October 25, 2005

Professor 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 faculty position (advertised in the September 2nd issue of Science) in the Department of Biology and Biocomplexity Insitute. I am currently a postdoctoral fellow working with Professor Hao Li in the Department of Biochemistry and Biophysics at the University of California in San Francisco. I received my Ph.D from the Physics Department at the University of Washington in Seattle, where I studied the theory of non-equilibrium systems.

My main research interests are gene regulation and comparative genomics. During my postdoctoral research at UCSF, I have developed new approaches to understand gene regulation using both bioinformatics and novel experimental techniques. My research in bioinformatics has included cross-species gene expression comparisons, transcription factor target predictions, and quantitative analyses of gene regulation using mechanistic models. I have also applied phylogenetic footprinting to globally measure the conservation of regulatory sequences across yeast species.

To understand gene regulation experimentally, it is essential to measure gene expression with a time resolution at least as fast as the underlying regulatory processes. I have developed a new automated system to measure real time protein abundances *in vivo*. Using this system, the dynamics of gene expression can be monitored at a time resolution of once per minute. This system will provide an enormous amount of information which I plan to use to quantitatively model gene regulation.

I have enclosed my *curriculum vitae* and a research statement, as well as copies of two recent papers. I have also arranged for letters of reference to be sent from Hao Li (UCSF), Joe DeRisi (UCSF), Marcel den Nijs (UW, Seattle), and Jeffrey Chuang (Boston College). Thank you for your consideration, and I look forward to hearing from you.

Sincerely,

Chen-Shan Chin

Chen-Shan Chin, PhD

University of California, San Francisco
1700 4th Street, Box 2542
San Francisco, CA 94143-2542
cschin@genome.ucsf.edu

Research Statement

Chen-Shan Chin

Introduction

Advances in technology in the post-genomic era have led to new opportunities for systems biology approaches to the study of gene regulation. A critical problem in systems biology is to model regulatory networks in a way that reflects the complex relationships between genes. At the moment, regulatory networks are usually modeled *qualitatively*. For example, gene networks are commonly drawn as a set of activating or repressing arrows between genes. This description usually fails to capture many of the combinatorial and dynamic relationships crucial to the functions of the network. A major challenge currently is to integrate experimental approaches with theoretical modeling to build a *quantitative* understanding of the structure, function, and design principles of gene regulatory networks.

I envision two complementary approaches to deciphering gene regulatory networks. The **bottom-up approach** focuses on characterizing the quantitative features of small modules within networks. Gene regulatory networks are analogous to modern electronic devices, which are made of many modular components such as logical gates and op-amplifiers. These components are wired in specific ways to yield higher-level functions. To understand the higher-level functions, we need both the wiring diagram and the quantitative characteristics, such as input-output relationships and the dynamic responses of each component under different conditions. Such quantitative characterization in regulatory networks is currently lacking. New experimental and computational tools are in urgent need for this task.

Parallel to the bottom-up approach is the **top-down approach**, in which one uses data from genome-wide studies to build an accurate wiring diagram describing the links between components and discover emerging properties of the regulatory systems. This requires novel computational tools to integrate various sources of data, such as multi-species genomic sequences, microarray gene expression data, protein interactions, and pathway information.

The long-term goal of my research is to build a comprehensive and quantitative picture of gene regulatory networks by **connecting the top-down and bottom-up approaches**. I will develop and apply both computational and experimental techniques to study gene regulation at a systems level. In particular, I am interested in the following areas:

(1) Development of new technologies to measure the dynamics of gene regulation.

(2) Development of new computational methods to systematically detect modules in gene regulatory networks using bioinformatics approaches.

(3) Study of the emergent properties and biological significance of regulatory modules by integrating the results from (1) and (2) with theoretical modeling based on physical principles.

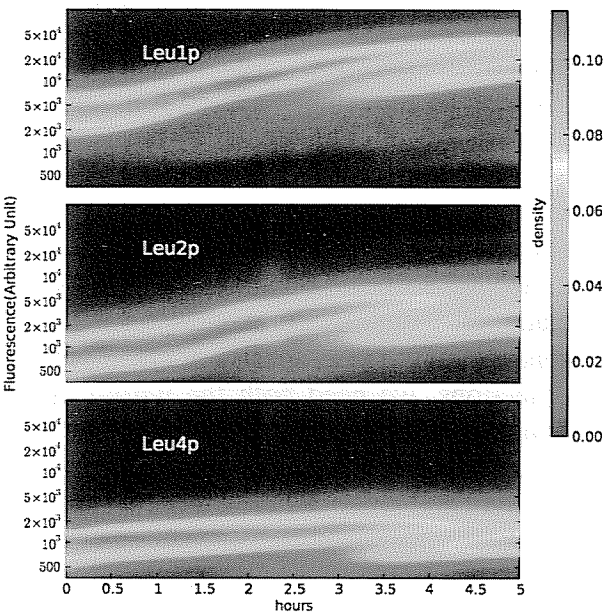
In this statement, I will summarize my research accomplishments followed by the outline of my future research plan.

