Łukasiewicz
PORT
Polish Center
for Technology
Development

ISBRA 2023

19th International Symposium on Bioinformatics Research and Applications

www.isbra2023.port.org.pl

ISBRA 2023

19th International Symposium on Bioinformatics Research and Applications

The International Symposium on Bioinformatics Research and Applications (ISBRA) provides a forum for the exchange of ideas and results among researchers, developers, and practitioners working on all aspects of bioinformatics and computational biology and their applications. Submissions presenting original research are solicited in all areas of bioinformatics and computational biology, including the development of experimental or commercial systems. Topics of interest include but are not limited to:

- Biomarker discovery
- Biomedical databases and data integration
- Biomedical text mining and ortologies
- Biomolecular imaging
- Comparative genomics
- Computational genetic epidemiology
- Computational proteomics
- Data mining and visualization
- Gene expression analysis
- Structural biology
- Genome analysis
- Systems biology

Łukasiewicz
PORT
Polish Center
for Technology
Development

Łukasiewicz – PORT Polish Center for Technology Development is a scientific and research institute. Our dedicated team of scientists engages in both fundamental research and the development of innovative technologies for various industries.

The institute's scientific and research endeavors are organized into three centers: Life Sciences and Biotechnology (LS&BC), Population Diagnostics (CPD) and Material Sciences and Engineering (MS&EC). These centers house specialized Research Groups, supported by state-of-the-art measurement laboratories and core facilities. Łukasiewicz – PORT also houses a biobank, Polish National Node (PNN) of the BBMRI-ERIC network.

The Life Science & Biotechnology Center is a forward-thinking institution with a strong focus on research, development, and implementation. It addresses significant societal challenges and cultivates expertise in neurobiology, oncology, and broad biotechnology applications.

Located within the historical Pracze Campus, Łukasiewicz – PORT features architecturally impressive buildings dating back to the turn of the 19th and 20th centuries, enveloped by lush green surroundings. Within these brick walls, modern laboratories equipped with world-class facilities enables both applied research projects and fundamental scientific investigations.

Since 2019, our institute has been a proud member of the Łukasiewicz Research Network, the third-largest research network in Europe. With a collaborative network spanning over 20 institutes across Poland, our shared objective is to foster stronger connections between academia and the business sector.

Committee

Steering Committee

Dan Gusfield *UC Davis, USA*

Ion Mandoiu *UConn, USA*

Yi Pan *SIAT, China and GSU, USA*

Marie-France Sagot *INRIA, France*

Zhirong Sun *Tsinghua, China*

Ying Xu *UGA, USA*

Aidong Zhang *SUNY, USA*

Zhipeng Cai *GSU, USA*

General Co-Chairs

Michal Malewicz *PORT, Poland*

Alexander Zelikovsky *GSU, USA*

Vice General Chair

Tianwei Yu *The Chinese University of Hong Kong – Shenzhen*

Program Co-Chairs

Xuan Guo *UNT, USA*

Murray Patterson *UConn, USA*

Serghei Mangul *University of Southern California (USC), USA*

Publicity Chairs

Olga Glebova *GSU, USA*

Local Arrangement Chair

Michal Malewicz *PORT, Poland*

Web Chair

Grigore Boldirev *GSU, USA*

Keynote Speakers

Teresa Przytycka

Senior Investigator
Algorithmic Methods in Computational and Systems Biology (AlgoCSB)
NLM/NCBI

Teresa Przytycka is a Senior Investigator and the Chief of the Computational Biology Branch at the National Center for Biotechnology Information (NCBI), which is part of the National Institutes of Health (NIH) in the United States. The research in her group focuses on developing computational methods for systems biology including application to cancer research and gene regulation and methods for analysis of new types of experimental data. She has worked on various topics, including protein structure prediction, functional annotation of genes, network biology, and cancer genomics. Throughout her career, Teresa Przytycka has published numerous scientific articles in reputable journals and has been actively involved in the bioinformatics research community. She has also served on program committees and advisory boards for several conferences and organizations related to bioinformatics and computational biology.

Abstract

Delineating relation between mutagenic signatures, cellular processes, and environment through computational approaches

Cancer genomes accumulate many somatic mutations resulting from carcinogenic exposures, cancer related aberrations of DNA maintenance machinery, and normal stochastic events. These processes often lead to distinctive patterns of mutations, called mutational signatures. However interpreting mutation patterns represented by such signatures is often challenging. This talk will focus on computational methods to elucidate the relations between mutational signatures and cellular and environmental processes developed in my group. In particular, I will discuss computational methods to untangle the contributions of DNA damage and repair processes to mutation signatures and network based approaches to uncover the interactions between mutational signatures and biological processes.

Anna Gambin

Faculty of Mathematics, Informatics and Mechanics
University of Warsaw

Anna Gambin is a Professor at the Faculty of Mathematics, Informatics and Mechanics at the University of Warsaw. Her research interests include computational molecular biology and medicine: bioinformatics for genome rearrangements, algorithms for mass spectra processing, mathematical modelling of signaling pathways, comparative genomics of transposable elements.

Abstract

Statistical modeling in proteomics

In this talk, I present a series of related topics in the mathematical and computational modeling of mass spectrometry data. The presentation opens with a presentation of two algorithms for the generation of the isotopic structure (BRAIN and ISOSPEC). Then I briefly touch on the problem of understanding electron-driven reactions, whose principal aim is to induce ion fragmentation. Here I apply the mathematical theory of reaction kinetics to estimate the reaction rates of the electron transfer reactions. Last, I present a breakthrough idea to use the mathematical theory of optimal transport for molecule identification.

Sagi Snir

University of Haifa
Israel

Sagi Snir is a Professor of computational evolution at the University of Haifa in Israel, where he has established and headed the Bioinformatics program for grad students. He is also the President of the Israeli Society for Bioinformatics and Computational Biology. His research combines algorithmic and combinatorial approaches to problems from evolution with focus on phylogenetic trees and networks. He has developed the Quartet MaxCut algorithm to combine conflicting signals between evolutionary trees and other fundamental results on maximum likelihood of trees and network. His papers have been published in both leading pure theoretical computer science venues and pure evolution venues. Ha has organised major national and international conferences at the University of Haifa.

Abstract

Assembling the Tree of Life in Light of Conflicting Signals

Mark Robinson

University of Zurich
Switzerland

Mark Robinson has been an Associate Professor since 2017 after joining the Department of Molecular Life Sciences of the University of Zurich (UZH) in 2011. He studied Applied Mathematics (BSc, Uni. Guelph) and Statistics (MSc, Uni. British Columbia), and did a PhD in statistical bioinformatics at the University of Melbourne. He has predoctoral experience at the Banting and Best Department of Medical Research (Uni Toronto) and postdoctoral experience in Cancer Epigenomics at the Garvan Institute in Sydney. The Robinson group at UZH develops statistical methods for interpreting high-throughput sequencing and other genomics technologies in the context of genome sequencing, gene expression and regulation and analysis of epigenomes, with a current focus on the analysis of single-cell and spatial datasets.

Abstract

On the care and feeding of (computational method) benchmarks

In an increasingly data-centric world, especially in biomedical data science, there has been an explosion of competing computational methods to process and model large-scale molecular datasets, and thus it is critically important to know their relative merits (e.g., performance). By introducing broader patterns of open science practices, including open data and especially open research code, we aim to make method performance assessments (benchmarks) more accessible and impactful to the community. This talk will take a deep look at how we currently do benchmarks, highlighting areas where we can improve as a community. Furthermore, our initiative OMNIBENCHMARK, provides a clear path forward to drastically improve all scientific aspects of benchmarking via the "open continuous benchmarkization" concept. We are poised to improve not only the efficiency and transparency of (computational) methods research, but also the robustness and usability of state-of-the-art methods for various end-users.

Ana Teresa Freitas

University of Lisbon
Portugal

Ana Teresa Freitas is a Full Professor at the Department of Computer Science and Engineering at Técnico Lisbon (IST), University of Lisbon, where she teaches the course of Computational Biology. Additionally, she serves as the Chairwoman of the School Assembly at IST and the Coordinator of the disciplinary field of Programming Methodology and Technology. As a senior researcher at the INESC-ID research institute, she also leads the „Life and Health Technology" thematic line. Currently, she holds the position of Head of Node of the ESFRI ELIXIR in Portugal. Previously, from 2013 to 2022, she was the CEO and co-founder of HeartGenetics, Genetics, and Biotechnology SA, a digital health and genetic testing company. She holds a PhD in Computer Engineering, a Master's degree in Computer and Electronics Engineering, and a degree in Innovation and Entrepreneurship. Her scientific expertise spans various areas, including Bioinformatics, Computational Biology, Human Genetics, Health Informatics, Algorithms, Data Mining, and AI.

Abstract

Turning Data into Genomic Medicine

In the realm of modern healthcare, the integration of genomics into medical practice holds the promise of revolutionizing the way we prevent, diagnose, and treat diseases. This talk, titled „Turning Data into Genomic Medicine," delves into the challenges and opportunities of harnessing the power of genomics for the betterment of public health.

I will begin by examining the hurdles faced when implementing genomic medicine, from ethical concerns to data privacy issues, and the need for robust infrastructure. One focal point will be the European Union's pioneering 1+ Million Genomes (1+ MG) initiative, which aims to amass a vast wealth of genomic data from diverse populations. I will also explore the follow-up B1MG project, designed to enhance the translation of this genomic treasure trove into tangible clinical benefits.

Additionally, I will shed light on the transformative Genomic Data Infrastructure (GDI) project, highlighting its pivotal role in securely storing, managing, and analyzing the burgeoning genomic datasets, thus enabling precision medicine on a previously unimaginable scale.

Moreover, this talk will emphasize the applications of genomic medicine in real-world scenarios, from tailoring treatments to an individual's genetic makeup to predicting disease susceptibility. A special focus will be placed on the use of polygenic risk scores, a cutting-edge tool that enables more accurate disease prevention strategies based on an individual's genetic predisposition.


Cancer/Health

Neurogenesis-associated Protein, a Potential Prognostic Biomarker in anti-PD-1 based kidney renal clear cell carcinoma patients therapeutics

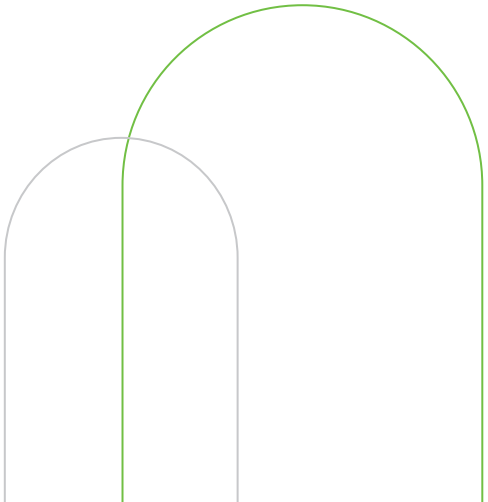
Rui Gao, Zixue Liu, Mei Meng, Jian He

Background: TKTL1 is an essential factor that has been found to perform an important role in brain development. Some studies have shown the influence of TKTL1 in cancers, but it is rarely reported in kidney cancer. Furthermore, the relationship of TKTL1 to prognosis potential and tumor infiltration immune cells in different cancers, especially kidney cancer, is still unclarified. Methods: TKTL1 expression and its clinical characteristics were evaluated on various databases. Also, the correlation between TKTL1 and TILs in the tumor and normal adjacent tissue of three types of renal patients respectively by using various types of bioinformatics approaches. Nextly, the association between TKTL1 and immune infiltrates of various types of cancer was investigated via TIMER. Furthermore, we studied the relationship between TKTL1 expression and response to PD-1 blocker immunotherapy in renal cancer and performed molecular docking to screen TKTL1 agonists. Results: We constructed a systematic prognostic landscape in various types of cancer and enclosed that TKTL1 significantly affects the prognostic potential in different types of kidney cancer patients. And the underlying mechanism might be the expression level of TKTL1 was positively associated with devious immunocytes in kidney renal clear cell carcinoma (KIRC) rather than in kidney renal papillary cell carcinoma (KIRP) and kidney chromophobe (KICH). Moreover, this recruitment may result from the upregulation of the mTOR signaling pathway affecting T-cell metabolism. We also found that TKTL1 might be an immunomodulator in KIRC patients' response to anti-PD-1 therapy. Finally, we found that 3-hydroxyflavone demonstrated a potential TKTL1 agonist compared to other flavonoids. Conclusions: Our discovery implies that TKTL1 appears to be a promising prognostic biomarker for KIRC patients that response to anti-PD-1 therapy. Moreover, flavonols might be a potential therapeutic combination to anti-PD-1-based immunotherapy.

Deep Learning Reveals Biological Basis of Racial Disparities in Quadruple-Negative Breast Cancer

Bikram Sahoo, Alex Zelikovsky

Triple-negative breast cancer (TNBC) lacks crucial receptors. More aggressive is quadruple-negative (QNBC), which lacks androgen receptors. Racial disparities emerge, with African Americans facing worse QNBC outcomes. Our study deploys deep neural networks to identify QNBC ancestral biomarkers. Achieving 0.85 accuracy and 0.928 AUC, the model displays robust learning, optimized through hyperparameter tuning. Top genes are chosen via ANOVA rankings and hypothesis testing, highlighting\emph{ABCD1} as significant post-correction. Effect sizes suggest important shifts in other genes. This approach enhances QNBC understanding, particularly racial aspects, potentially guiding targeted treatments.



Exploring Racial Disparities in Triple-Negative Breast Cancer: Insights from Feature Selection Algorithms

Bikram Sahoo, Alex Zelikovsky

Triple-negative breast cancer (TNBC) represents an aggressive and heterogeneous form of breast cancer with poor clinical outcomes. It lacks estrogen, progesterone, and human epidermal growth factor receptor, which limits treatment options. Notably, the incidence of TNBC is higher in African American (AA) women compared to European American (EA) women, resulting in worse clinical outcomes. The racial disparity observed in TNBC can be attributed to socioeconomic factors, lifestyle, and tumor biology. In this study, we explored feature selection algorithms, including filters, wrappers, and embedded methods, to identify significant genes associated with racial disparities. Our findings reveal that genes such as LOC90784, LOC101060339, XRCC6P5, and TREML4 were consistently selected by both correlation and information gain-based filter methods. Moreover, in our two-stage embeddedbased feature selection algorithm, we consistently identified LOC90784, STON1-GTF2A1L, and TREML4 as crucial genes across high-performing machine learning algorithms. Particularly noteworthy is the consistent selection of LOC90784 by all three filter selection methods. These comprehensive results, obtained through the implementation of three different feature selection algorithms, offer valuable insights to researchers studying racial disparities.

