Life in Translation

The challenges of research along the biophysical borderline

by

Jeremy L. England

What makes physics physics? More precisely, which kinds of scientific questions are physical ones, and which kinds are not? Already, the reader may suspect a trick; after all, aren't physicists the ones endlessly touting their 'theories of everything' and 'universal' laws? What sort of phenomenon could possibly lie beyond the scope of everything in the universe?

Perhaps not surprisingly, this turns out to be the wrong way of looking at things, for, while it *is* sensible to try to understand all the reproducible phenomena in the world from the standpoint of physics, it need not be the case that the physical account of a thing or an event is the only type of account one can give. Just ask a biophysicist!

Rather than being a simple subfield of the physical sciences, biophysics straddles a wide terminological and methodological gulf between two great scientific disciplines, and doing it right involves constantly switching back and forth between two wholly distinct frames of analysis. One may look at the living thing through the eyes of a biologist, and see a sensible, organized system with components whose forms provide the basis for the functions that they carry out within the context of the whole organism. At the same time, one may see the same system like a physicist might: as a lump of atoms and molecules, or of colloids and fluids. The task of the biophysicist is to identify the cases where one of these viewpoints can help to strengthen and enrich the other, and this turns out to be no mean feat. Biological systems do not operate outside the laws of physics, but they often are interesting for reasons that have little to do with their physics; we can, for example, predict the

inheritance of many traits from one generation to the next without having any idea what the physical basis for heredity even is. That being said, we are now beginning to discover many new cases in which the physical and biological understanding come hand in hand.

Proteins: Life's building blocks ... and builders

To wit, the molecular and cellular scale (0.1 nm to 1000 nm) is where rubber meets road, biophysically. Individual small molecules exhibit relatively simple, predictable dynamics from a physical standpoint, yet live cells, which are just giant, diverse mixtures and concatenations of such molecules, display an almost unfathomably complex array of interesting behaviors. Molecular and cellular biophysics therefore constitutes a particularly active scientific playground, situated right at the boundary between inanimate and living matter. Down at that range of scales, a multitude of different types of nano-sized events can be identified for their importance to the overall functioning and survival of an organism, and can be subsequently characterized in terms of their physical mechanisms based on well-established theories of intermolecular interactions. In this way, we can start to leverage the vast predictive power of physics to make better sense of how living things "work."

Understanding the physics of how protein architecture gives rise to biological function at the molecular level is one of the most exciting challenges in biophysics, and has broad implications not only for basic life science, but for health and medicine as well.

Proteins are a striking case in point. Much of our DNA-based genetic code acts as a template for the assembly of chains of amino acids called PROTEINS. There are twenty different types of amino acid that can be coded by our DNA, and they are distinguished by the chemical group, or side-chain, that sticks off of the so-called peptide backbone of the protein. Each kind of side-chain confers on the corresponding amino acid a unique portfolio of physicochemical properties, ranging from size to shape, and from electrostatic charge to flexibility of rotation about various internal chemical bonds. Biologically speaking, proteins are where it all begins: take a single amino acid in isolation, and all it can do for you is vibrate, rotate, and bounce around like a ping pong ball. String several hundred different amino acids together in a specific sequence along the polypeptide chain, however, and suddenly you have a 'macromolecule' whose natural three-dimensional shape renders it capable of sensing minute changes in its environment, relaying signals to its partners in biochemical networks, catalyzing chemical reactions, and even carrying out mechanical work like an engine! Not surprisingly, understanding the physics of how protein architecture gives rise to biological function at the molecular level is one of the most exciting challenges in biophysics, and has broad implications not only for basic life science, but for health and medicine as well.

The puzzle of how to predict a protein's three-dimensional structure from its amino acid sequence is often called the "protein folding problem," and really should be thought of as a family of problems related to different aspects of protein

biophysics. The basic formulation of the question came into focus after the landmark work of Christian Anfinsen, who was able to show in a test tube that a protein called ribonuclease could be reversibly 'folded' and 'unfolded' in and out of its functional conformation, without any other molecular components to help it. This result indicated that the specific sequence of amino acids out of which ribonuclease was built was sufficient to self-assemble into the correct, 'native' three-dimensional structure under appropriate conditions (*Figure 1*). In other words, the physical interactions arising from the so-called "primary sequence" of the chain (that is, its sequence of amino acids), were precisely balanced so as to bias the equilibrium shape of the macromolecule to be one that could carry out its catalytic function.

