Solving the protein folding problems

Robert Paul Bywater

Magdalen College Oxford England
and
Adelard Institute Manchester England.

Contact: romeobravo@adelard.org.uk

The protein folding problem persists in being the major unsolved problem in biophysics and molecular biology. The most immediate obstacle to its solution is said to be the immense combinatorial difficulty of finding the global minimum energy structure, due to the galactic number of possible rotamer states for the polypeptide backbone and its sidechains, the so-called Levinthal paradox\(^2\). Much attention has been paid to the topography of the free energy surface that the folding polypeptide has to negotiate in order to find this minimum and the notion of a funnel-shaped topography has been proposed\(^2\). While this has achieved wide acceptance, providing insight into the possible shape of the free energy surface, it still does not solve the problem. In this work it is suggested that, in order to proceed further with this line of enquiry, it would be useful to focus on not one but, rather, two of the free energy minima accessible to the folding polypeptide. These two minima are considered to represent the active and inactive, or resting, states of the protein. There is experimental evidence for this model and a well populated database of structures that is analysed in detail in this work.

When protein structures are determined by X-ray crystallography or NMR, it is a common practice to carry this out on a complex of the protein with some small ligand. For enzymes this may be the substrate, or rather, a close analogue that does not react easily, or an inhibitor. In the former case a justifiable assumption is that the protein is stabilised in its active form, while the converse would be true for the inhibitor case. When the approtein is the object of the structural study it may seem likely that it is the inactive form which is favoured, although there are examples of constitutively active proteins either naturally or that can be rendered constitutively active by mutation. This is common among G-protein coupled receptors for example\(^3\). Whichever is the case, it is reasonable to suppose that there are at least two low-energy, stable forms of each protein. Further, it has been shown\(^4\) that both the active and inactive configurations of the protein are visited repeatedly during the normal cycles of the dynamic behaviour of the protein. The ligand associated with activity binds opportunistically when the active configuration presents itself, and vice versa for the inhibitory ligand. According to these observations, the induced fit theory is not upheld. Both structures are accessible whether ligand is present or not. This means that we have two protein folding problems rather than one. Does this make the task easier or not?

A most valuable bioinformatics tool for investigating protein activation and the interchange between active and inactive states is the program Dyndom\(^5\). This program compares the crystal structures of the active and inactive forms of proteins by identifying domains and determining the rotation matrices that define the interdomain displacements. Users can either download the program for
personal use, or make use of the large database\textsuperscript{1} of structural pairs (active/inactive)\textsuperscript{6}.

In the present work, the entire DynDom database comprising 2035 protein pairs was examined in detail in order to reveal important structural features that are responsible for protein activation and folding. This database contains examples of many different protein classes. There is very good coverage of both structure and function space, but also a high degree of redundancy with many examples from closely related protein systems. Therefore it was considered satisfactory, and definitely more manageable, to select a representative subset of 21 protein pairs from the database. Although this constitutes only 1% of the entire database, there are examples from a wide variety of protein families including several functionally different enzymes with very different folds: a kinase, a DNA polymerase, a topoisomerase, a TIM barrel enzyme, a ubiquitin-conjugating enzyme, an insulin TK receptor domain, myosin, also: a lectin, an antibody domain, a membrane ion channel, GroEL chaperone, a ribosomal protein, a nuclear receptor and even a “helix destabilising protein”. The secondary structure distribution ranges from all-helix to all-strand with many intermediate cases, and one example with an unusually high $3_\text{_{10}}$ helix content. Thus, the chosen subset has the merit of having a wide coverage of structural type and functional repertoire while being nonredundant. This means that the results presented here can be regarded as a useful guideline as to what to expect in other members of these classes and they are sufficiently comprehensive for the conclusions to be valid for virtually all globular proteins.

