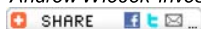


The protein structure prophet

02/01/2011

Andrew S. Wiecek

Using similar algorithms as his competitors, Yang Zhang's protein structure predictions remain in a class of their own. Andrew Wiecek investigates how he's taking the training wheels off computational protein modeling.



Forty-eight hours. That's how little time Yang Zhang needed to produce a model of a pair of variants of the enzyme pyruvate kinase M2—including a splice variant that is associated with rapidly growing cancer cells—after his colleague Gilbert Omenn walked into his office a few months ago and asked, "Have you ever thought about using your modeling tools for altered proteins?"

Although Zhang had never heard of computational modeling used to predict the structure of an alternative spliced protein, he agreed to try. "I'm used to people telling me that they have other priorities," says Omenn. But just two days later, Zhang showed Omenn some preliminary models.

Omenn was excited but not surprised. After all, he recruited Zhang as an associate professor of computational medicine and bioinformatics at the Center for Computational Medicine and Bioinformatics at the University of Michigan. And Zhang has taken top honors at the last three Critical Assessments of Techniques for Protein Structure Prediction (CASP) competitions. His computational models are offering proteomics some quick answers about protein structure and function, and he's getting ready to unveil a new method for template-free protein structure modeling.

Competition

Every other year since 1994, computational biologists have descended on the Asilomar Conference Center in Pacific Grove, CA for the CASP competitions. John Moult, professor of cell biology and molecular genetics at the University of Maryland, organized the first competition—which consisted of 35 teams—to determine how accurately a protein's structure could be predicted based on sequence information. Over 140 groups participated in the ninth competition, which concluded in December 2010.

Proteins are involved in almost every process in the cell. These macromolecules consist of a sequence of amino acids—defined by the sequence of that protein's corresponding gene—folded into a particular structure. The structure determines the protein's function, and how it interacts with other molecules in the cell. By understanding a protein's structure and function, biologists can begin to decipher how it may contribute to human disease.

To determine a protein's structure, structural biologists rely upon expensive and complicated X-ray crystallography or nuclear magnetic resonance (NMR) spectroscopy. Since each molecule has its own unique set of properties, researchers must use specific conditions to express, purify, and crystallize their targets. This makes high-throughput structure determination a trial-and-error process for determining the optimum conditions for each protein during each step.

In 1962, Max Perutz and John Kendrew shared the Nobel Prize in Chemistry for determining the first 3-D protein structures—those of hemoglobin and myoglobin. Over the past decade, large-scale structural genomics projects like the Protein Structure Initiative (PSI), which has received over \$1 billion in



Zhang has taken top honors at the last three Critical Assessments of Techniques for Protein Structure Prediction (CASP) competitions. His computational models are offering proteomics some quick answers about protein structure and function, and he's getting ready to unveil a new method for template-free protein structure modeling. Source: Yang Zhang

funding from the National Institutes of Health, have increased the number of determined protein structures by developing high-throughput methods. Today, the Protein Data Bank stores over 64,000 3-D protein structures from various organisms. But the number of possible proteins that may play a significant role in biology could reach beyond the millions.

To make a further dent in this protein universe, computational biologists are developing algorithms that can predict a protein structure. These calculations are based on how the atoms within these molecules interact with one another and the aqueous cellular environment. But these methods are far from replacing X-ray crystallography or NMR spectroscopy. They must become more accurate and provide higher-resolution structures. And that's the point of the CASP competitions.

The CASP organizers obtain determined protein structures from large-scale structural genomic projects—like the NIH's PSI—prior to publication in protein databases. Over several months, the year's targets are announced to competitors, starting with more basic structures and gradually increasing in difficulty. In 1994, there were 33 targets; in 2010, there were 196 targets.

The competition concludes with a meeting, limited to 200 participants, where the structures are revealed and the competitor's predictions assessed. During the meeting, the most successful predictors are invited to present their group's methods in an effort to share information and findings.

During the early CASP competitions, David Baker's group from the University of Washington in Seattle became the group to beat. Baker, a professor of biochemistry at the university, developed the Rosetta program, which computes the lowest energy structure of a particular protein to find the most stable structure for a molecule. His Rosetta@home program recruits the computing power from registered users' personal computers through the Internet to help with these calculations. Baker has also developed an online protein-folding video game—called FoldIt—that allows scientists and nonscientists alike to compete against one another to find the most stable protein structures.

