COLUMBIA UNIVERSITY
College of Physicians
and Surgeons

JINFENG LIU
Associate Research Scientist

Dept. of Biochem. & Mol. Biophys.
Center for Comput Biol. & Bioinformatics

JL840@columbia.edu

October 27, 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 Professor Brun:

I am writing to express my interest in the tenure-track Assistant Professor position you recently advertised through *Science* magazine.

I am currently an associate research scientist in Professor Burkhard Rost's laboratory in the Department of Biochemistry and Molecular Biophysics at Columbia University. My research at Columbia has focused on comparative genomics and structural bioinformatics with strong emphasis on large-scale sequence analysis and machine learning method development. I have also been actively involved in the North East Structural Genomics Consortium (NESGC), and have contributed significantly to the successful renewal of its multi-million dollar grant from NIH. My research has resulted in more than twenty papers published in peer-reviewed journals. I also received solid training in pharmacology, molecular and cellular biology. This, together with a keen awareness that the success of a computational biologist depends largely upon effective communication with experimental biologists, enabled me to closely collaborate with colleagues within and outside of the department, including the Whitehead Institute, RIKEN Institute in Japan, and NESGC. These fruitful collaborations have not only led to several publications, but also significantly enriched my scientific training by exposing me to leading scientists in the field. In recognition of my productivity and research independence, I was promoted to the rank of "Associate Research Scientist" only one year after starting my postdoctoral research.

I am convinced that my training and research experience in both experimental and computational biology have prepared me well for an independent career in interdisciplinary research focused on system biology. I look forward to the exciting opportunity to join the research community at Indiana University and contribute to its success in biological research. As requested, I have enclosed my curriculum vitae, statements of research and teaching interests, and representative publications. I have arranged for four letters of recommendation to be sent to you directly. Should you need any additional information, please call me at (212) 865-0682 or email me at JL840@columbia.edu.

Thanks for your consideration. I look forward to hearing from you.

Sincerely,

Jinfeng Liu, Ph.D.

Research Statement

Introduction

One characteristic of bioinformatics that distinguishes it from other branches of biology is its focus on data-driven research. In the post-genomics era, advances in experimental techniques have led to an explosion of biological data, including the sequences of many genomes and transcriptomes, and results from high throughput experiments such as microarray, mass spectrometry and large scale protein-protein interaction studies. Bioinformatics is playing a more and more important role in analyzing large-scale biological data and providing guidance for hypothesis-driven traditional biology.

During the last seven years, my research has followed three different but integrated directions to solve very diverse biological problems.

1. *Comprehensive data collection, integration, and efficient representation.* I characterized genome-wide structural and functional features of entirely sequenced genomes using various bioinformatics tools [1], and built a database of Predictions for Entire Proteomes (PEP) with a searchable web interface [2]. As a comprehensive data source, it not only provided the bedrock for my further computational research, but also can help experimental biologists to retrieve and integrate knowledge easily in order to formulate hypotheses more quickly and systematically.
2. *Large scale data analysis and interpretation.* Statistical analysis and data-mining on genome-wide data open the door to asking and answering biological questions at the genomic and system level, rather than focusing on individual biomolecules. My genome-wide sequence analysis led to the discovery of a class of proteins that have long regions of NO-Regular Secondary Structure (NORS) and appear to play important biological functions [3]. As a member of the bioinformatics team of the North East Structural Genomics Consortium (NESGC), I am responsible for selecting targets for structure determination by experimental structural biologists. I developed a homology-based protein domain dissection method and a protein family clustering method, and applied them and other bioinformatics tools to establish an automatic target selection pipeline for NESGC [4].
3. *Using machine learning techniques to predict unknown data.* Bioinformatics is not only useful in analyzing experimental data; more importantly, it has great power to go beyond the existing data to make predictions about unknowns using machine learning methods such as neural network and support vector machines, particularly in the areas where gathering large-scale experimental data is still infeasible or too laborious. I developed two methods in this regard: one for protein domain prediction called ChopNet, and the other for distinguishing protein-coding RNAs from non-coding RNAs (ncRNAs). ChopNet [5] is a *de novo* method to assign domain boundaries for a protein sequence using artificial neural network. It takes into account the information from secondary structure, solvent accessibility, multiple sequence alignment, and amino acid composition and flexibility. The support vector machines (SVMs)-based ncRNA predictor classifies transcripts according to features they would have if they were coding for proteins [6]. It predicted more than 24,000 ncRNAs from the 102,801 transcripts identified in the third phase of the "Functional Annotation of Mouse cDNAs" (FANTOM) project.