For the members of this subset, a protein from each pair was in turn examined for internal contacts, separated into those due to hydrogen bonding and those due to van der Waals (vdW) contacts. The program WHAT IF\textsuperscript{7} was used to conduct these studies. The WHAT IF command for hydrogen-bonded interactions is SSHBO for sidechain-sidechain interactions, BSHBO for backbone-sidechain interactions, both in the HBONDS menu while for the proximities due to vdW interactions it is CONTPS in the ANACON menu. In order to automate this procedure, WHAT IF scripts and two fortran programs that process the output of these scripts were written. (The raw data is available from the author and it can easily be reproduced for these proteins or any others using the scripts and programs which are also available on request. All computing was done on laptop computers running under Ubuntu v10.04 Linux. The computer programs are written in fortran and they compile under gfortran or g77).

The results are shown in Table 1 where the column headings are: \textbf{In both} = same pairs of residues and atoms within those residues are identical in both states of the protein. This is a fine-structure analysis, so that e.g. cases where CD1, CG1 atoms in isoleucine, CG1, CG2 in valine of CD1,CD2 in leucine are swapped are regarded as being rearrangements. \textbf{Active only} = number of contacts present in “active” but not in “inactive”, while \textbf{Inactive only} is the converse of this. In each category the scores for sidechain to sidechain hydrogen bond contacts, and those between backbone and sidechains, respectively, are recorded, as well as all hydrophobic contacts (involving residue types Ala, Val, Ile, Leu, Met, Phe, Tyr and Trp). Lastly, results for a subset of the latter, where only aromatic residue types (Phe, Tyr, Trp) are considered, are listed. The WHAT IF default cutoff distance of 0.25 nm is used throughout.

\textsuperscript{1} http://fizz.cmp.uea.ac.uk/dyndom/browse.do? 
\textsuperscript{6} type=WEBSITE&id=5EE751A622FAC3086265338661CDFFC43
The following conclusions can be made from the data in Table 1:

- Studies of hydrogen bonding patterns reveal major rearrangements between “active” and “inactive”, with relatively few interactions retained when the two structures interchange. This is apparent for both the sidechain-sidechain and backbone-sidechain cases. This trend seems to be independent of the relative prevalence of different secondary structure types. This reflects the fact that hydrogen bonding is mostly concerned with maintaining domain structure rather than inter-domain interactions. However, there are concerted realignments of the hydrogen bond connectivity.

- For hydrophobic interactions a similar trend is observed with rearrangements taking place between the two states of the protein. But the effect is much weaker and many more contacts are retained undisturbed after the transition between states. Hydrophobic interactions are critical for maintenance of “core” structure (that is to say, the core of the individual folding unit that forms a domain as normally defined in the context of protein structure) but they can also be critical for interfacial interactions between domains and some may need to be preserved.

- For the most part, there is a similar number of hydrogen bonding contacts exclusive to each state of the protein, while for hydrophobic interactions, there is in several instances a change in the number of interactions exclusive to each member of the pair. Given that, as is widely claimed, the “core” of protein domains is largely stabilised by hydrophobic interactions, this suggest that there is a change in stability in going from one structure to the other.

- Aromatic-aromatic interactions behave differently to the hydrophobic interactions taken as a whole. If anything, their behaviour resembles that of hydrogen bonding interactions. This is not surprising, since they, like the latter, exhibit considerable directionality in their interactions due to the quantum effects of orbital overlap, while vDW interactions are considerably more “smeared out”. It should be taken into account, from the point of view of statistics, that there are far fewer interactions of the aromatic type.

- The 1br4h/1brf pair (myosin) is a striking exception to the above, especially in regard to the behaviour of aromatic residues.

An overall conclusion is that there are very appreciable differences between the two states of the proteins in regard to the internal interactions that are supposed to maintain structural stability. The various regions of the protein that move seem to do so without very great changes in their local domain structure, most of the deformation is confined to a relatively short region of the polypeptide chain, while the secondary structures seem largely to remain undisturbed (see the columns labelled “helix/strand %” in Table 1). Readers are recommended to inspect the structures and their interconversions on the DynDom website to confirm this. So, it seems that the internal interactions studied here have more to do with the dynamical behaviour of the protein than with protein stability as such. Or, another way to put this is that it is domain structure which is preserved by the intra-residue interactions rather than stability of the global protein structure. This underlines the importance and validity of the notion of domain in protein structure which furthermore has its genetic origins in the exon sequences of the encoding DNA. Nota bene that many domains are constructed from noncontiguous stretches of polypeptide, so one has to exercise caution when utilising and discussing the domain concept.

