An outline of the ten lectures

Lecture 1: Introduction to Mathematical Molecular Bioscience and Biophysics
Biophysics, bioinformatics, and systems biology concern quantitative modeling, analysis, simulation, computation, and prediction of modern biological sciences. Among, them, biophysics emphasizes the application of physical laws to understand the molecular mechanisms of biological systems across all scales, from molecular to organismic. In contrast, bioinformatics focuses on statistical, mathematical and computer science based methodological developments for understanding biological data. Notably, systems biology concerns the computational and mathematical modeling and analysis of complex biological systems, particularly, their time evolution. It is fair to say that biophysics is mechanistic, bioinformatics is data-driven and systems biology is knowledge-based. While overlapping with biophysics, bioinformatics, and systems biology, mathematical molecular bioscience and biophysics (MMBB) is unique in emphasizing the application and development of mathematical approaches for solving challenging problems in molecular bioscience and biophysics, such as the understanding of biomolecular structure-function relationship, the determination of ion channel gating mechanism and the discovery of new drugs for curing human diseases. MMBB is also very keen on creating biology inspired mathematics as mathematical physics did to mathematics in the past century. This lecture starts with a historical review of biological science to elucidate its status, challenge, trend, and future. Discussions are given to the connection and distinction of various quantitative biology disciplines, including biophysics, bioinformatics, systems biology, and MMBB after an introduction to essential experimental methods and theoretical models.

Lecture 2: Differential Geometry Based Biomolecular Surface Modeling
Differential geometry concerns with the geometric structures, such as curves and surfaces, on differentiable manifolds. It utilizes techniques from calculus, variation, linear algebra and multilinear algebra and draws upon results from differential topology differential equation to study problems in geometry. In molecular biology, geometric modeling provides the structural representation of molecular structures and bridges the gap between molecular structural data and theoretical/mathematical models. One of the simplest molecular geometric models is the space-filling Corey-Pauling-Koltun (CPK) theory, which represents an atom by a solid sphere with a van der Waals (VDW) radius. With the 3D space-filling representation of a protein, various geometrical definitions, including VDW surface, solvent accessible surface, and solvent-excluded surface, have been introduced to distinguish a protein from its surrounding environment and to understand macromolecular interactions. The differential geometry of surfaces is a natural tool to describe biomolecular shapes and interactions. It avoids geometrical singularities such as cusps and tips, in many other commonly used surface definitions and thus provides a sound basis for advanced physical modeling, including multiscale models for electrostatics, charge and material transport. Multiscale curvature maps are utilized to identify the hot spots of protein-ligand and protein-protein interactions. Utilizing the Euler-Lagrange variation, a differential geometry based surface model, the minimal molecular surface, is introduced for biomolecular geometric modeling. Surface evaluation is employed to detect protein binding pockets and to analyze molecular topological changes.

Lecture 3: Differential Geometry Based Models for Electrostatics and Solvation

Under physiological condition, 65-90 percent of cellular mass is water and thus biomolecular solvation is the basic process that underpins the molecular mechanism for almost all other important biological processes, including protein folding, protein mutation, molecular recognition, protein-protein interaction, protein-ligand binding, transcription, post-translational modification, translation, enzyme catalysis, phosphorylation, and signal transduction. The biomolecules solvation involves bond construction/reconstruction, electrostatics and van der Waals forces due to the possible interactions with water molecules, aqueous ions, counterions, and other molecules. Explicit solvation models are computationally prohibited for large systems with solute molecules consisting of hundreds or millions of atoms, and at the same time, surrounded by millions of solvent molecules, which in turn rapidly change their positions and orientations. Implicit solvent models utilize a multiscale approach to reduce the complexity and the large number of degrees of freedom by a discrete atomistic description of the solute molecule while a continuum dielectric representation of the solvent. Some of the well-known issues in implicit solvent modeling include the ad hoc division between solvent and solute and the neglect of nonlinear effects due to solvent-solute mutual polarization. Geometric measure theory is introduced to represent the solute shape by the gradient of a hypersurface function. The variation principle is used to minimize the entropically and enthalpically unfavorable work of solute cavity formation and van der Waals interactions in a nonpolar solvent model. Since electrostatic effect is fundamental in nature and ubiquitous in all biomolecules, a polar component is indispensable. In a differential geometry based full solvation model, both polar component modeled by the Poisson-Boltzmann theory and nonpolar component are coupled with the solvent-solute interface via the total free energy variation. The density functional theory (DFT) is introduced for solute electrons to further account for the electronic rearrangement associated with the molecular surface reconstruction. The total free energy decay during variational simulations is proved rigorously.

