Protein analysis: the need for speed

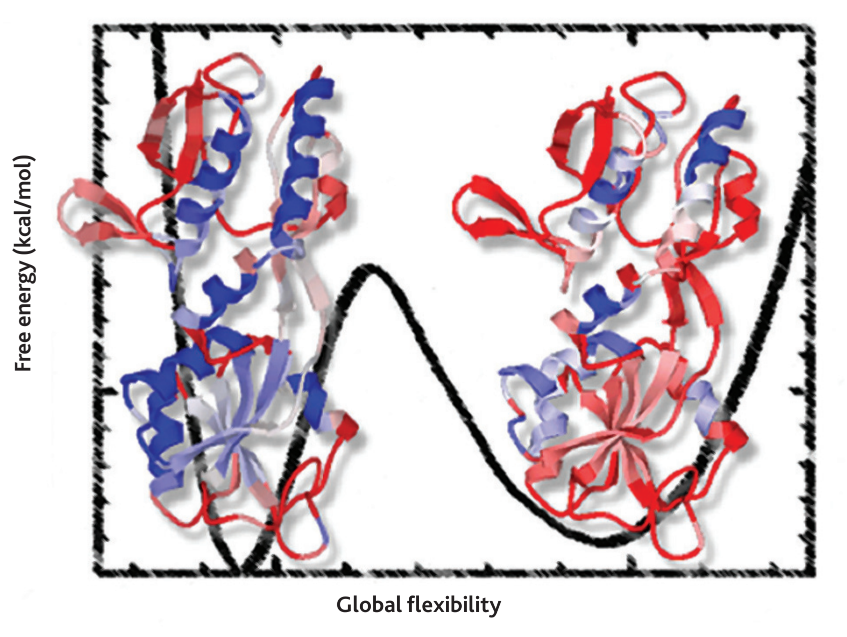
As Professor Donald Jacobs explains, a research team at the University of North Carolina at Charlotte (UNCC) is developing *FAST*, a high throughput protein analytical software platform which could transform the field of computer-aided macromolecular design

Could you explain the overall aims of the project?

The project consists of two aims: firstly, to develop theoretical concepts and methodologies that make it possible to accurately and rapidly calculate Quantified Stability/Flexibility Relationships (QSFR) of a protein; second, to release *FAST* software to provide a Flexibility and Stability Test on proteins to make possible unprecedented high throughput comparative QSFR analyses.

There are two misconceptions that linger among scientists. Firstly, erroneous additive models based on free energy decomposition are common. Secondly, those that know additive models fail in systems with strongly coupled interactions believe they must abandon the approach of free energy decomposition altogether, and resort to a molecular dynamics (MD) simulation. However, due to great computational expense, reliance on MD simulation remains unsatisfactory due to insufficient sampling problems for processes occurring on timescales of nanoseconds or more. Nanosecond time scales are only now being explored carefully. Over the last two decades, the promise of rational computer-aided design in biomolecular systems has failed, largely due to limitations in modelling and algorithm design, and reliance on faster computers to perform simulations on longer time scales and/or larger systems. However, a fundamentally different approach is needed to manage myriad high-throughput applications. *FAST* is designed specifically to meet this challenge, and thus it has great potential to revolutionise computer-aided macromolecular design.

QUANTITATIVE STABILITY/FLEXIBLE RELATIONSHIPS



QUANTIFIED STABILITY/FLEXIBILITY RELATIONSHIPS PRECISELY DESCRIBE THE GIVE-AND-TAKE BETWEEN PROTEIN STABILITY AND GLOBAL FLEXIBILITY

At intermediate temperatures, the free energy landscape is generally characterised by two competing minima, one for the natively-folded structure and another for the unfolded protein. In this example showing bacterial periplasmic binding proteins, the structures superimposed onto the free energy landscape are color-coded by their local flexibility profiles (blue = rigid and red = flexible)

Our alternative approach is motivated from the realisation that the problem of using an additive model is not associated with the free energy decomposition, but rather, on how to account for nonadditivity in the process of free energy reconstitution. This project is based on the pivotal hypothesis that network rigidity provides a mechanical mechanism for enthalpy-entropy compensation, which restores utility of a free energy decomposition approach.

What were the main obstacles that you had to overcome to develop the *FAST* software?

All obstacles of the project are related to answering one question: namely, how to capture the complexity of chemistry in order to arrive at an accurate model using a general mathematical formalism that is also efficient computationally? Two breakthroughs occurred while weighing tradeoffs. Firstly, a rigorous connection between conformational entropy and distance constraints was developed to establish a procedure for determining transferable parameters. Secondly, a self-consistent mean field calculation was developed, based on degrees of freedom and constraints as a probability flow within a network of links, which themselves have capacities defined by the molecular interactions subjected to thermal fluctuations.

