Abstracts of Invited Talks

Energy-guided iterative approach to computational prediction of ligand-RNA interaction

Shi-Jie Chen, University of Missouri

Abstract: Ribonucleic acid (RNA) molecules play critical roles in gene regulation and expression at the level of transcription, splicing, and translation. RNA-ligand interactions can lead to significant changes in RNA folding stability and RNA structure, which lead to RNA-mediated gene regulation. Present computational strategies for the prediction of ligand-RNA interactions are limited by the incomplete sampling of ligand binding sites and ligand poses. Here we describe RLDOCK, a novel energy-guided, step-wise iterative focusing algorithm for the computational prediction of ligand binding sites and poses. This new approach might be combined with RNA folding models to predict not only bound pose of a flexible ligand but also bound structure of a flexible RNA upon ligand binding.

Importance of quantum effects in biomolecular simulations

Jiali Gao, Shenzhen Bay Laboratory

Abstract: Traditionally, empirical force fields or molecular mechanics have been used in essentially all computer simulaitons of biological systems. This model is efficient and offers an overall excellent description of intermolecular interactions thanks to persistent parameterization for more than a half century. Can one move beyond this paradigm in the future? Here, I will present findings from studies of nuclear quantum effects on the importance of treating the potential energy surface by an electronic structural approach explicitly, and I will offer an alternative, a quantum mechanical force field, for describing intermolecular interactions of biological systems.

Opportunities and challenges for AI and Math in drug discovery

Duc Nguyen, University of Kentucky

Abstract: Drug discovery is one of the most challenging tasks in the biological sciences since it requires over 10 years and costs more than $2.6 billion to put an average novel medicine on the marketplace. The abundant availability of biological data along with the flourishing advanced AI algorithms opens a future with great hope for discovering new drugs faster and cheaper. Unfortunately, AI faces an enormous obstacle in drug discovery due to the intricate complexity of biomolecular structures and the high dimensionality of biological datasets. In our lab, these challenges have been tackled mathematically. We have introduced multiscale modeling, differential geometry, algebraic topology, and graph theory-based models to systematically represent the diverse biological datasets in the low-dimensional spaces. Combining these mathematical representations with cutting edge deep neural networks, we arrived at novel models not only perform well on virtual-screening targeting important drug properties but also have the ability to design new drugs at an unprecedented speed. Our team has emerged as a top

winner in D3R Grand Challenges, a worldwide annual competition series in computer-aided drug design, in the past few years.

Functional multi-body protein interaction supercomplex structure prediction

Xinqi Gong, Renmin University of China

Abstract: I will talk about our recent progress on predicting multi-body protein interaction supercomplex structure, including mechanism mining, deep learing method development and experimental applications.

Target search of a protein on DNA in the presence of position-dependent bias

Jinqiao Duan, Illinois Institute of Technology
Xi Chen, Xi'an University of Finance and Economics

Abstract: We study the target search on DNA for proteins in the presence of non-constant drift. This search is realized by the facilitated diffusion process. Existing works on this problem focus on the case of constant drift. Starting from a non-local Fokker–Planck equation with the α-order fractional Laplace operator and a 'sink' term, we obtain the possibility density function for a protein occurring at position x at time t. Based on this, we further compute the survival probability and the first arrival density in order to quantify the searching mechanisms. The numerical results show that in the linear drift case, there is an optimal α index for the search to be most likely successful (searching reliability reaches its maximum). This optimal α index depends on the initial position–target separation. It is also found that the diffusion intensity plays a positive role in improving the search success. The nonlinear double-well drift could drive the protein to reach the target with a larger possibility than the linear drag at the initial time period, but viewed over a long time duration, the linear drift is more beneficial for target search success. In contrast to the linear drift case, the search reliability and efficiency with nonlinear double-well drift have a monotonic relationship with the α index, that is, the smaller the α index is, the higher the likelihood of a protein finding its target.