Lecture 4: Variational Approaches to Membrane Transport

Lipid bilayer membrane is one of the most important biomolecular systems that establish the heterogeneous environments between intercellular and intracellular spaces and provide a platform for various essential transmembrane processes. Membrane transporters, including ion channels, are membrane transport proteins that regulate most cross-membrane material exchanges or information fluxes so as to sustain the regular functions of cells and subcellular organelles. Ion channels are pore-forming proteins that regulate signal transduction and action potential by gating the flow of ions across the cell membrane, controlling the flow of ions across secretory and epithelial cells, and modulating cell volume. Ion channels are prominent components of the nervous system for their role in transmitter-activated nerve impulse across neural synapses. They are responsible for numerous nervous and neuromuscular diseases and thus are some of the most important therapeutic targets. There are numerous ion channel models, ranging from simple-minded Hodgkin–Huxley model, Poisson-Nernst-Planck (PNP) theory, and molecular dynamics, to expensive quantum mechanical approaches. Variational multiscale models are formulated to reduce the number of degrees of freedom and simplify the model complexity of full-atom models while enhancing the description and predictive power of the classic PNP model. For proton transport, a variational density functional model is proposed to deal with the quantum effect in proton channels. It is shown that the proposed models reduce to appropriate simple models at various limits.

Lecture 5: Numerical Methods for Biomolecular Simulations

Mathematical modeling and simulation of biomolecular systems encounter a wide variety of computational challenges, such as the integration of molecular dynamics, the solution of coupled partial differential equations (PDEs) from multiscale models, the optimization of the loss function for biological predictions, the tracking of molecular free boundaries, the evaluation of biomolecular topological persistence barcodes and the diagonalization of the biomolecular graph/Hodge Laplacian matrices. Therefore, the development of efficient numerical methods and computational algorithms is an important task of mathematical molecular bioscience and biophysics. In fact, the numerical analysis of biomolecular systems furnishes rich mathematical development. For example, the electrostatic analysis may involve the Ewald method for system with periodic boundary conditions, multipole methods such as the treecode and fast multipole methods (FMM) for accelerating pairwise long-range integrations, interface methods such as immersed interface methods (IIM) and matched interface and boundary (MIB) for enforcing dielectric interface conditions on non-smooth solvent-solute interfaces, and boundary integral methods using induced surface charges to speed up electrostatic analysis. This lecture focuses on high-order MIB methods for solving elliptic interface problems arising from the Poisson-Boltzmann (PB) model. The problem is difficult due to its discontinuous interface, singular charge source, and nonlinearity. Methods and challenges for PB based molecular dynamics are also discussed.

Lecture 6: Graph Theory Based Modeling and Analysis

Arguably, graph theory is the most important subject in discrete mathematics. It concerns with the use of graphs as mathematical structures for modeling pairwise relations, i.e., edges, between vertices, nodes, or points. There are many different graph theories, such as geometric graphs, algebraic graphs, and topological graphs, that are versatile mathematical tools for analyzing biomolecular structure, function, dynamics and transport. For example, algebraic graph theory, particularly spectral graph theory, studies the algebraic connectivity via characteristic polynomials, eigenvalues, and eigenvectors of matrices associated with graphs, such as adjacency matrix or Laplacian matrix. This approach has been used in normal mode analysis and elastic network model, including Gaussian network model and anisotropic network model. Geometric graphs admit geometric objects as graph nodes or vertices and can significantly reduce the computational complexity of algebraic graph approaches for excessively large biomolecules. This lecture discusses the development of multiscale weighted colored graphs for biomolecular flexibility analysis, protein B factor prediction, protein domain classification, and protein hinge detection. It also deals with the functional prediction from massive biomolecular data arising from protein-ligand binding, protein-protein interaction, and protein folding stability changes upon mutation.

Lecture 7: Topology Based Modeling and Analysis

In mathematics, topology studies the topological invariants of continuous space or discrete space under continuous deformations. In chemistry, topological analysis based on the theory of atoms in molecules, electron localization function, and quantum chemical topology, provides powerful tools for characterizing chemical bonds and atoms, and for analyzing electron pair localization. In molecular biology, topology provides the ultimate abstraction of geometric complexity by concerning only the connectivity of different components in the space arising from biomolecular data and characterizing independent entities, rings, and higher-dimensional faces of the space in terms of Betti numbers. Topological analysis and modeling of biomolecular structures give rise

to intrinsic topological invariants of macromolecules, such as independent components (atoms), rings (pockets), and cavities. However, traditional topology oversimplifies biomolecular complexity and leaves too little information for analyzing biomolecular data. Persistent homology bridges between geometry and topology and offers a more powerful tool for data analysis. It turns out that persistent homology neglects chemical and biological information in diverse biomolecular data. Element-specific persistent homology, atom specific persistent homology, multi-level persistent homology, and electrostatic persistence are introduced to embed chemical and biological information into topological invariants during the topological abstraction of massive and diverse biomolecular data. Multiresolution induced multidimensional persistence is proposed to extract appropriate topological properties over a wide range of spatial scales.

