The proportion of polypeptide chains which generate native folds—part 2: theoretical studies

Royal Truman

Two decades ago Lau and Dill developed a computer program based on a two-dimensional lattice model. With this program they showed that between 10^{-6} to 10^{-10} random polypeptides, a hundred residues long, would produce stable protein-like folds. Ever after, their seminal work has been quoted as evidence for the claim that randomly formed polypeptides readily produce native-like folds and for a naturalistic origin of globular proteins. However, the model has never been calibrated nor validated against empirical data. Here, we show that their program/model is far too removed from the realities of protein chemistry to permit any kind of quantitative estimates.

Two decades ago Lau and Dill developed[1] a computer program, and the quantitative claims from the seminal paper are still quoted.[2] Backofen emphasized the importance of the problem being addressed:

> "The protein structure prediction problem is one of the most (if not the most) important problem in computational biology. This problem consists of finding the conformation of a protein with minimal energy. Because of the complexity of this problem, simplified models like Dill's HP-lattice model have become a major tool for investigating general properties of protein folding."[3]

Lau and Dill wrote [4] that theoretical estimations predict the fraction of random protein sequences that fold into stable, native-like structures to lie between 10^{-6} and 10^{-10}. The authors point out, correctly, that probabilities much lower than these would raise doubts whether a functional protein could arise naturally, and comment:

> "This general argument has become of some importance as support for the view that proteins could not have arisen from natural prebiotic chemical processes on earth and as support for creationism."[4]

We concur that a very low probability would fit better with creationism than materialism, but following good scientific methods, we should first collect the raw data and then decide about its significance.

The results of the model led Lau and Dill to claim,

> "On this basis, extrapolation shows that for chains of n = 100 monomer units, the fraction of these 2D sequences which fold is in the range of 10^{-6} to 10^{-10}, depending on the strictness of the criterion used to classify a sequence as a folding molecule."[4]

The authors conclude the paper with a sweeping statement about what they claim to have shown:

> "And it suggests how molecules as complex as catalytic globular proteins could have arisen so readily in simple prebiotic solutions, wherein only a virtually negligible fraction of all possible sequences would have been sampled during the origins of life."[5]

The experiment

Proteins are modeled as chains composed of only two elements: H (hydrophobic) or P (polar, meaning non-hydrophobic). Each element, which represents an amino acid, can occupy a single square within a two-dimensional lattice. Thus, each amino acid is connected to its chain neighbour(s) and can also be in contact with 'topological' neighbours adjacent in space (figure 1).

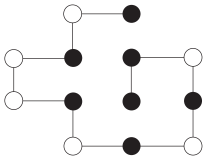

Figure 1. Two-dimensional model for a protein chain based on two kinds of amino acid: dark = hydrophobic; light = non-hydrophobic. Each amino acid can occupy only one position in the lattice.[6]

Every H-H contact between topological neighbours is assigned a contact energy <0 and every other interaction between neighbour pairs has energy equal to 0. For short chains of n = 10 to 25 elements,[7] every accessible conformation was found using an algorithm they programmed. The number of H-H contacts defined the energy of the whole chain, and sequences with lowest energy represent protein 'native states'.

The reasons for the project were stated in the Abstract:

> "… we address two questions regarding the evolution and origins of globular proteins. (i) How will protein native structures and stabilities be affected by single and double-site mutations? (ii) What is the probability that a randomly chosen sequence of amino acids will be compact and globular under folding conditions?"[8]

What is now required, based on goal (ii) is of paramount importance: a way to map the sequences of Hs and Ps, and their calculated 'energies', to the propensity of real proteins to fold, and remain folded, in a discrete native state. Without such calibration, the computer program is worthless. In addition, one must not forget that for proteins the conformation in the lowest-energy state will not automatically be a properly folded and soluble protein, based on well-defined alpha helices and beta coils. The protein could simply produce a tangled mess or be attached to other proteins in some amorphous manner.

