Dr. Marcos R. Betancourt

November 18, 2003

Biocomplexity 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. de Ruyter,

This letter is to express my interest in being considered for the Junior Faculty position at the Biocomplexity Institute.

I'm currently a Research Assistant Professor at the State University of New York Buffalo Center of Excellence in Bioinformatics working in computational protein modeling and bioinformatics problems. I am interested in obtaining a tenure track faculty position with research emphasis in molecular biophysics and that provides me with the opportunity for teaching, hiring and mentoring graduate students and postdocs, collaborating with experimentalists, and building a research group. For these reasons, I found that the IUB is an excellent place for pursuing my career and that I can in turn strongly contribute to the mission of the Biocomplexity Institute.

I did my Ph.D. in theoretical physics at the University of California, San Diego. While my initial interest and research efforts where in the area of fluid dynamics and chaos, I did my dissertation in the subject of protein folding kinetics. As an NSF posdoc, I contributed to the subjects of chaperonin mediated folding and *de novo* protein design, and in a second postdoc to the problem of protein structure prediction. Currently, I am developing and studying reduced protein models.

Some funding related activities that am I involved in or that I have available are:

- Recently submitted a research proposal to both NSF and NIH for studying protein folding kinetics and reduced model (results are still pending).
- Qualify for up to $50,000 in start up funds from my previous NSF postdoctoral fellowship if employed in a tenure track position.

Please find enclosed my curriculum vitae, my research plan, teaching statement, and references. If any other information is needed, please do not hesitate to contact me. Thank you for your consideration. I look forward to hearing from you in due course.

Sincerely yours,

Marcos R. Betancourt, Ph.D.
Research Assistant Professor
UB Center of Excellence in Bioinformatics
901 Washington Street, Suite 300
Buffalo, NY 14203
Phone; (716) 849-6712; E-mail: mb63@buffalo.edu

Research Interests

1 Introduction

My research has been focused in the area of the computational modeling of proteins and protein folding. Understanding and predicting the way proteins fold is a problem of great interest across many disciplines in science including biology, chemistry, and physics, with important applications to medicine and biotechnology. My approach has been the analysis of the problem through the use of simplified protein models and computer simulations. During my Ph.D. work, I studied the kinetics of protein folding using idealized proteins restricted to lattices. There, I studied the energetic and geometric factors affecting the folding times of these models, and studied ways of analytically solving the first passage time problem in simple cases (Betancourt, UCSD, 1995). For my first postdoctoral work, I carried out one of the first computational studies of folding assisted by molecular chaperonins (Betancourt & Thirumalai, J. Mol. Biol., 1999). In addition, I developed a theory for *de novo* protein design applied to lattice proteins (Betancourt & Thirumalai, J. Phys. Chem. B, 2002). After these studies and the insights I gained from idealized models, I moved into more realistic models, capable of reproducing real protein structures. One of my first projects in this area was to develop clustering algorithms to help analyze folding trajectories generated by state-of-the-art folding simulations (Betancourt & Skolnick, J. Comp. Chem., 2001), which was used in the Critical Assessment of Techniques for Protein Structure Prediction (Skolnick, *et al.*, Proteins, 2001).

My work as a research scientist has been focused towards two main goals: develop a simplified protein model simple enough to simulate folding events, but sufficiently accurate to reproduce the native structure without the use of homology modeling; and to develop better methods for native structure information extraction from threading for native structure prediction applications. These goals and some other results that I have obtained are explained in more detail in the Current Research section below. Developing a simple but realistic model is critical for my future research objectives, one of which is the study of protein kinetics and the kinetics of folding. To my knowledge, simulation of protein kinetics with quantitative accuracy is restricted to detailed atomic models. These models are limited by the enormous computational time required to fold even the smallest and fastest folding proteins. One goal is to adapt the simple model and the simulation techniques that I'm developing to achieve quantitative accurate folding simulations. Another goal is to extend the theory I developed for protein design of ideal models to realistic protein models. This theory allows designing protein sequences that fold to a unique native structure in a minimum time. Therefore, it requires the availability of a protein model that can simulate the native structure and folding kinetics to some degree of accuracy. More details on these objectives are described in the Future Research section below.