There has been considerable debate about whether hydrogen bonding or vdw interactions are most important for maintaining protein stability. Earlier work compared hydrophobic interactions with hydrogen bonding and tended to favour the former, while later, accurate thermodynamic studies supported the
case for hydrogen bonding. It is very hard to compare these two in any rigorous fashion because they arise from different mechanisms. Both are strengthened by the effects of desolvation, but here again, the desolvation mechanisms are very different due the different polarity of the surfaces being desolvated. It was one of the aims of this work to throw some light on this issue, since any reshuffling of the structures must involve a rearrangement of the participants in the various interactions.

In order further to illustrate the differences between a protein in its active conformation and its inactive counterpart an example, chosen from the list, is discussed in a little more detail. Consider the 2eck/4ake pair, adenylate kinase. In a sense, this was an arbitrary choice, but there was an element of deliberation in this choice given that this protein family is a member of the “alpha and beta” (αβ) class. This class is more complex in its structure than for example “all α” or “all β” proteins. They are shown as cartoons produced using the modelling program YASARA, in Figures 1a and 1b respectively. The difference between the active and inactive conformations is clearly discernible, the former being more closed and compact as the protein packs around the bound substrate (ligand not shown). In the active form, 2eck, the axis around which the beta-sheet is twisted is in perfect alignment with the axis of helix H2. In 4ake these helix axis no longer coaligns with the sheet axis. There may be interesting mechanical reasons for this. Figures 2a show the hydrogen bond networks while the vdW contacts are shown in Figures 2i. The very considerable rearrangements of these networks are clearly discernible.

Attention is now directed towards the question of whether these findings simplify or complicate the task of predicting the folded structure. Proteins are often described in terms of local structures along the polypeptide chain made up of so-called secondary structure elements (SSEs), within which stretches of polypeptide form a regular structure with repeating values of the Φ, ϕ backbone torsion angles. The accuracy of secondary structure prediction has improved somewhat over the years from a little over the 60% mark, flattening out to somewhere between 70 and 80%, depending on the algorithm used, although a combination of methods has been reported which gives accuracies in the range 67.5% to 93.2%. But the quality of the secondary structure prediction can hardly be expected to improve any further given that the existing methods have all been trained on protein structures as if each protein had only a single native structure. It is exactly this notion that we want to escape from.

There are many different secondary structure prediction programs available to researchers in bioinformatics. Prominent amongst these are the neural network programs PROF/PHD and Jpred3. In Table 2, line 6 contains the secondary structure predicted by the PROF/PHD program while line 7 shows the PROF/PHD prediction of solvent exposure for adenylate kinase. Line 8 shows the secondary structure predicted from Jpred3. Lines 2 and 4 show the secondary structure as determined by DSSP from the X-ray crystal structure of adenylate kinase, active (2eck) and inactive (4ake) respectively, while lines 3 and 5 give the corresponding calculated data for solvent exposure. These postdiction results can be contrasted with the PROF/PHD and Jpred3 predictions. It can readily be seen that the PHD and Jpred3 results represent some kind of middle ground between the two sets of DSSP postdictions.

Given that there may be considerable differences between active and inactive structures, beyond domain movements, including issues like unwinding of helices, accessibility etc. there is a major disadvantage with secondary structure prediction algorithms, including PHD and Jpred3, that they can hardly be expected ever to deliver an unequivocally “correct” result. This would require training the
program on only active, or only inactive structures, which may be difficult to do in any meaningful way. But these observations help to explain why secondary structure prediction may never get close to 100%. What PROF/PHD does predict rather well in fact is that the important hinge region, which can be inspected (regions coloured green) for all protein pairs on the DynDom website.

