cal bases of discrimination, such as total stimulus area. And finally, there has been no convincing demonstration that infants represent the ordinal relations among sets of 1, 2, and 3 elements.

How might nonhuman animals and prelinguistic infants represent number? Two classes of models for nonlinguistic numerical representational systems have received empirical support: object file models and analog magnitude models. In the object file model, the infant or monkey forms a representation with one symbol for each individual in the set and compares representations by computing one-to-one correspondences between sets. Such representations are limited to the number of individuals that can be held in short-term memory at any one time, which is 3 or 4. These representations contain no symbols that function as numerals, and there is no counting process. In analog magnitude models, number is represented by a continuous quantity, akin to a number line. Representations are compared by the same sorts of operations that compare lengths, durations, volumes, and other representations of continuous quantities. The process by which the analog magnitude is incremented for each item in the set is equivalent to counting (*6*), but analog magnitude models differ in many ways from integer list models (*7*).

Brannon and Terrace's data favor an analog magnitude model. Their monkeys represent numbers that exceed the limits of the object file model. Further, analog models correctly predict that number comparisons become easier when the differences between the numbers are greater (the distance effect). By contrast, for infants the evidence favors the conclusion that the object file model underlies the prelinguistic numerical representations in the events studied to date (*7*). There is also considerable indirect evidence that the integer list symbolic representation of number is built from object file representations, and not from analog magnitude representations (*4*), even though human adults certainly use the latter as well (*5*).

The upshot is that one evolutionary source of human number representation—the analog magnitude representations that Brannon and Terrace most probably are tapping in primates—is not the primary ontogenetic source of human symbolic number list representations, either in linguistic evolution or in individual development. Although this conclusion is controversial, our challenge is clear. We must specify the nature of nonlinguistic representations of number (there may be many) and characterize the process by which explicit symbolic representations are constructed in the history of each culture and again by each child.

References

1. E. M. Brannon and H. S. Terrace, *Science* **282**, 746 (1998).
2. K. Wynn, *Math. Cognit.* **1**, 35 (1995).
3. ———, *Cogn. Psychol.* **24**, 220 (1992).
4. J. R. Hurford, *Language and Number* (Oxford Univ. Press, Oxford, 1987).
5. S. Dehaene, *The Number Sense* (Oxford Univ. Press, Oxford, 1997).
6. C. R. Gallistel, *The Organization of Learning* (MIT Press, Cambridge, MA, 1990).
7. C. Uller, S. Carey, G. Huntley-Fenner, L. Klatt, *Cogn. Devel.*, in press.

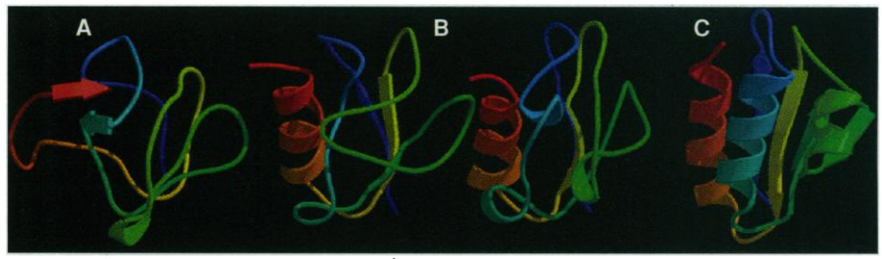
PERSPECTIVES: PROTEIN FOLDING

A Glimpse of the Holy Grail?

Herman J. C. Berendsen

The prediction of the native conformation of a protein of known amino acid sequence is one of the great open questions in molecular biology and one of the most demanding challenges in the new field of bioinformatics. Using fast programs and lots of supercomputer time, Duan and Kollman (*1*) report on page 740 of this issue that they have successfully folded a reasonably sized (36-residue) protein fragment by molecular dynamics simulation into a structure that resembles the native state. At last it seems that the folding of a protein by detailed computer simulation is not as impossible as most workers in the field believe. Or is this an overoptimistic view?

