- Bloche MG (2004) Race-based therapeutics. *N Engl J Med* **351**: 2035–2037
- Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA (2003) Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nat Genet* **33**: 518–521
- Carson P, Ziesche S, Johnson G, Cohn JN (1999) Racial differences in response to therapy for heart failure: analysis of the vasodilator-heart failure trials. Vasodilator-Heart Failure Trial Study Group. *J Card Fail* **5**: 178–187
- Cohn JN *et al* (1986) Effect of vasodilator therapy on mortality in chronic congestive heart failure. Results of a Veterans Administration Cooperative Study. *N Engl J Med* **314**: 1547–1552
- Cohn JN *et al* (1991) A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. *N Engl J Med* **325**: 303–310
- Couzin J (2004) Consensus emerges on HapMap strategy. *Science* **304**: 671–673
- Duster T (2005) Race and reification in science. *Science* **307**: 1050–1051
- Kahn J (2004) How a drug becomes 'ethnic': law, commerce, and the production of racial categories in medicine. *Yale J Health Policy Law Ethics* **4**: 1–46
- Lazarou J, Pomeranz BH, Corey PN (1998) Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. *J Am Med Assoc* **279**: 1200–1205
- Lee SS (2003) Race, distributive justice and the promise of pharmacogenomics: ethical considerations. *Am J Pharmacogenomics* **3**: 385–392
- Lee SS (2005) Racializing drug design: implications of pharmacogenomics for health disparities. *Am J Public Health* **95**: 2133–2138
- Rotimi CN (2004) Are medical and nonmedical uses of large-scale genomic markers conflating genetics and 'race'? *Nat Genet* **36** (Suppl): S43–S47
- Satel S (2002) I am a racially profiling doctor. *The New York Times Magazine*, May 5 <http://www.nytimes.com/pages/magazine>
- Tate SK, Goldstein DB (2004) Will tomorrow's medicines work for everyone? *Nat Genet* **36** (Suppl): S34–S42
- Taylor AL *et al* (2004) Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. *N Engl J Med* **351**: 2049–2057
- Tishkoff SA, Kidd KK (2004) Implications of biogeography of human populations for 'race' and medicine. *Nat Genet* **36** (Suppl): S21–S27
- Wilson JF, Weale ME, Smith AC, Gratrix F, Fletcher B, Thomas MG, Bradman N, Goldstein DB (2001) Population genetic structure of variable drug response. *Nat Genet* **29**: 265–269

Katrin Weigmann

doi:10.1038/sj.embo.7400654

Into the fold

Advances in technology and algorithms facilitate great strides in protein structure prediction

More than 40 years ago, US biochemist Christian Anfinsen published his thermodynamic hypothesis of protein folding: that the native, or tertiary, structure of a protein could, in principle, be determined from its amino-acid sequence alone (Anfinsen *et al*, 1961). On the basis of his theory that the conformation is in a thermodynamic equilibrium in its cellular environment, Anfinsen concluded that to determine protein structure, all one had to do was to calculate the sum of all interatomic interactions in each possible conformation, and find the one with the least internal energy. This revolutionary theory was exciting for biologists, and earned Anfinsen the 1972 Nobel Prize in Chemistry with Stanford Moore and William Stein. In practice, however, it is not that simple. Calculating the energy of all possible conformations—even for a small peptide—requires such colossal computing power that it still overwhelms the most sophisticated supercomputers. But recent advances, both in raw computational power and the development of algorithms to predict protein structure, are at last enabling Anfinsen's hypothesis to be realized *in silico*.

Computation is fast becoming as essential for biology as it has been for theoretical and high-energy physics. The rapid increase in computing power and new methods for performing algorithms on large distributed computer networks (Hadley, 2004) could enable biologists to determine the native form, and ultimately the function, of large numbers of proteins. Yet even with the massive supercomputing power available today—through machines such as IBM's Blue Gene or grids like Rosetta@home—the goals of structural genomics remain daunting.

Although most of the sequences in the GenBank DNA database encode proteins that have close homologues with known structures, thereby making it easier to derive their structure computationally, the

remainder still pose a considerable challenge. Even if only 1% of newly discovered protein sequences have no close homologues with known structures, there will be approximately 300,000 sequences released each year with no structural information—a situation that cannot be resolved using X-ray diffraction or other physical methods alone. At the same time, *in silico* methods are not yet able to achieve the higher resolutions necessary (1.5–2 Å) to specify the location and orientation of individual atoms. For average-sized proteins with no homologues of known structure, computer-based structure prediction methods range from 3 Å to 5 Å.

