Computational and Mathematical Bioinformatics and Biophysics (CMBB2019)

Tsinghua Sanya International Mathematics Forum

December 9-13, 2019

TSIMF

清华三亚国际数学论坛

The document in this publication may meet the minimum requirements of the American National Standard for Information Sciences — Permanence of Paper for Printed Library Materials, ANSI Z39.48–1984.

Organizers

Professor Guowei Wei, Department of Mathematics, Michigan State University
Professor Stephen S.-T. Yau, Department of Mathematical Sciences, Tsinghua University
Dr. Changchuan Yin, Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago
Professor Shan Zhao, Department of Mathematics, University of Alabama

清华三亚国际数学论坛
TSINGHUA SANYA INTERNATIONAL MATHEMATICS FORUM

Contents

Monday (Dec. 9)

Data-driven approaches to RNA computational biology

Shi-Jie Chen
Department of Physics, Department of Biochemistry,
and Institute of Data Sciences & Informatics
University of Missouri
Columbia, MO 65211, USA
Email: chenShi@missouri.edu

Abstract

RNA molecules play spectacularly versatile roles in living cells.Emerging biomedical advances such as precision medicine and synthetic biology, all point to RNA as the central regulators and information carriers. Furthermore, the ever-increasing database for non-coding RNAs inspire a great variety of RNA-based therapeutic strategies. RNA functions depend on precisely folded RNA structures. However, currently the number of available structures deposited in the structure database such as PDB is only a small fraction of all the structures that we would like to know. This gap has to be closed by computational methods. With the long-term goal of predicting three-dimensional structure from the nucleotide sequence and rational design of RNA-based drugs, we have systematically developed data-driven and data-drive/physics-based hybrid methods for important RNA biology problems such as the prediction of RNA three-dimensional structures from the sequence and metal ion-RNA interactions. I will discuss our recently developed new methods in addressing the above problems and the biomedical applications of these methods.

Jianshu Cao
Computational Science Research Center(CSRC)
Beijing, China
Email: jianshu@mit.edu

Prediction of RNA secondary structure using deep learning approach

Yi Xiao and Kangkun Mao
School of Physics
Huazhong University of Science and Technology
Wuhan 430074, Hubei, China
Email: yxiao@hust.edu.cn

Abstract

Non-coding RNA plays important roles in cell and their secondary structures are vital for understanding their tertiary structures and functions. There are several computational methods for the prediction of RNA secondary structure, but it is still challenging to reach high accuracy, especially for those with pseudoknots. Traditional prediction methods are mainly divided into two categories: single-sequence method and homologous-sequence method. Previous researches showed that the homologous-sequence method usually achieved higher precision, because the evolutionary information embedded within homologous sequences provided a rich source of data for inferring structural constraints. But its disadvantages are also obvious, many RNAs of interest lack sufficient numbers of related sequences, leading to noisy, error base pair predictions. Here we present an end-to-end deep learning-based approach to address this problem, filling the gap between these two different kinds of methods. Our method combines two famous neural network architecture bidirectional-LSTM and U-net. It could use both the structural knowledge and evolutionary information learning from similar RNA sequences to predict the secondary structure of a RNA. Benchmarks show that on the testing dataset, our method can achieve state-of-the-art performance compared to current most popular prediction methods.

References

1. He, Xiaoling and Mao, Kangkun and Wang, Jun and Zeng, Chen and Xiao, Yi (2019). Comparison of two algorithms of direct coupling analysis of protein. Communications in Information and Systems, 19(1).

2. He, Xiaoling and Li, Shuaimin and Ou, Xiujuan and Wang, Jun and Xiao, Yi(2019). Inference of RNA structural contacts by direct coupling analysis. Communications in

Information and Systems, (accepted).

Most probable transition pathways in stochastic dynamics of a gene regulation system

Jinqiao Duan
Illinois Institute of Technology
Chicago, IL 60616, USA
Email: duan@iit.edu

Abstract

Dynamical systems arising in biophysics are often subject to random fluctuations. The noisy fluctuations may be Gaussian or non-Gaussian, which are modeled by Brownian motion or α-stable Levy motion, respectively. Non-Gaussianity of the noise manifests as nonlocality at a ’macroscopic’ level. Stochastic dynamical systems with non-Gaussian noise (modeled by α-stable Levy motion) have attracted a lot of attention recently. The non-Gaussianity index α is a significant indicator for various dynamical behaviors.

The speaker will present recent work on analyzing and computing the most probable transition pathways, for stochastic dynamics of a gene regulation system.

Protein contact prediction and contact-assisted structure prediction

Wenzhi Mao, Wenze Ding, Yaoguang Xing and Haipeng Gong
School of Life Sciences
Tsinghua University
Beijing, China
Email: hgong@tsinghua.edu.cn

Abstract

Native contacts between residues could be predicted from the amino acid sequence of proteins, and the predicted contact information could assist the de novo protein structure prediction. Here, we present a novel pipeline of a residue contact predictor AmoebaContact and a contact-assisted folder GDFold for rapid protein structure prediction. Unlike mainstream contact predictors that utilize simple, regularized neural networks, AmoebaContact adopts a set of network architectures that are optimized for contact prediction through automatic searching and predicts the residue contacts at a series of cutoffs. Different from conventional contact-assisted folders that only use top-scored contact pairs, GDFold considers all residue pairs from the prediction results of AmoebaContact in a differentiable loss function and optimizes the atom coordinates using the gradient descent algorithm. Combination of AmoebaContact and GDFold allows quick modeling of the protein structure, with comparable model quality to the state-of-the-art protein structure prediction methods.