HetBiSyn: Predicting Anticancer Synergistic Drug Combinations Featuring Bi-perspective Drug Embedding with Heterogeneous Data

Yulong Li, Hongming Zhu, Xiaowen Wang, Qin Liu

Synergistic drug combination is a promising solution to cancer treatment. Since the combinatorial space of drug combinations is too vast to be traversed through experiments, computational methods based on deep learning have shown huge potential in identifying novel synergistic drug combinations. Meanwhile, the feature construction of drugs has been viewed as a crucial task within drug synergy prediction. Recent studies shed light on the use of heterogeneous data, while most studies make independent use of relational data of drug-related biomedical interactions and structural data of drug molecule, thus ignoring the intrinsical association between the two perspectives. In this study, we propose a novel deep learning method termed HetBiSyn for drug combination synergy prediction. HetBiSyn innovatively models the drug-related interactions between biomedical entities and the structure of drug molecules into different heterogeneous graphs, and designs a self-supervised learning framework to obtain a unified drug embedding that simultaneously contains information from both perspectives. In details, two separate heterogeneous graph attention networks are adopted for the two types of graph, whose outputs are utilized to form a contrastive learning task for drug embedding that is enhanced by hard negative mining. We also obtain cell line features by exploiting gene expression profiles. Finally HetBiSyn uses a DNN with batch normalization to predict the synergy score of a combination of two drugs on a specific cell line. The experiment results show that our model outperforms other state-of-art DL and ML methods on the same synergy prediction task. The ablation study also demonstrates that our drug embeddings with bi-perspective information learned through the end-to-end process is significantly informative, which is eventually helpful to predict the synergy scores of drug combinations.

Identification and functional annotation of circRNAs in neuroblastoma based on bioinformatics

Jingjing Zhang, Md. Tofazzal Hossain, Zhen Ju, Wenhui Xi, Yanjie Wei

Neuroblastoma is a prevalent solid tumor affecting children, with a low 5-year survival rate in high-risk patients. Previous studies have shed light on the involvement of specific circRNAs in neuroblastoma development. However, there is still a pressing need to identify novel therapeutic targets associated with circRNAs. In this study, we performed an integrated analysis of two circRNA sequencing datasets, the results revealed dysregulation of 36 circRNAs in neuroblastoma tissues, with their parental genes likely implicated in tumor development. In addition, we identified three specific circRNAs, namely hsa_circ_0001079, hsa_circ_0099504, and hsa_circ_0003171, that exhibit interaction with miRNAs, modulating the expression of genes associated with neuroblastoma. Additionally, by analyzing the translational potential of differentially expressed circRNAs, we uncovered seven circRNAs with the potential capacity for polypeptide translation. Notably, structural predictions suggest that the protein product derived from hsa_circ_0001073 belongs to the TGF-beta receptor protein family, indicating its potential involvement in promoting neuroblastoma occurrence.

Identifying miRNA-disease Associations based on Simple Graph Convolution with DropMessage and Jumping Knowledge

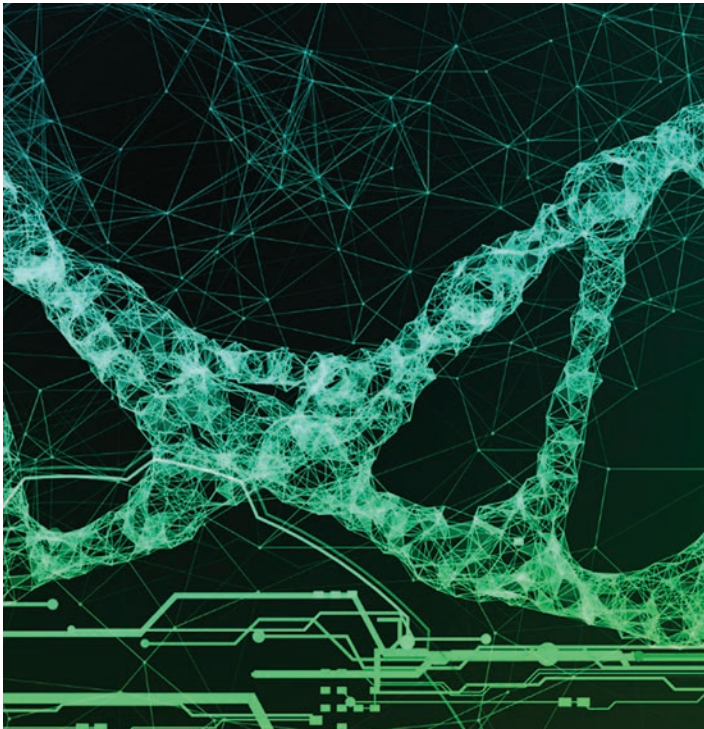
Xuehua Bi, Chunyang Jiang, Cheng Yan, Kai Zhao, Linlin Zhang, Jianxin Wang

MiRNAs play an important role in the occurrence and development of human disease. Identifying potential miRNA-disease associations is valuable for disease diagnosis and treatment. Therefore, it is very urgent to develop efficient computational methods for predicting potential miRNA-disease associations in order to reduce the cost and time associated with biological wet experiments. In addition, although the good performance achieved by graph neural network methods for predicting miRNA-disease associations, they still face the risk of oversmoothing and have room for improvement. In this paper, we propose a novel model named nSGC-MDA, which employs a modified Simple Graph Convolution (SGC) to predict the miRNA-disease associations. Specifically, we first construct a bipartite attributed graph for miRNAs and diseases by computing multi-source similarity. Then we adapt SGC to extract the features of miRNAs and diseases on the graph. To prevent over-fitting, we randomly drop the message during message propagation and employ Jumping Knowledge (JK) during feature aggregation to enhance feature representation. Furthermore, we utilize a feature crossing strategy to get the feature of miRNA-disease pairs. Finally, we calculate the prediction scores of miRNA-disease pairs by using a fully connected neural network decoder. In the five-fold cross-validation, nSGC-MDA achieves a mean AUC of 0.9502 and a mean AUPR of 0.9496, outperforming six compared methods. The case study of cardiovascular disease also demonstrates the effectiveness of nSGC-MDA.

Phylogenetic Information as Soft Constraints in RNA Secondary Structure Prediction

Sarah von Loehneysen, Thomas Spicher, Yuliia Varenyk, Hua-Ting Yao, Ronny Lorenz, Ivo Hofacker, Peter F. Stadler

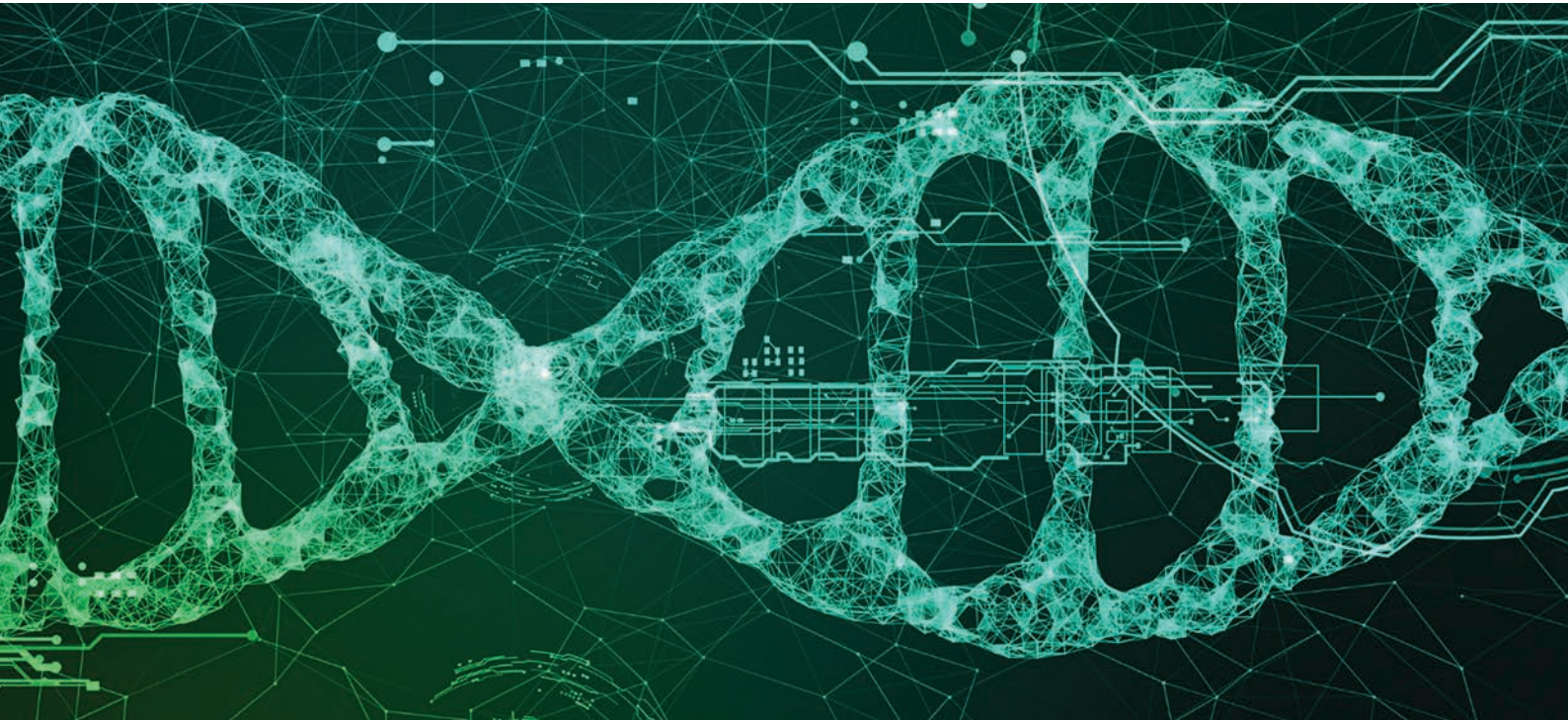
Pseudo-energies are a generic method to incorporate extrinsic information into energy-directed RNA secondary structure predictions. Consensus structures of RNA families, usually predicted from multiple sequence alignments, can be treated as soft constraints in this manner. In this contribution we first revisit the theoretical framework and then show that pseudo-energies for the centroid base pairs of the consensus structure result in a substantial increase in folding accuracy. In contrast, only a moderate improvement can be achieved if only the information that a base is predominantly paired is utilized.



The effect of transcriptomic annotations in breast cancer DGE study

Rafał Stępień, Joanna Szyda, Bartosz Czech, Magda Mielczarek

Gene expression profiling is crucial for understanding breast cancer biology and treatment individualization. The aim of this study was to elucidate the transcriptome annotation effect on differential gene expression (DGE) and breast cancer survival prognosis. DGE analyses were performed for MCF7 breast cancer (case) and normal tissues (control). The pipeline comprised quality control, quality-based data editing, transcript expression quantification, and DGE analysis. Two quantified transcripts expression outputs were used to apply four approaches defining DGE between (A1) case and control samples quantified based on GRCh37 assembly, (A2) case and control on GRCh38, (A3) case on GRCh37 and case on GRCh38 and (A4) control on GRCh37 and control on GRCh38. Identical Hallmark pathways resulted in Gene Set Enrichment Analysis for both A1 and A2, except Pancreas beta cells presented in A1 only. The Kyoto Encyclopedia of Genes and Genomes pathways presented only in one approach involved: Melanoma and Prostate cancer (A1) and ABC transporters, Acute myeloid leukaemia, Glycerophospholipid, and retinol metabolism, Hedgehog and p53 signalling (A2). Principal Component Analysis determined that the greatest variability (97%) was found between cancer and normal samples (A1, A2) and GRCh37 and GRCh38 annotations (A3). For A4 the variability determined by the annotations was lower (40%). The difference between the average expression of prognostic genes associated with survival in breast cancer (NADERI) between GRCh37 and GRCh38 was not statistically significant (P-value=0.91). The overall DGE outcomes were not identical between GRCh37 and GRCh38 annotations, however, the transcriptome annotation had no effect on survival prognosis in breast cancer.



The Ordered Covering Problem in Distance Geometry

Michael Souza, Nilton Maia, Carlile Lavor

This study is motivated by the Discretizable Molecular Distance Geometry Problem (DMDGP), a specific category in Distance Geometry, where the search space is discrete. We address the challenge of ordering the DMDGP constraints, a critical factor in the performance of the state-of-the-art SBBU algorithm. To this end, we formalize the constraint ordering problem as a vertex cover problem, which diverges from traditional covering problems due to the substantial importance of the sequence of vertices in the covering. In order to solve the covering problem, we propose a greedy heuristic and compare it to the ordering of the SBBU. The computational results indicate that the greedy heuristic outperforms the SBBU ordering by an average factor of 1,300x.

Approximating Rearrangement Distances with Replicas and Flexible Intergenic Regions

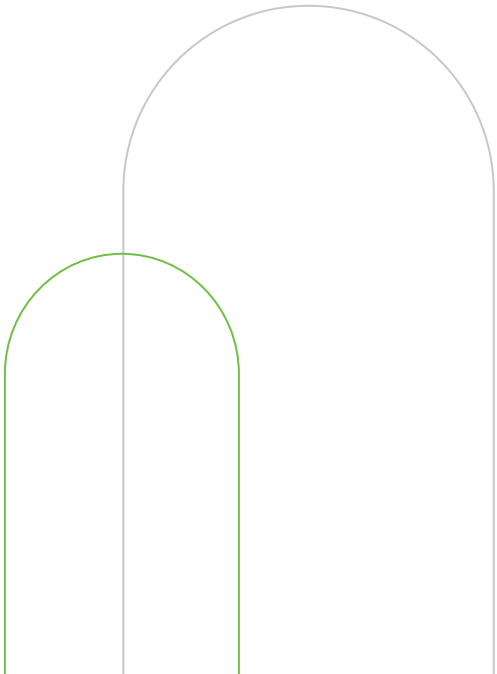
Gabriel Siqueira, Alexsandro Oliveira Alexandrino, Andre Rodrigues Oliveira, Géraldine Jean, Guillaume Fertin, Zanoni Dias

Many tools from Computational Biology compute distances between genomes by accounting the number of genome rearrangement events, such as reversals of a segment of genes. Most approaches to model these problems consider some simplifications such as ignoring nucleotides outside genes (the so-called intergenic regions), or assuming that just a single copy of each gene exists in the genomes. Recent works made advancements in more general models, considering replicated genes and intergenic region information. Our work aims at adapting those results by applying some flexibilization to the representation of intergenic region information. We propose the Signed Flexible Intergenic Reversal Distance problem, which seeks the minimum number of reversals necessary to transform one genome into the other and encodes the genomes using flexible intergenic region information while also allowing multiple copies of a gene. We show the relationship of this problem with the Signed Minimum Common Flexible Intergenic String Partition problem and use a 2k-approximation to the partition problem to show a 8k-approximation to the distance problem, where k is the maximum number of copies of a gene in the genomes.