...even with the massive supercomputing power available today...the goals of structural genomics remain daunting

Early attempts to calculate protein structure were hindered by limited computing power and by the inherent difficulties in assessing thermodynamic equilibrium. Achieving equilibrium involves jumping over local energy minima to reach the global minimum—just as a rock tumbling down a mountainside may bounce out of holes and hollows before falling to the valley floor. The apparent contradiction between the need for a protein to fold in finite time and the almost infinite number of possible conformations that the protein must search through to find the lowest energy structure is referred to as the Levinthal paradox. Cyrus Levinthal suggested that the native protein state might have a higher energy value than the theoretical free-energy minimum if the latter is not kinetically accessible—equivalent to the rock becoming permanently lodged in a gully halfway up the mountain.

In practice, however, proteins have ways of overcoming such local minima, in some cases with the help of chaperones, so

STRUCTURE PREDICTION METHODS

Structure prediction methods can usually be grouped into one of three general categories, although recent (and successful) methods may combine more than one approach.

Comparative, or homology, modelling

If a protein sequence (the target) has a close homologue with a known structure (the template), its own structure can be easily determined. The quality of the model is directly proportional to the degree of similarity between the target and template sequences.

Fold recognition

Even when a target sequence has no homologue of known structure, it may adopt a known fold. Also known as 'threading', this approach attempts to fit (or 'thread') the target sequence on to each structure in a library of known folds (the templates), and assesses the suitability of each fit by calculating its energy. The fit with the lowest energy is predicted to be the native structure.

Ab initio

Meaning 'from the beginning', *ab initio* methods rely on molecular dynamics calculations to determine the lowest-energy conformation. This is the only approach suitable for determining the structure of proteins with new topology.

computational methods also need to find a path to the global minimum. One of the key milestones in accomplishing this was the development of the Monte Carlo with Minimization method (MCM; Li & Scheraga, 1987) by Harold Scheraga, a pioneer in the field, at Cornell University (Ithaca, NY, USA). Unlike previous methods, MCM does not stop when a minimum-energy configuration has been reached, but continues evaluating the next best configurations until they no longer lead to lower energy levels. This allowed computers to 'climb out' of local minima, to mimic the actual folding process.

As early as 1976, Scheraga attempted to simulate protein folding *in silico*, in a seminal paper that set the scene for future work (Go & Scheraga, 1976). "If you consider the computational power available at the time and compare it with the power we now know is needed to fold a protein, his idea was truly visionary," commented Wolfgang Wenzel, a leading theorist on protein folding at the Institute for

Nanotechnology at the Research Centre Karlsruhe, Germany.

Indeed, at that time, detailed protein structures could only be obtained by X-ray crystallography and electron microscopy. But these methods cannot keep pace with the quantity of new sequences from recent genome projects, and also have technical limitations. High-quality crystals are not always possible to obtain; in particular, membrane-bound proteins are resistant to crystallization, as they tend to be dynamic. For example, G-protein-coupled receptors alter their shape to transduce signals into cells. In such cases, it is almost impossible to obtain crystals of the protein in every intermediate state—computational methods are the only viable option to determine their structure.

Scheraga and colleagues thus developed the United Residue (UNRES) technique for predicting structures, based on molecular dynamics (MD), in which Newton's equations of motion are solved for individual molecules as they progress towards the minimum-energy state (Liwo *et al*, 1997). So far, MD-based approaches are the only methods that analyse the dynamics of folding over time, whereas other methods bypass intermediate steps on their way to the structure with minimum energy. The UNRES method makes the assumption that each amino-acid residue in a protein chain interacts through only two points, one in the centre of the peptide bond and the other in a side chain represented as an ellipsoid. This approach enabled the prediction of larger protein structures, but, according to Wenzel, the ellipsoidal assumption limited the detail of the predicted interactions. As a result, UNRES has failed to attain high-resolution structures, although few other MD-based methods have fared better.

Recently, there has been considerable excitement in the protein structure community over a paper published by David Baker and colleagues at the University of Washington (Seattle, WA, USA) in September 2005, reporting that the prediction programme ROSETTA has been refined to calculate the structures of 17 small proteins comprising up to 100 amino acids with atomic-level accuracy (Bradley *et al*, 2005). The authors added a second stage to the existing ROSETTA algorithm, to represent the amino-acid side chains and the atomic forces acting in them, at the same level of accuracy as used for the

backbone atoms. Most researchers in the field have acknowledged that this is a step towards realizing Anfinsen's hypothesis for larger proteins, while pointing out that further progress is still required.

