Yu Xia, Ph.D.
Dept. of Molecular Biophysics & Biochemistry
Yale University
New Haven, CT 06511
Tel: (510) 292-6981
Email: yuxia@csb.yale.edu

October 17, 2005

Yves Brun
Systems Biology/Microbiology Faculty Search
Department of Biology, Indiana University
Jordan Hall 142, 1001 E 3rd Street
Bloomington, IN 47405-7005

Dear Dr. Brun,

I am writing to apply for the tenure-track position of Assistant Professor of Computational Systems Biology in the Department of Biology at Indiana University, as recently advertised in Science. I am currently a post-doctoral fellow in the Department of Molecular Biophysics and Biochemistry at Yale University, in the research group of Mark Gerstein. I received my Ph.D. from Stanford University in 2002.

As reflected in my curriculum vitae, I have extensive research experience and publication record in computational biology and bioinformatics. I am interested in applying computational techniques to study the structure, dynamics, and evolution of complex bio-molecular systems, such as proteins and protein networks, by combining the data-driven approach of integrated probabilistic modeling of diverse genome-wide information, together with the principle-driven approach of physical modeling and mechanistic simulation of complex systems. I believe that my experience and qualifications make me a strong candidate for the opening position.

In the application package, you will find the following items: my current curriculum vitae with contact information of four references, a statement of research interests, a statement of teaching interests, and selected reprints. I have arranged for the recommendation letters directly sent to you. Preprints and reprints for all my publications can be downloaded from the following URL:
http://www.csb.yale.edu/people/gerstein/yuxia/papers.html
I would be happy to forward you additional documents if they are desired. Thank you for your consideration and I look forward to hearing from you.

Sincerely,

Yu Xia, Ph.D.

Statement of Research Interests

Yu (Brandon) XIA, Ph.D.
Yale University, New Haven, CT

My research is focused on computational biology and bioinformatics. I apply computational techniques to study the structure, dynamics, and evolution of complex bio-molecular systems, such as proteins and protein networks, by combining the data-driven approach of integrated probabilistic modeling of diverse genome-wide information, together with the principle-driven approach of physical modeling and mechanistic simulation of complex systems.

Research Philosophy

With the completion of various genome sequencing projects, and the technological developments in functional, structural and chemical genomics, we are entering a new phase of genomic research where the bulk of the efforts will be shifted from data accumulation to data analysis. Currently, different data sets are often analyzed by experts from separate fields, such as biological sequence analysis, microarray analysis, computational structural biology, and computational chemistry, each revealing a different aspect of genome biology. The challenge now is to bring together these disjoint computational efforts to integrate large amounts of data from diverse sources, and to develop a unified and coherent probabilistic framework for data mining and knowledge discovery.

Another type of integration is to compare biological data across different time points, different cell types, different developmental stages, different individuals, and different species. Similarities and differences learned from such comparisons can give us insights into the dynamics and evolution of biological systems.

Integrated data mining provides us with a data-driven, statistical description of complex bio-molecular systems. In addition, these complex systems obey the laws of physics, and they have evolved to carry out specific biological functions, often in a reliable, efficient, and adaptable manner. By incorporating these statistical, physical, functional, and evolutionary constraints, it is possible to build increasingly quantitative biophysical models. Computer simulations of these models can begin to address how interactions among individual components lead to collective behavior of the system, thus providing us with additional insights into the structure, dynamics and evolution of the system. Predictions based on the model can be tested experimentally, which can in turn be used to refine the model.

The combined efforts of bioinformatics, biophysical modeling and simulation, together with experimental collaborations, will ultimately lead to a predictive, rather than descriptive theory of bio-molecular systems.

Current Research

My Ph.D. research focused on two aspects of computational structural biology, primarily using the principle-driven approach of biophysical modeling and simulation: predicting three-dimensional structure of a protein from primary sequence information, and physical and evolutionary simulation of protein folding. My postdoctoral research has been focusing on bioinformatics, primarily using the data-driven approach of statistical modeling: prediction and

analysis of bio-molecular networks; application of statistical machine learning techniques in genomics and proteomics.

Protein Tertiary Structure Prediction

I have developed a hierarchical method for *ab initio* protein tertiary structure prediction. This method was one of the best at the CASP3 meeting (Critical Assessment of Techniques for Protein Structure Prediction). The resulting protein structural models are sometimes accurate enough for inferring biological function. In addition, I developed a rigorous approach for deriving and assessing knowledge-based energy functions, which led to the design of improved energy functions. My other structure prediction projects included high resolution comparative modeling and structure-based prediction of dissociation kinetics for MHC-peptide complex, a system central to molecular immunology.

Physical and Evolutionary Simulation of Protein Folding

I have performed physical and evolutionary simulation of protein folding using a simplified model for proteins. This allowed me to characterize how protein folding thermodynamic and kinetic properties are distributed in sequence space. This global view of sequence space allowed me to study how a population of protein sequences evolves over time, and the distribution of protein stability and folding rate preferred by evolution.