Latest development of protein structure prediction by deep learning

Jinbo Xu, Toyota Tech Inst at Chicago

Abstract: We describe our latest study of the deep convolutional residual neural networks (ResNet) for protein structure prediction, including deeper and wider ResNets, the efficacy of different input features, and improved 3D model building methods. Our ResNet can predict correct folds (TMscore>0.5) for 26 out of 32 CASP13 FM (template-free-modeling) targets and L/5 long-range contacts for these targets with precision over 80%, a significant improvement over the CASP13 results. Although co-evolution analysis plays an important role in a successful structure prediction method, we show that when co-evolution is not used, our ResNet can still predict correct folds for 18 of the 32 CASP13 FM targets including several large ones. This

marks a significant improvement over the top co-evolution-based, non-deep learning methods at CASP13, and other non-coevolution-based deep learning models, such as the popular recurrent geometric network (RGN). With only primary sequence, our ResNet can also predict correct folds for all 21 human-designed proteins we tested. In contrast, RGN predicts correct folds for only 3 human-designed proteins and zero CASP13 FM target. In addition, we find that ResNet may fare better for the human-designed proteins when trained without co-evolution information than with co-evolution. These results suggest that ResNet does not simply denoise co-evolution signals, but instead is able to learn important sequence-structure relationship from experimental structures. This has important implications on protein design and engineering especially when evolutionary information is not available.
The web server and software are available at http://raptorx.uchicago.edu/ and https://github.com/j3xugit/RaptorX-3DModeling/

Energy-guided iterative approach to computational prediction of ligand-RNA interaction

Qunfeng Dong, Loyola University Chicago

Abstract: Ribonucleic acid (RNA) molecules play critical roles in gene regulation and expression at the level of transcription, splicing, and translation. RNA-ligand interactions can lead to significant changes in RNA folding stability and RNA structure, which lead to RNA-mediated gene regulation. Present computational strategies for the prediction of ligand-RNA interactions are limited by the incomplete sampling of ligand binding sites and ligand poses. Here we describe RLDOCK, a novel energy-guided, step-wise iterative focusing algorithm for the computational prediction of ligand binding sites and poses. This new approach might be combined with RNA folding models to predict not only bound pose of a flexible ligand but also bound structure of a flexible RNA upon ligand binding.

The framework for population epigenetic study

Yongshuai Jiang, Harbin Medical University

Abstract: At present, understanding of DNA methylation at the population level is still limited. Here, we first extended the classical framework of population genetics, such as single nucleotide polymorphism allele frequency, linkage disequilibrium (LD), LD block and haplotype, to epigenetics. Then, as an example, we compared the DNA methylation disequilibrium (MD) maps between HapMap CEU (Caucasian residents of European ancestry from Utah) population and YRI (Yoruba people from Ibadan) population (lymphoblastoid cell lines). We analyzed the differences and similarities between CEU and YRI from the following aspects: SMP (single methylation polymorphism) allele frequency, SMP allele association, MD, MD block and methylation haplotype (meplotype) frequency. The results showed that CEU and YRI had similar distribution of SMP allele frequency, and shared many MD block region. We believe that the framework of population genetics can be used in the population epigenetics. The population epigenetic framework also has potential prospects in the study of complex diseases, such as epigenome-wide association study.

Genetic risk for subsequent breast cancer among female survivors of childhood cancer

Zhaoming Wang, St. Jude Children's Research Hospital

Abstract: Due to advancement in cancer therapy, greater than 80% of children diagnosed with cancer will survivor five or more years. It is estimated there are half a million childhood cancer survivors currently living in the US. Survivors are at increased risk of subsequent neoplasms, mostly considered therapy related. Breast cancer is one of the most common subsequent neoplasms among female survivors. It is known that radiation dose and volume to the breast is a major risk factor, leading to 20-fold increased risk with rate of cumulative incidence reaching 30% by age of 50 years. For non-irradiated survivors, literature also reported the risk of chemotherapy agents including anthracyclines. However, the genetic contribution was largely unknown before the groundbreaking genomic project was launched for the St. Jude Lifetime Cohort (SJLIFE) a few years ago.

