], Eyyüb Selim Ünlü[5], Muhammet A. Çelik[6], Muhammed Miran Öncel[7], Hilal Hekimoğlu[6], Muhammad Farhan Sjaugi[2] and Mohammad Asif Khan[6,8]

[1]Beyond Limits SG Pte Ltd 13 Stamford Road 02-11 Capitol Singapore Singapore 178905
[2]Centre for Bioinformatics, School of Data Sciences, Perdana University, Kuala Lumpur, Malaysia
[3]Nuffield Department of Medicine, Oxford University
[4]Faculty of Tropical Medicine, Mahidol University, Bangkok
[5]İstanbul Faculty of Medicine, İstanbul University, Fatih, İstanbul, Turkey
[6]Bezmialem Vakıf University, Turkey
[7]Faculty of Medicine, Bezmialem Vakıf University, Fatih, İstanbul, Turkey
[8]Perdana University, Malaysia

Background: Pathogen sequence diversity is one of the major challenges in addressing viral infections. The diversity can be an outcome of a combination of underlying evolutionary processes (mutation, recombination, and assortment). A continuing goal is a greater understanding of viral proteome sequence diversity, the dynamics of substitutions, and effective strategies to overcome the diversity for drug or vaccine design. Herein, we present DiMA (https://github.com/PU-SDS/DiMA, also available via the Python Package Index, PyPI), a command-line tool designed to facilitate the dissection of protein sequence diversity dynamics for viruses.

Methodology: The tool accepts protein multiple sequence alignment (in aligned FASTA format) as an input and can be customized through various parameter settings. DiMA provides a quantitative measure of sequence diversity by use of Shannon's entropy (disorder), applied via a user-defined k-mer sliding window. For example, the k-mer length of nine is typically used for immunological applications, such as measure of antigenic diversity. Further, the entropy value is corrected for sample size bias by applying to the input alignment a statistical adjustment that estimates entropy values for an infinitely sized resampled alignments with analogous peptide distribution. Additionally, DiMA further interrogates the diversity by dissecting the entropy value at each k-mer position to various diversity motifs. The distinct k-mer sequences (k-mer_types) at the position are classified into the following motifs based on their incidence/frequency. Index is the predominant sequence, and all other sequences are referred to as total variants, sub-classified into major variant (the predominant variant), minor variants (comprising of k-mers with

incidence lower than major and higher than unique) and unique variants (k-mers seen once in the alignment). Moreover, the description line of the sequences in the alignment can be formatted for inclusion of meta-data, such as spatio-temporal information, among others, which can be annotated to the various k-mers comprising the different diversity motifs.

Results: DiMA outputs a JSON file that provides multiple facets of the sequence diversity. Sequence name, k-mer position, entropy, distinct k-mers at the position, and their incidence, motif classification and metadata (if available). The entropy values can be plotted to provide a panaromic overview of the protein sequence diversity, with low entropy representing conservation and high indicating variability. Separately, the diversity motifs enable diversity dynamics analyses of the virus protein.

Discussion: DiMA enables comparative sequence diversity dynamics analyses for a better understanding and insight within and between proteins of a virus species or proteomes of different viral species, whether at the genus, family, or higher lineage taxonomy rank. The tool can potentially be used for non-viral pathogens.

Convolutional neural network approach to distinguish and characterize tumor samples using gene expression data

Busra Nur Darendeli and Alper Yilmaz

Yildiz Technical University, Faculty of Chemical and Metallurgical, School of Bioengineering, 34225, Istanbul/ Turkey

Cancer is threatening millions of people each year and its early diagnosis is still a challenging task. Early diagnosis is one of major ways to tackle this disease and lower the mortality rate. Advancements in deep learning approaches and availability of biological data offer applications that can facilitate the diagnosis and characterization of cancer.

Here, we aimed to provide new perspective of cancer diagnosis using deep learning approach on gene expression data. In addition, we aimed to characterize the disease by identifying genes that are effective in cancer prediction.

In this study, The Cancer Genome Atlas (TCGA) dataset with RNA-Seq data of approximately 30 different types of cancer patients and GTEx RNA-seq data of normal tissues were used. The input data for the training was transformed to RGB format and the training was carried out with a Convolutional Neural Netowk (CNN). The trained algorithm is able to predict cancer with 90% accuracy, based on gene expression data. Moreover, we applied one-pixel attack on the trained model to determine effective genes for prediction of the disease.

In conclusion, our study shows that deep learning approach and biological data have a huge potential in diagnosis and characterization of tumor samples.

CoVrimer for SARS-CoV-2 primer prioritization

Merve Vural[1], Aslinur Akturk[2], Mert Demirdizen[2], Ronaldo Leka[2], Rana Acar[2] and Ozlen Konu[1,2]

[1] Interdisciplinary Neuroscience Program, Bilkent University, Ankara, Turkey 06800
[2] Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey 06800

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) genome rapidly accumulates mutations, interfering with the efficiency of primer binding to the variant genomes. It has been reported that some of the new variants have gained particular importance because of their potential for enhanced transmissibility, higher virulence, or decreased vaccination efficacy. Therefore, there is a great need for a web server that can help identify conserved primers effectively and develop degenerate versions of primers that exhibit frequent mutations. CoVrimer has been developed to fulfill the above-mentioned requirements and allows users to search for and align existing or newly designed conserved/degenerate primer pair sequences against the viral genome and assess the mutation load of both primers and amplicons. CoVrimer is implemented in R (version 4.0.2) as a Shiny app, providing interactive visualization of primer sequences and variant positions within the selected amplicon region. The mutation data, which is used to find conserved regions and mutation frequencies, are obtained from an online platform established by the National Genomics Data Center (NGDC) of the China National Center for Bioinformation (CNCB). Genomic regions with a mutation frequency of less than 0.1% are considered conserved, for which conserved primers are designed using the OpenPrimer package in R. Moreover, we have attained the genomic positions of all mutations that appear in the 10 most common SARS-CoV-2 lineages and generated primers containing degenerate versions of the nucleotides at these positions, making possible the detection of the presence of virus having any of the included variants. CoVrimer contains detailed information on the sequence and alignment of published primers and allows the users to enter their own primer pair sequence and inquire about available features of that primer pair via the CoVrimer pipeline. Accordingly, a degenerate version of the primer pair is created if the primer binding region contains any mutations, followed by the generation of a table presenting information on the primer pair such as length of primers, frequency of observed mutations in the primers, Tm and G/C content values, and amplicon length. In all modules, alignments of primers to the reference genome and the sequence of amplicons can be visualized in terms of mutation load. In conclusion, CoVrimer is the first conserved/degenerate primer selection and alignment web server for SARS-CoV-2 that considers viral evolution and is likely to help researchers with the challenges posed by viral evolution. It is freely available at http://konulabapps.bilkent.edu.tr:3838/CoVrimer/.

RNAseq based analysis for investigation of crosstalk among estrogen and drospirenone mediated signaling in breast cancer

Merve Vural, Kubra Calisir, Farid Ahadli, Ronaldo Leka, Irem Arici and Ozlen Konu

Department of Molecular Biology and Genetics, Bilkent University, Ankara, Turkey 06800

Hormone replacement therapy (HRT) is a medication containing estrogen along with progestins to relieve estrogen-deficiency symptoms and chronic diseases, including osteoporosis, heart disease, colorectal cancer, and depression [1]. Drospirenone is a member of progestins and is used with estrogen in HRT to relieve symptoms of menopause and in birth control pills to prevent unwanted pregnancy. With emerging NGS data methodologies including RNA-seq, Chip-seq, and Hi-C, and with the extensive bioinformatics methods and tools, the identification of transcriptome and transcriptional regulation across breast cancer, menopause, and HRT is now possible. For this purpose, we used MR overexpression models in MCF-7 breast cancer cells exposed to different combinations of aldosterone, drospirenone, and estrogen to test whether these drugs modulate similar and/or unique pathways, and we complemented our analyses with an R Shiny based user-interface to further analyze and visualize RNA-seq data with the incorporation of available related public data. qPCR was then used to validate findings, and future studies will focus on Chip-seq/ATAC-seq and/or Hi-Chip data to analyze DNA-RNA interactions simultaneously. Our project deals mainly with HRT used for menopause which embraces half of the human population, and it aims to discover the effects of HRT on breast cancer transcriptome and nucleome. This study has been funded by TUSEB (Grant No. 4405).

References

1. Minelli, Cosetta, et al. “Benefits and harms associated with hormone replacement therapy: clinical decision analysis.” Bmj 328.7436 (2004): 371.

An agent-based model to evaluate the effect of test kit usage of travelers in sparsely populated areas during pandemics

Baris Balaban, Erdem Berkay Bascura and Ugur Sezerman
Biostatistics and Bioinformatics Program, Institute of Health Sciences, Acibadem Mehmet Ali Aydinlar University, Istanbul, Turkey.

World Health Organization declared the global pandemic of COVID-19 on March 11, 2020. [1] Since then, the number of researches on pandemics greatly increased in various fields such as vaccination researches [2] and disease spread modeling. Such complex models help decision-makers to simulate macro outcomes by defining micro rules, especially in times of such infectious diseases. [3] For modeling such complex systems, agent-based models (ABMs) are used to simulate an environment with a specific set of micro rules defining individuals (agents) how to act including the probabilistic nature of actions. [4] Such work is crucial considering lowering the workload of the healthcare system is one of the main goals to serve every patient in need. In this study, we observe the effect of different strategies of travel restrictions on the disease spread in sparsely populated areas by using a multi-agent programmable modeling environment, NetLogo. [5] Our hypothetical simulation environment consists of 2 large and 4 small cities with a population of 500 and 50 residents, respectively. Each city is assumed to be fully connected in a given time point. Travel is restricted to only large city population that some of them may stay until the end of the simulation period. The simulation period is fixed to 4-months with 2 peaks (representing religious holidays) to simulate the holiday season in Turkey. The incubation period is assumed 4 to 6 days with equal probability and the period of isolation for travelers is set to 7 for obedient agents considering 2020 restrictions of the Turkish government. The infection rate of sparsely populated areas is assumed to be a stepwise function in relation to the current population of a particular city. The infection rate, the isolation percentage of travelers, the test kit usage of travelers and the probability of isolation of individuals showing symptoms for a given day are dynamic parameters to evaluate outcome measures in different conditions. Outcome measures to quantify the effect of test kit usage are area under the curve (AUC) of the infected percentage of small city native agents and maximum infected percentage of small city native agents in a given day. The model is tested to assure the mean values of applied rules are represented in the model and validated by estimating R0 which is a term to assess how contagious an infectious disease is, and comparing it to acceptable ranges seen in this pandemic. [6] R0 estimation of cities is done by calculating the total number of infected people by the ones who are infected now and dividing the number to previous ticks infected count excluding agents in incubation period. Local sensitivity analyses are conducted for outcome measures. The same set of random seeds are used to compare simulations, and paired t-tests are conducted with a 95% confidence interval using R version 3.6.3. [7] Local sensitivity analysis for dynamic parameters are conducted to select feasible parameters to investigate the effect of 5% increases in the test kit usage for travelers. 27 different parameter combinations for each kit-usage level is simulated for a test kit with a 10% false positive rate and a 97% true positive rate. For this particular test kit, we found that testing 15% of the travelers significantly diminishes the AUC of infected agents and the difference becomes insignificant when this

percentage increases. Mean (standard deviation) AUC of infected agents are 2.5% (2.2%) for 15% test kit usage simulations. On the other hand, not using test kits results in 23% of people that are in the infection period in a single day on average over 27 different parameter combinations. This percentage (the peak) drops to 4% when the test kit usage level is 25%. For future work, this study shows that ABM is applicable for decision-making processes and finding thresholds levels such as the percentage of the tested travelers.

References

1. Cucinotta D, Vanelli M. WHO Declares COVID-19 a Pandemic. Acta Biomed. 2020; 91(1) 157-160. doi:10.23750/abm.v91i1.9397. PMID: 32191675; PMCID: PMC7569573.

2. Polack, et al. Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. New England Journal of Medicine. 2020; 383 (27) 2603 – 2615. PMID: 33301246; doi 10.1056/NEJMoa2034577.

3. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. 2020; Nat Microbiol 5, 536–544. https://doi.org/10.1038/s41564-020-0695-z

4. Chowell G, Hyman J, Eubank S, Castillo-Chavez C. Scaling laws for the movement of people between locations in a large city. Physical Review E. 2003; 68:661021–661027. doi: 10.1103/PhysRevE.68.066102.

5. Wilensky, U. NetLogo. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL. 1999; http://ccl.northwestern.edu/netlogo/.

6. Petrosillo, et al. COVID-19 R0: Magic number or conundrum? Infectious Disease Reports. 2020; 12:8516 doi:10.4081/idr.2020.8516

7. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2020; https://www.R-project.org/

AnGenoV: A toolbox for analysis of genomic variants

Sevcan Doğramacı, Mert Doğan, Mehmet Gürsel Arslan, Süleyman Emre Çelik, Tuğba Nur Korkmaz and Arda Söylev

Department of Computer Engineering, Konya Food and Agriculture University, Konya,Turkey 42080

Evolution of genomes has been conducing to several variations because of mutations or other variation sources, causing generation of differentiations within a population. Since the genome is the source of hereditary and genetic information, these variants induce certain consequences on organisms, leading to genetic diseases, phenotypic abnormalities and more. Therefore, analyzing genomic variants is of crucial importance. General approach to analyse genomic variations include variant calling and variant annotation, where the first one involves the discovery of variants and the latter associates them with genes and functional impacts. Finally, manual curation of the variants is usually performed by utilizing visualization tools. There are several algorithms that use high-throughput sequencing technology to perform each of these tasks. However, one has to deal with each of them separately and the final conclusion is drawn by combining the results. Here we introduce AnGenoV (ANalysis of GENOmic Variants), which combines variant calling, variant annotation and visualization in one single application, thus enabling researchers to spend less time and effort with much higher accuracy. It has a modular structure and serves in a user-friendly graphical interface for easy usage, even for people with minor computational background.

Using AnGenoV, one can detect single nucleotide polymorphisms (SNPs), small insertions and deletions (Indels) and large structural variations (SVs) using short Illumina reads, as well as third generation long reads (PacBio or Oxford Nanopore). AnGenoV is configured with default state-of-the-art tools, however different algorithms can be added to its pipeline. Associating variations with several possible consequences is another critical point of AnGenoV, which is accessible from the variant annotation module. It incorporates a number of popular variant annotation databases including dbSNP, dbVar and Ensembl Variant Effect Predictor (VEP). On the other hand, we embedded Integrated Genomics Viewer (IGV) into AnGenoV so that selected variants can be visualized by tracking the reads that are mapped to the reference genome.

Another important aspect of AnGenoV that increases the accuracy of variant calling is the ability to run several variation discovery tools altogether, merge the results and output the most reliable set of variations. These final variations can be visualized in a user-friendly manner and desired variations can be searched without any UNIX command line knowledge (based on chromosome number, loci, variation type, etc). Then, the variants can be annotated by querying from multiple annotation databases. Finally, the variations can be visualized using IGV by just selecting the variation of interest among the list of possible candidates.

AnGenoV is implemented using JavaScript and Python and is freely available on https://github.com/SevcanDogramaci/AnGenoV.

On the path to reduce sugar intake: Sweet plant proteins

Nergiz Yuksel[1], Shokoufeh Yazdanıan Asr[2] and Burcu Kaplan Turkoz[3]

[1]Aydın Adnan Menderes University Germencik Yamantürk Vocational School Aydın, Turkey

[2]Ege University, Graduate School of Natural and Applied Sciences,Department of Food Engineering, İzmir,Turkey

[3]Ege University, Faculty of Engineering, Department of Food Engineering, İzmir, Turkey

Today, with the increase of conscious consumers, interest in healthy nutrition issues has intensified. As the negative effects of sucrose and artificial sweeteners are understood; The search for naturally derived alternatives, such as plant-derived sweeteners, has increased. Sweet plant proteins are several hundred or thousand times sweeter by weight and 10000 times sweeter on a molar basis than sucrose; calorie values are very low. Therefore; sweet plant proteins; It has attracted attention in terms of diabetes management, obesity control and oral health. So far, vegetable sweet proteins have been isolated from plants growing in tropical rainforests of Africa and Asia. Up to the present; Plant proteins of miraculin, thaumatin, monellin, pentadin, mabinlin, brazzein and curculin are known to provide sweetness. Since these proteins are produced by tropical plants and access to these plants is limited; focused on alternative methods such as recombinant protein production.

Our sweet sense; It is governed by the class C G-protein-coupled receptor (GPCR), which consists of 2 parts called tat type 1 Receptor 2 (T1R2) and tat type receptor 3 (T1R3) [6-7]. So far, comparison studies of amino acid sequences and 3-dimensional structures have been performed to detect conserved properties of sweet proteins. It has been suggested that these proteins may interact with the GPCR receptor. These proteins are thought to bind to the area surrounding the sugar binding site, tricking the receptor into thinking the sugar is binding.

In this study, it was aimed to examine homologous analogs of sweet proteins and to determine the similarity in gene regions. In our study, the protein family of Mabinlin and Thaumatin sequences was seen by BLAST search. Protein sequences obtained by BLAST were prepared with the COBALT multiple sequence alignment program and a phylogenetic tree was created. When the protein sequences were compared in detail, the proteins most similar to the mabinlin and thaumatin sequences were determined (respectively; XP_010539281.1, TXG69591.1). In addition, proteins were modeled with RaptorX and I-TASSER and high quality models were obtained according to quality criteria. A comparison of the structure model was made with Pymol. Our bioinformatics studies have shown that different plants also have genes encoding sweet protein-like proteins. Future studies; will focus on investigating the effects of different expression systems on recombinant protein production by cloning gene regions.

Assessment of the CASP14 assembly predictions

Burcu Ozden[1,2], Andriy Kryshtafovych[3] and Ezgi Karaca[1,2]

[1] Izmir Biomedicine and Genome Center, 35340, Izmir, Turkey

[2] Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, 35340, Izmir, Turkey

[3] Protein Structure Prediction Center, Genome and Biomedical Sciences Facilities, University of California, Davis, California, USA

In CASP14, 39 research groups submitted more than 2,500 3D models on 22 protein complexes. In general, the community performed well in predicting the fold of the assemblies (for 80% of the targets), though it faced significant challenges in reproducing the native contacts. This is especially the case for the complexes without whole-assembly templates. The leading predictor, BAKER-experimental, used a methodology combining classical techniques (templatebased modeling protein docking) with deep learning-based contact predictions and a fold-anddock approach. The Venclovas team achieved the runner-up position with template-based modeling and docking. By analyzing the target interfaces, we showed that the complexes with depleted charged contacts or dominating hydrophobic interactions were the most challenging ones to predict. We also demonstrated that if AlphaFold2 predictions were at hand, the interface prediction challenge could be alleviated for most of the targets. All in all, it is evident that new approaches are needed for the accurate prediction of assemblies, which undoubtedly will expand on the significant improvements in the tertiary structure prediction field.

AMULET: A novel read count-based method for effective multiplet detection from single nucleus ATAC-seq data

Asa Thibodeau[1,+], Alper Eroglu[1,+], Christopher S. McGinnis[2], Nathan Lawlor[1], Djamel Nehar-Belaid[1], Romy Kursawe[1], Radu Marches[1], Daniel N. Conrad[2], George A. Kuchel[3], Zev J. Gartner[2,4,5], Jacques Banchereau[1], Michael L. Stitzel[1,6,7], A. Ercument Cicek[8,9] and Duygu Ucar[1,6,7]

[1]The Jackson Laboratory for Genomic Medicine, Farmington, CT, 06032, USA [2]Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, 94158, USA [3]University of Connecticut Center on Aging, UConn Health Center, Farmington, CT, 06030, USA [4]Chan-Zuckerberg Biohub, San Francisco, CA, 94158, USA [5]NSF Center for Cellular Construction, San Francisco, CA, 94158, USA [6]Department of Genetics and Genome Sciences, University of Connecticut Health Center, Farmington, CT, 06030, USA [7]Institute for Systems Genomics, University of Connecticut Health Center, Farmington, CT, 06030, USA [8]Computer Engineering Department, Bilkent University, 06800, Ankara, Turkey [9]Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, 15213, USA [+]Equal Contribution

Detecting multiplets in snATAC-seq data is challenging due to data sparsity and limited dynamic range (0 reads: closed, 1: open in one chromosome, 2: open in both chromosomes). AMULET (ATAC-seq MULtiplet Estimation Tool) enumerates the regions with >2 uniquely aligned reads across the genome to effectively detect multiplets. We evaluated the method by generating snATAC-seq data in human blood and pancreatic islet samples. AMULET had high precision (estimated via donor-based multiplexing) and high recall (estimated via simulated multiplets) compared to alternatives and was the most effective when a certain read depth is achieved (25K median read count per nucleus).

Candidate antigen enrichment using scRNAseq data integration for CAR T cell therapy against non-small cell lung cancer

Mert Yıldız and Yasin Kaymaz

Bioengineering Department, Faculty of Engineering, Ege University, Izmir, Turkey.

Lung cancer is the most common disease worldwide and one of the deadliest diseases. Over 2.2 million new diagnoses are made annually, resulting in approximately 82% of deaths (WHO, 2021). Non-small cell lung cancer (NSCLC), which accounts for approximately 80% of lung cancer types, is the most common subtype. As for other cancers, there are various treatment options such as chemotherapy for lung cancer; however, survival rates are very low, and poor prognosis due to relapse or drug-resistant cases is still a major problem.

In the personalized medicine field, immuno-oncology approaches have gained attraction from the research community. One of the cutting-edge therapies is called chimeric antigen receptor (CAR) T cell therapy. With this approach, engineered chimeric receptors on cytotoxic T cells can increase binding specificity by recognizing the 3D structure of proteins on the outer surface of the tumor cell membrane as antigens. Thus, reprogrammed T cells that carry these so-called "specialized weapons" are expected to target and kill any tumor cell that carries these proteins on the cell surface. Even though it's a highly promising strategy, there is currently no FDA-approved CAR T cell therapy for NSCLC. One of the most important reasons is the low specificity of existing tumor antigens and their high side effects. Therefore, the determination of new target antigens and increasing the efforts to expand the candidate pool will open the doors of new immunotherapy-based clinical trials for NSCLC as well.

Expanding the candidate antigen pool is essential and more research efforts need to be channeled towards searches for new tumor-specific targets. Better bioinformatics workflows that can streamline target antigen identification and selection process are required. In the last 10 years, single-cell sequencing technologies have gained tremendous speed and have been the main approach used in more than 1200 studies. In these studies, more than 17 million individual cells have undergone scRNAseq for various purposes. Of these, roughly 10% represents various types of tumor cells providing new insights into tumor-specific transcriptome profiles at the single-cell resolution. Here in this study, we hypothesized that repurposing these datasets to identify alternative tumor-specific antigens for CAR T cell therapy for NSCLC can be achieved with appropriate data integration and analytical processing. To achieve this, we developed a target enrichment scheme that meticulously integrates various scRNAseq datasets, filters, normalizes, and prepares for differential gene expression analysis between multiple combinations of cell types. This complex analysis is successful only if cell type identities are correctly determined. Our analysis pipeline utilizes a machine learning approach, i.e., HieRFIT (Kaymaz et al., 2021 Bioinformatics), for hierarchical cell type classification of tumor and normal cell types. Thus, better tumor-specific antigens will be selected, and their toxigenic effects will be foreseen when used in CAR T cell therapy.

Here, we will introduce our analytical approach in detail and share our preliminary results leading to an effective antigen candidate. We believe that our work will contribute to a paradigm shift in the selection of candidate antigens for CAR T cell therapy by reprocessing existing single-cell transcriptome data. This datamining proof-of-concept using single-cell sequencing and metadata will help expand the scope of available treatments for other malignancies as well.

Novel approach for microbiome meta-analysis

Farid Musa and Efe Sezgin

Department of Food Engineering, Izmir Institute of Technology, Izmir, Turkey

Breakthroughs in DNA sequencing technology have pushed microbiome research to new limits, where previously unknown microbes not only can be taxonomically identified but also quantified. Recent studies on microbial interactions between the host genome and its internal gut microbiome have unveiled yet unknown effects of microbiota on our health, boosting the call for more studies. However, with the growing research interest for microbiome studies, demand for novel data analysis methods and software are increasing as well. Most of the currently available amplicon-based microbiome data analysis tools focus on processing raw sequence data and generating OTU-table aka. frequency table as the final product. However, the microbiome research community currently lacks comprehensive meta-analysis tools to process multiple OTU-tables that were originally processed differently; hence, impeding researchers to perform survey studies without raw sequences. Dissimilarities in frequency tables among independent microbiome studies are caused by several factors such as different sequencing platforms and library size, which are used to generate raw data, or OTU-picking tools and taxonomic classification databases, which are used to process the raw data and frequency tables. Here we present our microbiome meta-analysis framework developed to address aforesaid shortcomings. Phylogenetic Microbiome Analysis Framework(PhyloMAF) is based on the Python programming language with flexible and extendable design and is openly available on github.com/mmtechslv/PhyloMAF. Our framework is not limited to solely performing meta-analysis studies based on frequency tables and aims to provide much wider functionality. Particularly, the PhyloMAF was developed to provide a basic framework to work with biome data and taxonomic classification databases like Greengenes, SILVA, GTDB, etc. In addition to local databases, our framework also provides functionality to work with remote NCBI taxonomy. Moreover, the framework includes a "pipe" module that can be used to easily build data processing pipelines to interact with databases and work with biome data like taxonomy sequences, accession numbers or phylogenetic trees. In this poster, we demonstrate benchmark results of microbiome meta-analysis performed with both microbiota datasets and simulated 16S rRNA amplicon data. The alpha and beta diversity analysis of OTU-tables processed via PhyloMAF and via traditional tools like QIIME2 are compared. Our findings demonstrate that PhyloMAF proves to be an effective meta-analysis tool with no rivals in the literature.

