X-Sieve: CMU Sieve 2.2
To: Jeremy Bennett <jebennet@indiana.edu>
From: Yves Brun <ybrun@indiana.edu>
Subject: Fwd: application for the assistant professor in Systems Biology/Microbiology
Date: Sun, 20 Nov 2005 13:37:26 -0500
X-Mailer: Apple Mail (2.623)

Begin forwarded message:

> From: Jiajian Liu <jjliu@ural.wustl.edu>
> Date: November 20, 2005 1:19:11 PM EST
> To: Yves Brun <ybrun@indiana.edu>
> Subject: application for the assistant professor in Systems Biology/Microbiology
>
> Dear Dr. Brun:
> I am currently a post-doctoral research associate in computational biology in the laboratory of Dr. Gary Stormo, Department of Genetics, Washington University School of Medicine in St Louis. I am writing to apply for the assistant professor position in System Biology/Microbiology in your department as shown in the advertisement in Science (Sept 1, 2005). But I am not sure whether this position is still open or not.
>
> Over the past four years, I have mainly conducted two major research projects: One is to develop computational and experimental approaches to identify DNA recognition code for C2H2 Zinc finger transcription factors. This method will allow one to predict the transcription factor binding sites and their regulated genes within the genome based on the TF protein sequence. Two, I applied computational approaches to make genome-wide associations between transcription factors with their cognate binding sites in bacterial genomes. These studies have produced four publications in peer-reviewed journals (NAR, 2003, 2005, BMC bioinformatics, 2005, and a paper submitted to nature biotechnology).
>
> My Ph.D. research was performed under the supervision of Dr. Peter Zuber in the field of bacterial cellular differentiation and development in B. subtilis, Details about important contributions can be seen from my publications (J. Bacteriol., 1998, Mol. Microbiol, 1999, 2000a, 2000b).
>
> The research statement for my post-doctoral research and my future plan is attached for your review. If this position is still open, I will arrange my references to be sent to you, I will also mail you my application materials very soon.
>
> I look forward to hearing from you. Thanks for your kind assistance!
>
> Best regards,
>
> Jiajian

--

Jiajian Liu, Ph.D

Department of Genetics
Washington University Medical School
660 S. Euclid,Box 8232
St. Louis MO 63110
Phone: (314) 747-5535
Fax: (314) 362-7855
Email: jjliu@ural.wustl.edu
URL: http://ural.wustl.edu/~jjliu


application_systemBiol_microbiol_2005_IndianaU.pdf

--Yves

http://www.bio.indiana.edu/facultyresearch/faculty/Brun.html

Yves V. Brun
Professor and Director, Microbiology Program
Department of Biology, Indiana University
Bloomington IN 47405-7005
Phone: 812-855-8860 (office), 855-7239 (lab)
Fax: 812-855-6705

Dear Faculty Search Committee members:

I am writing to apply for the assistant professor position in Systems Biology/Microbiology in your department. I am currently a post-doctoral research associate in computational biology in the Laboratory of Dr. Gary Stormo, Department of Genetics, Washington University School of Medicine in St Louis. I believe that my extensive research experience combined with my strong interdisciplinary training in both biology and computational sciences make me a strong candidate for the position outlined in your notice.

One of the most challenging problems for post-genomic research is to associate on a genome-wide basis transcription factors (TFs) with their cognate transcription factor binding sites (TFBSs). Over the past four years, I have successfully developed a computational method to identify the DNA recognition code for eukaryotic C2H2 Zinc-finger TFs. This method will allow one to predict the TFBSs and their regulated genes within the genome based on the TF protein sequence alone for given C2H2 zinc finger TFs, and *vice versa*. In addition, I have also been working on the development of computational methods to make genome-wide associations between TFs with TFBSs in bacterial genomes. These studies have produced four publications in leading journals. Gains from these studies will help elucidate the global connectivity between TFs and TFBSs and construct potential regulatory networks on a genome-scale.

My Ph.D. research was performed under the supervision of Dr. Peter Zuber, entitled "Control of RNA polymerase sigma subunits that are required for developmentally regulated transcription in *Bacillus subtilis*". My Ph. D. studies made several important contributions to furthering the understanding of the molecular mechanisms controlling cellular differentiation and development in *B. subtilis*. Details can be seen from my publications (*J. Bacteriol.*, 1998, *Mol. Microbiol*, 1999, 2000a, 2000b).