Physical basis of protein liquid-liquid phase separation

Huan-Xiang Zhou
Department of Chemistry and Department of Physics
University of Illinois at Chicago
Chicago, IL 60607, USA
Email: hzhou43@uic.edu

Abstract

Intracellular membraneless organelles, corresponding to the droplet phase upon liquid-liquid phase separation (LLPS) of mixtures of proteins and possibly RNA, mediate myriad cellular functions [1]. Cells use a variety of biochemical signals such as expression level and posttranslational modification to regulate droplet formation and dissolution. Our study focuses on elucidating the physical basis of phase behaviors associated with cellular functions of membraneless organelles, using four complementary approaches. First, we use colloids and polymers, respectively, as models for structured and disordered proteins, to investigate both the common basis for protein phase separation and the unique characteristics of structured and disordered proteins in LLPS [2]. Disordered proteins are characterized by both extensive intermolecular attraction and low excluded-volume entropy, contributing to ready observation of phase separation. Second, we use multi-component patchy particles to investigate the wide range of effects of regulatory components on the droplet formation of driver proteins [3]. Third, the theoretical predictions have motivated our experimental work to define archetypical classes of macromolecular regulators of LLPS [4]. Lastly, we have developed a powerful computational method called FMAP for determining liquid-liquid phase equilibria [5]. By using fast Fourier transform to efficiently evaluate protein-protein interactions, FMAP enables an atomistic representation of the protein molecules. Application to -crystalins reveals how minor variations in amino-acid sequence, similar to those from posttranslational modifications and disease-associated mutations, lead to drastic differences in critical temperature. These studies contribute to both qualitative and quantitative understanding on the phase behaviors of membraneless organelles and their regulation and dysregulation.

References

1. S. Qin and H.-X. Zhou (2017). Protein folding, binding, and droplet formation in

cell-like conditions. Curr. Opin. Struct. Biol. 43, 28-37.

2. H.-X. Zhou, V. Nguemaha, K. Mazarakos, and S. Qin (2018). Why do disordered and structured proteins behave differently in phase separation? Trends Biochem. Sci. 43, 499-516.

3. V. Nguemaha and H.-X. Zhou (2018). Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation. Sci. Rep. 8, 6728.

4. A. Ghosh, K. Mazarakos, and H.-X. Zhou (2019). Three archetypical classes of macromolecular regulators of protein liquid-liquid phase separation. Proc. Natl. Acad. Sci. USA 116, 19474-19483.

5. S. Qin and H.-X. Zhou (2016). Fast method for computing chemical potentials and liquid-liquid phase equilibria of macromolecular solutions. J. Phys. Chem. B. 120, 8164-8174.

Exploiting ligand 3D shape similarity for computational structure based drug design

Kam Y. J. Zhang
Laboratory for Structural Bioinformatics
Center for Biosystems Dynamics Research
RIKEN, 1-7-22 Suehiro, Tsurumi
Yokohama, Kanagawa 230-0045, Japan
Email: kamzhang@riken.jp

Abstract

To predict the binding pose of a ligand in a target protein is a key step in virtual screening and computational structure based drug design. Both sufficient conformational sampling and accurate scoring function are required for successful pose prediction. To improve the accuracy of pose prediction by tackling the sampling problem, we have sought to exploit the ever-increasing amount of small molecule ligand and protein complexes in Protein Data Bank. We have developed a method of pose prediction using 3D shape similarity. It first places a ligand conformation of the highest 3D shape similarity with known crystal structure ligands into protein binding site and then refines the pose by various strategies. We have prospectively assessed our methods in several D3R Grand Challenges. Overall, our results demonstrated that ligand 3D shape similarity with the crystal ligand is sufficient to predict binding poses of new ligands with acceptable accuracy.

Computational methods for solving nonlinear systems arising from biophysics

Wenrui Hao
Department of Mathematics
Penn State University
University Park, PA 16802, USA
Email:wxh64@psu.edu

Abstract

This talk will cover some recent progress on computational methods to solve nonlinear systems arising from math biological models. I will start with homotopy methods for solving nonlinear PDEs with multiple solutions and bifurcations by coupling with domain decomposition and reduced basis methods. Examples from tumor growth models and pattern formation will be used to demonstrate the ideas. Then a homotopy method will be introduced for solving maximum entropy problem to reconstruct probability density functions based on empirical data. Applications to cardiovascular multi-scale modeling will be illustrated by coupling with machine learning techniques.