On Computing the Jaro Similarity Between Two Strings

Joyanta Basak, Ahmed Soliman, Nachiket Deo, Kenneth Haase, Anup Mathur, Krista Park, Rebecca Steorts, Daniel Weinberg, Sartaj Sahni, Sanguthevar Rajasekaran

Jaro similarity is widely used in computing the similarity (or distance) between two strings of characters. For example, record linkage is an application of great interest in many domains for which Jaro similarity is popularly employed. Existing algorithms for computing the Jaro similarity between two given strings take quadratic time in the worst case. In this paper, we present an algorithm for Jaro similarity computation that takes only linear time. We also present experimental results that reveal that our algorithm outperforms existing algorithms.



Using Generating Functions to Prove Additivity of Gene-Neighborhood Based Phylogenetics

Guy Katriel, Udi Mahanaymi, Christoph Koutschan, Doron Zeilberger, Mike Steel, Sagi Snir

Prokaryotic evolution is often described as the Spaghetti of Life due to massive genome dynamics (GD) events of gene gain and loss, resulting in different evolutionary histories for the set of genes comprising the organism. These different histories, dubbed as gene trees provide confounding signals, hampering the attempt to reconstruct the species tree describing the main trend of evolution of the species under study. The synteny index (SI) between a pair of genomes combines gene order and gene content information, allowing comparison of unequal gene content genomes, together with order considerations of their common genes. Recently, GD has been modelled as a continuous-time Markov process. Under this formulation, distance between genes along the chromosome, was shown to follow a birth-death-immigration process. Using classical results from birth-death theory, we recently showed that the SI measure is consistent under that formulation. In this work we provide an alternative, stand-alone combinatorial proof of the same result. By using generating function techniques we derive explicit expressions of the system's probabilistic dynamics in the form of rational functions of the model parameters. This, in turn, allows us to infer analytically the expected distances between organisms based on a transformation of their SI. Although the expressions obtained are rather complex, we establish additivity of this estimated evolutionary distance (a desirable property yielding phylogenetic consistency). This approach relies on holonomic functions and the Zeilberger Algorithm in order to establish additivity of the transformation of SI.

Reducing the impact of domain rearrangement on sequence alignment and phylogeny reconstruction

Sumaira Zaman, Mukul S. Bansal

Existing computational approaches for studying gene family evolution generally do not account for domain rearrangement within gene families. However, it is well known that protein domain architectures often differ between genes belonging to the same gene family. In particular, domain shuffling can lead to out-of-order domains which, unless explicitly accounted for, can significantly impact even the most fundamental of tasks such as multiple sequence alignment and phylogeny inference.

In this work, we make progress towards addressing this important but often overlooked problem. Specifically, we (i) demonstrate the impact of protein domain shuffling and rearrangement on multiple sequence alignment and gene tree reconstruction accuracy, (ii) propose two new computational methods for correcting gene sequences and alignments for improved gene tree reconstruction accuracy and evaluate them using realistically simulated datasets, and (iii) assess the potential impact of our new methods and of two existing approaches, MDAT and ProDA, in practice by applying them to biological gene families. We find that the methods work very well on simulated data but that performance of all methods is mixed, and often complementary, on real biological data, with different methods helping improve different subsets of gene families.

On Sorting by Flanked Transpositions

Huixiu Xu, Xin Tong, Haitao Jiang, Lusheng Wang, Binhai Zhu, Daming Zhu

Transposition is a well-known genome rearrangement event that switches two consecutive segments on a genome. The problem of sorting permutations by transpositions has attracted a great amount of interest since it was introduced by Bafna and Pevzner in 1995. However, empirical evidence has reported that, in many genomes, the participation of repeat segments is inevitable during genome evolution and the breakpoints where a transposition occurs are most likely accompanied by a triple of repeated segments. For example, a transposition will transform r x r y z r into r y z r x r, where r is a relative short repeat appearing three times and x and y are long segments involved in the transposition. For this transposition event, the neighbors of segments x and y remain the same before and after the transposition. This type of transposition is called flanked transposition. In this paper, we investigate the problem of sorting by flanked transpositions, which requires a series of flanked transpositions to transform one genome into another. First, we present an O(n) expected running time algorithm to determine if a genome can be transformed into the other genome by a series of flanked transposition for a special case, where each adjacency (roughly two neighbors of two element in the genome) appears once in both input genomes. We then extend the decision algorithm to work for the general case with the same expected running time O(n). Finally, we show that the new version, sorting by minimum number of flanked transpositions is also NP-hard.

USTAR: Improved Compression of k-mer Sets with Counters Using De Bruijn Graphs

Enrico Rossignolo, Matteo Comin

A fundamental operation in computational genomics is to reduce the input sequences to their constituent k-mers. Finding a space- efficient way to represent a set of k-mers is important for improving the scalability of bioinformatics analyses. One popular approach is to convert the set of k-mers into a de Bruijn graph and then find a compact representation of the graph through the smallest path cover. In this paper, we present USTAR, a tool for compressing a set of k-mers and their counts. USTAR exploits the node connectivity and density of the de Bruijn graph enabling a more effective path selection for the construction of the path cover. We demonstrate the usefulness of USTAR in the compression of read datasets. USTAR can improve the compression of UST, the best algorithm, from 2.3% up to 26,4%, depending on the k-mer size. The code of USTAR and the complete results are available at the repos- itory https://github.com/enricorox/USTAR

Predicting Comprehensive Drug–Drug Interactions by Magnetic Signed Graph Neural Network

Ming Chen, Bin Yao, Xiujuan Lei, Chunyan Ji, Zitao Hu, Yi Pan

Drug combination is a common means of clinical treatments, but detection and evaluation of drug–drug interactions (DDI) can be expensive. Graph neural network (GNN) is a popular method for DDI prediction, which has achieved encouraging performance in the scenarios of mono-type and multi-type DDI. However, most studies ignore the comprehensive information of DDI, such as the signs of DDI, asymmetric roles of drugs in pharmacological changes. In this article, we model DDI on signed and directed graphs with node attributes and define multiple tasks which include the sign direction prediction beyond individual tasks. Furthermore, we put forward a framework, called MSGNN-DDI, which uses spectral information of DDI networks and builds GNN models based on magnetic signed Laplacians. The framework not only facilitates the prediction of signs and directions in pharmacological changes but is also adaptable for their combination task. Our experiments use the drug data extracted from DrugBank and PubChem databases, and the results show the feasibility of our method on multiple tasks. The case study further verifies its effectiveness in sign direction prediction.

BiRNN-DDI: A Drug-drug Interaction Event Type Prediction Model based on Bidirectional Recurrent Neural Network and Graph to Sequence Representation

Hui Feng, Guishen Wang, Chen Cao

Drug-drug interactions (DDIs) prediction is helpful for better under- standing drug adverse reactions and drug combinations. Recent works reveal the importance of DDI event-type prediction. Hence, this paper proposes a Bidirec- tional Recurrent Neural Network for drug-drug interaction event type prediction (BiRNN-DDI). BiRNN-DDI model first constructs drug feature graphs based on drug feature similarity. To mine contextual information in DDI, the BiRNN-DDI model uses a graph-to-sequence model to transform drug feature homogeneous graphs into drug sequence representation. Then, a two-channel structure model consisting of the BiRNN is proposed to get contextual DDI sequence representa- tions. Finally, a feedforward neural network is used to predict the DDI event type. To test the effectiveness of our BiRNN-DDI model, representative state-of-the-art models are compared in two drug-drug interaction event type benchmarks. Ex- tensive experimental results show that our BiRNN-DDI model outperforms other compared models regarding precision, recall, and F1 value measures. In the mean- time, experiment results demonstrate that our model has lower parameter space. It indicates that our model is able to learn drug feature representations and predict possible drug-drug interaction event types more effectively.

PCPI: Prediction of circRNA and protein interaction using machine learning method

Md. Tofazzal Hossain, Md. Selim Reza, Yin Peng, Shengzhong Feng, Yanjie Wei

Circular RNA (circRNA) is an RNA molecule different from linear RNA with covalently closed loop structure. CircRNAs can act as sponging miRNAs and can interact with RNA binding protein. Previous studies have revealed that circRNAs play important role in the development of different diseases. The biological functions of circRNAs can be investigated with the help of circRNA-protein interaction. Due to scarce circRNA data, long circRNA sequences and the sparsely distributed binding sites on circRNAs, much fewer endeavors are found in studying the circRNA-protein interaction compared to interaction between linear RNA and protein. With the increase in experimental data on circRNA, machine learning methods are widely used in recent times for predicting the circRNA-protein interaction. The existing methods either use RNA sequence or protein sequence for predicting the binding sites. In this paper, we present a new method PCPI (Predicting CircRNA and Protein Interaction) to predict the interaction between circRNA and protein using support vector machine (SVM) classifier. We have used both the RNA and protein sequences to predict their interaction. The circRNA sequences were converted in pseudo peptide sequences based on codon translation. The pseudo peptide and the protein sequences were classified based on dipole moments and the volume of the side chains. The 3-mers of the classified sequences were used as features for training the model. Several machine learning model were used for classification. Comparing the performances, we selected SVM classifier for predicting circRNA-protein interaction. Our method achieved 93% prediction accuracy.

PDFll: Intrinsic protein disorder and function prediction from the language of life

Wanyi Yang, Chuanfang Wu, Jinku Bao

Identification of intrinsic disorder proteins and their function relies in large part on computational predictors, which demands that their quality should be high. Here we present a series of computational predictors, PDFll, that provide accurate disorder and disorder function predictions based on protein sequences. PDFll generated by two main steps, the first step relies on large protein language models (pLMS), which train on billions of protein sequences.The second step is to put the embeddings gained from pLMs into small and simple deep-learning models to get predictions.These predictions are substantially better than the results of the state-of-the-art predictors that predict disorder and funtion while training without evolutionary information.

Sequence-Based Nanobody-Antigen Binding Prediction

Usama Sardar, Sarwan Ali, Muhammad Sohaib Ayub, Muhammad Shoaib, Khurram Bashir, Imdadullah Khan, Murray Patterson

Nanobodies (Nb) are monomeric heavy-chain fragments de- rived from heavy-chain only antibodies naturally found in Camelids and Sharks. Their considerably small size (3-4 nm; 13 kDa) and favorable biophysical properties make them attractive targets for recombinant pro- duction. Furthermore, their unique ability to bind selectively to specific antigens, such as toxins, chemicals, bacteria, and viruses, makes them powerful tools in cell biology, structural biology, medical diagnostics, and future therapeutic agents in treating cancer and other serious illnesses. However, a critical challenge in nanobodies production is the unavail- ability of nanobodies for a majority of antigens. Although some com- putational methods have been proposed to screen potential nanobodies for given target antigens, their practical application is highly restricted due to their reliance on 3D structures. Moreover, predicting nanobody- antigen interactions (binding) is a time-consuming and labor-intensive task. This study aims to develop a machine-learning method to pre- dict Nanobody-Antigen binding solely based on the sequence data. We curated a comprehensive dataset of Nanobody-Antigen binding and non- binding data and devised an embedding method based on gapped k-mers to predict binding based only on sequences of nanobody and antigen. Our approach achieves up to 90% accuracy in binding prediction and is significantly more efficient compared to the widely-used computational docking technique.

Deep Learning Architectures For the Prediction of YY1-Mediated Chromatin Loops

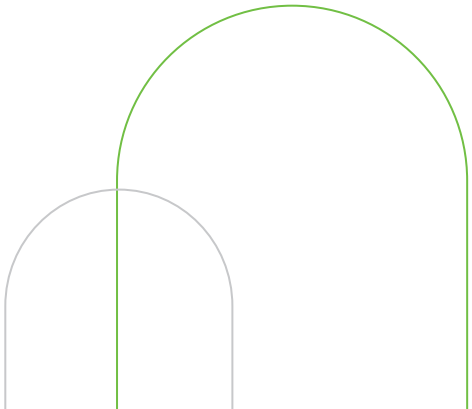
Ahtisham Fazeel, Muhammad Nabeel Asim, Johan Trygg, Andreas Dengel, Sheraz Ahmed