We analyzed the whole-genome sequencing data for 4402 survivors (2090 females, median age at follow up = 30 years, and median length of follow up = 22 years) from SJLIFE study. Pathogenic/likely (P/LP) variants were identified for 156 cancer predisposition genes and 127 DNA repair genes, respectively, with a small number of genes overlapped between the two gene sets. We also constructed a polygenic risk score based on 172 SNPs established from de novo breast cancer genetic association studies using the general population. Based on the multivariate piecewise exponential models, rate of subsequent breast cancer was significantly increased for female survivors who carried a P/LP variant in cancer predisposition genes and with or without chest irradiation. In addition, rate of subsequent breast cancer was significantly increased among all female survivors, especially survivors who carried a P/LP variant in homologous recombination DNA repair genes in conjunction with treatment exposures of chest irradiation (≥20Gy) and/or anthracyclines (doses in the second and third tertiles). Finally, we found survivors with high polygenic risk score had increased risk of developing subsequent breast cancer over and beyond the monogenic risk conferred by rare P/LP variants in breast cancer predisposition genes. Interestingly, African Americans seem to have the lowest prevalence of subsequent breast cancer among three different race groups: 1.3% for African Americans, 4.1% for European Americans and 5.9% for Aisan/Native-Indian Americans. On the other hand, African Americans seem to have higher risk in developing subsequent breast cancer among carriers of P/LP variants compared to European Americans. We anticipate utilizing common susceptibility loci in combination with rare high-risk P/LP variants and other risk factors may inform a better strategy for breast cancer risk stratification among survivors of childhood cancer.

A novel alignment-free method for HIV-1 subtype classification

He Lily, Beijing University of Civil Engineering and Architecture

Abstract: HIV-1 is the most common and pathogenic strain of human immunodeficiency virus consisting of many subtypes. To study the difference among HIV-1 subtypes in infection, diagnosis and drug design, it is important to identify HIV-1 subtypes from clinical HIV-1 samples. In this work, we propose an effective numeric representation called Subsequence Natural Vector (SNV) to encode HIV-1 sequences. Using the representation, we introduce an improved linear discriminant analysis method to classify HIV-1 viruses correctly. SNV is based

on distribution of nucleotides in HIV-1 viral sequences. It not only computes the number of nucleotides, but also describes the position and variance of nucleotides in viruses. To validate our alignment-free method, 6902 complete genomes and 11,668 pol gene sequences of HIV-1 subtypes were collected from the up-to-date Los Alamos HIV database. SNV outperforms the three popular methods, Kameris, Comet and REGA, with almost 100% Sensitivity and Specificity, also with much less time. Our subtyping algorithm especially works better for circulating recombinant forms (CRFs) consisting of a few sequences. Our approach is also powerful to separate unique recombinant forms (URFs) from other subtypes with 100% Sensitivity and Specificity. Moreover, phylogenetic trees based on SNV representation are constructed using full-length HIV-1 genomes and pol genes respectively, where viruses from the same subtype are clustered together correctly.

A GPU-Accelerated Fast Summation Method for Electrostatics of Biomolecules

Robert Krasny, University of Michigan

Abstract: We present a fast summation method based on barycentric Lagrange interpolation and dual tree traversal (BLDTT). The method is kernel-independent and runs on multiple GPUs. We demonstrate the performance of the BLDTT for a variety of particle systems and discuss its implementation into our boundary element TABI Poisson-Boltzmann solver. This is joint work with Leighton Wilson, Nathan Vaughn, and Weihua Geng.

Advances in Poisson–Nernst–Planck Ion Channel Models and Finite Element Solvers

Dexuan Xie, University of Wisconsin-Milwaukee

Abstract: In this talk, I will report the recent progresses that we made in the development of Poisson–Nernst–Planck ion channel (PNPIC) models and related finite element solvers. In particular, two improved PNPIC models and their finite element solvers will be presented. One of them involves periodic boundary value conditions to mimic an infinitely large ion channel membrane while the other one involves a membrane surface charge density and Neumann boundary conditions to reflect the effects of membrane charges and the influence of ion channels outside a simulation box. I then will show how we overcome the numerical difficulties caused by solution singularity, exponential nonlinearity, multiple physical domains, periodic/Neumann boundary conditions, and membrane charges. With advanced mathematical and numerical techniques, we have developed finite element algorithms for solving these two models, and implemented them as a finite element program package that works for real three dimensional ion channel protein molecular structures and mixtures of multiple ionic species, along with numerical schemes for computing ion channel kinetics such as Gibbs free energy, electric currents, transport fluxes, and membrane potential. In this talk, I will present them, and report numerical results to demonstrate their performance and application.

Dissimilar Ligands Bind in a Similar Fashion: A Guide to Ligand Binding Mode Prediction

Xiaoqin Zou, University of Missouri - Columbia
Abstract: To uncover binding fashions of dissimilar ligands on a protein, we developed a novel intercomparison strategy to compare the binding modes of the ligands with different molecular structures. The results revealed that a surprising number of very dissimilar ligands can bind in a similar fashion, based on which we developed a new template-guided method for predicting protein-ligand complex structures. With the use of dissimilar ligands as templates, our method significantly outperformed traditional molecular docking methods.