Whole genome phylogeny of giant viruses by Fourier transform

Changchuan Yin[1], Stephen S.-T. Yau[2*]
1. Department of Mathematics, Statistics, and Computer Science
The University of Illinois at Chicago, Chicago, IL 60607-7045, USA
2. Department of Mathematical Sciences
Tsinghua University, Beijing, China
*Corresponding author, Email: yau@uic.edu

Abstract

Dozens of giant viruses have been surprisingly discovered since 2003. Giant viruses are notably larger than typical bacteria and have extremely large genomes that encode thousands of genes. Because the giant viruses have super-sized shapes and genome sizes, they are distinguished from classical viruses and bacteria. The evolutionary origin of the giant virus is still an open question. The current phylogenetic studies on giant virus use specific genes or proteins, and can not elucidate the origins of the giant viruses from the global genome perspective. In this study, we perform the whole-genome phylogenetic analysis of the giant viruses using the Fourier transform-based alignment-free method. The phylogenetic analysis shows that the typical giant virus *Pandoravirus* and tailed giant virus *Tupanvirus* are closely related to archaea. This new finding suggests that giant viruses may origin from archaea and supports the reductive model on the giant virus evolution.

References

1.Yin, C., Chen, Y., & Yau, S. S. -T. (2014). A measure of DNA sequence similarity by Fourier Transform with applications on hierarchical clustering. Journal of Theoretical Biology, 359, 18-28.

2.Yin, C., & Yau, S. S. -T. (2015). An improved model for whole genome phylogenetic analysis by Fourier transform. Journal of Theoretical Biology, 382, 99-110.

3.Yin, C., & Yau, S. S. -T. (2019). Whole genome single nucleotide polymorphism genotyping of *Staphylococcus aureus*. Communications in Information and Systems, 19(1), 57-80.

Topological data analysis (TDA) based machine learning models for biomolecular data analysis

Kelin Xia
Nanyang Technological University, Singapore
Email: xiakelin@ntu.edu.sg

Abstract

Featurization or feature engineering is key to the success of machine learning models for chemical and biological systems. In this talk, we will discuss the topological data analysis (TDA) and its combination with machine learning models. Unlike traditional graph/network or geometric models, TDA characterizes only the intrinsic information, thus significantly reduces data complexity and dimensionality. In topology based machine learning models, biomolecular topological fingerprints are extracted by using persistent homology, which is the most important tool in TDA. These topological fingerprints are then transformed into feature vectors and inputted into machine learning models, including SVM, random forest, CNN, etc. We will discuss the application of TDA-based models in the analysis of molecular aggregations, hydrogen-bonding networks, molecular dynamic simulations, and drug design.

Neuro-network statistical energy functions of protein conformations and completely flexible protein backbone design

Haiyan Liu
School of Life Sciences
University of Science and Technology of China
Email: liu@math.psu.edu

Abstract

A general method to sample and optimize protein backbones without specific sequence information is much needed for computational protein design. A viable solution would be molecular simulations driven by an energy function that can faithfully recapitulate the characteristically coupled distributions of multiplexes of local and non-local conformational variables in designable backbones. It is desired that the energy surfaces are continuous and smooth, with easily computable gradients. We report an energy function named SCUBA, standing for Side-Chain-Unspecialized-Backbone-Arrangement. SCUBA uses neural networks (NN) learned from known protein structures to analytically represent high-dimensional statistical energy surfaces. Each NN term is derived by first estimating the statistical energies in a multi-variable space via neighbor-counting (NC) with adaptive cutoffs, and then training the NN with the NC-estimated energies. The weights of the different energy components are calibrated on the basis of SCUBA-driven stochastic dynamics (SD) simulations of natural proteins. We apply SCUBA SD simulated annealing to optimize artificially constructed polypeptide backbones of different fold classes. For a majority of the resulting backbones, structurally matching native backbones can be found with Dali Z-scores above 6 and less than 2 Å displacements of main chain atoms in aligned secondary structures. In one case, we have experimentally determined the structure of a protein with a SCUBA-designed backbone and an ABACUS-designed sequence. The results suggest that SCUBA-driven sampling and optimization can be a general tool for protein backbone design with complete conformational flexibility. In addition, the NC-NN approach can be generally applied to develop continuous, noise-filtered multi-variable statistical models from struc-

tural data. Linux executables with user instructions and demos of SCUBA-SD are available at http://biocomp.ustc.edu.cn/servers/download_scuba.php.

Tuesday (Dec. 10)

Reproducing ensemble averaged electrostatics with super-Gaussian-based smooth dielectric function:Application to polar solvation energy of proteins and electrostatic component of binding energy of protein complexes

Shailesh Panday[1], Mihiri Hewa[1], Arghya Chakraborty[1], Shan Zhan[1], Emil Alexov[2*]
1. Department of Mathematics,
The University of Alabama
Tuscaloosa, AL 35487-0350 USA
2. Department of Physics and Astronomy,
Clemson University, Clemson, SC 29634-0978 USA
* Email: ealexov@g.clemson.edu

Abstract

Proteins constantly sample various conformations as they carry their biological function, including interacting with their partners, and this should be taken into account in any numerical protocol aiming at computing their thermodynamic properties. Here we report an application of previously reported Super-Gaussian-based smooth dielectric function to reproduce ensemble averaged electrostatics. This is an important achievement, since it dramatically reduces the computational demand for MMPPBSA applications and bypasses the necessity of long molecular dynamics simulations. Instead, a single frame, typically energy-minimized structure, in conjunction of Super-Gaussian-based smooth dielectric function, as implemented in DelPhi, can deliver ensemble averaged quantities. The approach is tested against ensemble averaged polar solvation energy of monomeric proteins and electrostatic component of binding free energy of protein-protein complexes. It is demonstrated that Super-Gaussian-based smooth dielectric function reproduces ensemble averaged quantities, resulting in correlation coefficients of about 0.8 and slope of the fitting line of 1.0.