What makes the protein folding a difficult problem? Typically, the above mentioned Levinthal problem arising out of the huge number of conformers is invoked by many authors. This author however takes the view that proteins have evolved and their constituent aminoacid residue types selected in order to be able to fold into two stable structures, not just one. Otherwise, a protein may be regarded as “just” a rather complex organic polymer. It is after all made up from chemically rather simple building blocks. Small organic molecules find their global minimum conformation rapidly and their structures can be predicted accurately using the tools of conformational analysis and molecular mechanics (for conformational studies) and ab initio quantum mechanics (for accurate bond lengths and angles). While it is certainly possible to study the behaviour of protein molecules using molecular dynamics, these methods are not able to find the global minima within polynomial time unless a very good starting structure is available. “Good starting structures” is normally taken to mean those structures obtained from X-ray crystallography or NMR and indeed, molecular dynamics is used in the refinement of these experimentally derived structures. Nevertheless, it remains the goal of many to predict protein structures from sequence data alone. But just as for the experimental approaches, it is important to consider the active and inactive forms as separate, albeit related, structures. Attempts to predict protein folds from sequence alone should take into account the existence of two distinct possible conformations for the folded structure. In cases where a prediction is judged, for example in competitions such as CASP to be incorrect, the judgement may itself be in error because the structure predicted is not the one for which a crystal structure has been obtained but rather its (in/)active counterpart.

In this work, the emphasis is rather on discussing the problems how to predict not one, but two, folds for each protein representing, respectively, the active and inactive forms. This may at first sight appear to be an even more challenging problem than for a single “native” state. While most attempts at solving the protein folding problem aim at arriving at this single structure, it is often overlooked that the undulating energy surface near the bottom of the energy “funnel” will change its shape according to the prevailing conditions, especially if a ligand is attached, such that different states may correspond to the “global minimum” depending on circumstances. Further, it is known (vide supra) that proteins will cycle through these states all the time, the “circumstances” referred to here simply favour one or other of the structures.

In the light of these observations we now have to revise our approach to the solution of what have now become, in plural, the protein folding problems. Several approaches towards this are under investigation by the author.

One approach takes into account the evolutionary constraints on sequence variability that sustain the ability to still fold in two ways\textsuperscript{16}, corresponding to an active and inactive structure for each protein. The residues at different positions in a protein are responsible for different functions. These functions include protein folding, ligand recognition, recognition of other proteins which the protein in focus needs to interact with, surface regions that act as an “address label” to direct the protein to the correct locus in the cell and so on. Now we have to include the active/inactive “switch” regions. It has been shown\textsuperscript{17} that it is possible to partition a multiple sequence alignment, into disjoint (not contiguous in the primary
sequence) regions corresponding to these functions. The reason that these partitions can be found is that there is a different selection pressure operating on each region, corresponding to the function in question. In accordance with this observation, plots of Shannon entropy of the substitutions that have occurred and survived at any given site against the variability of residue type at that site show very distinct partitioning between different functions\textsuperscript{18}. This is revealed because entropy and variability express sequence disparity in different ways, the former has to do with what is required at a given residue position, while the latter measures what can be tolerated at that position. The balance between these two is different with different functional capacities. This results in the clustering of the residue positions into different “boxes” in the entropy/variability plots\textsuperscript{18}. Most of the functions mentioned above could be identified\textsuperscript{18}, but not of course the active/inactive switch, which is the focus of this paper. The question then is, do the “switch” residues reveal themselves at the level of entropy/variability plots? Figures 3a and 3b provide an affirmative answer to this. Both figures show identical entropy/variability 2D scatter plots for adenylate kinase, where the entropy and variability data are taken from a multiple alignment of 131 homologous sequences produced by the PROF/PHD program. The third dimension is displayed by colour coding where darker colouring indicates a variable derived from the crystal structures of the active (2eck) and inactive (4ake) forms respectively. For Figure 3a this is the relative displacement of each CA atom (a scalar quantity in Å, obtained using the WHAT IF program), while for Figure 3b it is the difference between the crystallographic B-factors at each CA atom, severally, in the two structures. The latter measures differences between thermal motions at these atoms. In both cases there is a clustering of residue positions with large values of these variables, CA-CA displacement and B-value difference, in the same “box” in the plot. This corresponds roughly to the “box 33” in the previously published work\textsuperscript{18}, for which no clear function was then identified. It appears from the present work, that the residues responsible for the “switch” function are to be found in this “box 33”. This remains to be confirmed since the dataset used here was small and only one protein family was considered.