Variational interface models for implicit solvation of biomolecules

Zhan Chen, Georgia Southern University
Abstract: Solute-solvent interactions are typically described by solvation energies (or closely related quantities). Implicit solvent methods are popular for the computation of solvation energies and many other applications in molecular simulation. An interface definition is required to indicate the separation of discrete solute atoms from the surrounding continuum solvent in implicit solvent models. In this talk, we will present a constrained variational interface model for implicit solvation of biomolecules. In our models, the optimal diffuse solute-solvent boundary, described by a characteristic function, is obtained by minimizing a total energy functional of the system to encompass energies of interest. The properties of the constrained solvation free energy functional and its minimizer have been carefully analyzed. In particular, the existence, uniqueness and boundedness of a global minimizer of the energy functional have been shown. Finally, with the rigorously derived PDE based diffuse interface model, we are able to compute the solvation free energies of nonpolar biomolecules accurately. This is a joint work with Prof. Yuanzhen Shao of University of Alabama, US.

Variational implicit-solvent predictions of the dry–wet transition pathways for ligand–receptor binding and unbinding kinetics

Shenggao Zhou, Soochow University
Abstract: Solvent fluctuations play a fundamental role in many water-mediated biological processes of importance. Dewetting and wetting transitions, induced by solvent fluctuations, take place in hydrophobic confinements. Based on a variational implicit solvent model, we combine the level-set method and the string method to study such dry-wet transitions, e.g., transition pathways and energy barriers. The resulting transition rates are then used in a spatially dependent multistate Brownian dynamics simulation and the related Fokker–Planck equation calculations of a ligand-receptor system. We find the hydration transitions to significantly slow down the binding process, in semiquantitative agreement with existing explicit-water simulations, but significantly accelerate the unbinding process. Moreover, our approach allows the characterization of nonequilibrium hydration states of pocket and ligand during the ligand movement, for which we find substantial memory and hysteresis effects for binding vs. unbinding. Our study thus provides a significant step forward toward efficient, physics-based interpretation and predictions of the complex kinetics in realistic ligand-receptor systems. This is a joint work with Dr. R. G. Weiß, L. Cheng, J. Dzubiella, J. A. McCammon, and B. Li.

Correlated Segment and Fuzzy Membrane Association of Intrinsically Disordered Proteins

Huan-Xiang Zhou, University of Illinois at Chicago

Abstract: Intrinsically disordered proteins (IDPs) account for a significant fraction of any proteome and are central to numerous cellular functions. Yet how sequences of IDPs code for their conformational ensembles, conformational dynamics, and ultimately, functions is poorly understood. I will report advances from our computational and experimental studies of two membrane proteins containing intrinsically disordered regions (IDRs). For ChiZ (a component of the cell division machinery in Mycobacterium tuberculosis), our NMR data revealed non-uniform backbone dynamics along the sequence of the 64-residue N-terminal IDR (NT). Our molecular dynamics (MD) simulations traced the origin to correlated segments, which are stabilized by polyproline II stretches, salt bridges, cation-pi interactions, and sidechain-backbone hydrogen bonds. Moreover, the extent of segmental correlation is sequence-dependent: segments in the first half of the NT sequence where internal interactions are more prevalent manifest elevated "collective" motions on the 5-10 ns timescale and suppressed local motions on the sub-ns timescale. Our NMR experiments found that NT associates with acidic membranes, but most residues remain dynamic, exception for a subset of Arg residues. MD simulations provided crucial details on the fuzzy membrane association, stabilized mostly by salt bridges between acidic lipids and Arg residues in the second half of the NT sequence. Lastly, we used MD simulations to investigate the mechanism of Ca2+-bound synaptotagmin-1 triggering membrane fusion, and produced a promising model that reconciles many conflicting experimental observations. Most importantly, a conserved acidic motif within an IDR competes with the vesicle membrane for interacting with the Ca2+-binding loops of the C2B domain, and flips C2B over for association with the plasma membrane, thereby bringing the two membranes closer for fusion. These findings serve as paradigms for sequence-conformation-dynamics-function relations of IDPs.