Robust prediction of genetic mutation effects by homology analysis

Alperen Taciroğlu[1], Yeşim Aydın Son[1] and Ogün Adebali[2]
[1]Middle East Technical University
[2]Sabancı University

Obscurin (OBSCN) along with Titin and Nebulin is a giant sarcomeric protein that is mainly expressed in human striated muscle cells. Tandem domain structure is a trademark of giant sarcomeric proteins. Tandem immunoglobulin domains (I-set) in the Obscurin sequence are thought to have evolved to endure mechanical stress caused by contractions and relaxations occurring in the sarcomere structure. Also, these I-set domains are the main contributor to the repetitive, large 8000 amino acid length structure of this protein and partially the reason why the function of Obscurin transcripts are poorly understood. Although Obscurin variants are associated with cardiomyopathies and cancer, causative variant effects are still being actively debated. In order to assign functions and predict the effect of mutations occurring in these transcripts, we conducted a detailed homology analysis, combined with domain-level annotation and clustering on the Obscurin B transcript. As a result, we obtained 35 different clusters containing sequences mapped to the corresponding transcript. 17 of these clusters contain sequences annotated as Pfam domain families (Pfam clusters) and 18 of the clusters contain sequences that are not annotated as any domain by Pfam (non-Pfam clusters). However, conservation analysis revealed that like Pfam clusters non-Pfam clusters also contain significantly conserved sub-sequences. This finding lead us to believe that some of these sequences might correspond to functional unannotated linker domains or unexplored functional residues that are not part of the Pfam domain families. Therefore, we believe resulting clusters are important assets to find out highly conserved sub-sequences in Obscurin and can help us predict the effect of mutations occur at the residue level.

Support Vector Machine supported by Disease Ontology (SVM-DO) to identify mRNA signatures discriminating tumour cells

Mustafa Erhan Ozer, Pemra Ozbek Sarica and Kazim Yalcin Arga
Department of Bioengineering, Marmara University, Istanbul, Turkey

Finding tumour discriminative genes is challenging due to the complex nature of the disease. The transcriptome-based supervised classification by Support Vector Machine (SVM) is recently gaining popularity in this field. The existence of less significant variables in the construction of classification models can result in misclassification errors. To this end, application of feature selection methods such as enrichment analysis extracts useful sets of variables for better model performances. Gene Ontology (GO) terms are mainly preferred despite having annotation problems. On the other hand, cancer and chronic disorders have repeatedly been linked, hinting a possible connection. But there is still a few genetic studies on cancer occurrence in this field. Hence, discovering disease-associated genes could be a viable strategy for identifying a gene set that distinguishes tumour and normal states. To this end, we used SVM classification of Disease Ontology (DO) enriched differentially expressed genes to construct an algorithm. Wilk's Lambda Criterion filtration is used to remove non-discriminative features at the data level among disease-associated genes prior to the classification method. The algorithm is applied on RNA-Seq data by extracting the selected number of most up-regulated and down-regulated genes. The algorithm's classification model is found to perform well when it came to predicting cancer patients. According to the study-wise comparisons, our generated gene sets with lesser number of variables predicted cancer samples with good accuracy. Acquired gene sets are also shown to be useful in distinguishing between tumour and normal states in various expression datasets and cancer prognosis. Our approach may contribute to future cancer studies for more targeted clinical applications by unifying gene sets in both diagnosis and prognosis. The developed technique will be displayed in a graphical user interface called by a single R package.

Bayesian networks for inter-omics analysis

Muntadher Zahid Jihad and İdil Yet

Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey

Single cell sequencing technologies have begun to take a role in solving serious biological and medical problems that remained elusive in earlier time. Single-cell omics are providing data with a high resolution on cellular phenotypes and different communication networks between and inside cells. Integrating these different omics provides a valuable insight about the cellular mechanism of complex diseases such as cancer. Here, we proposed the use of Bayesian Networks (BN) as a method of single cell multi-omics integration. We tested different BN models in order to explore the causal relationship between the genomic, transcriptomics and epigenomics in 25 Hepatocellular Carcinoma (HCC) single cells. Three BN models were constructed to represent the alternative hypotheses of the causal relations between copy number variation (CNV), DNA methylation and gene expression. We first analyzed the datasets on their own. We used Hidden Markov Methods (HMM) for the estimating of CNVs. Considering the CNV as the starting point of the causal model, the three models are: INDEP, CME, and CEM. In "INDEP" model, CNV affects independently DNA methylation and gene expression. In "CME" model, CNV affects DNA methylation, which in turn affects gene expression. In "CEM" model, CNV affects gene expression, which then affects DNA methylation. The parameters of these networks were estimated by using maximum likelihood estimation (MLE). The three alternative models were applied to all the detected genes over the 25 HCC single cells. Then, the best fitted model for each gene were chosen according to Akaike information criterion (AIC) and Bayesian information criterion (BIC) scores. We only kept the models when the best model is at least ten times more likely to be than the second-best model. Thereafter, genes with a verified model were further investigated by using cBioportal database. We showed the HCC heterogeneity by showing different genes in different or the same signaling pathways have followed different BN models. Moreover, we have defined the BN model for 207 genes (CEM: 169, CME: 34, INDEP: 4) that have been previously reported in at least one HCC study in cBioportal database. Lastly, the results have shown that HLA gene which follows CME model can be a candidate gene that might play a major role in HCC development. By following CME model, we could say that the role of HLA gene in HCC is dependent on the DNA methylation levels.

Bioinformatic analysis of Bifidobacterium breve TIR domain

Bahar Bakar[1], Dicle Dilara Akpınar[2] and Burcu Kaplan Türköz[3]

[1]Ege University, Graduate School of Natural and Applied Sciences, Department of Food Engineering, Izmir,TR

[2]Yildiz Technical University, Graduate School of Science and Engineering, Department of Food Engineering, Istanbul, TR

[3]Ege University, Faculty of Engineering, Department of Food Engineering, Izmir, TR

Toll-like receptor (TLR) signaling pathway is one of the main pathways in innate immune system where dimerization of Toll-interleukin-1 receptor (TIR) domains play a key role [1]. All organisms including some bacteria and plants possess TIR domain proteins. Animal and plant TIR domain proteins function in immunity and defense whereas the function of bacterial TIR domain proteins are intriguing as they are shown to interact with mammalian TIR domains. Several TIR domain proteins from pathogenic bacteria were shown to mimic mammalian TIR domains and manipulate TLR signaling [2,3]. Recently bacterial TIR domain proteins were also shown to function in NAD metabolism [4].

Probiotics are beneficial microorganism and many probiotics were shown to modulate host signaling pathways, including TLR signaling. In an attempt to identify putative probiotic TIR domain proteins we found a TIR domain protein encoded in Bifidobacterium breve genome. The protein, which we named BbTIR, is highly similar to bacterial and mammalian TIR domains as shown by multiple sequence alignments. The structure of BbTIR was modeled using I-TASSER and RaptorX and high quality models were obtained based on the quality criteria. Structural alignment was carried out using PyMOL and Chimera and RMSD values were evaluated between the BbTIR model and other known TIR domains structures. Sequence and structure based analysis verified the high similarity of BbTIR to other TIR domains. Furthermore, our experimental studies confirmed that BbTIR is a functional TIR domain as the protein interacted with human TIR domains in pull down assays. Future research will focus on in vivo experiments to elucidate the interaction mechanism between probiotic TIR proteins and TLR signaling.

References

1. O'Neill, and Bowie, (2007). The family of five: TIR-domain-containing adaptors in Toll-like receptor signalling. Nature Reviews Immunology, 7(5), 353-364.

2. Chan, et al., (2009). Molecular mimicry in innate immunity: crystal structure of a bacterial TIR domain. Journal of Biological Chemistry, 284(32), 21386-21392.

3. Kaplan-Türköz, et al., (2013). Structure of the Toll/interleukin 1 receptor (TIR) domain of the immunosuppressive Brucella effector BtpA/Btp1/TcpB. FEBS letters, 587(21), 3412-3416.

4. Coronas-Serna, et al.,. (2020). The TIR-domain containing effectors BtpA and BtpB from Brucella abortus impact NAD metabolism. PLOS Pathogens, 16(4), e1007979.

Comparison of the performances of in silico pathogenicity prediction tools on cancer-related variants

Metin Yazar[1,2] and Pemra Özbek Sarıca[1]

[1]Department of Bioengineering, Marmara University, Göztepe, İstanbul, Turkey
[2]Department of Genetics and Bioengineering, Istanbul Okan University, Tuzla, Istanbul, Turkey

Single nucleotide variations (SNVs) are the most widespread cause of alterations in human DNA and they could play exclusive functional roles in the cell processes such as gene expression, disease susceptibility, and protein-protein interactions. The phenotypic consequences of SNVs can be neutral or negative effect on protein function or structure. Experimentally classification of variants can be slow, inefficient and inconvenient, hence in silico tools and algorithms have become popular for the prediction of the variants' effects from physical, biochemical and biological parameters. Cancer is a complex and heterogenous disorder including abnormal cell growth with a possibility of spreading to other parts of human body. Genomic instability and variations in human DNA have been demonstrated as a hallmark of cancer progression, thus cancer-related SNVs have been getting more attention in clinical oncology. Cancer-related SNVs have different classifications such as somatic-germline variants or passenger-driver variants but prediction and identification of their phenotypic effect on protein function and structure are important for drug responses and treatment options in clinical oncology. There are several consensus guidelines for the interpretation and classification of variants suggesting the usage of in silico tools and algorithms for prediction. ClinVar is the most common public variation database, involving clinical consequences with genotype and disease data. In ClinVar, there are many cancer-related missense variants, however interpretation and classification of a substantial amount of these variants remains uncertain or elusive. In literature, there are many studies on the evaluation of in silico prediction tools' performances with benchmark datasets; however none of these have targeted the assessment of the performance on cancer related variant datasets from a public database. In this study, we aimed to evaluate and compare the performance of 13 different in silico tools or algorithms, including SIFT, CADD, FATHMM-weighted, FATHMM-unweighted, GERP++, MetaSVM, Mutation Assessor, MutationTaster, MutPred2, PolyPhen-2, Provean, Revel and VEST4 for investigation of functional effects of the variants from 8 different cancer datasets (bladder, breast, colon, colorectal, kidney, liver, lung and pancreas cancer) obtained from ClinVar. For this purpose, datasets were retrieved from ClinVar and filtered based on their review status, molecular consequences and variation types. Then, prediction scores of each variant were obtained using in silico tools' own websites and Variant Effect Predictor (VEP). All prediction scores were transformed into binary forms such as "Benign" and "Pathogenic" according to each tools' pathogenicity thresholds. Performance evaluation and statistical analysis were performed on the normalized prediction scores . For the statistical classification of variants in different cancer datasets, confusion matrix was created with several parameters. Receiver operating characteristic (ROC) curve analysis and correlation analysis of normalized prediction scores obtained from in silico tools were also conducted. As a result, MetaSVM,

VEST4 and SIFT were found to have the closest prediction distribution frequencies when compared to ClinVar's distributions in all cancer datasets. Investigation of features and sources used by these in silico tools revealed that sequence conservation/identity score were the common categories among those tools. Statistical performance analysis and ROC curve results have demonstrated that MetaSVM has the highest discriminatory power for all cancer datasets.

Phylogeny-aware amino acid substitution scoring

Nurdan Kuru, Onur Dereli, Emrah Akkoyun, Aylin Bircan, Öznur Taştan and Ogün Adebali

Sabancı University

With the advancement in high throughput sequencing technologies, our ability to detect genetic variation, and to predict the effect of a variant in the clinical diagnosis have been revolutionized. Single nucleotide polymorphisms (SNPs) in coding regions might cause the change of a single amino acid into another in the resulting protein (i.e., missense mutations). These mutations might have no effect on protein function, or it can alter protein function which might result in a disease. Understanding the effect of a missense mutation with respect to whether it has a neutral or disease-causing effect on protein function helps to diagnose rare diseases. Although the cost of genome sequencing has decreased, it is still a challenging task to assess the functional consequences of variations. For this purpose, several conservation-based statistical and machine learning approaches have been proposed in the literature to predict the potential consequence of a variant. SIFT and PolyPhen-2 are the most widely used tools of such approaches. Although these methods do not yield the desired level of accuracy and as a result are not suggested to be used in clinical studies, the clinicians use them to prioritize and reduce the number of variants to be analyzed.

Here, we introduce a novel phylogeny-dependent probabilistic approach, Phylas (Phylogeny-Aware Amino Acid Substitution Scoring) to predict the functional effects of missense mutations. Our approach exploits the phylogenetic tree information to measure the deleteriousness of a given variant. Independent evolutionary events and phylogenetic relationship among species are driven from the gene-based phylogenetic trees. With the help of ancestral reconstruction, we obtain the probability distribution at each internal node of the phylogenetic tree. Starting from the queried specie, which is human in our algorithm, we travel through the tree and record the probability change for each amino acid. Although the positive change in probability is a result of an alteration, the negative changes are observed as the effect of a substitution that belongs to the previously covered part of the tree. To include each dependent change at once, the negative changes are ignored in the computation. In addition to taking the dependent alterations into account, the effects of independent substitutions observed during the evolutionary time repeatedly should be included. To account for this effect, we present a correction over the score by considering the count of the independent substitutions guaranteeing that the harmfulness of the related alteration is decreased. It has been previously hypothesized that a variant in the human gene is more likely benign when it is observed in closely related species, whereas it is more likely deleterious when it only exists in distant ones. Thus, during the travel through the phylogenetic tree, all positive changes on amino acid probabilities are summed by a weighting approach inversely proportional to the distance between related node of change and human. After completion of the travel, we obtain substitution scores for each of the 20 amino acids for the given position of the query sequence. The normalized resulting values give us the probability of observing any amino acid at the given position of the protein in question.

This probability is used to measure the pathogenicity of a possible amino acid substitution. Although the query sequence is human in our experiments, the approach can be easily used for other species by changing the starting point and the direction of the travel through the tree.

We compare the predictive performance of our algorithm against SIFT and PolyPhen-2. We generated the benchmark datasets by combining variants from Clinvar, Humsavar, Gnomad, and four other datasets proposed in Varibench. Our algorithm outperforms SIFT and PolyPhen-2 in predicting the pathogenicity of missense mutations by improving the AUROC values by 3 and 7% respectively (see Figure 1).

Discovering coding lncrnas using deep learning training dynamics

Afshan Nabi, Berke Dilekoğlu, Ogun Adebali and Öznur Taştan
Sabanci University, Faculty of Engineering and Natural Sciences

Genome-wide transcriptome analyses have revealed that the vast majority of the human genome is transcribed; but only 2% of the human genome is annotated as protein coding. A considerable fraction of transcripts are annotated as ncRNAs and lncRNAS constitute the largest category of ncRNAs. While lncRNAs studied are known to play vital roles in cellular processes, the functions of most lncRNAs remain unknown. Moreover, although lncRNAs - by definition- do not code for proteins, recent studies have shown that short the open reading frames (sORFs) within some lncRNAs are translated into micropeptides of a median length of 23 amino acids. The translation events of lncRNAs were overlooked previously because the open reading frames (ORFs) present in lncRNAs do not meet the conventional criteria of an ORF: that it encodes at least 100 amino acids in eukaryotes. Despite this, recent studies have shown that micropeptides translated from lncRNAs perform vital functions across species, including bacteria, flies and humans. Therefore, identifying misannotated lncRNAs is a necessary step towards the functional 1 characterization of this large class of transcripts.

We present a framework that leverages deep learning models' training dynamics to determine whether a given lncRNA transcript in the dataset is misannotated. In the first step, we train convolutional neural network (CNN), long short term memory (LSTM) and Transformer architectures to predict whether a given nucleotide sequence is non-coding or coding. Each input RNA sequence is represented with 3-mer 'words' are obtained by using a window that slides by 1 nucleotide at each step and for each 3-mer 'word', 100-dimensional embeddings are used as input. Our models learn to distinguish between coding and non-coding RNAs with average AUC scores >91% (Table 1).

In the second step, we inspect the training dynamics of these deep sequence classifiers to identify possible misannotated lncRNAs. By inspecting lncRNAs where the model consistently and with high confidence predicts as coding through all training epochs, we identify the possibly misannotated candidates. Example of training epochs for different RNAs are given in Figure 1.

Through this inspection, we identify candidate lncRNAs. Our results show a significant overlap with previous methods that use riboseq data to identify misannotated lncRNAs as well as with a set of experimentally validated misannotated lncRNAs. Moreover, we search proteins similar in sequence to the candidates and curate a subset with high similarity to known proteins. This work represents the first instance where deep learning model training dynamics are successfully applied to identify misannotated lncRNAs from nucleotide sequences. This approach can be applied to better curate datasets for training coding potential prediction models and can be applied alongside ribo-seq data to identify misannotated lncRNAs with high confidence.

G-Protein selective activation mechanisms in GPCRs

Berkay Selcuk[1], Ismail Erol[2], Serdar Durdagi[2] and Ogun Adebali[1]

[1]Sabanci University

[1]Bahcesehir University

G-protein coupled receptors (GPCRs) induce signal transduction pathways through coupling to four different subtypes of G-proteins (Gs, Gi, Gq, G12/13). Each G-protein induces a different subcellular pathway resulting in distinct physiological responses. While a single receptor can couple to multiple G-proteins, receptors having diverse evolutionary backgrounds can couple to the same G-protein. Therefore, it is important to understand the features required for receptors to couple to different G-proteins to develop better therapeutic approaches. There has been previous research to identify determinants of coupling selectivity within the interface between receptor and G-protein. However, very little attention was given to receptor level G-protein selective coupling mechanisms and residual determinants. Here, we propose a model that conserved G-protein specific activation mechanisms exist within a subfamily of GPCRs and they determine receptors' ability to couple to a particular G-protein. In our study, we worked with aminergic receptors due to availability of mutational data, structural structures and coupling profiles. Phylogenetic analyses were conducted to identify orthologs of each aminergic receptor with a novel methodology. Then we used orthologous protein sequences from various eukaryotic organisms to calculate conservation scores for each position. By grouping receptors based on previously known coupling profiles, we identified a pool (n=51) of specifically conserved residues that are likely to determine the coupling selectivity for nine different G-protein subtypes and 22 residues that were conserved in all aminergic receptors. For different G-proteins the functional distribution of specifically conserved residues varies in terms of their spatial and functional clusters, indicating the mechanistic differences required for a successful G-protein engagement. Initially, we were able to annotate some of these specifically conserved residues into function clusters such as ligand binding or G-protein interacting. Moreover, to gain more insight into their roles in receptor activation, we benefited from a previously proposed methodology to calculate residue-residue contact scores (RRCS) for a given experimental structure. We calculated differences in contact scores (ΔRRCS) between inactive-state and G-protein coupled active-state structures for eight different receptors and four G-protein subtypes. Through statistical comparison of ΔRRCSs calculated for each G-protein subtype, we identified statistically higher or lower contact changes ($\Delta RRCS \neq 0$, p 0.05) indicating the presence of G-protein specific activation mechanisms for Gs, Gi1, Gio and Gq. Using the specific activation mechanism for Gs, we hypothesized that the increasing contact between 7x41 and 6x48 is a key mediator of differential TM6 tilt that has been observed previously within the structures of Gs-coupled receptors. The existing mutational information supports our hypothesis and shows that 7x41 Glycine is intolerant to any mutational change. We are the first group to identify the functional role of this residue for Gs coupling. To further validate our hypothesis and understand the molecular impact of the identified position, we performed all-atom molecular dynamics (MD) simulations. First, we modeled missing residues for ADRB2 and its orientation in the membrane calculated with the

OPM web server. We created four different systems containing three mutated and one wild-type ADRB2 (PDB ID: 3SN6) in a monomeric active-state. Simulations were performed with Gromacs version 2020 using CHARMM36m force field for the protein, POPC lipid, and TIP3 water model. 7 replicates of simulations for 500 ns (in total 3.5 µs) revealed that the WT receptor maintained its active-state more than its mutated variants. Additionally, we aim to perform in-vitro mutational analysis for Gi1, Gio, and Gq and measure the change in coupling levels of the receptors to validate the networks that we obtained for selective G-protein coupling. Our results indicate that GPCRs contain an evolutionary conserved, subfamily, and G-protein specific set of specifically conserved residues which determine coupling selectivity. With the increasing number of experimental structures and known coupling profiles, our approach is applicable to all GPCR subfamilies. We plan to expand our analyses for other subfamilies of GPCRs in the future.

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources).

Analysis of structural and functional impact of SNVs in hAKT1 gene using in silico tools

İlayda Üzümcü and Elif Uz Yıldırım

Bursa Uludag University, Faculty of Arts & Science, Department of Molecular Biology and Genetics, Bursa, Turkey

Background/Aim: AKT1 (AKT Serine/Threonine Kinase 1) gene plays a role in fundamental cellular processes such as apoptosis, metabolism, migration, proliferation, cell survival, gene expression, and differentiation in order to maintain homeostasis. AKT1 has a role in PI3K/AKT pathway as a main effector. Variations in this gene have been reported to contribute to the pathogenesis of Cowden syndrome, breast, colorectal, ovary cancer, Proteus syndrome and schizophrenia. Our aim in this study is to evaluate the functional effects of SNVs using different in silico tools, investigate evolutional significance of its SNVs, and find out disease-related SNVs and to summarize the allele frequencies in AKT1.

Methods: The exonic, splice region, 3 prime UTR, 5 prime UTR, frameshift, inframe deletion, inframe insertion, splice acceptor, and stop gained SNVs of AKT1 gene were filtered using VariationViewer. Functional consequences of of exonic SNVs were evaluated using PolyPhen2, SIFT and Mutation Taster web tools. A combined score for each SNV was determined using a combined approach from these three in silico prediction tools. The allele frequencies of the variants were obtained using GnomAD dataset. The regulatory effect of the SNVs were determined using RegulomeDB tool. Finally, the post-translational consequences of the variants were evaluated using Phosphosite Plus web tool.

Results: A total of 251 SNVs, of whom 239 are exonic were obtained from Variation Viewer. 107 (45%) of them were determined as benign and 132 (55%) were damaging according to our combined approach. The most number of variants are located on exon 11, from where part of the protein kinase domain is expressed. According to the RegulomeDB analysis, 44 of the SNVs have a score of 2a and 2b, which means these regions have important regulatory effects. Allele frequencies of 163 SNVs were obtained from GnomAD database and none of them was above 0.05. On AKT1 protein; 14 phosphorylation, 6 ubiquitination, and one acetylation sites were to be determined using Phosphosite Plus web tool. The variations may have effects on these post-translational modification sites.

Conclusion: Using various in silico platforms, whether SNVs contribute to the development of the disease, their evolutionary importance, functional effects, and prevalence in the population were investigated in AKT1. Since most SNVs do not occur in a protected area in the evolutionary process, they can be tolerated. Since they are rarely seen in the society, their presence can cause serious damage and cause various diseases.

Bioinformatic analyses of Heparinase HepIII from Azospirilum brasilanse

Seyhan İçier[1] and Burcu Kaplan Türköz[2]

[1]Ege University, Graduate School of Natural and Applied Sciences, Department of Biotechnology, İzmir, Turkey

[2]Ege University, Faculty of Engineering, Department of Food Engineering, İzmir, Turkey

Heparan sulfates are glycosaminoglycans, which are components of extracellular matrix and function in cell-cell associations, cell signaling, cell growth and differentiation in mammals [1,2]. Heparan sulfate interacts with fibroblast growth factor (FGFs) and receptors (FGFRs), and generates FGF-HS-FGFR signaling complexes [2]. Heparinase, is an enzyme which degrades heparan sulfate and is produced from animals and bacteria. Heparinase was shown to have a role in modulation of FGF signaling as antithrombotic agent [2,3]. Heparinases are important enzymes for the production of heparin and heparan sulfates as an antithrombotic agent for pharmaceutical industry [3, 4]. Microbial heparinases have potential for production of low molecular weight heparins with anticoagulant effect and therefore are valuable for biotechnology research[5]. Microbial heparinase enzymes are classified in three groups based on their recognition of different sulfation levels; heparinase I, heparinase II and heparinase III. Heparinase III generally cleaves at the sulfate rare regions of heparan sulfate resulting in the antithrombotic effect [2]. Azospirillum brasilanse is a nitrogen-fixing bacterium which functions as a plant growth factor. The genome of A. brasilanse encodes for a heparinase III (HepIII) enzyme based on sequence comparisons. Following these, we hypothesized that heparinase from A. brasilanse could be a promising candidate for antithrombotic applications. In order to verify this, here a structural based approach is undertaken for structure-function analysis of heparinase III from A. brasilense and other microbial heparinase enzymes. Briefly, heparinase III is modeled using I-TASSER and RaptorX, and structural models were compared with known structures based on rmsd values. This initial bioinformatic analysis will help identification of novel microbial heparinase enzymes.

References

1. Hashimoto, W., Maruyama, Y., Nakamichi, Y., Mikami, B., & Murata, K. (2014). Crystal Structure of Pedobacter heparinus Heparin Lyase Hep III with the Active Site in a Deep Cleft. Biochemistry, 53(4), 777–786.

2. Dong, W., Lu, W., McKeehan, W. L., Luo, Y., & Ye, S. (2012). Structural basis of heparan sulfate-specific degradation by heparinase III. Protein cell, 3(12), 950-961.

3. Singh, V., Haque, S., Kumari, V., El-Enshasy, H. A., Mishra, B. N., Somvanshi, P., & Tripathi, C. K. M. (2019). Isolation, purification, and characterization of heparinase from Streptomyces variabilis MTCC 12266. Scientific reports, 9(1), 1-8.

4. Rivera, D., Revale, S., Molina, R., Gualpa, J., Puente, M., Maroniche, G., ... & Cassán, F. (2014). Complete genome sequence of the model rhizo-

sphere strain Azospirillum brasilense Az39, successfully applied in agriculture. Genome announcements, 2(4), e00683-14.

5. Zhang, C., Yang, B. C., Liu, W. T., Li, Z. Y., Song, Y. J., Zhang, T. C., & Luo, X. G. (2019). Structure-based engineering of heparinase I with improved specific activity for degrading heparin. BMC biotechnology, 19(1), 59. 3201

Co-expression networks from transcriptome data reveal molecular mechanisms playing roles in the progression of Parkinson's disease

Tunahan Çakır and Elif Emanetçi
Gebze Technical University

Parkinson's disease (PD) is one of the most common progressive neurodegenerative diseases, and it affects physical functions. PD is characterized by the loss of dopaminergic neurons and Lewy body aggregation in substantia nigra region in brain. The aggregation disturbs synaptic communication, and it also has other effects on brain tissue such as proteasome dysfunction, mitochondrial dysfunction, and oxidative stress. Progression of PD is defined by six Braak stages. Early stages (Braak 1-2) are pre-symptomatic stages and characterized by loss of non-motor functions. In mid-stages (Braak 3-4), patients loss motor functions. At the final stages (Braak 5-6), patients have all other symptoms and the disease moves into the other brain parts. To understand the underlying mechanisms of PD and the effect of Braak stages on progression of PD, we used transcriptome data from PD patients. Here we created a co-expression interaction networks from the Braak-stratified PD transcriptome data to identify the altered mechanisms while the disease progresses. First, we created a coexpression network with decreasing correlation pattern from control to Braak 5-6 or from Braak 5-6 to control stages by using Pearson correlation. Further, we created more specialized networks from these co-expression networks by using Weighted gene correlation network analysis (WGCNA) algorithm and divided these specialized networks to functional communities using Leiden algorithm. We have two decreasing co-expression networks from Braak 5-6 stages to control and one co-expression decreasing network from control to Braak 5-6 stages. Enrichment analysis of network genes reveals PD related mechanism such as mitochondrial mechanisms, synaptic signaling, and oxidative phosphorylation. WGCNA networks were more specialized, and they have all previous terms in the enrichment analysis in addition to motor function abnormalities, abnormal reflex, coordination, dopamine deficiency, bradykinesia, and involuntary movement. Dividing the networks to smaller functional units by Leiden algorithm leads to three major communities with specific mechanisms related to PD: (i) signaling mechanism and axon maintenance, (ii) energy and mitochondrial mechanisms and (iii) PD related mechanisms and mitochondrial functions. There are also other genes and other mechanisms that has not yet been associated with PD based on the literature. With the network-based approach, we show that it is possible to capture unknown mechanisms and genes that trigger PD and cause its progression at molecular level.

