WEIQUN PENG

Department of Physics and
The Center for Theoretical Biological Physics
University of California, San Diego
9500 Gilman Drive, MC 0374
La Jolla, California 92093-0374

January 2, 2004

Faculty Search Committee
c/o Prof. Rob de Ruyter van Steveninck
Biocomplexity Institute
Indiana University
Swain Hall West 117
Bloomington, IN 47405-7105

Dear Prof. Rob de Ruyter,

I wish to apply for the tenure track faculty position in theoretical biocomplexity/biophysics in the Biocomplexity Institute, Department of Physics, as advertised in the website for the career center of the American Physical Society. Enclosed are a copy of my curriculum vitae, a list of references and the statement of research interests.

My researches have always been of interdisciplinary nature. As a Ph. D student, in addition to works in soft condensed matter and statistical physics, I engaged in collaboration with an experimental biophysics group, developing a model that quantifies the idea of probing protein dynamics using fluorescence resonance energy transfer with donors of different lifetimes. Currently, my research focuses on applying theoretical and computational approaches to study systems biology from physical, molecular and evolutionary perspectives. As a faculty member in your institute, I will establish a strong interdisciplinary research program on qualitative understanding and quantitative modeling of gene regulation and evolution, which meshes very well with two of the central themes of the institute: modeling biological networks and bioinformatics. Cutting across the traditional boundaries of statistical physics, molecular biology, theoretical biology and bioinformatics, my research program will foster collaborations among local institutions in those research areas.

I am committed to pursuing a research and teaching career in academia, and I am very excited about the prospect of working at your institute. With my broad research interests and theoretical and computational expertise, I will add new dimensions to the existing research activities in your institute. I look forward to hearing from you regarding my application. If you require additional materials or wish to discuss my application, please email me at w-peng@physics.ucsd.edu or call me at (858) 534-7256.

Sincerely,

Weiqun Peng

Current and Future Research

Weiqun Peng

The past decade has witnessed explosive advances in molecular biology and genomics. A great challenge in this new era is to link the in-depth knowledge of molecular mechanisms to understanding of system characteristics and functions, and physics could play a major role in establishing the underlying principles. My research focuses on using statistical physics to study complex behaviors that arise from interplay of constitutive components of biological systems. More specifically, my research develops and uses theoretical and computational tools to explore the two intimately related aspects of biocomplexity, gene regulation – the main control system of a living cell and evolution – the driving force behind the emergence of biocomplexity.

Gene Regulation

Transcription regulation plays a dominant role in the spatial and temporal control of the expression of various genes in a biological cell. The control information is encoded in the DNA sequences upstream of genes, referred to as cis-regulatory elements. Genes are turned on or off when the regulatory elements are bound to transcription factor proteins, which in turn interact with the transcription machinery and affect gene expression. Cracking the "cis-regulatory code" is a major challenge in the post-genome biology. My researches in this area develop novel tools by combining methods of statistical physics, evolutionary biology and bioinformatics, and apply those tools to understand the physical, functional and evolutionary aspects of gene regulation.

Physical and functional study of gene regulation: In eukaryotes, the regulatory DNA sequences which transcription factors bind to are often short (5-8 bases) and fuzzy. It remains a big puzzle how transcription factors find and bind to target sites given a genome background full of spurious sites, enabling gene regulations in a specific and reliable manner. Studies of exemplary systems suggest that the regulatory information is organized into modules, each consisting of multiple binding sites, which allow for multiple transcription factors to act cooperatively, therefore enhancing specificity. The critical issues are the organizational principle for the modules and the resulting functional capacity.

To address this issue, I intend to explore the idea that the organization of the target binding sites is associated with the higher-order nucleosomal structure: They are in close juxtaposition to be within the same nucleosome, with the cooperativity arising from competition between histone octomer and transcription factor proteins to bind the same stretch of sequence[1]. The merit of this mechanism is that the cooperativity is insensitive to both the spacing between target sites and the intricate interactions between transcription factors. This generic mechanism is particularly appealing in light of the recent bioinformatic works on drosophila[2], where it is found that, simply by looking for genomic regions of unusually high concentrations of predicted binding sites, a significant fraction of known regulatory modules is recovered.

Considering a target module embedded in the genomic background consisted of a rugged binding-energy landscape full of spurious binding sites, I will study the functional capacity of this mechanism via a quantitative model under the statistical mechanical framework for DNA-protein binding[3]. In particular, I plan to investigate the following physical and functional issues: (a) The condition (i.e., the number of sites needed in the nucleosome, the fuzziness allowed for each site) under which the target sites to be bound with much higher probability than the rest of the genome;

(b) The dynamic time scale for the transcription factors to find their target sites, given the rugged binding energy landscape the genome confers; (c) The degree of gene expression changes achievable by simply changing the sequence of the target sites. Furthermore, I plan to make use of the sequences of known regulatory modules to look for signals of such a mechanism as work, such as whether regions of regulatory modules are also energetically preferred positions of nucleosomes along the DNA.