Yet another way that this problem is being attacked is to consider the entire scheme of protein folding/unfolding/misfolding and to try to determine the transition functions that govern each step in the pathways. See scheme in Box 1. Each node in the pathway represents a state of the protein expressed, for mathematical convenience, as a NxN square matrix where N is the number of atoms in the protein. These matrices may be for example 2D atom-atom contact maps, or a version of the 3xN coordinate matrix padded with zeros to create a NxN square matrix. The transitions between states is governed by NxN transition matrices C such that a transition from, say, state $A$ (e.g. active) to state $J$ (inactive) may be expressed as a matrix product $CA = J$. It is the prediction of the elements in these matrices, and their inverses $C^{-1}$, which, in this approach, now constitutes the “solution to the protein folding problems”. The determination of these transition matrices has been accomplished for test systems when the structures are known (results to be published) but for proteins where there is no crystal structure it is necessary to resort to the use of numerical inference methods, which are also under development.

It is still not clear whether the fact that there are multiple “native” states makes the task of fold prediction easier or not. In mathematics, one would not normally attempt to solve a problem unless it could be shown that a solution exists, as is manifestly the case with protein folding. But here there are (at least) two solutions. Again, there examples in mathematics when there are two solutions and where there is a known relation between these solutions, it becomes much easier to solve both problems.
Acknowledgements

Gert Vriend is thanked for interesting discussions over many years on protein modelling and structure prediction, including the entropy/variability approach pioneered by him. He and Elmar Krieger are thanked for supplying the WHAT IF and YASARA programs.

Author information

This work was inaugurated during the tenure of a Visiting Fellowship at Magdalen College, Oxford. The College and its Fellows are hereby thanked for providing a stimulating environment and ambience for conducting this work. There has been no funding nor sponsorship of any kind and therefore there are no competing interests to be declared.
References