The research I intend to pursue in the future will follow three directions: One, I will use the computational method that I developed to identify the DNA-binding model for C2H2 zinc-finger TFs in eukaryotic organisms. By combining genome sequence data, microarray expression profiles, and experimental approaches, this information will be applied to further characterize the biological functions for novel C2H2 zinc-finger TFs. Two, as no recognition model exist for TF families other than C2H2 zinc-finger TFs, I will apply computational and experimental approaches to identify the DNA recognition codes for Homeodomain proteins, the second largest TF family in eukaryotes. Three, I will develop new computational methods to make genome-wide association between TFs and TFBSs in bacterial genomes of interest based on different sources of data.

Besides my research in the lab, I would like to develop collaborative connections with members in the department, as well as those within the University. In addition, I hope to bring my interests in these fields into the classroom. I feel confident in my ability to teach genomics, bioinformatics, biochemistry, molecular biology and microbiology to undergraduates, and I also have an interest in teaching more advanced courses on genomics, computational biology and molecular biology. I believe that I am ready for a mentorship role.

Through my Ph. D. research, I have been very well trained in the fields of biochemistry, molecular biology and microbiology. With my part-time graduate studies in the Department of Computer Science and Engineering at Oregon Graduate Institute of Science and Technology, I developed my strong skills in computer programming. With my postdoctoral research, I have gained a solid background in the field of computational biology through hands-on experience.

Taken together, I am confident that I have developed strong interdisciplinary skills that will serve me well in tackling the new challenges in the post-genomic era.

Please let me know if I can provide any additional information. I look forward to hearing from you.

Respectfully,

Jiajian Liu

Area of Interest

The ability to associate on a genome-wide basis transcription factors (TFs) with their cognate transcription factor binding sites (**TFBSs**) is a central problem for post-genomic research. Current experimentally based methods for identifying TF-TFBS associations, such as high throughput ChIP-chip technology, can link TFs with their cognate TFBSs under specific growth conditions at particular times. However, these approaches are labor intensive, impractical for all possible growth conditions and cell/tissues. Computational approaches, however, hold the potential to make such connections. My research plan uses a 'TF family-wise' approach by combining structural information and sequence data to predict the DNA recognition codes for novel TFs from the same structural family. Out of 54 TF structural families (*Luscombe et al, 2000*), I focus my studies on the TFs from the two largest families in eukaryotes: Zinc-finger proteins and Homeodomain proteins. In addition, I will continue to apply different sources of data to develop improved computational methods to make connections between TFs and TFBSs in bacterial genomes. All of these efforts are aimed toward identifying a cell's complete genetic network in both eukaryotes and prokaryotes.

Approach and Postdoctoral Work

Computational and experimental approaches to elucidate the DNA recognition code for C2H2 Zinc- finger TFs

The C2H2 zinc-finger protein is the largest TF family in sequenced eukaryotic genomes. Analysis of DNA-Zinc finger protein complexes has revealed that the binding specificity of each Zinc-finger domain is determined by residues at four key positions. Several quantitative methods are used to predict DNA-zinc finger protein recognition codes that specify DNA base-amino acid interactions. At present two major issues greatly limit the accuracy of these predictions: 1) All existing models assume independence of the contributions to binding between the positions in DNA-protein interactions (called the additivity model), ignoring interactions among positions, and 2) limited data exist upon which to infer model parameters for statistical approaches. To address these problems, I first used experimental and statistical approaches to assess the validity of the additivity model and found that the additive models are good approximations for each test protein and DNA alone, but position dependence was clearly observed when both protein and DNA vary simultaneously (*BMC Bioinformatics, 2005*). Based on these results, I developed a non-linear network model that takes position effects into account to predict DNA-binding profiles for Zinc-finger TFs (*submitted*). Comparisons of my predictions to all published quantitative affinity data (over 450 pairs of DNA-EGR proteins) and known DNA-binding motifs for dozens of Zinc-finger proteins revealed strong correlation. This method outperforms all existing methods. Finally, to acquire quantitative data necessary for the development of more accurate models for DNA-Zinc-finger protein interactions, I developed a systematic experimental method by combining affinity chromatography-SELEX with a quantitative binding assay (*NAR, 2005*).

