## 1. INTRODUCTION

Early genomic sequencing efforts promised to revolutionize biology and medicine. Subsequently, large-scale post-genomic efforts have provided an even greater profusion of biological data. However, the abundance of information currently being generated is creating nearly as many problems as it provides answers. This information avalanche can only be stemmed through increasingly accurate **bioinformatic** and **biophysical** methods for storing, annotating, and analyzing the data. My broad research goals are to help resolve these challenges on two fronts:

- Development and application of novel biophysical techniques, especially as related to protein design,
- Development and application of accurate and robust post-genomic bioinformatic techniques.

Specifically, my lab is trying to decipher the subtle sequence/structure/function relationships within protein families and superfamilies. Using a combination of bioinformatic and biophysical approaches, we are trying to understand the evolutionary origins of protein families and conserved function. The general rules that we uncover are later applied to our broader research goals. For example, in my early work (with S. Subramaniam, UCSD), I observed that electrostatically and functionally important sequence regions are more likely, compared to the intervening space, to conserve the overall familial phylogeny. My lab has successfully utilized this phenomenon to **accurately predict protein functional sites from sequence using *phylogenetic motifs*.**

In addition to sequence/structure/function relationships, I am increasingly interested in stability/flexibility/function relationships. In collaboration with Don Jacobs (Cal State Northridge), we are currently developing a powerful distance constraint model (DCM) that efficiently calculates both protein flexibility and stability profiles. Using the DCM, we can harmoniously quantify both at any given temperature. While coarse-grained, the DCM possesses most of the essential physics involved in protein folding/unfolding. **In fact, the DCM is the first computational method that can quantitatively reproduce protein unfolding heat capacity curves.** Because the DCM is so fast, **it's more than $10^{10}$ times faster than molecular dynamics simulations**, we believe it will become a ubiquitous structural genomics tool. Our future work will use the DCM to assess the stability and functional efficiency (through analysis of catalytic normal modes) of designed protein mutants. Examples of recent and ongoing investigations include:

- Understanding how electrostatics can mediate familial conserved function (w/ S. Subramaniam, UCSD),
- The role of electrostatics in molecular recognition, specifically antibody-antigen (w/ S. Subramaniam, UCSD),
- Development of a novel weighted-ensemble Brownian dynamics algorithm (w/ S. Subramaniam, UCSD),
- Sequence and structure differences between mesophilic and thermophilic orthologs,
- Conferring thermostability to mesophilic proteins through optimized electrostatic surfaces,
- Development of *phylogenetic motif* approaches for predicting protein functional sites from sequence,
- Development of a web-based phylogenetic motif identification server (called MINER),
- The evolutionary and catalytic importance of familial conserved electrostatic networks,
- Development and application of a novel (coarse-grained) protein stability and flexibility biophysical model,
- Protein family Quantified Stability/Flexibility Relationships (QSFR),
- Protein stability/flexibility changes on substrate binding.

Future work will seek to maintain our research momentum in several high impact areas. Our short-term efforts will concentrate on our current research strengths: **(1)** protein electrostatics, **(2)** the Distance Constraint Model, **(3)** large-scale sequence/structure comparisons, and **(4)** phylogenetic motifs. My lab is at the cusp of a very productive period. Considerable effort has been invested to develop our techniques, but we are now at a point where we can apply our methods to a variety of systems in ways similar to my previous work. At the same time, several extensions/improvements of our current methods (listed below) are planned in the near future. In the long term, I also plan to expand into systems biology. In short, I plan to incorporate improved models of protein stability and function into whole-cell models, which should improve their accuracy. A brief description of my labs main research efforts follows; funding history can be found within my curriculum vitae.

- Extension of current DCM to include: rotamer-dependent conformational entropies, sequence dependence, hydrophobic effects and a transferable parameterization,
- Marriage between Poisson-Boltzmann continuum electrostatic theory and the DCM,
- Development of phylogenetic motif identification algorithms that: (1) do not rely on an underlying multiple sequence alignment and (2) improve phylogenetic motif tree significance.
- Development of a functional site prediction *pipeline* that uses varying levels of bioinformatic sophistication.