"Refinement to high resolution is still an open problem," pointed out Andrew Kolinski, Professor of Chemistry and Head of the Laboratory of Theory of Polymers at the University of Warsaw, Poland. "Baker's results are encouraging, although this is just one example, and the computational cost was extreme so it is not a road to genomic-scale refinement." Janus Bujnicki, Head of the Laboratory of Bioinformatics and Protein Engineering at the International Institute of Molecular and Cell Biology in Warsaw, while hailing Baker's work as a significant achievement, is also waiting to see confirmation. "I would rather wait until the rate of success and the average accuracy of predictions made by the new version of ROSETTA are validated in the course of the 'blind' prediction contest, the Critical Assessment for Protein Structure Prediction [CASP]." In CASP, held every two years since 1994, participants are given the sequences of proteins for which structures have already been determined experimentally—but kept secret—with the aim of determining the structures computationally. Methods are grouped according to whether they rely on homologous sequence or structural information, or calculate conformation from atoms alone (see sidebar). CASP assessors then rate their success to find the most accurate method in each category for predicting the structure of a protein from its primary sequence.

Although CASP assesses the performance of these methods, it is not sufficiently comprehensive to gauge their accuracy and reliability when predicting structures of newly discovered proteins. Consequently, there is a need for methods that predict structures in a reproducible way when using random starting assumptions, as the aim is to avoid the need to use experimental methods for confirmation. Furthermore, given the computational costs involved, it should be possible to predict in advance that the correct conformation will be reached, before committing resources. Progress has been made recently by Wenzel and colleagues with the reproducible folding of the three-helix headpiece of the HIV accessory protein (Herges & Wenzel, 2005).

As Wenzel noted, two things must happen to make folding predictable. First, the process must be reproducible and must achieve the same structure repeatedly given varying starting assumptions or different trajectories through folding space. Second, there must be a criterion for judging that the predicted structure is sufficiently close to the native structure in the absence of experimental results. "For us this criterion is free energy. If we reproducibly reach the lowest-energy conformation, then Anfinsen's hypothesis stipulates that the conformation we have reached is the native state," said Wenzel.

...there is a need for methods that predict structures in a reproducible way when using random starting assumptions...

Wenzel's method, based on an analysis of the free energy in a system driving the folding process, complements two main approaches: MD and the heuristic methods embodied in ROSETTA, which identify possible structures and then refine the most probable candidates to locate the energy minimum. Wenzel's method is faster than methods based on MD, but does not provide the dynamic information to reconstruct the folding pathway. Combining the two approaches, said Wenzel, might allow proteins to be folded reproducibly, which MD alone cannot do, while reconstructing the dynamic information about folding, which is only possible with MD at present. At the same time, free-energy methods can be used to discriminate between structures predicted by heuristic methods, to identify the native conformations. As Kolinski commented, the most significant advances in protein structure prediction have been made by combining several methods.

Although identifying the structures of all proteins would mean enormous progress for biology, it would not necessarily answer many questions relating to disease. For that, the folding process itself must be understood. Many diseases result from protein misfolding or the failure of proteins to denature when their time is up—in both cases, proteins develop when they should not. Some diseases are also caused by folding that takes place correctly but at the wrong speed: for instance, a

hereditary form of emphysema develops when a mutation slows down the normal folding process.

The folding process is similar to a maze with many dead ends and alternative routes leading to one exit: the centre of minimum energy. Finding the most probable route can only be determined by considering several paths in a computer model. According to Ajay Royyuru, Senior Manager of the Computational Biology Center at IBM's Thomas J. Watson Research Center (New York, NY, USA), this requires huge computing power to simulate the large number of folding events needed to obtain statistically significant predictions of where the folding process will go next.

"You need a statistical ensemble of conformational states," said Royyuru. At each stage of the fold there will be a cluster of options for the next step, and it is only after several simulations, perhaps 100, that it becomes clear which direction is the most relevant to the folding process. "Only then can we progress to the next step," said Royyuru. Not surprisingly, even simulating a single folding trajectory is a large computational task; multiple simulations require much more power, but at least they can be done in parallel. For this reason, protein-folding simulation can be distributed among unconnected computers, such as Rosetta@home.

Researchers also need massive parallel computing power to study the folds of specific proteins. This was the motivation for IBM (White Plains, NY, USA) to build its Blue Gene supercomputer. Europe's largest Blue Gene installation, at the John von Neumann Institute for Computing at the Research Centre Jülich, Germany, has made significant progress in the folding field with a string of recent publications. A team under Ulrich Hansmann is using Blue Gene to study protein misfolding in disease pathogenesis. They observed that a particular peptide has a tendency to form a β -strand in the vicinity of other β -strands, when otherwise an α -helix would be formed (Peng

...the most significant advances in protein structure prediction have been made by combining several methods

& Hansmann, 2003). This suggests how an incorrectly folded protein may induce misfolding in other proteins in their vicinity, as appears to happen in several neurological diseases, including prion diseases. Hansmann hopes that his recently acquired Blue Gene supercomputer will enable him to scale up this work to study medically relevant proteins.