New Players

At the 2006 CASP7 competition, newcomer Zhang's I-TASSER approach ended Baker's dominance. Zhang triumphed again at the 2008 CASP8 competition. "In 2010, he put in two entries like the Kentucky Derby, a paired entry, and he finished one and two," says Omenn. "The guy's really onto something."

Before being recruited for the Center for Computational Medicine and Bioinformatics at the University of Michigan in 2009, Zhang studied theoretical physics and bioinformatics at institutes around the world. He earned his doctorate in physics from the Central China Normal University and spent two years as a research fellow in the Physics Department of the Free University of Berlin, Germany. After a subsequent postdoctoral research in the Chinese Academy of Sciences, he joined the Jeffrey Skolnick Group at Danforth Plant Science Center and SUNY Buffalo, where he changed his research focus from theoretical physics to bioinformatics. In 2005, he became an assistant professor at the Center for Bioinformatics at the University of Kansas.

Zhang's I-TASSER program generates a 3-D model from a protein's amino acid sequence using multiple threading alignments and repeated assembly simulations. The program then infers the protein's function by comparing the predicted structure with previously determined structures in existing databases. The program even grades itself, providing users with an accuracy score on its predictions.

To advance the structure prediction field, Zhang has published the I-TASSER method and hosts an online server where the research community can generate their own structure and function models. It has swept the field; over 15,000 labs from 90 countries are using the server. "One of the major goals for us to develop a method is really a desire to serve the community," says Zhang, "so that people can use your method to study their own projects."

Even though the community has access to the I-TASSER server, Zhang continues to make better predictions than the other CASP participants. Omenn chalks this up to Zhang's intuitive nature when it comes to modeling these protein structures; he makes better decisions throughout the process that add up to better predictions than those of his competitors' models. "This is not simply a cookie-cutter process," says Omenn. "It's a combination of a powerful algorithm and making the right decisions."

And Zhang is constantly looking for new ways to improve his protein modeling methods. He downplays the competitive aspect of the CASPs, choosing to focus instead on the collaborative atmosphere of the community. "The goal is not to win the task; the goal is really to push ourselves to develop a better method," he says. "I really don't think that winning the test is that important, but I will be really proud if we have a method that could help solve the problem of protein folding."

Removing the training wheels


The CASP competition is most useful when the targets challenge the competitors, but Zhang's group finds most to be relatively easy to model. "In this last one, I think 70–80% were easy targets, so only about 20–30% were hard targets," says Zhang. "It's not just our lab; most labs will find them easy since they can all find the templates."

For the past 20 years, most computational protein structure determination methods—including Zhang's I-TASSER—have relied on comparative modeling, using templates from previously determined structures of proteins with similar amino acid sequences. If the target protein has a template, researchers use current methods to refine the template to improve its accuracy. If a template does not exist for a target protein, it's very hard to get an accurate and reliable structure prediction.


As more and more proteins are solved by X-ray crystallography or NMR spectrometry and placed in the Protein Data Bank, protein structure prediction via templates will continue to get easier. Ten years ago, almost any structure that the CASP

Recent Activity


Sign Up Create an account or **log in** to see what your friends are doing.



BioTechniques - Videos
55 people recommend this.



BioTechniques - Lab Grammys 2012: And the Winner for "Science Parody of the Year" Is...
18 people recommend this.



BioTechniques - Videos
49 people recommend this.

Facebook social plugin

organizers found was considered difficult because there were so few templates available to the community. Organizers these days, however, are having a difficult time identifying targets without templates. "We have to try to find more targets—harder targets—so that we can test this modeling measure," says Zhang. "That is part of the challenge for this competition."

Hard targets are the only way that Zhang and his competitors can develop better predictive methods. "It's engineering, not predictive science," says Zhang of the easy ones. "There are too many proteins that have been solved with templates, and there's not enough work on developing a method that doesn't use a template."

While older methods are good for template-based models, they are difficult to improve upon. Furthermore, they only highlight the field's dependency on templates to determine structures. "This tells us that we really don't know much about how a protein folds, how it interacts with water, and lots of other things involved," says Zhang.

But this may change in the near future. Zhang's lab has developed a new computer algorithm—called QUARK—that does not rely upon templates. QUARK constructs the 3-D protein model using only the amino acid sequence. The models are built from small fragments of 1–20 residues by replica-exchange Monte Carlo simulation under the guide of an atomic-level knowledge-based force field.

In the free-modeling category at the CASP9, the QUARK server came in first place. "We're excited about a new method we've developed and how it may help with these questions," says Zhang.