## 2. OPTIMIZATION OF PROTEIN ELECTROSTATIC SURFACES

Thermophiles are organisms that live at high ambient temperatures (60 to 100$^{o}$C). Enzymes of such organisms have adapted to thrive in such hostile environments, whereas proteins from organisms that live in standard conditions (mesophilic) would quickly denature. Many attempts to identify the most efficient method of conferring enhanced

thermostability to mesophilic structures are reported in the literature. Recent experimental studies have successfully increased mesophilic protein stability through mutagenesis of a single solvent exposed residue, presumably through optimization of the protein's electrostatic surface. These results confirm that surface electrostatics are intimately related to overall protein stability, and, mutation of only few surface residues is generally sufficient for conferring thermostability to mesophilic proteins.

We have developed and applied a simple theoretical model to quickly screen mutant structures for increased thermostability through optimization of the protein's electrostatic surface. Our results are able to reproduce the experimental observation that elimination of like-charge repulsions and/or creation of opposite-charge attractions on the protein surface is an efficient method of conferring thermostability. Using Poisson-Boltzmann electrostatics, we calculate relative protein stabilities for the exhaustive surface mutagenesis of several protein structures. Comparison with 25 experimentally characterized cold shock protein mutants reveals an average correlation of 0.86. Our model is also quantitatively accurate when reproducing the experimental D49A and D49H mutant stabilities of RNase T1 (see Fig. 1). **This work represents the first comprehensive *in silico* screening of mutant candidates likely to confer thermostability through optimization of electrostatic surfaces.**

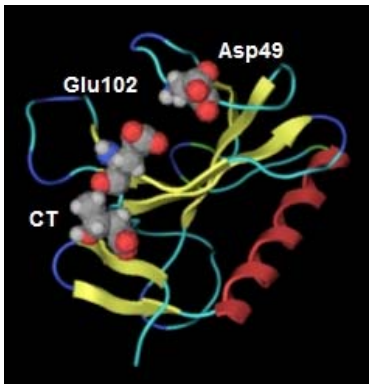


**Fig. 1** – The D49H mutation is one of the most stabilizing screened because it eliminates the unfavorable repulsions between it and an additional pair of anionic residues (Glu102 and the C-terminus).

Additionally, we have recently compared the charge-charge (++, --, and +-) correlation functions (G(r)) within 100+ orthologous mesophilic/thermophilic structural families. Our systematic structural comparisons clearly indicate that Nature also employs optimization of the electrostatic surface to increase the stability of thermophilic structures. However, to conserve functional rates, observed differences are generally isolated from active site regions.

Taken together, we expect our biophysical and empirical results will enable experimentalists to make more informed decisions when attempting to do determine mutants likely to confer thermostability to mesophilic proteins. Nick Pace and coworkers (Texas A&M University) are currently experimentally testing our predictions on RNase's (RNase T1 and several RNase Sa isoforms), and Lisa Alex (Cal Poly Pomona) is doing the same on CheY mutants. The bulk of this work was performed by six Cal Poly Pomona students (four undergraduate and two graduate students). Research manuscripts detailing the above work have recently been published in the *Biophysical Journal* and *Protein Engineering*.