Research Accomplishments

1. High resolution real time protein dynamics measurements

In collaboration with the DeRisi and Weissman labs at UCSF, I have built a prototype robotic system to measure protein abundance. This system allows simultaneous measurement of multiple protein abundance levels in single cells with high temporal resolution. It automatically delivers samples of cells with GFP-tagged genes from a bank of chemostat reactors to a flow cytometer for fluorescence measurements at a time resolution of once per minute. Such high precision data are crucial for quantitative modeling of the dynamics of gene regulation.

This new system allows us to study the single cell dynamics of gene expression with precision and temporal resolution that far exceed what can be achieved by DNA microarray or standard proteomic methods. The plot shown on the right is an example of measurements taken with this new tool. It shows the dynamic changes of three leucine synthesis enzymes, Leu1, Leu2 and Leu4, fused with EGFP in *S. cerevisiae* under starvation conditions over 5 hours. Each time slice represents the protein abundance distribution within the cell population and the color code indicates the fraction of cells in each bin. More than 5×10^4 cells *in vivo* are measured every minute for multiple strains simultaneously. The width of the distribution at each time slice conveys information about the level of gene regulatory noise. These data show intriguing complexities of gene regulation and pose new challenges for incorporating such data into models.



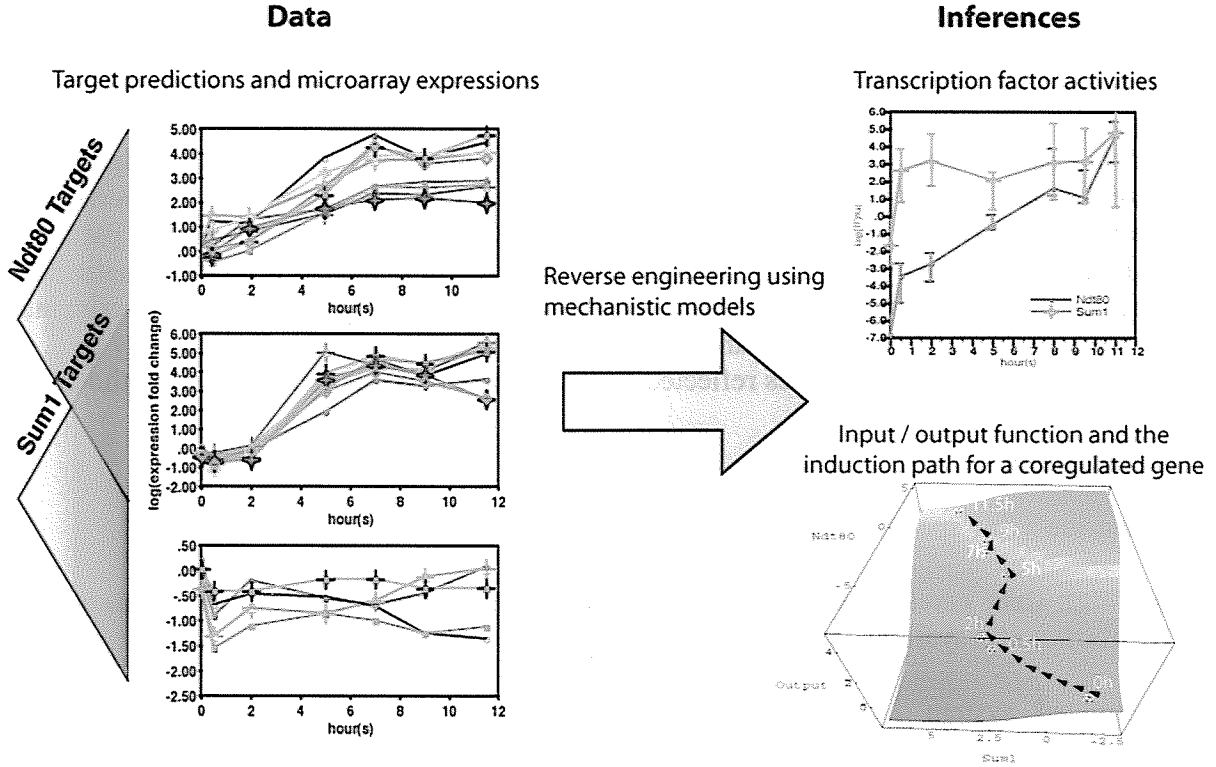
A GFP-tagged library of all *S. cerevisiae* proteins is available [1]. The library and this automatic measurement system make possible systematic studies of the dynamics of cellular responses to a wide range of conditions, which will be described in the Future Directions section below.

