Reflections on the protein-folding problem

2003 J. Phys.: Condens. Matter 15 S1779

(http://iopscience.iop.org/0953-8984/15/18/311)

INSTITUTE OF PHYSICS PUBLISHING JOURNAL OF PHYSICS: CONDENSED MATTER

J. Phys.: Condens. Matter **15** (2003) S1779–S1786 PII: S0953-8984(03)54971-2

Reflections on the protein-folding problem

Per-Anker Lindgård

Materials Research Department, Risø National Laboratory, 4000 Roskilde, Denmark

Received 11 October 2002
Published 28 April 2003
Online at stacks.iop.org/JPhysCM/15/S1779

Abstract

A brief overview is given of the nature and importance of proteins. Current equilibrium theories concerning the folding process are discussed and an alternative non-equilibrium model proposed. This so-called hydrophobic hinge model can predict protein-folding classes and provide a mechanism by which proteins may already choose the right path to the native structure in the extended state. Numerical simulation data support the idea of hinge forces stabilizing particularly crucial parts of the protein in an early stage of the folding process.

(Some figures in this article are in colour only in the electronic version)

1. Introduction

The 'miracle' that proteins on their own can fold from a linear string to a unique three-dimensional structure is mentioned as a puzzling example of a perhaps unexplained phenomenon of nature, in the article 'The middle way' [1]. There, the thought-provoking idea of there possibly being a *new organizing principle* at the nanoscale is propounded.

The final pieces of the human genome are being determined. The next problem is to understand and ultimately manipulate it, for example in efforts to cure inherited diseases. It is written on the DNA string which is about 1 m long. This has patches with the information needed for the assembly of the right sequences of the 20 different amino acids, which are to be connected into the polypeptide chains of proteins. The amino acids act like letters in a text with a 20-letter alphabet. Apparently, by far the majority of the DNA contains pure nonsense with no information content that we can understand[1]. The rest codes for the about 100 000 different proteins which are needed for making a human being function.

Some proteins occur in cell membranes, where they regulate transport of energy or material, control adhesion to other cells and so on. Others, the globular ones, which we shall focus on here, occur freely in water inside and outside the cells. They perform specific functions of all kinds, from catalysis of either uniting or cutting molecules to providing mechanical movement. However, the function is completely dependent on the correct folding of the

[1] Perhaps it is just as in any information medium—for example, like a book, where the information-carrying letters constitute only a small part as compared to the paper and cover. Old book covers occasionally contain fragments of ancient scripts—such reuses appear in DNA too.

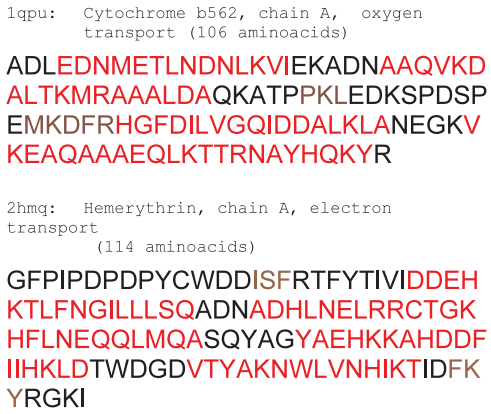

Figure 1. Examples of two protein domains and the corresponding different amino acid sequences (capital letters). From this it is, with 80% accuracy, possible to predict that the red (grey) regions should fold into α-helices; the black (other) regions are loops. By mutation, it is possible to turn parts of the loops, the dark red (dark grey) region, into the α-helix structure. The proteins are structurally related, having four α-helices. The PDB codes—*1qpu* and *2hmq*—relate to the Brookhaven Databank, where the structure and more information can be obtained.