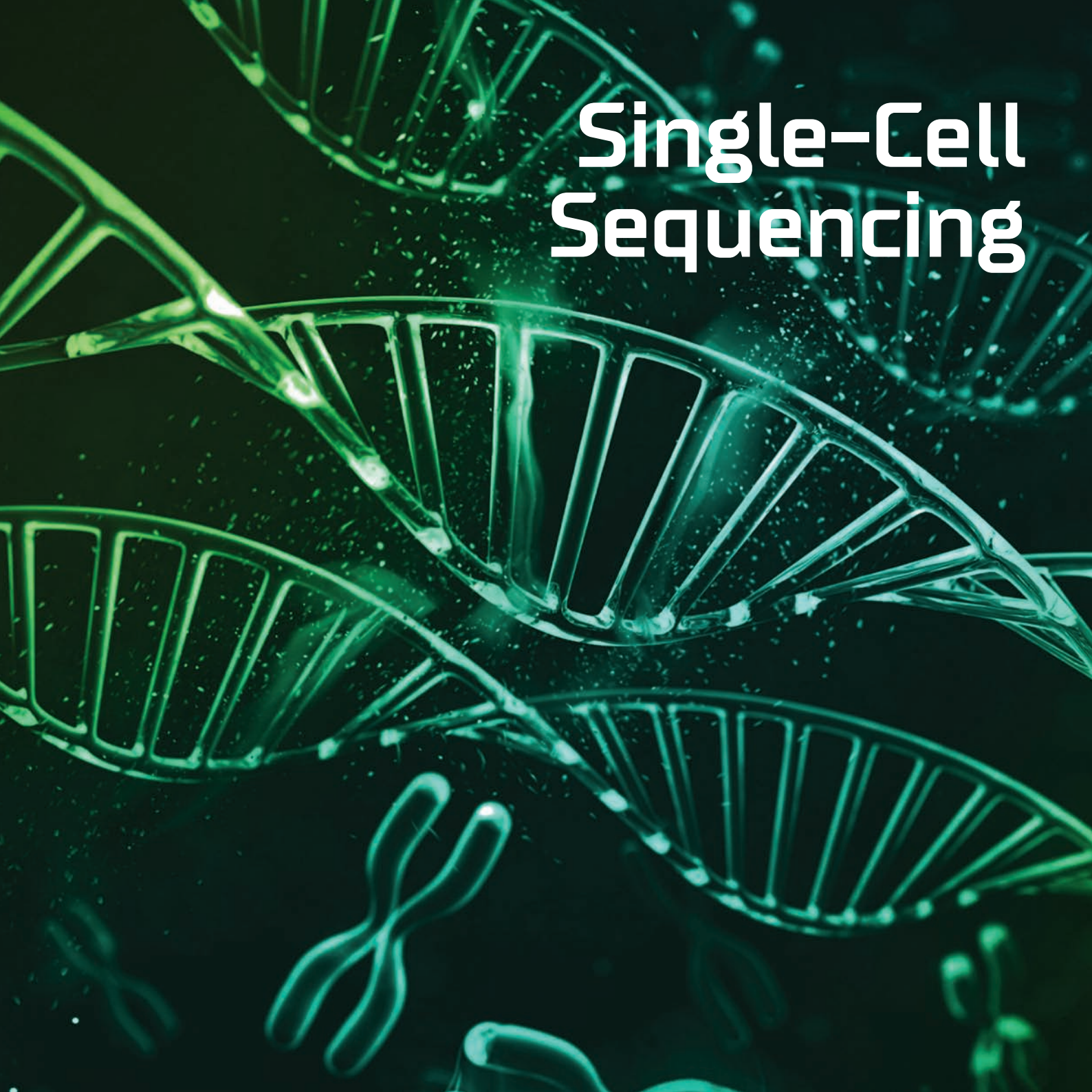
YY1-mediated chromatin loops play substantial roles in basic biological processes like gene regulation, cell differentiation, and DNA replication. YY1-mediated chromatin loop prediction is important to understand diverse types of biological processes which may lead to the development of new therapeutics for neurological disorders and cancers. Existing deep learning predictors are capable to predict YY1-mediated chromatin loops in two different cell lines however, they showed limited performance for the prediction of YY1-mediated loops in the same cell lines and suffer significant performance deterioration in cross cell line setting. To provide computational predictors capable of performing large-scale analyses of YY1-mediated loop prediction across multiple cell lines, this paper presents two novel deep learning predictors. The two proposed predictors make use of Word2vec, one hot encoding for sequence representation and long short-term memory, and a convolution neural network along with a gradient flow strategy similar to DenseNet architectures. Both of the predictors are evaluated on two different benchmark datasets of two cell lines HCT116 and K562. Overall the proposed predictors outperform existing DEEPYY1 predictor with an average maximum margin of 4.65%, 7.45% in terms of AUROC, and accuracy, across both of the datases over the independent test sets and 5.1%, 3.2% over 5-fold validation. In terms of cross-cell evaluation, the proposed predictors boast maximum performance enhancements of up to 9.5% and 27.1% in terms of AUROC over HCT116 and K562 datasets.

ABCAE: Artificial Bee Colony Algorithm with Adaptive Exploitation for Epistatic Interaction Detection

Boxin Guan, Anqi Wang, Yahan Li, Feng Li, Jin-Xing Liu, Junliang Shang

The detection of epistatic interactions among multiple single-nucleotide polymorphisms (SNPs) in complex diseases has posed a significant challenge in genome-wide association studies (GWAS). However, most existing methods still suffer from algorithmic limitations, such as high computational requirements and low detection ability. In the paper, we propose an artificial bee colony algorithm with adaptive exploitation (ABCAE) to address these issues in epistatic interaction detection for GWAS. an adaptive exploitation mechanism is designed and used in the onlooker stage of ABCAE. By using the adaptive exploitation mechanism, ABCAE can locally optimize the promising SNP combination area, thus effectively coping with the challenges brought by high-dimensional complex GWAS data. To demonstrate the detection ability of ABCAE, we compare it against four existing algorithms on eight epistatic models. The Adaptive exploitation · Artificial bee colony · Complex disease Epistatic interactionexperimental results demonstrate that ABCAE outperforms the four existing methods in terms of detection ability.




Single-Cell
Sequencing

scGASI: A graph autoencoder-based single-cell integration clustering method

Tian-Jing Qiao, Feng Li, Shasha Yuan, Ling-Yun Dai, Juan Wang

Single-cell RNA sequencing (scRNA-seq) technology offers the opportunity to study biological issues at the cellular level. The identification of single-cell types by unsupervised clustering is a basic goal of scRNA-seq data analysis. Although there have been a number of recent proposals for single-cell clus-tering methods, only a few of these have considered both shallow and deep potential information. Therefore, we propose a graph autoencoder-based sin-gle-cell integration clustering method, scGASI. Based on multiple feature sets, scGASI unifies deep feature embedding and data affinity recovery in a uniform framework to learn a consensus affinity matrix between cells. scGASI first constructs multiple feature sets. Then, to extract the deep poten-tial information embedded in the data, scGASI uses a graph autoencoder (GAEs) to learn the low-dimensional latent representation of the data. Next, to effectively fuse the deep potential information in the embedding space and the shallow information in the raw space, we design a multi-layer kernel self-expression integration strategy. This strategy uses a kernel self-expression model with multi-layer similarity fusion to learn a similarity ma-trix shared by the raw and embedding spaces of a given feature set, and a consensus learning mechanism to learn a consensus affinity matrix across all feature sets. Finally, the consensus affinity matrix is used for spectral clus-tering, visualization, and identification of gene markers. Large-scale valida-tion on real datasets shows that scGASI has higher clustering accuracy than many popular clustering methods.

Integrative analysis of gene expression and alternative polyadenylation from single-cell RNA-seq data

Shuo Xu, Liping Kang, Xingyu Bi, Xiaohui Wu

Single-cell RNA-seq (scRNA-seq) is a powerful technique for assaying tran-scriptional profile of individual cells. However, high dropout rate and overdisper-sion inherent in scRNA-seq hinders the reliable quantification of genes. Recent bioinformatic studies switched the conventional gene-level analysis to APA (al-ternative polyadenylation) isoform level, and revealed cell-to-cell heterogeneity in APA usages and APA dynamics in different cell types. The additional layer of APA isoforms creates immense potential to develop cost-efficient approaches for dissecting cell types by integrating multiple modalities derived from existing scRNA-seq experiments. Here we proposed a pipeline called scAPAfuse for en-hancing cell type clustering and identifying of novel/rare cell types by combing gene expression and APA profiles from the same scRNA-seq data. scAPAfuse first maps gene expression and APA profiles to a shared low-dimensional space using partial least squares. Then anchors (i.e., similar cells) between gene and APA profiles were identified by constructing the nearest neighbors of cells in the low-dimensional space, using algorithms like hyperplane local sensitive hash and shared nearest neighbor. Finally, gene and APA profiles were integrated to a fused matrix, using the Gaussian kernel function. Applying scAPAfuse on four public scRNA-seq datasets including human peripheral blood mononuclear cells (PBMCs) and Arabidopsis roots, new subpopulations of cells that were unde-tectable using the gene expression or APA profile alone were found. scAPAfuse provides a unique strategy to mitigate the high sparsity of scRNA-seq by fusing gene expression and APA profiles to improve cell type clustering, which can be included in many other routine scRNA-seq pipelines.

Inferring Boolean networks from single-cell human embryo datasets

Mathieu Bolteau, Jérémie Bourdon, Laurent David, Carito Guziolowski

This study aims to understand human embryonic development and cell fate determination, specifically in relation to trophectoderm (TE) maturation. We utilize single-cell transcriptomics (scRNAseq) data to develop a framework for inferring computational models that distinguish between two developmental stages. Our method selects pseudo-perturbations from scRNAseq data since actual perturbations are impractical due to ethical and legal constraints. These pseudo-perturbations consist of input-output discretized expressions, for a limited set of genes and cells. By combining these pseudo-perturbations with prior-regulatory networks, we can infer Boolean networks that accurately align with scRNAseq data for each developmental stage. Our publicly available method was tested with several benchmarks, proving the feasibility of our approach. Applied to the real dataset, we infer Boolean network families, corresponding to the medium and late TE developmental stages. Their structures reveal contrasting regulatory pathways, offering valuable biological insights and hypotheses within this domain.

CHLPCA: Correntropy-Based Hypergraph Regularized Sparse PCA for Single-cell Type Identification

Tai-Ge Wang, Xiang-Zhen Kong, Sheng-Jun Li, Juan Wang

Over the past decade, high-throughput sequencing technologies have driven a dramatic increase in single-cell RNA sequencing (scRNA-seq) data. The study of scRNA-seq data has widened the scope and depth of researchers' understanding of cellular heterogeneity. A prerequisite for studying heterogeneous cell populations is accurate cell type identification. However, the highly noisy and high-dimensional nature of scRNA-seq data poses a challenge to existing methods to further improve the success rate of cell type identification. Principal component analysis (PCA) is an important data analysis technique that is widely used to identify cell subpopulations. On the basis of PCA, we propose correntropy-based hypergraph regularized sparse PCA (CHLPCA) for accurate cell type identification. In addition to using correntropy to reduce the effect of noise, CHLPCA also considers higher-order relationships between samples by constructing the hypergraph, which compensates for the lack of local structure capture ability of PCA. Furthermore, we introduce the L2,1/5-norm into the model to enhance the interpretability of principal components (PCs), which further improves the model performance. CHLPCA has superior clustering accuracy and outperforms the best comparative method by 5.13% and 8.00% for ACC and NMI metrics, respectively. The results of clustering visualization experiments also confirm that CHLPCA can better perform the cell type recognition task.

Simulating tumor evolution from scDNA-seq as an accumulation of both SNVs and CNAs

Zahra Tayebi, Akshay Juyal, Alex Zelikovsky, Murray Patterson

Ever since single-cell sequencing (scDNA-seq) was coined 'method of the year' in 2013, it has provided many insights into the evolution of tumors, viewed as a branching process of accumulating can- cerous mutations that initiated with a single driver mutation — a model of clonal evolution which has been theorized almost half a century ago (Nowell, 1976). With this, is seen an explosion of methods for inferring the histories of such evolution, often in the form of a phylogenetic tree, from single-cell sequencing data. While the first methods modeled such evolution as an accumulation of point mutations (SNVs), copy number aberrations (CNAs, i.e., duplications or deletions of large genomic re- gions) are an important factor to consider. As a result, later methods began to bolster cancer phylogeny inference with bulk sequencing data, to account for CNAs. Despite the dozens of such inference methods avail- able, there still does not exist much in the form of a unified benchmark for all such methods.

This paper moves to initiate such a benchmark, which can be built upon, by proposing a simulator which models both SNVs and CNAs jointly in generating an evolutionary scenario which can be interpreted as a scDNA-seq/matched bulk sample pair. The simulator models the ac- cumulations of SNVs, and the duplication or deletion of chromosomal segments. We test this simulation on three methods: (a) a method which accounts for SNVs only, and under the infinite sites assumption (ISA), (b) a second more general method which models only SNVs, but allows for relaxations to the ISA, and (c) a third most general method which accounts for both SNVs and CNAs (and violations to the ISA). Results are consistent with the generality of these methods. This work is a step in the direction of developing a de-facto benchmark for cancer phylogeny inference methods.

Multi-Class Cancer Classification of Whole Slide Images through Transformer and Multiple Instance Learning

Haijing Luan, Taiyuan Hu, Jifang Hu, Ruilin Li, Detao Ji, Jiayin He, Xiaohong Duan, Chunyan Yang, Yajun Gao, Fan Chen, Beifang Niu

Whole slide image (WSI) produces images of high resolution, which are rich in details; however, it lacks localized annotations. WSI classification can be treated as a multiple instance learning (MIL) prob- lem while only slide-level labels are available. We introduce a approach for WSI classification that leverages the MIL and Transformer, effectively eliminating the requirement for localized annotations. Our method con- sists of three key components. Firstly, we use ResNet50, which has been pre-trained on ImageNet, as an instance feature extractor. Secondly, we present a Transformer-based MIL aggregator that adeptly captures con- textual information within individual regions and correlation information among diverse regions within the WSI. Our proposed approach effectively mitigates the issue of high computational complexity in the Transformer architecture by integrating linear attention. Thirdly, we introduce the global average pooling (GAP) layer to increase the mapping relationship between WSI features and category features, further improving classifi- cation accuracy. To evaluate our model, we conducted experiments on the CPTAC dataset. The results demonstrate the superiority of our ap- proach compared to previous MIL-based methods. Our proposed method achieves state-of-the-art performance in WSI classification without re- liance on localized annotations. Overall, our work offers a robust and effective approach that overcomes challenges posed by high-resolution WSIs and limited annotation availability.

Efficient Sequence Embedding For SARS-CoV-2 Variants Classification

Sarwan Ali, Usama Sardar, Imdadullah Khan, Murray Patterson

Kernel-based methods, such as Support Vector Machines (SVM), have demonstrated their utility in various machine learning (ML) tasks, including sequence classification. However, these methods face two primary challenges: (i) the computational complexity associated with kernel computation, which involves an exponential time requirement for dot product calculation, and (ii) the scalability issue of storing the large $n \times n$ matrix in memory when the number of data points (n) becomes too large. Although approximate methods can address the computational complexity problem, scalability remains a concern for conventional kernel methods.

This paper presents a novel and efficient embedding method that overcomes both the computational and scalability challenges inherent in kernel methods. To address the computational challenge, our approach involves extracting the k-mers/nGrams (consecutive character substrings) from a given biological sequence, computing a sketch of the sequence, and performing dot product calculations using the sketch. By avoiding the need to compute the entire spectrum (frequency count) and operating with low-dimensional vectors (sketches) for sequences instead of the memory-intensive $n \times n$ matrix or full-length spectrum, our method can be readily scaled to handle a large number of sequences, effectively resolving the scalability problem.