Lecture 8: de Rham-Hodge Theory Based Modeling and Analysis

The de-Rham-Hodge theory is an important landmark of the 20th Century mathematics that coherently connects differential geometry, algebraic topology and partial differential equation (PDE). The de-Rham-Hodge theory provides a rigorous foundation for Maxwell's theory, quantum mechanics, quantum field theory and Yang-Mills theory, among many others. This lecture describes the application of the de Rham-Hodge theory for biomolecular analysis and modeling. Based concepts, including exterior derivatives, differential forms, closed form, exact form, (co)cycle, (co)boundary, de Rham (co)chain complexes, and Hodge duality are reviewed. Discussions are given to de Rham cohomology, Hodge star operator, and Hodge decomposition theorem. These concepts and techniques are applied to biomolecular manifolds arising from biomolecular structural data. Discrete exterior calculus tools in three dimensional (3D) space are developed for Hodge decomposition on an arbitrarily complex geometry with appropriate boundary conditions. Applications are considered for cryo-electron microscopy (cryo-EM) maps and protein structures. It is shown that the eigenvectors from Hodge Laplacian operators shed light on anisotropic motion of proteins and cryo-EM maps. Additionally, when de Rham-Hodge theory is applied to protein flexibility analysis and it offers some of the best predictions of B factors. In associated with machine learning, element-specific de Rham-Hodge theory and de Rham-Hodge persistence provide a new paradigm for the prediction of biomolecular data.

Lecture 9: Mathematics for Biomolecular Data

Thanks to the rapid advance in gene sequencing, X-ray crystallography, nuclear magnetic resonance (NMR) and cryo-EM technologies, the GenBank has collected over two hundred million sequences, while the Protein Data Bank (PDB) has accumulated more than 150,000 biological macromolecular structures. Meanwhile, the availability of high-performance graphics processing unit (GPU) and stochastic gradient descent (SGD) algorithm makes various deep learning algorithms feasible for large datasets. The developments in biological data and machine learning algorithms have given rise to an unprecedented opportunity for bioinformatics and biomolecular data analysis. However, biomolecular datasets contain complex macromolecular structures for which brute force 3D image representations involve excessively large machine learning dimensions and are not scalable from different structures. The need to analyze massive diverse and complex biomolecular datasets calls for innovative mathematical methods for macromolecular structural complexity reduction. This lecture provides a brief summary of the problems in biomolecular data analysis and major machine learning algorithms. The design and construction of various mathematical techniques discussed in the early lectures, namely, differential geometry, algebraic topology, graph theory, and de Rham-Hodge theory, are

highlighted to provide new powerful tools for simplifying macromolecular structural complexity and for generating scalable feature vectors for machine learning predictions. The mathematical representations of biomolecular datasets are coupled with many machine learning and deep learning architectures to predict protein-protein and protein-ligand binding affinities, protein folding stability change upon mutation, drug toxicity, solvation, solubility, permeability, and partition coefficient.

Lecture 10: Mathematics for drug discovery

One of the ultimate goals of modern biological science is to reveal the secret of life and cure human diseases. Drug discovery is an expensive and time-consuming process which takes over 10 years and about US$2.6 billion to bring an average drug to market. The discovery process involves disease identification, target hypothesis, lead discovery, lead optimization, preclinical development, clinical trials, and drug efficacy optimization. Although much progress has been made in computer-aided drug design and discovery, the process is still labor intensive and essentially depends on trial and error. The recent success of Google's AlphaFold in winning the Critical Assessment of Structure Prediction competition ushered in a new era of scientific discovery. It holds great promise to discover new drugs significantly faster and cheaper, which could be particularly beneficial to patients with rare medical ailments, for whom drug discovery is currently not profitable or for those whose medical ailments currently cannot be effectively treated with drugs, such as Alzheimer's disease. However, drug discovery is much more complex and challenging than predicting protein folds. The structural complexity of protein-drug complexes and their intricate interactions is one of the major obstacles in AI-based high-throughput screening, hit to lead, and lead optimization. This lecture reviews the concepts and challenges for drug discovery. Emphasis is given to the development of mathematical methods, including differential geometry, persistent homology, algebraic and geometric graphs, and multiscale partial differential equations, for reducing biomolecular structural complexity and for extracting protein-drug interactions. These mathematical approaches have been integrated with advanced machine learning algorithms, such generative adversarial network (GAN), to win many contests in D3R Grand Challenges, a worldwide competition series in computer-aided drug design. Mathematical methods are devised for quantitative systems pharmacology (QSP) modeling of drug pharmacokinetics, pharmacodynamics, and efficacy.