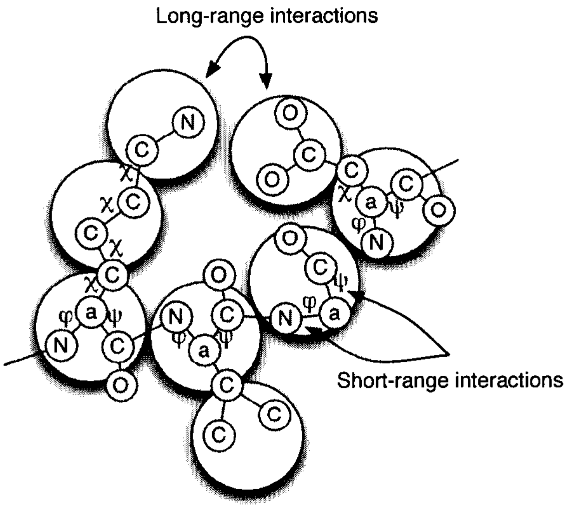

Figure 1: Geometry for the pseudo-atomic model. The relevant dihedral angles for the backbone and side-chain are shown. In the case of proline (not shown), ω is used instead of ϕ. The big spheres represent the pseudo atoms. Up to three pseudo-atoms per residue are defined.

2 Current Research

2.1 Reduced Protein Models

Most protein models that describe the molecular interactions in enough detail to investigate the physical behavior in a tractable manner are too complicated to study the folding of typical proteins because of the computational time required to carry out this simulations (Duan & Kollman, Science, 1998; Snow, et al., Nature, 2002). Techniques and models used today for native fold prediction rely heavily on homology modeling, even for simple monomeric proteins (Strauss, et al., Proteins, 2003; Skolnick, et al., Proteins, 2003). The protein representation that I have been developing aims to overcome these obstacles. In general, it uses a coarse grained representation of the atoms in residues, an implicit solvent, and interaction energies based uniquely on knowledge based potentials (KBP).

The protein model geometry consists of a hybrid combination of a detailed model for local energy calculations with a coarser model for non-local energy calculations. A sample protein fragment showing the model geometry is illustrated in Fig. 1. The model keeps track of the backbone dihedral angles (ω, ϕ, ψ) and most of the side chain dihedral angles (χ_i). The dihedral angles, with the exception of the peptide plane dihedral angles, are free to rotate, while bond distances and bond angles[1] are fixed. Potentials depending on the dihedral angles affect the local geometry of the backbone and side chains. The model also contains a coarser description, in which groups of atoms are combined into pseudo-atoms with effective non-bonded interactions. For each residue, 1 to 3 pseudo-atoms are assigned depending on the side chain size. One of the pseudo-atoms is assigned to the backbone,

[1] Exceptions are the C_α bond angles that are allowed to fluctuate, as will be explained shortly.

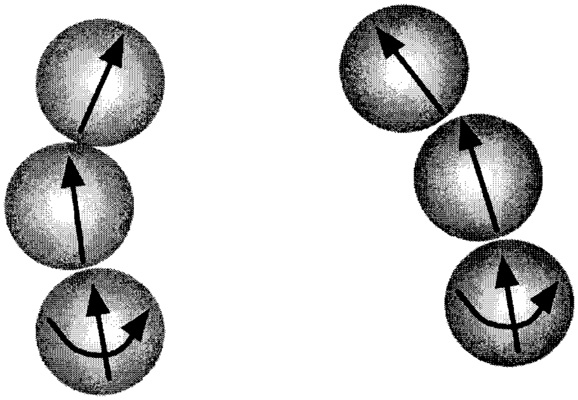

Figure 2: Pseudo atom orientations.

which is placed at the average position for the backbone atoms, including the oxygen and the beta carbon. Only the side chain dihedral angles needed to reconstruct the pseudo-atoms are considered.