polypeptide chain. The importance of the correct folding is exemplified by the protein, called a prion, which causes mad-cow disease (BSE). It can occur in two folded forms, one harmless and the other deadly. A typical unit of a protein, a domain, contains around a hundred amino acids linked together in a polypeptide chain: the backbone, with various side chains sticking out. A protein corresponds to the number of letters and information content in, for example, a *Phys. Rev. Lett.* abstract. The problem is that of how it can, in all but such exceptional cases as the prion, find a unique structure in a matter of seconds based on (known) physical principles alone: how can it build a specific 3D structure, called *native*, from brief 1D information? Examples of sequences are given in figure 1. Less than 1000 protein structures are known from x-ray scattering, neutron scattering or NMR examination of crystallizable proteins, whereas for many more the sequences are known but not the structure. Discovering the secret of going from sequence to structure is considered like finding the 'Holy Grail' in biological physics. The 'joker' in biology is the role played by Darwinian evolution at the protein level; it may have reduced the phase space in a way unfamiliar to statistical physics. The sequences useful for making proteins appear to be exceedingly complex (see figure 1), but are no more random than the letters in an abstract.

The 20 amino acids can be grouped into two roughly equally sized groups of water-loving (hydrophilic) and water-hating (hydrophobic) ones, with small differences in charges and sizes. However, the main point is that from an energetic point of view no particular pair formation is obviously better than any other. For example, oppositely charged amino acids should attract each other indiscriminately. But this appears not to be the case—and actually it is a big problem to understand how formation of seemingly favourable pairs can be prevented. By statistically analysing the known structures and the corresponding sequences one finds, surprisingly, that some pairs occur as close neighbours more frequently than others. If one *were to assume* that the structure is in *thermodynamic equilibrium*, this knowledge could be transformed into specific interactions (a potential of mean force) between the 20 amino acids. Ignoring their differences in size, these can be called 'beads'. The problem is then reduced to finding the minimum-energy structure for a string of about 100 beads with various interactions by testing all configurations or folds. For simplicity, one can put the string on a lattice. This is called the

bead model for protein folding. Levinthal [2] pointed out that testing all pair combinations, amounting to $\sim z^{100}$ cases, where z is at least 2, would take longer than the age of the Universe. Something is badly wrong!

2. Equilibrium theories

If one considers the energy of the various dense folds as an energy landscape, it is clear that this must be very rugged. Hence it is extremely difficult to find the true minimum. This would be like the problem encountered in *spin glasses*. Even worse, there may be different dense folds with almost the same energy, yet very far apart in configuration space (i.e. one needs perhaps an almost total unfolding to go from one to another). To ameliorate the situation, Wolynes [3] has suggested that somehow nature, during the evolution, has formed the landscape as an overall *funnel*, which gives a higher probability for finding the native state. The problem is, how? Others [4] have suggested, also on the basis of the bead model, that nature has only utilized certain highly *designable* structures. Supporting this idea is the fact that no known protein structure contains knots. So a whole subset of dense folds is not used. The problem, but not the structure, may be a Gordian knot.

One of the oddest things about protein folding is that one can make large changes in the amino acid sequence and still retain the structures with only small, sometimes even desirable, changes in properties. This is the basis for the usefulness of gene manipulations, where one wants only to change the response of cells to certain diseases, but not to make them completely malfunction. It is hard to reconcile this fact with the spin glass paradigm, even when limited by the funnel and designability restrictions. The article 'The middle way' [1] raises the question of the possible existence of *protectorates* at the mesoscale, for example a funnel protectorate. However, a further question arises—it is perhaps not a thermodynamical equilibrium problem at all (contrary to the fundamental assumption behind all the above considerations).

It is not quite clear what is meant by a protectorate. As an example, reference [1] suggests that the crystal state constitutes a protectorate in which phonons live; but no specific example is given for a protectorate for proteins. What could be a protectorate for—let us rather say—protein folding? By (maybe too simple an) analogy, it could be water, slightly salty and only in a rather narrow temperature interval slightly over room temperature. That is where proteins live, fold and keep floating without unduly sticking together, coagulating or flocculating. Both at higher and lower temperatures, proteins tend to unfold, and to denature. Water in itself, and in particular salty water with potentially both monovalent and polyvalent ions (e.g. Na^+ and Ca^{2+}), is an extraordinarily capricious protectorate, which is very far from being understood. It is even harder to predict the behaviour of simple heteropolymers (chains of organic molecules with varying charges and responses to water, of which proteins are an important subgroup) in that medium. Many examples of surprising effects are illustrated in the recent symposium on *Colloidal Physics and Physics of Colloids* [2].