Multi-scale models for ESCRT driven membrane remodeling

Qiang Cui, Boston University

Abstract: TBA

Final size relation and control for epidemic model with heterogeneous mixing

Jing-an Cui, Beijing University of Civil Engineering and Architecture

Abstract: This talk focuses on the basic reproduction number and final size of infectious disease outbreaks. In view of heterogeneous multi-population epidemic model, the relationship between basic reproduction number and final scale is discussed and applied to the case study of infectious diseases.

A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended

Shaojun Pei, Tsinghua University, China

Abstract: Chaos Game Representation (CGR) was first proposed to be an image representation method of DNA and have been extended to the case of other biological macromolecules. Compared with the CGR images of DNA, where DNA sequences are converted into a series of points in the unit square, the existing CGR images of protein are not so elegant in geometry and the implications of the distribution of points in the CGR image are not so obvious. In this study, by naturally distributing the twenty amino acids on the vertices of a regular dodecahedron, we introduce a novel three-dimensional image representation of protein sequences with CGR method. We also associate each CGR image with a vector in high dimensional Euclidean space, called the extended natural vector (ENV), in order to analyze the information contained in the CGR images. Based on the results of protein classification and phylogenetic analysis, our method could serve as a precise method to discover biological relationships between proteins.

Fast stochastic compression algorithms for biological data analysis

Duan Chen, University of North Carolina at Charlotte

Abstract: Our recent work is motivated by two types of biological problems. One is inferring 3D structures of chromatins based on chromosome conformation capture (3C), such as Hi-C, which is a high-throughput sequencing technique that produces millions of contact data between genomic loci pairs. The other problem is computational deconvolution of gene expression data from heterogeneous brain samples, for extracting cell type-specific information for patients with Alzheimer's Disease (AD). Both problems involve large volumes of data, thus fast algorithms are indispensable in either direct optimization or machine learning methods. A central approach is the low-rank approximation of matrices. Conventional matrix decomposition methods such as SVD, QR, etc, are expensive, so not suitable for repeated implementation in these biological problems. Instead, we develop fast stochastic matrix compressions based on randomized numerical linear algebra (RNLA) theories. In this talk, we will emphasize on a recently developed stochastic kernel matrix compression algorithm. In this algorithm, samples are taken at no (or low) cost and the original kernel matrix is reconstructed efficiently with desired accuracy. Storage and compressing processes are only at O(N) or O(NlogN) complexity. These stochastic matrix compressing can be used to the above-mentioned biological problems to greatly improve algorithm efficiency, they can also be applied to other kernel based machine learning algorithms for scientific computing problems with non-local interactions (such as fractional differential equations), since no analytic formulation of the kernel function is required in our algorithms.

Immunotherapy Modeling: Molecular Interaction and Recognition of MHC/peptide/TCR Complexes

Ruhong Zhou, Zhejiang University and Columbia University

Abstract: Cancer immunotherapy has been among the most promising breakthroughs in oncology, particularly in the case of immune check point inhibitors, however, the effective

response rate remains quite low, only about 20-30%. In this talk, I will talk about our recent collaborative work which solves one mystery behind this low response rate. We found that patients with certain HLA genotype (HLA-B44) have consistently higher survival rate, while patients with some other type (HLA-B15) have much poorer survival rates. It's also shown that patients harboring tumors with very high mutation rates responded disproportionately well to these immune checkpoint inhibitor treatments. Large scale molecular dynamics simulations further reveal that HLA-B15 proteins with poorer therapeutic outcomes had structural appendages (HLA bridges with residues Arg62, Ile66, and Leu163) that closed over the cancer neoantigens with much less flexibility. The same techniques have also been applied to the design and development of HIV vaccines, which has been of great interest as well in recent years due to the wide spread infection of AIDS. With a combined in silico and in vivo approach, we studied the TCR/peptide/HLA interactions from multiple clonotypes specific for a well-defined HIV-1 epitope, and found that effective and ineffective clonotypes bind to the terminal portions of the peptide-HLA through similar salt bridges, but their hydrophobic side-chain packings can be very different, which accounts for the major part of the differences among these clonotypes. Together with state-of-the-art free energy perturbation calculations for point mutations on viral peptide, our results clearly indicate a direct structural basis for heterogeneous T cell antiviral function.