Moving forward, I would like to continue my research efforts along these three directions. I will rely on my expertise in large-scale data analysis and machine learning to establish a research program focused on understanding the relationships between protein sequence, structure, and function, and their implications in biological systems as a whole, and human disease in particular. Deeply aware that the success of computational biologists depends critically upon understanding biological problems and effective communication with experimental biologists, I will establish collaborations with wet-lab colleagues within and outside of the department, and will integrate data-driven science with traditional hypothesis-driven research. More specifically, the three projects proposed herein are examples of how computational approaches can be used to analyze and answer important biological questions. Separately, these projects will tackle three distinct biological problems: study of the protein tyrosine phosphatase family, alternative splicing, and prediction of post-translational modifications; yet put together, they will all contribute to our understanding of protein tyrosine phosphorylation-regulated events, e.g., how different tyrosine phosphatases are related to each other structurally and functionally, how alternative splicing affects the function of phosphatases and kinases, and what are the phosphorylation sites in target proteins.

Project 1: Systematic bioinformatics analysis of protein tyrosine phosphatases

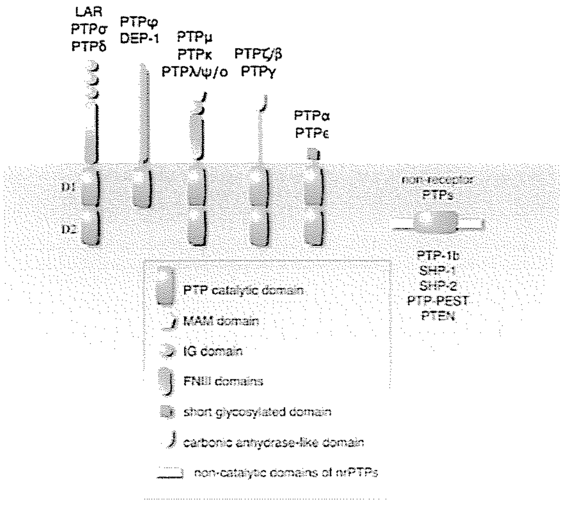

Figure 1. Schematic of the PTP family. Adopted from Stoker [7]

Significance

Protein tyrosine phosphorylation, regulated by protein tyrosine kinases (PTKs) and protein tyrosine phosphatases (PTPs), is a cornerstone of many cellular functions, including cell-cell adhesion, cytokine signaling, metabolic homeostasis and neural development (reviewed in [7]). In vivo, tyrosine phosphorylation is reversible and dynamic; disruption of balance between counteracting PTKs and PTPs has been implicated in human diseases such as cancer, diabetes and inflammation (reviewed in [8]). Therefore, understanding the regulation and function of tyrosine phosphatases in the human proteome is a prerequisite to defining the molecular mechanisms that underlie signal transduction in health and disease and ultimately to revealing opportunities for therapeutic intervention. Hundreds of PTP sequences have been identified and documented in protein databases, and many structural and functional studies have been done. Recently, an on-line database of PTPs was published [9]; however, it does not contain detailed function annotation for individual PTPs. Moreover, because it is not designed to be searchable and lacks the ability for user interaction, it is more like a web-accessible collection of information than a database in the truest sense. A database of PTPs that catalogues comprehensive experimental annotations and in-depth bioinformatics analysis and has a user-friendly interface would greatly benefit both hypothesis-driven experimental research and systematic bioinformatics study.

Specific aims

1. Construct a database of protein tyrosine phosphatases with a web searchable interface.
2. Apply bioinformatics tools to systematically study protein tyrosine phosphatases with regard to their sequence-structure relationship, biological pathways, and to identify new members in the superfamily.

Feasibility

Protein sequences and brief functional annotation can be obtained from SWISS-PROT database [10], and structural information from Protein Data Bank [11]. Detailed functional information is available from the literature. From my past experience with building the PEP database [2] and other projects, I am very familiar with both database construction and bioinformatics tools, including sequence alignment, protein family analysis, structural alignment, and homology modeling.

