Volume 307, number 1, 10–13 FEBS 11317 July 1992

Minireview

Concepts in protein folding

David J. Thomas

European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, D-6900 Heidelberg, Germany

Received 18 May 1992

Certain concepts and misconceptions in the field of protein folding are discussed from the viewpoint of a theoretical physicist. It is argued that there can be no protein folding code and that perceived correlations between sequence or composition and three-dimensional structure are more likely to be an artefact of a limited database than a real result. Attempts at using molecular dynamics algorithms are also likely to produce artefactual results because results depend critically on the unknown hamiltonian energy function. Correct calculations of configurational entropy are thought to be the most likely next step in understanding how and why proteins fold.

Hamiltonian energy calculation; Configurational entropy; Protein folding; Molten globule; High polymes

1. SOLVING THE PROTEIN FOLDING PROBLEM

In the strict sense of being the successful prediction of the three-dimensional structure of a protein from its amino-acid sequence, the 'Protein Folding Problem' is as much a problem for theoretical physicists as for molecular biologists. Theoretical physics is not, however, a particularly approachable field. The aim of this minireview is, therefore, to try to familiarize the more biologically oriented reader with some of the concepts dominant in a physicist's view of protein folding. This is important not just as a matter of general interest, but because resolving the protein folding problem will demand extensive collaboration and understanding between scientists as disparate as physicists, biologists, chemists and mathematicians.

2. PROTEIN FOLDS ARE NOT CODED

The genesis of the protein folding problem can be traced to early observations of the reversibility of the denaturation of proteins [1–3], though it could not be cast into its presently understood form before the discovery that proteins have a definite sequence 'coding for' a more or less well-defined three-dimensional structure. Much of the modern literature makes an unsupported conceptual jump from these observations and from the existence of a genetic code to the idea that there exists a 'Protein Folding Code'. But 'code' is a term with a simple and specific meaning. It would imply that local sequences would specify local structures simply and uniquely. This is in clear disagreement with observation. It would be equally wrong to infer, instead, the existence in the sequence of 'encrypted' data determining the fold, for much the same reason of implying too high a degree of determinacy and the corollary of well-defined 'decryption' machinery. There is little sense in such a concept, since a protein molecule capable of unaided folding must be its own decryption machine, and every case becomes special. The view must be more that the sequence holds data describing the fold in neither a coded nor an encrypted form, but more simply that evolutionary pressure selects sequences which, following normal physical laws, fold into something useful. This is opportunism, not information-based determinism, and the real problem in protein folding must be unravelling the complexities of the physics. The solution itself, though based on diverse physical theories and much experimental evidence, must take the form of an algorithm.

3. INTERACTIONS AND STRUCTURAL TENDENCIES

Arguments that single amino acids have varying tendencies to form the extended or the helical conformation, and even that interactions between residues are unimportant, have been overstated in the literature [4] and have led to much misunderstanding. At first sight, this error appears to arise from underestimating the difference between the amorphous, relatively open structure of 'ordinary' high-polymers, where side-chain interactions are relatively non-specific, and the dense

Correspondence address: European Molecular Biology Laboratory, Meyerhofstraße 1, Postfach 10.2209, D-6900 Heidelberg, Germany.

 Published by Elsevier Science Publishers B.V.

well-packed and ordered structure of folded proteins, where side-chain interactions can be highly specific. Acceptance of this simplifying error and a vague sense of the sequence coding the structure have led to numerous data-based attempts on the protein folding problem [5–8], even though a straightforward statistical attempt to determine a relationship between sequence and secondary structure shows it to be indeterminate for lack of data in the forseeable future [9]. A knowledge of sequence variability alleviates this problem only partially [10]. The present author's (unpublished) work in this area shows explicity the degree to which the secondary structure is determined by physical interactions between pairs, triplets and higher multiplets of residues. As a rule, the result can be summarized by saying that higher multiplets have an extra effect which cannot be predicted from those of the underlying lower multiplets, but generally does not outweigh them, at least for the proteins in the database [11]. This observation accords well with physical intuition. The same analysis shows that some sequences have strong structural preferences, whilst others clearly do not, which facts are well known to aficionados of the field. Irritatingly, but perhaps inevitably, the portions of the sequence for which too few data exist to make a good prediction concentrate in the very regions that give proteins their specific characteristics.