Furthermore, conventional kernel methods often rely on limited algorithms (e.g., kernel SVM) for underlying ML tasks. In contrast, our proposed fast and alignment-free spectrum method can serve as input for various distance-based (e.g., k-nearest neighbors) and non-distance-based (e.g., decision tree) ML methods used in classification and clustering tasks. By applying our method solely to real-world biological sequences, specifically those of the coronavirus spike/Peplomer, we achieve superior predictive performance without the need for full-length genome sequences. Moreover, our proposed method outperforms several state-of-the-art embedding and kernel methods in terms of both predictive performance and computational runtime.

Unveiling the Robustness of Machine Learning Models in Classifying COVID-19 Spike Sequences

Sarwan Ali, Pin-Yu Chen, Murray Patterson

In the midst of the global COVID-19 pandemic, a wealth of data has become available to researchers, presenting a unique opportunity to investigate the behavior of the virus. This research aims to facilitate the design of efficient vaccinations and proactive measures to prevent future pandemics through the utilization of machine learning (ML) models for decision-making processes. Consequently, ensuring the reliability of ML predictions in these critical and rapidly evolving scenarios is of utmost importance. Notably, studies focusing on the genomic sequences of individuals infected with the coronavirus have revealed that the majority of variations occur within a specific region known as the spike (or S) protein. Previous research has explored the analysis of spike proteins using various ML techniques, including classification and clustering of variants. However, it is imperative to acknowledge the possibility of errors in spike proteins, which could lead to misleading outcomes and misguide decision-making authorities. Hence, a comprehensive examination of the robustness of ML and deep learning models in classifying spike sequences is essential. In this paper, we propose a framework for evaluating and benchmarking the robustness of diverse ML methods in spike sequence classification. Through extensive evaluation of a wide range of ML algorithms, ranging from classical methods like naive Bayes and logistic regression to advanced approaches such as deep neural networks, our research demonstrates that utilizing k-mers for creating the feature vector representation of spike proteins is more effective than traditional one-hot encoding-based embedding methods. Additionally, our findings indicate that deep neural networks exhibit superior accuracy and robustness compared to non-deep-learning baselines. To the best of our knowledge, this study is the first to benchmark the accuracy and robustness of machine-learning classification models against various types of random corruptions in COVID-19 spike protein sequences. The benchmarking framework established in this research holds the potential to assist future researchers in gaining a deeper understanding of the behavior of the coronavirus, enabling the implementation of proactive measures and the prevention of similar pandemics in the future.

MPFNet: ECG Arrhythmias Classication Based on Multi-Perspective Feature Fusion

Yuxia Guan, Ying An, Fengyi Guo, Jianxin Wang

Arrhythmia is a common cardiovascular disease that can cause sudden cardiac death. The electrocardiogram (ECG) signal is often used to diagnose the state of the heart. However, most existing ECG diagnostic methods only use information from a single perspective, ignoring the extraction of fusion information. In this paper, we propose a novel Multi-Perspective feature Fusion Network (MPFNet) for ECG arrhythmia classification. In this model, two independent feature extraction modules are first deployed to learn one-dimensional and two-dimensional ECG features from the original one-dimensional ECG signals and its corresponding recurrence plots. At the same time, an interactive feature extraction module based on bidirectional encoder-decoder is designed to further capture the interrelationships between one-dimensional and two-dimensional perspectives, and combine them with independent features from two different perspectives to enhance the completeness and accuracy of the final representation by utilizing the correlation and complementarity between perspectives. We evaluate our method on a large public ECG dataset and the experimental results demonstrate that MPFNet outperforms the state-of-the-art approaches.

Hist2Vec: Kernel-Based Embeddings for Biological Sequence Classification

Sarwan Ali, Haris Mansoor, Prakash Chourasia, Murray Patterson

Biological sequence classification is vital in various fields, such as genomics and bioinformatics. The advancement and reduced cost of genomic sequencing have brought the attention of researchers for protein and nucleotide sequence classification. Traditional approaches face limitations in capturing the intricate relationships and hierarchical structures inherent in genomic sequences, while numerous machine-learning models have been proposed to tackle this challenge. In this work, we propose Hist2Vec, a novel kernel-based embedding generation approach for capturing sequence similarities. Hist2Vec combines the concept of histogram-based kernel matrices and Gaussian kernel functions. It constructs histogram-based representations using the unique k-mers present in the sequences. By leveraging the power of Gaussian kernels, Hist2Vec transforms these representations into high-dimensional feature spaces, preserving important sequence information. Hist2Vec aims to address the limitations of existing methods by capturing sequence similarities in a high-dimensional feature space while providing a robust and efficient framework for classification. We employ kernel Principal Component Analysis (PCA) using standard machine-learning algorithms to generate embedding for efficient classification. Experimental evaluations on protein and nucleotide datasets demonstrate the efficacy of Hist2Vec in achieving high classification accuracy compared to state-of-the-art methods. It outperforms state-of-the-art methods by achieving $>76\%$ and $>83\%$ accuracies for DNA and Protein datasets, respectively. Hist2Vec provides a robust framework for biological sequence classification, enabling better classification and promising avenues for further analysis of biological data.

SGMDD: Subgraph Neural Network-Based Model for Analyzing Functional Connectivity Signatures of Major Depressive Disorder

Yan Zhang, Xin Liu, Panrui Tang, Zuping Zhang

Biomarkers extracted from brain functional connectivity (FC) can assist in diagnosing various psychiatric disorders. Recently, several deep learning-based methods are proposed to facilitate the development of biomarkers for auxiliary diagnosis of depression and promote automated depression identification. Although they achieved promising results, there are still existing deficiencies. Current methods overlook the subgraph of braingraph and have a rudimentary network framework, resulting in poor accuracy. Conducting FC analysis with poor accuracy model can render the results unreliable. In light of the current deficiencies, this paper designed a subgraph neural network-based model named SGMDD for analyzing FC signatures of depression and depression identification. Our model surpassed many state-of-the-art depression diagnosis methods with an accuracy of 73.95%. To the best of our knowledge, this study is the first attempt to apply subgraph neural network to the field of FC analysis in depression and depression identification, we visualize and analyze the FC networks of depression on the node, edge, motif, and functional brain region levels and discovered several novel FC feature on multi-level. The most prominent one show that the hyperconnectivity of postcentral gyrus and thalamus could be the most crucial neurophysiological feature associated with depression, which may guide the development of biomarkers used for the clinical diagnosis of depression.

TCSA: A Text-guided Cross-view Medical Semantic Alignment Framework for Adaptive Multi-view Visual Representation Learning

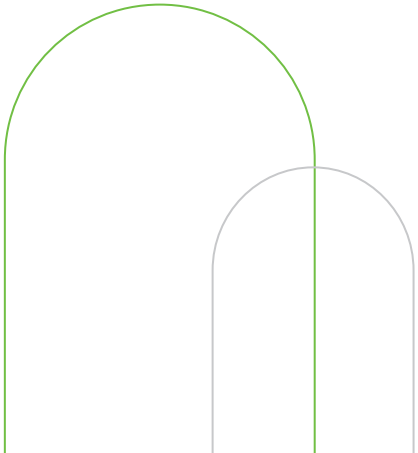
Hongyang Lei, Huazhen Huang, Bokai Yang, Guosheng Cui, Ruxin Wang, Dan Wu, Ye Li

Recently, in the medical domain, visual-language (VL) representation learning has demonstrated potential effectiveness in diverse medical downstream tasks. However, existing works typically pre-trained on the one-to-one corresponding medical image-text pairs, disregarding fluctuation in the quantity of views corresponding to reports (e.g., chest X-rays typically involve 1 to 3 projection views). This limitation results in sub-optimal performance in scenarios with varying quantities of views (e.g., arbitrary multi-view classification). To address this issue, we propose a novel Text-guided Cross-view Semantic Alignment (TCSA) framework for adaptive multi-view visual representation learning. For arbitrary number of multiple views, TCSA learns view-specific private latent sub-spaces and then maps them to a scale-invariant common latent sub-space, enabling individual treatment of arbitrary view type and normalization of arbitrary quantity of views to a consistent scale in the common sub-space. In the private sub-spaces, TCSA leverages word context as guidance to match semantic corresponding sub-regions across multiple views via cross-modal attention, facilitating alignment of different types of views in the private sub-space. This promotes the combination of information from arbitrary multiple views in the common sub-space. To the best of our knowledge, TCSA is the first VL framework for arbitrary multi-view visual representation learning. We report the results of TCSA on multiple external datasets and tasks. Compared with the state of the art frameworks, TCSA achieves competitive results and generalize well to unseen data.

A Convolutional Denoising Autoencoder for Protein Scaffold Filling

Jordan Sturtz, Richard Annan, Binhai Zhu, Xiaowen Liu, Letu Qingge

De novo protein sequencing is a valuable task in proteomics, yet it is not a fully solved problem. Many state-of-the-art approaches use top-down and bottom-up tandem mass spectrometry (MS/MS) to sequence proteins. However, these approaches often produce protein scaffolds, which are incomplete protein sequences with gaps to fill between contiguous regions. In this paper, we propose a novel convolutional denoising autoencoder (CDA) model to perform the task of filling gaps in protein scaffolds to complete the final step of protein sequencing. We demonstrate our results both on a real dataset and eleven randomly generated datasets based on the MabCampath antibody. Our results show that the proposed CDA outperforms recently published hybrid convolutional neural network and long short-term memory (CNN-LSTM) based sequence model. We achieve 100% gap filling accuracy and 95.32% full sequence accuracy on the MabCampth protein scaffold.



Enhancing t-SNE Performance for Biological Sequencing Data through Kernel Selection

Prakash Chourasia, Taslim Murad, Sarwan Ali and Murray Patterson

The genetic code for many different proteins can be found in biological sequencing data, which offers vital insight into the genetic evolution of viruses. While machine learning approaches are becoming increasingly popular for many ``Big Data'' situations, they have made little progress in comprehending the nature of such data. One such area is the t-distributed Stochastic Neighbour Embedding (t-SNE), a general-purpose approach used to represent high dimensional data in low dimensional (LD) space while preserving similarity between data points. Traditionally, the Gaussian kernel is used with t-SNE. However, since the Gaussian kernel is not data-dependent, it only determines each local bandwidth based on one local point. This makes it computationally expensive, hence limited in scalability. Moreover, it can misrepresent some structures in the data. An alternative is to use the isolation kernel, which is a data-dependent method. However, it has a single parameter to tune in computing the kernel. Although the isolation kernel yields better performance in terms of scalability and preserving the similarity in LD space, it may still not perform optimally in some cases. This paper presents a perspective on improving the performance of t-SNE and argues that kernel selection could impact this performance. We use 9 different kernels to evaluate their impact on the performance of t-SNE, using SARS-CoV-2 ``spike'' protein sequences. With three different embedding methods, we show that the cosine similarity kernel gives the best results and enhances the performance of t-SNE.

PDB2Vec: Using 3D Structural Information For Improved Protein Analysis

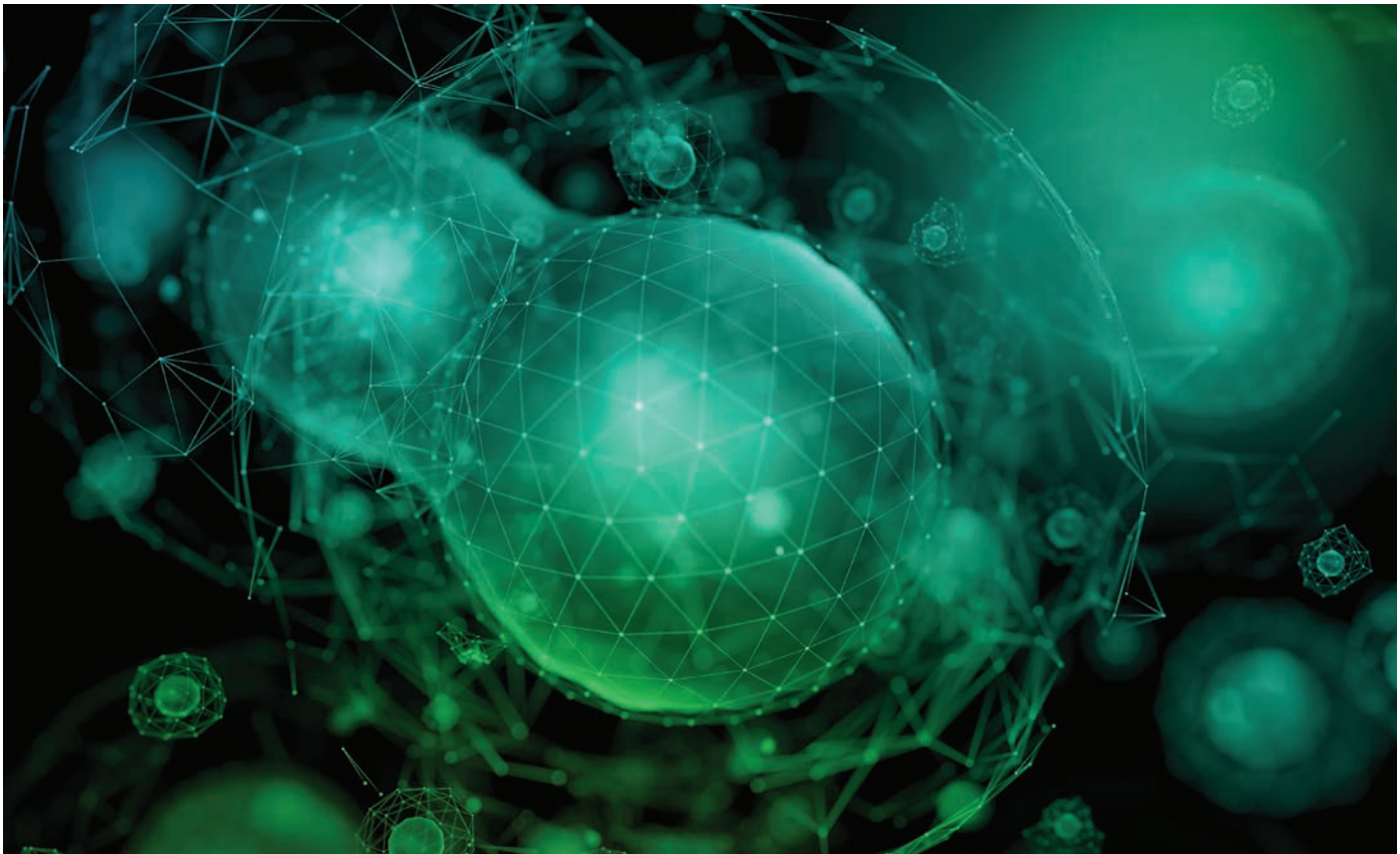
Sarwan Ali, Prakash Chourasia, Murray Patterson

In recent years, machine learning methods have shown remarkable results in various protein analysis tasks, including protein classification, folding prediction, and protein-to-protein interaction prediction. However, most studies focus only on the 3D structures or sequences for the downstream classification task. Hence analyzing the combination of both 3D structures and sequences remains comparatively unexplored. This study investigates how incorporating protein sequence and 3D structure information influences protein classification performance. We use two well-known datasets, STCRDAB and PDB Bind, for classification tasks to accomplish this. To this end, we propose an embedding method called PDB2Vec to encode both the 3D structure and protein sequence data to improve the predictive performance of the downstream classification task. We performed protein classification using three different experimental settings: only 3D structural embedding (called PDB2Vec), sequence embeddings using alignment-free methods from the biology domain including on k-mers, position weight matrix, minimizers and spaced k-mers, and the combination of both structural and sequence-based embeddings. Our experiments demonstrate the importance of incorporating both three-dimensional structural information and amino acid sequence information for improving the performance of protein classification and show that the combination of structural and sequence information leads to the best performance. We show that both types of information are complementary and essential for classification tasks.