Finding cis-regulatory elements via comparative genomics: The completion of genomes of related organisms has made comparative genomic analysis a powerful tool for understanding both the function and the evolution of a genome. My research employs this approach to find regulatory elements and to study the underlying principles of gene regulation at the system level, with special focus on integrating comparative genomic analysis with evolutionary study, which includes the two complementary aspects: (a) Incorporating the appropriate evolutionary model for sequence alignment, the key ingredient of any comparative genomic analysis; (b) Identifying regions that are subject to additional evolutionary forces or constraints, in the aligned sequences.

In an ongoing project, I have investigated the revelation of human cis-regulatory elements, by pair-wise and multiple comparison of genomic sequences of human, mouse and rat.
The novel feature of my approach is the incorporation of the most accurate model of the background evolutionary processes of human, mouse and rat[4], which exhibit non-equilibrium, irreversibility and a prominent neighbor-dependent mutational process (i.e., the dinucleotide methylation-assisted transition: CG → CA/TG). The new feature enables the algorithm to better distinguish the sequence homology due to functional constraints from that due to the common evolutionary ancestry (which is already about 70%). Preliminary dataset analysis shows that our tool set is promising.

Based on previous achievement, I will explore the following directions in the future: (a) Collaborating with molecular biologists working on transcription regulations in human, mouse or rat. By applying our tool sets to specific gene loci under experimental investigation, I will provide a list of potential regulatory elements assessed with significance, which can be tested via standard experimental techniques. (b) Improving the sensitivity of the algorithm by incorporating a more realistic evolutionary model for the functional regulatory elements[5], and developing algorithms that detects regulatory elements consisting of multiple transcription factor binding sites; (c) Integrating results of this approach with the DNA microarray gene expression data and the transcription factor binding loci data from ChIP experiments; (d) Using this tool set to investigate the pattern of conservation and divergence in gene regulation across the three species.

Evolution

Evolution is the guiding principle for the entire biological world. Understanding evolution is crucial for every aspect of biology, including gene regulation. Mathematical modeling has a long tradition in the study of evolution. In contrast to traditional theoretical works studying models without any physical or biological underpinning, my research concentrates on investigation of evolutionary models motivated by laboratory evolutionary experiments, including directed *in vitro* evolution of biomolecules[6] as well as long term laboratory evolutionary experiments with microorganisms[7]. Having seen extraordinary advances in recent year, these experimental approaches have well controlled setup and the ability to track the evolutionary dynamics via sequencing and bioinformatic approaches. They provide perfect systems for quantitative study and testing of theoretical predictions.

<u>Characterization of evolutionary dynamics:</u> Evolution is a dynamical process involving mutation, recombination, selection and reproduction operations with a finite number of individuals. My research applies theoretical methods of non-equilibrium statistical physics to investigate explicit evolutionary models.

I have already worked on a couple of issues arisen from *in vitro* directed molecular evolution. In one project[8] I investigated the evolutionary of DNA sequences selected via competitive binding to a transcription factor protein. This is a molecular system where the genotype (i.e., the sequence), phenotype (i.e., the property of the molecule, which in this case is the binding affinity of the DNA sequence), and selection are linked simply but firmly by well-characterized thermodynamics. Based on thermodynamic and dynamical considerations, I described the competitive evolutionary process by a Schröedinger-like equation driven by a self-consistent, time-dependent potential. Treating the evolutionary process as a dynamical system, I borrowed ideas from the front propagation studies in non-equilibrium systems. I found that the equation admits a multitude of decelerating pulses, the correct one selected by a marginal stability criterion. The pulse description agrees well with simulation results in the range of parameters under study. In a further work, I studied DNA shuffling from an evolutionary perspective[9]. DNA shuffling is an evolutionary protocol that uses recombination in addition to mutation to generate diversity in the sequence. Practices have shown that recombination greatly facilitates the progress of the *in vitro* evolution. However, a quantitative and predictive description was lacking. Theoretical study of recombination is difficult because it is a highly non-local operation in the genomic space. I proposed a simplified model that includes all the essential ingredients, which provides not only an adequate description of the dynamics of DNA shuffling but also an explicit understanding of the evolutionary benefit of recombination. Incidentally, this is one of the rare multi-locus evolutionary models that are solvable. It can serve a role in evolutionary modeling of recombination similar to the random energy model in disordered systems.