Identification of major depression related transcriptional changes through integration of multiple datasets

Berkay Selcuk, Tuana Aksu and Ogun Adebali
Sabanci University

Major Depressive Disorder (MDD) is a commonly observed psychiatric disorder that affects more than 2% of the world population with a rising trend. However, pathways and biomarkers that are associated with MDD are yet to be clarified. In this study, we analyzed previously generated RNA-seq data across six brain regions from three different studies to identify genes that are differentially and co-expressed for MDD by considering covariates of gender, age, postmortem interval, brain region, and the dataset they belonged to. Differential gene expression (DGE) analysis revealed a significant downregulation in immediate-early genes (IEGs), especially for NPAS4, FOS, and FOSB. Furthermore, we conducted a co-expression analysis and identified gene modules that play a role in glutamatergic signaling which controls pathways for repeated-learning and circadian entrainment. We combined both DGE and co-expression analyses and associated a novel pathway to MDD. The results suggest that disruption in glutamatergic signaling causes downregulation of mainly NPAS4 and other IEGs that decrease synaptic plasticity in patients with MDD. We anticipate that our study will open doors to develop better therapeutic approaches targeting glutamatergic receptors.

A story of an online internship in computational structural biology

İrem Yılmazbilek[1,2] and Ezgi Karaca[2,3]
[1]Middle East Technical University, Ankara, Turkey 06800
[2]Izmir Biomedicine and Genome Center, Izmir, Turkey 35340
[3]Dokuz Eylül University, Izmir, Turkey

As a molecular biology and genetics graduate student, you are commonly expected to perform a wet lab internship. However, there is another wide and different world of biology that I have had the opportunity to experience outside the laboratory door. With our ability to adapt, which is passed down in our genes for years, together with my supervisor, Dr. Ezgi Karaca and her group, we changed the internship rules and rewrote them to adapt to the current pandemic conditions.

Our six weeks online internship program is designed to teach the basics of computational structural biology to undergraduate students. When this introductory session is finished, the intern gets the chance to be assigned to a specific project in the lab. The first phase of our curriculum is oriented around the grounds of computational structural biology and why it is crucial to dissect molecular mechanisms. For this, the intern goes over the "Protein-Protein Interaction Interfaces and Their Functional Implications" section of the "Protein-Protein Interaction Regulators" book by Gideon Schreiber[1]. Then, the principles of structure determination and acquiring PDB data are thought through the online RCSB-PDB and EMBL-EBI resources (specifically: "Guide to Understanding PDB Data[2] and Biomacromolecular structures; an introduction to EMBL-EBI resources"[3]). This didactic learning stage is followed by running the tutorials of PyMOL and the molecular modeling program HADDOCK. Following these tutorials create an active learning environment for students to help them learn by themselves. The internship is finally finished by learning UNIX and shell scripting tools upon following online UNIX resources. Throughout this internship program, each intern is assisted by an experienced member of the group, which helps tracking the learning curve of the intern.

So, in this poster, our aim is to reflect on our experience of the online internship program we offered, from which four biology-centered students benefited during the 2020-2021 term. By presenting what we have learned from our experiences, we also aim to discuss the developments in online internship programs performed at different universities.

References

1. G. Schreiber, CHAPTER 1:Protein–Protein Interaction Interfaces and their Functional Implications , in Protein–Protein Interaction Regulators, 2020, pp. 1-24 DOI: 10.1039/9781788016544-00001

2. PDB101: Learn: Guide to Understanding PDB Data: Introduction. RCSB. (n.d.). https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/introduction.

3. Argasinska J, Symmons MF, Gutmanas A, Kleywegt GJ, . Biomacromolecular structures. European Bioinformatics Institute (EMBL-EBI); 2011. DOI: 10.6019/tol.bms.2011.00001.1

Comparison and assessment of speed and accuracy of AutoDock Vina and AutoDock CrankPep for short peptide docking

Sefer Baday and Numan Nusret Usta
Istanbul Technical University

Protein-peptide interactions (PPI) are one of most interesting area of molecular biology due to their importance in biological processes. In computational biology, molecular docking is commonly used technique for investigating the PPI. However, due to the flexible structure of peptides, docking of peptides to proteins are very challenging due to high conformational degrees of freedom. Various molecular docking tools are available with different search and scoring algorithms for different docking needs such as protein-protein, protein-ligand, nucleic acid-ligand and etc. For investigating the effectiveness of molecular docking tools, some benchmark studies were done by different research groups. But, especially for protein-peptide docking, these studies could seem inadequate in terms of number of studied complexes and peptide sequence length diversity of complexes. By considering these circumstances, we choose PepBDB which is a comprehensive structural database of biological protein-peptide complexes. PepBDB has more than 13000 complexes that their peptide lengths ranged from 1 to 50. In this study, we worked with two different tools of most common free docking softwares AutoDock CrankPep (ADCP) and AutoDock Vina. Accuracy of results are determined with RMSD. We performed local docking for all complexes and evaluated the top pose of each of them. We analyzed docking results of the peptides by their sequence length. Docking results of shorter peptides are better compared to longer ones for both of the docking tools. Also, time consumption of longer peptides is much more in terms of peptide per minute. Success rate of ADCP is noticeably better than Vina for shorter peptides. However, the gap between these programs is decreasing as peptide length increases. We also determined the optimal parameters for both programs by the sequence length.

---

miRModuleNet: Detecting miRNA-mRNA regulatory modules

Malik Yousef[1], Gokhan Goy[2] and Burcu Bakir Gungor[3]

[1]Zefat College

[2]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri

[3]Abdullah Gul University

MicroRNAs have been demonstrated to have a key role in carcinogenesis in a growing number of studies, therefore understanding the regulation mechanism of miRNAs in the gene-regulatory network is crucial. Understanding the interactions of miRNA and mRNA will help to elucidate the complex biological processes that occur during malignancy. This fact has led to the development of different tools that capture those interactions. Our previous tool miRCorrnet has utilized this fact and suggested a machine learning approach that is based on grouping and scoring (ranking) of groups, where each group is a miRNA and its members are the genes that correlate with this specific miRNA. The input to miRCorrnet tool is two -omics data, i.e. miRNA and mRNA expression profiles. The same input is now subject to our new tool called miRModuleNet. miRModuleNet is based on grouping and scoring by grouping mRNA (genes) for each miRNA to form the star shape, which is the miRNA-mRNA regulatory module. Then the scoring procedure is applied on each module to rank those modules in terms of classifications. One of the useful outputs of miRModuleNet is the list of significant miRNA-mRNA regulatory modules. The tool was validated on external datasets and its output was validated against known databases; additionally functional enrichment analysis was performed. miRModuleNet could aid in the identification of functional relationships between these biomarkers, revealing essential pathways involved in cancer pathogenesis.

Predicting side effects of chemotherapy using drug-induced gene expression profiles and a random forest-based strategy

Ozlem Ulucan

Department of Genetics and Bioengineering, Istanbul Bilgi University, 34060 Eyupsultan, Istanbul, Turkey

About 90% of the drug candidates fail in clinical trials, with the majority of failures due to safety or efficacy issues despite in depth preclinical testing practices. Late-stage failures have serious implications such as site closures, job losses and reduced research and development budgets, inflation in prices for successful drugs and discouraging development of innovative drugs for pharmaceutical industry as well as for general public. Identification of side effects in this regard has great potential to improve the probability of introduction of new drug candidates and thus avoiding the consequences mentioned above.

Chemotherapeutic agents are notorious for yielding side effects that significantly reduce the quality of life for the patients. Among the most common side effects induced by chemotherapy are alopecia (hair loss), oedema, and diarrhea, although their severity depends on the drug, dosage of the drug and frequency of the treatment. While there have been proposed mechanisms by which anticancer agents cause these side effects, the exact underlying mechanisms have not been entirely identified. It is vital to understand the mechanism of the side effects to pave the way for new therapeutic agents with little or more tolerable side effects. In this study, we addressed the common side effects of chemotherapeutic agents, namely alopecia, diarrhea and oedema. Our aim was twofold: 1) to develop a model that uses drug-induced transcriptome response to predict side effects accurately and 2) to explain the mechanism of the side effects of interest. We employed an iterative strategy based on Random Forest algorithm and unveiled an expression signature involving 40 genes that predicted these side effects with an accuracy of 89%. We further characterized the gene expression signature and its association with the side effects using functional enrichment analysis and protein-protein interaction networks. The approach that we employed in this study can be generalized to other side effects. This work contributes to the ongoing efforts in drug development for early identification of side effects to use the resources more effectively.

Identification novel inhibitors targeting putative Dantrolene binding site for Ryanodine receptor 2

Cemil Can Saylan[1] and Sefer Baday[2]

[1]Computational Science and Engineering Department, Informatics Institute Istanbul Technical University

[2]Applied Informatics Department, Informatics Institute Istanbul Technical University

Ryanodine receptors (RyRs) are large (20 MDa) homotetrameric intracellular ion channels. RyRs are located in the membrane of sarcoplasmic reticulum (SR). RyRs play a central role in the excitation-contraction coupling by regulating Ca2+ release from the SR to the cytosol. There are three isoforms sharing 70% sequence similarity; RyR1, RyR2 and RyR3 predominantly expressed in skeletal, cardiac muscles and neurons, respectively. Dysregulation of RyRs leads to abnormal cellular activity. More than 300 mutations have been associated with muscle and neuronal diseases. Previously, there have been identified several heart diseases caused by aberrant RyR2 activity such as catecholaminergic polymorphic ventricular tachycardia, cardiomyopathies, and cardiac arrhythmias. This unwanted RyR2 activity can be modulated by drugs. Dantrolene is an approved muscle relaxant used for treatment of malignant hyperthermia. However, dantrolene has poor water solubility and dantrolene exerts its effect on RyR2 only in the presence of regulators such as Mg2+ and calmodulin. In addition to this, although the amino acids 590-609 in RyR1 (601-620 RyR2 equivalent) were identified as dantrolene binding sequences, dantrolene bound complex structure has not been elucidated yet. Here, we aim to identify novel inhibitors targeted to the putative binding sequence of dantrolene to regulate RyR2 function. While most of the structure of RyR2 has been solved recently, some of the regions are still missing. To predict missing regions, we used trRosetta which is a developed deep-learning-based method and was ranked as the second group after AlphaFold2 at CASP13. After that to predict dantrolene binding pose, blind docking approach was performed using three different docking programs; Vina, LeDock and Glide. Three binding poses were selected based on the scores and binding pose similarity of docking results. 200ns molecular dynamics (MD) simulation was performed on these three binding poses. Two orientations of dantrolene remained in their binding position. Afterwards, we focused on the dantrolene binding region and applied a high-throughput screening of 3 million molecules retrieved from the ZINC15 database. For virtual screening a series of docking calculations were performed. The first initial screen of 3 million molecules were carried out using AutoDock Vina program with exhaustiveness parameter set to 8. Next, top-ranked 200K molecules were redocked with LeDock and Vina programs (with 24 exhaustiveness). Molecules that are ranked in top-10K for both Vina and LeDock results were selected and redocked with GlideXP. In the last step, top-100 molecules were selected and ADME properties were evaluated using QikProp. Among the ADME properties, the best 10 were suggested as candidates that regulate RyR2 activity. As a future work, predicted molecules will be tested experimentally by our collaborators.

Piperidine-including natural drug discovery for inhibition Type 4 Pili's (T4P) in P. aeruginosa and N. meningitidis

Aslıhan Özcan Yöner[1], Özlem Keskin Özkaya[2], Berna Sarıyar Akbulut[1,3] and Pemra Özbek Sarıca[1,3]

[1]Institute of Pure and Applied Sciences, Department of Bioengineering, Marmara University, Istanbul, Turkey

[2]College of Engineering, Chemical and Biological Engineering, Koc University, Istanbul, Turkey

[3]Faculty of Engineering, Department of Bioengineering, Marmara University, Istanbul, Turkey

The rapidly increasing resistance to available antibiotics and the reduced rate in new antibiotic discovery lead to different therapeutic approaches. An attractive strategy is to identify molecules that target bacterial virulence as an alternative to traditional antibiotics with low efficacies. Anti-virulence therapies cleanse the pathogens from their weapons instead of killing the pathogens that cause infections in humans [1], [2].

The current work undertakes the effort to target the type 4 pili (T4P) in P. aeruginosa and N. meningitidis. T4P is an important virulence factor in both and is the main target of this study. The selected two microorganisms are among the World Health Organization's list of pathogenic bacteria for which there is urgence to develop new therapeutics. To this end, PilB of P. aeruginosa and PilF of N. meningitidis have been studied in detail. As a novel strategy, piperidine-containing natural product solutions were screened for their inhibition of pilB and pilF. Piperidine is an incredibly significant building block in the production of pharmacological substances and is the most regularly utilized heterocycle among US FDA approved medications [3].

The strategy followed in this work started with homology modelling of pilB and pilF of P. aeruginosa and N. meningitidis, respectively, since their structures were not available. The binding sites in the target structures were determined by metaPocket 2.0. Here the ADP binding regions have been observed to play an important role in the to inhibit T4P. As the drug database, COlleCtion of Open Natural prodUcTs (COCONUT) resource was used and for filtration FAF-Drugs4 was used. Finally, in order to select natural products to inhibit T4P by binding pilB and pilF, virtual library screening was performed. Ligands with binding energies better than -9.0kcal/mol for both structures were accepted to be potential inhibitors. Molecules selected in this work might have a potential to be used in novel therapeutic applications in combination with existing drugs and/or other virulence factor inhibitors.

References

1. D. A. Rasko and V. Sperandio, "Anti-virulence strategies to combat bacteria-mediated disease," Nature Reviews Drug Discovery, vol. 9, no. 2. Nature Publishing Group, pp. 117–128, Feb. 18, 2010, doi: 10.1038/nrd3013.

2. K. Denis et al., "Targeting Type IV pili as an antivirulence strategy against invasive meningococcal disease," Nat. Microbiol., vol. 4, no. 6, pp. 972–984, 2019, doi: 10.1038/s41564-019-0395-8.

3. “Piperidine-Based Drug Discovery - 1st Edition.” https://www.elsevier.com/books/piperidine-based-drug-discovery/vardanyan/978-0-12-805157-3 (accessed Jun. 14, 2021).

Studying complex human diseases using time-series ancient DNA data: Obesity and Type 2 Diabetes in Anatolia

İdil Taç[1], Ulaş Işıldak[2], Kıvılcım Başak Vural[2], Ezgi Altınışık[3], Yılmaz Selim Erdal[3], Mehmet Somel[2], Füsun Özer[1,4] and Gülşah Merve Kılınç[1,2]

[1]Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, 06100, Ankara, Turkey

[2]Department of Biological Sciences, Middle East Technical University (METU), Ankara, Turkey

[3]Department of Anthropology, Hacettepe University, Ankara, Turkey

[4]Human G Lab, Department of Anthropology, Hacettepe University, Ankara, Turkey

Studying the evolutionary processes behind complex human diseases is of great interest in evolutionary genetics. Two tantalizing questions are (i) how and which evolutionary forces have influenced the genetic variation associated with complex human diseases and (2) how the long-term dietary shifts of ancient humans have affected the susceptibility of modern humans to these diseases. In order to shed light on these questions, we focus on two complex metabolic diseases; obesity and type 2 diabetes, and study the evolutionary dynamics behind these two diseases in Anatolia, where the prevalence of both diseases is comparably high today. To investigate the temporal dynamics of evolutionary processes behind the genetic variation associated with obesity and type 2 diabetes we analyzed published modern and ancient genomes from Anatolia using ancient genomics approaches. We compiled all single nucleotide polymorphisms (SNPs) associated with these diseases from GWAS catalog which results in 144 SNPs (p ¡ 0.005). We produced time-series genotype datasets for these SNPs associated with obesity-and type 2 diabetes as well as for 1,000 neutral SNPs using two different approaches including genotype likelihoods with ANGSD and pseudohaploid genotypes using pileupcaller. We investigated the allele frequency trajectories over the last 10,000 years of Anatolia which reveals that risk allele frequencies for a set of SNPs fluctuate between different time periods in the region (e.g. the frequency of rs1111875 on HHEX gene in Anatolia is 0.46 for present-day; 0.35 for present-day-3,500 cal BCE, and 0.12 for 3,500-8,500 cal BCE Anatolia (Dunn test for pairwise comparisons have p value ¡ 0.05). We further investigated the linkage disequilibrium structure around these SNPs over time. Our preliminary analysis indicates that the fluctuating allele frequencies over time could be due to demography and/or change of diet, or social structure and requires further investigation.

Sequence diversity of envelope protein of Dengue virus serotype 1

Gökçen Şahin[1], Li Chuin Chong[1], Erdem Aybek[1] and Asif M. Khan[1,2]

[1]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Turkey

[2]School of Data Sciences, Perdana University, Malaysia

Background: Dengue virus (DENV), a member of the Flaviviridae family, is transmitted to the human host by mosquito vectors, in tropical and subtropical regions of the world, causing infection of a wide spectrum in the host, ranging from mild fever to potentially fatal severe dengue hemorrhagic fever. DENV evolvability is constrained by reliance on both the host and the vector for its propagation. There are four distinct, but closely related serotypes of the virus, which pose a challenge to vaccine development and infection management. The virus genome is a positive sense, single-stranded RNA of approximately 11 kb length, which encodes three structural and seven non-structural proteins. The structural protein, envelope (E), typically of length 493-495 amino acids, is critical for entry into the host cell. Herein, the E protein sequences of dengue virus serotype 1 (DENV-1) isolated from human and mosquito were studied to examine the sequence diversity between the host and the vector.

Methodology: DENV-1 E protein sequences were retrieved from the NCBI Virus database. The host sources for the sequences were examined and those isolated from human and mosquito were observed to outnumber those from other sources, such as bats and primates, and thus, were downloaded separately. The E protein, UniProt record of accession P17763 was used as a reference for BLAST search against the downloaded sequences, to generate the DENV-1 E protein datasets for human and mosquito. Duplicates were removed from each dataset by use of CD-HIT V4.8.1. The datasets were merged for a co-alignment by use of MUSCLE and the results were manually inspected for misalignment and corrected by use of the UGENE analysis suite. The co-alignment was then split, resulting in a dataset each for the host/vector, allowing a comparative analysis. Viral sequence diversity dynamics analysis was performed by use of the DiMA (https://github.com/PU-SDS/DiMA), which utilises Shannon's entropy to quantify diversity for every overlapping k-mer positions of the alignment. The k-mer length of nine was chosen for immunological applications.

Results: The number of DENV-1 E protein sequences retrieved from the NCBI Virus database was 11,025 for human and 254 for mosquito (various species), which collectively covered nearly all (89.7%) of the host/vector sequences in the database. The numbers decreased to 2,131 and 50, respectively, after deduplication, indicating high (80%) redundancy in the dataset. The average entropy across the overlapping, aligned nonamer (9-mer) positions was 0.70 and 0.51 for human and mosquito viruses, respectively, suggesting high conservation across the DENV-1 E protein within the host/vector. Peak entropy for human viruses was at position 337 with an entropy value of 2.10, while for mosquito viruses, the peak value of 1.99 was observed at position 153, postulating different evolutionary patterns between the human host and the mosquito vector.

Discussion: The diversity analysis of DENV-1 E protein provides an insight into the evolutionary differences between the human host and the mosquito vector. A caveat to be noted is that it is not clear how many of the isolated human and mosquito viruses underwent both the host/vector environments. This study serves as a pilot for similar analyses of other proteins for the four dengue serotypes.

Investigating potential interplay between R-loops and nucleotide excision repair

Sezgi Kaya and Ogun Adebali
Sabanci University

Non-canonical structures on DNA have been attracting attention in the context of gene regulation and genome instability. R-loops, a class of these structures, are formed on DNA when an RNA molecule anneals on its complementary DNA strand, leaving the opposite DNA strand single-stranded (ssDNA). To date, R-loops have been associated with transcription, protection of promoters from methylation, immunoglobulin class switch recombination, as well as double-strand break (DSB) formation and genome instability. However, although there is increasing evidence about their importance and potential functions, how their mechanisms of action or why they cause DSBs are still not clear. In this project, we aimed to investigate potential effects of R-loops on UV-induced DNA damage and nucleotide excision repair (NER), and to understand more about R-loop functions and genomic distributions. For that purpose, first, we analyzed the distributions of these structures on different states and regions of the genome, using qDRIP-seq and RR-ChIP-seq datasets that provided the genomic coordinates of R-loops. We have integrated previously published RNA-seq, GRO-seq, BLESS-seq and ChIP-seq datasets in our analysis to gain more insight about the relationship between R-loops and transcription and DSBs, as well as protein-binding on R-loops. In order to find out about the repair profiles on R-loop-forming regions on DNA, we have analyzed Damage-seq and XR-seq datasets which provide the locations of UV-induced damage and excision products, respectively. Our results have shown that the UV-induced damages on R-loop-containing regions were repaired more efficiently than other DNA regions. When the two strands of R-loops were compared, DNA:RNA hybrid strands were more efficiently repaired than ssDNA strands. In addition, RPA, an ssDNA-binding protein that plays role in damage recognition steps of NER, had a better binding on R-loop-containing regions than other regions, which might be a clue for the reason why R-loops are repaired more efficiently. Moreover, RPA-bound regions that intersects with an R-loop were repaired better than other RPA-binding regions, suggesting a potential combined effect of R-loops and RPA on repair efficiency. To sum up, R-loops might promise a lot in terms of DNA damage repair and further investigation of R-loop functions and effects will lead us understand more about their positive and negative impacts on our genomes.

PersonaDrive: A Computational Approach for Prioritization of Patient-specific Cancer Drivers

Cesim Erten, Aissa Houdjedj, Hilal Kazan and Ahmed Amine Taleb Bahmed
Antalya Bilim University

A major challenge in cancer genomics is to distinguish the driver mutations that are causally linked to cancer from passenger mutations that do not contribute to cancer development. The majority of the methods proposed for this problem provide a single driver gene list for the entire cohort of patients. On the other hand, it is well-known that the mutation profiles of patients from the same cancer type show a high degree of heterogeneity. Since each patient has a distinct set of driver genes, a more ideal approach is to identify patient-specific drivers.

In this study, we propose a novel method that integrates genomic data, biological pathways, and protein connectivity information for personalized identification of driver genes. The method is formulated on a personalized bipartite graph which consists of the mutated genes of the patient in one partition, and the set of patientspecific dysregulated genes in the other. Our approach provides a personalized ranking of the mutated genes of a patient based on the sum of weighted 'pairwise pathway coverage' scores across all the patients, where an appropriate pairwise patient similarity score with respect to the sets of dysregulated genes of the pair is employed as a weighting factor of pathway coverage scores of mutant-dysregulated pairs. We compare our method against three state-of-the-art patient-specific cancer gene prioritization methods; Prodigy [1], SCS [2] and DawnRank [3]. The comparisons are with respect to a novel evaluation method which takes into account the personalized nature of the problem; different from previous evaluation methods employed in literature, we assume the results are reported as average values for the entire cohort as a function of the top k ranked genes, where k is dependent on the size of the personalized reference set and individuals with less than k ranked genes are excluded from the evaluation. Two main data sets are considered for the evaluations; TCGA and cell-line data (DepMAP [5]) for colon and lung cancers. For the TCGA data, the reference sets are defined based on the overlaps of mutated genes of the patient and the Cancer Gene Census (CGC) [6], Network of Cancer Genes (NCG) [7], and CancerMine[8] databases of known driver genes. For the cell-line data, we define novel reference gene sets by compiling the targets of drugs that are found to be sensitive based on data from GDSC [4] and DepMAP [5] databases. We show that our approach outperforms the existing alternatives for both the TCGA and cell-line evaluations. Additionally, we show that the KEGG/Reactome pathways enriched in our ranked genes and those that are enriched in cell lines reference sets overlap significantly when compared to the overlaps achieved by the rankings of the alternative methods. Fig 1 shows the performance of all methods on TCGA and CCLE cohorts. The findings of our approach can lead to the development of personalized treatments and therapies.

References

1. Gal Dinstag, et al. (2020) Bioinformatics. 1831–1839

2. Guo WF, et al. (2018). Bioinformatics. 1;34(11):1893-1903.

3. Hou, J.P., Ma, J. (2014). Genome Med 6, 56.

4. Wanjuan Yang, et al. (2013). Nucleic Acids Research, V.41, Pages D955–D961.

5. Steven M Corsello, et al. (2019). bioRxiv.

6. John G Tate, et al. (2019). Nucleic Acids Research, V.47, P.D941–D947.

7. Repana D., et al. (2019). Genome Biol.

8. Lever J. et al. (2019). Nat Methods, 16:505–507

PROT-ON: A Python package for redesigning the protein-protein interfaces by using EvoEF1

Mehdi Koşaca[1,2], Berçin Barlas[1,2] and Ezgi Karaca[1,2]

[1]Izmir Biomedicine and Genome Center, Dokuz Eylul Health Campus, Izmir
[2]Izmir International Biomedicine and Genome Institute, Dokuz Eylul University, Izmir

Nonsynonymous single nucleotide polymorphisms (SNPs) in the coding regions (exons) can affect the stability or functionality of a protein. While the effect of these SNPs on the protein structure can be calculated by experimental methods, they can also be estimated by computational tools, such as EvoEF1 [1,2], FoldX [3], and MutaBind2 [4]. All these tools aim to interpret the impact of an interfacial protein mutation on the already known protein protein interactions by using classical force field terms. The general function of these algorithms usually aims to predict the binding affinity or stability change upon mutation. In this work, we utilized the best performer of all these tools, EvoEF1, to redesign the known protein protein interactions.