## 3. PROTEIN STABILITY/FLEXIBILITY RELATIONSHIPS

We are very encouraged by the above results, but the approach is limited to mutations on the surface (where conformational entropy effects are most likely self-canceling). From a protein design point of view, we would like to be able to efficiently model the effects of mutants in the core. This task requires an accurate assessment of protein flexibility and thermodynamic stability, both done in a computationally efficient way. Together with Don Jacobs (Cal State Northridge), we are developing a sophisticated biophysical model that combines network rigidity, protein electrostatics, and informatics into improved methods for *in silico* screening of stability, flexibility, and functionality (see below). Several experimentalists, including Frank T. Robb (University of Maryland Biotechnology Institute), Maria Luisa Tasyaco (City College of New York), and Nick Pace (Texas A&M University), have expressed an interest in experimentally testing our predictions. A key research goal is to design and engineer protein mutants with increased structural stability. However, thermophilic proteins often lose function at lower temperatures, presumably through decreased flexibility within the active site region. **Therefore, the Holy Grail of our work involves broadening the temperature ranges of <u>both</u> stability and functionality.** This will be achieved through engineering stabilizing interactions, paying special attention to maintain key allosteric motions.

The core of our combined research is built around a powerful Distance Constraint Model (DCM) that is able to harmoniously model thermodynamic and flexibility properties at a given temperature. The DCM is based on the quantifiable hypothesis that network rigidity is an underlying mechanical interaction that provides enthalpy-entropy compensation, critical to stability and molecular cooperativity in native protein structures. Although the total enthalpy is additive, the entropy is not. The non-additive property of component entropies derives from not knowing which degrees of freedom in the system are independent or redundant. The DCM resolves this by using a mechanical representation of the protein structure (network rigidity) that can efficiently identify both. **This formulation yields a precise mathematical description of how to calculate total enthalpies and entropies of a protein from a table of known component enthalpies and entropies associated with individual residues** (see Fig. 2). A further consequence is

that flexibility and rigidity are precisely calculated by rules governing the mechanical constraint network, which provides key insights to the correlated (allosteric) intramolecular motions.

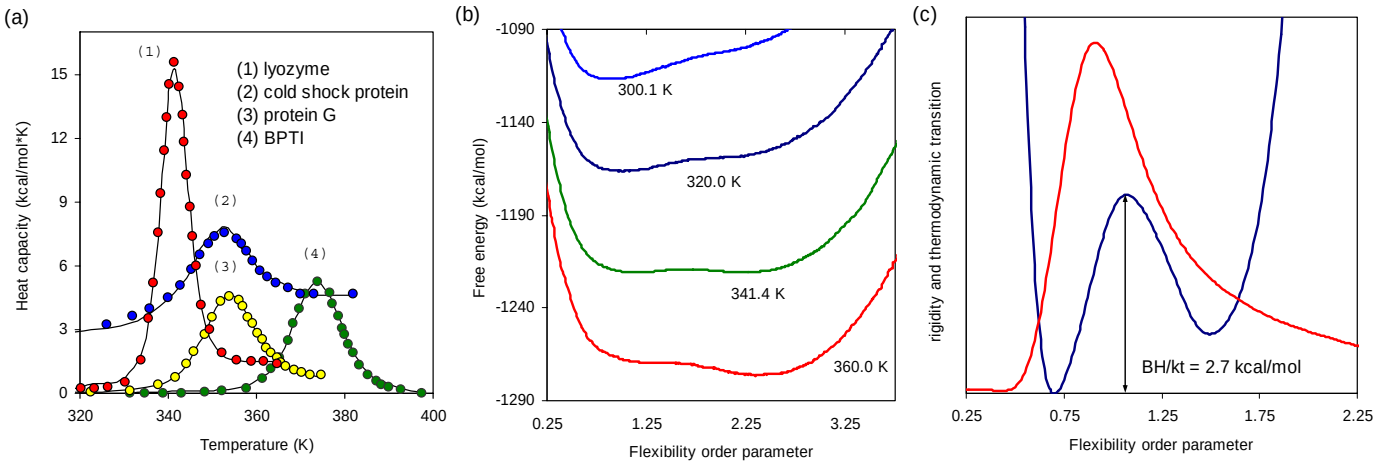


**Fig. 2** – **(a)** Typical best-fits to four heat capacity curves. **(b)** Landau free energies for lysozyme where $T_m$ = 341.4 K. **(c)** Magnified look at G(T,θ) (blue line) for lysozyme at $T_m$ highlighting the barrier separating the two phases. (Note: $T_m$ does not correspond to two minima of precisely equal depth whenever the two local wells have different shape.) The rigid cluster size susceptibility for lysozyme (red line) locates a percolation threshold, which describes the mechanical transition from rigid to floppy.