The criterion used for whether folding occurred was described as follows:

> "To decide whether a sequence is a 'folder' or not, we use the compactness $(1\text{-}3)n = t/t_{max}$, where t is the total number of topological neighbours in a given conformation and t_{max} is the maximum possible number of topological neighbours which could be achieved by any conformation of a given chain length. Below we consider different criteria for globularity, ranging from the most strict ($t = t_{max}$) to less strict ($t >= t_{max} - 2$); a conformation is considered folded if it satisfies this criterion under folding conditions."[8]

No empirical justification for the '$(1\text{-}3)n = t/t_{max}$' assumption was provided, based on known protein chemistry realities, and the reference cited[9] was not helpful. Essentially, the authors assume that if enough hydrophobic residues (linked amino acids) are neighbours in space (but not adjacent along the chain sequence) in the two-dimensional lattice, then a native-like fold will form.

The assumption was then used to answer the question posed in the Abstract:

> "What fraction of sequence space corresponds to compact globular molecules under folding conditions? Let N(n) equal the number of sequences of chain length n in the sequence space, and let $N_f(n)$ equal the number of sequences which fold—i.e., those in which the conformations of lowest energy (native states) are maximally compact or nearly so. The simulations show that the fraction of sequence space corresponding to folding molecules diminishes approximately as
>
> $$N_f(n)/N(n) = ka^n, \qquad (1)$$
>
> where the constants are k = 2.04 and a = 0.792 for $t = t_{max}$, k = 4.87 and a = 0.813 for $t >= t_{max} - 1$, and k = 4.06 and a = 0.860 for $t = t_{max} - 2$."[4]

Table 1. Proportion of stable, native folds, based on an extrapolation from a computer model using the relationship ka^n, for a hypothetical two-dimensional lattice of n amino acids.[4]

	k	a	ka^n (n = 50)	ka^n (n = 100)	ka^n (n = 150)	ka^n (n = 300)
$t = t_{max}$	2.04	0.792	1.8E-05	1.5E-10	1.3E-15	8.5E-31
$t = t_{max-1}$	4.87	0.813	1.6E-09	5.0E-09	1.6E-13	5.2E-27
$t = t_{max-2}$	4.06	0.86	2.2E-03	1.1E-06	6.1E-10	9.1E-20

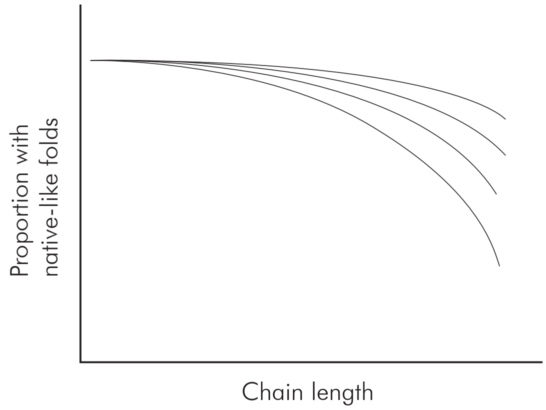

Figure 2. Proportion of random polypeptides with native-like folds decreases with chain length but the quantitative relationship is unknown.

Now, everyone agrees that the proportion of polypeptides leading to native-like folds decreases with chain length, but it remains to be clarified experimentally, or based on sound theoretical principles, how rapidly. Figure 2 illustrates various possibilities.

The parameters for the equation of form ka^n were determined empirically, based only on the speculative criterion for folding mentioned above. We used the equation with several values of *n*, leading to the results given in table 1. There we find for *n* = 100 the range of 10^{-6} to 10^{-10}, which the authors claimed[4] to be a realistic estimate for the proportion of stable, native-folded proteins in a random sequence of amino acids that length.

Evaluation

The authors clearly believe they have a useful tool to understand biological proteins, and use the computer program to make a variety of broad statements about the quantitative effects of single and double mutations; the proportion of lower-energy degenerate states and when they occur[10]; and about the nature of the internal protein core. They use phrases like, "the simulations show" and "the simulation results" many times.