For this, we developed the package PROT-ON (PROTein mutatiON), uses the EvoEF1 algorithm as the mutation and binding energy calculation tool. The algorithm of PROT-ON follows: interface_residues.py script finds all amino acids of a given monomer in contact with its partner within 5A cut-off and mutates them to all other 19 amino acid combinations. Afterwards, energy_computation.py script builds mutation models with BuildMutant command of EvoEF1 and calculates the binding affinities of each mutation with an optimized scoring function of EvoEF1. Finally, detect_outliers.py script analyzes the distribution of all mutations with boxplot, as well as a heatmap statics to focus on negative and positive outliers to offer most enriching and depleting mutations as output. Here, our assumption is that these mutations are the most prone ones to enhance or abolish the known binding pattern. Running PROT-ON takes two minutes on a standard laptop. For me, please check our GitHub page: https://github.com/CSB-KaracaLab/find-designer-mutations

References

1. Robin Pearce, Xiaoqiang Huang, Dani Setiawan, Yang Zhang. EvoDesign: Designing Protein–Protein Binding Interactions Using Evolutionary Interface Profiles in Conjunction with an Optimized Physical Energy Function. Journal of Molecular Biology (2019) 431: 2467-2476.

2. Xiaoqiang Huang, Robin Pearce, Yang Zhang. EvoEF2: accurate and fast energy function for computational protein design. Bioinformatics (2020), 36:1135-1142. https://doi.org/10.1093/bioinformatics/btz740

3. Buß, O., Rudat, J., & Ochsenreither, K. (2018). FoldX as Protein Engineering Tool: Better Than Random Based Approaches?. Computational and structural biotechnology journal, 16, 25–33.
https://doi.org/10.1016/j.csbj.2018.01.002

4. Zhang, N., Chen, Y., Lu, H., Zhao, F., Alvarez, R. V., Goncearenco, A., ... Li, M. (2020). MutaBind2: Predicting the Impacts of Single and Multiple

Mutations on Protein-Protein Interactions. IScience, 23(3).
https://doi.org/10.1016/j.isci.2020.100939

GeNetKEGG: Gene expression based KEGG pathway grouping and ranking

Malik Yousef[1], Fatma Ozdemir[2], Amhar Jabeer[3], Jens Allmer[4] and Burcu Bakir-Gungor[5]

[1]Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel
Galilee Digital Health Research Center (GDH), Zefat Academic College, Israel
[2]Department of Electrical and Computer Engineering, Graduate School of Engineering & Science, Abdullah Gul University, Kayseri, Turkey
[3]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey
[4]Institute of Measurement and Sensor Technology, Hochschule Ruhr West University of Applied Sciences, 45479 Mülheim an der Ruhr, Germany
[5]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey

Kyoto Encyclopedia of Genes and Genomes (KEGG) analyzes systematically gene functions genes and molecules as a knowledge base. The PATHWAY database is the main component of KEGG. It includes graphical diagrams of biochemical pathways consisting of common metabolic pathways and some of the common regulatory pathways. KEGG supply estimating gene regulatory networks from the gene expression profiles and reconstructing biochemical pathways from the complete genome sequence. In this study, we propose a new approach called GeNetKEGG, which is based on gene expression as the KEGG Pathway grouping and ranking function. In our approach, we utilized the KEGG pathway as grouping (term) information and inserted this information into a machine learning algorithm for selecting the most significant groups (terms) of KEGG. Those groups are utilized to perform the machine learning model for the classification task. In our experiments, we observed that the approach successfully obtained the essential KEGG terms that would be utilized as a classification model. This study was tested on 13 gene expression datasets including various diseases. A list of important KEGG pathways has been provided, including genes that can separate data classes. The biology researcher will use the list for in-depth analysis and better interpretability of the role of KEGG pathways in the data. We compare the performance of the approach to SVM-RCE, CogNet, maTE which is similar in their merit. The results indicate that we outperform maTE in most cases and it uses less gene than SVM-RCE-R and CogNet.

How Epstein-Barr virus envelope glycoprotein gp350 tricks the CR2? A molecular dynamics study.

Ilgaz Taştekil[1], Cansu Yay[2], Nursena Keskin[2], Elif Naz Bingöl[1] and Pemra Ozbek Sarıca[2]

[1]Department of Bioengineering, Institute of Pure and Applied Sciences, Marmara University

[2]Faculty of Engineering, Department of Bioengineering, Marmara University

Epstein-Barr virus (EBV) is a latent type of viruses that is related with severe diseases such as Burkitt Lymphoma, Hodgkin disease, multiple sclerosis, systemic lupus erythematosus (SLE) and different B-cell lymphomas. It infects human B-cells surface receptor type 2 protein CR2 through binding EBV envelope glycoprotein gp350, the most attractive protein for vaccine development studies against EBV infections. As a result, many mutational studies have been applied for the determination of crucial active site residues in binding of gp350 and CR2 especially for discovering inhibition mechanisms. Although active site residues are known experimentally, how gp350 and CR2 interacts structurally is not clearly understood. On the other hand, understanding the binding mechanism underlying gp350-CR2 complex formation is vital for effective vaccine development. Experimental studies are usually challenging for understanding such complex biological processes while computational approaches provide easier, faster and cost-effective discoveries of how protein-protein interactions occur.

Here, in this study, it is aimed to characterize the interactions between gp350 and its receptor CR2. Several mutations were performed on gp350 structures prior to molecular docking. Each mutated and wild type gp350 structures were docked with CR2 based on experimental active site data. Then, molecular dynamics simulations were performed on these docked complexes. Molecular dynamics simulations revealed the interesting relationship between the distant sites of gp350, including linker-2, domain 3 (D3) and CR2. Together with that, more structural information elucidating the binding mechanism between two molecules were obtained. This computational study can guide vaccine development strategies to regulate the gp350 activity and prevent the entrance of EBV to the B cells.

Metabolic network-driven analysis of yeast metabolic cycle through the incorporation of RNA-seq and ATAC-seq datasets

Müberra Fatma Cesur, Tunahan Çakır and Pınar Pir

Department of Bioengineering, Gebze Technical University, Kocaeli, Turkey

Saccharomyces cerevisiae, which is a well-established model organism in many industrial and medical applications, undergoes robust oscillations to regulate its physiology for adaptation and survival under nutrient-limited conditions. The rhythmic alterations in gene expression pattern and cell metabolism coordinate responding to environmental cues. Yeast metabolic cycle (YMC) is a remarkable example of the coordinated and dynamic yeast behaviour, which is regulated through metabolic oscillations. It is divided into three phases based on periodic alterations in gene expression across varying oxygen consumption levels: quiescence-related reductive charging (RC) phase, growth-related oxidative (OX) phase, and cell division-related reductive building (RB) phase [1]. Thus, YMC tracks the life cycle of yeast via an interplay among growth, proliferation, and quiescent phases.

Genome-scale metabolic network (GMN) models have been extensively used platforms to analyse yeast metabolism since 2003 [2]. They are stoichiometry-based mathematical representations of metabolism with all known chemical reactions, metabolites, and genes. GMN models provide a powerful platform for the systems-based understanding of metabolic processes within an organism. To date, different omics data integrated models have been developed for phenotypic characterizations and metabolic engineering. Despite the common use of transcriptome in the contextualization of GMN models, incorporation of epigenetic information is still a gap in the field of metabolic modelling. Besides, a clear interaction between metabolism and epigenetics was highlighted in many studies. Here, we investigated the contribution of combinatory use of transcriptomic and epigenomic information in the simulation of cellular metabolism via a recent yeast model, Yeast8 (3,991 reactions, 2,691 metabolites, and 1,147 genes) [3]. To this aim, we first employed hierarchical clustering for both RNA-seq and ATAC-seq datasets [4] dedicated to each YMC phase. Thus, the pathways associated with each YMC phase were identified (data-based approach). We subsequently reconstructed diverse GMN models through mapping these datasets [4] in both individual and combinatorial fashions. This facilitated the simulation of early RC, mid OX, and late RB phases. Thus, we evaluated the performance of each model using the experimental flux data derived from 13C metabolic flux analysis [5]. We also characterized differential flux profiles and pathways through comparative analyses (model-based approach). Lastly, we compared the results obtained via data- and model-based approaches to each other and validated based on the literature.

Comparative analysis of the predicted and measured fluxes revealed that the use of ATAC-seq data considerably improved model performances. The pathways dedicated to each YMC phase were elucidated using data- and model-based approaches. As expected, over-representation of the growth-related processes (e.g., biosynthesis of amino acids and nucleotides) and tricarboxylic acid cycle were shown in mid OX phase. On the other hand, early RC and late RB phases were found to ex-

hibit similar characteristics. Over-representation of various glycolytic processes, NADP metabolism, and pentose-phosphate shunt were determined in late RB phase in agreement with literature. To our knowledge, this is the first attempt to use chromatin accessibility data in the reconstruction of context-specific GMN models, despite the increasing popularity of ATAC-seq method. Thus, we demonstrated that integration of epigenomic data with transcriptomic profiling can pave the way for more realistic metabolic simulations.

References

1. Rao AR, Pellegrini M. Regulation of the yeast metabolic cycle by transcription factors with periodic activities. BMC Syst Biol. 2011;5(1):160.

2. Förster J, Famili I, Fu P, Palsson B, Nielsen J. Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res. 2003;13(2):244–53.

3. Lu H, Li F, Sánchez BJ, Zhu Z, Li G, Domenzain I, Marci S, Anton PM, Lappa D, Lieven C, Beber ME, Sonnenschein N, Kerkhoven EJ, Nielsen J. A consensus S. cerevisiae metabolic model Yeast8 and its ecosystem for comprehensively probing cellular metabolism. Nat Commun. 2019;10(1):3586.

4. Gowans GJ, Schep AN, Wong KM, King DA, Greenleaf WJ, Morrison AJ, Gowans GJ, Schep AN, Wong KM, King DA, Greenleaf WJ, Morrison AJ. INO80 Chromatin Remodeling Coordinates Metabolic Homeostasis with Cell Division. Cell Rep. 2018;22(3):611–23.

5. Zhang J, Martinez K, Elmar G, Sebastian H, Wahl A, Malate M. Metabolic switches from quiescence to growth in synchronized Saccharomyces cerevisiae. Metabolomics. 2019;15(9):1–13.

Peptide - gold (111) interactions: mechanisms and design

Didem Ozkaya[1,2], Busra Demir[1,2], Caglanaz Akin[1,2], Zeynep Koker[2] and Ersin Emre Oren[1,2]

[1]Department of Materials Science & Nanotechnology Engineering, TOBB University of Economics and Technology, Ankara, Turkey.

[2]Bionanodesign Laboratory, Department of Biomedical Engineering, TOBB University of Economics and Technology, Ankara, Turkey.

Proteins and peptides are known to have a role in mineral nucleation, growth, and structure creation, as well as providing molecular scaffolding for the formation of hard tissues such as bone and dental tissues [1,2]. Using biopanning protocols, it has been shown that one can generate peptides for inorganic material synthesis, formation, and assembly [2]. A deeper understanding of the interaction between peptide and inorganic surfaces, also binding affinities or specificities may aid in the development of novel peptides with desired features in engineering and medicine [3,4]. Gold is a highly conductive inorganic and amenable to surface modification. When gold is functionalized with peptides it has practical applications such as bioimaging, biosensors, cancer therapy and drug delivery, [5-7]. In this research, we use atomistic methods to understand how individual amino acids and small peptides interact with Au (111) surface that will allow us to design novel moieties for gold functionalization. Peptides which are used in this research are chosen systematically by induction starting from the best binding amino acids among 20, up to the chain of 7 amino acid peptides. Structures of chosen peptides were optimized on the gold surface by conformational search. Then, best conformation(s) used as starting structure for molecular dynamics (MD) simulations. Peptide stability and binding affinities were analyzed with the MD trajectories. Also, binding free energy calculations were done by using MM-GBSA method [8].

References

1. Mann, S. (1988). Molecular recognition in biomineralization. Nature, 332(6160), 119–124.

2. Paine, M. L., & Snead, M. L. (1997). Protein Interactions During Assembly of the Enamel Organic Extracellular Matrix. Journal of Bone and Mineral Research, 12(2), 221–227.

3. Hnilova, M. (2008) Effect of molecular conformations on the adsorption behavior of gold-binding peptides, Langmuir, 24, 12440-12445.

4. Oren, E.E. (2007) A novel knowledge-based approach to design inorganic-binding peptides, Bioinformatics, 23, 2816-2822.

5. Chen, S., Liu, L., Zhou, J., & Jiang, S. (2003). Controlling Antibody Orientation on Charged Self-Assembled Monolayers. Langmuir, 19(7), 2859–2864.

6. Park, T. J., Lee, S. Y., Lee, S. J., Park, J. P., Yang, K. S., Lee, K.-B., ... Choi, I. S. (2006). Protein Nanopatterns and Biosensors Using Gold Binding Polypeptide as a Fusion Partner. Analytical Chemistry, 78(20), 7197–7205.

7. Sarikaya, M., Tamerler, C., Schwartz, D. T., & Baneyx, F. (2004). Materials Assembly and formation using engineered polypeptides, Annual Review of Materials Research, 34(1), 373–408.

8. Tsui, V. (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations, Biopolymers, 56, 275-29.

The mutation profile of SARS-CoV-2 is primarily shaped by the host antiviral defense

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

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

Tracing the evolution of novel coronavirus (SARS-CoV-2) is a crucial task for coping with the COVID-19 pandemic. The evolution of the virus is driven by mutagenesis and the selection mechanisms, that are the major factors of genetic variation across the virus isolates. Distinct mutational patterns can be embedded in the genetic contents of the isolates. Mutational patterns and mechanisms that are contributing to these patterns should be uncovered to understand the evolution of SARS-CoV-2. Here, we show the contribution of potential biological mechanisms to the mutational profile of SARS-CoV-2. Using a group of representative genomes, we generated a phylogenetic tree, and identified the independent mutations based on the mutational events at each node of the tree. This method allowed us to retrieve independent mutational events, that cannot be captured through conventional methods such as pairwise sequence comparison relative to the reference genome. As a result, we found that the heterogeneous mutation patterns of SARS-CoV-2 genomes reflect the RNA editing activity of the host antiviral mechanisms named APOBEC (Apolipoprotein B mRNA-editing enzyme, catalytic polypeptide), ADAR (Adenosine deaminases acting on RNA), and ZAP (Zinc-finger antiviral protein). There is also a probable adaptation against the ROS (Reactive oxygen species). These results suggest that the primary contribution to the genomic diversity of SARS-CoV-2 genomes originates from the mutation events triggered by the antiviral defense agents of the host.

Application of machine learning algorithm for the accurate diagnosis of breast cancer

Rumeysa Fayetörbay and Uğur Sezerman

Department of Biostatistics and Bioinformatics, Graduate School of Health Sciences, Acıbadem MAA University, İstanbul, Turkey

With 2.26 million new cancer cases in 2020, breast cancer is the most common cancer type [1]. Globally, it is the most frequently observed malignancy for females, corresponding to the approximately 1 in 4 cases among women [2]. Tumorigenic samples were taken from breast mass by applying the fine needle aspiration (FNA) biopsy technique [3]. 10 distinct features, relevant to the diagnostic accuracy, which are evaluated from FNA's digitized image, are area, compactness, concave points, concavity, fractal dimension, perimeter, radius, smoothness, symmetry and texture [3]. These features refer to the characteristics of the cell nuclei illustrated in the digital images [3]. Each feature has 3 subcategorical information about the 'mean', 'standard error' and 'worst' of the images [3]. In total, there are 30 different features calculated in the dataset which is also available at the UCI's machine learning repository.

Machine learning algorithms have been performed for feature extraction, classification and clustering approaches. To prevent the overfitting in the logistic regression models, the least absolute shrinkage and selection operator (LASSO), penalized regression method, which is mainly used for variable selection and regularization, is an alternative to apply. Lasso regression performs L1 regularization which basically parameterizes the shrinkage of estimates and penalizes certain regression coefficients with zero weight unless they are significant in order to enhance the prediction accuracy [4].

In this study, our main purpose is to reveal the essential characteristics of the breast mass which directly indicate the tumorigenicity (benign/malignant) of the breast cells. For this purpose, we partitioned our data into the training and the test sets. To further implement, we assigned binary values for determining the level of the oncogenicity. After making predictions on the test data, we checked the performance of our model with lasso regression, our model was predicted the response variable with 97.2% accuracy. Out of 30 characteristic features, there were 13 significant coefficients for variable selection which were concave points mean, symmetry mean, compactness standard error, fractal dimension standard error, radius standard error, smoothness standard error, texture standard error, concave points worst, concavity worst, radius worst, smoothness worst, symmetry worst and texture worst. Since other 17 estimated coefficients were not significant, they were shrunk to zero by our model. The sensitivity and the specificity of our model were approximately 96% and 98%, respectively. To evaluate the performance of our lasso regression model, ROC curve plot was drawn; we were able to distinguish between the groups and obtained a clear separation which was close to the ideal. The results suggested the application of such machine learning algorithms had a high potential to determine the features that are relevant to the diagnostic accuracy of the breast cancer.

References

1. Sung, Hyuna et al. “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries.” CA: a cancer journal for clinicians vol. 71,3 (2021): 209-249.

2. Momozawa, Yukihide et al. “Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls.” Nature communications vol. 9,1 4083. 4 Oct. 2018.

3. Wolberg, W H, and O L Mangasarian. “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.” Proceedings of the National Academy of Sciences of the United States of America vol. 87,23 (1990): 9193-6. doi:10.1073/pnas.87.23.9193.

4. Veen, Kevin M et al. “A clinician’s guide for developing a prediction model: a case study using real-world data of patients with castration-resistant prostate cancer.” Journal of cancer research and clinical oncology vol. 146,8 (2020): 2067-2075.

A time-efficient and user-friendly tool for molecular dynamics analysis

Halil İbrahim Özdemir, Elif Naz Bingöl and Pemra Özbek Sarıca
Department of Bioengineering, Marmara University, Goztepe, Istanbul, Turkey

Molecular dynamics (MD) simulations of biological molecules have been developing for about 45 years and have become more powerful and accessible every year (McCammon, Gelin & Karplus, 1977). Extracting the useful information from simulations requires interpretation in light of all available experimental data for the molecular system under study. The need for the development of analysis methods and tools has arose as a result of the popularization of MD simulations. (Hospital, Battistini, Soliva, Gelpí, & Orozco, 2019).

Analyzing MD simulation results can be difficult for several reasons. A classic simulation can track the positions and velocities of 100,000 atoms over billions of time steps. It is difficult to make sense of the biologically important findings of these data. The analysis process requires a careful combination of numerical values calculated via existing software as well as visual analysis using quantitative analysis. Existing software can be used for some common analyzes as well as making extensive use of specialized analysis programs or scripts. Although software libraries for MD data processing and analysis are developed by academic groups, this creates a barrier between software packages and non-developer researchers. On the other hand, these software packages are rather cumbersome for routine data processing. Therefore, there is a high demand for a platform which would make MD analysis simpler, more efficient, and suitable for routine use by experts and non-specialists alike.

In response to this demand, MolDynAnalyze is offered, an easy-to-use and extensible Graphical User Interface (GUI) program that simplifies complex operations in data analysis by combining powerful MD analysis libraries and software packages that can present them visually.

References

1. McCammon, J., Gelin, B. & Karplus, M. Dynamics of folded proteins. Nature 267, 585–590 (1977). https://doi.org/10.1038/267585a0

2. Hospital, A, Battistini, F, Soliva, R, Gelpí, JL, Orozco, M. Surviving the deluge of biosimulation data. WIREs Comput Mol Sci. 2020; 10:e1449. https://doi.org/10.1002/wcms.1449

PhosProViz: A web-based tool to generate and interactively explore phosphoproteomics networks

Irene Font Peradejordi[1], Shreya Chandrasekar[1], Berk Turhan[2],Selim Kalayci[3], Jeffrey Johnson[3] and Zeynep H. Gümüş[3]

[1]Cornell Tech, Cornell University
[2]Sabanci University
[3]Icahn School of Medicine at Mount Sinai

To better understand the human cellular systems and how diseases influence cells and protein expression, researchers are profiling human cells before and after infection, vaccination, or treatment, collecting massive -omics data sets. Some of these studies involve protein phosphorylation interactions and perturbations. Protein phosphorylation is a reversible post-translational modification that involves the phosphorylation of a substrate protein residue by a kinase, which is a vital modification for cellular processes and signaling networks. Recent advances in mass spectrometry based quantitative proteomics have led to the rapid generation of massive protein phosphorylation datasets at multiple states. These are typically visualized as networks, with kinases and substrates represented as nodes, and their interactions as edges. It is also important to visualize the states of all phosphorylation sites altered within each substrate. Currently available visualization tools are not optimized for large and complex phosphoproteomics networks with their associated phosphorylation data. Furthermore, there are currently no dedicated tools to generate, explore and share interactive visualizations of these systems. There is thus a need for a tool that facilitates user-intuitive and interactive exploration, visualization, and communication of phosphoproteomics datasets.

Here, we present PhosProViz, a tool that empowers users to easily generate shareable interactive 3D phosphoproteomics network visualizations with multiple phosphorylation sites and states, by simply uploading their data as a comma separated file. PhosProViz is an integrated platform to explore human phosphoproteomics data sets in multiple cell types or cohorts, in the context of their rich phosphorylation information. The landing page of PhosProViz provides the main interface for uploading user datasets. Users can either directly upload their datasets themselves; or they can get help for formatting their data by answering several questions on edge directionality; source color and size attributes; target color and size attributes; phosphorylation sites; and edge color and weight attributes. After answering the questions, users can inspect their data in PhosProViz input format. A dialog window also includes guidelines to properly format and upload datasets. After successful completion of data upload, users can simultaneously explore multiple network datasets that belong to different states (time point, treatment, other) or query and explore in detail their protein of interest and its interactions. Furthermore, PhosProViz has several user-intuitive features. These include: i) enabling users to visualize all relevant data simultaneously (e.g., nodes, directed edges, and their userdefined attributes), which include proteins, their connections and phosphorylation sites; ii) 3D interactivity, including pan, zoom, rotate, and drag nodes; iii) threshold selection, where users can filter their network by applying different thresholds (e.g., log fold change values); iv) annotations, where each node contains a link

that redirects to annotation-specific information to outside websites; v) modularity which enables users to build more specialized interfaces with specific requirements; and vi) easy accessibility through a standard web browser making it very fast and easy to use; vi) enabling cross-states or time-points comparisons; and vii) detailed FAQ and tutorials that guide users at every step of interaction with the tool. After visualization, users can manually adjust their network view and take snapshots. Note that user datasets are not transferred to any remote server or third-party site and all operations are handled completely within the local browser of the user with multiple client-side Javascript libraries (e.g., Three.js and 3d-force-graph) for visualization and user interaction in real time.

PhosProViz will facilitate researchers from a wide spectrum of computational skill levels to conduct their own analyses and share their results. Furthermore, its flexibility also allows its utilization for other applications that involve network data even without any phosphorylation site info (i.e., biomolecular interaction networks), expanding its use to a wide variety of research questions and investigators. Code is open source on GitHub, and the tool itself is also publicly available with hosting on GitHub Pages which can be accessed from: irenefp.github.io/PhosProViz-temp/

Metatranscriptome analysis of human gut microbiome by ASAIM workflow

Ceyda Demirtaş and Seda Koldaş
Acıbadem University

Microbiomes play a critical role in human health in terms of understanding various disease relations with microorganismal functional features. Metatranscriptomic analysis is being used to define and study the molecular activities and taxonomy of microbial communities using Next-generation sequencing (NGS) platforms. We performed a Metatranscriptome analysis on pre-runed RNA-seq data derived from a collected stool from a research of Harvard School of Public Health (Run Accession code: SRR6038203) and data was accessed via NCBI. Our aim is to understand the microbiome's response to its environment by investigating human gut microbiota to determine the most abundant gene families as well as molecular function abundances while obtaining taxonomic profiles. We used ASaiM which is a Galaxy based workflow to do metatranscriptomic analysis on Galaxy Europe /Metagenomics. Our results rely on the raw data which was pre-filtered in terms of human reads and low-quality and FASTQ data was accessed from ENA Browser. During the ASAIM workflow, the quality control of data was checked with the FASTQC tool. Then, the Trimmomatic tool was used to trim bad quality reads and remove too short reads. In order to make the downstream functional annotation faster, we sorted rRNA sequences with the SortMeRNA tool which removes rRNA reads. In order to understand the community structure, we did taxonomic profiling and we used cDNA converted rRNAs for being good marker genes. We visualized our community structure results by GraPhlAn for the indication of phylogenetic trees and Taxonomy. Taxonomic analysis was performed by MetaPhIAn2 tool while HUMAnN2 tool was used for identification of genes, their transcripts functions, and build pathways to examine their contribution to the microbial community. Moreover, we combined HUMAnN2 and MetaPhIAn2 results to relate gene family abundances and functional pathway abundances to species/genus found on the microbiota. Here, we used UniRef50 gene families. According to our MetaPhlAn2 results, we found as abundances 67.9% viruses, 27.3% bacteria, 3.1% Achaea, Eukaryote 1.3%, viroid 0.26%. Approximately 22% of reads assigned to a gene family have not been assigned to a pathway i.e. these were UNINTEGRATED data. But we found the most abundant normalized pathways; glycolysis IV (plant cytosol) that Bacteroides pectinophilus and Prevotella copri are involved in production. We detected largely unmapped reads which covered approximately 76% of our reads that are not assigned to gene families but most abundant gene families were Gemmata, Faecalibacterium, Corynebacterium respectively. Our goal is to identify the role of the microbiome in disease phenotype and treatment. We also want to investigate whether microorganisms play a role in the resistance that is encountered in the treatment response in some systemic diseases. In further analysis, our aim is to investigate microbial metabolite exchange and its impacts on the functions.

mirDisNet: A novel approach for cancer classification using mir-disease associations

Amhar Jabeer[1], Burcu Bakir-Gungor[1] and Malik Yousef[2]

[1]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University, Kayseri, Turkey

[2]Department of Information Systems, Zefat Academic College, Zefat, 13206, Israel

miRNAs (microRNAs) are a family of short non-coding RNAs that regulate gene expression post-transcriptionally in diverse species. They repress protein production by translational silencing, binding to the 3'-UTR (untranslated region) of their target mRNAs, and destabilizing them. Growing evidence shows that miRNAs exhibit a variety of crucial regulatory functions in all mammals; and because of their potential role related to cell growth, development, and differentiation, while being associated with a wide variety of human diseases. They have been proposed to be good candidates for cancer therapy since they have been associated with cancer biology: metastasis, angiogenesis, and proliferation. Conversely, considering the inherent time-consuming and expensive method of traditional in vitro experiments, the need for feasible and efficient computational methods to predict miRNA and diseases association have become apparent. We propose mirDisNet, a novel approach that detects the biomarker of miRNAs genes that is associated with diseases. In mirDisNet, biological domain data that incorporates the knowledge about miRNAs association with disease is used to serve as the grouping function for the tool. Each of the groups created have a disease name with a corresponding set of miRNAs related to the disease. We ranked the groups by scoring them on their importance in the two-class classification task. By integrating miRNA-disease associations, mirDisNet showed promising results of 95% in accuracy, 92% in sensitivity, 96% in specificity, and 98% in AUC across 11 datasets obtained from TCGA. Additionally, the most significant miRNAs and disease groups which were ranked by robustrankaggreg, were validated by external datasets, databases, and through literature. We hypothesize that mirDisNet has the prospect to understand disease prognosis as well as diagnosis by finding potential biomarkers and disease relationship networks.