Project 2: Impact of alternative splicing on protein structure and function

Significance

Alternative splicing is a process in eukaryotic pre-mRNA processing where different sets of exons are joined resulting in different mature mRNAs from the same pre-mRNA. Although it was originally estimated that alternative splicing only accounted for a small percentage of splicing events, recent EST data and transcriptome studies indicated that more than 60% of the transcription units in human and mouse are alternatively spliced [12,13], making it more the rule than the exception. It has been suggested that alternative splicing plays an important role in diversifying gene products given the surprisingly low number (~30,000) of genes identified by the Human Genome Project [14,15]. Most alternative splicing events affect protein coding sequence, with half of them resulting in frame shift. Structures of several alternatively-spliced proteins showed only small changes in local structures without overall changes of protein fold (reviewed in [16]). Functionally, alternative splicing mostly results in subtle functional modulations, such as altering the allosteric regulation sites and the subcellular localization of proteins (reviewed in [17]). In the case of receptor protein tyrosine phosphatases, alternative splicing appears to regulate ligand binding specificity and spatial and temporal expression patterns [18]. Despite the growing interests in understanding the importance of alternative splicing, experimental knowledge of its implications for protein structure and function has remained limited, and there is no comprehensive computational study to date.

Specific aims

1. Analyze effects of alternative splicing on protein domain organization, local secondary structure, and topology of transmembrane helices
2. Investigate impacts of alternative splicing on protein functions, such as subcellular localization and disruption of functional motifs, and potential implications for human diseases.

Feasibility

Recently, a mouse transcriptome study published by the FANTOM consortium identified more than 180,000 transcripts, with 65% of the transcription units containing multiple splicing variants [13]. The availability of such a big data set provides an excellent opportunity to study the implication of alternative splicing on a genomic scale. Numerous accurate computational methods are available across the spectrum of protein structure and function, including prediction methods for secondary structure [19], transmembrane helix [20], and subcellular localization [21], and sensitive search algorithms for protein domain and functional motifs [22-24]. I am very familiar with these tools, and in fact have used them for a comparative study on genome-wide structure and function features [1].

Project 3: Predicting protein post-translational modifications

Significance

Post-translational modifications (PTMs) are covalent modifications of peptide chains by either the addition of modifying groups to some amino acids or by proteolytic cleavage. They occur almost ubiquitously in eukaryotic proteins, and play critical roles in modulating protein functions. For example, phosphorylation on serine, threonine, or tyrosine residues is essential for cell signaling; ubiquitination on lysine residues regulates protein turnover; and proteolysis is required for activation of many enzymes. Detection of PTMs thus becomes an important biological problem. Traditionally, PTMs have been studied by standard molecular techniques such as site-directed mutagenesis and purification of modified proteins, approaches which are often laborious. Recently, high throughout proteomics methods like mass spectrometry have shown promise in determining PTMs on a large scale (reviewed in [25]). However, these methods are still in their early stages of development due to technical limitations. Accurate computational prediction methods, therefore, can play important roles in quickly identifying potential modification sites, or at least providing a short list of such sites for ultimate experimental verification.

Specific aims

1. Develop a prediction method to identify protein phosphorylation sites.
2. Develop a method to predict preferred sites for protein ubiquitination.
3. Develop a prediction method to identify sites for glycophosphatidlyinositol (GPI) anchor attachments.

Feasibility

A significant amount of experimentally verified data is a prerequisite for any machine learning task. Protein phosphorylation and GPI-anchor sites have been extensively documented in several databases [10,26]; precise ubiquitination sites for a significant number of yeast proteins have also been available from proteomics studies [27]. Several previous computational methods based on protein sequence features have been developed for both phosphorylation and GPI-anchor sites more than five years ago; however, they didn't incorporate either the evolution information or local secondary structure features, which have been shown to significantly improve prediction accuracy in related methods. The rapid increase of annotated data in recent years also warrants revisiting these problems using new data and new learning techniques as well.

Reference

1. Liu J, Rost B (2001) Comparing function and structure between entire proteomes. Protein Sci 10: 1970-1979.
2. Carter P, Liu J, Rost B (2003) PEP: Predictions for Entire Proteomes. Nucleic Acids Res 31: 410-413.
3. Liu J, Tan H, Rost B (2002) Loopy proteins appear conserved in evolution. J Mol Biol 322: 53-64.
4. Liu J, Hegyi H, Acton TB, Montelione GT, Rost B (2004) Automatic target selection for structural genomics on eukaryotes. Proteins 56: 188-200.
5. Liu J, Rost B (2004) Sequence-based prediction of protein domains. Nucleic Acids Res 32: 3522-3530.
6. Liu J, Gough J, Rost B (2005) Distinguish Protein-Coding from Non-Coding RNAs Using Support Vector Machines. PLoS Genetics: submitted.
7. Stoker AW (2005) Protein tyrosine phosphatases and signalling. J Endocrinol 185: 19-33.

