Research Interests and Accomplishments

One of my research interests is developing new multiscale techniques for simulations of biological macromolecules in solution. Because the time scales for atomic motions and large scale motions of biomolecular complexes may differ by more than a factor of 10^{10}, it is important to develop methods that bridge these scales in order to provide detailed mechanistic insight into the molecule's function.

I have developed a fast multiscale algorithm that combines a novel fast multipole method with a boundary element formulation of the linear Poisson-Boltzmann equation to calculate the electrostatic forces on macromolecules in an ionic solution. This method introduces several innovations: a fast multipole method for continuous charge distributions that may be used with a boundary element representation of the potential, direct evaluation of the force and torque on a molecule through its boundary element representation, and a fast one-step method to calculate the polarization charge on one molecule due to the electrostatic field of another molecule. The method was tested by calculating the electrostatic forces between two actin monomers. I would like to incorporate this method in a molecular dynamics program and apply it towards studying actin nucleation and polymerization dynamics and protein-protein docking.

I am also interested in developing hybrid explicit/implicit solvation models for simulations. Much of the computational effort in molecular dynamics simulations with explicit water molecules is expended in calculating the dynamics of the distant water molecules even though their detailed dynamics are unimportant for the protein's motion. At the other extreme are methods that model the solvent as a continuum, however these methods neglect important short range water-protein interactions. A solution to this problem is to use explicit water molecules in the region near the protein's molecular surface and to use a continuum model for the remaining distant region. The challenge is to match these two complementary descriptions of the solvent in a consistent manner.

One major limitation of molecular dynamics simulations of large biomolecules is that the largest accessible times scale is too short to study functionally important large scale motions. This is due to the large range of characteristic times for dynamics from the atomic scale to the domain or protein scale, as mentioned above. One solution that I would like to pursue is a multiscale method using stochastic dynamics for the local "fast" degrees

of freedom combined with deterministic dynamics for the remaining "slow" degrees of freedom. This would enable the use of larger time steps for the "slow" degrees of freedom. The diffusional parameters could either be periodically updated using short atomic level simulations or predetermined as conformation dependent parameters. The resulting method would be tested using experimental results on large scale motions of proteins due to allosteric regulation or on the mechanical and dynamical properties of structural proteins such as actin and tubulin.

My most recent research project has been the prediction of protein-protein interaction interfaces. The sequencing of entire genomes has facilitated a global system-wide approach to biological processes. One such global view is the interactome or the complete interaction network of biological molecules, such as proteins. However experiments to discover interacting proteins, such as yeast two hybrid experiments, have large false positive and negative rates, and do not provide detailed information on the structure of the complex that could be used to infer its function. X-ray crystal structures or NMR structure determination are the only means to obtain this information. However, these methods are time consuming, particularly for novel protein complexes. Therefore a method using the available structures of the two component proteins to predict their complex is very useful for making functional predictions and suggesting further experiments, such as mutagenesis experiments, to confirm their prediction.

As a first step in this direction, I have developed a method to predict protein-protein interfaces given the structure of one of its component proteins and a multiple sequence alignment. A support vector machine trained on multiple sequence alignments for proteins in a carefully prepared database of biologically relevant complexes was used for this task. Cross-validation demonstrates that the method performs well with the predicted interface overlapping the actual interface for 97% of the complexes. Furthermore, the method was able to identify interfaces not present in the crystal structure for intracellular signalling proteins that have multiple binding partners. One immediate application of this prediction is to use it to guide all-atom protein-protein docking simulations. I would also like to use this data to develop a hybrid empirical/physical scoring function for use in protein-protein docking.

The evolutionary conservation of residues in a protein is an important indicator of their functional roles in tertiary structure and folding, enzymatic activity, ligand binding, and interaction with other proteins. I developed a robust Bayesian method for calculating evolutionary site rates in proteins, in

collaboration with Ruben Abagyan. Bayesian estimation avoids the problem of overfitting and the inclusion of alignment reliability makes the results less sensitive to alignment errors. I would like to continue exploring improved methods for calculating evolutionary conservation as well as applications to the prediction of biologically relevant protein-protein and protein-small molecule binding interactions. The large scale prediction of protein-protein interactions from correlated mutations and the prediction of the structures of protein complexes are of particular interest.