<table>
<thead>
<tr>
<th>Protein</th>
<th>Dyndom pairs</th>
<th>Active Sidechain H-bonds</th>
<th>Inactive Sidechain H-bonds</th>
<th>Active Backbone H-bonds</th>
<th>Inactive Backbone H-bonds</th>
<th>Hydrophobic interactions</th>
<th>Aromatic interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>2i41</td>
<td>72.0</td>
<td>99.3</td>
<td>1hrm</td>
<td>74.0</td>
<td>0</td>
<td>12 10</td>
<td>22 25</td>
</tr>
<tr>
<td>1dkf</td>
<td>65.3</td>
<td>100</td>
<td>1lbd</td>
<td>63.3</td>
<td>1</td>
<td>14 16</td>
<td>33 25</td>
</tr>
<tr>
<td>2a0l</td>
<td>64.0</td>
<td>99.6</td>
<td>1orc</td>
<td>75.0</td>
<td>0</td>
<td>1 3</td>
<td>2 12</td>
</tr>
<tr>
<td>2csb</td>
<td>60.1</td>
<td>100</td>
<td>2csd</td>
<td>61.3</td>
<td>2</td>
<td>62 54</td>
<td>94 93</td>
</tr>
<tr>
<td>1xkn</td>
<td>50.0</td>
<td>100</td>
<td>1aon</td>
<td>48.13</td>
<td>0</td>
<td>37 77</td>
<td>66 77</td>
</tr>
<tr>
<td>1arop</td>
<td>47.5</td>
<td>99.7</td>
<td>1h8a</td>
<td>53.8</td>
<td>0</td>
<td>35 64</td>
<td>10 116</td>
</tr>
<tr>
<td>1kx5</td>
<td>47.0</td>
<td>100</td>
<td>1kx5</td>
<td>49.0</td>
<td>0</td>
<td>9 9</td>
<td>14 12</td>
</tr>
<tr>
<td>1br4</td>
<td>47.3</td>
<td>100</td>
<td>1br1f</td>
<td>43.3</td>
<td>0</td>
<td>9 10</td>
<td>27 26</td>
</tr>
<tr>
<td>2bp</td>
<td>46.9</td>
<td>100</td>
<td>1bp</td>
<td>47.9</td>
<td>1</td>
<td>18 9</td>
<td>23 23</td>
</tr>
<tr>
<td>2eck</td>
<td>43.0</td>
<td>100</td>
<td>4eka</td>
<td>45.16</td>
<td>3</td>
<td>22 15</td>
<td>30 17</td>
</tr>
<tr>
<td>5stim</td>
<td>39.0</td>
<td>100</td>
<td>1tpd</td>
<td>39.15</td>
<td>0</td>
<td>28 25</td>
<td>57 64</td>
</tr>
<tr>
<td>1z2u</td>
<td>36.0</td>
<td>90.7</td>
<td>1x23</td>
<td>35.15</td>
<td>4</td>
<td>5 10</td>
<td>29 29</td>
</tr>
<tr>
<td>1fw0</td>
<td>34.0</td>
<td>100</td>
<td>1ftja</td>
<td>35.15</td>
<td>4</td>
<td>11 16</td>
<td>36 42</td>
</tr>
<tr>
<td>1ewk</td>
<td>33.0</td>
<td>100</td>
<td>1ew</td>
<td>33.19</td>
<td>6</td>
<td>64 64</td>
<td>85 88</td>
</tr>
<tr>
<td>1ypit</td>
<td>33.0</td>
<td>99.6</td>
<td>1lys</td>
<td>32.20</td>
<td>8</td>
<td>28 31</td>
<td>73 78</td>
</tr>
<tr>
<td>2auh</td>
<td>31.0</td>
<td>100</td>
<td>2b4s</td>
<td>32.17</td>
<td>0</td>
<td>22 24</td>
<td>57 66</td>
</tr>
<tr>
<td>2bw0a</td>
<td>24.0</td>
<td>92.8</td>
<td>1s3i</td>
<td>24.26</td>
<td>4</td>
<td>30 34</td>
<td>49 42</td>
</tr>
<tr>
<td>1je5</td>
<td>14.0</td>
<td>100</td>
<td>1je5</td>
<td>14.35</td>
<td>1</td>
<td>16 13</td>
<td>19 23</td>
</tr>
<tr>
<td>1ulk</td>
<td>10.0</td>
<td>100</td>
<td>1ulk</td>
<td>6.10 (12)</td>
<td>0</td>
<td>8 5</td>
<td>28 29</td>
</tr>
<tr>
<td>1ay1l</td>
<td>5.51</td>
<td>91.4</td>
<td>1dqd</td>
<td>2.49</td>
<td>4</td>
<td>18 17</td>
<td>8 41</td>
</tr>
<tr>
<td>4phv</td>
<td>4.51</td>
<td>99.0</td>
<td>3phv</td>
<td>4.40</td>
<td>0</td>
<td>10 2</td>
<td>10 9</td>
</tr>
</tbody>
</table>
Legend to Table 2.

1st line – Peptide sequence in standard single-letter code.
3rd line – Solvent exposure postdiction for active conformer. A = more than 10Å² exposed to solvent or more than 33% of its accessibility in the unfolded state.
4th line – Secondary structure postdiction for inactive conformer
5th line – Solvent exposure postdiction for inactive conformer.
6th line – Secondary structure prediction from PHD program H=helix, E = extended (i.e. strand).
7th line – Predicted solvent exposure. e = exposed, b = buried, i = intermediate.
8th Line – Secondary structure prediction from Jpred3. H = helix, E = extended (i.e. strand).
WHAT IF