The reader of their paper is confronted with the obvious question of the relevance to real proteins. Amazingly, the results from their model were not calibrated in any manner against real biological data. Although they comment on their model and approach that "It has no adjustable parameters or additional *ad hoc* assumptions",[8] this claim is misleading. For example, the criterion for whether a sequence is a folder or not has not been verified quantitatively in any manner. As we'll see below, the authors examined between *n = 10* to *25* points on a lattice. Still, many refer to this model with very little semblance to known protein chemistry, while extrapolating to $n = 70$ and higher.[2] The readers who have not examined the original source for such quantitative values are led to believe that the values have a biological basis in reality; when they are purely speculative models based on minimal assumptions.

The authors concluded that globular proteins arose readily in simple pre-biotic solutions,[5] but they did not address any of the chemical and thermodynamic issues which preclude this bold extrapolation. Additional considerations must include the need for homochiral amino acids; avoidance of side chain reactions; intra-molecular peptide bond formation; instability of growing polypeptides in water; the presence of other reactive chemicals; and so on.

To be as generous as possible to the authors' model, we shall assume that optically pure amino acids only were available, react to form only linear proteins, and do not agglomerate with other chains. Nevertheless, the effects of some Neglected Realities (NR1–NR5 below) need to be estimated if any value is to be extracted from their report. Then an estimate for the proportion of random polypeptides of chain length *n* amino acids which lead to a native, stable fold can be obtained.

Neglected Realities (NR) for biological proteins

NR1: folding degrees of freedom

In the computer model, the next member of each chain can theoretically assume any of three conformations (figure 3).

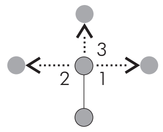

Figure 3. For a chain in two dimensions, an additional monomer can theoretically assume any of three positions in the lattice.

Those who have played the computer game 'Pacman'[11] should recognize the principle: one can only move forward, left, or right. Given that at least two residues are needed for a polypeptide and that the end of any chain has a maximum of three possibilities at which the next residue could be placed on the lattice, the maximum number of chains for the computer simulation is easily seen[12] to be given by:

$$\text{maximum number of chains, C} = 3^{(n-2)} \qquad (1)$$

for *n* residues, where *n* is two or greater.

However, a little thought shows that as chain size increases in the two-dimensional lattice, the maximum number of variants possible becomes far less than implied by formula (1). This is illustrated in figure 4**.** The quantitative effect is significant, since every chain end with less than three subsequent alternatives limits the entire branch of possibilities starting from that point.

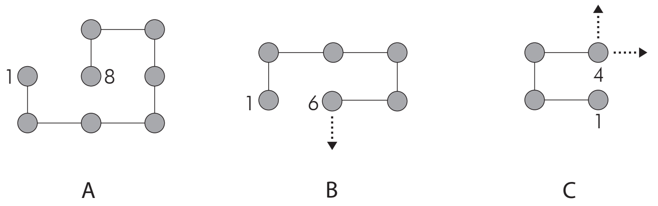

Figure 4. As the two-dimensional lattice becomes larger, ever more ends of the chain will have fewer degrees of freedom for additional growth. **A**: zero chain extension possibilities. **B**: one degree of freedom at the end position. **C**: two degrees of freedom at the end position.

In real proteins, the potential to assume erroneous conformations is much greater than implied by the lattice model,[13] and the requirements to ensure funneling into the native fold configuration increases accordingly. There are two ways to calculate the number of folding conformations of proteins. Using *n* for the number of amino acids, some researchers[13] use the formula 3^{2n-2} and others[14] use 8^n. For most practical purposes it does not too much matter which approach is used, since the resulting discrepancy calculated rarely affects the conclusions made.[15]

For an average-sized 150-amino-acid domain of a protein,[16] the potential number of folds is vastly greater than the two-dimensional model suggests, even ignoring the fact that eq. (1) overestimates the maximum number of variants by many orders of magnitude:

$$3^{2n-2} / 3^{n-2} = 3^{298} / 3^{148} = 4\text{x } 10^{71} \qquad (2)$$

This fact is not properly captured by the authors through limiting the computer runs to between 10 and 25 'circles' on the two-dimensional lattice. Note that the computer model only permits exact 90° angles, whereas protein can adjust angles in three dimensions over several residues to accommodate crowded folds.