2. Analysis of input-output function and the dynamics of combinatorial *cis*-regulation

The gene regulatory network consists of various combinatorial control modules, where the output of a single gene depends on the combination of input signals. The details of the function of these modules are largely unknown. There are two important aspects of these combinatorial controls that need to be scrutinized in order to understand the associated functions: (1) the relationship between the regulatory input (transcription factor activities) and the regulatory output (mRNA production rates) of the genes in the module, and (2) how different transcriptional factors interact to perform combinatorial regulation. I have developed a framework based on a

physical model [2, 3] to infer transcription factor activities and protein-protein/protein-DNA interactions from gene expression data.

The figure below shows an example of such an analysis applied to genes regulated by the transcription factors, Sum1 and Ndt80, which are involved in regulating sporulation in *S. cerevisiae*. We first identified three different groups of genes: those regulated only by Sum1, those regulated only by Ndt80, and those co-regulated by both factors. We use the sporulation gene expression profiles together with the wiring diagram reconstructed from bioinformatics analysis to infer the dynamic activities of these two regulators as well as the protein-protein/protein-DNA interactions on the promoters. The input-output functions for each gene controlled by the regulators are reconstructed using the inferred protein-protein/protein-DNA interactions. The same method has been applied to other combinatorial modules in amino acid synthesis pathways to study dynamic responses to starvation. These inferences will help to interpret the functions of combinatorial modules and shed light on the underlying regulatory mechanism.



3. Measuring the regulatory complexity of yeast by phylogenetic footprinting

In this research, published in Genome Research [4], I explored the global mutational patterns and functional sequence conservation in gene promoters of the *Saccharomyces* genus using comparative genomics. Mutation rates across the *Saccharomyces* genome were measured and the conserved functional sequences in the promoter regions identified. This work had several important conclusions. The mutation processes in *S. cerevisiae* were found to be uniform. This finding allowed us to precisely gauge the amount of functional sequences in the non-coding regions. I found that, for a typical gene in *S. cerevisiae*, about 90 base pairs of promoter sequence are functionally conserved. Each promoter was further characterized by a hidden

Markov model to identify the highly conserved functional regions. This analysis provided rich information about the complexity of transcriptional regulation and the extent of the combinatorial control.

4. Predictions of transcription factor targets using multiple information sources

Although a number of data types relevant to inferring regulatory modules are available, accurate predictions of the targets of each transcription factor remain a challenge. I developed a statistical method that integrates multiple information sources, e.g. binding affinity measurements, sequence conservation, and microarray data, to predict transcription factor targets [5]. The novelty of this method is that it allows us to estimate the total number of targets and the false negative rate of our predictions. Thus, the scale of regulatory modules can be quantified. The method has been applied successfully to the transcription factor, Ndt80, in *S. cerevisiae*, and it can be easily extended to other factors.

5. Cross-species gene expression comparisons and protein interaction network analysis

In collaboration with the Bargmann and Jan labs at UCSF, I have compared gene expression across species to identify conserved transcriptional programs, particularly in the aging processes of *Drosophila* and *C. elegans* [6]. In another project, I developed a graph-theoretical algorithm to analyze the global connectivity properties of the protein interaction network in *S. cerevisiae* [7]. I defined a few quantities to measure (1) how strongly a protein ties with the other parts of the network, and (2) how significantly an interaction contributes to the integrity of the network. Using these measurements, I found that the interaction data obtained from different experimental methods such as immunoprecipitation and two-hybrid techniques contributed differently to network integrities. Such differences reflected the systematic bias inherent in these methods.

Future Directions

Here I describe a few specific directions that I plan to pursue.

1. Using high time resolution protein abundance measurements to study the dynamics of gene regulation at a systems level

The automated system I described above allows high temporal resolution single cell measurement of multiple genes and thus offers a unique opportunity for discovering novel dynamic features of gene regulation that are otherwise unavailable by conventional methods. As an exploratory tool, a systematic survey of the time courses of induction and suppression of all genes responding to a change in environment may reveal interesting information. The timing and amplitude of gene expression and coordination between genes in different pathways can be measured accurately. I propose to expand and improve on this system to include multiple automated reactor banks that would enable sets of genes representing whole pathways to be measured as the cellular environment is systematically perturbed. This system, combined with a large collection of genetic mutants, will allow me an unprecedented ability to accurately quantify cellular expression control networks.