Controversy detection on health-related tweets

Emine Ela Küçük[1], Selçuk Takır[2] and Dilek Küçük[3]

[1]Giresun University, Faculty of Health Sciences, Department of Nursing, Giresun, Turkey
[2]Giresun University, Faculty of Medicine, Department of Pharmacology, Giresun, Turkey
[3]TÜBİTAK, Marmara Research Center, Energy Institute, Ankara, Turkey

Social media analysis is a recent and active research area. Many researchers and practitioners of natural language processing and information retrieval are conducting research on social media posts for different text processing purposes to extract useful information. One relevant topic is social media analysis for the purposes of health surveillance, where social media posts (mostly tweets) are analyzed in order to extract health-related information.

Controversy detection is an important research problem within the context of Web mining and social media analysis [1-7]. Controversial posts are defined as those pieces of published texts which receive both positive and negative feedback from people [1]. Sample controversial topics include climate change, evolution, and political debates as in the case of discussions regarding referendums and elections [5]. It is reported in the related literature that controversy detection is an important stage towards uncovering the public sentiment and influence assessment regarding events or issues that have impact on society [1, 3]. Using Wikipedia editing histories together with the related controversy scores is one of the state-of-the-art methods for automatic controversy detection [4].

Within the context of health informatics, controversial discussions especially on social media include complementary and alternative medicine (CAM) [8], vaccination [5], and the use of other medicine such as hydroxychloroquine for the treatment of Covid-19.

Automatic and continuous detection of controversial discussions on Twitter is important on health-related topics due to a number of reasons. First of all, detecting controversial topics and presenting those topics conveniently is important for information retrieval systems [1-8], including topics related to aforementioned public health concerns. Thereby, search engine results about health topics could be presented to related users (and to general public use) more conveniently. Secondly, public health surveillance systems can detect and track the evolution of health-related controversial topics by introducing controversy detection into their processing pipelines. And additionally, sentiment and stance of the community towards these controversial and health-related discussions can be automatically detected [9].

To sum up, in this study we will discuss the contribution of controversy detection to the topic of health-related information detection, retrieval, and monitoring on Twitter. Controversy detection is an important and recent research topic in social media analysis and similarly, public health monitoring on social media is again a significant and practical research area in health informatics. We believe that high-performance and automatic controversy detection schemes can be used to improve the performance of public health surveillance systems.

References

1. Hessel, J., & Lee, L. (2019). Something's Brewing! Early Prediction of Controversy-causing Posts from Discussion Features. In Proceedings of An-

nual Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT) (pp. 1648-1659).

2. Popescu, A. M., & Pennacchiotti, M. (2010). Detecting Controversial Events from Twitter. In proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM) (pp. 1873-1876).

3. Zhong, L., Cao, J., Sheng, Q., Guo, J., & Wang, Z. (2020). Integrating Semantic and Structural Information with Graph Convolutional Network for Controversy Detection. arXiv preprint arXiv:2005.07886.

4. Jang, M., & Allan, J. (2016). Improving Automated Controversy Detection on the Web. In Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (pp. 865-868).

5. Jang, M., Dori-Hacohen, S., & Allan, J. (2017). Modeling Controversy within Populations. In Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (pp. 141-149).

6. Jang, M., Foley, J., Dori-Hacohen, S., & Allan, J. (2016). Probabilistic Approaches to Controversy Detection. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management (CIKM) (pp. 2069-2072).

7. Garimella, K., Morales, G. D. F., Gionis, A., & Mathioudakis, M. (2018). Quantifying Controversy on Social Media. ACM Transactions on Social Computing, 1(1), 1-27.

8. Zhang, S., Qiu, L., Chen, F., Zhang, W., Yu, Y., & Elhadad, N. (2017l). We Make Choices We Think Are Going to Save Us: Debate and Stance Identification for Online Breast Cancer CAM Discussions. In Proceedings of the 26th International Conference on World Wide Web Companion (WWW) (pp. 1073-1081).

9. Küçük, D., & Can, F. (2020). Stance Detection: A Survey. ACM Computing Surveys (CSUR), 53(1), 1-37.

Application of machine learning for the identification of novel diagnostic biomarkers for COVID-19 by using transcriptomic data

Didem Okmen[1], Athanasia Pavlopoulou[1] and Eralp Dogu[2]

[1]Izmir Biomedicine and Genome Center, Izmir, Turkey 35340

[2]Department of Statistics, Muğla Sıtkı Koçman University, Mugla, Turkey, 48000

COVID-19, caused by the Severe Acute Respiratory Syndrome (SARS-CoV-2) virus is an ongoing epidemic all over the world with a high death and mutation rate and remains a global priority to develop effective treatments. Therefore, elucidating the molecular mechanisms underlying the human host response to lung coronavirus infection is essential for the diagnosis and monitoring of patients with COVID-19, as well as for the rational design of effective anti-coronaviral therapeutic strategies. At the onset of the pandemic, data on gene expression related to COVID-19 were scarce. Over the past year, high-throughput (HTP) "omics" technologies, such as microarrays and RNA-Seq, have enabled the generation of data on human host transcriptome response following SARS-CoV-2 infection. Today, healthcare professionals and scientists are intensively researching new computational and machine learning technologies and methodologies for efficient processing of COVID-19-related data, and accordingly, the amount of gene expression related to COVID-19 is gradually increasing. In order to contribute to these studies, the main goal of our study is to investigate the mechanisms/pathways that mediate the transcriptional response of human SARS-CoV-2 infected lungs. To this end, we obtained HTP gene expression (transcriptome) data from publicly accessible databases. The first datasets we obtained included both mock-treated and SARS-CoV-2-infected pluripotent stem cells, as well as transcriptomic profiles of lung tissues from deceased COVID-19 patients. We designed an interactive workflow to extract, process and analyze biologically meaningful information. In addition to generalized linear models, we also applied bootstrap analysis, one of the statistical methods specifically designed to analyze small samples, since our datasets contain relatively few unequal samples per group. However, the traditional methods (filtering, clustering, etc.) we have used in the preprocessing procedures in the datasets we have obtained so far, in determining the gene expression data, in both COVID-19 infected and uninfected normal lung cells and tissues, the pathological structure of the disease, age and environmental factors are to cause inconsistency, increased variation and complexity due to the presence of other diseases. The inaccessibility of epithelial lung cells of COVID-19 patients, particularly in the early stages of the disease, points to the need to develop human model systems, such as in vitro lung cell-based models, that can adequately recapitulate the pathogenesis of COVID-19; these models also enable global testing of antiviral drugs. For this reason, it is still important to compare the effects of SARS-CoV-2 to those of other respiratory tract viruses, together with the biological processes occurring in the early stages of the infection, in terms of effective treatment methods and drug development. In clinical setting, machine learning methods such as decision trees, regression, clustering, neural networks etc., have been successfully applied to complex "omics" data for the discovery of powerful biomarker genes for better detection, prognosis and management of various diseases. Motivated by this,

we focus on comparing diagnostic gene signatures derived from lung tissue samples and lung cells using hybrid models for efficient filtering and feature selection, in addition to machine learning techniques such as regression-based methods. This would enable us to determine to what extent the results derived from the lung cell lines and lung tissues investigated in this study will reflect the genomic complexity of lung tissues infected with SARS-CoV-2. The results of this comparison may be important for the rational design of cell-based in vitro models for SARS-CoV-2 studies to identify molecular markers that can be used in the biomedical and clinical setting and target to develop widely applicable treatments.

MicroBiomeNet: Machine learning analysis of metagenomics datasets: Colon cancer dataset

Malik Yousef[1], Anas Nadifi[2], Amhar Jabeer[2] and Burcu Bakır[2]

[1]Galilee Digital Health Research Center (GDH), Zefat Academic College
[2]Department of Computer Engineering, Faculty of Engineering, Abdullah Gul University

Before the accomplishment of the Human Genome Project in April 2003, multiple researchers foresaw that around 100,000 genes would be found. To the surprise of many, the actual results consisted of only around 20,000 rotein-coding genes contained in the human genome. The living microbes or as generally referred to by the term 'microorganisms', which live both on and inside humans in an environment referred to by the scientific community as the 'microbiota', are roughly estimated to surpass the number of somatic and germ cells by 10 folds. Together, the genomes of these microbial symbionts (collectively defined as the microbiome) provide traits that humans did not need to evolve on their own, the gut microbiome specifically has attracted considerable scientific attention due to its potential role in disease susceptibility of a given host.

As a consequence of the completion of the human genome project, more emphasis is being directed towards studying the human gut microbiome by scientists, it is a known scientific fact that the gut microbiota has a major role in humans' health, in terms of affecting the host's immune, metabolic and endocrine functions, as a consequence to this fact, any dysbiosis (gut microbiota imbalance) is directly linked to have an important role in notably extra intestinal disorders such as cardiovascular diseases, type 1 diabetes, obesity, metabolic syndrome, asthma and allergies, as well as intestinal disorders such as IBD (inflammation or destruction of the bowel wall), IBS (irritable bowel syndrome), Hemorrhoids, Constipation, stenosis, diverticular disease and colon cancer. The recent advances in the NGS technologies (next generation sequencing) enables further and deeper understanding of the gut microbiota, by unraveling information on microbiota composition in the human gastrointestinal tract which is generally achieved by analysing human faeces. By combining the data gathered from NGS analysis and the state of the art machine learning algorithms, scientists can now unravel previously undiscovered biological relations between specific gut microbiome species and specific diseases. Multiple studies attempted successfully to establish an association between certain gut microbiome species and mainstream diseases such as Irritable bowel syndrome. This study will focus solely on Colon Cancer by meticulously depicting which microbiome species are directly related to the disease.

Colon cancer, known also as Colorectal cancer, bowel cancer and rectal cancer, is a type of cancer specifically targeting the colon and rectal area of the human's organs, risk factors contributing to the possibility of colorectal cancer development include but are not limited to old age, lifestyle habits, diet (mainly consumption of red meat, processed meat and alcohol) , obesity, smoking and lack of physical activity, in addition to pre-existing disease such as inflammatory bowel disease, Crohn's disease and ulcerative colitis. Colon cancer is a well known disease responsible for the deaths of around 1 million every year, and the fatality rate related to colon cancer is increasing, as it went from 490,000 recorded deaths in 1990 to 715,000 in 2010

and is expected to increase even further due to the risk factors being widely ignored by the world's population. The popularity of the disease is evident by the fact that in the United States the month of March is colon cancer awareness month. As like any type of cancer, early diagnosis and treatment is vital for a higher chance of survival. In this study we aim to pave the way for early diagnosis and eventually precision treatment of the disease, which is done by exploring Colon cancer related BioMarkers by the use of classification on our dataset with feature selection. Our sequencing dataset consists of 48 colon cancer patients and 60 healthy individuals, the data was taken from a metagenome-wide association study and categorised into disease states based on the associated metadata. After associating the reads to taxa, we identified 1455 species and thus used them as input to our classification model, with test and train sets. In order to obtain optimal results, we applied grouping and ranking of the features, the features were grouped based on the taxonomic information of the microbiome species, we explored grouping and ranking features by 3 taxa levels (order, genus and family), we then applied classification based on the highest ranked features in each group retrieved.

Concisely, this study makes use of state of the art machine learning algorithms, through supervised and unsupervised learning, to accurately diagnose Colorectal cancer (CRC), by pinpointing exactly which gut microbiota species is mostly related to the disease, using the taxa information available.

Protein sequence diversity dynamics of primate Erythroparvovirus 1

Pendy Tok[1], Li Chuin Chong[2] and Mohammad Asif Khan[2,3]

[1]Faculty of Information Science and Technology, Multimedia University, 75450 Bukit Beruang, Melaka, Malaysia

[2]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Beykoz 34820, Turkey

[3]Centre for Bioinformatics, School of Data Sciences, Perdana University, Damansara Heights, Kuala Lumpur 50490, Malaysia

Background: Primate erythroparvovirus 1, generally known as parvovirus B19 (B19V), is a human viral pathogen of the genus Erythrovirus from the family Parvoviridae. The virus commonly infects children, causing erythema infectiosum. Studies reporting on the sequence diversity of the virus are limited, in particular utilizing all the available sequences in public repositories. Herein, we report the proteome sequence diversity dynamics of B19V.

Materials and Methods: All reported protein sequences of the virus, either full or partial length, isolated from the human host were retrieved from the public repositories, NCBI Entrez Protein and Virus databases, by use of the taxonomy identifier, "1511900". A total of 6,562 protein sequences (as of January 2021) were retrieved, deduplicated (CD-HIT), then processed into separate protein datasets (BLASTp), and each aligned (Clustal Omega). Only three (VPs: VP1 and VP2; and NS1) of the six proteins had sufficient sample size (>30) for further analyses. Shannon's entropy calculations and quantitative pattern analysis of sequence diversity motifs (index and its (total) variants: major, minor and unique) were carried out for each of the overlapping nonamer positions of the proteins. The sliding window size (k-mer) of nine was chosen for immunological applications. The completely and highly conserved index nonamers (incidence $\geq$ 90%) were concatenated to form a longer sequence if they overlapped by at least one amino acid or were adjacent to each other.

Results: The B19 proteome was observed to exhibit a low mean entropy of $\sim$ 0.5 cross the proteins studied, with only nine positions exhibiting total variants close to 58%, and thus indicating a high conservation, generally. Distribution of index, the sequence with the highest incidence, and its variant motifs (major, minor and unique variants) across the nonamer positions of each protein illustrated a distinctive pattern of sequence change dynamics. The incidence of the index nonamer ranged from 100% (completely conserved) to $\sim$ 42%, corresponding to the high conservation observed, with $\sim$ 10% of the nonamer positions being completely conserved. The major variant, the second most predominant sequence at a given nonamer position, exhibited a rough pyramidal pattern, with a peak incidence of $\sim$ 43%. Minor variants, the nonamer sequences that were observed more than once but of lesser frequency than the major variant, were strikingly absent in $\sim$ 77% of the nonamer positions. Thus, most of the protein nonamer positions lacked minor variants. Notably, unique variants, which are singleton nonamer sequences, were observed for nearly all the protein nonamer positions. The variants of the index originated mostly from $\sim$ 26% of the protein positions, with total variants of $\geq$ 10%. The

index sequence of the 156 completely conserved and 899 highly conserved nonamer positions were concatenated into 30 distinct sequences.

Conclusion: This study provided insight to the Primate erythroparvovirus 1 protein sequence diversity. The 30 concatenated sequences of high conservation are candidates for further studies relevant to vaccine target discovery.

Protein sequence diversity of human respiratory syncytial virus

Faruk Üstünel[1] and Asif M. Khan[2,3]

[1]Department of Biotechnology, Institute of Health Sciences, Bezmialem Vakif University, 34093 Fatih, Istanbul, Turkey

[2]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Beykoz 34820, Turkey

[3]Centre for Bioinformatics, School of Data Sciences, Perdana University, Damansara Heights, Kuala Lumpur 50490, Malaysia

Background: Human respiratory syncytial virus (hRSV) is one of the most contagious viruses worldwide, which can lead to lower respiratory tract infections, such as bronchiolitis and pneumonia. There is no licensed prophylactic RSV vaccine. The virus can be divided into two groups, A and B, although there is only one serotype. A large number of hRSV sequences are available in public databases, which serve as a treasure trove for a comprehensive sequence diversity study. Herein, we describe a large-scale proteome-wide sequence diversity analysis of hRSV.

Methods: This study focussed on all available protein sequences of the virus, covering both groups A and B. Sequence data was collected (as of 31 March 2021) from publicly available NCBI Virus and Virus Pathogen Database and Analysis Resource (VIPR) databases. Protein sequence data of both groups were pooled, and duplicates were removed by use of CD-HIT. BLASTp search was performed on the collected sequences using UniProt reference records of the 11 hRSV proteins to generate individual protein datasets. Each protein dataset was multiple sequence aligned using MAFFT. Sequence diversity was measured by use of Shannon's entropy for each overlapping 9-mer (nonamer) (1–9, 2–10, etc.) position across the length of the protein.

Results: The total number of sequences collected from VIPR and NCBI was 93,382. After the removal of duplicates, the number decreased to 12,413 (an $\sim$ 87% reduction). The peak entropy value observed for hRSV was $\sim$4.2 in the M2-2 protein at starting nonamer position 43. However, the proteome-wide average entropy was low ($\sim$0.8) and thus, indicating high conservation, with values ranging from $\sim$0.5 (protein N) to $\sim$2.3 (protein M2-2). Most of the variants ($\sim$55%) were between the entropy range of 0 to 1, with the remaining distribution as follows: $\sim$36% (entropy range 1-2), $\sim$6% (entropy range 2-3), $\sim$2% (entropy range 3-4), and $\sim$0.2% (entropy range 4-5). Approximately half ($\sim$51%) of the nonamer positions were classified as highly conserved (index incidence $\geq$ 90%), with the rest ($\sim$49%) as mixed-variable (index incidence $<$ 90% & $\geq$ 20%); highly diverse (index incidence $<$ 20%) positions were not observed.

Discussion: This study provides information about sequence diversity across the proteome of hRSV. The high conservation of the proteome merits further investigation for vaccine target discovery.

Prediction of regulatory network interactions with CNN model using human RNA-Seq data

Gülce Çelen and Alper Yılmaz

Department of Bioengineering, Yildiz Technical University, Istanbul, Turkey

Gene regulatory networks (GRNs) are useful models to elucidate complex relationships between transcription factors (TFs) and their targets. Identifying GRNs for every TF-target pair using experimental approaches is infeasible. Therefore, many computational approaches have been developed to infer GRNs from different types of biological data. However, uncovering regulatory interactions from gene expression data more accurately is still a challenge in computational biology. A recent study involving Arabidopsis thaliana microarray data which used Convolution Neural Network approach successfully predicted TF-target pairs [1]. In this study we applied similar approach to human RNA-Seq data.

Human TF-target interactions were retrieved from TRRUST database [2] and human RNA-Seq data for tumor and normal samples were retrieved from UCSC Toil [3] which contains data from TCGA [4] and GTEx [5] projects, respectively. For each TF-target pair, expression values were extracted to assemble training data (Figure 1). Two separate models were generated for normal and tumor samples.

As a result, two separate models were developed to predict TF-target pairs in normal and tumor samples. The models demonstrated high accuracy. We also identified pairs which were predicted as pair in normal samples but not in tumor samples or vice versa. Our approach has potential to extend the existing and ever growing human GRN. Also, separate TF-target predictions for normal and tumor samples may suggest TF-target interactions presence or absence of which has role in tumor mechanisms.

References

1. MacLean, D. "A Convolutional Neural Network for Predicting Transcriptional Regulators of Genes in Arabidopsis Transcriptome Data Reveals Classification Based on Positive Regulatory Interactions". bioRxiv. 2019. 618926. doi: 10.1101/618926

2. Han H, Cho JW, Lee S, et al. "TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions". Nucleic Acids Research, 2017. 46(D1), D380-D386 doi: 10.1093/nar/gkx1013

3. Vivian J, Rao AA, Nothaft FA, et al. "Toil enables reproducible, open source, big biomedical data analyses". Nat Biotechnol. 2017 Apr 11;35(4):314-316. doi: 10.1038/nbt.3772

4. https://www.cancer.gov/tcga

5. Ardlie KG, Deluca DS, Segre AV, et al. "The genotype-tissue expression (GTEx) pilot analysis: Multitissue gene regulation in humans". Science, 2015, 348 (6235), 648–660. doi: 10.1126/science.1262110.

Integrating multi-omics data and deep learning for discovering new subtypes of breast cancer

Huseyin Uyar and Ozgur Gumus
Department of Computer Engineering, Ege University, Izmir, Turkey

Molecular subtypes of breast cancer are mainly luminal A, luminal B, HER2-enriched, basal and normal-like [1]. Identification of these subtypes is of great importance to select the most appropriate treatment for the patient. Hence, discovering new subtypes opens the door for new and more precise treatment strategies and allows us to give the most suitable treatments to patients. In this preliminary study, we used the breast cancer (BRCA) project data of The Cancer Genome Atlas [2] program, with the aim of discovering new subtypes of breast cancer. Dataset consists of multi-omics, such as SNP, CNV, mRNA, miRNA and methylation data of breast cancer patients to capture different aspects of carcinogenesis. We filtered the data to obtain only the patients who have molecular subtype information (924 patients). For each omic data, we trained a 1D convolutional neural network [3] by using subtype labels to reduce the original feature space and extract new features. Then, we concatenated the extracted features of each omic data and performed t-SNE algorithm on to be able to visualize the patients in 2D graph. Patients who are the same subtype were grouped together in 2D space which is what we expected. After analyzing the graph, we noticed that two new subgroups had emerged and decided to cluster the patients into seven subgroups by using hierarchical clustering algorithm. However, we set the cluster size parameter to nine to be able to cluster outlier patients to their subgroups and discard them for further analysis. Clustering results were highly concordant with the original subtypes but our focus was on the two new subgroups. We performed differential gene expression analysis between each new subgroup and every other subgroup to identify upregulated genes in the new subgroups. By giving identified upregulated genes to functional enrichment analysis tool g:Profiler [4], we found that PPAR signaling pathway is enriched in one of the new subgroups and cilium term is enriched in the other new subgroup. These are already associated with breast cancer in the literature [5,6,7]. We conclude that these pathways may be used as biomarkers for newly discovered subgroups, however, more extensive analysis is needed to validate these initial results.

References

1. Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J., & Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. 15.

2. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Współczesna Onkologia, 1A, 68–77.

3. Mostavi, M., Chiu, Y.-C., Huang, Y., & Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. BMC Medical Genomics, 13(S5), 44.

4. Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., & Vilo, J. (2019). g:Profiler: A web server for functional enrichment analysis

and conversions of gene lists (2019 update). Nucleic Acids Research, 47(W1), W191–W198.

5. Zaytseva, Y. Y., Wallis, N. K., Southard, R. C., & Kilgore, M. W. (2011). The PPAR Antagonist T0070907 Suppresses Breast Cancer Cell Proliferation and Motility via Both PPAR-dependent and -independent Mechanisms. ANTICANCER RESEARCH, 11. , Y. Z., Xue, J. Y., Chen, C. M., Yang, B. L., Xu, Q. H., Wu, F., Liu, F., Ye, X., Meng, X., Liu, G. Y., Shen, Z. Z., Shao, Z. M., & Wu, J. (2012). PPAR signaling pathway may be an important predictor of breast cancer response to neoadjuvant chemotherapy. Cancer Chemotherapy and Pharmacology, 70(5), 637–644.

6. Wang, B., Liang, Z., & Liu, P. (2021). Functional aspects of primary cilium in signaling, assembly and microenvironment in cancer. Journal of Cellular Physiology, 236(5), 3207–3219.

Survival prediction of sepsis patients in an intensive care unit

Beste Kaysi and Ozgur Gumus
Department of Computer Engineering, Ege University, Izmir, Turkey 35030

Infection is one of the most important causes of death in patients treated in intensive care units in hospitals. In cases where the source of infection cannot be eliminated due to inadequate treatment, sepsis syndrome may occur. Sepsis takes place as a result of an exaggerated response of the immune system due to severe infection in any part of the body [1]. If the immune system is not taken under control to destroy the infection, a clinical picture that can lead to death in tissues and organs occurs. The incidence of sepsis continues to be the most important health problem all over the world, despite the developments in the field of health. Regardless of age, gender and income, sepsis syndrome can be seen in everyone. It ranks first among the causes of death especially in intensive care patients and is among the top ten causes of death due to all diseases [2]. According to the World Health Organization (WHO) estimates, sepsis affects more than 49 million people worldwide and causes about 11 million deaths annually [3]. Today, many studies, particularly in the field of medicine, estimate survival from diseases by using traditional statistical methods. In this study, using various machine learning algorithms, the survival of sepsis patients hospitalized in the intensive care unit was estimated with the data obtained from the Medical Information Mart for Intensive Care database. As a result of our initial findings, the most successful prediction accuracy was obtained with the Extreme Gradient Boosting algorithm with 82.7%. In future studies, it will be tried to increase the prediction success by adding more clinical features that may affect survival from diseases to the model.

References

1. Singer M., Deutschman C. S., Seymour C. W., et al. (2016). The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA, 315(8):801–810.

2. Martin G. S., Mannino D. M., Eaton S., Moss M. (2003). The epidemiology of sepsis in the United States from 1979 through 2000. N Engl J Med, 348:1546-54.

3. World Health Organization (WHO). (2021). Improving the Prevention, Diagnosis and Clinical Management of Sepsis. https://www.who.int/sepsis/en/.

Potential implementation of amino acid conjugates as novel micronutrient fertilizers

Emre Aksoy

Dept. of Biological Sciences, Middle East Technical University, Ankara, Turkey

Essential micronutrients are required in many vital processes in plants. Proper plant development and yield are strongly affected by micronutrient availability in the soil. Micronutrient deficiencies decline the crop yield and, in turn, cause extensive revenue losses in the world. Fertilizer expenses are among the top cost items for any agricultural production system. Therefore, the development of novel fertilizers is at the forefront to decrease the agriculture cost. Recent studies have demonstrated the conjugation of different amino acids with micronutrients, especially in the divalent form. Among them, negatively charged amino acids stand out with their chemical properties to complex with iron, zinc, magnesium, and manganese, etc. in aqueous solutions. Some of these amino acids were shown to correct the micronutrient deficiencies in plants meanwhile their usage in the biofortification of crops is increasing in the last few years. This review will focus specifically on the physical structures of amino acids conjugated with micronutrients and the potential of their utilization as ground-breaking fertilizers in the correction of mineral deficiencies in plants.

