HARVARD UNIVERSITY
BAUER CENTER FOR GENOMICS RESEARCH

BAUER LABORATORY
7 DIVINITY AVENUE
CAMBRIDGE, MASSACHUSETTS 02138

November 1, 2005

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

Dear Prof. Brun,

I would like to be considered for the tenure-track assistant professor position in systems biology advertised by your department. I have a Ph.D. degree in applied mathematics and have had extensive experience in developing mathematical and statistical methods. Currently I am a joint postdoctoral fellow at the Center for Genomics Research and the Department of Statistics of Harvard University. I believe that my analytical and biological background is well-suited for the advertised position.

My postdoctoral research at Harvard University has been focused on developing bioinformatics methods for investigating the structure and functions of chromatin. In collaboration with Drs. Oliver Rando and Steve Altschuler, I developed a hidden Markov model approach for analyzing tiling microarray data. By applying my model to analyze nucleosome array data, we were able to identify high-resolution nucleosome positioning information at a genomic-scale in yeast, whereas traditional methods can only generate two or three high-resolution nucleosome positions at a time. By searching public databases for correlations with our data, I discovered that transcription factor binding sites tend to be depleted of nucleosomes at a genomic scale.

I am currently developing statistical methods in order to understand the global gene regulatory role of chromatin. As a first step, I have been investigating the global association between gene expression and histone acetylation levels in collaboration with Dr. Ping Ma and Prof. Jun Liu. I have built linear regression models to estimate the regulatory effects of histone acetylation while controlling the confounding effect of sequence information. Our analysis has offered new insight into the complexity of the histone code.

Prior to joining CGR, my research was focused on analyzing nonlinear dynamical systems and their applications in ocean and atmospheric modeling. I developed computational tools to analyze various dynamic properties in spatial and temporal systems. I will apply dynamical systems techniques to understand the dynamics and control of gene expression patterns.

Attached please find copies of my curriculum vitae, statement of research, teaching interests, and recent paper. Letters of recommendation from Profs. Jun Liu, Tim Mitchison, Andrew Murray, Steve Altschuler, and Jim Yorke will be sent separately.

Please do not hesitate to contact me if you have any questions. I can be reached at the phone numbers 617-496-2997 (Office) and 617-519-8335 (Home). My email address is gyuan@cgr.harvard.edu. Thank you for your attention and consideration of my application.

Sincerely,

Guo-Cheng Yuan

Research Statement

Guo-Cheng Yuan

My research interests lie in the understanding of systems biology by using computational methods. My long term goal is two-faceted. First, I will continue working closely with experimental biologists to investigate the establishment, maintenance, and inheritance of gene expression patterns with particular interest in the role of chromatin. Secondly, I will develop new mathematical and statistical methods that not only are useful for biological data analysis but also may have broader applications.

I have a Ph.D. degree in applied mathematics and have had extensive experience in analyzing complex dynamical systems. During my postdoctoral training at Harvard, I have been working hand-in-hand with biologists and statisticians on several chromatin related projects. In the following I will explain in detail my accomplishments and future plan.

1. Problem

DNA in eukaryotic cells is packaged into chromatin. Chromatin plays important roles in the regulation of gene transcription. It can dynamically control the overall accessibility of the genome sequences, signal to enhance or inhibit the recruitment of specific regulatory proteins, and facilitate epigenetic inheritance by preserving its structure during cell divisions. However, due to the insufficiency of high-throughput data on chromatin, its important roles have been ignored in current computational studies.

With the rapid development of experimental technologies, especially the microarray technology, it is now a great time to investigate genome-wide chromatin structure and functions and to incorporate this knowledge to enhance our understanding of the gene regulatory mechanisms at a systems level. There are a number of fundamental questions that need to be addressed. What is the global structure of chromatin? What is the global regulatory role of chromatin? How is the chromatin structure regulated to adapt to changing environments? What are the mechanisms to achieve dynamical control? How is chromatin information inherited to the offspring? In multi-cellular organisms, how is cell differentiation controlled during development? What is the relation between of proper regulation of chromatin and clinical diseases such as cancer?

I will take three levels of approaches to addressing some of the above questions. First, I will develop data analysis tools to help biologists retrieve and process high-throughput data efficiently. Second, I will investigate the regulatory mechanisms for gene expression and epigenetic inheritance by integrating multiple experimental information, including sequence information, gene expression, protein-DNA, and

protein-protein interactions. Third, I will develop dynamical systems approaches to investigate the dynamic and epigenetic control of gene expressions. The key to this success is to establish computational-experimental partnership. Having already been closely working with bench biologists for nearly two years, I am in a good position to continue existing and forming new collaborations with bench biologists in the future.