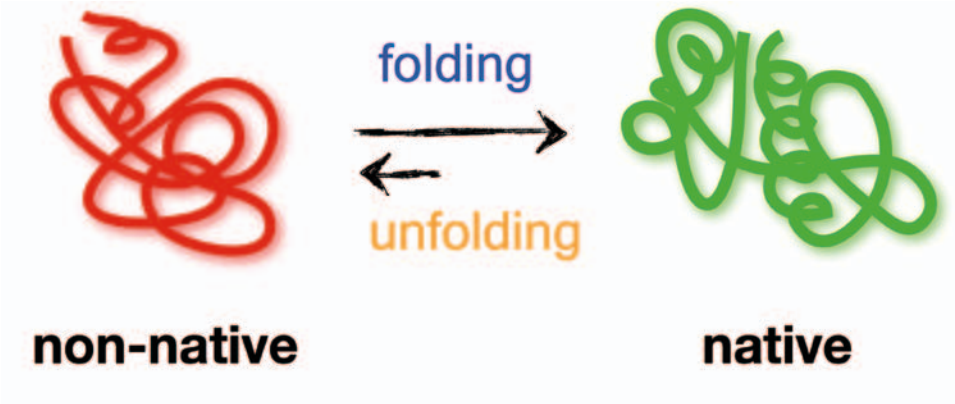


FIGURE 1

Proteins are polypeptide chains *assembled out of specific sequences of amino acids. Under physiological conditions, a protein's sequence produces a subtly balanced set of forces that guide it into a shape, called its "native" conformation, that provides the basis for its function in the cell.*

The Anfinsenian paradigm has spurred decades of theoretical and experimental work on the protein folding problem, which has turned out to be a devilishly hard nut to crack. While it has long been possible to obtain structural information about some proteins at atomic resolution using methods such as X-ray crystallography and nuclear magnetic resonance (NMR), the general question of how structure may be predicted from sequence has proven more difficult for several reasons. First, although it is reasonable to assume that a protein's native structure will minimize its energy (or, more accurately, its 'free energy'), many different physical factors contribute to the outcome for a given sequence with roughly the same strength, and this suggests that a brute force, high-resolution numerical computation of the energies involved is necessary to tabulate and weigh all the forces. The problem here is that the number of different arrangements for a modestly-sized protein containing merely thousands of atoms is already astronomically vast, and that makes the calculation costly and slow, and also causes the result to depend on a large number of modeling parameters. Thus, while supercomputing approaches have shown great success in predicting structure in particular cases, they bring with them significant caveats and disadvantages.

A second challenge to the Anfinsenian paradigm comes from the biology itself. It is convenient to suppose that proteins adopt single, functional native structures in the biological context, but the real story is much more complicated. Many proteins

are capable of adopting more than one functional conformation, and it is often the switching among these conformations that underlies some aspect of that protein's function. Still other proteins do not adopt any particular structure at all, and rather remain disordered even as they function in the cell. Finally, many proteins can only fold into the proper shape with help from a class of proteins called molecular "chaperones," which bind to unfolded proteins and guide them along the way as they fold. Chaperones, moreover, are particularly important because protein structure is highly context-dependent, and the intracellular environment is so densely crowded that it is almost meaningless to ask how a protein would fold without chaperones being there to constantly prevent catastrophic off-pathway misfolding and aggregation with its neighbor.

Protein structure gives rise to protein function, but often it is the way a protein's structure **changes** in response to an environmental cue that allows it to carry out its proper role in the cellular context.

A game of musical chairs

All of these biological facts confound the physical simplicity of the one sequence, one structure idea. Perhaps surprisingly, though, they also light the way to a new approach of remarkable theoretical simplicity. What conformational change, structural disorder, and misfolding-driven aggregation all have in common is that they force one to conceive of the protein not as a single shape, but as a blur of different shapes that the polypeptide chain explores as thermal energy causes it to fluctuate. The nice thing about a blur, of course, is that lots of different high-definition images look the same once they get blurred out, and this suggests that if we conceive of *all* proteins as blurry—that is, if we only go for a rough, approximate representation of structure contained in some ensemble of different individual possible shapes—then we may be able to get away with ignoring many of the physical forces that only influence the finer details of how a protein folds.