In the next couple of years, I plan to focus on the dynamics of gene regulation in metabolic pathways, such as pathways for amino acid synthesis. More specifically, I would like to address the following questions: (a) Is there coordinated regulation of protein synthesis/stability for the genes on the same metabolic pathway and, if so, what is its relationship to transcriptional regulation? (b) How do combinatorial controls and the network topology affect the dynamics of the regulated genes and their noise levels? (c) What is the importance of the crosstalk between different metabolic pathways? (d) What is the significance of the detailed temporal, amplitude, and noise controls of genetic switches?

For the study of metabolic pathways, I will systematically measure the dynamics of the genes in the pathways under a variety of nutrient conditions. To study the relationship between network structure and quantitative dynamics, I will knock out various regulators involved and mutate the binding sites of transcription factors to alter the regulatory relationships, and then carefully measure the dynamics of the perturbed networks.

I will further develop this technology by improving the throughput and temporal resolution and developing a user-friendly interface to facilitate its use by other researchers. I also plan to combine the system with other quantitative methods such as microarray, qPCR, and microfluidic devices. These other methods will generate data complementary to the protein abundance data that can be integrated by theoretical modeling.

2. Characterization of regulation modules using theoretical and computational analysis

The experimental approaches I have proposed will generate large-scale quantitative data reflecting the output of gene regulatory networks. To understand the underlying mechanism and design principles, we need a theoretical framework to integrate different sources of data and to derive a few general rules governing the systems properties of the networks. There are two important components of this theoretical modeling I would like to pursue. One is to use the bioinformatics tools developed in my previous research to accurately reconstruct the wiring diagram of the networks. The other is to focus on specific subnetworks for quantitative modeling using mechanistic models. The predictions of the theoretical model will guide experiments and experiments in turn can be used to refine models in an iterative process.

For network reconstruction using a bioinformatics approach, I will further develop the computational tools for predicting the binding sites and regulatory targets of transcription factors [5]. In these methods I will combine genome sequence information with functional genomics data such as DNA microarray gene expression data and ChIP-chip data (chromatin immunoprecipitation followed by microarray to map the genomic location of transcription factors). I will also use the comparative genomics method for analyzing functional sequences [4] to improve the sensitivity and specificity of the predictions. Together with the knowledge of the regulatory pathways in the literature, I expect to get a wiring diagram as complete as possible with high confidence.

For mechanistic modeling, I will first focus on the regulatory system of metabolic pathways. Previously, I have developed a computational approach based on a statistical mechanics model to infer the regulatory logic and input signals for genes controlled by multiple transcription factors.

This approach allows us to infer the parameters for protein-DNA, protein-protein interactions on the promoters of each individual gene that are necessary for characterizing the input-output function at each node on the regulatory networks. Combining this information at each node with the larger network diagram and experimental data, I will be able to build theoretical models to (a) extract the kinetic parameters, (b) test our understanding of the network architectures, and (c) find emerging properties and design principles of the networks.

3. Use comparative genomics to study mutational process and functional sequences in multiple species

In an independent direction, I would like to generalize the comparative genomic tool developed in [4] to analyze higher eukaryotes including mouse, chimpanzee and human. I have demonstrated that this method, based on careful calibration of neutral mutation rate, is highly sensitive and capable of identifying functional sequence with a resolution of less than 10 base pairs. I will systematically analyze the mutation processes and characterize the genome-wide regulatory complexity by mapping out the functional non-coding sequences in other organisms.

1. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, Dephoure N, O'Shea EK, Weissman JS: Global analysis of protein expression in yeast. *Nature* 2003, **425**(6959):737-741.
2. Ackers GK, Johnson AD, Shea MA: Quantitative model for gene regulation by lambda phage repressor. *Proc Natl Acad Sci U S A* 1982, **79**(4):1129-1133.
3. Buchler NE, Gerland U, Hwa T: On schemes of combinatorial transcription logic. *Proc Natl Acad Sci U S A* 2003, **100**(9):5136-5141.
4. **Chin CS**, Chuang JH, Li H: Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res* 2005, **15**(2):205-213.
5. Jolly E, **Chin CS**, Herskowitz I, Li H: Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis. *BMC Bioinformatics* 2005, **in press**.
6. McCarroll SA, Murphy CT, Zou S, Pletcher SD, **Chin CS**, Jan YN, Kenyon C, Bargmann CI, Li H: Comparing genomic expression patterns across species identifies shared transcriptional profile in aging. *Nat Genet* 2004, **36**(2):197-204.
7. **Chin CS**, Samanta MP: Global snapshot of a protein interaction network-a percolation based approach. *Bioinformatics* 2003, **19**(18):2413-2419.