Development of computational methods to make genome-wide TFBSs ↔TFs in bacterial genomes

Until recently little was known about the genome-wide correspondence between a single TF and its regulated genes, even in *E. coli*. By integrating comparative genomics and microarray expression approaches, I developed a computational method to identify members of a bacterial regulon (*NAR, 2003*). Our method was verified by an independent ChIP-chip study by the Losick group. However, it is difficult to extend this approach to all TFs within a genome, as most TFs in the genome lack experimental studies. The goal of one project is to use *Shewanella oneidensi*, a model organism

for environmental bioremediation, to develop an improved computational method for genome-wide TFs ↔TFBSs. We have identified as many as 300 conserved regulatory motifs in *S. oneidensis*. Following the lead of Tan and Stormo (*Tan and Stormo, 2005*), I am using the information in genomic location and phylogenetic correlation and structural constraints to aid in the identification of TFs↔TFBSs in *S. oneidensis*.

Future Research Plan

Despite having numerous sequenced genomes containing thousands of putative TFs, only little is currently known about their DNA-binding specificities. My research plans as an independent investigator focus on three main areas: One, I will use my computational method coupled with experimental approaches to identify the DNA-binding models for all C2H2 zinc-finger TFs in eukaryotic organisms. This information will be applied to characterize the biological functions for uncharacterized C2H2 zinc-finger TFs. Two, as no reliable recognition model exist for TF families other than C2H2 zinc-finger TFs, I will apply computational and experimental approaches to identify the DNA recognition codes for Homeodomain proteins, the second largest TF family in eukaryotes. Three, I will use computational approaches based on different sources of data to make connections between TF with TFBSs in bacterial genomes and to characterize *cis*-regulatory modules in bacteria.

1) TF-family wise identification of the DNA recognition code for C2H2 zinc-finger TFs in *S. cerevisiae*

Although *S. cerevisiae* have been extensively studied for many years, as many as 50% of TFs have not been characterized for binding specificity. I propose to identify the DNA-binding motifs and target genes for all C2H2 Zinc-finger TFs in this organism. I will apply the lessons learned from yeast to other eukaryotic genomes including worm, fly and mammals, ultimately human.

a) Identifying DNA-binding models for all C2H2 Zinc-finger TFs in *S. cerevisiae*

Initially I will use the C2H2 zinc-finger protein HMM-profile model (Hidden Markov Model) to identify all C2H2 Zinc-finger proteins in *S. cerevisiae*. For each identified TF, I will then use its amino acid sequence to predict its DNA-binding motif represented as a weight matrix by the method I developed previously. The specificity and sensitivity of our predictions will be assessed with a collection of published data. Novel predictions will be validated with the *in vitro* "band-shift" assay and the *in vivo* "yeast one-hybrid" assay. For false negatives, I will employ the affinity-SELEX procedure that I developed (*NAR*, 2005) to identify their binding models.

b) Characterization of biological functions for uncharacterized C2H2 Zinc-finger TFs in *S. ceverisiae*

The derived DNA-binding models for C2H2 Zinc-finger TFs will be applied to scan the regulatory promoter regions of all *S. cerevisiae* genes to search for potential binding sites for that TF. I will then filter out those that are not conserved in orthologous promoters of related *Saccharomyces* species. Next we will infer the biological functions for novel TFs from gene annotation and expression data. Further verification of the predicted functions will be performed through *in vivo* analyses of the phenotypes of their mutants.

2) Elucidating the DNA recognition codes for Homeodomain (HD) TFs.

HD proteins in eukaryotes have been demonstrated to play an essential role for cellular growth, differentiation and development through their selective binding to DNA target sites. However, the DNA-binding specificities for most HD TFs have not been characterized. Although evidence suggests a DNA recognition code for HD proteins exists, identification of such a code is beset with a much more challenging problem, because the DNA-HD protein physical contact model has not defined well.

a) Co-evolution analysis to map key residues in HD responsible for DNA-base interactions

Although structure analyses have shown that four amino acid residues in the recognition helix of the HD dictate sequence-specific DNA recognition at the major groove, the positions in the N-terminus which dominate binding of DNA bases within the minor groove remain unclear. In addition, DNA-HD protein interactions complicate the model. I will apply a mutual information analysis between DNA positions and protein positions to identify key residues and their associations with DNA bases. This information will then be used to generate the general DNA-HD protein contact model.

b) Experimental and computational apaches to elucidate the DNA recognition codes for HD TFs

To quantitatively model DNA-HD protein interactions, I will first generate a large collection of data using newly developed bacterial one-hybrid system (*Menget al, 2005*) and phage display. By combining these data with previously published ones, I will then use a 'look-up table' to qualitatively represent the base-amino acid preferences for HD proteins. The log-odds for each possible pairs will be computed to give a relatively accuracy measure. A more sophisticated recognition code will be computed with an approach similar to that use for zinc-finger TFs.