This is the approach we have taken in the England group at MIT. We start by assuming that the protein may be thought of as a string of beads or marbles scrunched up in a ball. In other words, we have a polymer (a string of amino acid beads), and it is reasonable to presume that this polymer will tend under folding conditions to collapse into a "globule," *i.e.*, into a ball. The reason we say we are dealing with beads or marbles here, though, is that each amino acid takes up some volume in this ball, meaning that the "steric repulsion" between one bead and another becomes very strong at very close range and prevents the beads from overlapping in space. The question now simply becomes: Why would our chain prefer one scrunching of the beads over another?

For our answer to this question, we turn to what empirically may perhaps be the single most important force in determining protein structure: the HYDROPHOBIC EFFECT. Some amino acids are very "oily," or hydrophobic, in their physicochemical properties, *e.g.*, phenylalanine, which strongly resembles benzene; and some are very "watery," or hydrophilic, *e.g.*, aspartate, which resembles vinegar. The play-by-play here is therefore going to look familiar to anyone who has watched salad

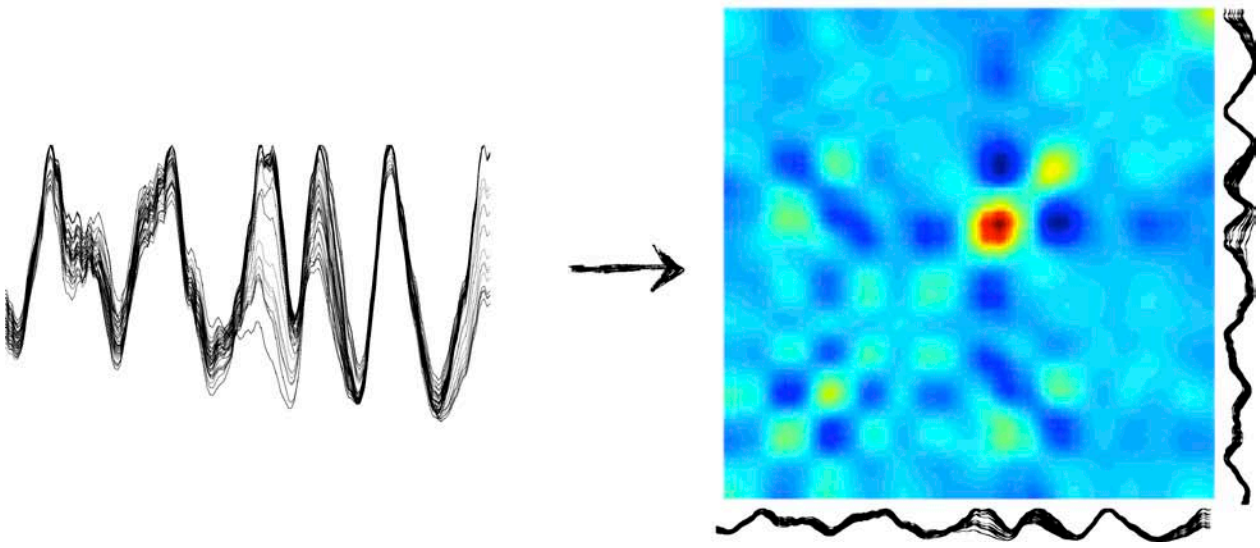
dressing separate: the protein is surrounded by water, and the oily parts are going to try to get buried on the inside, while the watery parts try to get exposed on the outside. (This is also the same principle by which detergents help solubilize little microdroplets of grease when we wash dishes.) However, two factors make this situation more complicated. First, the steric repulsion in the core of the protein turns things into a molecular game of musical chairs: there are lots of oily amino acids, but there is only so much room in the crowded core of the protein globule; thus, not every oily amino acid is going to be able to find a seat. Second, at the same time, this is no ordinary game of musical chairs, because our amino acids are linked together in a polymer! So try imagining what musical chairs would look like when combined with a conga line: in the race to the core, some watery amino acids are going to get dragged along with their oily neighbors, and vice versa, and the result is that a highly non-trivial ensemble of three-dimensional arrangements is going to do the best job burying its hydrophobic amino acids given the steric and polymeric constraints.