The backbone potential is derived from the observed population of dihedral angles in proteins. A set of high-resolution non-homologous proteins is used for obtaining the statistics. The potentials are derived for each of the 20 amino acids and depend on the type and conformation of the adjacent residues. That is, the potentials are a function of residue triplets and their dihedral angles. Because of the limited number of proteins and the large number of possible triplets and conformations, the potentials are divided in independent terms depending on two dihedral angles at a time for all 8,000 amino acids combinations. Each term (seven in total) depends on neighbors and nearest neighbors dihedral angle combinations, excluding the dipeptide bond angle, ω, which remains approximately fixed. An exception is made for proline, in which ω takes the place of proline's ϕ angle. This potential has been tested and showed to improve the selectivity of the dihedral angles by more than 20% (Betancourt & Skolnick, in preparation, 2003). Many amino acid triplets show significant deviation in their Ramachandran density plot from the traditional plots that depend on the identity of a single residue. In addition, the potential terms involving angles between adjacent amino acids show cooperative behavior, which for example enhance the probability of one residue to be in the extended or coiled state if its neighbors are in the extended or coiled state, respectively.

While the backbone potential helps to determine the local geometry and to reduce the number of conformations to physiological ones, a non-local pairwise potential is responsible for guiding the protein to the native topology. I derived a pairwise KBP for the pseudo atoms shown in Fig. 1. In the original model (Betancourt, Proteins, 2003), the potential depended on the pseudo atom types, distance, and separation along the main chain. It was shown that this model is more accurate than residue based KBPs and as accurate as atomic detailed KBPs in identifying protein native structures by threading, including gapped decoy sets and the "Decoys R Us" database (Samudrala & Levitt, Protein Science, 2000). In the latest version of the model, relative orientations between the pseudo atoms have been added to the potential (unpublished results). The orientations are described in Fig. 2.

Because of the limited data, only those orientations believed to be most significant are included. The pair-wise potential depends on the relative direction between the pseudo atom direction, the latter being measured from the previous pseudo atom. In addition, a dependence of the backbone ψ angle is included between the interactions of the backbone pseudo atoms. This accounts in part for the orientation of the backbone oxygen and possibly for its hydrogen bond effects. Preliminary tests have shown improvements in energy ranking, in particular of beta dominated native structures. In addition to the pair-wise potential, a side-chain torsion potential is determined by statistical methods.

The pair-wise pseudo atomic potential, side chain torsion angle potential, and backbone potentials are combined for simulating the "pseudo" dynamics of proteins. This is achieved by a new and efficient Monte Carlo (MC) method that I developed (Betancourt, in preparation, 2003). The MC method consists primarily in moves that change only an arbitrarily long segment of the protein, without affecting bond distance, bond angle, and the peptide dihedral angle constraints, with the exception of the C-α bond angles, which are varied over a natural range. Each move changes a minimal number of torsion angles. MC methods like this have appeared before, but the simplicity in the amount of the algebra required to perform a move makes my method much more efficient than others. Final evaluation of the model and the method is being carried out.

2.2 Threading and Structure Prediction

The prediction of the native structures for medium and large proteins ($>$ 150 residues) by folding simulations is still a challenging problem even for reduced models. A faster determination of the native fold can be achieved by extracting information from known native structures by threading. I have developed an algorithm for this purpose. The information is obtained in the form of the average and dispersion of the inter-residue distances, which can be used as biases in folding simulations. The method uses a gapless threading Monte Carlo approach combined with a generalized dynamic programming method that selects threaded structural clusters with optimal properties. The clusters are determined by the residue's proximity in space, which are not necessarily contiguous along the sequence. Distance averages are obtained by weighted averages of the clusters inter-residue distances. The weights are estimated from a general correlation between the Z-score and a size-independent distance root mean square deviation (DRMSD). This correlation depends on the pseudo atomic potential, which is currently being optimized. The results show that a reasonable good correlation function can reconstruct a protein topology from many structures of low sequence similarity.