We believe that the key player—or perhaps rather playground—in the protein folding is water and as a key rule: non-equilibrium behaviour may be allowed and used.

3. Hydrophobic hinge model

First-come-first-served (FCFS) is a well-known principle—perhaps not regarded as fundamental in physics. But could this be the principle by which the building blocks of life, the proteins, are organized? *It is not (necessarily) an equilibrium principle*. The newly born protein string of the about 100 amino acids is randomly extended for entropy reasons. On its way to lower energy, suppose it only locally forms pairs of the first and best suited

partners, and as a first step begins to form the well-known secondary structures. These are called an α-helix, a strand for a β-sheet or the intervening loops. It is actually possible from the sequences, with about 80% accuracy, to predict [6] the amino acids of a string that will belong to one of these groups. By the FCFS principle the string could undergo a rough partitioning and begin building up bonds and structure in such units (see figure 1) before the dense packing is established. Unfortunately, so far neither experiments nor computer simulations have been able to probe the relevant time window (typically 0.1–100 μs) and witness this happening. These units are typically about ten amino acids long (the averages and variances are [5] 10 ± 5 for α-helices, 5 ± 3 for β-strands and 5 ± 3 for loops). If this happens, the problem is reduced to packing only about 10–20 tied-together sticks (or 'bones') with the main characteristics that they do not like water. We call this scenario the hydrophobic hinge (HH) model [5]; see figure 2, right: 'hydrophobic' because that is the main driving force and 'hinge' because a guiding force is also needed—more later on that. The problem is still big, but one particular part of it is solved—namely, that part concerning preventing wrong, but energetically favourable, accidental contacts between distant parts of the chain. The combined interaction of the 2×10 amino acids on two different α-helices will be small, due to cancellation. It is approximately true that the helices do not interact at a distance. The problem is then just that of packing the 10–20 bones in a sack as densely as possible to avoid contact with water. This can be done only in a relatively limited number of ways. For simplicity, the problem can again (and with more justification) be put on a lattice.

4. Fold classes

The number of ways such units can be close packed as a function of the number of units can be precisely counted [5], and each fold is known and can be named. A four-helix protein, as in figure 2, has six hinge spins—therefore five interaction constants are needed to determine the optimal relative orientations. The sequence of these constants can be used as the name. In a mnemonic code [5] the name of the fold class indicated is *irili* (there are five other directional, dense folds, with the systematic names *obubo, iruri* and the reverse *iliri, ubobu, iloli*). It is found that there exist folds with numbers of secondary structures, which can be packed in particularly few ways. These have especially small configurational entropy, and can be argued to be particularly stable and independent of the sequence [5]. There should accordingly be certain 'magic' number proteins, which are abundant, i.e. with many different species—very similar to the phenomenon encountered for stable small atomic clusters and stable nuclei. In figure 3(a) we show the number of distinct folds with clear minima at the indicated secondary structure content. Figure 3(b) shows an analysis of the abundance of different proteins in a representative database [7]. There is a clear correlation. These fold classes can be uniquely named and may therefore form a basis for providing a precise structural classification of the protein folds—at least in principle; the details need still to be worked out. If so, there should be only around 4000 different domain fold classes—enormously reduced from the astronomic number z^{100}. It is close to the *ad hoc* guess of Chothia [8] of about 1000—in particular, if Nature has only used a fraction for evolutionary easy-to-fold reasons.

5. Hinge forces

So far we have argued that it is possible from the sequence for a particular domain to predict the subdivision into secondary structures, and subsequently a relatively small number, of the order of 10–100, of known folding classes to which they may fold; see figure 3(a). These folds