Shape shifting

The nice thing is that we have succeeded in mapping the problem as described above to a highly tractable, rapid computation called a LINEAR PROGRAMMING PROBLEM [1]. This means we can take a protein's amino acid sequence, and in seconds compute a "burial trace," a rough picture of how buried and exposed different parts of the protein chain are to the surrounding water. More importantly, we can use the same technique to make a map of the *fluctuations* in shape experienced by the protein chain at low energy, *i.e.*, near its native conformation. These fluctuations turn out to contain fascinating information about protein function, as well as about protein malfunctions that have been linked to disease (*Figure 2*).

FIGURE 2

Burial mode analysis *can be used to compute the conformational fluctuations of a protein from sequence. Here, an ensemble of burial traces has been generated for the sequence of a protein called LFA-1* (Left). *Each trace measures how far a piece of the protein is expected to be from the polymer's buried core (y-axis) as a function of position along the chain (x-axis). By cross-correlating the ups and downs of different parts of the chain across this whole, low-energy ensemble, a correlation map may be constructed that tells us which parts of the chain tend to move in tandem* (Right: red positive correlation, blue negative correlation).



Protein structure gives rise to protein function, but often it is the way a protein's structure *changes* in response to an environmental cue that allows it to carry out its proper role in the cellular context. Conformational change in proteins is frequently mediated by a phenomenon known as ALLOSTERY. Allosteric motion happens when a small perturbation, such as a the binding of a drug or the covalent modification of an amino acid, can take place at one location in a protein and nevertheless bring about a large structural rearrangement at a distant location elsewhere in the protein. This type of mechanism can be important for sensing and transducing signals, as when a signaling molecule triggers a change in a protein's ability to bind to another protein and thereby activate some downstream cascade of molecular events (*Figure 3*). It turns out that it is possible to use linear programming to explain these mysterious avalanches in protein structure. By computing the ensemble of low-energy burial traces for a given sequence, we can start to see how it is that certain regions of the protein chain become correlated in their motion with other regions that may be located in a relatively distant part of the globule. The coupling between them arises because of the hydrophobic game of musical chairs described above: when one cluster of amino acids vacates the protein core, another cluster elsewhere on the chain is 'next-in-line' to fall into the now more roomy core and keep the energy of the system as low as possible. This theoretical breakthrough therefore enables us to predict which parts of the protein should be targeted in order to produce responses at other locations, a finding which has potentially beneficial implications for future computational approaches to drug design.

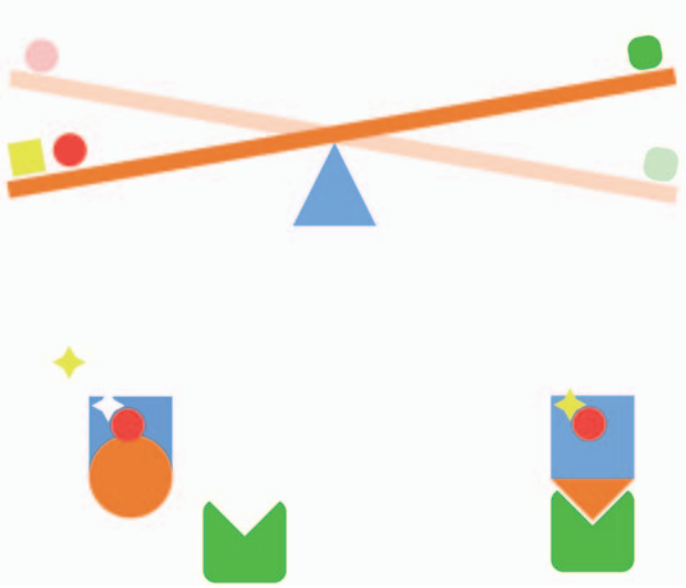


FIGURE 3

Just like a small change in weight *can tip the balance on a see-saw, a small molecular binding event can bring about a large allosteric change in shape in a protein. On the bottom left, you can see one protein (blue-red-orange), that is unable to bind to another protein (green). On the bottom right, a small molecule (yellow) binds in a pocket on the first protein, and induces an allosteric shape change which now enables the two proteins to interact. Thus, allostery enables a protein to sense a molecular signal in its environment and transduce it into a new action.*