A network-based integrated framework for predicting virus-host interactions

Fengzhu Sun, PhD
Quantitative and Computational Biology
University of Southern California
Los Angeles, CA 90089-2910, USA
http://www-rcf.usc.edu/ fsun/

Abstract

Viruses play important roles in controlling bacterial population size, altering host metabolism, and have broader impacts on the functions of microbial communities, such as human gut, soil, and ocean microbiomes. However, the investigations of viruses and their functions were vastly underdeveloped. Identifying the hosts of viruses is an essential problem in virus studies. We previously used alignment-free sequence comparison statistic d2∗ to identify the hosts of viruses. Recently we developed an integrative Markov random field based approach to combine alignment-free sequence comparison, alignment, CRISPR, etc. to predict virus hosts. We applied our host-prediction tool to three metagenomic virus datasets: human gut crAss-like phages, marine viruses, and viruses recovered from globally-distributed, diverse habitats. Host predictions were frequently consistent with those of previous studies, but more importantly, this new tool made many more con_dent predictions than previous tools, up to 6-fold more (n>60,000), greatly expanding the diversity of known virus-host interactions.

References

1.Ahlgren, N. A., Ren, J., Lu, Y. Y., Fuhrman, J. A., & Sun, F. (2016). Alignment-free oligonucleotide frequency dissimilarity measure improves prediction of hosts from metagenomically-derived viral sequences. Nucleic Acids Research, 45(1), 39-53.

2.Wang, W., Ren, J., Ahlgren, N. A., Fuhrman, J. A., Braun, J., & Sun, F. (2018). A network-based integrated framework for predicting virus-host interactions with applications. bioRxiv, 505768.

Pharmacological Modeling in Thrombosis

Limei Cheng[1], Guo-Wei Wei[2] and Tarek Leil[1]
1. Clinical Pharmacology and Pharmacometrics
Bristol-Myers Squibb, NJ 08540, USA
2. Department of Mathematics
Michigan State University
East Lansing, MI 48824, USA

Abstract

Hemostasis and thrombosis are often thought of as two sides of the same clotting mechanism whereas hemostasis is a natural protective mechanism to prevent bleeding and thrombosis is a blood clot abnormally formulated inside a blood vessel, blocking the normal blood ow. The evidence to date suggests that at least arterial thrombosis results from the same critical pathways of hemostasis. Analysis of these complex processes and pathways using quantitative systems pharmacological model-based approach can facilitate the delineation of the causal pathways that lead to the emergence of thrombosis. In this paper, we provide an overview of the main molecular and physiological mechanisms associated with hemostasis and thrombosis and review the models and quantitative system pharmacological modeling approaches that are relevant in characterizing the interplay among the multiple factors and pathways of thrombosis. Emphasis is given to computational models for drug development. Future trends are discussed.

Efficient sampling and optimization on manifolds for macromolecular docking

Dmytro Kozakov
Department of Applied Mathematics & Statistics
Stony Brook University
Stony Brook, NY 11794-3600, USA
Email: dmytro.kozakov@stonybrook.edu

Abstract

Three-dimensional structure prediction of macromolecular interaction complex is an important component in small molecular and biologics drug discovery. The search space includes the 6D rotational/translational space of mutual rigid body orientations of receptor and ligand, as well as additional degrees of freedom that represent the flexibility of the two molecules. Solving this problem requires detailed sampling and optimization of an energy-based scoring function.

Since the energy function has a large number of local minima separated by high barriers, the minimization problem is extremely challenging. The search space includes the 6D rotational/translational space as well as additional degrees of freedom that represent the flexibility of the macromolecules and is a manifold. Here we present effective approaches for different steps of docking protocols, which effectively use manifold geometry to significantly speed up the search. Specifically we will describe Fast Manifold Fourier Transform (FMFT) approach for effective global grid based sampling for macromolecular docking, and local and medium range optimization using exponential map parametrization for docking refinement. The methods described above have been blindly validated in international docking competitions CAPRI (protein docking) and D3R (protein-ligand docking) and were among the best performers in both.

Jie Wu
Department of Mathematics
National University of Singapore, Singapore
Email: matwuj@nus.edu.sg

Multimer protein-protein complex topology and structure prediction

Xinqi Gong 龚新奇
Institute for Mathematical Sciences
Renmin University of China
No.59 Zhongguancun Street, Haidian Distric
Beijing 100872, P.R. China
中国人民大学数学科学研究院
北京市海淀区中关村大街59号，100872
邮箱：xinqigong@ruc.edu.cn cn