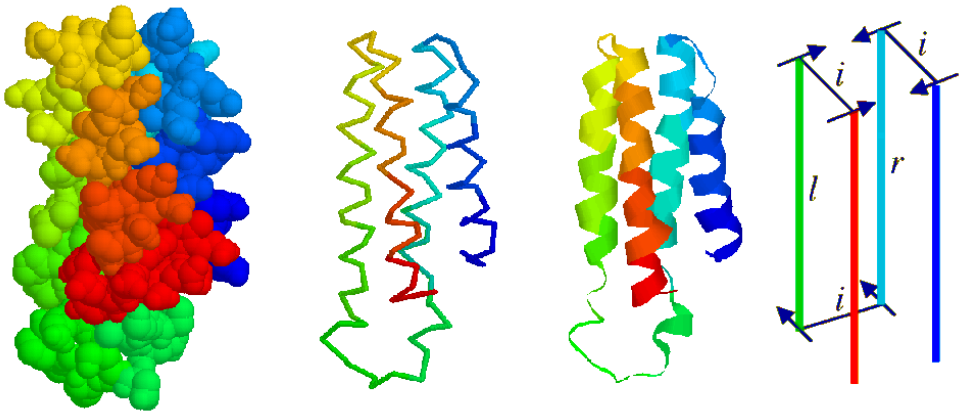

Figure 2. The structure of *1qpu*, cytochrome b562, chain A, determined by means of NMR in a solution. Left: the all-atom view, which is obviously quite dense. Even when immersed in water, proteins contain only about 2% in the interior. Next: the underlying polypeptide backbone. In the *bead model*, a bead is added at each kink. Here there are 106 interacting beads. Then: the ribbon representation showing helices and loops. Notice that the rather flimsy looking loops are actually quite dense in the all-atom view. Right: the corresponding *HH model* [7]. The coloured and black lines represent a lattice simplification of the structure elements, helices and loops; arrows indicate the hinges. A structure element is supposed to carry the information about the optimal direction of the hinge spins giving the name *irili*. The structures of *2hmq* and *1qpu* are identical (hence belonging to the same fold class) in the HH model representation—although they are somewhat different in the ribbon representation.

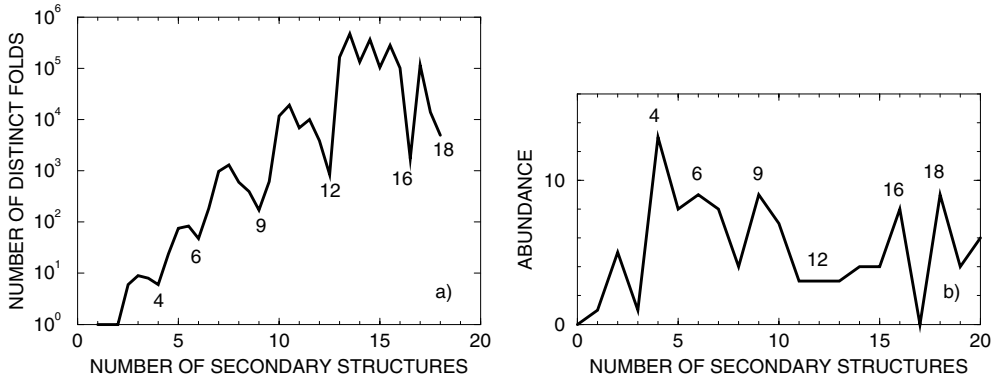

Figure 3. (a) The number of distinct folds obtained by an exhaustive count using the *hydrophobic hinge model* [5]. The minima correspond to small configurational entropy, which one can argue leads to higher probability of abundance. (b) The corresponding abundance found in the representative, diverse protein data set of 136 proteins with sequence similarity less than 25%, selected by Rost and Sanders [9]. A substantial correlation between the indicated 'magic' numbers is evident.

are far apart in configuration space. In an energy landscape picture, this would correspond to a landscape with 10–100 major minima or funnels. These are formed by minimizing the interaction with water—and should therefore be quite similar in depth and shape. However, for a given protein sequence, only one of these minima is important: that containing the *native*

state. The protein is able to find this with no errors. In interaction-based models (such as the bead model) this can only be tested by comparing the relative energy of the various dense, but very distant folds. That means that a protein has to go down into each of the up to 100 minima, compare, go up again and back into the best!