3 Future Research

One of my long-term goals is to simulate the long-time scale dynamics of protein folding using reduced protein models and Monte Carlo methods. The fundamental approach consists of systematically establishing an appropriate energetic and kinetic equivalence between the reduced protein model described and other models of increased complexity and accuracy.

My other goal involves *de novo* protein design.

3.1 Statistical Potentials and Protein Kinetics Modeling

A statistical potential will be extracted from an all-atom protein model of short polypeptides. This will be used to introduce temperature and solvent effects to the reduced model potential, as well as interactions with other solutes. Comparing this statistical potential to those obtained from known protein structures will allow to make modifications to the all-atom model potentials.

A connection between Langevin (Brownian) dynamics and Monte Carlo dynamics will be established by adapting a dynamic Monte Carlo theory to proteins. The Monte Carlo dynamics will incorporate the viscosity and diffusion effects present in the Langevin dynamics.

3.2 *De Novo* Protein Design

Some time ago, I developed a theory for *de novo* protein design that given a target structure and target temperature selects a protein sequence that folds into it with optimal folding time and structural stability. This theory was developed for contact interactions and tested on idealized lattice models and produced extremely good results (Betancourt & Thirumalai, J. Phys. Chem. B, 2002). After the development of the pseudo atomic model and the kinetic modeling technique just described, I will extend the protein design theory to more realistic models. This will also require close collaboration with experimentalists to corroborate results for newly designed polypeptides and hetero-polymers with protein-like properties.

Teaching Goals

I strongly feel that teaching is an integral part of a scientist's career. Richard Feynman once wrote: *I don't believe I can really do without teaching. The reason is, I have to have something so that when I don't have any ideas and I'm not getting anywhere I can say to myself, "At least I'm living; at least I'm doing something; I am making some contribution" it's just psychological. When I was at Princeton in the 1940s I could see what happened to those great minds at the Institute for Advanced Study , who had been specially selected for their tremendous brains and were now given this opportunity to sit in this lovely house by the woods there, with no classes to teach, with no obligations whatsoever. These poor bastards could now sit and think clearly all by themselves, OK? So they don't get any ideas for a while: They have every opportunity to do something, and they are not getting any ideas. I believe that in a situation like this a kind of guilt or depression worms inside of you, and you begin to worry about not getting any ideas. And nothing happens. Still no ideas come. Nothing happens because there's not enough real activity and challenge: You're not in contact with the experimental guys. You don't have to think how to answer questions from the students. Nothing!* (R. Feynman, 1986)

While by having any ideas he probably meant having no good ideas leading to a solution, I do certainly agree with him in that being in contact with experimentalist and answering questions from students is essential for the development of a successful research career. Besides these benefits, teaching is a fulfilling experience by itself. I have always dreamed about the opportunity of teaching science to young students. As a student, I had the privilege of learning introductory physics from a teacher who taught physics as a fascinating story, one full of excitement, expectancy, and surprises. I want to be able to teach science and generate the same level of interest and fascination that I experienced in that class. One in which getting good grades comes primarily as a consequence of the student's interest in learning and not only from the student's interest in getting good grades.

To accomplish this, one must be as fascinated about the subject as one expects the student to be. One must integrate the students and involve them as much as possible during lectures. One must simplify the ideas and select supporting examples and demonstrations. Also, one must listen to the students and use their constructive criticism for making improvements to the course. My experiences as a teacher assistant during my Ph.D. and as a leader in the National Guard have given me a wide perspective on teaching techniques. Allowing the students to rationalize the subject by themselves and provide them with enough interactions and practical exercises are keys in the learning process.

My interest is to teach courses in undergraduate level physics, biophysics, and computer science. At the advanced level, I am interested in teaching specialized courses in statistical physics, molecular biophysics, bioinformatics, and numerical analysis. In addition, teaching other subjects not directly related to my research interest can also broaden my knowledge and could benefit my research in many ways.