2. Developing Tiling Array Data Analysis Tools for Probing Global Chromatin Structure

The basic repeating unit of chromatin is the nucleosome. It consists of 146 base pairs of DNA wrapped 1.7 times around an octamer of histone proteins (two of each of the histones H2A, H2B, H3, and H4). The N-terminal tails of each of the four core histones are highly conserved in their sequence. There are at least two classes of complexes that can modify the nucleosome structure: one class remodels the nucleosome positions in an ATP-dependent manner and another covalently modifies various amino acid residues on the histone tails.

Traditional methods for identifying nucleosome positions are limited either by their scope or mapping resolution. We developed a novel tiling array approach (in collaboration with Oliver Rando, Harvard University, CGR) to identifying a genome-scale, high resolution mapping of nucleosome positions in yeast (Fig. 1, reproduced from Yuan *et al.*, **Science**, 2005, *vol.* 309, *p.*626-630). More recently, this approach has also been applied to obtain high resolution histone acetylation and methylation patterns in yeast. Such information has offered new biological insights into understanding the regulatory role of nucleosome positioning and histone modifications.

Currently the tiling array technology is still in its infancy. I am among the first to develop methods for analyzing tiling array data. In particular, I developed a hidden Markov model (HMM) for objective detection of nucleosome/linker boundaries (Fig. 1D-G). HMMs use observable data to infer hidden states responsible for generating the signal. Here, observable signals are hybridization values at the tiled probes (Fig. 1C, D), and the hidden states are nucleosome and linker DNA (Fig. 1E). A major challenge for data analysis is to distinguish nucleosome signals from random peaks due to experimental noise. This challenge cannot be easily resolved using competing methods such as the Wilcoxon test. However, our HMM method easily facilitates the separation of the real and random signals by incorporating topological constraints that are consistent with biological knowledge (Fig. 1E). The HMM method yields both deterministic and probabilistic predictions of nucleosome positions: the deterministic prediction estimates the most likely configuration of nucleosome positions; whereas the probabilistic prediction is an estimate of the posterior probability of nucleosome occupy at each location given observed data. Using this method, I analyzed our nucleosome array data and identified the positions of more than two thousand nucleosomes over half a megabase in yeast genome. Surprisingly, I found that most nucleosomes were well-positioned (meaning there was little cell-to-cell variation), even

though nucleosome binding seems to have little sequence specificity. Such information certainly cannot be achieved using low resolution methods.