In addition, the spatial relationship between residues involved in alpha helices[17] and beta sheets[18] shows no resemblance to the flat two-dimensional shapes assumed for the lattice. For example, a regular alpha helix has 3.6 residues per turn[19] due to hydrogen bonds formed between residues i and i+4. A variant helix called 3_{10} is based on hydrogen bonds formed between residues i and i+3.[20] There are even examples of hydrogen bonds formed between residues i and i+5, known as the pi helix.[20]

Beta sheets, parallel and anti-parallel, also involve hydrogen bonds.[21] In addition, there are many kinds of turns which also represent secondary structure.[22] And all these secondary structures, fundamental to producing native folds, have little to do with whether the residue is polar or hydrophobic.

NR2: packing geometries

In the model, only two elements are allowed to build the chain, 'H' or 'P'. In reality, the side chains of real amino acids differ considerably in size, shape and electrical charge.[23] The overall environment of the folded protein

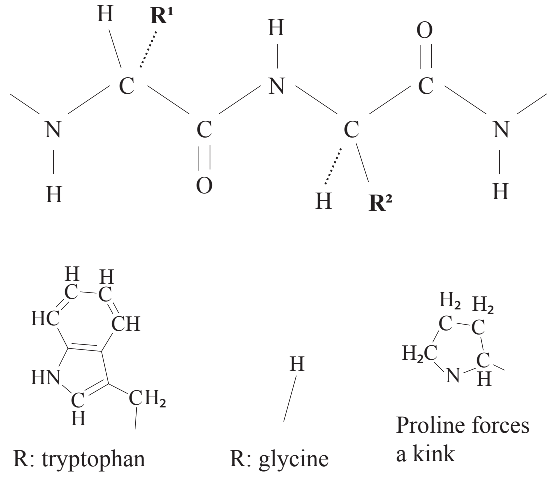

Figure 5. Dipeptide formed from the condensation of two amino acids. The twenty natural amino acids each have a different side group (R^1 and R^2); these differ considerably in size and chemical characteristics.

brings different side chains together, and to pack properly various constraints must be met. If the side groups brought together in the protein core are too small, then cavities are introduced into which water can penetrate, producing an instable conformation. If some of the side chains are too large, stable packing is also hindered.

Figure 5 illustrates how different amino acids can be. For example, compare the size of the side chain, R, for glycine with that for tryptophan. The residues classified as H (hydrophobic) can have dramatically different effects on the ability to pack into a native-like fold, depending on which residues are located nearby.

Very different residues, like the three shown in figure 5, were indiscriminately treated as equivalent ‘H’ in the computer model. But large side groups, like tryptophan in figure 5, cannot be compressed near other large side chains in the core. Proline, the third hydrophobic amino acid shown in figure 5, forces a sharp ‘kink’ in the protein, severely disturbing the geometry of most protein cores except when the rest of the chain has been appropriately designed.

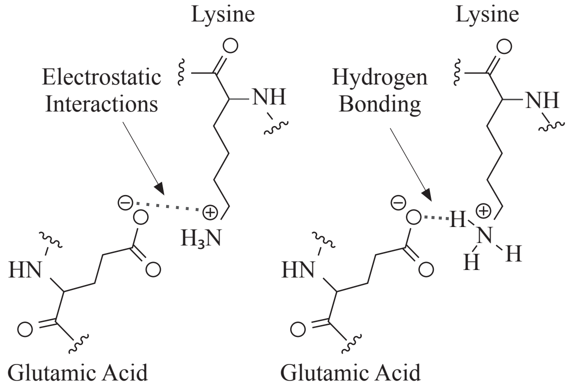

Figure 6. Example of salt bridges with the amino acids glutamic adic and lysine.[24]

Salt bridges[24] (figure 6) can form with some residues, contributing to the stability of native folds, and so can disulfide bonds[25] formed between the thiol groups of cysteine residues (figure 7).