Abstract

Theoretical understanding of the structurally determining factors of interaction sites will help to understand the underlying mechanism of protein-protein interactions. Taking advantage of advanced mathematical methods to correctly predict interaction sites will be useful. Although some previous studies have been devoted to the interaction interface of protein monomer and the interface residues between chains of protein dimers; very few studies about the prediction of protein multimers, including trimers, multimer and even more proteins in a large protein complex. Many proteins function with the form of multibody protein complexes. And the complexity of the protein multimers structure causes the difficulty of interface residues prediction on them. So, we hope to build a method for the prediction of protein multimer interface residue pairs. We developed a new deep network based on graph convolutional network combining LSTM network to predict protein multimers interaction interface residue pairs. On the account of the protein structure data is not the same as the image or video data which is well-arranged matrices, namely the Euclidean Structure mentioned in many researches. Because the Non-Euclidean Structure data can’t keep the translation invariance, and we hope to extract some spatial features from this kind of data applying on machine learning, a Graph Convolutional Network algorithm was developed to predict the interface residue pairs of protein interactions based on a topological graph building a relationship between vertexes and edges in graph theory combining multilayer Long Short-Term Memory network. First, selecting the training and test samples from the Protein

Data Bank, and then extracting the physicochemical property features and the geometric features of surface residue associated with interfacial properties. Subsequently, we transform the protein multimers data to topological graphs and predict protein interaction interface residue pairs using the Graph Convolutional model. In addition, different types of evaluation indicators verified its validity. Furthermore, we developed a conditional Generative Adversarial Network (GAN) direct coupling analysis method. It takes predicted contact map from other methods as initial input, which can learn from the predicted dca-based contact map and generate a new improved one more close to that of the native inter-protein. Comparisons against other methods show that our GAN network can enhance the performance of contacts map prediction very much.

Haibao Duan
The Institute of Mathematics
Chinese Academy of Sciences
Email: dhb@math.ac.cn

Yanying Wang
College of Mathematics and Information Science
Hebei Normal Universityy
Email: yywang@hebtu.edu.cn

Comparisons of BRAT1 gene variants

En-Bing Lin
Department of Mathematics, Department of Biology
Central Michigan University
Mt. Pleasant, MI 48859, USA
Email: lin1e@cmich.edu

Abstract

BRAT1 (BRACA1 Associated ATM Activator 1) is a protein coding gene that interacts with and activates the tumor suppressor gene BRCA1, a protein complex that repairs DNA damage due to ionizing radiation. In this paper, we obtain approximation and detail information of the numerical representation, i.e. wavelet transformation with Daubechies 2 and Coiflet, of the gene variants. We compare the computational and graphical results of these gene variants to each other, as well as to those obtained from BLAST (Basic Local Alignment Search Tool). Thus, we can compare the results of an alignment free search method (wavelet) to that of an alignment dependent search method (BLAST). In general, both methods located the same areas of similarity but the results of the wavelet analysis provide numerical and visual comparisons. We also conclude with some other advantages of using wavelet method.

Sparse representation of Gaussian molecular surface

Minxin Chen
Department of Mathematics
Soochow University
Suzhou, Jiangsu, China
Email: chenminxin@suda.edu.cn

Abstract

In this talk, we propose a model and algorithm for sparse representing Gaussian molecular surface.By solving a nonlinear L1 optimization problem, the original Gaussian molecular surface is approximated by a relatively small number of radial basis functions (RBFs) with rotational ellipsoid feature. Experimental results demonstrate that the original Gaussian molecular surface is able to be represented with good accuracy by much fewer RBFs using our L1 model and algorithm. The sparse representation of Gaussian molecular surface is useful in various applications, such as molecular structure alignment, and this method in principle can be applied to sparse representation of general shapes and surfaces.

Wednesday (Dec. 11)

Phylogenetic inferences of viral and bacteria genomes: ancient taxons are distinct from their extant sister taxons in fast mutating adaptive alleles but share slow evolving alleles that are phylogenetically informative

Shi Huang
Center for Medical Genetics and Hunan Key Lab of Medical Genetic
Central South University
Changsha, Hunan, China
Email: huangshi@sklmg.edu.cn

Abstract

Present popular methods of molecular phylogenetic inferences are based on the neutral theory and the molecular clock hypothesis, and regards nearly all of the genome to be equally informative in revealing phylogenetic affinities. This is however intuitively unsound as many variants are involved in adaptation. Thus, shared variants may not be due to common ancestry but due to shared adaptation. Such variants must be excluded in phylogenetic inferences but are largely ignored by the field, leading to numerous absurd conclusions, which are further exaggerated by ancient DNA studies. Ancient DNAs differ from extant samples in numerous adaptive variants as ancient environments are very different from today's. If such variants are not excluded, ancient taxons would be found to be extinct today or as outgroup to extant species. We here demonstrate this by using viruses and bacteria. We studied the published ancient genomes of HBV, Parvovirus, and the Plague bacteria Yersinia pestis. While studies not excluding the fast mutating adaptive variants all consistently showed these ancient species to be extinct, our study removed those variants and found the ancient samples and extent taxon to be genetically continuous. The results invalidate the unrealistic assumptions of the presently popular phylogenetic methods and confirm the notion that genetic variations are largely at optimum balance maintained by natural selection as described by our maximum genetic diversity (MGD) theory.

References

Hu, T., Long , M., Yuan D., Zhu Z., Huang, Y., and Huang, S. (2013) The genetic equidistance result: misreading by the molecular clock and neutral theory and reinterpretation nearly half of a century later. Sci China Life Sci, 56:254-261.

Huang, S. (2012) Primate phylogeny: molecular evidence for a pongid clade excluding humans and a prosimian clade containing tarsiers. Sci China Life Sci, 55: 709-725.