With the number of known gene sequences increasing at an accelerating pace (the complete genomes of 13 bacteria and of yeast are now known, the first multicellular animal will follow soon, three plants and the fruit fly are in the pipeline, and the human genome sequence can be expected at the beginning of the next century), the quest for the structure and function of the coded proteins becomes pressing. The obvious route to that goal is by homology modeling: use as much information as you can get from the database of known structures. But at the present level of sophistication, such methods are effective for only about 25% of the proteins for which the amino acid sequence is known; if sequence homology drops below 25%, the reliability of database-oriented methods drops to nearly zero.



Unfolding is easier than folding. Four snapshots from the simulation of an unfolding protein called HPr (a phosphate-transferring protein): **(A)** native conformation, **(B)** partly unfolded conformations that still contain most of the secondary structure, and **(C)** an unfolded (or randomly folded) structure.

Still, most small proteins fold spontaneously in seconds into their native conformations; secondary structure elements like α helices or β turns fold in tens of nanoseconds to microseconds. Such folding is thermodynamically downhill and is just a result of the physical interactions between atoms, including those of the solvent. These atomic interactions are elementary and well-known, so why can't we use this knowledge to mimic the native folding process? Well, for two reasons: First, existing computers cannot sample enough configurations in a reasonable time to come up with the thermodynamically stable native structure; second, we are not too sure that the available force field descriptions, which we need to compute the energy of each configuration, are accurate enough to come up with a reliable free energy of a conformation. The trouble resides in the enormously large positive and negative contributions that nearly cancel in the computation of the total energy.

The sampling problem can be summarized as Levinthal's paradox: If we assume three possible states for every flexible dihedral angle in the backbone of a 100-residue protein, the number of possible backbone configurations is 3^{200}. Even an

The author is in the Department of Biophysical Chemistry and the BIOSON Research Institute at the University of Groningen, Netherlands. E-mail: berendsen@chem.rug.nl

CREDIT: K. ANTON FEENSTRA, GRONINGEN

incredibly fast computational or physical sampling in 10^{-15} s would mean that a complete sampling would take 10^{80} s, which exceeds the age of the universe by more than 60 orders of magnitude. Thus, real proteins fold in a more clever way than by random trial, presumably by specific pathways and starting at specific nucleation patterns, and some information on the pathway must be present in unfolded states as well. Computers should try to do it even more cleverly because at the present state of the art detailed molecular dynamics simulations of proteins including explicit solvent cover real times on the order of 10 ns. Although in 10 years time, this will increase to microseconds, simulations will still be six orders of magnitude short of reality, which must be bridged by methodological simplifications.

Simplifications abound, but they are ineffective. Most effort has gone into the use of lattice models; with residues only allowed on regular lattice sites, these models are caricatures of the real world. They can—and are often meant to—teach us principles of folding, but they yield no solutions to real folding problems. The required properties of the free energy landscape for folding have been extensively discussed (*2*), and several rules have been formulated. However, in a thoughtful article on the folding of a simplified protein-like model, Crippen and Ohkubu (*3*) have shown most of these rules to be inadequate. Simplified force fields have been invented in variety: elimination of solvent, reduction of each residue to a few pseudoatoms, and hamiltonians derived from the database of folded structures. But structural aspects are extremely sensitive to details of force fields [in one example from our laboratory (*4*), we found that a specific fold of a tetrapeptide in water, observed by nuclear magnetic resonance, could only be reproduced by simulation with one popular model for water and not with another, slightly different, but equally popular model], and it is unlikely that reduced force fields can come up with the required precision. No one knows how models that are precise enough can be applied to short-cut the folding process such that available computational power suffices to reach the desired goal. Many despair: The application of force field–based methods in the "critical assessment of methods of protein structure prediction" (CASP) contest tends to worsen rather than improve predictions (*5*).

Thus, one of the "grand challenges" (*6*) of high-performance computing—predicting the structure of proteins—acquires much of the flavor of the Holy Grail quest of the legendary knights of King Arthur: It is extremely desirable to possess but extremely elusive to obtain.

Now Duan and Kollman have not only succeeded in applying molecular dynamics simulations to a solvated protein (small, but still 12,000 atoms) over a full microsecond but also saw the chain fold during 150 ns into a compact structure resembling the native state (known from nuclear magnetic resonance). It then unfolds again and refolds for a shorter period toward the end of the simulation. Apart from computational details, their molecular dynamics method is quite standard and, as far as the treatment of long-range interactions is concerned, even somewhat below standard. This gives hope that "brute force dynamics" can go a long way toward protein folding in the future.