RNA secondary structure prediction by Evolutionary Profile and Mutational Coupling

Yaoqi Zhou, Griffith University

Abstract: The recent discovery of numerous non-coding RNAs (long non-coding RNAs, in particular) has transformed our perception of the roles of RNAs in living organisms. Our ability to understand them, however, is hampered by our inability to solve their secondary and tertiary structures in high resolution efficiently by existing experimental techniques. Computational prediction of RNA secondary structure, on the other hand, has received much-needed improvement, recently, through deep learning of a large approximate data, followed by transfer learning with gold-standard base-pairing structures from high-resolution 3-D structures. Here, we expand this single-sequence-based learning to the use of evolutionary profiles and mutational coupling. The new method allows large improvement not only in canonical base pairs (RNA secondary structures) but more so in base-pairing associated with tertiary interactions such as pseudoknots, noncanonical and lone base pairs. In particular, it is highly accurate for those RNAs of more than 1000 homologous sequences by achieving >0.8 F1 score (harmonic mean of sensitivity and precision) for 14/16 RNAs tested. The method can also significantly improve base-pairing prediction by incorporating artificial but functional homologous sequences generated from deep mutational scanning without any modification. The fully automatic method (freely available as server and standalone software) should provide the scientific community a new powerful tool to capture not only secondary structure but also tertiary base-pairing information for building three-dimensional models. It also highlights the future of accurately solving the base-pairing structure by using a large number of natural or artificial homologous sequences.

Collective dynamics of active particles on surfaces

Qi Wang, University of South Carolina
Abstract: We study collective motion of active particles on three prescribed surfaces with distinct topological and geometrical properties. Kinematics of the active particles on the surfaces is driven by selfpropelling, particle-particle interaction, surface constraining and under-damped stochastic forces described by Ornstein-Uhlenbeck processes. We demonstrate the prevailing collective patterns in the active particle systems on the three types of surfaces: a sphere, a torus and a hill and valley landscape with distinct topological and geometrical properties. We note that all the sustainable, spatial-temporal patterns are profoundly affected by the curvature of the surfaces as well as their symmetry. In particular, we find that the large magnitude of curvature in the hill and valley landscape coupled with certain surface symmetry warrants a spatial-temporal periodic traveling rings pattern which synchronizes the collective movement of the active particles with the symmetry in the landscape. However, the large magnitude of curvature alone without the necessary surface symmetry is not sufficient to sustain such a periodic, spatial-temporal pattern, instead collective motion settles into cyclic rotation.

Spatial signaling in single-cell data via optimal transport

Zixuan Cang, University of California Irvine
Abstract: One of the most interesting abilities of cells in a multicellular organism is their ability to acquire and change fate. This process is controlled by many factors including the environment and communications with other cells. A recent technological breakthrough (single cell RNA sequencing) allows us to see the expression of thousands of genes at single-cell resolution providing unprecedented information. This data makes it possible to identify heterogeneous population of cells in a tissue and to infer cell development trajectories. However, in single cell RNA sequencing experiments, tissues are dissected into single cells leading to loss of spatial information which is crucial to the analysis of communications among cells through space. With in situ gene expression data (usually containing only a few genes), we are able to retain some spatial information for single cells. We propose to use optimal transport theory to connect these two types of data. This connection allows us to reconstruct spatial expression of genes in single cell data and infer spatial distances among single cells. Communication among the single cells is inferred by using the retained spatial information and single cell RNA sequencing data based on a constrained optimal transport. These new insights can shed light on mechanisms of cell fate acquisition and changes as a collaborative process.

Topological data analysis, machine learning, drug design

Kelin XIA, Nanyang Technological University
Abstract: Effective molecular representation is key to the success of machine learning models for molecular data analysis. In this talk, we will discuss a series of persistent representations, including persistent homology, persistent spectral models, and persistent Ricci curvature and their combination with machine learning models. Unlike traditional graph/network or geometric models, these filtration-induced persistent models can characterize the multiscale intrinsic information, thus significantly reduces molecular data complexity and dimensionality. Feature vectors are obtained from various persistent attributes and inputted into machine learning

models, in particular, random forest, gradient boosting tree and CNN. Our persistent representations based molecular fingerprints can significantly boost the performance of learning models in drug design.

Abstracts of the Special Issue

A special issue entitled "Computational and Mathematical Bioinformatics and Biophysics" is dedicated to this conference, and will be published in Communications in Information & Systems. Six manuscripts have been accepted/submitted to this special issue. For your information, the abstracts of these manuscripts are given below.