Integrated Prediction of Protein-Protein Interaction Networks

I have developed a method to predict yeast protein-protein interactions by integrating diverse features based on sequence, function, localization, abundance, regulation, and phenotype information. These features include functional similarity, co-localization, total and relative protein abundance, total and relative mRNA expression, mRNA expression correlation in time-course experiments, co-transcriptional regulation, total and relative essentiality, synthetic lethality, similarity of phylogenetic profiles, conservation of gene neighborhood, and gene fusion in another genome. I developed a logistic regression method to integrate these features, which improves upon the popular Naïve Bayes method by automatically dealing with potential feature correlation and redundancy. The method has been applied to predict membrane protein interactions in yeast, as membrane proteins are particularly difficult to study experimentally. I was able to predict ~4,000 interactions among helical membrane proteins, with an estimated accuracy level comparable to those of the current experimental interactome mapping efforts. I have validated many of these predictions by literature search. The method has also been applied to predict interactions among soluble proteins as well.

Genomic Determinants of Protein Evolutionary Rate

I have carried out an initial study to predict the evolutionary rate of yeast proteins by integrating diverse features based on sequence, structure, function, localization, abundance, regulation, phenotype, and interaction information. These features include amino acid and codon composition, chromosomal location, gene duplication, mRNA expression, protein abundance, essentiality, subcellular localization, functional classification, secondary structure composition, fold assignment, number of interactors, and number of transcriptional regulators. Various machine learning methods, such as logistic regression and Bayesian networks, were used to train the classifier. This integrated probabilistic modeling revealed the otherwise hidden structure of inter-dependencies among diverse properties. I was able to distinguish those features that are strong predictors for evolutionary rate, from those that are at most weak predictors. Furthermore, in some cases I was able to distinguish features that directly correlate with

evolutionary rate, from those that indirectly relate to evolutionary rate through intermediate factors. For example, protein abundance directly correlates with evolutionary rate; codon adaptation index and number of interactors indirectly correlate with evolutionary rate through protein abundance. These demonstrate the power of integrated genome-wide data mining.

Collaborations in Genomics and Proteomics

I have collaborated with various experimental researchers on several aspects of computational genomics and proteomics, such as analyzing signal sequences and protein binding sequences, predicting single nucleotide polymorphisms (SNP), and analyzing sequence, structural, functional, and network properties of subsets of the proteome (such as all chaperone substrates).

Future Research

Integrated Prediction of Bio-Molecular Networks

I will extend my current work on integrated prediction of yeast protein-protein interaction networks in two directions. First, I will reconstruct protein-protein interaction networks for different organisms, by combining three sources of information: sequence and functional genomic data on the organism, experimental protein-protein interaction data on the organism based on mass spectrometry and two-hybrid system, and interolog mapping (i.e. the transfer of interaction annotation from another organism using comparative genomics). Second, I will make integrated prediction for other types of bio-molecular networks as well, such as transcriptional regulatory network. This will be done by integrating the following sources of information: sequence, function, localization, phenotype, expression, interaction, and experimental protein-DNA binding data based on chromatin immunoprecipitation.

Structural, Biophysical Modeling of Protein-Protein and Protein-DNA Interactions

In addition to statistical integration of sequence and functional genomic information, I will also use structural and physical principles to predict protein-protein interaction and transcriptional regulatory networks. The many experimentally determined three-dimensional structures of protein-protein and protein-DNA complexes provide a set of templates upon which I will model the structure and energetics of other homologous protein-protein and protein-DNA interactions. The models will be constructed using energy minimization with full side-chain flexibility and restricted main-chain flexibility. The energetics of the models will be evaluated using a combination of standard molecular mechanics force fields, implicit solvation treatment, and knowledge-based energy functions. Energy calculations on these structural models will allow me to predict effects of mutation on binding specificity, and serve as additional evidences to be integrated into protein-protein and transcriptional regulatory interaction predictions.

Analysis of Bio-Molecular Networks: Topology, Function, Dynamics, Evolution

In addition to protein-protein interaction and transcriptional regulatory networks, there are other types of bio-molecular networks such as metabolic networks and signal transduction networks. Each individual network is scale-free in topology, and can be decomposed into modules and network motifs. Importantly, these different networks need to be modeled together for a proper understanding of cell behavior. I will develop methods to characterize the topology, and detect modules and motifs for the combined network. Secondly, I will study the functional significance of network topology by comparing various topological measures with corresponding functional genomic data. Thirdly, I will study how network topology changes in different conditions (such as different time points, or different cell types) by combining network topological analysis with

condition-specific functional genomic data. Finally, I will study the evolution of networks by combining network topological analysis with comparative genomic analysis.