DCNN: Dual-Level Collaborative Neural Network for Imbalanced Heart Anomaly Detection

Ying An, Anxuan Xiong, Lin Guo

The electrocardiogram (ECG) plays an important role in assisting clinical diagnosis such as arrhythmia detection. However, traditional techniques for ECG analysis are time consuming and laborious. Recently, deep neural networks have become a popular technique for automatically tracking ECG signals, which has demonstrated that they are more competitive than human experts. However, the minority class of life-threatening arrhythmias causes the model training to skew towards the majority class. To address the problem, we propose a dual-level collaborative neural network (DCNN), which includes data-level and cost-sensitive level modules. In the Data Level module, we utilize the generative adversarial network with Unet as the generator to synthesize ECG signals. Next, the Cost-sensitive Level module employs focal loss to increase the cost of incorrect prediction of the minority class. Empirical results show that the Data Level module generates highly accurate ECG signals with fewer parameters. Furthermore, DCNN has been shown to significantly improve the classification of the ECG.



Proteoform identification for top-down tandem mass spectra: efficient algorithms for global and local alignments with peak error correction

Lusheng Wang, Zhaohui Zhan

Proteoform identification is an important problem in proteomics. The main task is to find a modified protein that best fits the input spectrum. Proteoform mass graph and spectrum mass graph are used to represent the protein database and the spectrum, respectively. Peak error correction is an important issue for computing an optimal alignment between the two input mass graphs. Error correction alignment was defined to give each matched peak a precise (corrected) position. However, the algorithms are slow so that only local alignments with pre-fixed starting positions can be computed for real data sets. In this paper, we propose a faster algorithm for the error correction alignment of spectrum mass graph and proteoform mass graph problem, and produce a program package TopMGFast in C++. The newly designed algorithms require less space and running time so that we are able to compute global optimal alignments for the two input mass graphs in a reasonable time. For the local alignment version, experiments show that the running time of the new algorithm is reduced by 2.5 times. For the global alignment version, experiments show that the maximum mass errors between any pair of matched nodes in the alignments obtained by our method are within a small range as designed, while the alignments produced by the state-of-the-art method, TopMG, have very large maximum mass errors for many cases. The obtained alignment sizes are roughly the same for both TopMG and TopMGFAst. Therefore, our new algorithm can obtain more reliable global alignments within a reasonable time. This is the first time that global optimal error correction alignments can be obtained using real data sets. The software package and test data sets are available at https://github.com/Zeirdo/TopMGFast.

SaID: Simulation-aware Image Denoising Pre-trained Model for Cryo-EM Micrographs

Zhidong Yang, Hongjia Li, Dawei Zang, Renmin Han, Fa Zhang

Cryo-Electron Microscopy (cryo-EM) is a revolutionary technique for determining the structures of proteins and macromolecules. Physical limitations of the imaging conditions cause a very low Signal-to-Noise Ratio (SNR) in cryo-EM micrographs, resulting in difficulties in downstream analysis and accurate ultrastructure determination. Hence, the effective denoising algorithm for cryo-EM micrographs is in demand to facilitate the quality of analysis in macromolecules. However, lacking rich and well-defined dataset with ground truth images, supervised image denoising methods generalize poorly to experimental micrographs. To address this issue, we present a Simulation-aware Image Denoising (SaID) pre-trained model for improving the SNR of cryo-EM micrographs by only training with the accurately simulated dataset. Firstly, we devise a calibration algorithm for the simulation parameters of cryo-EM micrographs to fit the experimental micrographs. Secondly, with the accurately simulated dataset, we propose to train a deep general denoising model which can well generalize to real experimental cryo-EM micrographs. Extensive experimental results demonstrate that our pre-trained denoising model can perform outstandingly on experimental cryo-EM micrographs and simplify the downstream analysis. This indicates that a network only trained with accurately simulated noise patterns can reach the capability as if it had been trained with rich real data. Code and data will be available at https://github.com/ZhidongYang/SaID.

Attention-Guided Residual U-Net with SE Connection and ASPP for Watershed-based Cell Segmentation in Microscopy Images

Jovial Niyogisubizo, Zhao Keliang, Jintao Meng, Yi Pan, Didi Rosiyadi, Yanjie Wei

Time-lapse microscopy imaging is an important method used in biomedical studies to observe how cells behave over time. This technique provides valu-able cell numbers, sizes, shapes, and interaction data. Manual analysis of hundreds or thousands of cells is impractical, necessitating automated cell segmentation approaches. Due to their success, deep learning (DL) based methods, particularly those using U-Net-based networks, have gained popu-larity in medical and microscopy image segmentation. However, accurately segmenting touching cells in images with low signal-to-noise ratios remains challenging. Existing methods often simplistically combine low-level and high-level features, leading to model confusion. To address these issues, we propose a novel framework called RA-SE-ASPP-Net, which incorporates Residual Blocks (RB), Attention Mechanism (AM), Squeeze and-Excitation (SE) connection, and Atrous Spatial Pyramid Pooling (ASPP) for precise and robust cell segmentation. We evaluate our proposed architecture using an in-duced pluripotent stem (iPS) cell reprogramming dataset, which has received limited attention in this field. Additionally, we compare our model with dif-ferent ablation experiments to demonstrate its robustness. Our dataset achieves mean Jaccard scores of 0.835, 0.854, 0.846, 0.862, 0.871, 0.889, and 0.89 for U-Net, Att-U-Net, ResU-Net, ResAtt-U-Net, ResU-Net-SE, ResU-Net-ASPP, and RA-SE-ASPP-Net, respectively. The proposed architec-ture outperforms the baseline models in all evaluated metrics, providing the most accurate semantic segmentation results. Finally, we applied the water-shed method to the semantic segmentation results to obtain precise segmentations with specific information for each cell. The source code is publicly available at https://github.com/jovialniyo93/cell-segmentation.

Multi-modality MRI Feature Interaction for Pseudoprogression Prediction of Glioblastoma

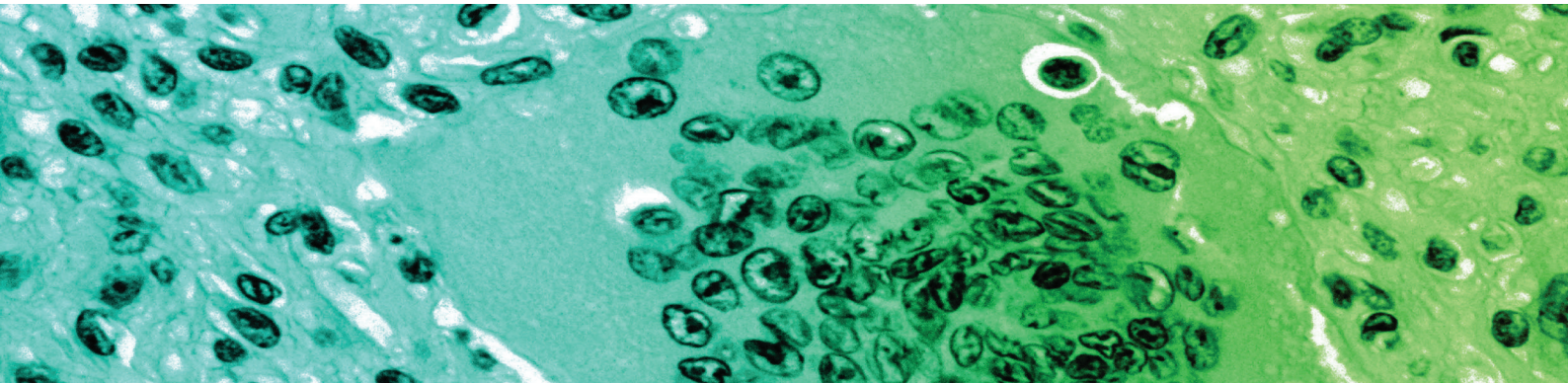
Ya Lv, Jin Liu, Pei Yang, Yi Pan

Pseudoprogression (Psp) prediction of glioblastoma (GBM) is a challenging task in clinical practice. Currently, Psp prediction of GBM is mainly performed by scanning different magnetic resonance imaging (MRI) modalities. However, how to effectively make use of the complementary information between the multi-modality to improve Psp prediction of GBM is still a challenge. To address these challenges, we propose a multi-modality MRI feature interaction method for Psp prediction of GBM, using T1 and T2 MRI. To mine multi-modality multi-scale features, we design a multi-scale feature extraction network based on a three-branch asymmetric convolution (TAC) block. Particularly, to make full use of the complementary information between T1 and T2 MRI, we propose a multi-modality MRI feature interaction (MMFI) module. Our proposed method is evaluated on a private dataset from Hunan Cancer Hospital including 10 subjects with Psp and 42 subjects with relapse. The experimental results show that the average accuracy (ACC) and area under the receiver operating characteristic curve (AUC) of the proposed method are 0.954 and 0.929, respectively. Compared with some existing methods, the proposed method can obtain better results. In summary, our proposed method has the potential for Psp prediction of GBM in clinical practice.

NeoMS: Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry

Shaokai Wang, Ming Zhu, Bin Ma

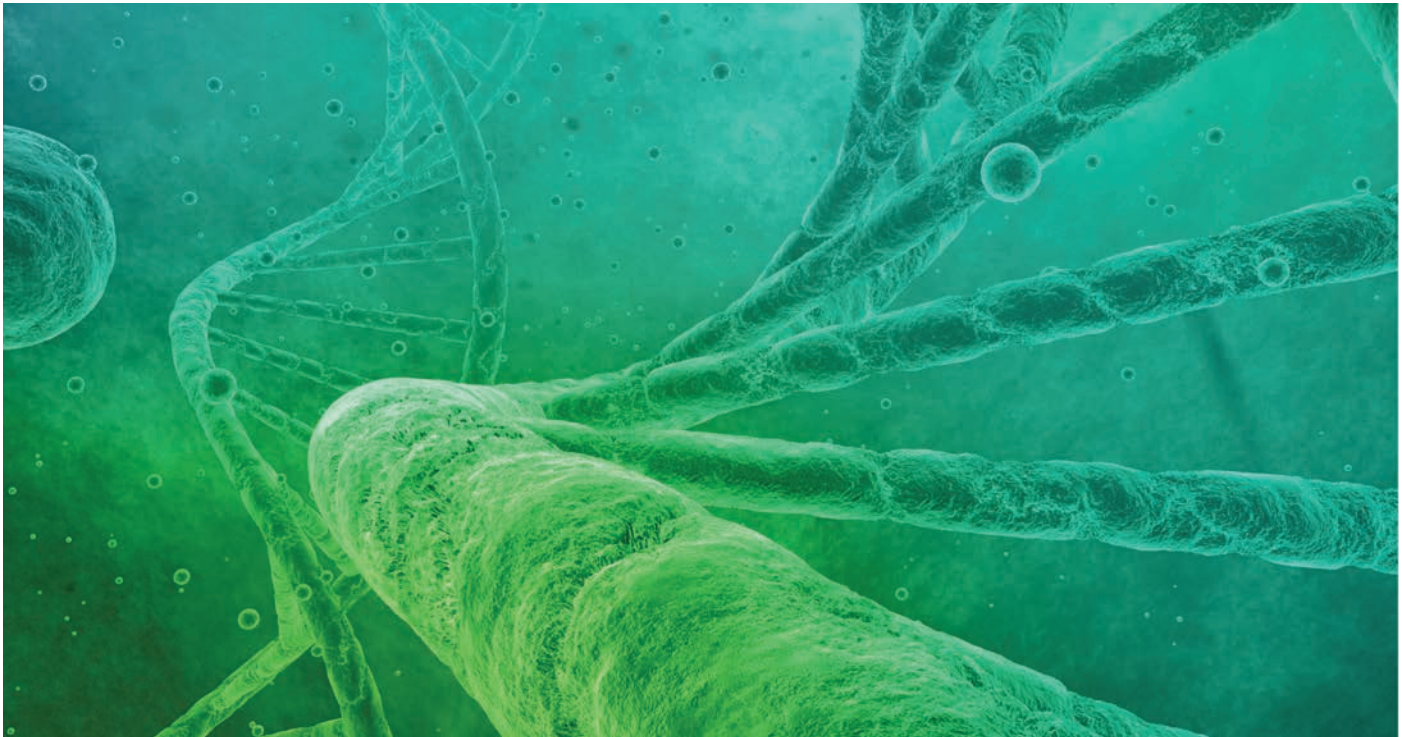
The study of immunopeptidomics requires the identification of both regular and mutated MHC-I peptides from mass spectrometry data. For the efficient identification of MHC-I peptides with either one or no mutation from a sequence database, we propose a novel work- flow: NeoMS. It employs three main modules: generating an expanded sequence database with a tagging algorithm, a machine learning-based scoring function to maximize the search sensitivity, and a careful target- decoy implementation to control the false discovery rates (FDR) of both the regular and mutated peptides. Experimental results demonstrate that NeoMS both improved the identification rate of the regular peptides over other database search software and identified hundreds of mutated pep- tides that have not been identified by any current methods. Further study shows the validity of these new novel peptides.