4. THE CONSEQUENCES OF ENERGIES BEING INCALCULABLE

It is commonly asserted that the failure of algorithms to predict secondary structure is obvious because the folding of the backbone is 'context dependent', where context is taken to mean 'local spatial environment'. Unfortunately, an acknowledgement of this weakness is not the same as knowing how to do better. A recent paper aimed directly at molecular biologists [12] describes a relatively new and interesting attempt to broach this impasse. Apparently, it ties together approximated but respectable physics and database analysis using a technique known as "associative memory Hamiltonians" identifying tertiary as well as secondary structure. An earlier paper aimed at physicists presented the argument clearly [13], saying that the method is to "use the spatial statistics of a database of known structures to determine an energy function", i.e. the associative memory Hamiltonian. The consequences of this deduced energy function are then explored using a molecular dynamics (MD) algorithm. The real point here is that the true Hamiltonian (energy) of a protein is so unimaginably complicated that we have neither means nor hope of determining it properly. Accurate knowledge of it is, however, an absolute requirement for credible simulations of molecular dynamics. This popular technique has found many applications, and when constrained by structural data from X-ray or nuclear magnetic resonances (NMR) studies can be very useful, if unnecessarily expensive, since well-designed Monte Carlo (random trial) algorithms can achieve comparable results much more efficiently without giving the false impression that the dynamics of the molecule are understood [14]. It is now known that all MD algorithms could be seriously flawed for subtle but basic computational reasons even if the energy function were known exactly [15]. Not all of the reasons for the failings of present-day calculations of molecular dynamics are clear, however, though certain approximations likely to lead to problems, for example the neglect of electrostatic interactions beyond a certain distance, are being avoided in newer implementations running on massively parallel computers [16]. There seems to be no hope that three- and higher-body interactions can ever be accommodated, and the well-known argument that correct or even approximate calculations of molecular dynamics will always be too expensive to solve the protein folding problem still seems irrefutable.

5. STATISTICAL MECHANICS GIVES A CLEAR PICTURE

A bizarre feature of the literature is a recurrent deprecation of the obvious fallacy that a protein molecule achieves its final fold by performing an exhaustive search of all possible conformations. This is opposed by an even less helpful concentration on experimental 'proof' that proteins fold via a well-defined pathway instead, though this is a dangerously naive interpretation of macroscopic observations. Both of these extreme viewpoints are based on an inadequate appreciation of statistical physics [17]. In that language, a classical system in equilibrium must explore all possible states with probability dictated by the Boltzmann distribution, which decays exponentially fast with increasing energy. One molecule might take an infinitely long time to do this, and an artifice to achieve the same mathematical results is to average an infinite number of molecules for an infinitesimal time instead. But protein molecules are not in equilibrium anyway, so for them a finite time suffices to perform the necessary sampling of configurations; also, the vast majority of configurations are either inaccessible from the starting state or are energetically so unfavourable that they can be neglected. This leaves only accessible, dominantly low-energy conformations. Some of these conformations will lie in densely represented local minima of energy and would consequently be observable experimentally as kinetic traps. However, the presence of interconvertible (i.e. topologically mutually accessible) kinetic traps will lead to exactly the type of multiply connected routes of folding that have been observed [18]. This complicated interconverting maze is not described appropriately as a pathway in the biochemical sense, and attempting to redefine 'pathway' (like 'code') is to obfuscate.

6. EXPERIMENTS IN UNFOLDING CONDITIONS

Experimentally, protein folding is often studied in the reversed sense of unfolding native proteins. Objections have been raised to this [19], because it cannot be assumed without proof that the denaturation of a protein in strongly unfolding conditions proceeds by the same route at the microscopic level as folding in more normal conditions. Some authors claim that no problem exists, but cite as proof a dubious invocation of a principle of microscopic reversibility and experiments which, though novel, are incapable of resolving this question [20,21].

7. A LABILE STATE CAN STILL HAVE STRONG TENDENCIES TO THE NATIVE STATE

The chains of unfolded proteins are nowhere near as straight and open as is commonly imagined or illustrated [4,22,23], but claims that they follow a path typical of a random walk are also overstated, denying the so-called 'excluded volume' effects of chain-to-chain collisions. Ordinary high polymers can be observed in special conditions (called the theta-point) where the effects of these interactions cancel to a large degree, as far as some macroscopically observable properties are concerned, but a true theta-point is impossible for a protein because it would need different conditions at each point in the sequence. This means that partially unfolded proteins in the so-called 'molten-globule state' still display structural tendencies clearly related to their native structures, and it is unlikely that these tendencies ever vanish completely, even under strongly unfolding conditions. Molten globules themselves have become a dangerous concept lately (simply because they are far too popular), and there is a risk of not distinguishing clearly enough between true molten globules and more general lability, which may often be functional in normal physiological conditions.