Huang, S. (2016) New thoughts on an old riddle: what determines genetic diversity within and between species. Genomics, 108: 3-10. doi:10.1016/j.ygeno.2016.01.008

Lei, X., Yuan, D., and Huang, S. (2018) Collective effects of common SNPs and risk prediction in lung cancer. Heredity, doi:10.1038/s41437-018-0063-4

Yuan, D., Lei, X., Gui, Y., Zhu, Z., Wang, D. Yu, J., and Huang, S. (2017) Modern human origins: multiregional evolution of autosomes and East Asia origin of Y and mtDNA. bioRxiv. doi: https://doi.org/10.1101/101410

John Zhang
School of Chemistry and Molecular Engineering
East China Normal University
Email: zhzhang@phy.ecnu.edu.cn

Generative network complex (GNC) for drug discovery

Christopher Grow, Kaifu Gao, Duc Duy Nguyen and Guo-Wei Wei
Department of Mathematics
Michigan State University
East Lansing, MI 48824, USA
Email: weig@msu.edu

Abstract

It remains a challenging task to generate a vast variety of novel compounds with desirable druggable properties. In this work, a generative network complex (GNC) is proposed as a new platform for designing novel compounds, predicting their physical and chemical properties, and selecting potential drug candidates that fulfill various druggable criteria such as binding affinity, solubility, partition coefficient, clearance, etc. We combine a SMILES string generator, which consists of an encoder, a drug-property controlled or regulated latent space, and a decoder, with verification deep neural networks, a target-specific three-dimensional (3D) pose generator, and mathematical deep learning networks to generate new compounds, predict their drug properties, construct 3D poses associated with target proteins, and reevaluate druggability, respectively.

New compounds were generated in the latent space by either randomized output, controlled output, or optimized output. In our demonstration, 2.08 million and 2.8 million novel compounds are generated respectively for Cathepsin S and BACE targets. These new compounds are very different from the seeds and cover a larger chemical space. For potentially active compounds, their 3D poses are generated using a state-of-the-art method. The resulting 3D complexes are further evaluated for druggability by a championing deep learning algorithm based on algebraic topology, differential geometry, and algebraic graph theories. Performed on supercomputers, the whole process took less than one week.

This work was supported by NIH, NSF, Pfizer, Bristol-Myers Squibb, and Michigan State University.

Progress on protein structure prediction by deep learning

Jinbo Xu
Toyota Technological Institute at Chicago
Chicago, IL 60637, USA
Email: jinbo.xu@gmail.com

Abstract

Accurate description of protein structure and function is a fundamental step towards understanding biological life and highly relevant in the development of therapeutics. Although greatly improved, experimental protein structure determination is still low-throughput and costly, especially for membrane proteins. As such, computational structure prediction is often resorted. Predicting the structure of a protein without similar experimental structures is very challenging and usually needs a large amount of computing power. This talk will present the deep learning method (i.e., deep convolutional residual neural network) we have invented for protein contact and distance prediction that won the CASP (Critical Assessment of Structure Prediction) in both 2016 and 2018 in the category of contact prediction. In this talk we show that by using this powerful deep learning technique, even with only a personal computer we can predict the structure of a protein much more accurately than ever before. In particular, we predicted correct folds for the 3 largest hard targets ($\sim$ 350 amino acids) in CASP13 (2018) and generated the best 3D models for two of them among all the human and server groups including DeepMind's AlphaFold. Inspired by our success in CASP12 in 2016, this deep learning technique has been adopted widely by the structure prediction community and thus, resulted in the widespread, largest progress in the history of CASP , which will also be discussed in this talk.

From dinucleotide to chromatin, a phase separation perspective for chromatin structure change in development, differentiation, senescence and certain diseases

Yi Qin Gao
College of Chemistry & Molecular Engineering,
and Biomedical Pioneering Innovation Center
Peking University
Beijing, 100871, China
Shenzhen Bay Lab, Shenzhen, China
Email: gaoyq@pku.edu.cn

Abstract

The high-order chromatin structure plays an important role in gene regulation. The mechanism, especially the sequence dependence for the formation of varied chromatin structures in different cell states remain to be elucidated. In this talk, we try to touch on three questions: (1) What is the sequence dependence and chemical structure basis in the formation of high order chromatin structure, such as compartments? (2) How does the chromatin structure reflect the biological function of different cellular states and tissue-specificity? (3) How does this sequence-dependent chromatin structure formation manifest in different species? We identified CGI (CpG island) forest and prairie genomic domains based on CGI densities, and divided the genome into two sequentially, epigenetically, and transcriptionally distinct regions. These two megabase-sized domains were found to spatially segregate, and to different extents in different cell types. Forests and prairies show enhanced segregation from each other in development, differentiation, and senescence, meanwhile the multi-scale forest-prairie spatial intermingling is cell-type specific and increases in differentiation, helping to define cell identity. We propose that the phase separation of the 1D mosaic sequence in the 3-D space serves as a potential driving force, and together with cell type specific epigenetic marks and transcription factors, shapes the chromatin structure in different cell types. Specifically, based on the analysis of latest published Hi-C data of post-implantation

stages, we present a consistent view of the chromatin structural change and the corresponding sequence dependence. The forests and prairies show systematic and overall increase of spatial segregation during embryonic development, but with notable mixing occurring at two stages, ZGA and implantation. The segregation level change largely coincides with the change of genetic and epigenetic properties, leading to a possible mechanism of functional realization during implantation was proposed. Interestingly, body temperature changes coincide with the change in chromatin segregation, implying that temperature is a possible factor influencing chromatin phase separation and global chromatin structure formation.