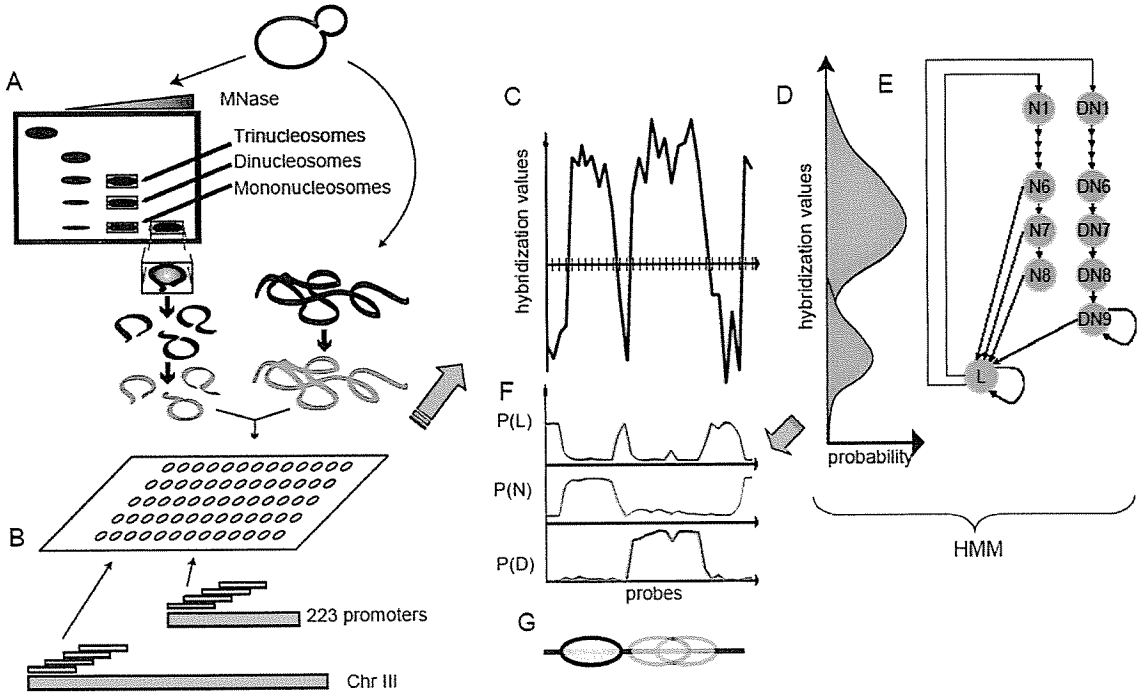

Figure 1

Future work will obtain a genome-wide mapping of nucleosome positions in yeast and also investigate how nucleosome positioning is regulated under different environments using tiling arrays. I am also interested in investigating the chromatin structure in higher eukaryotic cells and understand the similarities and differences among different organisms. An undergoing project (in collaboration with Bob Kingston's group in MGH) is to identify nucleosome positions in human cells and test whether different classes of chromatin remodeling complexes cause distinct chromatin structural patterns.

The microarray technology is advancing rapidly. Soon many labs will be able to probe interactions between DNA and nucleosomes and other proteins at high resolution using tiling arrays. On the other hand, currently there are very few analytical methods for data analysis. By developing the HMM method for analyzing nucleosome array data, I have made a first step towards developing an efficient, flexible, and user-friendly data analysis software for tiling arrays. The immediate goal is to improve the current HMM method in the following four directions: data trends removal, allowing for general noise distributions to improve robustness, maximum likelihood search, and user-friendly software design.

3. Developing an Integrated Approach to Investigating the Regulatory Role of Chromatin

Experimental studies have implicated chromatin in important but incompletely understood roles in gene regulation. Plenty of questions still remain. Does nucleosome occupancy play an instructive or permissive role in gene regulation? Histone tails are subject to several types of covalent modifications, including acetylation, methylation, phosphorylation, and ubiquitination, and each type of modification can occur at different positions. Do specific combinations of histone modification pattern form distinct "codes" that are recognized by specific molecules, as proposed by Allis, or do different patterns have redundant functions? Do histone modification patterns regulate gene expression levels as a binary switch or in a graded fashion?

To gain insight into the regulatory role of nucleosome positioning, I conducted bioinformatics data mining, searching public databases for correlations with the nucleosome positioning information obtained from our own experiments (see Section 2). I discovered that transcriptional factors binding sites tend to be depleted of nucleosomes at the genomic scale. In contrast, other regions containing similar sequence information do not have this tendency. This implies that nucleosome positioning is important for regulating transcription factor binding. I intend to follow up this study by investigating how nucleosome positions are changed under different conditions and correlate such information with changes of transcription factor binding patterns known in the literature.

I am currently investigating the regulatory code of histone acetylation in *Saccharomyces cerevisiae* by bioinformatics methods integrating genome-wide acetylation, nucleosome occupancy, gene expression, and sequence information (Yuan *et al.,* in preparation). I built linear and partially linear models using step-wise regression and regularized sliced inverse regression methods to relate gene expression levels with histone acetylation data. Compared to conventional methods, our model is more realistic in the sense that the confounding effect of promoter sequence is modeled explicitly using linear regression models. Due to the large number of candidate motifs, only the significant ones were selected by regularized sliced inverse regression. Preliminary results suggested that multiple acetylation positions have cumulative effects in regulating gene expression and that acetylation at different histones play different roles in gene regulation. These data call a strict "histone code" model into question, and suggest instead different acetylation patterns play functionally redundant roles.

In future work I will further investigate the regulatory role of nucleosomes using genome-wide data from yeast and human cells. Current methods on the computational prediction of transcription factor binding sites are primarily based on sequence information only. I will apply Bayesian data analysis methods to incorporate the nucleosome positioning and modification information to better predict transcription factor binding sites. A long term goal is to develop analytical methods to identify the structure and functions of gene regulatory networks.

4. Using Dynamical Systems Approach to Investigate the Dynamics and Epigenetic Inheritance of Gene Expression Patterns

Gene expression patterns in living cells are constantly fluctuating in response to changing intra- and extra-cellular environments. Computational and experimental investigations of dynamical mechanisms have offered new insights into understanding biological principles such as robustness, feedback, and adaptation that would not be appreciated by examining the equilibrated patterns alone. The role of chromatin in dynamically controlling gene expression is still poorly understood. I am especially interested in investigating the following problems. Does the chromatin structure encode on or off information for gene transcription? What are the mechanisms allowing the chromatin structure, including nucleosome positioning and specific modifications, to be faithfully inherited through cell divisions? What are the mechanisms controlling the reversible change between different epigenetic states? How are long term cellular memories maintained across generations?