We currently have a crude representation of the DCM. At this point, DCM parameterization is achieved by fitting to experimental data. Despite its simplicity, DCM results reproduce a wide variety of experimental heat capacity protein folding curves, **the first biophysical model or free energy decomposition scheme to do so**. After parameterization, the DCM can be used to calculate the free energy landscape of the protein (lysozyme is given as an example in Fig. 2b). The calculated free energy landscapes acts as a bridge to the underlying mechanical transitions taking place. An important feature of all free energy landscapes is that at the $T_m$ there are two nearly equal minima separated by a free energy barrier, indicative of a first-order (two-state) phase transition. Further, the DCM is able to capture small, realistic energy differences (Fig. 2c), between very large numbers (Fig. 2b). In all cases, the rigidity percolation threshold, which describes the mechanical transition, parallels the thermodynamic transition, although they are never exactly simultaneous (Fig. 2c). The relative location of the mechanical and thermodynamic transitions provides realistic predictions of transition state "compactness". For example, the DCM reproduces the experimental result that the transition state of cold shock protein is "remarkably native-like". These results were recently published in *FEBS Letters*.

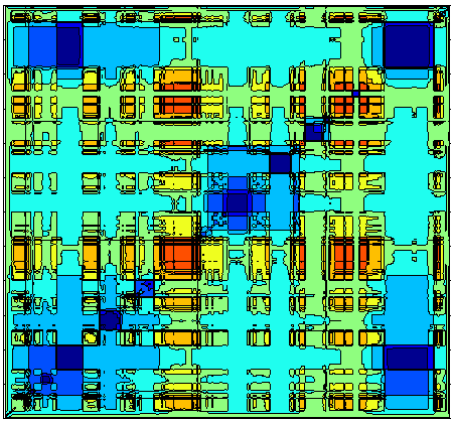

**Fig. 3** – This plot shows regions that are flexibly/rigidly correlated regions within the histidine binding protein. It is used to identify allosteric effects present in a protein. That is, application of a constraint at one location can produce an effect on conformational flexibility far removed from that location. A single number ranging from [+1=red; -1=blue] represents the degree of cooperativity in flexibility as visually shown. Mutant structures of the same protein can be analyzed to give similar plots. Differences between correlation plots will give an indication of how much the mutant protein changes its flexibility cooperativity, and this information will be correlated with known biochemical function. **We can also construct movies that demonstrate how the correlation between flexible/rigid regions changes with temperature**.

The ability to quantify stability/flexibility relationships is especially critical to our protein design efforts. Identification of correlated motions within the protein structure highlight key flexibility requirements (i.e., substrate association induced fit, allosteric conformational changes, etc.). An example of a cooperativity correlation plot (histidine binding protein) is shown in Fig. 3. Analysis of the cooperativity correlation plots from a family of orthologous proteins allows us to identify critical functional motions. Once these allosteric motions have been identified, mutant structures can be screened vis-à-vis changes thermodynamic stability and functional motions. Stabilizing mutant structures that restrict catalytic normal modes are excluded from future consideration. **This strategy meets our design mandate of conferring increased thermostability to a given protein structure without compromising function at lower temperatures.**

Currently, we are applying the DCM to a variety of protein examples, including HIV protease. In this study we are attempting to quantify the stability and flexibility changes that occur on binding of different HIV protease inhibitors. We are also currently investigating pairs of mesophilic/thermophilic orthologs (specifically RNase H and Cytochrome P551) in order to compare the stability/flexibility profiles **at their respective optimal growth temperature**. Papers describing these investigations should be submitted before the end of the year. We have recently submitted an NIH-SCORE renewal proposal to continue to fund our familial stability/flexibility profile comparisons. Additionally, an NIH-R01 proposal will be re-submitted in February in order to fund a substantial upgrade of the current theory. The two most important outcomes of the R01 research will result in **(1)** a more robust parameterization, eventually eliminating the need for fitting to experimental data, and **(2)** incorporation of long range electrostatic effects. The R01 was originally submitted in 2004, while not recommended for funding, the referee's comments were encouraging.