Thursday (Dec. 12)

Machine learning of the rate constants for the reaction between alkanes and hydrogen/oxygen atom

Junhui Lu[1,2†], Jinhui Yu[1,2†], Hongwei Song[1], and Minghui Yang[1,2,3,*]
1. Key Laboratory of Magnetic Resonance in Biological Systems,
State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics,
National Center for Magnetic Resonance in Wuhan,
Wuhan Institute of Physics and Mathematics,
Chinese Academy of Sciences, Wuhan 430071, China
2. University of Chinese Academy of Sciences, Beijing 100049, China
3. Wuhan National Laboratory for Optoelectronics,
Huazhong University of Science and Technology, Wuhan
[†]They made equal contributions to this work.
[*]Corresponding author, Email:yangmh@wipm.ac.cn

Abstract

The reaction rate constant is of important meaning in modeling combustion and biochemistry reaction network. The rate constants could be determined by means of experimental measurement or theoretical computation. This work presented a machine learning approach to construct neural network (NN) models for the training and predicting of rate constants and employed the approach to two kinds of hydrogen abstraction chemical reactions: hydrogen + Alkanes (HR) and Oxygen + Alkanes (OR). Each reaction is described with five parameters: three parameters were used to describe the molecular structure of reactant alkane, one parameter is the location of the breaking C-H bond and one is the reaction temperature. The obtained NN models showed satisfying results in training and validating calculation for eight HR and eleven OR reactions. We also tested the prediction ability of these NN model and found the derivations are close to that with transition state theory. This work showed that machine learning could be a good choice for the calculation of chemical reaction rate constant

Combining molecular dynamics and machine learning for membrane protein studies

Chen SONG
Center for Quantitative Biology
& Peking-Tsinghua Center for Life Sciences
Peking University,
Beijing 100871, China
Email: c.song@pku.edu.cn

Abstract

Membrane proteins are essential for signal transduction and substrate transport across cell membranes. Predictions of the structure and function mechanisms of membrane proteins are of particular importance. We have been trying to combine molecular dynamics (MD) simulations and deep learning to investigate the structure and dynamics of membrane proteins. Our preliminary results indicate that MD simulations can provide structural and dynamic information to characterize the unique feature of membrane proteins. The MD results can be used as the training data for deep learning-based property prediction, which can significantly improve the prediction accuracy of membrane protein structures.

A novel approach to clustering genome sequences using inter-nucleotide covariance

Rui Dong, Stephen S.-T. Yau *
Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China
*Corresponding author, Email: yau@uic.edu

Abstract

Classification of DNA sequences is an important issue in the bioinformatics study, yet most existing methods for phylogenetic analysis including Multiple Sequence Alignment (MSA) are time-consuming and computationally expensive. The alignment-free methods are popular nowadays, whereas the manual intervention in those methods usually decreases the accuracy. Also, the interactions among nucleotides are neglected in most methods. Here we propose a new Accumulated Natural Vector (ANV) method which represents each DNA sequence by a point in R18. By calculating the Accumulated Indicator Functions of nucleotides, we can further find an Accumulated Natural Vector for each sequence. This new Accumulated Natural Vector not only can capture the distribution of each nucleotide, but also provide the covariance among nucleotides. Thus global comparison of DNA sequences or genomes can be done easily in R18. The tests of ANV of datasets of different sizes and types have proved the accuracy and time-efficiency of the new proposed ANV method.

Fast and accurate genome comparison using genome images: The extended natural vector method

Shaojun Pei, Stephen S.-T. Yau [*]
Department of Mathematical Sciences
Tsinghua University
Beijing 100084, China
[*]Corresponding author, Email: yau@uic.edu

Abstract

Using numerical methods for genome comparison has always been of importance in bioinformatics. The Chaos Game Representation (CGR) is an effective genome sequence mapping technology, which converts genome sequences to CGR images.To each CGR image,we associate a vector called an Extended Natural Vector(ENV). The ENV is based on the distribution of intensity values. This mapping produces a one-to-one correspondence between CGR images and their ENVs. We define the distance between two DNA sequences as the distance between their associated ENVs. We cluster and classify several datasets including Influenza A viruses, Bacillus genomes, and Conoidea mitochondrial genomes to build their phylogenetic trees. Results show that our ENV combining CGR method (CGR-ENV) compares favorably in classification accuracy and efficiency against the multiple sequence alignment (MSA) method and other alignment-free methods. The research provides significant insights into the study of phylogeny, evolution, and efficient DNA comparison algorithms for large genomes.