Sometimes, however, conformational changes in proteins are not part of their natural functioning, but rather are catastrophic events that can lead proteins to clump, or aggregate together in ways that are toxic to cells, and consequently to whole organisms. Many neurodegenerative diseases, such as Alzheimer's, Parkinson's, and Huntington's diseases, for example, have had protein misfolding and aggregation implicated as a possible cause. Through our new technique of burial trace analysis from linear programming, we can now study which mutations in protein sequences are most likely to lead to increased misfolding and aggregation, and thereby get a better handle on the physical, molecular mechanisms of various harmful pathologies. For example, it is known experimentally [2] in the case of human lysozyme that a particular disease-causing mutation which substitutes a threonine for an isoleucine at amino acid position 56 leads to increased structural disorder in a particular region of the protein (*Figure 4*). When we study the conformational fluctuations of these two sequences using our model, we are able to

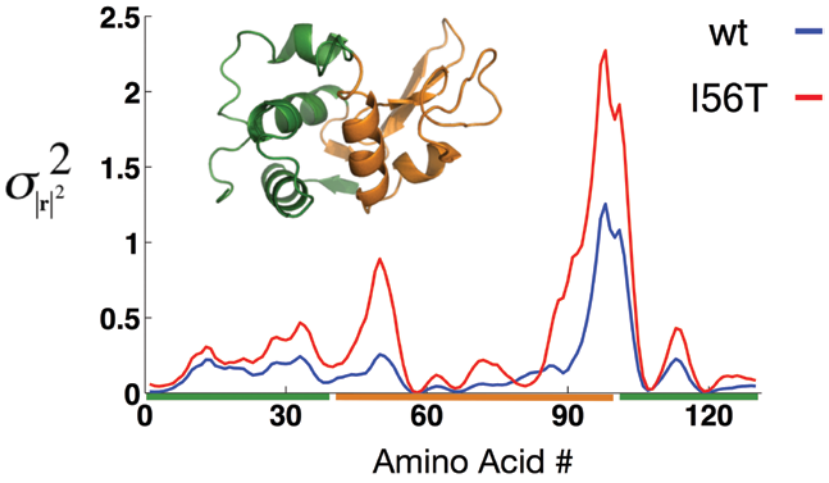


FIGURE 4

Human lysozyme is a protein that can mutate *and cause a fatal disease known as amyloidosis. It is known that for a particular single amino acid mutation (isoleucine 56 turning to threonine), a whole region of the protein (orange) becomes more structurally disordered than it is in the "wild-type" (WT) sequence. This is precisely the region of the protein chain predicted by burial mode analysis to experience increased structural variability (y-axis) in the presence of the mutation.*

reproduce the pathological effect of this mutation, suggesting it may be possible in the future to develop a more detailed understanding of how and why the protein becomes destabilized, and how this dangerous result might be prevented.

We have referred to only one example of how physical models can provide insight into events at the molecular level that are not only biologically important, but can also be medically significant. Happily, the future of research at the biophysical borderline in general looks just as bright. Advances in DNA sequencing, super-resolution live-cell fluorescence microscopy, mass spectrometry, and other nano-scale sensing methods have combined in recent years to have a transformative impact on the kind and quantity of biophysical data that may be generated out of living organisms, opening whole new frontiers in the relationships between form and function that make up the molecular foundations of life. Physics and biology both have rolls to play in driving these new discoveries, and the new generation of biophysicists will more than ever need to be fluent in both languages so as to take the greatest advantage of the many opportunities nature offers us to translate between them.

REFERENCES

[1] J. L. England, *Structure*, 2011.

[2] F. Chiti and C. Dobson, *Nat Chem Biol*, 2009.

JEREMY L. ENGLAND *joined the MIT Physics Department as an Assistant Professor in September 2011. Born in Boston, Jeremy grew up in a small college town near the New Hampshire seacoast. After earning a bachelor's degree in biochemistry from Harvard University in 2003, he began his graduate studies at the University of Oxford, and subsequently completed his doctorate in physics at Stanford University in 2009. Before coming to MIT, he spent two years as a lecturer and research fellow at Princeton University.*