November 15, 2005

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

Dear Dr. Yves Brun

I am writing to apply for the position of assistant professor with an emphasis in systems biology using biological network modeling. I am a post-doc fellow at The University of Texas at Austin, working with Dr. Edward Marcotte. I believe that my interdisciplinary research experience makes me a strong candidate for the position.

As my curriculum vitae shows, I have extensive formal training and research experience in a variety of biological problems using a wide spectrum of scientific approaches—from biochemistry, to molecular genetics, to biological data mining, to network analysis. My current scientific interest is in network modeling of biological systems and its application in systems biology in which both computational skills and biological insight are essential.

During my post-doc work, I developed a new biological network model for systems biology which I call *'Probabilistic Functional Gene Network (Pfungene)'* (**Science 306:1555, 2004**). *Pfungene* model provides a very extensive (covers > 94% of proteome) view of the yeast cellular system with highly accurate predictions in system level behaviors. *Pfungene project* is still evolving and I am extending it to other organisms—modeling is in progress for worm (*C. elegans*), human, and pathogenic bacteria (e.g. *M. tuberculosis*)—with a focus on medical or commercial applications. My future research will focus on improving *Pfungene* and analysis of the *Pfungene* model to discover new biological hypotheses—for example, unraveling system organization, assigning new gene functions, discovering new drug targets, predicting phenotypic behavior for a given genotype, integrating *Pfungene* with other biological system models, and comparative system model analysis.

With the increasing importance of interdisciplinary efforts in biomedical research, reliable data analysis is critical for experimental or clinical scientists, likewise quality raw data are indispensable for data analysts. Currently, I am seeking a research position where I can contribute my bioinformatics and systems biology knowledge to the scientific community and benefit from collaborations with inspiring experimental or clinical scientists. I believe your institute is a great place where I can accomplish my scientific goals.

Respectfully,

Insuk Lee

REFERENCES

Insuk Lee (Lee-micro@mail.utexas.edu)

Dr. Edward Marcotte
Director, Center for Systems and Synthetic Biology
Associate Professor, Department of Chemistry and Biochemistry
Institute for Cellular and Molecular Biology, MBB 3.210
2500 Speedway
The University of Texas at Austin
Austin, TX 78712
Phone: 512-471-5435
E-mail: marcotte@icmb.utexas.edu

Dr. Rasika M. Harshey
Professor, Molecular Genetics and Microbiology
The University of Texas at Austin
Molecular Genetics & Microbiology
1 University Station A5000
Austin TX 78712-0162
Phone: 512-471-6881
E-mail: rasika@uts.cc.utexas.edu

Dr. Vishwanath Iyer
Director, Center for Systems and Synthetic Biology
Assistant Professor, Molecular Genetics and Microbiology
Institute for Cellular and Molecular Biology, MBB 3.212AA
2500 Speedway
The University of Texas at Austin
Austin, TX 78712
Phone: 512-232-7833
E-mail: vishy@mail.utexas.edu

Dr. Andrew Fraser
Investigator
The Wellcome Trust Sanger Institute
Wellcome Trust Genome Campus,
Hinxton, Cambridge, CB10 1SA, UK.
Phone: 44-1223-496854
E-mail: agf@sanger.ac.uk

RESEARCH STATEMENT

Insuk Lee (Lee-micro@mail.utexas.edu)

Introduction

Understanding the system-level organization and behavior of an organism is the ultimate goal of the new discipline of *systems biology*. In this statement, I will briefly describe my past and current research in this area, including a new biological network model for systems biology called *probabilistic functional gene networks (Pfungene; see* ***Science 306:1555***) developed during my post-doc research, and I will propose my vision for systems biology *via* network modeling and other data mining approaches.

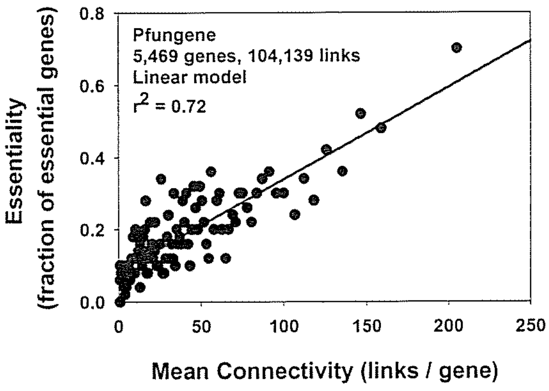
Past & Current research

Probabilistic Functional Gene Network (Pfungene): A New Network Model for Systems Biology

Pfungene is a representation of genes and their functions based upon modeling the coupling between genes as a network. In this model, genes (or proteins) are represented as nodes and functional couplings between genes as weighted edges. Pfungene has two important merits over other protein or gene network models. First, using a generalized notion of gene association (the functional coupling between two genes), we can provide a reasonably *complete view* of a cellular system. Second, because functional coupling includes many specific types of gene associations, we can easily integrate incomplete heterogeneous genomics data.