## 4. PREDICTING PROTEIN FUNCTIONAL SITES WITH PHYLOGENETIC MOTIFS

Several recent efforts have attempted to use sequence motifs as bioinformatic tools to predict functional sites. Unfortunately, these efforts suffer from exceedingly large numbers of false positives. One of the primary objectives of our current bioinformatic efforts is to develop and apply accurate functional site prediction schemes using evolutionary information. The motivation for the proposed work stems from the growing list of anecdotal evidence indicating that motifs conserving phylogeny are often directly related to structure and/or function, including the copper, zinc superoxide dismutase and enolase families (see Livesay et al, *Biochemistry*, **42**:3464-3473).

We have reversed the above scenario and recently demonstrated that sequence fragments approximating the complete familial tree (termed *phylogenetic motifs*) represent good functional site predictions. We briefly highlight the key results of our previous report here. Across a structurally and functionally diverse protein family dataset, **phylogenetic motifs (PMs) consistently correspond to functional sites defined by surface loops, active site clefts, and partially buried regions interacting with prosthetic groups.** In all instances, the functional importance of the identified PMs is verified through structural comparisons (see Fig. 4 & 5). PMs structurally cluster around known functionality despite little overall sequence proximity. Similarity between traditional and phylogenetic motifs is generally observed. However, there are instances (i.e. cytochrome P450) when PMs are not overall well conserved in sequence. **This point is enticing because it implies that PMs are able to functionally annotate regions where traditional motifs fail.** The PM approach is similar *in spirit* to the evolutionary trace method, and as expected, the results from the two methods are consistent. However, PMs ostensibly identify sequence clusters of evolutionary trace residues, which significantly improves prediction accuracy. Finally, tree significance, especially in the PM regions, has been demonstrated using bootstrapping. A manuscript describing these results is currently published online in *Proteins*, *Structure, Function, and Bioinformatics.* Additionally, we have just been informed that our paper will be featured on the cover when published in hardcopy form. The bulk of this work has been performed by two Cal Poly Pomona students.