Physical and Evolutionary Simulation of Bio-Molecular Networks

I will continue to work on methods for simulating thermodynamics, kinetics, and evolutionary dynamics of complex bio-molecular systems. During my Ph.D. work I focused on the complex system of protein folding. Now I plan to extend the simulation methodology to bio-molecular networks. I will study how fundamental physics and network structure determine network dynamics and function, and how mutations and perturbations in the network affect its dynamics and function. In addition, I will study the population dynamics of network evolution, and the design principles of evolved networks. Such principle-driven studies of bio-molecular networks, combined with the data-driven studies outlined in the last section, will provide deeper insights into the interplay among fundamental physics, structure/topology, function/dynamics, and evolution of bio-molecular networks.

Genomic Determinants of Disease-Related Proteins

In addition to protein evolutionary rate, I will apply our integrated probabilistic modeling approach to study other important protein properties as well. For example, I will study the genomic determinants of cancer-related proteins, defined as yeast proteins with human homologs implicated in cancer. I will systematically determine the sequence, structural, functional, and network characteristics that differentiate cancer-related proteins from other proteins. I plan to collaborate with experimentalists to apply this approach to other specific biological systems.

Methods for Integrated Analysis in Genomics and Proteomics

Functional genomic datasets are complex, heterogeneous, and noisy. They involve many types of molecules (DNA, RNA, proteins, small molecules, etc.), properties (sequence, structure, expression, function, etc.), and relations (interaction, transcriptional regulation, enzyme catalysis, etc.). I plan to develop new ways to represent and integrate these complex entities and relations, and new machine learning methods to model and exploit the statistical structure of these datasets for more powerful knowledge discovery.

Structural and Functional Annotation of Proteomes

I will continue my Ph.D. work on protein structure prediction and structure-based function prediction. For a protein sequence with unknown structure and function, I will break it down into predicted sequence domains, and make structure predictions for each domain using a combination of homology modeling, threading, and *ab initio* structure prediction. Furthermore, I will attempt to predict the function of a protein by looking for binding sites in the resulting structural model. With recent advances in structural genomics and protein structure prediction methodology, it is now possible to build structural models useful for function prediction for a significant fraction of the proteome.

Future Collaborative Efforts

I have collaborated extensively with experimental researchers. My philosophy is to cooperate and collaborate with many different experimentalists, so as to facilitate new interdisciplinary insights. As a bioinformatician, I will continue to collaborate closely with research groups conducting large-scale functional genomic experiments. I will help set up the bioinformatics core infrastructure, and facilitate the design of the experiments. Finally, such collaborative efforts will allow me to incorporate the resulting first-hand data into my own work.

Statement of Teaching Interests

Yu (Brandon) XIA, Ph.D.
Yale University, New Haven, CT

Teaching Philosophy

My teaching philosophy stems from my past teaching experiences. I was a teaching assistant for one graduate-level computational biology course ("Computational Structural Biology"), two undergraduate-level physical chemistry courses ("Physical Chemistry Principles"), and one chemistry laboratory course ("Chemical Separations"). In addition, I have supervised the research of several undergraduate and first-year graduate students at Stanford University and Yale University. I have also tutored several undergraduate students at Beijing University and UC Berkeley on the subjects of mathematics, physics, and chemistry.

The first goal of my teaching is to stimulate the interaction with students. Courses on computational biology and bioinformatics are special because they are inter-disciplinary, and students tend to have diverse background ranging from physics, computer science, to molecular biology and medicine. They speak different scientific language, and they tend to approach the same problem from different viewpoints. By encouraging students to interact, they learn to always keep an open mind, and learn to approach scientific problems from a more comprehensive perspective. Quite often such "cultural clashes" can also lead to creative ideas.

The second goal of my teaching is to get students involved in hands-on research. In addition to theory and algorithms, I will also teach students practical online databases, tools, and resources for genomic and proteomic analysis. Students will be asked to use these tools to study a biological problem of their own interest, and submit a term paper describing the research. I found these term papers to be often stimulating, and serve as a starting point for further independent research.

Teaching Plan

I am able to teach a module in general biochemistry and molecular biology. I will cover the following areas: sequence and structure of bio-molecules, physical chemistry principles in biology, overview of bioinformatics and computational genomics.

I can also teach an upper-division or introductory graduate course on computational biology and bioinformatics. The goal of the course is two-fold: to provide students a general overview of the field, and to teach them hands-on bioinformatics skills that can be used in their own research. This course can also be compressed into a module and offered as part of a graduate-level course. The following topics will be included: tools and databases, sequence comparison and classification, protein structure comparison and classification, energetics of protein structure, protein structure prediction, molecular simulations, analysis of microarray data, analysis of protein-protein interaction and regulatory networks, statistical and machine learning methods in genomics and proteomics.

I am also interested in offering an advanced seminar course in computational biology for students in the bioinformatics track. Recently published research papers in leading journals will be discussed in detail in the course.