Evolutionary dynamics of cancer: from epigenetic regulation to cell population dynamics

Jinzhi Lei
Zhou Pei-Yuan Center for Applied Mathematics
Tsinghua University
Beijing, China
Email: jzlei@tsinghua.edu.cn

Abstract

Cancer is a group of diseases involving abnormal cell growth with the potential to invade or spread to other parts of the body. Cancer development is a long-term process which remains mostly unknown; predictive modeling of the evolutionary dynamics of cancer is one of the major challenges in computational cancer biology. In this talk, I introduce a general mathematical framework for understanding the behavior of heterogeneous stem cell regeneration, and the application of the model framework to study the evolutionary dynamics of cancer. The proposed model framework generalizes the classical G0 cell cycle model, incorporates the epigenetic states of stem cells that are represented by a continuous multidimensional variable, and the kinetic rates of cell behaviors, including proliferation, differentiation, and apoptosis, which are dependent on their epigenetic states. The random transition of epigenetic states is represented by an inheritance probability that can be described as a conditional beta distribution. Moreover, the model framework can be extended to investigate gene mutation-induced tumor development. The model equation further suggests a numerical scheme of multiscale modeling for tissue growth where a multiple cell system is represented by a collection of epigenetic states in each cell. We applied the numerical scheme to model the two processes of inflammation-induced tumorigenesis and tumor relapse after CD19 chimeric antigen receptor(CAR) T cell therapy of acute B lymphoblastic leukemia (B-ALL). Model simulations reveal the multiple pathways of inflammation-induced tumorigenesis, and the a mechanism of tumor relapse due to leukemic cell plasticity induced by CAR-T therapy stress.

References

1.Lei, J. (2019). A general mathematical framework for understanding the behavior of

heterogeneous stem cell regeneration. arXiv preprint arXiv:1903.11448.
2.Liu, J., Song, Y., & Lei, J. (2019). Single-cell entropy to quantify the cellular transcriptome from single-cell RNA-seq data. bioRxiv, 678557.
3.Lei, J., Levin, S. A., & Nie, Q. (2014). Mathematical model of adult stem cell regeneration with cross-talk between genetic and epigenetic regulation. Proceedings of the National Academy of Sciences, 111(10), E880-E887.
4.Guo, Y., Nie, Q., MacLean, A. L., Li, Y., Lei, J., & Li, S. (2017). Multiscale modeling of inflammation-induced tumorigenesis reveals competing oncogenic and oncoprotective roles for inflammation. Cancer research, 77(22), 6429-6441.

Similarity analysis of protein sequences using a reduced k-mer amino acid model

Jia Wen[1*], Yuyan Zhang[2]
1. School of Information Engineering,
Suihua University, Suihua 152061, China
2. School of Agriculture and hydraulic Engineering,
Suihua University, Suihua 152061, China
*Email: wenjia198021@126.com

Abstract

Based on the properties of amino acid side chain, the 20 natural amino acids are divided into a simplified feature space, and the original protein sequence could be represented by a reduced amino acid sequence, which contains only four residues. Associating with this reduced protein sequence representation, the k-mer natural vector is defined and utilized to describe the similarity analysis of protein sequences, in which the frequencies and positional information of k-mers appearing in a reduced amino acid sequence are characterized by a feature vector. The similarity analysis of protein sequences can be easily and fast performed without requiring evolutionary models or human intervention. In order to show the utilities of our new method, it is applied on the real protein datasets for similarity analysis, and the obtaining results demonstrate that our new approach can precisely describe the similarities of protein sequences, and also strengthen the computing efficiency, compared with multiple sequence alignment. Therefore, our reduced k-mer amino acid representation model is a very powerful tool for analyzing and annotating protein sequence.

Vector bundles and characteristic classes

Fei Han
Department of Mathematics
National University of Singapore, Singapore
Email: mathanf@nus.edu.sg

Abstract

Vector bundle is such construction that on every point in a topological space, a vector space is assigned and in a neighborhood of any point, such assignment is just a product with the vector space. In this talk, I will briefly review such structure and introduce the Characteristic classes of a vector bundle, a system of topological invariants taking values in the cohomology of the space that can be used to distinguish the vector bundles topologically.

Fengchun Lei
School of Mathematical Sciences
Dalian University of Technology
Email: fclei@dlut.edu.cn

Identifying zero−inflated distributions with a new R package iZID

Lei Wang*, Hani Aldirawi*, and Jie Yang †
Department of Mathematics, Statistics, and Computer Science
The University of Illinois at Chicago, Chicago, IL 60607-7045, USA
*These two authors contributed to this paper equally.
†Email: jyang06@uic.edu

Abstract

Count data with a large portion of zeros arise naturally in many scientific disciplines. When conducting one-sample Kolmogorov−Smirnov (KS) test for count data, the estimated p-value is biased due to plugging in sample estimates of unknown parameters. As a consequence, the result of a KS test could be too conservative. In the newly developed R package iZID for zero inflated count data,we use bootstrapped Monte Carlo estimates to overcome the bias issue in estimating p−values, as well as bootstrapped likelihood ratio tests for zero inflated model selection. Our new package also provides miscellaneous functions to simulate zero-inflated count data and calculate maximum likelihood estimates of unknown parameters. Compared with other R packages available so far, our package coverer more types of zero-inflated distributions and provides adjusted p −value estimates after incorporating the influence of unknown model parameters. To facilitate potential users, in this paper we provide detailed descriptions of functions in iZID and illustrate the use of them with executable R code.

Xiwu Yang
Department of Mathematics
Liaoning Normal University
Email: cinema@lnnu.edu.cn