8. THE MOST ADVANCED POLYMER PHYSICS HAS ITS LIMITATIONS

The polymer chain analogy has been studied intensively, often using the most advanced available mathematical models, like dimensional regularization and renormalisation group theory [23–27]. Dimensional regularization is a trick to extrapolate results evaluated in, say, four dimensions back to the everyday three dimensions where they are otherwise incalculable. It is a technique used widely throughout physics but a proof of its validity cannot be found [28]; the author believes this to be a strong indication that it violates the symmetry of nature. An absolute proof of its impossibility, on the other hand, would imply that there is an error in established physics, which is quite likely given the current plethora of insoluble problems. Renormalisation group theory is more complicated, but also more rigorous, and involves establishing relationships which remain valid if certain parameters (e.g. length scale) change. Calculations using these methods suggest that the transition from a collapsed globule to a so-called unfolded state may be thermally equivalent to a liquid–gas phase transition, but is of first order (i.e. with a latent heat) only for chains of infinite length [29]. For chains of finite length the calculated order is indeterminate between first and second (i.e. without a latent heat), which may be relevant to observations of the lack of a discernible activation barrier between the compact intermediate and the unfolded state. It has also been shown recently, using a functional integral approach (which approximates a finite random chain by an infinitely more detailed random curve), that in a presumably well-solvated high polymer the formation of any looped region tends to favour the formation of the next in a nearby location [30]. However, all of these advanced mathematical methods of polymer physics make approximations which are inappropriate in the special case of proteins, clearly limiting their applicability.

9. MELTING WITHOUT GETTING WET

The molten globule-to-native fold transition in proteins is generally held to be analogous to a (first order) liquid–solid phase transition, but what characterizes it most strongly is a large change in the specific heat capacity [31], which is directly proportional to the change in exposed apolar surface area [32], at least for soluble globular proteins. The volume of the molecule increases by ca. 15% going from the native fold to the molten globule. This increase in volume can be explained even in the absence of any extra interior water, being typical of a normal solid–liquid transition in which dense ordered packing gives way to dense disordered packing [33].

10. A ROLE FOR ENTROPY

There is a consensus growing that the molten globule is a general precursor [34], or even that the native fold might be a relic of the molten globule [35], but either way it is necessary to avoid creating a vitalistic view whereby the fold is deemed to be determined by highly specific inter-residue interactions at a time before they could possibly act. We are thus forced to conclude that whereas the fine details of the native fold may depend on such specific interactions, the fold at a more general level must be determined by something less specific. This means that a so-called 'coarse-grained' (i.e. lower resolution) picture is sought, which usually takes the form of a mean-field formalism [7,17,36–38], and preferably a self-consistent one [26]. Attempts to determine mean fields (which yield a sort of averaged force field)

typically produce the well-worn result for soluble globular proteins that hydrophobicity is the dominant force involved [39]. However, hydrophobicity is not a true force, and if we know anything about it, it is that it cannot be defined precisely [40]. Similarly, attempts to determine which forces are dominant in protein folding seem misguided [41], since most of the well-defined distinguishable contributions are of roughly the same strength. Folding is held to be a stochastic (i.e. random) process within the guiding mean-field, and as such must be subject to strong entropic effects. Configurational entropy must have an important effect, but we do not have a formula for it. Arguments about entropic loop tensions are only a very early step along this path [42].

11. THERE IS ONLY ONE WAY FORWARD

If the following sounds like a jest, forgive me: an irrefutable logician would say that the fact that we (who are made of proteins) exist proves that an algorithmic solution of finite complexity must exist because it is found easily enough by a finitely connected set of brainless particles. I agree, but first find your irrefutable logician! More seriously, we really should think very hard about the entropy of molecular systems. It is our only hope.

Acknowledgements: The author is grateful to Drs. Chris Sander, Andrew McLachlan FRS, Tom Creighton and Prof. Sir Sam Edwards FRS for helpful comments on the manuscript.