Scheraga has also made significant progress on the folding problem. An ideal folding model would simulate the changing energy states of all atoms in both the protein itself and the surrounding solvent to match real conditions, but this is not computationally feasible because of the huge number of degrees of freedom involved. Scheraga has therefore applied Langevin dynamics to UNRES, to incorporate the extra degrees of freedom within Newton's equations of motion (Liwo *et al.*, 2005). This approach has been credited with opening a new era in protein-folding simulation, by beginning to answer fundamental questions such as whether a protein collapses before the appearance of the α -helices and β -sheets or vice versa, or if both events coincide. Scheraga's work suggests that it might be possible to dismiss the second option, at least for the proteins he has studied: he has found that collapse either precedes

or coincides with the appearance of secondary structures.

Until recently, however, the main successes of computational structural biology have been in understanding protein-to-protein docking, rather than working out native conformations using molecular dynamics. "Docking's a much easier problem than prediction from the initial sequence because you already know the structure of the proteins and you just have to fit them together," noted Baker. He and colleagues applied ROSETTA to study the docking of the three constituent proteins of the anthrax toxin, and achieved reliable models of the resulting complex (Lacy *et al*, 2005). However, as Wenzel pointed out, current methods cannot fully resolve protein-to-protein docking in large complexes because this requires greater computational power than is available to allow for backbone flexibility in the peptide chains. "It is obvious that protein flexibility is a key to the puzzle since docking implies changes in function that are mediated by changes in structure," he explained.

At present, it is possible to simulate the folding of small proteins, and the docking of larger ones, at close to the resolution of experimental methods, while progress has been made in visualizing aspects of the folding process itself. Further progress towards the ambitious goal of structural genomics will probably proceed through initiatives such as the World Community Grid, which supports the Human Proteome Folding Project in its

bid to generate low-resolution structural models for the human proteome. This will provide clues about the structure and function of individual proteins, providing the basis for more accurate models through a combination of experiment, modelling and refinement.

...the main successes of computational structural biology have been in understanding protein-to-protein docking, rather than working out native conformations...

After a lifetime of work, Scheraga at last believes that a solution is in sight. "I hope that the protein folding problem will be solved in the next few years," he said. For this reason, he suggests that aspiring biologists should expand their interests beyond the folding problem itself. "I think they would be better advised to turn to other problems if they don't want to be scooped in the middle of their research."

Still, computational structural biology will remain a fertile field. Kolinski believes that the focus of computational structure prediction will shift towards large macromolecular assemblies, and then to molecular models of simple metabolic or regulatory pathways. In this way, structure prediction may eventually converge with systems biology, which is concerned more with state than with form.

REFERENCES

- Anfinsen CB, Haber E, Sela M, White FH Jr (1961) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. *Proc Natl Acad Sci USA* **47**: 1309–1314
- Bradley P, Misura KM, Baker D (2005) Toward high-resolution *de novo* structure prediction for small proteins. *Science* **309**: 1868–1871
- Go N, Scheraga HA (1976) On the use of classical statistical mechanics in the treatment of polymer chain conformation. *Macromolecules* **9**: 535–542
- Hadley C (2004) Biologists think bigger. *EMBO Rep* **5**: 236–238
- Herges T, Wenzel W (2005) *In silico* folding of a three helix protein and characterization of its free-energy landscape in an all-atom force field. *Phys Rev Lett* **94**: 018101
- Lacy DB, Lin HC, Melnyk RA, Schueler-Furman O, Reither L, Cunningham K, Baker D, Collier RJ (2005) A model of anthrax toxin lethal factor bound to protective antigen. *Proc Natl Acad Sci USA* **102**: 16409–16414
- Li Z, Scheraga HA (1987) Monte Carlo-minimization approach to the multiple-minima problem in protein folding. *Proc Natl Acad Sci USA* **84**: 6611–6615
- Liwo A, Khalili A, Scheraga HA (2005) *Ab initio* simulations of protein-folding pathways by molecular dynamics with the united-residue model of polypeptide chains. *Proc Natl Acad Sci USA* **102**: 2362–2367
- Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA (1997) A united-residue force field for off-lattice protein-structure simulations. I. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. *J Comput Chem* **18**: 849–873
- Peng Y, Hansmann UH (2003) Helix versus sheet formation in a small peptide. *Phys Rev E Stat Nonlin Soft Matter Phys* **68**: 041911

Philip Hunter

doi:10.1038/sj.embo.7400655