8. Andersen JN, Jansen PG, Echwald SM, Mortensen OH, Fukada T, et al. (2004) A genomic perspective on protein tyrosine phosphatases: gene structure, pseudogenes, and genetic disease linkage. Faseb J 18: 8-30.
9. Andersen JN, Del Vecchio RL, Kannan N, Gergel J, Neuwald AF, et al. (2005) Computational analysis of protein tyrosine phosphatases: practical guide to bioinformatics and data resources. Methods 35: 90-114.
10. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, et al. (2005) The Universal Protein Resource (UniProt). Nucleic Acids Res 33: D154-159.
11. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
12. Hanke J, Brett D, Zastrow I, Aydin A, Delbruck S, et al. (1999) Alternative splicing of human genes: more the rule than the exception? Trends Genet 15: 389-390.
13. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, et al. (2005) The transcriptional landscape of the mammalian genome. Science 309: 1559-1563.
14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921.
15. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. Science 291: 1304-1351.
16. Stetefeld J, Ruegg MA (2005) Structural and functional diversity generated by alternative mRNA splicing. Trends Biochem Sci 30: 515-521.
17. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, et al. (2005) Function of alternative splicing. Gene 344: 1-20.
18. Johnson KG, Van Vactor D (2003) Receptor protein tyrosine phosphatases in nervous system development. Physiol Rev 83: 1-24.
19. Rost B, Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 232: 584-599.
20. Rost B, Casadio R, Fariselli P, Sander C (1995) Transmembrane helices predicted at 95% accuracy. Protein Sci 4: 521-533.
21. Nair R, Rost B (2005) Mimicking cellular sorting improves prediction of subcellular localization. J Mol Biol 348: 85-100.
22. Hulo N, Sigrist CJ, Le Saux V, Langendijk-Genevaux PS, Bordoli L, et al. (2004) Recent improvements to the PROSITE database. Nucleic Acids Res 32: D134-137.
23. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138-141.
24. Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. J Mol Biol 313: 903-919.
25. Mann M, Jensen ON (2003) Proteomic analysis of post-translational modifications. Nat Biotechnol 21: 255-261.
26. Kreegipuu A, Blom N, Brunak S (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. Nucleic Acids Res 27: 237-239.
27. Peng J, Schwartz D, Elias JE, Thoreen CC, Cheng D, et al. (2003) A proteomics approach to understanding protein ubiquitination. Nat Biotechnol 21: 921-926.

Jinfeng Liu

Statement of Teaching Plans

Teaching Philosophy

Teaching, whether in the form of lecturing in classrooms or mentoring graduate students and postdocs, will be an indispensable part of my research career. Built upon my own experiences from both ends of the education process, as a trainee for many years and as a supervisor for several rotation students and interns, my teaching philosophy is based on the following principles:

1. It is the teacher's responsibility to motivate students to learn, by both stimulating their interests and challenging them to their full potentials. I will strive to relate to my students by starting from what they know and building upon it, and by providing real-life examples about how bioinformatics tools can help their other research projects.
2. Education is not only about passing existing knowledge along to students, but also about teaching them to think critically and independently. For example, when introducing a computational method, it is important to first ask and then allow students to think through questions such as: "What is the underlying biological problem?", "Why are computational methods needed?", and "What is the most suitable method to use?"
3. Presentation matters. I will try to present information in an interesting and easy-to-understand way by extensive use of visual aids and computer labs. In particular, there are many bioinformatics resources on the internet. By relating algorithms in textbooks to real-world web services that implement them, plain materials can suddenly become stimulating.

Teaching Interests

A general course on bioinformatics and computational biology would capitalize on my expertise and give me the opportunity to introduce students to this emerging and critical subject. In such a course, topics ranging from sequence alignment and motif finding, to structure prediction, functional genomics and data integration will be covered to introduce students to the colorful world of bioinformatics.

Another topic that I would love to teach is an introductory course on biological sequence analysis. It can be tailored to give experimental biologists hands-on experience running bioinformatics programs.

I also would like to teach a more advanced course on machine learning and its applications in bioinformatics. The most appropriate audience would likely be graduate students who are interested in doing research in that direction.