There are at least three major analytical challenges associated with the systematic studies of dynamic control mechanisms. First, biological pathways are intricately wired. However, it seems to be common that many links are redundant and not essential to the function of a pathway, whereas a few links are critical. Can we develop analytical methods to guide experimentalists' searching for these critical links? Second, in many cases, different organisms use different mechanisms to achieve similar functions. Can we build simple computational models to identify the common principles that are independent of the detailed mechanisms? Third, even simple pathways in model organisms may display complex dynamical properties such as multi-stability, oscillation, stochastic switch, robustness, etc. Can we develop analytical methods to systematically investigate these properties?

I have had extensive and productive research experience in dynamical systems (see my publication list). The analytical techniques I used in studying various chaotic dynamical systems, including stability analysis, Lyapunov exponents, stable and unstable manifolds, and bifurcation theory, are suitable for attacking the analytical challenges mentioned above. For example, treating each component in a complex pathway as a state variable, the critical links may be identified by searching for the basins of attractors. A long term goal is to investigate how chromatin and transcription factors cooperatively achieve proper dynamic control of gene expression patterns. Some of the specific questions are mentioned in the above. Various aspects of these questions can be analyzed using mathematical models with ordinary or partial differential equations followed by dynamical systems analyses. Such a deterministic approach may also be complemented by statistical methods such as stochastic simulation in order to better estimate the noise effects. Finally, it is clear that analytical studies cannot be isolated from experimental inputs and validations. A key element of my future work will be continuing and expanding close collaboration with experimental biologists on systems biology problems.

Teaching Interests

Guo-Cheng Yuan

Learning is a life-long journey for me and probably for many other people. Thanks to a number of great teachers I have met in my life, I have found learning a very enjoyable activity. I am passionate about teaching not because I am eager to distribute my own knowledge but because I think it is a truly respectable duty to inspire young students to enjoy learning.

In my mind, the most important mission of a teacher is to let the students discover that learning is fun. I have seen many people, who did not like learning when they were young, have become enthusiastic when they grow up. I think grown-ups usually have a specific goal in mind and enjoy learning because it provides useful knowledge or skills. However, the enjoyment of learning does not have to be adults' privilege, and the most important question a teacher faces is how to convey the beauty or utility of a particular subject to the young people. I have taught mathematics courses at several levels. Mathematics can be enjoyed as an exciting intellectual game or a key to solving difficult problems. I would like to encourage students to do sufficient exercises on their own because I think exercises are not only necessary for mastering a new concept or technique but also important for developing a passion toward mathematics. There is nothing more intellectually satisfying than solving a challenging problem on your own.

How do we teach students to enjoy learning? I think that students enjoy learning skills that can be applied later in their life rather than simple facts that must be memorized in their brain. Based on my own teaching experience, I have found it helpful to relate a subject to real world applications. In 2002, I taught an ocean dynamics course together with Prof. Chris Jones at the Brown University. We encouraged the students to select a real world problem that would interest them, apply the theory learned in the lectures to analyze the problem, and then present the results in a poster format as the final project. The students were enthusiastic about working on their projects and some came up with very creative ideas. One student who loved surfing searched as many resources as he could possibly access to learn how ocean data were gathered and assimilated and summarized all in a beautifully designed flow chart. Another gathered satellite images from the internet and made a brave and creative attempt to predict the trajectories of warm rings that were spun off the Gulf Stream.

It is often also important to teach basic facts as well, but that does not imply that we cannot explain why these facts are necessary. Even if a subject that may sound dull in the beginning, the students will be inspired once they discover its hidden power. I taught MATLAB as a computational tool in biology at Marine Biology Laboratory in the last summer. Although most of the students had little programming experience, they were impressed by the power that MATLAB could offer. Some students wrote codes to analyze their own biological data and were amazed by how much easier it was to use

MATLAB instead of other software. One student even learned to do computational modeling using MATLAB in his summer research project. Everybody was happy to sit in the computer lab until midnights.

When I teach, I would like to think from students' perspective. As a group, students tend to have enormous uncertainty and anxiety. They need guidance and assurance to stay in the right direction. I think it is important to encourage students even if they make mistakes. In fact, if we were students and had to study the subject we now teach, we would probably also make similar mistakes as our students'. The goal of teaching is not to prove that we, as teachers, are smarter than the students. On the contrary, it is our duty to let the students overcome the challenge of mastering the unknown. I think it is important to point out that they are on the right track if their answers are partially correct. I can imagine that, given time and experience, the students will eventually know the subject as well as the teachers. Even if their answers are totally wrong, it is important to offer constructive criticisms rather than chides, although I do not believe that it is fair to complement every student invariably and extravagantly.

In summary, as a teacher, I will inspire students to enjoy learning, to know why learning is useful, and to be confident about learning. I will do my best to be an inspiring teacher, as the many great teachers who have inspired me to pursue an academic career.