This Pfungene approach has been applied to yeast to generate a network covering ~81% of the yeast proteome with accuracy comparable to that of small scale experiments (a gold standard set which covers only ~26% of proteome; *see* ***Science 306:1555***). Clustering of genes in the network has successfully modeled functional modules, and PRP43's dual-functions (mRNA splicing and ribosome biogenesis) have been predicted from the network and experimentally validated as well.

The latest version of yeast Pfungene is much improved, covering ~94% of proteome. The model is strongly predictive: (1) it correctly predicts gene function in cross-validation tests; (2) it provides a predictive regression model for genes' essentiality from their network connectivity; (3) it predicts function for unclear genes—we have experimentally validated 8 such predicted functions in ribosome biogenesis.



A predictive regression model for genes' essentiality (measured by a fraction of essential genes in a bin of 50 genes) from their network connectivity (measured by the average number of links connected to a gene) in a yeast Pfungene covering more than 94% of proteome with 104,139 links.

Pfungene can model highly complex organisms as well. As an initial effort, I have constructed a Pfungene model for the nematode *C. elegans*. The worm Pfungene covers ~70% of the worm proteome with accuracy significantly higher than a random model. Experimental verification using RNAi gene inhibition is underway by collaborators in England (Andrew Fraser and Ben Lehner, The Wellcome Trust Sanger Institute).

Research Vision

1. Implementation of New Methods in Biological Data Mining and Integration

Discovery of functional linkages between genes from various genomics data is critical for modeling gene networks. The current version of yeast Pfungene is constructed using information from 11 different types of experimental and computational data analyses. The improvement of current methods in linkage discovery (e.g. extending coverage and reducing false discovery rate) and the implementation of new methods are still major issues. I am interested in developing methods to discover gene functional linkages from unexploited genomics data sets (e.g. genome-scale protein quantitative profiles, genome-scale RNAi knock-out screening, and other genome wide biological features).

Learning new biology using *prior* biological knowledge is another major interest of mine. Supervised learning (learning from reference examples) has proved quite successful in learning biology. Selection, evaluation, and optimization of reference sets are critical, because their quality and usage are major determinants in success of supervised learning. The new yeast Pfungene was improved by using an optimized reference set, arguing that *prior* biological data need to be treated appropriately for better supervised learning in systems biology.

Data integration has recently drawn considerable attention because information about cellular systems often incomplete represented by heterogeneous genomics data with widely varying levels of accuracy. I have developed a new data integration method (a variant of naïve Bayesian integration that is simple but still handles data correlation), and its initial application to Pfungene demonstrated many advantages over other methods (e.g. the Bayesian Net approach). I would continue to develop new data integration algorithms.

2. Inference of New Biological Hypotheses from a Network Model Analysis

For highly complex systems, even simplified network models are too complicated to interpret *via* manual investigation. Therefore extraction of biological information from networks in a simple manner that biologists can appreciate is critical. Two approaches have been applied with success: (1) finding functional annotation enrichment among network neighbors of a given gene to assign its new function, and (2) clustering of genes in the network to discover functional modules and their organization. These two approaches still have huge room for improvement. I am especially interested in developing methods identifying overlapping functional modules from Pfungene, because it would provide more biologically plausible models.

Network topology analysis is another promising approach for studying a biological system with network models, and its application to the latest version of yeast Pfungene has already provided many striking insights about yeast cellular system organization. I would like to improve applications of network topology analysis to learn new biology from network models.

3. Development of a Public Data base of Pfungene models

One of the important requirements in the systems biology area is the need for extensive and reliable data sets. In this regard I have already completed initial versions of Pfungene models for yeast, worm (*C. elegans*), human, and the pathogen *M. tuberculosis*, and have plans to work on other organisms of scientific interest. Incorporating Pfungene data in the public domain will greatly benefit the scientific community.

4. Comparative Systemics

As we obtain system-level models from many different organisms, we can begin to carry out a comparative study of systems (e.g. comparing topological properties or organization of functional modules). I think of this field as *comparative systemics*. As comparative genomics has provided powerful tools in understanding genome organization and function, comparative systemics would be an effective approach to understand system-level organization and behaviors of organisms. I would like to apply my network models to comparative studies between organisms and begin to develop this exciting new field.

5. Multi-dimensional Gene Annotation and System Modeling

Better understanding of an organism requires high-dimensional information like gene functions, gene interactions, temporal (e.g. different cell cycle time or different developmental stages) and spatial (e.g. different environments or different tissues) data. Most current gene functional annotations contain only minimal consideration of the dynamics of gene functions. For a given spatial or temporal condition, each gene has different interacting partners forming different functional modules to carry out different biological processes. Therefore multi-dimensional functional annotation—with dimension of time, space, interacting genes, and so on—is ultimately mandated for systems biology.