We believe that the protein must 'know' its native state already in the unfolded or partly folded state. This can be achieved by 'hinge forces', which specify a preferred direction for the units to be joined. Each unit could carry this information in its own linear sequence. It can thus be assigned a hinge label. As a physical mechanism for such a force, one may consider an α-helix in the z-direction. Suppose the joining unit at the bottom enters in the x-direction; then the probability for the top outgoing unit will depend strongly on the number of turns, or simply on the number of helix-forming amino acids. This information can very easily be stored in the sequence, thus yielding the hinge information. For example: ingoing along x and outgoing along y provides a right turn (r), whereas outgoing along $-y$ provides a left turn (l) in figure 2. In other cases, the information could be hidden in the sequence of amino acids near the joints of the units in a code that we are so far unable to read, but which could still be known to the proteins. In this scenario the folding should be strongly sensitive only to some very specific mutations (as is known to be the case), namely in the parts containing the hinge information: for example, by modifying the length of the α-helices, the properties of the loops or the secondary structure content.

In the example, figure 2, there are two obviously equally dense folds, which therefore should be degenerate with respect to the close packing in water: the one shown, *irili*, and the reverse, *iliri*, with the red–green helix pair placed on the other side of the blue–violet pair. But only one fold is the native one. A fold inverting (i) the string and bringing the α-helices parallel is rather obviously favourable for packing. Any of the pairs (the red–green, the green–blue or the blue–violet) could form first by the FCFS principle. However, it is crucial that the blue helix favours the left (l) sense of junctions and the green helix the right (r) sense of junction to position the remaining chain on the correct side of the plane defined by the first-formed pair of helices. The opposite choice would give the reverse, non-native form, *iliri*. The final form of this is clearly far away from *irili* in configuration space.

The hinge forces could operate in the extended state, where the secondary structures are about to be formed. They would then guide the fold gently (since they are probably rather weak compared to the local and the hydrophobic forces) towards the right region in configuration space, which contains the native fold. This is consistent with the fact that a small guiding field is able to resolve the degeneracy in the magnetic ordering of the Heisenberg model for a spin system. There, the orientational degeneracy of the ground state is infinite. In the protein-folding problem the hinge forces only have to resolve the ground state degeneracy (10–100) of the fold classes, which are determined by the indiscriminate hydrophobic forces. After being guided to the right fold class, the best minimization of the interaction energies can begin. (It is not a funnel, because the fold class need not correspond to the overall dominant energy valley, or even possess the lowest-energy minimum.) However, any *major* changes in the already forming secondary structures or loops are prohibited. It is in the spirit of the FCFS principle that the structure building may start around a nucleus of potentially well-matched units. This seems to be what is observed in small proteins, where the folding process can be followed.

Clearly, with this scenario, the overall energy (free energy) need not be minimal—and the native state need not be in a state of minimal energy or of minimal frustration, as suggested on the basis of the spin glass paradigm [2]. It could be so by chance or as a result of evolution; but the robustness to mutations seems to indicate that this is not the case. Anyway, it is very hard to reconcile the glass paradigm, which is characterized by not yielding a specific structure, and a single (native) structure, which is more like a particular crystal.

6. Simulation results

These ideas were advanced some years ago [7], when very limited information was available. Now, several results from simulations [10] and mutation studies [11] have actually supported the HH model. In the most extensive protein-folding simulation to date, of duration 200 ns, a *villin* headpiece subdomain of 36 residues (three α-helices) was studied. It showed, surprisingly, that the loop or turn at residues 20–23 was stable even at 1000 K. This of course limits the folding room for the rest of the chain considerably, just as the hinge forces are expected to do. Next, helix 2 is formed and then, partly, the two others. It is generally found that the α-helices are not fully stable up to the final stage, but rather fold dynamically with moving defects or with an incorrect pitch. Such disorder retains the residues involved, in a limited spatial region—and may in fact be instrumental in moving the rather large unit (as an α-helix is) more efficiently. Hence it supports, rather than contradicts, the above picture. It has also been found that for the folding of β-sheets, the folding of the turn (β-hairpin) is where the folding initiates. This is again consistent with the hinge picture. It has been concluded that it is the hydrophobic groups which are important in the early stages, whereas, curiously enough, the hydrogen bonds seem to slow down the folding process. This is perhaps because they indiscriminately also form wrong connections, which have to be broken again later.