We should be careful, however. Folding to the stable native state has not (yet) occurred, and the simulations do not contain any relevant statistics on the process. The real protein will fold and refold hundreds to thousands of times until it stumbles into the stable conformation with lowest free energy. Because this hasn't happened (and couldn't happen) in the simulations, we still cannot be sure of the full adequacy of the force field.

A prudent approach to the simulation of folding would be to choose a simpler system on which sufficient statistics can in fact be obtained. This is just what Daura *et al.* (*7*) at the Eidgenössische Technische Hochschule, Zürich, have recently done. These authors studied the folding of a β-heptapeptide in methanol (5400 atoms, helical structure in the native state) over a total of 0.25 μs in several simulations and established a statistical folding-unfolding equilibrium at various temperatures. Such simulations allow precise comparison with experiment and validation of force fields.

Molecular dynamics simulation is back in place on the road toward protein folding. Improving its physical and computational performance is worth the trouble, but for the time being we also need to augment the ab initio physics with all the experimental knowledge we can lay our hands on to unravel the protein-folding problem.

"The Grail had many different manifestations throughout its long history, and many have claimed to possess it or its like" (*8*). We might have seen a glimpse of it, but the brave knights must prepare for a long pursuit.

References

1. Y. Duan and P. A. Kollman, *Science* **282**, 740 (1998).
2. See the review by C. L. Brooks III, *Curr. Opin. Struct. Biol.* **8**, 222 (1998), and references therein.
3. G. M. Crippen and Y. Z. Ohkubu, *Proteins* **32**, 425 (1998).
4. D. Van der Spoel, A. R. van Buuren, D. P. Tieleman, H. J. C. Berendsen, *J. Biomol. NMR* **8**, 229 (1996).
5. For CASP see *Proteins* (Suppl. 1) (1997); C. Venclovas, A. Zembla, K. Fidelis, J. Moult, *ibid.*, p. 7 (1997).
6. *Grand Challenges 1993: High Performance Computing and Communications*, The FY 1993 U.S. Research and Development Program (supplement to the President's Fiscal Year 1993 Budget).
7. X. Daura, B. Jaun, D. Seebach, W. F. Van Gunsteren, A. E. Mark, *J. Mol. Biol.* **280**, 925 (1998).
8. J. Matthew, *The Grail, Quest for the Eternal* (Thames and Hudson, London, 1981), p. 72.

NOTA BENE: TRIPLET REPEAT DISEASES

Innocent Inclusions

In a curious set of neurodegenerative diseases, a long string of the nucleotide triplet CAG lodges within genes, causing the death of subsets of neurons and ultimately disease. Exactly how these strings of repeats cause cell death is not known, but they do not simply disrupt the function of their target gene. Rather, the long CAG string has a deadly—but undefined—effect of its own.

One popular idea is that the CAG repeats cause the protein to form a toxic aggregate in the nucleus of cells. These so-called nuclear inclusions are common in the brains of patients with these disorders. But in two recent papers in *Cell*, this explanation is called into question. One group shows, in a cultured cell model system for Huntington's disease (*1*), that cells may die even without the presence of nuclear inclusions. In the most dramatic experiment, expression of a fragment of the mutant huntingtin protein containing a 68-repeat insertion, together with an inhibitory form of the ubiquitin-conjugating enzyme, resulted in far fewer intranuclear inclusions. The mutant huntingtin actually triggered more cell death in this situation than it would have in the presence of inclusions, leading the authors to the bold suggestion that the inclusions may actually be protective. A second group made transgenic mice that mimicked the disorder spinocerebellar atrophy type 1 (*1*), in which the repeat-containing protein ataxin-1 lacked a self-aggregating region. These mice had no nuclear inclusions, but still showed the characteristic degeneration of cerebellar Purkinje cells. The field may now have to look elsewhere for the mechanism by which these repeats do their damage to the cell.

–Katrina L. Kelner

References

1. F. Saudou *et al.*, *Cell* **95**, 55 (1998); A. Klement *et al.*, *ibid.*, p. 41.