Functional stratification of small molecule drugs through integrated network similarity

Seyma Unsal Beyge[1] and Nurcan Tuncbag[2]

[1]Informatics Institute, Middle East Technical University
[2]School of Medicine, Koc University

Recovery of an optimal cancer treatment strategy is challenging due to the inter- and intra-heterogeneity of tumor samples. Modulations in signaling pathways and interactions between various bio-entities are critically important in the multi-stage tumor cell formation. Hence, multi-omic data integration is vital for understanding the molecular interactions happening in the cancer cell and for development of an optimum treatment strategy. Since development and approval of new drugs is both expensive and time-consuming, drug repurposing is an advantageous strategy in cancer treatment. Classification of currently available therapeutic agents is important but also a complex procedure and necessary to index possible drugs for drug repurposing approaches. Moreover, determination of molecular mechanisms of available drugs in different cancer types deciphers the possibilities in drug repurposing given the heterogeneity of cancer. Conventionally, drugs are classified based on their primary targets, therapeutic actions, target specificity, nature of interaction, molecular type and chemical structure similarity. The effects of small drug molecules are highly dependent on the cellular and physiological factors. Many drugs have multiple targets with variable binding affinities. Depending on the features selected, groups of similar drugs may change. Even though two drugs are present in the same group, they may modulate different signaling mechanisms within the cell.

In this study, the transcriptomic and phosphoproteomic data of cell lines treated with small molecule drugs are used to reconstruct networks by integrating the data with drug targetome and human interactome. Data integration is a challenging task since the proper integration method changes depending on the nature of the data and the network reconstruction principles. In this study transcriptomic data is used to back trail the regulatory elements that acts on the experimental hits. Phosphoproteomic hits allowed the selection of functionally active proteins that may be closely related to the drug modulatory effects. Also, human interactome is referenced by the network reconstruction software, Omics Integrator, to map the seed proteins and find the optimum connected subnetwork. Human interactome is known with its incompleteness and also it has bias to well-studied proteins. Therefore, it has many false negatives and false positives. In this study, human interactome is processed to eliminate these drawbacks. First, high-degree nodes eliminated and for each cell line – drug condition, it is filtered for very low expressed genes by the aid of associated transcriptomic data. After, each interactome is enriched with a link prediction approach followed by localization filters. The final processed interactome used for each cell line – drug condition has turned into both tissue and drug specific.

Total of 250 cell line and drug specific networks are reconstructed covering 70 drugs and six cell lines. A rigorous topological and pathway analysis of these reconstructed networks provided insight into the drug modulations occurring in different cell lines, including mostly cancer models. It is found that chemically and functionally different drugs may modulate overlapping networks. Moreover, the target

selectivity of the drug is an important factor leading to separate networks for drugs with same mode of action. Network-based analysis coupled with multi-omic data integration helped to reveal cell line and drug specific hidden modulated pathways such that drugs having overlapping networks generally modulate transcriptional misregulation in cancer pathway. Next, topological distance and active pathways of drug networks may guide the use of efficient drug combinations. Finally, separation between networks of a drug across cell lines can help to infer their resistance or sensitivity or no response to that drug.

Towards integrative mechanistic models of mammalian cell responses to extracellular perturbations: growth factors, hormones, and cytokines

Cemal Erdem[1], Sean M. Gross[2], Laura M. Heiser[2] and Marc R. Birtwistle[1,3]

[1]Department of Chemical and Biomolecular Engineering, Clemson University, Clemson, SC

[2]Department of Biomedical Engineering, Oregon Health Science University, Portland, OR

[3]Department of Bioengineering, Clemson University, Clemson, SC

A critical missing capability in current cancer research is the ability to predict how a particular single cancer cell will respond to microenvironmental cues or a drug cocktail. Yet, it is not even possible to perform this task well for normal healthy cells. This work builds on the hypothesis that first principles, mechanistic models of how cells respond to different perturbagens, will ultimately improve drug combination response predictions. However, building such single-cell models of complex, large-scale, and incompletely understood systems remains an extremely challenging task. To address this issue, we defined an open-source pipeline for scalable, single-cell mechanistic modeling from simple, annotated input files (structured lists of species, parameters, and reaction types). The input files are converted into an SBML (Systems Biology Markup Language) model file. Using this pipeline, we:

1. Re-created one of the largest pan-cancer signaling models in the literature (774 species, 141 genes, 8 ligands, 2400 reactions)

2. Enlarged the model to include Interferon- (IFN) signaling pathway (950 species, 150 genes, 9 ligands, 2500 reactions)

3. Re-parametrized the model to test and prioritize candidate mechanisms for experimental observations

Specifically, we used the enlarged model to test alternative mechanistic hypotheses for the experimental observations that IFN inhibits epidermal growth factor (EGF)-induced cell proliferation. We ran stochastic single-cell simulations for two different crosstalk mechanisms and looked at the number of cycling cells in each case. Our model-based analysis suggested, and experiments support that these observations are better explained by IFN-induced SOCS1 expression sequestering activated EGF receptors, thereby downregulating AKT activity, as opposed to direct IFN-induced upregulation of p21 expression. Finally, our new modeling format is available online (github.com/birtwistlelab/SPARCED) and compatible with high-performance (Kubernetes) computing platforms, enabling us to study virtual cell population responses. Overall, our new model enables easy modification of large mechanistic models and simulation of thousands of single-cell responses to multiple ligands and drug combinations.

Sequence diversity of M proteins of Influenza A (H7N9) virus

Gizem Yilmaz[1], Li Chuin Chong[1], Hasiba Karimi[1], Eyyub Selim Unlu[2], Muhammed Miran Oncel[1] and Mohammad Asif Khan[3,4]

[1]Beykoz Institute of Life Sciences and Biotechnology, Bezmialem Vakif University, Turkey

[2]Istanbul University, Faculty of Medicine, Turkey

[3]Bezmialem Vakif University, Turkey

[4]Perdana University, Malaysia

Background: Influenza A H7N9 virus belongs to the Orthomyxoviridae family and is mostly transmitted from avian reservoir to human host and causes respiratory illness, such as pneumonia or acute respiratory distress syndrome. The first human infections were reported in China in February 2013 and to date hundreds of cases have been confirmed, with a high fatality rate, and a pandemic potential. The non-structural protein M1 (matrix protein) has been reported as a potential target for antiviral therapy. M1 plays a crucial role at many stages of the virus life cycle. The M2 membrane protein is a proton channel that is mostly expressed on virus-infected cells, causing dissociation of the viral RNA from the matrix proteins and fusion of the viral membranes. This study focuses on sequence diversity analyses of M1 and M2 proteins of the virus.

Methodology: M1 and M2 protein sequences of avian and human influenza A (H7N9) virus were retrieved from both the FLUDB and EpiFlu GISAID databases (as of January 2021). Duplicate sequences were removed for each host/reservoir dataset by use of CD-HIT. The sequences of the host/reservoir were then merged for each protein, co-aligned using MAFFT, and then separated according to host/reservoir. Shannon's entropy as a measure of sequence diversity was calculated for each separated host/reservoir dataset of the M1 and M2 proteins. Entropy calculation was carried out for each aligned overlapping nonamer position of the proteins by use of DiMA; the length of nine was chosen for immunological applications.

Results: The total downloaded sequences for M1 protein from FLUDB and EpiFlu GISAID were 1088 for avian viruses and 1494 for human viruses, collectively 2582. Similarly, for M2 protein, the numbers were 1641 for avian viruses and 1066 for human viruses, collectively 2707. Removal of duplicates reduced the M1 dataset to 198 sequences (avian: 105; human:93) and M2 dataset to 293 sequences (avian:159; human:134). M1 protein was highly conserved with a mean entropy of $\sim$0.35 and $\sim$0.61 for the human host and the avian reservoir, respectively. The peak entropy of M1 protein was $\sim$1.27 at aligned nonamer position 33 for human viruses, and was higher for avian viruses, of $\sim$1.64 at aligned nonamer position 161. The M2 protein exhibited higher diversity than M1, with a mean entropy of $\sim$1.18 for the human viruses and $\sim$1.46 for the avian viruses. The peak entropy of M2 was $\sim$2.79 at aligned nonamer position 20 for human viruses and significantly higher for avian viruses, of $\sim$3.12 also at nonamer position 20.

Discussion: According to Swan et al. (2021), their 2014 dataset showed M1 peak entropy of $\sim$0.30 for the human viruses and of $\sim$1.20 for the avian viruses. The M2

peak entropy was ∼1 for the human viruses and of ∼2.10 for the avian viruses. The peak values for the current 2021 dataset described herein are significantly higher than that of 2014 peak values (human M1, increase of 0.97; avian M1, increase of 0.44; human M2, increase of 1.79; avian M2, increase of 1.02). This is indicative of continued evolution of the virus with increasing diversification. Given the pandemic potential of the virus, continued surveillance is necessary.

Explainable artificial intelligence perspective to the computational drug discovery process

Kevser Kübra Kırboğa[1] and Ecir Uğur Küçüksille[2]

[1]Süleyman Demirel University and Research Center for Innovative Technology Application, Bioengineering Laboratory, 32260 Isparta, Turkey
[2]Süleyman Demirel University, Engineering Faculty, Computer Engineering Department, 32260 Isparta, Turkey

With the tools developed thanks to the latest technology, the difficult and costly aspects of the drug discovery process have evolved to be faster and more economical over time. Approximately 10 billion drug candidate molecules are identified in each project, and only a few of the molecules that are eliminated in silico can reach cell culture and animal experiments. For this reason, computer-assisted studies allow rapid screening and examination of many drug candidates at once. Undoubtedly, the volume and diversity of data has increased significantly with the development of new technologies and new data initiatives. Big data is used the entire drug discovery process, from targeting and mechanism of action to identifying new leads and drug candidates. Chemoinformatics tools offer enormous potential to advance the process of in silico drug design discovery, as they serve to increase the reliability of data results and integrate information at various levels. Chemical structure similarity research, data mining/machine learning, and bioactivity spectrum-based algorithms continue to be routinely and successfully implemented. Drug discovery software is used to develop new pharmaceutical drugs and to test whether a newly created drug will be effective in treating a particular disease. It automates and uses innovative technology that significantly reduces the demanding drug development, testing, and time to market. Most drug discovery solutions offer scanning, predictive analytics, modeling, simulation, and computation capabilities. These functions provide reproducibility as well as assist with tasks such as image analysis and presentation of clinical trial results. Researchers and scientists use drug discovery software to leverage advances in drug design and synthesis, combat evolving and adapting diseases, and protect and manage data integrity as drugs move from discovery to clinical trials. Python is an open-source programming language used in drug discovery. It plays a major role in the calculation of the bioactivity data of the targets. A successful drug discovery process includes many additional bioanalyses, such as efficacy in functional analysis, ADME, toxicity, and physicochemical properties required for a successful drug. In drug discovery, it is important to understand the range of critical compound properties that are best associated with the clinical survival or failure of these compounds to develop computational methods that can best predict the physicochemical properties of compounds. For this reason, scientists have developed the simple rules necessary for the target molecule to be a drug. These rules are used to identify possible active drug candidates with chemical and physical properties. They help determine whether a biologically active chemical has the chemical and physical properties for bioavailability orally. Various 'artificial intelligence' (AI) concepts have been successfully adopted for computer-assisted drug discovery over the past few years. This progress is mostly due to the ability of deep learning algorithms, namely artificial neural networks with multiple processing layers, to model

complex nonlinear input-output relationships and perform pattern recognition and feature extraction from low-level data representations. The ability to capture complex nonlinear relationships between input data and associated output often comes at the expense of the limited intelligibility of the resulting model. Although some methods attempt to explain complexity, deep neural network models fall short of being accessible. Therefore, mechanically interpretable and highly accurate models could be the key to accelerated drug discovery with XAI.

In this study, the computational drug discovery process, how to evaluate the transition of molecules to the drug process in Python software with statistical and graphical data, and how XAI's changes in the input to the drug discovery model affect the result, the process of determining the factors affecting the success of the model, the determination of the special decision points used by the model for decision, why and how this decision was made, how to learn what type of errors the model is sensitive to, and finally how it will play a role in learning how these errors can be corrected will be explained. Accordingly, in time- and cost-sensitive scenarios such as drug discovery, deep learning practitioners within the scope of computational drug discovery will detail the entire process from modeling choices to the responsibility of carefully examining and interpreting the predictions obtained.

Classifying antibiotic resistance mechanisms in dihydrofolate reductase by tracking dynamical shifts in hydrogen bond occupancies

Ebru Cetin, Ali Rana Atilgan and Canan Atilgan
Sabancı University

Antibiotic resistance is a global health problem in which mutations occurring in functional proteins render the drugs ineffective. In this study, the working mechanisms of nine trimethoprim resistant single mutants of dihydrofolate reductase (DHFR) determined by morbidostat experiments [1] are categorized into four distinct groups using hydrogen bonding occupancies. DHFR catalyzes DHF reduction by NADPH into tetrahydrofolate (THF) and NADP+, producing the precursor for purine/thymidylate synthesis. Loss of THF from the ternary complex (DHFR: THF: NADPH) is the rate-limiting step [2]. The structure of DHFR is comprised of an α/β arrangement of eight β strands and four α helices. Residues 38-104 are considered as Adenosine Binding Domain whereas residues 1-37 and 105-159 are named as the Loop Domain [3]. Loops are categorized as the FG Loop (residues 116-132), the GH Loop (residues 142 – 150), and the CD Loop (residues 64-71) [3]. Reduction of DHF occurs through the transfer of two protons. At the hydride transfer step, one proton is supplied by the cofactor NADPH, and the other is extracted from the environment.

A significant limitation of studying DHFR is due to the large timescales relevant for enzyme dynamics, as they vary for many proteins in the range from femtoseconds to seconds [4]. Interconversion time scales of the reactive substates, however, are distributed in the millisecond to microsecond range. These milli to microsecond distributed substate compositions are collapsed into the small portion of reactive substates either by ligand binding or catalysis [5]. Regarding the catalysis mechanism, DHFR experiences multiple conformational changes, including domain rotation and the motion of structural loops. One motion of the structural loops is defined as the occluded state, where the tip of the M20 loop closes over the DHF binding site [3].

In order to study the working mechanisms for DHFR, two replicates of WT and nine mutant molecular dynamics trajectories are generated, each of 210 ns length. For all the generated trajectories for the point mutants, we have studied several properties that would characterize the effect of the mutation. We find that the solvent accessible surface area of the binding site and the salt bridges do not significantly change upon the introduction of mutations; an exception to the latter is the formation of a salt bridge at R30-E139 for W30R [6]. We then turn to an analysis of all the hydrogen bonds formed in the protein. In particular, we focus on significant shifts in the hydrogen bond occupancies in comparison to those in the WT protein. Since slight changes in the tertiary structure can reflect on the hydrogen bond formation, this provides an excellent opportunity to classify the mechanisms of the single mutants.

In this work, we tracked those hydrogen bonds on single mutants of DHFR whose occupancy deviate from that of the WT by $\pm 30\%$. Of the total of 3019 hydrogen bonds between all pairs of residues in the WT and the nine mutants, only 18 show large differences in their occupancies. We classify the bonding patterns into four

distinct groups. The first group (L28R, W30G) relates to mutants directly working on DHF binding regions and consist of those displaying the highest fitness in the presence of the inhibitor trimethoprim in morbidostat experiments. The second group (W30R) is a particular case working on binding cavity tension happening due to tight binding and also displays high fitness. The third group (D27E, I94L, F153S) enables an interesting hydrogen bond network at the distant CD loop. The last group (I5F, A26T, R98P) does not have any significant changes in hydrogen bond occupancies; these mutants are also never observed as a first mutation, but rather appear as epistatic showing up after one of the other mutants is fixed. Armed with the knowledge of these categories, alternative strategies to block the development of extreme resistance to trimethoprim can be developed, as we have recently demonstrated for the L28R mutant [7].

References

1. Toprak, E.; Veres, A.; Yildiz, S.; Pedraza, J. M.; Chait, R.; Paulsson, J.; Kishony, R., Building a morbidostat: an automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nature Protocols 2013, 8 (3), 555-567.

2. Fierke, C. A.; Johnson, K. A.; Benkovic, S. J., Construction and evaluation of the kinetic scheme associated with dihydrofolate reductase from Escherichia coli. Biochemistry 1987, 26 (13), 4085-4092.

3. Sawaya, M. R.; Kraut, J., Loop and Subdomain Movements in the Mechanism of Escherichia coli Dihydrofolate Reductase: Crystallographic Evidence. Biochemistry 1997, 36 (3), 586-603.

4. Boehr, D. D.; McElheny, D.; Dyson, H. J.; Wright, P. E., The Dynamic Energy Landscape of Dihydrofolate Reductase Catalysis. Science 2006, 313 (5793), 1638.

5. Benkovic, S. J.; Hammes, G. G.; Hammes-Schiffer, S., Free-Energy Landscape of Enzyme Catalysis. Biochemistry 2008, 47 (11), 3317-3321.

6. Tamer, Y. T.; Gaszek, I. K.; Abdizadeh, H.; Batur, T. A.; Reynolds, K. A.; Atilgan, A. R.; Atilgan, C.; Toprak, E., High-Order Epistasis in Catalytic Power of Dihydrofolate Reductase Gives Rise to a Rugged Fitness Landscape in the Presence of Trimethoprim Selection. Molecular Biology and Evolution 2019, 36 (7), 1533-1550.

7. Manna, M. S.; Tamer, Y. T.; Gaszek, I.; Poulides, N.; Ahmed, A.; Wang, X.; Toprak, F. C. R.; Woodard, D. R.; Koh, A. Y.; Williams, N. S.; Borek, D.; Atilgan, A. R.; Hulleman, J. D.; Atilgan, C.; Tambar, U.; Toprak, E., A trimethoprim derivative impedes antibiotic resistance evolution. Nature Communications 2021, 12 (1).

Expression profile survey of circular RNAs and their parent genes in context of tissue specificity

Elif İrem Keleş and Alper YILMAZ

Department of Bioengineering, Yildiz Technical University, Istanbul, Turkey

NcRNA (non-coding RNA), which covers a large part of the human genome, is involved in regulation of gene expression at the transcriptional and post-transcriptional levels. Circular RNAs (circRNAs) are non-coding RNA molecules that a class of newly-identified and single-stranded covalently closed RNA rings. CircRNAs were firstly thought as an deduced splicing default that is responsible for regulation of genes by interacting with other non-coding RNAs. Linear mRNA is generated from precursor mRNA by splicing, in contrast, circular RNA is originated from protein coding and non-coding genes by backsplicing mechanism between 3' end and 5' end. As a result of this mechanism, circRNAs are resistant to cellular exonucleases and have enhanced stability due to the lack of free ends. CircRNAs are found in various organisms from archaea to eukaryotes and their stability is very high; however, their biological functions have not been fully determined. They have tissue or cell specific expression levels. Thus, it is predicted that circRNAs, because of their highly stable nature and significant tissue specificity, will emerge as reliable biomarkers of disease. There is no comprehensive research about tissue specific expression patterns of circRNAs. We aim to search the relationship between the expression profile of the host gene and its circRNA expression profile.

In this study, circRNA and parent gene expression data is retrieved from the circAtlas 2.0 [1] which integrates over one million circRNAs across 6 species (human, macaca, mouse, rat, pig, chicken) and a variety of tissues. Genes solely expressed in one tissue are selected. CircRNA and gene data that are only in one tissue are joined and analyzed with R programming language.

As a result, we observed that some genes are highly expressed in any tissue, their circRNAs are expressed in different tissue. Although hsa-CELA2_0002 circRNA is only expressed in brain, its parent gene is highly expressed in pancreas. Hsa-TNNT1_0006 circRNA is expressed in heart,its parent gene is highly expressed in skeletal muscle in the same way. The results suggest that contrary to the research so far, circRNA could have been generated by a mechanism yet to be discovered.

References

1. Wu W, Peifeng J, Zhao F. "CircAtlas: an integrated resource of one million highly accurate circular RNAs from 1070 vertebrate transcriptomes". Genome Biol. 2020. 21(1):101. doi: 10.1186/s13059-020- 02018-y

Investigation of radicals present in biological systems by molecular modeling methods

Busra Bas and Cenk Selcuki
Ege University

Free radicals are essential for most of the reactions, especially the oxidation, occurring in biological systems but there should be a maintained balance between the production and consumption of free radicals. Because of their high reactivity, they attack macromolecules such as proteins, lipids, or nucleic acids rapidly. They are also considered as the reason for aging, mutations, neurological disorders such as Alzheimer's, Parkinson's diseases, and cancer. In the human body, so many different kinds of reactive species can be generated. Creatine is known as a bioenergetic molecule but recently it's been found to be potentially therapeutic for diseases such as Alzheimer's and Parkinson. Taurine is the most abundant free amino acid in the brain and the researches show that taurine could be associated with neurological disorders just like creatine. And tyrosine is a crucial amino acid. It's the precursor of neurotransmitters including epinephrine, norepinephrine, and dopamine. The radical form of tyrosine, in this case tyrosyl, can be found free or in the protein structure and it helps various types of enzyme to perform their catalytic activity, for instance ribonucleotide reductase class 1 enzyme has a stabile tyrosyl radical in its active site.

The computational chemistry methods can be a source of information for behaviors of target molecules in vitro. This study aims to investigate the conformational properties and the lowest energy geometries of radical forms of creatine, taurine, and tyrosine by molecular modeling approaches. Two different kinds of basis sets are selected for DFT calculations and also MP2 method will be used for calculations. Although the radical forms are related to many diseases, mechanisms remain unknown. Thus investigation of carbon centered radical forms by quantum mechanical calculations could be enlightening for further studies.

This work is supported by Ege University Research Funds (BAP) Project No: FYL-2020-22417. Some of the calculations were performed on TUBITAK ULAKBIM TRUBA Resources.

Human inbreeding has decreased in time through the Holocene

K. Gürün[1,+], F.C. Ceballos[1,+], N.E. Altınışık[2], H.C. Gemici[1], C. Karamurat[1], D. Koptekin[1], K.B. Vural[1], I. Mapelli[1], E. Sağlıcan[1], E. Sürer[1], Y.S. Erdal[2], A. Götherström[3], F. Özer[2], Ç. Atakuman[1] and M. Somel[1]

[1]Middle East Technical University, Ankara, Turkey
[2]Hacettepe University, Ankara, Turkey
[3]Stockholm University, Stockholm, Sweden
[+]Equal Contribution

Runs of homozygosity (ROH) are long homozygous stretches of the genome, presence of which indicates inbreeding due to small population sizes and genetic drift, and/or mating between close relatives, i.e., consanguinity [1]. We developed a method to optimize the parameters of PLINK to detect ROH, which relies on a model-free, observational approach that does not require a reference panel [2]. We were able to tune ROH calling parameters to suit low genomic coverages and correct for ROH overestimation. Our method works efficiently down to 3x SNP coverage and reliably calls ROH $>$ 1 Mb in genomes across $>$ 1 million SNPs [3]. We confirmed the power and accuracy using simulations and by comparison with a recently published method, which relies on a reference haplotype panel [4]. We used our approach to study the controversial history of human inbreeding by systematically analyzing for the first time the ROH levels in 411 published ancient genomes of the last 15,000 years from West and Central Eurasia. The Neolithic Transition to food production and the development of sedentary and/or agricultural societies may have influenced overall inbreeding levels, relative to those of hunter-gatherer communities. This transition could have had opposite effects on average ROH levels: ROH might decrease by increasing population size [5], or ROH might increase due to higher consanguinity and endogamy in farming communities compared to forager communities [6-7]. We estimated the genomic inbreeding coefficient FROH per genome as the sum of ROH $>$ 1.5 [8] and showed that the frequency of inbreeding, as measured by FROH, has decreased over time throughout the Holocene. This result was robust to the SNP list used and was reproducible in downsampling experiments. The result was also supported by a multiple regression model that includes time (age), cultural groups, and technical covariates. Some ancient individuals showed high FROH, but were rare in the sample and they included both hunter-gatherers and farmers. The main shift in FROH happens after the Neolithic, but the trend has since continued, indicating a population size effect on ROH and inbreeding prevalence. Post- Neolithic increase in population admixture [9] may also play role. We further show that most inbreeding in our historical sample can be attributed to small population size and drift, instead of consanguinity. Such high drift individuals were mainly hunter-gatherers. Extreme consanguineous matings did occur, but were rare and only observed among agriculturalist members of farming societies in our sample, in line with ethnographic work [6-7]. Despite the lack of evidence for common consanguinity in our ancient sample, consanguineous traditions are today prevalent in various modern-day Eurasian societies, suggesting that such practices may have become widespread within the last few millennia.

References

1. Ceballos et al. 2018 Nat Rev Genet
2. Chang et al. 2015 Gigascience
3. Ceballos et al. 2020 Biorxiv
4. Ringbauer et al. 2020 Biorxiv
5. Gignoux et al. 2011 PNAS
6. Walker 2014 Evol & Human Behav
7. Hill et al 2011 Science
8. McQuillan et al. 2008 AJHG
9. Lazaridis et al. 2016 Nature

Biomarker prediction for Parkinson's disease by transcriptome mapping on a genome-scale metabolic model

Ecehan Abdik and Tunahan Çakır
Department of Bioengineering, Gebze Technical University

Parkinson's disease (PD) is the most common neurodegenerative motor function disorder, and it influences 1% of the population older than 60 years old [1]. Pathological indications of the disease appear at the molecular level several years before the clinical symptoms. Therefore, the investigation of preclinical biomarkers of PD is essential to diagnose the disease in the early stages and develop target-based therapeutical agents. Biomarker prediction by integrating computational models and omics data is an active research field [2,3]. TIMBR (Transcriptionally Inferred Metabolic Biomarker Response) [4] is an algorithm that predicts metabolite biomarkers from the constraint-based analysis of metabolic networks. TIMBR basically compares the network demands for the production of external metabolites between two states as a function of gene expression changes. Before, it was successfully applied to predict early markers for toxicant-induced organ damages in rats [5-7] and to analyze metabolite level changes of different experimental mouse models of PD [8].