Radiology Report Generation via Visual Recalibration and Context Gating-aware

Xiaodi Hou, Guoming Sang, Zhi Liu, Xiaobo Li and Yijia Zhang

The task of radiology report generation aims to analyze medical images, extract key information, and then assist medical personnel in generating detailed and accurate reports. Therefore, automatic radiology report generation plays an important role in medical diagnosis and healthcare. However, radiology medical data face the problems of visual and text data bias:medical images are similar to each other, and the normal feature distribution is larger than the abnormal feature distribution; second, the accurate location of the lesion and the generation of accurate and coherent long text reports are important challenges. In this paper, we propose Visual Recalibration and Context Gating-aware model (VRCG) to alleviate visual and textual biases for enhancing report generation. We employ a medical visual recalibration module to enhance the key lesion features' extraction. We use the context gating-aware module to combine lesion location and report context information to solve the problem of long-distance dependence in diagnostic reports. Meanwhile, the context gating-aware module can identify text fragments related to lesion descriptions, improve the model's perception of lesion text information, and then generate coherent, consistent medical reporting. Extensive experiments demonstrate that our proposed model outperforms existing baseline models on publicly available IU X-Ray datasets.



CSA-MEM: Enhancing Circular DNA Multiple Alignment through Text Indexing Algorithms

André Salgado, Francisco Fernandes, Ana Teresa Freitas

In the realm of Bioinformatics, the comparison of DNA sequences is essential for tasks such as phylogenetic identification, comparative genomics, and genome reconstruction. Methods for estimating sequence similarity have been successfully applied in this field. The application of these methods to circular genomic structures, common in nature, poses additional computational hurdles. In the advancing field of metagenomics, innovative circular DNA alignment algorithms are vital for accurately understanding circular genome complexities. Aligning circular DNA, more intricate than linear sequences, demands heightened algorithms due to circularity, escalating computation requirements and runtime. This paper proposes CSA-MEM, an efficient text indexing algorithm to identify the most informative region to rotate and cut circular genomes, thus improving alignment accuracy. The algorithm uses a circular variation of the FM-Index and identifies the longest chain of non-repeated maximal subsequences common to a set of circular genomes, enabling the most adequate rotation and linearisation for multiple alignment. The effectiveness of the approach was validated in five sets of mitochondrial, viral and bacterial DNA. The results show that CSA-MEM significantly improves the efficiency of multiple sequence alignment, consistently achieving top scores compared to other state-of-the-art methods. This tool enables more realistic phylogenetic comparisons between species, facilitates large metagenomic data processing, and opens up new possibilities in comparative genomics.

Genetic Algorithm with Evolutionary Jumps

Hafsa Farooq, Daniel Novikov, Akshay Juyal, Alex Zelikovsky

It has recently been noticed that dense subgraphs of SARS-CoV-2 epistatic networks correspond to future unobserved variants of concern. This phenomenon can be interpreted as multiple correlated mutations occurring almost simultaneously, resulting in a new variant relatively distant from the current population. We refer to this phenomenon as an evolutionary jump and propose to use it for enhancing genetic algorithm. Evolutionary jumps were implemented using CliqueSNV algorithm which find cliques in the epistatic network. We have applied the genetic algorithm with evolutionary jumps (GA+EJ) to the 0-1 Knapsack problem, and found that evolutionary jumps allow the genetic algorithm to escape local minima and find solutions closer to the optimum.

A Brief Study of Gene Co-Expression Thresholding Algorithms

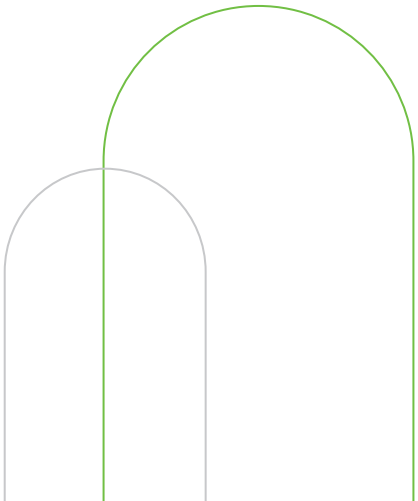
Carissa Bleker, Stephen Grady, Michael A. Langston

The thresholding problem is considered in the context of high-throughput biological data. Several approaches are reviewed, implemented, and tested over an assortment of transcriptomic data.

Graph-Based Motif Discovery in Mimotope Profiles of Serum Antibody Repertoire

Hossein Saghaian, Pavel Skums, Yurij Ionov, Alex Zelikovsky

Phage display technique has a multitude of applications such as epitope mapping, organ targeting, therapeutic antibody engineering and vaccine design. One area of particular importance is the detection of cancers in early stages, where the discovery of binding motifs and epitopes is critical. While several techniques exist to characterize phages, Next Generation Sequencing (NGS) stands out for its ability to provide detailed insights into antibody binding sites on antigens. However, when dealing with NGS data, identifying regulatory motifs poses significant challenges. Existing methods often lack scalability for large datasets, rely on prior knowledge about the number of motifs, and exhibit low accuracy. In this paper, we present a novel approach for identifying regulatory motifs in NGS data. Our method leverages results from graph theory to overcome the limitations of existing techniques.



ricME: long-read based mobile element variant detection using sequence realignment and identity calculation

Huidong Ma, Cheng Zhong, Hui Sun Haixiang Lin

Mobile element variant is a very important structural variant, accounting for a quarter of structural variant, and it is closely related to many issues such as genetic diseases and species diversity. However, few detection algorithms of mobile element variants have been developed on the third-generation se-quencing data. We propose an algorithm ricME that combines sequence rea-lignment and identity calculation for detecting mobile element variants. ric-ME first performs the initial detection to obtain the positions of insertions and deletions and extracts the variant sequences; then applies sequence rea-lignment and identity calculation to obtain the transposon classes related to the variant sequences; finally, adopts a multi-level judgment rule to achieve accurate detection of mobile element variants based on the transposon clas-ses and identities. Compared with a representative long-read based mobile element variant detection algorithm rMETL, ricME improves the F1 scores by 11.5% and 21.7% for the experimental results run on simulated datasets and real datasets, respectively. The proposed algorithm ricME is available freely at https://github.com/mhuidong/ricME.

Reconciling Inconsistent Molecular Structures from Biochemical Databases

Casper Asbjørn Eriksen, Jakob Lykke Andersen, Rolf Fagerberg, Daniel Merkle

Information on the structure of molecules, retrieved via biochemical databases, plays a pivotal role in various disciplines, such as metabolomics, systems biology, and drug discovery. However, no such database can be complete, and the chemical structure for a given compound is not necessarily consistent between databases. This paper presents StructRecon, a tool for resolving unique and correct molecular structures from database identifiers. StructRecon traverses the cross-links between database entries in different databases to construct what we call an identifier graph, which offers a more complete view of the total information available on a particular compound across all the databases. In order to reconcile discrepancies between databases, we first present an extensible model for chemical structure which supports multiple independent levels of detail, allowing standardisation of the structure to be applied iteratively. In some cases, our standardisation approach results in multiple structures for a given compound, in which case a random walk-based algorithm is used to select the most likely structure among incompatible alternates. We applied StructRecon to the EColiCore2 model, resolving a unique chemical structure for 85.11 % of identifiers. StructRecon is open-source and modular, which enables the potential support for more databases in the future.

Clique-based topological characterization of chromatin interaction hubs

Gatis Melkus, Sandra Siliņa, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms, Juris Viksna

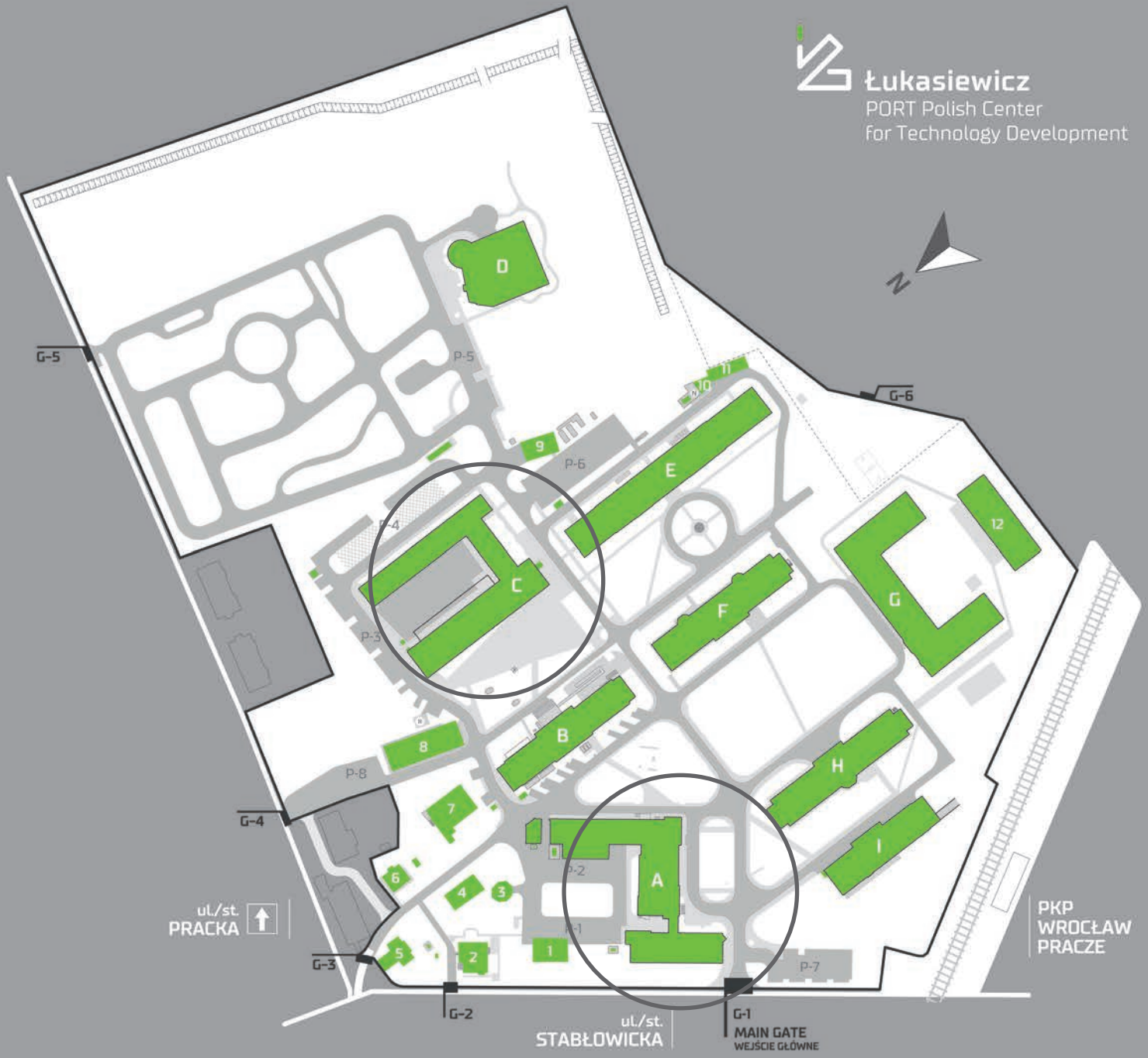
Chromatin conformation capture technologies are a vital source of information about the spatial organization of chromatin in eukaryotic cells. Of these technologies, Hi-C and related methods have been widely used to obtain reasonably complete contact maps in many cell lines and tissues under a wide variety of conditions. This data allows for the creation of chromatin interaction graphs from which topological generalizations about the structure of chromatin may be drawn. Here we outline and utilize a clique-based approach to analyzing chromatin interaction graphs which allows for both detailed analysis of strongly interconnected regions of chromatin and the unraveling of complex relationships between genomic loci in these regions. We find that clique-rich regions are significantly enriched in distinct gene ontologies as well as regions of transcriptional activity compared to the entire set of links in the respective datasets, and that these cliques are also not entirely preserved in randomized Hi-C data. We conclude that cliques and the denser regions of connectivity in which they are common appear to indicate a consistent pattern of chromatin spatial organization that resembles transcription factories, and that cliques can be used to identify functional modules in Hi-C data.

On the Realisability of Chemical Pathways

Jakob Lykke Andersen, Sissel Banke, Rolf Fagerberg, Christoph Flamm, Daniel Merkle, Peter F. Stadler

The exploration of pathways and alternative pathways that have a specific function is of interest in numerous chemical contexts. A framework for specifying and searching for pathways has previously been developed, but a focus on which of the many pathway solutions are realisable, or can be made realisable, is missing. Realisable here means that there actually exists some sequencing of the reactions of the pathway that will execute the pathway. We present a method for analysing the realisability of pathways based on the reachability question in Petri nets. For realisable pathways, our method also provides a certificate encoding an order of the reactions which realises the pathway. We present two extended notions of realisability of pathways, one of which is related to the concept of network catalysts. We exemplify our findings on the pentose phosphate pathway. Lastly, we discuss the relevance of our concepts for elucidating the choices often implicitly made when depicting pathways.