Is the *FAST* software applicable to any other fields other than protein engineering and, if so, could it unlock potential in these areas?

The theoretical developments are based on fundamental physical principles, and thus have general applicability ranging from molecular biophysics to condensed matter physics and beyond. We are in an embryonic stage, injecting profound ideas and new concepts into the physics community that are powerful, yet, very foreign. The paradigm put forth that combines constraint theory and free energy decomposition could be as revolutionary as the Landau theory of phase transitions, BCS theory for superconductivity, Fermi liquid theory and renormalisation group theory. Hopefully young researchers will be inspired to further develop this paradigm and realise an impact of revolutionary caliber.

What is the future of the project? Will the recent developments in processing abilities have a significant impact on any future *FAST* developments?

This project has focused on the thermodynamic and mechanical response of aqueous proteins. More development will continue to accurately predict mixed solvent and pressure effects. Extension to membrane bound proteins has been foreseen in the initial modelling scheme. Utilities are being added to investigate allosteric regulation. We will expand capabilities to include ligand-binding kinetics. *FAST* will serve as a platform for many years of development, and we hope others will come on board and make use of this innovative approach. The processing capability of *FAST* for end-user and developer functionality is key to establishing credibility and significance throughout the scientific community. It is necessary that the number of satisfied *FAST* users rapidly grows to justify future developments. *FAST* must sell itself based on how successful its predictions prove to be in diverse biochemical/biological applications.

The current processing ability of FAST is estimated to take several minutes of CPU time on a single processor to obtain a complete 3D free energy landscape for moderate size proteins. Is there a desire to reduce this processing time?

We are doing our best to make sure this estimate is realised. Already extraordinarily fast for an all-atom model, parallel computing increases the speed further. Over time, we expect the calculation will slow down as we focus on refining the model to improve accuracy. In fact, we do not mind if the programme eventually slows down by an order of magnitude over the years going forward because the steady increase we enjoy in speed of computers will compensate. On the other hand, we can consider a more coarse-grained model to handle much larger systems (such as viruses). In this approach, the calculations can be made much faster, but then it will be less accurate. The tradeoff is always the balance between accuracy and speed.

Can you offer an update on the release of *FAST*?

We are doing our best to have *FAST* released by December of 2010. Our NIH funded project has a no-cost extension until February 2011. We will not publicly release *FAST* until it provides useful functionality to an end-user, because quality is important. Pre-release versions of *FAST* will be available to end-users within collaborations.

Unlocking the molecular basis of disease

Professor Dennis Livesay from the UNCC research team behind *FAST* is in no doubt that despite the difficulties of measuring protein flexibility, the exciting potential for therapeutic applications such drug screening heralds a bright future for the project

Can you outline the basics and complexities of focusing on conformational flexibility?

As related to us, the primary complexities associated with focusing on conformational flexibility are two-fold. Firstly, it is very difficult to measure protein flexibility in general. While there are several experimental methods to do so (e.g. NMR, FRET), these methods require much time, skill, and investment. As such, there is a strong need to develop computational methods that can act as surrogates to these experimental methodologies.

The second set of complexities is related to the computational difficulty of doing just that, which is directly related to the sheer number of conformational states that must be considered. A computational model of protein flexibility must account for an astronomical number of states, while doing so in a time that is computationally tractable. The way to do this is to bin (coarse-grain) similar conformations together; however, if taken to the extreme then all of the important details are lost. For example, consider a number continuum between 0 and 6 - if the problem requires that the precision be in the thousandth decimal place, one must consider 6000 states. To speed up the calculation, one could round and group states, but if taken too far (e.g. just using a 6-sided dice), the results are no longer measured in detail. We believe that compared to existing methods, the *FAST* optimally balances these two competing characteristics.

What is the critical operational window in which an enzyme is flexible enough to mediate a reaction pathway and yet remains rigid enough to achieve molecular recognition?

This is another very important difficulty in measuring protein flexibility. The timescales of the motions within a protein typically span ~12 orders of magnitude! For example, the highest frequency motions are related to bond vibrations, which occur on the scale of femto- to picoseconds, whereas diffusive motions can take as long as seconds to minutes. Critical elements of protein structure and function occur at both extremes of the scale, and everything in between. Based on this tremendous diversity within timescales, no single method (computational or experimental) is able to accurately describe them all. As such, particular methods must focus on specific timescale regimes. Based on the common belief that they are most important to protein structure and function, we focus on long timescale, quasi-stationary, motions. While time does not exactly map to the *FAST* approach, the timescales that we are probing are primarily nanoseconds and beyond.