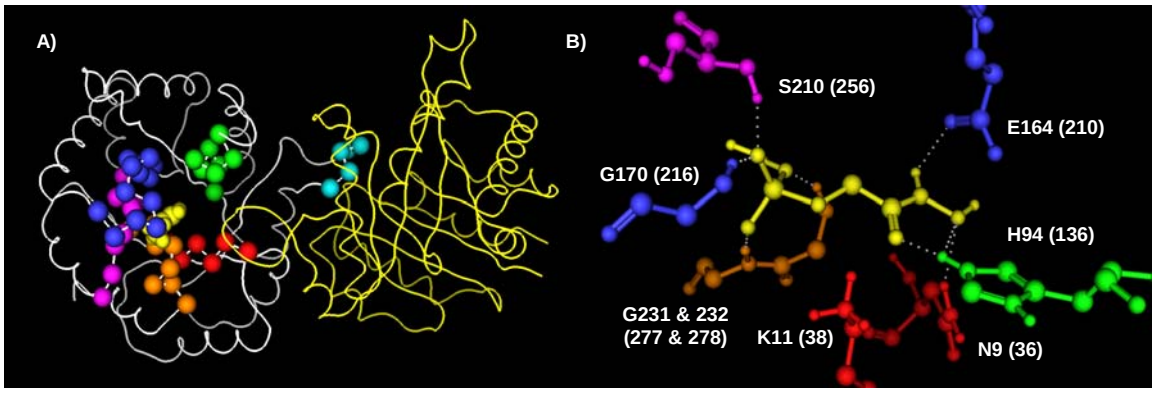


**Fig. 4** – **(a)** A *Saccharomyces cerevisiae* triosephosphate isomerase structure annotated with the identified phylogenetic motifs. The cyan phylogenetic motif corresponds to the catalytically important "flexible lid", which covers the active site on substrate binding. All of the remaining phylogenetic motifs are directly interacting with the substrate. In fact, the best scoring region completely covers the PROSITE active site definition. The substrate analogue is colored yellow. **(b)** Blowup image of the active site region. All H-bonds and salt bridges between the enzyme and substrate are identified as PMs. Coloring in (b) is the same as (a)

Currently, we are also attempting to exploit the "motif-ness" of PMs. We believe that, in addition to an accurate functional site prediction scheme, PMs are a promising approach to (globally) assigning function to ORFans. Additionally, as we have demonstrated previously (La et al. *Biochemistry*, **42**(30):8988-8998), there are significant differences between sequence comparisons based on motifs and complete alignments in large scales analyses. Our

current PM identification algorithm relies on an underlying alignment, which is potentially problematic. In order to alleviate alignment quality concerns from our multi-genome analyses, we will soon be developing a second PM identification algorithm that is based on pre-computed motifs. Other recent phylogenetic motif investigations include:

- Using phylogenetic motifs, along with Poisson-Boltzmann electrostatics, to investigate the evolutionary and catalytic importance of conserved electrostatic networks (this work has been submitted to *Protein Science*),
- Development of an automated phylogenetic motif identification algorithm based on k-mediods clustering of the pairwise phylogenetic similarity scores (a manuscript describing this work is currently being finalized),
- Complete functional annotation of the COG database,
- Implementation of a web-based phylogenetic motif identification server (called MINER), which can be accessed at http://www.pmap.csupomona.edu/MINER/