In this study, TIMBR algorithm was used to predict candidate biomarkers for PD by utilizing thirteen different PD transcriptome datasets from post-mortem human subjects, which are available in Gene Expression Omnibus (GEO) [9] and ArrayExpress [10]. The datasets were from substantia nigra and prefrontal cortex Broadmann Area 9 (BA9) regions, the most affected brain regions in PD. They were separately mapped on a generic human genome-scale metabolic model (Human-GEM) with 3628 genes and reactions through TIMBR to analyze the main effects of the disease on metabolism. Metabolites predicted as biomarkers for the majority of the datasets were suggested as reliable biomarkers with a meta-analysis approach. Enrichment analysis was also applied for metabolites whose secretion pattern were predicted to be changed in PD. Enriched terms for the substantia nigra are mainly related to amino acid metabolism, catecholamine biosynthesis, and mitochondrial electron transport chain.

References

1. Dawson, T. M. & Dawson, V. L. Molecular pathways of neurodegeneration in Parkinson's disease. Science (80-.). 302, 819–822 (2003).

2. Shiri Stempler, K. Y. & Ruppin, E. Integrating transcriptomics with metabolic modeling predicts biomarkers and drug targets for Alzheimer's disease. PLoS One 9, (2014).

3. Ozerov, I. V et al. In silico Pathway Activation Network Decomposition Analysis (iPANDA) as a method for biomarker development. Nat. Commun. 7, 1–11 (2016).

4. Blais, E. M. et al. Reconciled rat and human metabolic networks for comparative toxicogenomics and biomarker predictions. Nat. Commun. 8, 14250 (2017).

5. Pannala, V. R. et al. Metabolic network-based predictions of toxicant-induced metabolite changes in the laboratory rat. Sci. Rep. 8, 1–18 (2018).

6. Pannala, V. R. et al. Genome-scale model-based identification of metabolite indicators for early detection of kidney toxicity. Toxicol. Sci. 173, 293–312 (2020).

7. Rawls, K. D. et al. Genome-Scale Characterization of Toxicity-Induced Metabolic Alterations in Primary Hepatocytes. Toxicol. Sci. 172, 279–291 (2019).

8. Abdik, E. & Cakir, T. Systematic Investigation of Mouse Models of Parkinson's Disease by Transcriptome Mapping on a Brain-Specific Genome-Scale Metabolic Network. Mol. Omi. (2021).

9. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 30, 207–210 (2002).

10. Brazma, A. et al. ArrayExpress—a public repository for microarray gene expression data at the EBI. Nucleic Acids Res. 31, 68–71 (2003).

Constraint-based modelling and machine learning identifies metabolic alterations in the Substantia nigra in Parkinson's disease

Regan Odongo and Tunahan Çakır
Department of Bioengineering, Gebze Technical University, Gebze/Kocaeli

Perturbed metabolism is a major molecular feature in Parkinson's disease (PD) pathology; however, the complete compendium of altered metabolites/metabolic pathways remains elusive. Genome scale metabolic modelling can be used to systematically simulate condition-specific flux distributions in healthy and disease conditions. Machine learning methods can then be used to identify features associated with phenotypes from such high-dimensional data. Here, we used iBrain671, the current human brain specific genome-scale metabolic network containing 994 reactions governed by 671 genes, to obtain personalized metabolic models for healthy control and PD using post-mortem human brain transcriptome of the Substantia nigra tissues. We applied iMAT, an algorithm that tailors genome-scale metabolic models based on a given transcriptome data, to generate personalized models and separately used two algorithms (MOMA and LseiFBA) to predict PD flux distributions from the personalized models. Population-wide biomarker identification using ElasticNet Machine Learning model identified perturbations in important metabolic reactions that were consistent with previous literature reports and were related to mitochondrial transport and oxidative phosphorylation, GABA and Glutamate-Glutamine cycle, glycolysis, ROS, TCA cycle, amino acids (methionine, leucine, phenylalanine-tyrosine metabolism), histamine and polyamine metabolisms. Reactions in these metabolic pathways were found to be important in differentiating between the two conditions. This study demonstrates how machine learning can be used to augment genome-scale metabolic modeling to prioritize metabolic features for subsequent validation studies and lays the foundation for increased application of machine learning techniques for the exploration of fluxome data in the context of systems medicine.

Reconstruction and transcriptome-based analysis of rat brain specific genome scale metabolic network model for Parkinson's disease

Orhan Bellur and Tunahan Çakır
Gebze Technical University Bioengineering Department

Parkinson's disease (PD) is the second most common neurodegenerative disorder in the world. It is characterized by the loss of dopaminergic neurons in the midbrain's substantia nigra (SN) region. Several studies have shown that metabolic dysregulations are associated with the neurodegeneration in PD. Changes in redox homeostasis and energy metabolism are thought to lead to dopaminergic cell homeostasis impairment. Genome-scale metabolic models (GEMs) include all known metabolic processes active in an organism, and they are commonly used to simulate metabolic alterations and to establish a mechanistic connection between genotype and phenotype. Only a portion of known metabolic reactions is active in a specific condition (eg. disease state, healthy state). GEMs can be integrated with transcriptome data to build condition-specific metabolic models. Rats are one of the most often used preclinical model organisms for neurodegenerative disorders, and the interpretation or rat-based transcriptome datasets of PD is important for disease diagnosis and treatment of neurodegenerative diseases. In this study, a rat brain-specific genome-scale metabolic model (iBrain766-Rn) was reconstructed based on a homology-based approach for the first time in literature. 11 comparisons from six rat PD-associated transcriptome datasets were then used to reconstruct condition-specific GEMs. Transcriptome datasets were integrated into the newly reconstructed iBrain766-Rn model by iMAT and INIT algorithms, and 44 condition-specific models (11 iMAT-based models, 11 INIT-based models separately for control and disease conditions) were created. Four constraint-based modeling methods including FBA, MOMA, LseiFBA, and modified linear LseiFBA (mLseiFBA) were used to obtain flux predictions for healthy and Parkinson's disease cases. The analysis on 44 condition-specific models revealed that the iBrain766-Rn model can simulate brain metabolism correctly in terms of flux prediction in healthy and PD conditions. The results between constraint-based modeling methods were substantially different, and the mLseiFBA method was found to perform better in terms of flux prediction for PD. Difference in flux predictions between algorithms indicates that the model reconstruction approach (iMAT or INIT) has an important role on the flux prediction.

Discovery of latent drivers from double mutations in pan-cancer data reveal their clinical impact

Bengi Ruken Yavuz[1], Chung-Jung Tsai[2], Ruth Nussinov[2] and Nurcan Tuncbag[3]
[1]Middle East Technical University
[2]National Cancer Institute at Frederick
[3]Koç University

Despite massive advancements in cancer genomics, to date driver mutations whose frequencies are low, and their observable translational potential is minor have escaped identification. Yet, when paired with other mutations in cis, such 'latent driver' mutations can drive cancer. Here, we discover potential 'latent driver' double mutations. We applied a statistical approach to identify significantly co-occurring mutations in the pan-cancer data of mutation profiles of ∼80,000 tumor sequences from the TCGA and AACR GENIE databases. The components of same gene doublets were assessed as potential latent drivers. Merging significant double mutation with drug response data of cell lines and patient derived xenografts (PDXs) allowed us to link the potential impact of double mutations to clinical information and discover signatures for some cancer types. Our comprehensive statistical analysis identified 228 same gene double mutations of which 113 are cataloged as latent drivers. Oncogenic activation of a protein can be through either single or multiple independent actions. Combinations of a driver mutation with either a driver, a weak driver, or a strong latent driver have the potential of a single gene leading to a fully activated state and high drug response rate. Evaluation of the response of cell lines and patient-derived xenograft data to drug treatment indicate that in certain genes double mutations can increase oncogenic activity, hence obtain a better drug response (e.g. in PIK3CA), or they can promote resistance to the drugs (e.g. in EGFR).

ProFAB – Open Protein Functional Annotation Benchmark

Ahmet Samet Özdilek[1], Ahmet Atakan[1], Tunca Doğan[2], Rengül Çetin-Atalay[3], Mehmet Volkan Atalay[1] and Ahmet Süreyya Rifaioğlu[4]

[1]Middle East Technical University
[2]Hacettepe University
[3]University of Chicago
[4]İskenderun Technical University

As the number of protein sequences in protein databases increases, accurate computational methods are required to annotate the available data. For this purpose, several machine learning methods have been proposed in recent years [1]. However, two main issues in the evaluation of computational prediction methods are the construction of reliable positive and their negative training/validation datasets and the fair evaluation of performances based on predefined experimental settings. Recently, several benchmarking platforms have been proposed in various fields to overcome similar issues. For example, Therapeutics Data Commons provides ready-to-use biomedical datasets for drug, toxicity, screening, antibody development along with the appropriate evaluation metrics [2]. Open Graph Benchmark is a platform that provides social and biological network datasets, and experimental settings for fair comparison of algorithms [3]. In the field of protein function prediction, Critical Assessment of Functional Annotation (CAFA) challenge [4] is an important initiative where the aim is to evaluate the performances of automated protein function prediction methods. CAFA challenge is organised about every two years; however, it is a one-time challenge and it is not trivial to repeat the challenge with the same experimental settings afterwards. In addition, CAFA challenge does not provide any training dataset, machine learning models or different data splitting strategies (e.g., temporal, similarity-based and random split settings). To overcome these limitations, here, we propose an open-source Python package called ProFAB, Open Protein Functional Annotation Benchmark platform. The aim of ProFAB is to create a fair comparison platform for protein function prediction methods based on Gene Ontology (GO) (861,299 proteins annotated to 8360 function terms) and Enzyme Commission (EC) numbers (563,554 proteins annotated to 269 enzymatic functions). ProFAB supplies both positive and negative datasets separately for each function. To obtain numerical features from protein sequence data, various protein descriptors are provided: amino acid composition (AAC), pseudo amino acid composition (PAAC), sequence order coupling number (SOCNumber), conjoint triad (CTriad) and grouped amino acid composition (GAAC). ProFAB consists of 5 main functionalities: (i) Training dataset construction where several training/test/validation datasets are created using UniProtKB and UniRef databases, (ii) Data splitting where random, temporal and similarity-based splitting methods are provided, (iii) Standardization where several methods are used to scale the input features, (iv) Training and tuning of the classifiers which constitute support vector machine (SVM), random forest, k-nearest neighbor (KNN), decision trees, naive bayes, multilayer perceptron, gradient boosting, logistic regression and the ridge classifier (tuning of these classifiers is done automatically by the modules with the predefined parameters that are specific and modifiable for each machine learning algorithm), (v) Evaluation where several

metrics are used to assess the predictive performance of trained models. To perform above functionalities, ProFAB uses Python modules and interfaces such as NumPY, scikit-learn, RDKit, pickle and tqdm. To convert protein sequences into numerical feature vectors, iLearn web tool [5] was used. With these implementations, we believe that ProFAB is useful for both computer scientists to find ready-to-use biological datasets, and wet-lab researchers to utilise ready-to-use machine learning algorithms to gain pre-knowledge about the functional view of proteins. ProFAB is available at https://github.com/Sametle06/ProFAB.git. To access the use case, please see: https://github.com/Sametle06/ProFAB/blob/master/test_file.ipynb.

References

1. Tiwari, A. K., et al. (2014). International Journal of Proteomics, 2014, 1–22. https://doi.org/10.1155/2014/845479

2. Huang K., et al. 2021, arXiv:2102.09548v1.

3. Hu W., et al. 2021, arXiv:2005.00687.

4. Zhou, N., et al. Genome Biol 20, 244 (2019). https://doi.org/10.1186/s13059-019-1835-8.

5. Zhen C., et al. 2020, Briefings in Bioinformatics, 21(3): 1047–1057. https://doi.org/10.1093/bib/bbz041

Language models can learn complex functional properties of proteins

Serbulent Unsal[1], Heval Ataş[2], Muammer Albayrak[1], Kemal Turhan[1], Aybar Acar[2] and Tunca Doğan[3]

[1]Karadeniz Technical University
[2]Middle East Technical University
[3]Hacettepe University

Proteins are essential macromolecules for life. To understand and manipulate biological mechanisms, functions of proteins should be understood, and this is possible through studying their relationship with the amino acid sequence and 3-D structure. So far, only a small percentage of proteins could be functionally characterized (currently ∼0.5% according to UniProt) due to cost and time requirements of wet-lab-based procedures. Lately, protein function prediction (PFP), which can be defined as the annotation of proteins with functional definitions using statistical/computational methods, gains importance to explore the uncharacterized protein space and/or protein variants carrying function altering changes. Among many different algorithmic approaches proposed so far, machine learning (ML), especially deep learning (DL), techniques have become popular in PFP due to their high predictive performance. The input data used by these ML/DL methods are numerical feature vectors representing the protein (i.e., protein representations), and they are mostly generated from amino acid sequences of proteins which are readily available in databases (e.g., UniProt).

Early protein representation construction methods built upon evolutionary relationships, the composition of amino acids in the sequence, and/or the physicochemical features of amino acids, which are directly related to the biochemical function of the protein. These methods can be considered classical "model-driven" approaches. Recently, researchers started to utilize ML models to automatically learn protein representations from available protein data (e.g., protein sequence, protein-protein interactions, biomedical literature/texts), in the context of a "data-driven" approach called protein representation learning (PRL). Most of the novel PRL models are based on algorithms from the natural language processing (NLP) field (e.g., word2vec, doc2vec, LSTM/transformerbased BERT, XLNet, etc.), which are originally developed to model languages for automated and simultaneous translation, context-based text generation, etc. with elevated performances. In recent years, the number of PRL methods has multiplied and they are starting to be used in various areas from biomedicine to biotechnology. However, there is no comprehensive study and tool available to assess these representation methods in the context of modeling the functional properties of proteins, to help the researcher choose the suitable method for the task at hand.

In this study, we evaluated protein representation methods for the prediction of functional attributes of proteins and benchmarked these methods in 4 challenging tasks, namely: (i) Semantic similarity inference (we calculated pairwise semantic similarities between human proteins using their gene ontology annotations and compared them with representation vector similarities to observe the correlation in-between), (ii) Ontological protein function prediction (we built GO term categories

based on term specificities and the sample sizes which reflects different levels of predictive difficulty and evaluated representation methods by training/validating ML models on these datasets), (iii) Drug target protein family classification (five major target families are selected and methods are evaluated in terms of classifying proteins to families via ML models), and (iv) Protein-protein binding affinity estimation (we used the SKEMPI dataset to evaluate methods in estimating protein-protein binding affinity changes upon mutations). We evaluated 23 protein representation methods in total, including both classical approaches and cutting-edge representation learning methods, to observe whether these novel approaches have advantages over classical ones, in terms of extracting high-level/complex properties of proteins that are hidden in their sequence. Finally, we provide an open-access tool, PROBE (Protein RepresentatiOn BEnchmark), where the user can assess new protein representation models over the abovementioned benchmarking tasks with only a line of code.

The results of benchmarks showed that numerous DL-based PRL methods, especially large-scale protein language models, performed significantly better than classical representation methods on function and structural feature prediction-related tasks. Also, results indicated that the model architecture and training data types/sources are the two key factors affecting the performance. We also inspected possibilities of data leak from training to test, the cases where the tasks used during the pre-training are biologically related to the benchmark tasks (e.g., models that are constructed using Pfam protein family annotation data are good at predicting structural features since these two are directly related). Furthermore, we discussed current challenges in PRL such as differences between problems in the NLP domain and the ones in protein informatics, in the context of data structures and model interpretability. Finally, we discussed future applications of PRL in the fields of automated protein design and engineering. The details of methodology and results can be found in our preprint (https://doi.org/10.1101/2020.10.28.359828), which will be examined in detail and discussed further. PROBE/ProtBench is available at https://github.com/serbulent/TrainableRepresentationAnalysis.

Consensus clustering analysis as a sample selection method in biomarker discovery: Lung cancer case-study

Nehir Kızılilsoley and Emrah Nikerel

Yeditepe University, Genetic and Bioengineering Department

Lung cancer is the most lethal cancer type in both men and women, as it comes first in cancer-related deaths worldwide; with a survival rate of 15% in the first 5 years and 7% in 10 years after diagnosis [1]. The high rates of mortality arise from not only the lack of early diagnosis strategies but also the lack of efficient treatments specialized for the stage and subtype of lung cancer the patients are suffering from. These limitations reflect an urgent need for biomarkers that allow for early stage diagnosis and prognosis of lung cancer, which may improve the treatment patients receive [2]. Along with the advance of omic technologies, transcriptomics has assisted greatly in the identification of biological markers for not only lung cancer but numerous phenotypes of interest. Finding gene markers that allow for accurate and non-invasive diagnosis that are also sensitive and reproducible is of great interest in precision medicine.

The sample selection is one of the steps of biomarker discovery that tunes the quality of the analysis. The feature selection algorithms and classifiers are built by the contribution of each sample that is included in the pipeline. As a result, each sample affects the quality of the output i.e. gene marker candidates. Consensus clustering analysis is a technique used to run a (collection of) selected clustering algorithm (k-means, hierarchical, etc.) recursively on sub-samples of the datasets to obtain a consensus result from all the results of each iteration; to e.g. determine the optimum number of clusters, to evaluate the stability of the found clusters, to reduce data dimension while keeping the information content etc. [3, 4].

In this study, consensus clustering analysis was used to select samples that are most representative (core samples) of the 2 main subtypes of lung cancer: adenocarcinoma (AD) and squamous cell carcinoma (SC). For this work, TCGA data of 65048 genes and 1145 lung cancer samples were used as a case study. The consensus clustering analysis has shown that the lung cancer data (AD: 535, SC: 502, healthy: 108 samples) reaches its maximum cluster stability when divided into three, which coincides with the number of actual groups of the samples. Using silhouette width as a measure for representative capacity of samples, 900 out of 1145 samples were selected as core samples and 245 remaining samples were discarded. The efficiency of using consensus clustering analysis as a sample selection method was assessed by naive comparison of marker genes from the initial set (1145 samples) and core set (900 samples). The features (genes) common to both lists are further discussed vis-a-vis lung cancer biomarker context.

References

1. Ferlay J, Ervik M, Lam F, Colombet M, Mery L, Piñeros M, Znaor A, Soerjomataram I, Bray F (2020). Global Cancer Observatory: Cancer Today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today, accessed [03 May 2021].

2. Indovina, P., Marcelli, E., Maranta, P., & Tarro, G. (2011). Lung cancer proteomics: recent advances in biomarker discovery. International journal of proteomics, 2011.

3. Monti, S., Tamayo, P., Mesirov, J., & Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine learning, 52(1), 91-118.

4. Şenbabaoğlu, Y., Michailidis, G., & Li, J. Z. (2014). Critical limitations of consensus clustering in class discovery. Scientific reports, 4(1), 1-13.

Interaction energy analysis of lidocaine and papaverine with the drug carrier pectin

Nesrin Işıl Yaşar[1], Tuğçe İnan[2], Ayşe Özge Kürkçüoğlu Levitas[2] and Fethiye Aylin Sungur[1]

[1]Computational Science and Engineering Division, Informatics Institute, Istanbul Technical University, Istanbul, Turkey

[2]Department of Chemical Engineering, Istanbul Technical University, Istanbul, Turkey

Pectin is a complex heteropolysaccharide found in the cell walls of all land plants. While the pectin forms water-soluble mixtures with monovalent cations due to its chemical structure, it also forms a water-insoluble gel with divalent and trivalent cations. In the gelation of pectin chains, the hydrogen bond formed by the carboxylic acid and hydroxyl groups in its structure and the hydrophobic interactions, together with the cation bridge between the chains of +2 charged ions such as calcium are effective. Low methoxy pectin gels with divalent cations forms the egg box structure, and the egg box structure becomes more stable with neighboring pectin chains with the help of hydrogen bonds and Van Der Waals interactions. Pectin is widely used in pharmaceutical applications due to its highly biocompatible, biodegradable, and non-toxic features. The hydrophilic structure of pectin allows the removal of wound fluid and helps to form a barrier against bacteria due to its acidic environment. In addition, since it binds to molecules such as drugs and growth factors, it takes an active role in the controlled release of these structures. In this context, the interaction ability of pectin with lidocaine, a local anesthetic, and papaverine, which is used to treat muscle spasms, were investigated by molecular dynamics simulations and quantum mechanical methods, respectively. In the MD calculations, the simulation boxes were minimized by steepest descent and the systems were equilibrated in the NVT ensemble using shake algorithm with gradual heating from 290 K to 298 K, and simulated for 50 ns at a constant temperature of 298 K in the NPT ensemble. In quantum mechanical studies, interaction energies were calculated from the density functional theory with the WB97XD functional using the 6- 311+G(d,p) basis set in the water as implemented in G16. The basis set superposition error was calculated with counterpoise method [1]. Charge calculations were performed with ESP and charge distributions were also examined to reveal the interactions between molecules under investigation.

References

1. S. F. Boys and F. Bernardi, Mol. Phys. 19, 553 (1970)

DebiasedDTA: Model debiasing to boost drug-target affinity prediction

Rıza Özçelik[1], Alperen Bağ[1], Berk Atıl[1], Arzucan Özgür[1] and Elif Özkırımlı[2]
[1]Bogazici University
[2]F. Hoffmann-La Roche AG, Switzerland

Prediction of drug-target affinity (DTA) in silico can significantly accelerate drug discovery process. Many in silico models rely on the drug-target interaction datasets, since they aim to learn the binding mechanisms between biomolecules through the information the datasets contain. However, the datasets on which the models rely also contain surface patterns, or dataset biases, that prevent models to generalize novel biomolecules [3, 2, 1]. Here we present DebiasedDTA, a model debiasing approach to boost the performance of DTA models on novel biomolecules. DebiasedDTA comprises a weak learner and a strong learner to identify and avoid dataset biases during training. With the fact that the non-complex models can memorize dataset biases easily, a weak learner is used to quantify the biases in the training samples. Once the bias is quantified, the strong learner avoids the biases by adjusting the training sample weights during training.

We experiment with 2 different weak learners to identify different bias sources: ID-DTA and BoW-DTA. ID-DTA is an identity-based weak learner that represents the biomolecules with one-hot encoding. On the other hand, BoW-DTA is a biomolecule-word-based approach that vectorizes the biomolecules with bag-of-words method, based on their chemical and protein words. Both weak learners concatenate chemical and protein representations to represent the interaction and use decision tree for regression.

We also experiment with 3 strong learners to observe the effect of debiasing with different strong learner architectures: DeepDTA, BPE-DTA, and LM-DTA. DeepDTA is a character-level-convolution based DTA model which was frequently used in the literature. DeepDTA uses SMILES strings of ligands and amino-acid sequences of proteins for biomolecule representation. We also design BPE-DTA, which uses the same model architecture as DeepDTA but segments sequences with Byte-Pair-Encoding instead of characters. Finally, we design LM-DTA that uses pre-trained biomolecule language model embeddings to represent the chemicals and proteins. Afterward, those representations are concatenated and fed into a 2 layered fully connected network.

We test DebiasedDTA on two datasets and evaluate the effect of debisasing on known and novel molecules sep- arately. The results show that DebiasedDTA improves prediction performance on 44 of 48 experiments, suggesting that the proposed approach is useful in most scenarios. Both the known and novel biomolecules benefit from the performance boost and the boost is amplified when the test biomolecules are dissimilar to training set. The experiments also highlight that both identity and word based biases are prevalent in the datasets and each experimented strong learner can leverage the novel training scheme in DebiasedDTA, indicating that the proposed approach is generalizable across models. As such, we believe that DebiasedDTA will be an in uential work for drug-target affinity prediction and will be used to debiase many future models.

References

1. L. Chen, A. Cruz, S. Ramsey, C. J. Dickson, J. S. Duca, V. Hornak, D. R. Koes, and T. Kurtzman. Hidden bias in the dud-e dataset leads to misleading performance of deep learning in structure-based virtual screening. PloS one, 14(8):e0220113, 2019.

2. J. Scantlebury, N. Brown, F. Von Delft, and C. M. Deane. Data set augmentation allows deep learning-based virtual screening to better generalize to unseen target classes and highlight important binding interactions. Journal of Chemical Information and Modeling, 60(8):3722-3730, 2020.

3. J. Yang, C. Shen, and N. Huang. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. Frontiers in Pharmacology, 11:69, 2020.

Methylation deviation as a marker of intratumor heterogeneity and cancer progression

Ersin Onur Erdoğan[1], Ömer Çinal[2] and Mehmet Baysan[2]

[1]Department of Computer Engineering, Istanbul University-Cerrahpaşa, Istanbul, Turkey
[2]Department of Computer Engineering, Istanbul Technical University, Istanbul, Turkey

Cancerous cells undergo consecutive molecular alterations that leads to several distinct clones within the same tumor. Intra-Tumor Heterogeneity (ITH) refers to the structural and/or functional differences among these different clones. ITH has been found to be associated with survival, tumor progression rate, risk of metastasis, immune system activity, therapeutic resistance, drug response and so forth. There have been many studies focusing on the explanation of ITH-associated prognostic features, based on copy number, mutation, and expression profiles. On the other hand, epigenetics based ITH predictions are not studied extensively in the literature. Epigenetic features such as DNA methylation and histone modifications play important regulatory roles and recent improvements in technology allow high-throughput measurement of these features.

In this study, we propose Methylation Deviation Index that is to be a prognostic biomarker explaining survival as the most significant ITH-associated feature. DNA methylation is among the most important epigenetic mechanisms that allows cells to differentiate for specific roles in different tissues. Aberrant methylation profiles may distort the expression profiles and lead to genomic instability which is a hallmark of many diseases including cancer. In this scope, we developed a methylation deviation index which measures the total deviation in methylation profile for each sample. We utilized TCGA pan-cancer data that includes DNA expression, Methylation, Copy Number, and Mutation profiles alongside the clinical data for 33 distinct tumor types with 6000+ samples. To assess the effect of methylation deviation on disease progression, we applied detailed survival analyses which include potential confounding variables such as age, subclone count, copy number changes, mutations, and expression deviation.

Survival analysis for pan-cancer analyses indicated that Methylation Deviation is significantly associated with Overall Survival and Disease Specific Survival (p-value < 0.005). We found out that our novel Methylation Deviation is associated with poor prognosis in pan-cancer dataset. This might be explained with the increased aggression of tumor due to high level of genomic instability. When we studied cancers individually, Methylation Deviation is positively associated with the risk of death for most of the cancers which complies with the pan-cancer results. The figure below displays the survival profiles for samples with different methylation deviation levels.