Very recently, after this paper was submitted, an important study was published on the folding of the small artificial 23-residue mini-protein BBA5 [12] (having one α-helix and two β-strands). It has a strong propensity to form secondary structures and a small hydrophobic core. By means of a true *tour de force*—using 30 000 computers, distributed all over the world, and taking several months of CPU time on each—Snow *et al* [12] were able to follow a statistical distribution of folding trajectories. (A main criticism of the above work on *villin* has been that just one possible trajectory was explored.) As a result, it has been possible to piece together trajectories as long as 700 μs and thousands of shorter ones, which show folding patterns from which the behaviour at longer times can be deduced. They found much higher folding rates at large temperatures $T = 338$ K than at 278 K, and further that the overall folding time is 6 μs (comparing well to the experimental value 7.5 ± 3.5 μs); and that the α-helix and β-hairpin fold in 0.8 and 1.5 μs, respectively, at $T = 298$ K. Additionally, it was found that a single, particular mutation was detrimental to the β-hairpin formation. These results support strongly—even for a mini-protein—the assumptions underlying the HH model.

The folding of BBA5 is efficient and fast because of the initial folding of the secondary structures. This is contrary to the commonly held picture of folding of *small proteins*, namely that it happens as a concerted motion in which all parts fold simultaneously in accordance with the funnel paradigm. For references and more discussion of this, see the paper by Irbäck [13].

7. Concluding remarks

Summarizing, the sought organizing principle [1]—or protectorate—at the mesoscopic scale is perhaps that one cannot rely on thermodynamic equilibrium on the scales of time and space relevant to life. This seems to happen for certain mesoscopic cases found in metallurgy (martensitic transformations [14]) and the partial oxygen ordering [15] in high-T_c superconducting $YBa_2Cu_3O_{6+x}$. The systems simply cannot reach equilibrium, but are self-organized in small-scale domains by the *FCFS principle*. The HH model for protein folding described provides a possible way to go from the sequence to the structure without invoking any unknown physical principles.

References

[1] Laughlin R B, Pines D, Schmalian J, Stojković B P and Wolynes P G 2000 *Proc. Natl Acad. Sci. USA* **97** 32
and see also
Laughlin R B and Pines D 2000 *Proc. Natl Acad. Sci. USA* **97** 28

[2] Levinthal C J 1968 *Chem. Phys.* **65** 99

[3] Wolynes P G, Onuchic J N and Thirumalai D 1995 *Science* **267** 1619

[4] Li H, Helling R, Tang C and Wingreen N 1996 *Science* **273** 666

[5] *Colloidal physics and physics of colloids* 2002 *J. Phys.: Condens. Matter* **14** R859, 7551–779 (Special Issue)
In particular, the history article on p7769 is interesting and perhaps not irrelevant to the protein-folding problem.

[6] Petersen T N, Lundegaard C, Nielsen M, Bohr H, Bohr J, Brunak S, Gipperst G P and Lund O 2000 *Protein Struct. Funct. Genet.* **41** 17

[7] Lindgård P-A and Bohr H 1996 *Phys. Rev. Lett.* **77** 779
Lindgård P-A and Bohr H 1997 *Phys. Rev.* E **56** 4497

[8] Chothia C 1992 *Nature* **357** 543

[9] Rost B and Sanders C 1993 *J. Mol. Biol.* **232** 584

[10] Duan Y and Kollman P A 2001 *IBM Syst. J.* **40** 297

[11] Baltzer L, Nilsson H and Nilsson J 2001 A review on experimental *de novo* design of proteins *J. Chem. Rev.* **101** 3153

[12] Snow C D, Nguyen H, Pande V S and Gruebele M 2002 *Nature*
doi:10.1038/nature01160—www.nature.com/nature

[13] Irbäck A 2003 *J. Phys.: Condens. Matter* **15** S1797

[14] Vives E, Castán T and Lindgård P-A 1996 *Phys. Rev.* B **53** 8915

[15] Mønster D, Lindgård P-A and Andersen N H 1999 *Phys. Rev.* B **60** 110