REFERENCES

[1] Michaelis, L. and Rona, P. (1910) Biochem. Z. 29, 494–500.
[2] Spiegel-Adolf, M. (1926) Biochem. Z. 170, 126–172.
[3] Anson, M.L. and Mirsky, A.E. (1931) J. Phys. Chem. 35, 185–193.
[4] Flory, P.J. (1988) Statistical Mechanics of Chain Molecules, Carl Hanser Verlag, Munich.
[5] Kaden, F., Koch, I. and Selbig, J. (1990) J. Theor. Biol. 147, 85–100.
[6] Selbig, J., Kaden, F. and Koch, I. (1992) FEBS Lett. 297, 241–246.
[7] Sippl, M.J. (1990) J. Mol. Biol. 213, 859–883.
[8] Bowie, J.U., Lüthy, R. and Eisenberg, D. (1991) Science 253, 164–170.
[9] Rooman, M.J. and Wodak, S.J. (1988) Nature 335, 45–49.
[10] Benner, S.A. and Gerloff, D. (1991) Adv. Enzyme Reg. 31, 121–181.
[11] Bernstein, F.C., Koetzle, T.F. Williams, G.J.B., Meyer Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535–542.
[12] Friedrichs, M.S., Goldstein, R.A. and Wolynes, P.G. (1991) J. Mol. Biol. 222, 1013–1034.
[13] Sasai, M. and Wolynes, P.G. (1990) Phys. Rev. Lett. 65, 2740–2743.
[14] Parak, F. and Knapp, E.W. (1984) Proc. Natl. Acad. Sci. USA 81, 7088–7092.
[15] Yee, H.C., Sweby, P.K. and Griffiths, D.F. (1991) J. Comp. Phys. 97, 249–310.
[16] Windemuth, A. and Schulten, K. (1990) Mol. Simulation 5, 353–361.
[17] Finkelstein, A.V. and Reva, B.A. (1991) Nature 351, 497–499.
[18] Creighton, T.E. and Goldenberg, D.P. (1984) J. Mol. Biol. 179, 497–526.
[19] Buchner, J. and Kiefhaber, T. (1990) Nature 343, 601–602.
[20] Matouschek, A., Kellis Jr., J.T., Serrano, L., Bycroft, M. and Fersht, A.R. (1990) Nature 346, 440–445.
[21] Sancho, J., Meiering, E.M. and Fersht, A.R. (1991) J. Mol. Biol. 221, 1007–1014.
[22] Treloar, L.R.G. (1975) The Physics of Rubber Elasticity, Clarendon Press, Oxford.
[23] Freed, K.F. (1987) Renormalization Group Theory of Macromolecules, Wiley-Interscience, New York.
[24] de Gennes, P.-G. (1972) Phys. Lett. A 38, 339–340.
[25] des Cloizeaux, J. (1980) J. Phys. (Paris) 41, 223–238.
[26] Edwards, S.F. (1965) Proc. Phys. Soc. London 85, 613–624.
[27] Wilson, K.G. and Fisher, M.E. (1972) Phys. Rev. Lett. 28, 240–243.
[28] Schäfer, A. and Müller, B. (1986) J. Phys. A: Math. Gen. 19, 3891–3902.
[29] Kholodenko, A.L. and Freed, K.F. (1984) J. Phys. A: Math. Gen. 17, 2703–2727.
[30] Chan, H.S. and Dill, K.A. (1990) J. Chem. Phys. 92, 3118–3135.
[31] Privalov, P.L. (1990) Crit. Rev. Biochem. 25, 281–305.
[32] Livingstone, J.R., Spolar, R.S. and Record Jr., M.T. (1991) Biochemistry 30, 4237–4244.
[33] Waldram, J.R. (1985) The Theory of Thermodynamics, pp. 173–175., Cambridge University Press, Cambridge.
[34] Ptitsyn, O.B., Pain, R.H., Semisotnov, G.V., Zerovnik, E. and Razgulyaev, O.I. (1990) FEBS Lett. 262, 20–24.
[35] Thomas, D.J. (1991) J. Mol. Biol. 222, 805–817.
[36] Bryngelson, J.D. and Wolynes, P.G. (1990) Biopolymers 30, 177–188.
[37] Garel, T. and Orland, H. (1988) Europhys. Lett. 6, 597–601.
[38] Garel, T. and Orland, H. (1988) Europhys. Lett. 6, 307–310.
[39] Casari, G. and Sippl, M.J. (1992) J. Mol. Biol. 224, 725–732.
[40] Cornette, J.L., Cease, K.B., Margalit, H., Spouge, J.L., Berzofsky, J.A. and DeLisi, C. (1987) J. Mol. Biol. 195, 659–685.
[41] Dill, K.A. (1990) Biochemistry 29, 7133–7155.
[42] Thomas, D.J. (1990) J. Mol. Biol. 216, 457–463.