My current Pfungene model is a comprehensive but static network (in other words, a network containing all possible gene relations in many conditions). I am also interested in extracting multiple network models with tissue or developmental stage specific manner from a Pfungene of higher eukaryote such as human or plant.

6. Integration of Heterogeneous System Models

Due to the unmanageable complexity of biological systems, a wide spectrum of modeling approaches is mandated to understand system-level behaviors. Pfungene is an extensive but simple model, and it needs to be dynamically connected to other system models to provide more insightful models. In other words, if we integrate information from different system models, we can see multiple types of system-level properties for a given system in a single integrated model. For example, Pfungene is highly informative about system organization (identifying functional modules and their organization) but not about system control architecture (identifying which gene controls which set of genes). In contrast, transcription factor regulatory networks capture information flow from transcription factors to downstream targets. If we can dynamically integrate gene regulatory models with functional organization models, we build an integrated model which captures even more of the cell's behavior.

Concluding Remarks

I believe that systems biology is the future of the life sciences. Network modeling approaches have been very successful in understanding system-level properties of organisms, and Pfungene is a new network model with huge potential. My future research will focus on developing and optimizing a variety of data mining techniques for modeling Pfungene, on network analysis to learn new biology, and on publishing Pfungene data to benefit the scientific community. Upon completion of Pfungene for different organisms, I would help develop the field of comparative systemics to use comparative studies to explore cellular system organization. Furthermore, stratifying Pfungene into multiple conditional models or integrating Pfungene with other types of system models would provide more informative and insightful system-level models, eventually enabling us to predict system-level behaviors of organisms for a given condition, which is the ultimate goal of systems biology.

Selected publication

1. **Insuk Lee**, Rammohan Narayanaswamy, Edward Marcotte. Bioinformatic prediction of yeast gene function, in **Yeast Gene Analysis** (ed M. Tuite and A. Brown) Elsevier. **In press**.
2. **Insuk Lee,** Shailesh V. Date, Alex T. Adai, Edward Marcotte. A Probabilistic functional network of yeast genes. **2004 Science** 306:1555-1558
3. Bork, P., Jensen, L.J., Von Mering, C., Ramani, A.K., **Lee I**, Marcotte, E.M. Protein interaction networks from yeast to human. **2004 Curr. Opin. Struct. Biol.** 14:292-9
4. Manuscripts about yeast Pfungene version 2 and worm Pfungene are in preparation.

TEACHING STATEMENT

Insuk Lee (Lee-micro@mail.utexas.edu)

Teaching Philosophy

I believe teaching is the best way to learn. As a professor in a teaching institute, continuous learning is mandatory to teach a fast-moving field, such as biological science. I believe scientific communication with students via course teaching and mentoring would nurture not only their intelligence but also my research ability, because many of their scientific perspectives out of fresh mind would be inspiring.

Teaching Experience

I have a scientific background in microbiology from my graduate school. Thus, most of my class teaching experience is related to microbiology. As a graduate teaching assistant I taught two semesters of General Microbiology at The University of Texas at Austin. I typically gave a 30 minute lecture and demonstration, and supervised 1-2 hours of laboratory class. For these courses, I was fully responsible for designing class exams (except midterm and final exams), grading all class work, and assigning the final grades for the semester. Although these courses were at an undergraduate level, I was able to learn general skills about managing college-level course work.

During my senior graduate years, I trained junior graduate students in our research group. This training involved mostly introducing our research field and laboratory techniques, as well as planning their research projects.

Teaching Plans

Because my current scientific expertise is in computational biology, I am interested in teaching opportunities in that field. Since I took formal courses across different fields, including molecular biology, microbiology, biochemistry, computer science, mathematics, and bioinformatics, I can design bioinformatics courses for students with diverse academic backgrounds. Main topics for the general bioinformatics course would include a biology introduction (for non-biology majors), a programming introduction (for biology majors), sequence analysis, information theory, microarray data analysis, machine learning applications, biological networks, and other topics related to the field.

For team teaching opportunities, I am interested in network biology and protein-protein interaction. I would also be interested in seminar style courses with a narrower scope for network biology or systems biology.

In addition to teaching advanced courses in my research area, I am also interested in teaching introductory undergraduate courses in biological science—for example, biochemistry, genetics, microbiology, cell biology, or molecular biology—to support the department's teaching efforts for undergraduate level education.

I believe, as a faculty in a teaching institute, mentoring individual graduate students is the most important teaching responsibility. This academic interaction with graduate students, however, would be the biggest benefit for professors because of the endless supply of fresh ideas.

I was fortunate to have great mentors who really care about my career and life during my graduate school and post-doc training. I look forward to creating the same atmosphere for my graduate students.