Inverted repeats in coronavirus SARS-CoV-2 genome and implications in evolution

Changchuan Yin, Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Stephen S.-T. Yau, Department of Mathematical Sciences, Tsinghua University, Beijing 100084, China

Abstract: The coronavirus disease (COVID-19) pandemic, caused by the coronavirus SARS CoV-2, has caused 60 millions of infections and 1.38 millions of fatalities. Genomic analysis of SARS-CoV-2 can provide insights on drug design and vaccine development for controlling the pandemic. Inverted repeats in a genome greatly impact the stability of the genome structure and regulate gene expression. Inverted repeats involve cellular evolution and genetic diversity, genome arrangements, and diseases. Here, we investigate the inverted repeats in the coronavirus SARS-CoV-2 genome. We found that SARS-CoV-2 genome has an abundance of inverted repeats. The inverted repeats are mainly located in the gene of the Spike protein. This result suggests the Spike protein gene undergoes recombination events, therefore, is essential for fast evolution. Comparison of the inverted repeat signatures in human and bat coronaviruses suggest that SARS-CoV-2 is mostly related SARS-related coronavirus, SARSr-CoV/RaTG13. The study also reveals that the recent SARS related coronavirus, SARSr-CoV/RmYN02, has a high amount of inverted repeats in the spike protein gene. Besides, this study demonstrates that the inverted repeat distribution in a genome can be considered as the genomic signature. This study highlights the significance of inverted repeats in the evolution of SARS-CoV-2 and presents the inverted repeats as the genomic signature in genome analysis.

SARS-CoV-2 becoming more infectious as revealed by algebraic topology and deep learning

Jiahui Chen [1], Rui Wang [1], and Guo-Wei Wei [1;2;3]

[1] Department of Mathematics, Michigan State University, MI 48824, USA.
[2] Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA.
[3] Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA.

Abstract: Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) caused by coronavirus disease 2019 (COVID-19) has led to tremendous human fatality and economic loss. SARS-CoV-2 infectivity is a key reason for the widespread viral transmission but its rigorous experimental measurement is essentially impossible due to the on-going genome evolution around the world. We show that artificial intelligent (AI) and algebraic topology (AT) offer an accurate and efficient alternative to the experimental determination of viral infectivity. AI and AT analysis indicates that the on-going mutations make SARS-CoV-2 more infectious.

A Bayes-inspired theory for optimally building an efficient coarse-grained folding force field

Travis Hurst [1], Dong Zhang [1], Yuanzhe Zhou [1], and Shi-Jie Chen [1,2,3]

[1] Department of Physics, University of Missouri Columbia, Columbia, MO 65211, USA.
[2] Department of Biochemistry, University of Missouri Columbia, Columbia, MO 65211, USA
[3] MU Institute for Data Science and Informatics, University of Missouri Columbia, Columbia, MO 65211, USA

Abstract: Because of their potential utility in predicting conformational changes and assessing folding dynamics, coarse-grained (CG) RNA folding models are appealing for rapid characterization of RNA molecules. Previously, we reported the iterative simulated RNA reference state (IsRNA) method for parameterizing a CG force field for RNA folding, which consecutively updates the simulation force field to reflect marginal distributions of folding coordinates in the structure database and extract various energy terms. While the IsRNA model was validated by showing close agreement between the IsRNA simulated and experimentally observed distributions, here, we expand our theoretical understanding of the model and, in doing so, improve the parameterization process to optimize the subset of included folding coordinates, which leads to accelerated simulations. Using statistical mechanical theory, we analyze the underlying, Bayesian concept that drives parameterization of the energy function, providing a general method for developing predictive, knowledge-based, polymer force fields on the basis of limited data. Furthermore, we propose an optimal parameterization procedure, based on the principal of maximum entropy.