One versus 1000 proteins

Although CASP has helped the protein modeling community, they are still a long way off from making X-ray crystallization and NMR spectroscopy obsolete. But structure prediction might have real applications, and Omenn and Zhang are currently interested in exploiting predictive algorithms.

Omenn is interested in cancer proteomics. As a research associate at the NIH in the late 1960s, he studied under biochemist Christian B. Anfinsen, who won a Nobel Prize for demonstrating that amino acid sequence determines the structure and function of proteins. Since then, he has studied genetics at the University of Washington in Seattle, WA, has become a director of the pharmaceutical company Amgen, Inc., and has been involved with the Plasma Proteome Project of the Human Proteome Organization. In 2008, President Obama selected him as one of his campaign's science advisors.

Before becoming director of the UM Center for Computational Medicine and Biology, he was the executive vice president for medical affairs at UM and started to build a bioinformatics infrastructure at the university. "It was clear to me that we had to have ways of dealing with organizing, mining, modeling, and annotating the huge high throughput data output from genomics and proteomics, and other related fields."

Omenn believes that genome-wide association studies—which compare the genomes of several individuals to identify genetic variation that may be associated with a predisposition for disease—have explained little about disease because the model is incomplete. For one thing, he says, they don't account for the environmental and behavioral factors that can influence phenotype. The combination of environmental and genetic factors can be studied only by the expression, modifications, and functions of proteins. Omenn's lab is particularly interested in protein isoforms that result from alternative splicing—in which RNA exons are reconnected in different ways during splicing—and that could become a new class of cancer biomarkers.

But proteomics faces a major dilemma: researchers can identify thousands of proteins in a specimen, knowing little about each one's structure and function. The alternative is for researchers to focus on a single protein, painting an intimate picture of how it interacts within the cell. "That's the traditional model, which requires a full lifetime from each investigator per protein," says Omenn. "That's one extreme. The other extreme is just a list."

So Omenn's hope, then, when he walked across the hall to Zhang's office that day in 2010, was that computational modeling of proteins could fill in the gap between these two extremes. Indeed, it could. In two days' time, Zhang had a model for the PKM2 variants that showed not only the structural changes around the exon swap site but also at other sites far away from the exon swap site. One structural change in the splice variant was the location of the ATP-binding site, which is involved in intracellular energy transfer and essential for cellular metabolism.

Omenn and Zhang, along with their colleagues Rajasree Menon and Srayanta Mukherjee, have just submitted a paper together on three other pairs of proteins with splice variations, using computational modeling to infer both structural and functional changes. "This is a hot new development that might actually be rather useful," says Omenn. "The PKM example could be tremendously important."

The endgame

The endgame for Zhang is to reverse the protein structure prediction process. Instead of extrapolating structural information from an existing amino acid sequence, he wants to be able to create an amino acid sequence that would produce a desired 3-D structure and function. This would open a new door to designing proteins for therapeutic applications.

Most drugs attempt to inhibit or activate a particular protein that is involved with a disease. Knowing the structure of that target protein would help researchers develop a drug that can bind tightly to the protein. Without the structure, developing a drug is like shooting in the dark. And current computational modeling cannot help because it cannot achieve the proper accuracy or resolution. "There's only a very small portion—just a tiny percent—of cases where the computational prediction of the structure can be useful for drug design," says Zhang.

"Drug design is a very difficult business," says Omenn. Pharmaceutical companies have spent billions of dollars and decades of research attempting to find new drugs. Even when computational approaches can provide rational drug designs, these drugs still may not have a selective effect. For example, a target for a particular protein in the kinase family—proteins that modify other proteins via phosphorylation—may have effects on other substrate proteins. And protein structure can change as the protein moves between different environments, such as from the nucleus to the cell membrane.

While computational modeling will not design any drugs in the near future, Zhang continues to further our understanding of protein folding through his algorithms. "The problem is not yet solved although progresses have been consistently claimed,"

says Zhang. "The model that you use is only as good as the template, and we are unable to fold big proteins without using templates, which are hurdles we have to overcome."

Additional reporting by Suzanne Winter.

Keywords: [protein structure](#) [computational biology](#) [proteomics](#)

[submit papers](#) |

[permissions](#) |

[terms & conditions](#) |

© 1983-2010 BioTechniques

[submit covers](#) |

[sitemap](#) |

[contact us](#) |

[reprints](#) |

[subscriptions](#) |

[adv](#)

[privac](#)