Can you give some examples of the genetic diseases that are affected by amino acid mutations within a protein?

Most diseases are complex associations between many different mutations that lead to a set of propensities for it, which has greatly confounded genome association studies. Nevertheless, there are some 'simple' diseases whose molecular origins can be traced back to a single amino acid mutation. For example, sickle cell anemia is caused by a point mutation where a hydrophilic (water-loving) glutamate residue on the surface of one of the hemoglobin proteins is mutated to hydrophobic (water-hating) valine. As a consequence, the hemoglobin proteins stick together, and the valines thus become hidden from water. It is this precipitation that causes red blood cells to become sickle-shaped, thus leading to the disease pathology.

What are the practical medical applications of the *FAST* software?

A proper understanding of biophysics is imperative to describe how diseases emerge from their molecular underpinnings. From these mechanisms, new therapeutic strategies can be uncovered. *FAST* is expected to be used frequently at the more basic levels of biomedical research. However, once a therapeutic approach has been selected, *FAST* will also be able to computationally screen for drug candidates that satisfy the therapeutic approach.

What have been the major results of this project to date?

In addition to myriad theoretical advances that are unique to *FAST* (i.e., reproducing experimental protein folding heat capacity curves), the primary application success is based on a high-dimensional assessment of the Quantitative Stability/Flexibility Relationships (QSFR) within a target protein. QSFR data describes both the mechanical and thermodynamic properties of the protein, and their relationships. Moreover, *FAST* provides this QSFR data very quickly. As such, we can actually compare QSFR descriptions across groups of evolutionarily related proteins. Taken as a whole, our results demonstrate that protein families have evolved such that local (per residue) descriptions of protein flexibility are generally conserved. However, how flexibility and rigidity propagate through the protein's structure (called allostery) varies significantly, which is consistent with a growing consensus regarding the susceptibility of allosteric mechanisms.

Will the development of the *FAST* software require specific training or has it been designed to be easily accessible?

FAST has been developed to be a generalised DCM (distance constraint model) solver, representing our underlying biophysical approach. We have spent considerable effort to make *FAST* extendable so that it will support several model generations. Given a particular model, *FAST* should be accessible to anyone within a general background in protein biophysics within a few hours of training.

How do you predict the field will progress and are there significant discoveries that still remain to be made?

From a technical viewpoint, I expect that both computational and experimental innovations will continue at roughly the same pace for years to come. However, in my opinion, the real advance will be in what people do with the information that is being generated. The past ~20 years has clearly established the importance of protein stability and flexibility, and provided many deep insights into how they control functional mechanisms. Yet, without trying to sound too negative, most of the advances to date have been descriptive. In the next 20 years, the advances will be based on what is done with this information. In fact, we have invented (patent pending) a procedure to computationally design proteins and other macromolecules based on the specific protein stability and flexibility signatures. For example, one can imagine improving an enzyme's functional efficiency by amplifying catalytic normal modes. Our design process will focus on a specific set of correlated motions that were a priori known to be important, and attempt to stabilise the protein without losing the mechanical response that is important for function. The designed mutant would function well while being more thermodynamically reproducible.

INTELLIGENCE

PREDICTING PROTEIN FLEXIBILITY AND STABILITY

OBJECTIVES

The project has two aims: firstly, to develop theoretical concepts and methodologies that make it possible to accurately and rapidly calculate Quantified Stability/Flexibility Relationships (QSFR) of a protein; secondly, to release *FAST* software to provide a Flexibility And Stability Test on proteins to enable unprecedented high throughput comparative QSFR analyses.

FUNDING

This project is supported by NIH GM073082

CONTACT

Donald Jacobs
Associate Professor

Department of Physics and Optical Science
University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223
Grigg Hall 343

T 1 704 687 8143
F 1 704 687 8197
E djacobs1@uncc.edu

DONALD J JACOBS (Don) is Associate Professor of Physics at UNC Charlotte. His expertise is in statistical and computational physics (BS 1985, Union College; PhD 1992, Purdue University). After postdocing at the Institute of Theoretical Physics of the University of Utrecht, and at Michigan State University, his first faculty position was at California State University, Northridge. With more than 35 papers and 3 patents on the subject, Don has been modelling protein flexibility and stability since 1997.

DENNIS LIVESAY is an Associate Professor of Bioinformatics and Genomics at UNC Charlotte. His research interests are in the area of protein family sequence/structure/function relationships. Prior to UNC Charlotte, he was an Associate Professor of Chemistry at the California State Polytechnic University. All of his education is in chemistry, including a PhD in physical chemistry from the University of Illinois at Urbana-Champaign.

UNC CHARLOTTE