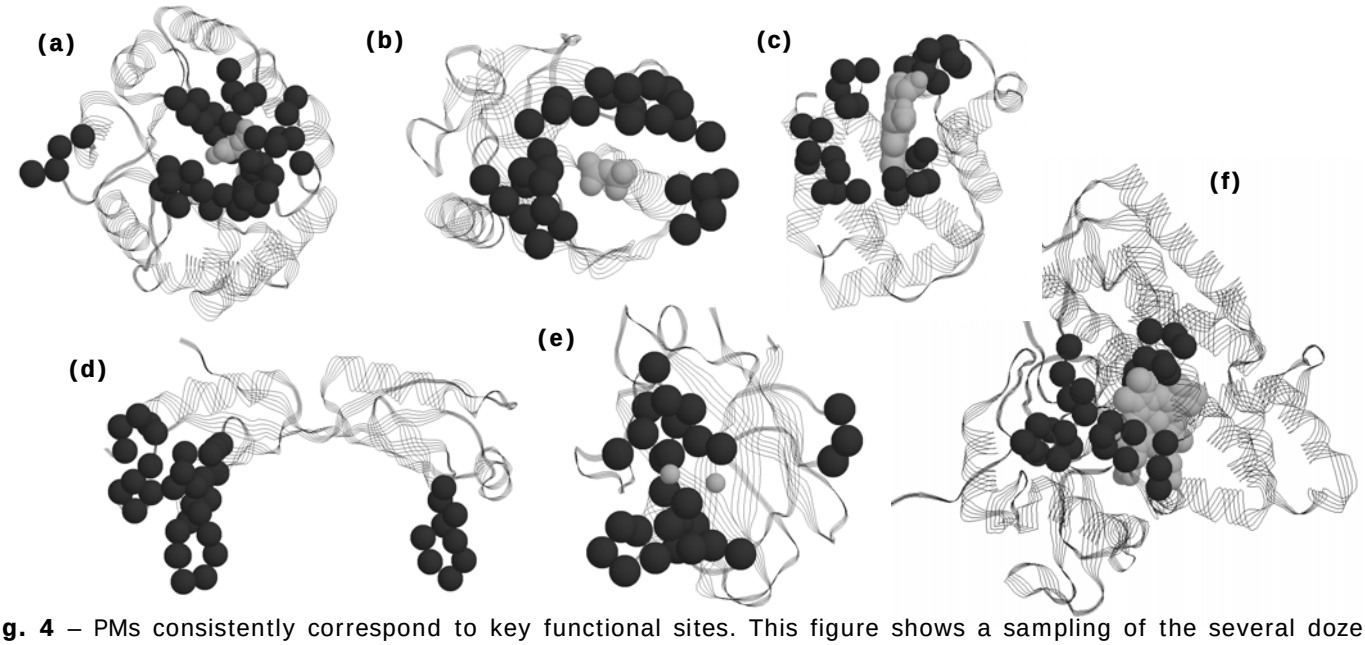


**Fig. 4** – PMs consistently correspond to key functional sites. This figure shows a sampling of the several dozen structurally diverse examples currently investigated. Identified PMs are structurally clustered and correspond to functional sites in a wide variety of proteins, including: **(a)** TIM **(b)** inorganic pyrophosphatase, **(c)** myoglobin, **(d)** TATA-box binding protein, **(e)** CuZn superoxide dismutase, and **(f)** cytochrome P450. Dark spheres represent phylogenetic motif $\alpha$-carbons; light spheres represent substrate analogues.

## 5. SYSTEMS BIOLOGY AND THE DYNAMIC PROTEOME (a future research focus)

In addition to continuing our two main research focuses, I plan (in the long term) to also develop and apply computational methods for understanding proteome dynamics. Specifically, I am interested in understanding the time-dependent properties of the gene products expressed and how expression is effected by particular physio-chemical properties of the underlying species. My group will be involved in the development of software and databases for cataloging and analyzing the proteins expressed in any given organism (and for any given cell in eukaryotic organisms) at any given time, including isoforms and post-translational modifications. In close collaboration with experimentalists, we will seek to develop systems-based libraries of the proteins present within the cell under varying conditions. Through efficient database cataloging and analysis of the empirical time- and stimuli-dependent protein concentrations, we should be able to develop improved hypotheses and conclusions concerning the *in vivo* processes governing protein flux. As data-mining improves, so will our understanding of the complex multi-variant data, thus we should eventually be able to design new experimentation that will test our hypotheses further.

The complexity of biology is built upon physics and chemistry. As such, we will also try to incorporate our understanding of protein sequence/structure/function relationships in order into cellular models and probe their global effects. For example, scientists have studied enzyme kinetic processes since the dawn of biochemistry. However, very little is known about how enzyme efficiency is affected by changes in the cellular milieu, which of course is related to the (in)accuracy of kinetic models of the cell. Local concentration of certain species may inhibit (or activate) other cellular species through non-specific, crowding effects. Therefore, biophysical approaches are necessary to provide insight into the origins of cellular complexity. By working with both the empirical (database) data and the theoretical models, we should be better equipped to understand the proteome. As our understanding improves, this information will be incorporated into systems biology (kinetic, stochastic, and graph-theoretic) models of the cell.