3) Computational and experimental approaches to make TF↔TFBS in bacterial genomes.

Despite several decades of intense research, only ~1/3 of estimated ~300 TFs have been characterized even in *Escherichia coli* and *Bacillus subtilis*. Some people posit that we can use the same 'TF family-wise' approach to identify the DNA recognition codes for bacterial HTH TFs. However, variation of DNA docking arrangements for TFs in different subfamilies makes it difficult for this approach. Here I will adopt a different approach to associate TFs with TFBSs as described below. I use *B. subtilis*, because the availability of its genome sequence and that of six closely related organisms, considerable expression data that would facilitate the method development, while a wealth of experimental data in this organism would allow for prediction validation. In addition, I have considerable experimental experience with this organism (*J. Bacteriol.* 1998, *Mol. Microbiol.* 1999, 2000a, 2000b. and *NAR* 2003).

a) Identification of a complete set of evolutionarily conserved regulatory motifs in *B. subtilis*

The first step necessary for making connections between TFs and TFBSs is to identify two key components, the complete set of conserved regulatory motifs and all TFs in *B. subtilis*. The computational pipeline developed previously for the *Shewanella* project will be directly applied here to identify the database of evolutionarily conserved sequence motifs, while the full repertoire of putative TFs will be identified through family and functional domain search combined with ortholog identification. The sensitivity for the DNA-binding motif predictions will be assessed with a collection of known motifs to examine the predictive coverage. The validated weight matrix models will then be used to scan the complete set of regulatory sequences for all predicted operons to identify additional members of bacterial regulons in *B. subilis*, by the integration of microarray expression data.

b) Computational approaches to associate TFs with their cognate TFBSs in *B. subtilis*

To our knowledge, only one computational method has been developed to make TFs ↔TFBSs associations in bacterial genomes (*Tan and Stormo, 2005*). Two factors could limit their application of that approach here: First, two out of three types of information that they rely upon for predictions are only partially valid. e.g. structural family constraints are not true for many cases. Second, the structure of the bacterial operons was not fully considered. In addition to the phylogenetic correlation and distance constraints used by Tan and Stormo, I will combine expression correlation and operon organization to compute the probabilities of association for any given a TF with different DNA motifs. Second, I will apply Bayesin network model to associate TFs with their target genes based on expression data, TF, and genes bearing TFBS. Sensitivity and specificity of our predictions will be

assessed with motifs for TF that have been characterized, while novel predictions will verified with the *in vitro* "band-shift" assay and the *in vivo* "yeast one-hybrid" assay. Finally, the associations between TFs and TFBSs will be applied to generate genetic networks in *B. subtilis* using a bottom-up approach.

c) Characterization of *cis*-regulatory modules involved in the cellular sporulation process in *B. subtilis*

Even given a complete set of DNA-protein interactions for a given genome, it is still difficult to predict gene expressions, as we have limited understanding of coordination between *cis*-elements (regulatory modules) in the promoter regions of genes. I propose to use the sporulation process in *B. subtilis* as a model to investigate the nature of bacterial regulatory modules. The sporulation process in *B. subtilis* is controlled by six sigma factors, a number of master TFs and a set of specific TFs. Different combinations of these TFBSs occur in different sporulation-involved genes.

There are several algorithms available to predict regulatory modules for eukaryotes. However, relative to higher eukaryotes, less is known about the module constraints in bacteria. To deal with this challenge, I will adopt a *de novo* approach to identify sporulation-related modules. First, I will collect all known sporulation genes and their orthologs in related species and use them to identify a full set of binding motifs with PhyloCon (*Wang and Stormo, 2003*). After eliminating the redundant motifs, the weight matrix models will then be used to scan the promoter sequences of sporulation genes. The identified clusters of DNA-binding sites are putative regulatory modules if the distances between two nearby binding sites for any given clusters follow our constrints. Statistical analysis and experiments will be applied to assess the prediction accuracy. Next goal is to develop a quantitative model to describe physical nature of bacterial regulatory modules.