My previous works focused on mean-field studies. I intend to explore the fluctuation effect due to the finite population size in the future. Due to the special stochastic reproductive nature of the evolutionary dynamical system, fluctuation effects can accumulate, resulting in qualitative difference in the evolutionary dynamics from the mean-field results. Indeed, the fluctuation effect makes a dramatic difference even in laboratory evolution of E. Coli[7] with population size reaching 10^7. I will focus on the finite-population evolution of a phenotype that depends on multiple genes, an issue deeply related to the propagation of a noise-dominated front in non-equilibrium statistical physics. A general framework to quantitatively characterize such an evolutionary process has been lacking. I plan to develop such a framework by generalizing a modified heuristic cut-off approach[10], which is found to agree surprisingly well with stochastic simulation results in models I have tried. The project will include

- Developing theoretical underpinning of the heuristic approach, by joining a stochastic description of the noisy propagating front with mean-field description in the bulk of the population distribution[11].
- Applying the cut-off approach to evolutionary dynamics of mutator systems. Mutator alleles, with mutation rate 100 fold higher than that of the wild type, have been found to emerge and take over the entire population during adaptation. The study will help understand the recent evolutionary experiments on mutator alleles in asexual E. Coli populations[12], which found dramatic difference of evolutionary dynamics from theoretical predictions based on mean-field treatment.
- Applying the cut-off approach to sexual populations. Theoretical analysis shows that range of parameters where recombination is beneficial is limited for a sexual population of infinite size[13]. Recent simulation studies[14] suggest that a finite

population could enhance the benefit of recombination. To examine this issue, I intend to integrate our heuristic approach with the model of recombination developed for DNA shuffling; aiming to shed some light on one of the big puzzles for evolutionary biology, i.e., the prevalence of sex and how the advantage of sex can offset its two-fold cost (due to half of the population being non-reproductive).

Evolution of simple genetic circuits: The general principles of transcriptional regulation have traditionally been studied via dissection of exemplary promoters and cross-species comparison. I plan to explore these principles from the novel evolutionary and engineering perspective, via *in silico* evolutions of regulatory sequences that implement desired regulatory functions (e.g., gene expression at a required level, bi-stable switch and oscillator). My attempts will be directed towards a quantitative understanding the evolutionary process of a gene acquiring new function and genes obtaining coordination. Insights gained in this study will also be helpful to bioengineering applications, such as making designer gene circuits for drug delivery. In the *in silico* evolution, I will adopt the molecular model of transcription regulation based on DNA-protein binding and combinatorial control[15]. Sequences are selected by the level of gene expression. The gene expression level is set to be proportional to the binding affinity of the RNA polymerase to the promoter, which is determined by the regulatory sequence and the transcription-factor concentrations. I will concentrate on the following aspects:

- Establishing principles for finding a selection scheme that would efficiently guide the evolution towards a solution.
- Probing the evolutionary robustness of various implementations of a function. I will find out the evolutionary choice given a function that has multiple implementations. I will also investigate the dependence of the choice on the selection scheme, mutation rate and population size. In addition, I will explore the possibility of a polymorphic population that maintains the diversity of implementations.

[1] J. A. Miller and J. Widom, Mol. Cell. Biol. **23**, 1623 (2003).
[2] B. P. Berman, Y. Nibu, B. D. Pfeiffer, P. Tomancak, S. E. Celniker, M. Levine, G. M. Rubin and M. B. Eisen, Proc. Natl. Acad. Sci. **99**, 757 (2002).
[3] U. Gerland, J. D. Moroz and T. Hwa, Proc. Natl. Acad. Sci. **99**, 12015 (2002).
[4] P. F. Arndt, D. A. Petrov, and T. Hwa, Mol. Biol. Evol. **20**, 1887 (2003).
[5] A. M. Moses, D. Y. Chiang, M. Kellis, E. S. Lander and M. B. Eisen, BMC Evolutionary Biology **3**, 19 (2003).
[6] E. T. Farinas, T. Bulter and F. H. Arnold, Curr. Opin. Biotechnol. **12**, 545 (2001).
[7] S. F. Elena and R. E. Lenski, Nat. Rev. Genet. **4,** 457 (2003).
[8] W. Peng, U. Gerland, T. Hwa and H. Levine, Phys. Rev. Lett. **90**, 088103 (2003).
[9] W. Peng, H. Levine, T. Hwa and D. A. Kessler, to appear in Phys. Rev. E. (2003).
[10] E. Brunet and B. Derrida, Phys. Rev. E. **56**, 2597 (2003).
[11] I. M. Rouzine, J. Wakeley and J.M. Coffin, Proc. Nat. Acad. Sci. **100**, 587 (2003).
[12] A. C. Shaver, P. G. Dombrowski, J. Y. Sweeney, T. Treis, R. M. Zappala and P. D. Sniegowski, Genetics **162,** 557 (2002).
[13] S. P. Otto and T. Lenormand, Nat. Rev. Genet. **3**, 252 (2002).
[14] S. P. Otto and N. H. Barton, Evolution Int. J. Org. Evolution. **55**,1921(2001)
[15] N. E. Buchler, U. Gerland and T. Hwa, Proc. Natl. Acad. Sci. **100**, 5136 (2003).