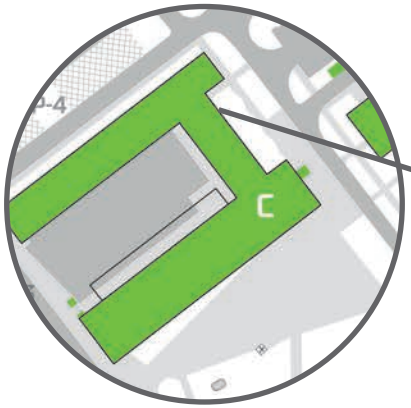
Łukasiewicz
PORT Polish Center
for Technology Development



Session A

Buliding A

Session B



Buliding C

Timetable

ISBRA 2023

19th International Symposium on Bioinformatics Research and Applications

MONDAY 9.10.2023

19:00–21:00 Welcome Dinner

TUESDAY 10.10.2023

A

08:30–09:00 **Registration**

09:00–09:30 **Session 1: Official Opening of the Conference**

09:30–10:30 **Session 2: Keynote Talk**

Teresa Przytycka *Delineating relation between mutagenic signatures, cellular processes, and environment through computational approaches*

10:30–10:50 **Coffee Break**

10:50–12:20 **Session 3A: Cancer/Health**

10:50 **Rui Gao, Zixue Liu, Mei Meng and Jian He**
Neurogenesis-associated Protein, a Potential Prognostic Biomarker in anti-PD-1 based kidney renal clear cell carcinoma patients therapeutics

11:10 **Bikram Sahoo and Alex Zelikovsky**
Deep Learning Reveals Biological Basis of Racial Disparities in Quadruple-Negative Breast Cancer

B

10:50–12:20 **Session 3B: RNA/Transcriptomics**

10:50 **Jingjing Zhang, Md. Tofazzal Hossain, Zhen Ju, Wenhui Xi and Yanjie Wei**
Identification and functional annotation of circRNAs in neuroblastoma based on bioinformatics

11:10 **Xuehua Bi, Chunyang Jiang, Cheng Yan, Kai Zhao, Linlin Zhang and Jianxin Wang**
Identifying miRNA-disease Associations based on Simple Graph Convolution with DropMessage and Jumping Knowledge

11:25 **Bikram Sahoo and Alex Zelikovsky**
Exploring Racial Disparities in Triple-Negative Breast Cancer: Insights from Feature Selection Algorithms

11:40 **Yulong Li, Hongming Zhu, Xiaowen Wang and Qin Liu**
HetBiSyn: Predicting Anticancer Synergistic Drug Combinations Featuring Bi-perspective Drug Embedding with Heterogeneous Data

12:20–13:00 **Lunch Break**

13:00–14:00 **Matchmaking event | Registration required**

14:00–15:00 **Session 4: Keynote Talk**
Anna Gambin *Statistical modeling in proteomics*

15:00–15:20 **Coffee Break**

15:20–17:40 **Session 5A: Theory**

15:20 **Michael Souza, Nilton Maia and Carlile Lavor**
The Ordered Covering Problem in Distance Geometry

15:40 **Gabriel Siqueira, Alexsandro Oliveira Alexandrino, Andre Rodrigues Oliveira, Géraldine Jean, Guillaume Fertin and Zanoni Dias**
Approximating Rearrangement Distances with Replicas and Flexible Intergenic Regions

16:00 **Joyanta Basak, Ahmed Soliman, Nachiket Deo, Kenneth Haase, Anup Mathur, Krista Park, Rebecca Steorts, Daniel Weinberg, Sartaj Sahni and Sanguthevar Rajasekaran**
On Computing the Jaro Similarity Between Two Strings

16:20 **Guy Katriel, Udi Mahanaymi, Christoph Koutschan, Doron Zeilberger, Mike Steel and Sagi Snir**
Using Generating Functions to Prove Additivity of Gene-Neighborhood Based Phylogenetics

16:40 **Sumaira Zaman and Mukul S. Bansal**
Reducing the impact of domain rearrangement on sequence alignment and phylogeny reconstruction

17:00 **Huixiu Xu, Xin Tong, Haitao Jiang, Lusheng Wang, Binhai Zhu and Daming Zhu**
On Sorting by Flanked Transpositions

17:20 **Enrico Rossignolo and Matteo Comin**
USTAR: Improved Compression of k-mer Sets with Counters Using De Bruijn Graphs

18:00–21:00 **Foyer on Kampus Pracze**

11:30 **Sarah von Loehneysen, Thomas Spicher, Yuliia Varenyk, Hua-Ting Yao, Ronny Lorenz, Ivo Hofacker and Peter F. Stadler**
Phylogenetic Information as Soft Constraints in RNA Secondary Structure Prediction

11:50 **Rafał Stępień, Joanna Szyda, Bartosz Czech and Magda Mielczarek**
The effect of transcriptomic annotations in breast cancer DGE study

15:20–17:40 **Session 5B: Interaction/Binding/Function Prediction**

15:20 **Ming Chen, Bin Yao, Xiujuan Lei, Chunyan Ji, Zitao Hu and Yi Pan**
Predicting Comprehensive Drug-Drug Interactions by Magnetic Signed Graph Neural Network

15:40 **Hui Feng, Guishen Wang and Chen Cao**
BiRNN-DDI:A Drug-drug Interaction Event Type Prediction Model based on Bidirectional Recurrent Neural Network and Graph to Sequence Representation

16:00 **Md. Tofazzal Hossain, Md. Selim Reza, Yin Peng, Shengzhong Feng and Yanjie Wei**
PCPI: Prediction of circRNA and protein interaction using machine learning method

16:20 **Wanyi Yang, Chuanfang Wu and Jinku Bao**
PDFll: Intrinsic protein disorder and function prediction from the language of life

16:40 **Usama Sardar, Sarwan Ali, Muhammad Sohaib Ayub, Muhammad Shoaib, Khurram Bashir, Imdadullah Khan and Murray Patterson**
Sequence-Based Nanobody-Antigen Binding Prediction

17:00 **Ahtisham Fazeel, Muhammad Nabeel Asim, Johan Trygg, Andreas Dengel and Sheraz Ahmed**
Deep Learning Architectures For the Prediction of YY1-Mediated Chromatin Loops

17:20 **Boxin Guan, Anqi Wang, Yahan Li, Feng Li, Jin-Xing Liu and Junliang Shang**
ABCAE: Artificial Bee Colony Algorithm with Adaptive Exploitation for Epistatic Interaction Detection

WEDNESDAY 11.10.2023

A

09:00–10:00 **Session 6: Keynote Talk**

Mark Robinson *On the care and feeding of (computational method) benchmarks*

10:00–10:20 **Coffee Break**

10:20–12:20 **Session 7A: Single-Cell Sequencing**

10:20 **Tian-Jing Qiao, Feng Li, Shasha Yuan, Ling-Yun Dai and Juan Wang**
scGASI: A graph autoencoder-based single-cell integration clustering method

10:40 **Shuo Xu, Liping Kang, Xingyu Bi and Xiaohui Wu**
Integrative analysis of gene expression and alternative polyadenylation from single-cell RNA-seq data

11:00 **Mathieu Bolteau, Jérémie Bourdon, Laurent David and Carito Guziolowski**
Inferring Boolean networks from single-cell human embryo datasets

11:15 **Tai-Ge Wang, Xiang-Zhen Kong, Sheng-Jun Li and Juan Wang**
CHLPCA: Correntropy-Based Hypergraph Regularized Sparse PCA for Single-cell Type Identification

11:30 **Zahra Tayebi, Akshay Juyal, Alex Zelikovsky and Murray Patterson**
Simulating tumor evolution from scDNA-seq as an accumulation of both SNVs and CNAs

12:20–13:00 **Session 8: Lunch**

13:00–14:00 **Matchmaking event | Registration required**

14:00–15:00 **Session 9: Keynote Talk**

Sagi Snir *The Department of Evolutionary and Environmental Biology, University of Haifa*

15:00–15:20 **Coffee Break**

15:20–17:40 **Session 10A: Learning**

15:20 **Yan Zhang, Xin Liu, Panrui Tang and Zuping Zhang**
SGMDD: Subgraph Neural Network-Based Model for Analyzing Functional Connectivity Signatures of Major Depressive Disorder

15:40 **Hongyang Lei, Huazhen Huang, Bokai Yang, Guosheng Cui, Ruxin Wang, Dan Wu and Ye Li**
TCSA: A Text-guided Cross-view Medical Semantic Alignment Framework for Adaptive Multi-view Visual Representation Learning

B

10:20–12:20 **Session 7B: Classification**

10:20 **Haijing Luan, Taiyuan Hu, Jifang Hu, Ruilin Li, Detao Ji, Jiayin He, Xiaohong Duan, Chunyan Yang, Yajun Gao, Fan Chen and Beifang Niu**
Multi-Class Cancer Classification of Whole Slide Images through Transformer and Multiple Instance Learning

10:40 **Sarwan Ali, Usama Sardar, Imdadullah Khan and Murray Patterson**
Efficient Sequence Embedding For SARS-CoV-2 Variants Classification

11:00 **Sarwan Ali, Pin-Yu Chen and Murray Patterson**
Unveiling the Robustness of Machine Learning Models in Classifying COVID-19 Spike Sequences

11:20 **Yuxia Guan, Ying An, Fengyi Guo and Jianxin Wang**
MPFNet: ECG Arrhythmias Classication Based on Multi-Perspective Feature Fusion

11:40 **Sarwan Ali, Haris Mansoor, Prakash Chourasia and Murray Patterson**
Hist2Vec: Kernel-Based Embeddings for Biological Sequence Classification

15:20–17:40 **Session 10B: Imaging/Signal Processing**

15:20 **Lusheng Wang and Zhaohui Zhan**
Proteoform identification for top-down tandem mass spectra: efficient algorithms for global and local alignments with peak error correction

15:40 **Zhidong Yang, Hongjia Li, Dawei Zang, Renmin Han and Fa Zhang**
SaID: Simulation-aware Image Denoising Pre-trained Model for Cryo-EM Micrographs

16:00 **Jordan Sturtz, Richard Annan, Binhai Zhu, Xiaowen Liu and Letu Qingge**
A Convolutional Denoising Autoencoder for Protein Scaffold Filling

16:15 **Prakash Chourasia, Taslim Murad, Sarwan Ali and Murray Patterson**
Enhancing t-SNE Performance for Biological Sequencing Data through Kernel Selection

16:30 **Sarwan Ali, Prakash Chourasia and Murray Patterson**
PDB2Vec: Using 3D Structural Information For Improved Protein Analysis

16:45 **Ying An, Anxuan Xiong and Lin Guo**
DCNN: Dual-Level Collaborative Neural Network for Imbalanced Heart Anomaly Detection

18:00–21:00 **Dinner on Odra River and Cruise**

16:00 **Jovial Niyogisubizo, Zhao Keliang, Jintao Meng, Yi Pan, Didi Rosiyadi and Yanjie Wei**
Attention-Guided Residual U-Net with SE Connection and ASPP for Watershed-based Cell Segmentation in Microscopy Images

16:20 **Ya Lv, Jin Liu, Pei Yang and Yi Pan**
Multi-modality MRI Feature Interaction for Pseudoprogression Prediction of Glioblastoma

16:40 **Shaokai Wang, Ming Zhu and Bin Ma**
NeoMS: Identification of Novel MHC-I Peptides with Tandem Mass Spectrometry

17:00 **Xiaodi Hou, Guoming Sang, Zhi Liu, Xiaobo Li and Yijia Zhang**
Radiology Report Generation via Visual Recalibration and Context Gating-aware

THURSDAY 12.10.2023

A

09:00–10:00 **Session 11: Keynote Talk**

Ana Teresa Freitas *Turning Data into Genomic Medicine*

10:00–10:20 **Coffee Break**

10:20–12:20 **Session 12A: Algorithms**

10:20 **André Salgado, Francisco Fernandes and Ana Teresa Freitas**
CSA-MEM: Enhancing Circular DNA Multiple Alignment through Text Indexing Algorithms

10:35 **Hafsa Farooq, Daniel Novikov, Akshay Juyal and Alex Zelikovsky**
Genetic Algorithm with Evolutionary Jumps

10:50 **Carissa Bleker, Stephen Grady and Michael A. Langston**
A Brief Study of Gene Co-Expression Thresholding Algorithms

11:05 **Hossein Saghaian, Pavel Skums, Yurij Ionov and Alex Zelikovsky**
Graph-Based Motif Discovery in Mimotope Profiles of Serum Antibody Repertoire

12:20–12:30 **Session 13: Closing Remarks**

12:30–13:30 **Lunch**

B

10:20–12:20 **Session 12B: Misc**

10:20 **Huidong Ma, Cheng Zhong, Hui Sun and Haixiang Lin**
ricME: long-read based mobile element variant detection using sequence realignment and identity calculation

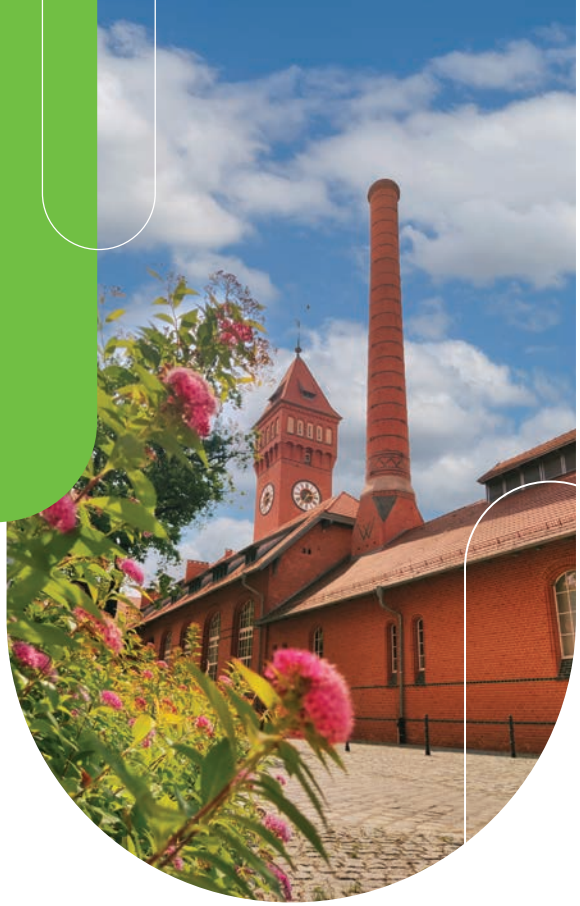
10:40 **Casper Asbjørn Eriksen, Jakob Lykke Andersen, Rolf Fagerberg and Daniel Merkle**
Reconciling Inconsistent Molecular Structures from Biochemical Databases

11:00 **Gatis Melkus, Sandra Siliņa, Andrejs Sizovs, Peteris Rucevskis, Lelde Lace, Edgars Celms and Juris Viksna**
Clique-based topological characterization of chromatin interaction hubs

11:15 **Jakob Lykke Andersen, Sissel Banke, Rolf Fagerberg, Christoph Flamm, Daniel Merkle and Peter F. Stadler**

On the Realisability of Chemical Pathways

Łukasiewicz
PORT
Polish Center
for Technology
Development



Łukasiewicz Research Network – PORT
Polish Center for Technology Development

147 Stabłowicka Street
54-066 Wrocław, Poland

port.lukasiewicz.gov.pl

Media

Biotechnologia.pl

Partners

NSF

Georgia State University

ISCB International Society for Computational Biology
Affiliated Conference

Industry Contact Point
Medical Technologies and Health

DOLNY ŚLĄSK
PATRONAT HONOROWY MARSZAŁKA WOJEWÓDZTWA DOLNOŚLĄSKIEGO CEZAREGO PRZYBYLSKIEGO