Another research project that I recently completed is the prediction of the protein structure and stability changes for single point mutations. A short Monte Carlo simulation of the side-chains of the mutated residue as well as nearby residues was first performed in order to predict the conformation of the mutated protein. A large database of protein structures that differ by a single point mutation was used to validate this procedure. Next, an empirical potential was optimized using experimental stability data and used to predict stability changes. The method was shown to be statistically stable and accurate by cross validation. I am interested in applying this method to predict the stability effects of single nucleotide polymorphisms on human proteins implicated in disease. Initially the correlation of the stability with deleterious mutations would be studied and later novel computational predictions would be compared with experiments conducted by a collaborator to llmeasure the stability of the mutant protein.

I have also been working on using molecular mechanics simulations to perform molecular mechanics docking and binding energy calculations for peptide-MHC binding. A biased-probability Monte Carlo method is used to sample the torsion angles of the peptide using an all-atom force field for the peptide and a grid potential for the MHC. Grid potentials for multiple MHC conformations are used in order to account for flexibility. An empirical potential energy function is then used to the predicting binding affinity. I have also studied docking into homology models of MHC allotypes for which an X-ray structure is unavailable, and have gotten accurate results, particularly for allotypes in which only a few residues in the binding interface differ. This method is designed to be fast enough to screen a large number of candidate peptides and is expected to be useful for predicting epitopes in query proteins, particularly those without standard anchor residues or those which bind allotypes without available X-ray structures. I am interested in further improving the use of homology models for MHC proteins in peptide docking as well as the prediction of the T-cell-peptide-MHC ternary complex

geometry using protein-protein docking calculations.

Recent genome sequencing and expression analysis research has yielded reasonably reliable information on the location of genes and their expression levels in different environments and cellular states. However, much less is known about the regulation of gene expression through transcription factors. The location regulatory sequences in the genome and the function of the corresponding transcription factors is necessary for understanding the control of gene expression levels and consequently for a system-wide understanding of biological processes. The prediction of regulatory sequences is an open problem since current methods predict a large number of false positives. I am interested in developing more accurate methods to predict regulatory sequences in genomes by integrating information from local sequence motifs, larger scale sequence patterns such as CpG islands, phylogenetic footprinting, and microarray data. Combining data from a number of different signals should substantially increase the prediction selectivity. It is also important to investigate prediction methods for specific subclasses of promoters since they may have specific features that make their prediction easier.

Teaching Interests and Experience

My teaching interests are in the general areas of molecular simulations, phylogenetic and sequence analysis, and machine learning methods in bioinformatics. I am interested in teaching at both the undergraduate and graduate levels.

I would be particularly interested in teaching a class on the theory and methods used in molecular simulations. The course material would include the relevant background in statistical mechanics, Monte Carlo methods, molecular dynamics methods, free energy calculations, Langevin dynamics, and more advance techniques such as biased sampling. I could also teach a class on the general use of computers in the physical sciences in which the students would work on individual projects.

I would also be interested in teaching classes on sequence and phylogenetic analysis. An introductory class would be oriented towards students in biology and medical sciences and would teach the use of sequence and structural databases available on the Internet, the use of sequence alignment tools such as ClustalW and BLAST, and the basics of machine learning methods for sequence analysis and classification. I could also teach a more advanced class that covers the dynamic programming algorithms for sequence alignment, multiple alignment algorithms, heuristic algorithms for sequence database searches, clustering algorithms, and algorithms used for phylogenetic analysis.

Finally, I could teach a class on machine learning methods in bioinformatics. The class would study the theory behind Hidden Markov Models, Support Vector Machines, and Artificial Neural Networks as well as the applications of these methods to biological problems such as sequence classification, gene prediction, structure prediction, and large scale expression analysis from microarray experiments.

My previous teaching includes a total of three years of teaching experience as a graduate student teaching assistant. This includes teaching three different classes: introductory physics for non-science majors (non-calculus physics), introductory physics for scientists and engineers, and electronics laboratory. I taught both the laboratory and discussion portions of the introductory physics classes.