Predicting the impact of cancer somatic mutations on protein-protein interactions

Ibrahim Berber[1], Cesim Erten[2] and 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

An existing challenge in cancer genomics is to distinguish between driver and passenger mutations. Most of the existing methods count only the number of somatic mutations that occur in a gene across a set of patients, without considering the specific effect of each mutation. In this study, we focus on those somatic mutations that appear on the interface regions of the protein and predict the interactions that would be affected by a mutation of interest. To this end, we use IntAct's Mutations Influencing Interactions dataset as our training data where we filter for Single Nucleotide Variant (SNV)s and Homo Sapiens as organism [1]. This dataset is curated from experimental studies that investigate the effects of mutations on protein-protein interactions. Each entry contains information on sequence change and position of the mutation as well as the set of participating proteins within the interaction and the observed effect due to the mutation. We classify the effects into two main groups: disruptive group and increasing or no-effect group.

Next, we utilize the ELASPIC Web Server with mutations from our training data to extract several sequence and structure based features [2]. Then, we create a subset of this data where we sample a single mutation from each protein. This is done to ensure that we do not have two data instances with same values for some or all features as this would result in inflated accuracies in cross-validation setting. This results in 164 mutations across a total of 164 proteins where 106 mutations belong to disruptive group and 58 mutations belong to increasing or no effect group. We then train a random forest model to predict whether the mutation disrupts an interaction or not. We perform feature selection based on the SHAP (SHapley Additive exPlanations) values of features [3]. Using the selected set of 11 features, we obtain a balanced accuracy value of 0.7 and AUROC value of 0.8 in 10-fold cross-validation. We then compile features for 2292 number of mutations that we obtained from the TCGA BRCA dataset. Using our trained random forest model, we predict which interactions would be disrupted due to these mutations. Our analysis reveals that the BRCA cohort mutations on the PIK3CA gene often appear on the interface region across the patients (124 patients vs 35 patients). Another interesting gene is H3C1, which is mutated in 14 patients in total and in 13 of these patients, the mutation maps to the interface region. Also, in all these patients, the corresponding mutation is predicted to disrupt all the interactions of this protein.

Lastly, we check whether our predictions can improve the recovery of driver genes. To this end, we compare a baseline method which, for each protein, counts its existing number of interactions for each observed interface mutation to a more sophisticated approach that counts only those interactions that are predicted to be disrupted according to our random forest model. We then rank the proteins according to these counts and calculate AU-ROC values comparing our prioritization against a reference set of known cancer genes compiled from CancerMine database

[4]. As a reference, we also include a model (coverage) which simply counts the number of patients that the gene is mutated. Considering the effect of mutations on interactions results in an improvement in AU-ROC values (Figure 1).

To summarize, we developed a method to predict the effect of interface mutations and apply this on the TCGA BRCA dataset. Our predictions reveal interesting patterns where for some genes interface mutations dominate. Also, we show that driver identification methods can improve by utilizing the disruptiveness predictions of mutations.

References

1. Kerrien, S., et al. (2012). The IntAct molecular interaction database in 2012. NAR, 40(D1).

2. Witvliet, D. K., et al. (2016). ELASPIC web-server: proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics, 32(10).

3. Lundberg, S. M., et al. (2020). From local explanations to global understanding with explainable AI for trees. Nature machine intelligence, 2(1).

4. Lever, J., et al. (2019). CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. Nature methods, 16(6).

Drug-target interaction prediction using transfer learning

Alperen Dalkıran[1], Ahmet Süreyya Rifaioğlu[2], Aybar Can Acar[1], Tunca Doğan[3], Rengul Atalay[4] and Volkan Atalay[1]

[1]Middle East Technical University
[2]Iskenderun Technical University
[3]Hacettepe University
[4]The University of Chicago

Drugs are bioactive chemical compounds used to treat diseases. In the drug discovery process, the aim is to find a compound that regulates cellular functions in pathological conditions for medical treatment purposes, by interacting with targeted protein(s). New opportunities have emerged in the field of drug discovery with the introduction of computational methods whose aim is to reduce the duration and cost of drug development. Virtual screening of compounds against a target cell or protein is now a widely used step at the beginning of the drug discovery process, in which promising results have been obtained especially with deep learning-based models, in terms of drug-target interaction prediction and compound property prediction. However, its impact has been limited as the majority of the proposed deep learning models so far require large-scale training data. This large-scale data is not available for many target proteins and therefore no prediction models are available for these targets. In computer vision, transfer learning methods have been widely used to learn from a small amount of data. Transfer learning is a machine learning method where a model is trained for a source task and this pretrained model is reused as the starting point for another model on a target task. Transfer learning within the scope of drug-target interaction prediction has not been extensively studied so far [1].

In this study, we developed three transfer learning based methods to predict binding affinity values for compounds against proteins with low numbers of bioactivity data using MDeePred [2]. In MDeePred, multi-channel 2-D protein features and ECFP4 fingerprints are fed to the pairwise input neural network and binding affinity value is predicted for the given drug–target pair. In Method 1, the bottom few layers of the trained source model were frozen and not updated during back-propagation. In Method 2, the only difference from Method 1, a shallow regression model is used for training. Method 3 is full fine-tuning; learnable parameters of the source model are saved and the target model is trained using the saved parameters instead of initializing with random parameters. The training datasets and test datasets are generated from the ChEMBL database by applying the data filtering protocol that has been developed in our previous studies. Kinase family was selected to explore transfer learning on proteins with low number of bioactivity data which has nine subfamilies (see Table 1), and Tyrosine Kinase (TK) was selected as the source dataset, since it has the most number of drug–target interactions. In order to generate datasets containing lower numbers of drug-target interaction data points in a controlled manner, we employed Butina clustering algorithm. Four datasets were constructed where the number of bioactivities were 500, 250, 100, and 50, respectively. 5-fold cross validation is used to evaluate the models. Root mean square error (RMSE) is used to measure the performance of the models. As an example,

average RMSE values for 5-fold cross-validation are given in Table 1 for the 8 subfamilies having 100 drug-target interactions. Method 3 achieved better results for most of the subfamilies. The average RMSE values for Method 1 and Method 3 are better than the Baseline Method, therefore we can say that transfer learning results are promising. Results for the subfamilies that have 50 drug-target interactions are similar to Table 1. During our experiments, when transfer learning is used, training loss starts from very low values. Therefore, lower numbers of epochs were sufficient in training and the training time decreased significantly. We continue to investigate and explore transfer learning and low-shot learning for drug-target interaction prediction.

References

1. Lee,M. vd. 2019." Multi-channel PINN: investigating scalable and transferable neural networks for drug discovery", J. Cheminform., 11, 46.

2. Rifaioglu, A.S., Cetin Atalay, R., Cansen Kahraman, D., Doğan, T., Martin, M., Atalay, V. 2021. "MDeePred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery", Bioinformatics 37:693–704, https://doi.org/10.1093/bioinformatics/btaa858

Prediction of resistance to drugs in triple negative breast cancer based on gene expression levels

Bengisu Karaköse, Berk Gürdamar and Uğur Sezerman
Acıbadem Mehmet Ali Aydınlar University School of Medicine

Introduction: Triple-negative breast cancer is characterized by decreased expression of progesterone and estrogen receptors and the absence of HER-2. They develop in a shorter time than other breast cancers with a shorter survival time. Preoperative neoadjuvant chemotherapy is used in the treatment of locally advanced stages. Only 30-60% of tumors treated with neoadjuvant therapy achieve a complete response and prolong survival. In this study, we aim to detect transcriptomic differences in tumors that respond or not to neoadjuvant therapy, at the level of differentially expressed genes and affected pathways, and to develop drug resistance models by using locally advanced TNBC biopsy samples.

Methods: In this project, 17 biopsy samples taken from locally advanced TNBC patients and resistance information of 52 drugs were analyzed. Firstly, differential gene expression analysis was performed by using "DESeq2" tool with R software. A regular logistic regression technique, elasticnet model was established by using 10 genes with the lowest padj values. Secondly, separate models were established for all drugs with 10 differentially expressed genes. The accuracy, sensitivity, and specificity of the predictions were calculated, and Root Mean Squared Error value was indicated. Additionally, multitask learning analysis was performed using all differentially expressed genes. Pathway analyzes of differentially expressed genes were made with the pathfindR tool.

Results: Accuracy of predictions made with elasticnet models and multitask learning models were found as 0.80 and 0.69 respectively. Most upregulated genes were HSP1A, IL1A, PRKCB, FCAR, PLA2G4E and most downregulated genes were TUBB, CALML3, CACNA1S, MICB, COL1A1. Pathways that are mostly affected were Estrogen, MAPK and Ras Signaling Pathways, Antigen Processing and Presentation pathways.

Discussion: In the literature review, we saw that genes and pathways that are differentially expressed in our study were found to be important in drug resistance mechanisms in different cancers by other researchers as well. Examples of these are HSP1A, which plays a role in drug resistance in breast and prostate cancers, IL1A, which has a strong positive correlation with cancer stem-cell-positive phenotype, estrogen signaling pathway with increased expression in the chemotherapy-resistant luminal subtype of TNBC, MAPK signaling pathway which plays a role in drug resistance in mice with breast cancer. In the literature, it was seen that HER2 expression in breast cancer has been analyzed with lasso and elasticnet models, but no study has been done to analyze drug resistance of TNBC with the elasticnet model. Therefore, our work is pioneering. In conclusion, we desire to integrate clinical data with omic data and to implement these methods in medicine for personalized diagnosis and treatments.

Predicting oral health using machine learning

Emrah Kırdök[1] and Andres Aravena[2]

[1]Department of Biotechnology, Mersin University, Turkey

[2]Department of Molecular Biology and Genetics, Istanbul University, Turkey

The oral microbiome is the set of commensal species that reside in the oral cavity. It contains a diversity of bacterial species. In general, healthy oral microbiota is characterized by a balanced taxonomic composition. Changes in the taxonomic composition results in a microbial imbalance called dysbiosis. In a dysbiotic state, a group of bacterial species becomes abundant while others reduce their relative presence. Two well studied examples of oral dysbiotic states are periodontitis and caries. In these conditions, certain microbes colonise teeth, gums, and periodontal pockets inside the gums—as in periodontitis case.

These conditions are usually limited to the oral cavity. However, eventual ruptures in the capillary system could enable microbes to enter the circulatory system and colonise other parts of the body. Several studies show that it is possible to associate oral bacteria to systemic disease. Thus, understanding the oral dysbiosis could be used as an indicator for caries and periodontitis, as well as for some systemic diseases.

In this study, we aimed to predict the health condition of a person based on the taxonomic profile of their oral microbiota. We look for the microbial abundance signatures of each dysbiosis state. To do so, we used machine learning on previously-published oral microbiota data. We trained a set of supervised classifiers. Each training example, labeled as "healthy", "caries" or "periodontitis", is a vector with the relative abundances of each taxa. The components of these vectors are usually called "features". In this case the features are the relative abundances of each microbial species. In our training dataset there are many features and few examples, thus we devised a strategy to select a small subset of features that could be enough to train an accurate machine learning model.

We compiled DNA sequences from two studies of salivary samples. We identified the taxonomic composition of each DNA library using classic metagenomic classification methods. Each DNA library was represented by a vector of relative abundances for each microbial species.We trained two pools of classifiers: one separating healthy versus periodontitis cases, and another separating healthy versus caries cases. Our implementation used Random Forests classifiers, which —as the name suggests— include a stochastic selection of features. To avoid spurious results, we created a pool of classifiers, each one with a different random seed. We used the mean decrease of Gini impurity to determine the smallest set of features that predicted oral health with high probability. We discarded the features that do not contribute to the overall accuracy using an iterative feature elimination method. The minimal set of health-predictive species is consistent with the specialized literature. The best models predicted the correct health condition with an error rate under 14%.

Archaeogenetic analysis of Neolithic sheep from Anatolia

Erinç Yurtman[1], Onur Özer[1,2], Eren Yüncü[1], Nihan Dilşad Dağtaş[1], Dilek Koptekin[3], Yasin Gökhan Çakan[4], Mustafa Özkan[1], Ali Akbaba[5], Damla Kaptan[1], Gözde Atağ[1], Kıvılcım Başak Vural[1], Can Yümni Gündem[6], Louise Martin[7], Gülşah Merve Kılınç[1,8], Ayshin Ghalichi[1,9], Sinan Can Açan[1], Reyhan Yaka[1], Ekin Sağlıcan[1], Vendela Kempe Lagerholm[10], Maja Krzewinska[10], Torsten Günther[11], Pedro Morell Miranda[11], Evangelia Pişkin[12], Müge Şevketoğlu[13], C. Can Bilgin[1], Çiğdem Atakuman[12], Yılmaz Selim Erdal[14,15], Elif Sürer[16], N. Ezgi Altınışık[14,15], Johannes Lenstra[17], Sevgi Yorulmaz[1], Mohammad Foad Abazari[18], Javad Hoseinzadeh[19], Douglas Baird[20], Erhan Bıçakçı[4], Özlem Çevik[21], Fokke Gerritsen[22], Rana Özbal[23], Anders Götherström[10], Mehmet Somel[1], İnci Togan[1] and Füsun Özer[14,15]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey [2]Emmy Noether Group Evolutionary Immunogenomics, Max Planck Institute for Evolutionary Biology, Plön, Germany [3]Department of Health Informatics, Middle East Technical University, Ankara, Turkey [4]Department of Prehistory, Istanbul University, Istanbul, Turkey [5]Department of Anthropology, Ankara University, Ankara, Turkey [6]Department of Archaeology, Batman University, Batman, Turkey [7]Institute of Archaeology, University College London, London, UK [8]Department of Bioinformatics, Graduate School of Health Sciences, Hacettepe University, Ankara, Turkey [9]Department of Archaeogenetics, Max-Planck Institute for the Science of Human History, Jena, Germany [10]Archaeological Research Laboratory, Department of Archaeology and Classical Studies, University of Stockholm, Stockholm, Sweden [11]Department of Organismal Biology, Human Evolution Research Program, Uppsala University, Uppsala, Sweden [12]Department of Settlement Archaeology, Middle East Technical University, Ankara, Turkey [13]Centre for Archaeology, Cultural Heritage and Conservation, Cyprus International University, Haspolat, Cyprus [14]Department of Anthropology, Hacettepe University, Ankara, Turkey [15]Hacettepe University Molecular Anthropology Group (Human_G), Ankara, Turkey [16]Department of Modeling and Simulation, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey [17]Faculty of Veterinary Medicine, Utrecht University, Utrecht, Netherlands [18]Research Center for Clinical Virology, Tehran University of Medical Sciences, Tehran, [Iran19]Department of Archaeology, University of Kashan, Kashan, Iran [20]Department of Archaeology, Classics, and Egyptology, University of Liverpool, Liverpool, UK [21]Department of Archaeology, Trakya University, Edirne, Turkey [22]Netherlands Institute in Turkey, Istanbul, Turkey [23]Department of Archaeology and History of Art, Koç University, Istanbul, Turkey

Even though sheep was among the first domesticated animals, the demographic history of sheep has yet been little investigated using ancient DNA. In this study we investigated sheep domestication history by analysing single nucleotide polymorphism (SNP) and mitochondrial DNA (mtDNA) data obtained for Epipaleolithic and Neolithic period Anatolian sheep. We compared these data with published Neolithic and Bronze Age sheep genomes from central Asia and with modern-day breeds. We found that, Anatolian Neolithic sheep were genetically distinct from all modern breeds, with genetic affinity to present-day European breeds than to Asian (including southwest Asian) breeds. In contrast, central Asian sheep showed higher genetic affinity to present-day Asian breeds. These results suggest that east-west genetic structure of the present-day sheep breeds had already been emerged by 6000

BCE. This also indicate that there may have been multiple sheep domestication events with different centers in Western and Central Asia, or alternatively, early introgression of wild sheep in southwest Asia. Overall, we conclude that the gene pools of European and Anatolian domestic sheep have been strongly reshaped since the Neolithic.

Potential inhibitor identification for deoxyhypusine synthase

Ayşenur Öztürk and Fethiye Aylin Sungur
Computational Science and Engineering Division, Informatics Institute, Istanbul Technical University, Istanbul, Turkey

Deoxyhypusine synthase (DHPS) is one of the major proteins in hypusination pathway. Malfunctions of the hypusination pathway, such as those produced by mutations in the DHPS encoding gene, have been linked to diseases like cancer and neurodegeneration. There are known inhibitors of DHPS, but the search for more efficient inhibitors continues. In this study, known inhibitor structures were taken as the starting point for these new inhibitors. Our aim is finding candidate drug molecules for our target DHPS. The binding scores obtained by docking with the molecules in the ligand library created from Zinc15 and PubChem databases[1,2]. The created library follows the Lipinski Rule of Five. The crystal structure of DHPS in complex with 6-[(2R)-1-Amino-4-Methylpentan-2-yl]-3-(pyridin-3-yl)-4H,5H,6H,7H-thieno[2,3- C]pyridin-7-one at 2.12 Å resolution is retrieved from the Protein Data Bank (PDB ID: 6WL6) [3]. Molecular docking of 6WL6 against the candidates is performed using AutoDock Vina software. First, rigid docking will be done and then flexible docking work will be done according to amino acid interactions. Compounds with the highest vina docking scores down to 15selected. Then MM-GBSA calculations is performed for the selected candidates. (Scheme 1). In order to design pharmacophores, energy decomposition calculations per residue is done to figure out key binding residues stabilizing the ligands. From the results, the leading molecules were selected to be potential inhibitors of DHPS for further experimental studies.

References

1. Sterling and Irwin, J. Chem. Inf. Model, 2015 http://pubs.acs.org/doi/abs/10.1021/acs.jcim.5b00559.

2. RiKim S, Chen J, Cheng T, et al. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res. 2021;49(D1):D1388–D1395. doi:10.1093/nar/gkaa971.

3. Tanaka Y, Kurasawa O, Yokota A, Klein M, Saito B & Matsumoto S et al. (2020). New Series of Potent Allosteric Inhibitors of Deoxyhypusine Synthase. ACS Medicinal Chemistry Letters, 11(8), 1645-1652. doi: 10.1021/acsmedchemlett.0c00331.

Inter-tissue convergence of gene expression and loss of cellular identity during ageing

Hamit İzgi[1], DingDing Han[2,+], Ulas Isildak[1], Shuyun Huang[2+], Ece Kocabiyik[1], Philipp Khaitovich[3], Mehmet Somel[1] and Handan Melike Dönertaş[4]

[1]Department of Biological Sciences, Middle East Technical University, Ankara, Turkey
[2]CAS Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai, China
[3]Center for Neurobiology and Brain Restoration, Skolkovo Institute of Science and Technology, Moscow, Russia
[4]European Molecular Biology Laboratory, European Bioinformatics Institute EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, UK
[+]Present Address: Department of Clinical Laboratory, Shanghai Children's Hospital, Shanghai Jiaotong University, Shanghai, China

Studying gene expression changes during ageing gives insight to identifying age-related molecular and cellular processes. Recent molecular studies including these two periods have reported a reversal of the ageing transcriptome towards pre-adult levels in primate brain and mouse liver and kidney. Several major questions remain to be answered. Prevalence of reversal phenotypes across tissues is yet unclear and is not studied extensively, as most research has been conducted in the brain. To address this, we generated RNA-seq transcriptomes of 16 mice from four tissues; cortex, lung, liver and muscle, covering development and ageing intervals. First, we revealed that approximately %50 of the genes showed expression reversals in each tissue, although these proportions are not significant in permutation tests, suggesting that the expression trajectories of the genes do not necessarily continue linearly into the ageing period. Functional analysis of the genesets showing reversal patterns identified pathways related to development, metabolism and inflammation. Next, we asked whether different tissues show similarities in their reversal patterns. We found no significant overlap in reversal genes among tissues suggesting that expression reversals might be tissue-specific. In concordance with the tissue-specific reversals, we showed that during development, tissues become more distinct (diverge) and interestingly during ageing, tissues become more similar (converge) in their gene expression levels. We confirmed this observation using two other independent RNA-seq (human) and microarray datasets (mouse). Moreover, divergence-convergence pattern is enriched among tissue-specific genes which show either decreased expression in native tissue or gaining expression in non-native tissue during ageing. Further, using publicly available single-cell transcriptome data, we showed that divergence-convergence pattern is driven both by alterations in cell type proportions and also by cellautonomous expression changes. This supports our previous hypothesis that loss of cellular identity during ageing might be a general phenomenon in mammalian ageing.

Ensemble learning approach for computational drug repurposing

Ismail Denizli, Oguzhan Sahin, Ozgur Dogan, Tugba Suzek and Baris Suzek
Muğla Sıtkı Koçman Universitesi

Drug development carries risks such as large sampling of possible protein-molecule pairs, failure of experimental processes (e.g, toxicity), and adverse effects in clinical trials of successful pairs. These risks cause serious time and financial losses for pharmaceutical companies. In recent years, pharmaceutical companies have resorted to various methods in order to reduce their risks, reduce development costs and improve their processes. One of these methods is drug repurposing. With the help of drug repurposing, drugs that have been tested for safety in humans and available in the market in the treatment of a disease that was not intended for the original disease they were designed for.

Many computational approaches to predict drug-target interactions(DTI) for drug repurposing have been recently proposed in literature. Convolutional neural networks based approaches, DeepDTA and DeepConvDTI, learn the semantic meanings of the characters that make up the sequences by backpropagation with a deep network-based architecture, and compose these semantic meanings with different layers, creating protein and molecule representations and classifying them. The most recent work in this field is a study called MolTrans, which includes novel techniques that model interactions between sub-segments of peptide and molecule sequences.

In this study, we investigated whether the classification performance metrics increase by combining different models on a combined data set, namely BindingDB, DAVIS, BIOSNAP. BindingDB has 32,601 DTI pairs, The DAVIS has 11,103 DTI pairs. BIOSNAP has 18,690 DTI pairs. The combined dataset contains 11724 drugs, 3067 proteins and 63605 DTI pairs. The DeepDTA, DeepConv-DTI and MolTrans models were used with different ensemble techniques such as stacking, ensembling, voting. The combined set is divided into slices for training(70% of all data), testing (20% of all data), and validation (10% of all data). The predictions of DeepDTA, DeepConv-DTI and MolTrans models were used as inputs to the ensemble models. The hyperparameters of the ensemble model were learned from the validation set, and ensemble results were obtained for the test set. The ensemble model improved the single best model by a 2.46%, 1.8%, 1.5%, 0.9% increase in AUPRC, ACCURACY, PRECISION, AUROC metrics respectively, with a promise of reducing the margin of error of ensemble learning in in-silico drug repurposing.

Extraction of herb-drug interactions

Erkan Yaşar[1], Remzi Çelebi[2] and Özgür Gümüş[1]
[1]Ege University
[2]Maastricht University

Drugs taken for therapeutic purposes may cause unexpected adverse reactions when taken with other drugs or herbs. These interactions can pose serious health risks to patients and even cause death. In particular, cancer patients may have greater risk to their health when they consume various herbs in combination with their cancer medications without consulting their doctors. It is critical for a patient's health to identify such drug-herb interactions before a risk occurs.

Although there have been many studies on predicting drug interactions such as drug-drug interaction prediction, drug-target interaction prediction, there are only a few studies on drug-herb interaction prediction and no comprehensive database is available for research. These interactions are mostly reported on literature sources such as PubMed on the web. Creating a comprehensive drug-herb interaction database can be used for the prediction of future interactions.

The process of creating such a database is a time-consuming and costly process due to the fact that it requires expert knowledge. In this work, we aimed to construct a comprehensive database of drug-herb interaction using NLP techniques (i.e., entity recognition model). For this purpose, first of all, the drug-herb interactions listed in the Memorial Sloan Kettering (MSKCC) website were retrieved and their literature references were automatically converted to a corpus of text containing herbs and drugs. Then, the herb and drug mentions in these retrieved texts are recognized using the Named Entity Recognition (NER) model. An annotated corpus was generated and trained with the Prodigy annotation tool. The trained model achieves 0.84 accuracy. We hope this model can serve as a basis for future herb-drug relation extraction studies and prediction of yet-unknown herb-drug interactions; it will be re-used for many diverse objectives.

Identification of autophagy-related miRNA–mRNA regulatory network in calorie-restricted mouse brain

Atakan Ayden[1], Elif Yılmaz[1], Bilge G. Tuna[2], Ayşegül Kuskucu[3], Ömer F. Bayrak[3], Andrés Aravena[4] and Soner Doğan[5]

[1]Department of Biotechnology, Yeditepe University, Istanbul, Turkey
[2]Department of Biophysics, Yeditepe University, Istanbul, Turkey
[3]Department of Genetics, Yeditepe University, Istanbul, Turkey
[4]Department of Molecular Biology and Genetics, Istanbul University, Turkey
[5]Yeditepe University

In recent years, research on autophagy as a key regulator of neurodegeneration has increased. It was primarily investigated in the brain in neurons, where the transport of hazardous chemicals and organelles to the lysosome via autophagy is critical for neuronal health and survival.

Caloric restriction is an anti-aging regimen that stimulates autophagy and has been shown to extend longevity in various organisms. Caloric restriction (CR) is implemented in experimental models to lower calorie intake without causing a nutritional deficiency. There are two types of calorie restriction commonly used: chronic calorie restriction and intermittent calorie restriction. A complete understanding of miRNA expression change following calorie restriction could reveal how calorie restriction prevents neurodegeneration via autophagy.

In the current study, ten-week-old female C57BL/6 mice were separated into three groups and fed three diets: ad libitum, chronic calorie restriction (15% CR compared to ad libitum group), or intermittent calorie restriction (three weeks of ad libitum diet followed by one week 60% CR).

We sacrificed mice at week 49/50 and extracted RNA samples from the brain tissue of these mice. Afterwards, we used a miRNA microarray to evaluate the changes in miRNA expression levels in the brain in response to different calorie-restricted diets.

In total, 25 miRNA showed a significant differential expression in all dietary groups. Among them, seventeen are down-regulated and eight are up-regulated. We predicted their target genes using miRNAtap, which uses rank aggregation from five different prediction algorithms. We also predicted their target lncRNAs using LncBase.

These tools predicted 5954 target genes and 2566 lncRNAs associated with the differentially expressed miRNAs. In the target analysis, we identified miRNA, lncRNA, and target gene pairs related to autophagy. Afterwards, we constructed interaction networks of upregulated and downregulated miRNAs using STRING and Cytoscape. We used mirPath to identify the pathways enriched in differentially expressed miRNAs targeting autophagy-related genes.

Based on our findings, we propose a putative autophagy-related ncRNA-mRNA regulation network, which can be helpful for investigating novel pathways affecting brain health in aged organisms.

This work was financially supported by the Scientific and Technological Research Council of Turkey (TUBITAK, 119S238).