Adaptive pseudo-time methods for the Poisson-Boltzmann equation with Eulerian solvent excluded surface

Benjamin Jones [a], Sheik Ahmed Ullah [b], Siwen Wang [a], Shan Zhao [a]

[a] Department of Mathematics, University of Alabama, Tuscaloosa, AL 35487, USA
[b] Department of Mathematics, Stillman College, Tuscaloosa, AL 35401, USA

Abstract: This work further improves the pseudo-transient approach for the Poisson Boltzmann equation (PBE) in the electrostatic analysis of solvated biomolecules. The numerical solution of the nonlinear PBE is known to involve many difficulties, such as exponential nonlinear term, strong singularity by the source terms, and complex dielectric interface. Recently, a pseudo-time ghost-fluid method (GFM) has been developed in [S. Ahmed Ullah and S. Zhao, Applied Mathematics and

Computation, 380, 125267, (2020)], by analytically handling both nonlinearity and singular sources. The GFM interface treatment not only captures the discontinuity in the regularized potential and its flux across the molecular surface, but also guarantees the stability and efficiency of the time integration. However, the molecular surface definition based on the MSMS package is known to induce instability in some cases, and a nontrivial Lagrangian-to-Eulerian conversion is indispensable for the GFM finite difference discretization. In this paper, an Eulerian Solvent Excluded Surface (ESES) is implemented to replace the MSMS for defining the dielectric interface. The electrostatic analysis shows that the ESES free energy is more accurate than that of the MSMS, while being free of instability issues. Moreover, this work explores, for the first time in the PBE literature, adaptive time integration techniques for the pseudo-transient simulations. A major finding is that the time increment should become smaller as the time increases, in order to maintain the temporal accuracy. This is opposite to the common practice for the steady state convergence, and is believed to be due to the PBE nonlinearity and its time splitting treatment. Effective adaptive schemes have been constructed so that the pseudo-time GFM methods become more efficient than the constant dt ones.

New Variational Analysis on the Sharp Interface of Multiscale Implicit Solvation: General Expressions and Applications

Elizabeth Hawkins [1], Yuanzhen Shao [2], Zhan Chen [1]

[1] Department of Mathematical Sciences, Georgia Southern University, Statesboro, GA, USA
[2] Department of Mathematics, University of Alabama, Tuscaloosa, AL, USA

Abstract: The interface definition between regions of different scales becomes a key component of a multiscale model in mathematical biology and other fields. Differential geometry based surface models have been proposed to apply the theory of differential geometry as a natural means to couple polar-nonpolar and solute-solvent interactions. As a consequence, the variational analysis of such models heavily relies on the variation of the interface. In this work, we provide a new variational approach to conduct the variational analysis on the sharp interface of multiscale implicit solvation models. It largely simplifies the computations of variations of the area and volume functionals. Moreover, general expressions of the second variation formulas of the solvation energy functional are obtained and used for the stability analysis of the equilibrium interface. Finally, we establish a reasonable concept of stability which generalizes the well-known results in minimal surfaces with constant volume and then the necessary and sufficient condition for stability. Our work paves the way to conducting stability analysis for a general energy functional especially with constant volume.

Fast random algorithms for manifold based optimization in reconstructing 3D chromosomal structures

Duan Chen [a], Shaoyu Li [a], Xue Wang [b], Kelin Xia [c;d]

[a] Department of Mathematics and Statistics, University of North Carolina at Charlotte, Charlotte, NC 28223, USA
[b] Department of Health Science Research, Mayo Clinic, Jacksonville, FL, USA

[c] School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore
[d] School of Biological Sciences, Nanyang Technological University, Singapore

Abstract: Inferring 3D structures of chromatins and chromosomes from experimental data is critical to understand biological functions of DNAs, and various computational algorithms have been developed in the past a few years. All algorithms are subject to the challenge of high computational cost if the number of loci in the target chromosome is large. In this paper, we tackle this difficulty and develop a set of fast algorithms for the manifold-based optimization (MBO) model, which is a popular method to reconstruct 3D chromosomal structures from Hi-C data. The proposed algorithms are based on random projection theory. We first approximate the column (row) space of the original data in reduced dimension. Then interpolative decomposition technique is used to decompose the data matrix into a product of two matrices, each of which has a much smaller dimension comparing to the number of degree of freedom of the problem. With this low-rank approximation, all components in the gradient descent method of the optimization, including calculating gradient, line search, and solution updating, have the linear complexity, with respect to the total number of loci in the target chromosome. At last, a randomly perturbed gradient descent method is adopted so one can effectively escape saddle points of the non- convex optimization. In simulations, a synthetic simple helix and a simulated chromosomal structure are used to validate our algorithms, suggesting its highly enhanced efficiency and desired ability to recover structures from data subject to random lost and mild contamination of noises.