## TEACHING STATEMENT

When I began the search for my current position, I focused specifically on primarily undergraduate institutions because of the importance they attribute to good teaching. Despite the fact that I now wish to move to a more research oriented institution, being a quality educator is still very important to me. As a researcher, my work is at the interface of chemistry, biology, physics, and computer science. Although it is much harder to implement in the classroom, I strive to maintain this same interdisciplinary viewpoint as an educator. At Cal Poly Pomona, I have been actively involved in the creation and implementation of a new computational chemistry option (*Molecular Modeling & Simulation*). This new degree will train scientists to employ computational methods to address a wide variety of chemical problems. In addition to my normal *Biochemistry* teaching responsibilities, I have also (personally) developed three original courses, two of which are in support of our new degree. Each of these courses, *Bioinformatics*, *Macromolecular Modeling*, *Proteomics*, reflect my interdisciplinary bent, and are routinely populated by undergraduate and graduate students from multiple departments. (Note: curricula materials for each of the above upper division courses can be found at http://www.csupomona.edu/~drlivesay/courses.html.) I have also recently been involved in a collaborative effort developing and delivering a "team-taught" Bioinformatics course within the Biology Department. This course was taught by five Cal Poly Pomona faculty from three different departments (Biology, Chemistry, and Computer Science). Below is a list (with brief descriptions) of the courses that I have taught since arriving at Cal Poly Pomona.

| Course # | Course title | # of times | Course description |
|---|---|---|---|
| CHM 121 | General Chemistry I | 2 | First quarter general chemistry |
| CHM 321 | Elements of Biochemistry | 5 | 1 quarter survey course |
| CHM 321L | Elements of Biochemistry Lab | 4 | Lab for CHM 321 |
| CHM 327 | Biochemistry I | 4 | Proteins, carbohydrates, lipids, and kinetics |
| CHM 327L | Biochemistry I Lab | 4 | Lab for CHM 327 |
| CHM 328 | Biochemistry II | 2 | Anaerobic & aerobic metabolism |
| CHM 328L | Biochemistry II Lab | 1 | Lab for CHM 328 |
| CHM 329 | Biochemistry III | 5 | Nucleic acids and genetics |
| CHM 329L | Biochemistry III Lab | 4 | Lab for CHM 329 |
| CHM 416 | Macromolecular Modeling | 2 | Molecular dynamics, Monte Carlo, Poisson-Boltzmann electrostatics, etc. |
| CHM 417 | Bioinformatics | 2 | Theoretical foundation and application of sequence & structural analysis algorithms |
| CHM561 | Protein MS / Proteomics | 2 | Protein and whole cell MS, yeast-2-hyrbrid, co-affinity assays, 2D-electrophoresis, etc. |
| BIO 499 | Introduction to Bioinformatics | 1/5 | my portion covered protein sequence and structure comparison algorithms |

I am currently teaching CHM416 (*Macromolecular Modeling*) for a second time. The course was very well received previously, but I was unsatisfied with the lab exercises that I had developed. As such, this time around I am having the class do a group, course-long research project. The project uses a combination of molecular dynamics simulations and Poisson-Boltzmann continuum electrostatics theory to investigate the dynamical nature of residue pKa values. The methods employed are the same as those covered the last time I taught the course, but instead of disparate lab exercises, they are used in a sequential research project. The *process* of research has really engaged and excited the students much more than traditional lab exercises. While designed to be a "safe" project, the results are quite interesting -- the students will present a poster describing their results at the 2005 Annual Meeting of the California State University Program for Education and Research in Biotechnology. Additionally, I am planning on writing and submitting a research manuscript (with the students as coauthors) detailing the work in the spring.

As a faculty member at a predominantly undergraduate institution, one of my most important teaching responsibilities (and most rewarding!) is mentoring student research. Since beginning at Cal Poly Pomona, twelve undergraduate and four graduate students have worked in my laboratory. Their work has led to four published peer-reviewed journal articles (one more is in review and two more are in preparation), all of which are co-authored by CPP students from my lab. (Cal Poly Pomona students are the first two authors in all six of the seven manuscripts). Additionally, my students have presented their results at National and local meetings, where they have been very well received (students from my lab won awards at the 2002 and 2004 annual meetings of the Protein Society and California State Program for Education and Research in Biotechnology, respectively). Six students from my lab have moved on to Ph.D. programs, and two others are currently applying.