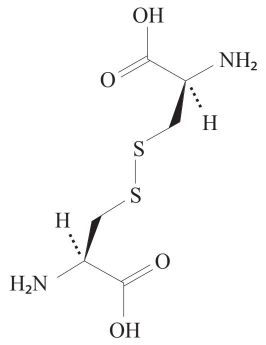

Figure 7. Disulfide bond formed between two cysteine amino acids.[25]

NR3: minimum energy necessary to remain folded

The proportion of random *real* polypeptides able to fold, and to remain folded reliably, was not addressed in this paper. We quoted above what the authors did: they compared the total number of topological neighbours in a given conformation with the maximum possible number of topological neighbours which could be achieved, and then concluded, on the basis of two-dimensional hydrophobic interactions, that a large number would be ‘folders’.[8]

The model is too crude to be quantitatively relevant, even as a rough indicator. The actual distance between the hydrophobic side chains and strength of the specific interactions will determine the stabilizing contribution. Treating all non-‘H-H’ interactions as neutral is wrong, since electrically charged amino acids like aspartic acid and glutamic acid (negatively charged); and lysine, arginine and histidine (positively charged) will generally provide a stabilizing contribution if in contact with water instead of in the core.

NR4: conformations with local minimum energies

The authors assume that the most stable conformation *will inevitably be found* and ignore the critically important question of how long this may take. In fact, there are far more possible conformations than could ever be examined by trial and error. This protein-folding problem is well known and referred to as the Levinthal Paradox,[13] which we shall revisit (later).

We saw in eq. (2) that the lattice model understates absurdly the number of incorrect ways a protein could fold. Finding the native-like fold among random sequences in a time span relevant for biological purposes would be prohibitive.

The fundamental issue is that acceptable sequences must guide the folding process, like a funnel, into the correct conformation. This means that alternative, low-energy conformations along the path to the correct fold must be prevented. Locally minimum energy states could be stable enough to hinder unfolding to permit additional trials in the search for a better conformation. Another perspective is that the protein could spend too much time in countless alternative low-energy, but functionally worthless, topologies for that sequence to be useful (figure 8).

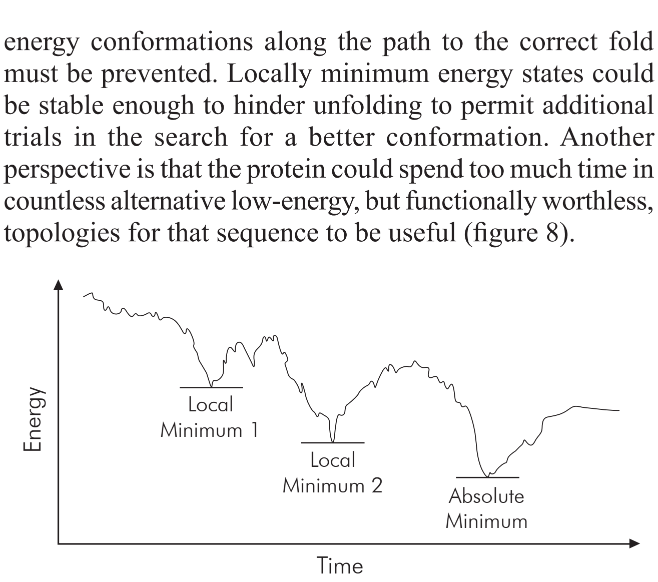

Figure 8. A protein can remain trapped in a local energy minimum along the multitude of paths leading to an absolute lowest energy conformation.

One can use the two-dimensional lattice to illustrate the difficulty. During chain folding, the next circle assumes up to three possible positions. This will often result in variants with a very stable local topology (figure 9). Those variants would include a portion locked into a particular arrangement and cannot unfold easily (especially under biologically relevant time scales) to search for the best-of-all conformation. A collection of random sequence will have folded members distributed throughout a vast range of conformations. But the pathway to getting there is littered with pitfalls. Therefore, useful real proteins must hinder the formation of undesired false minima.

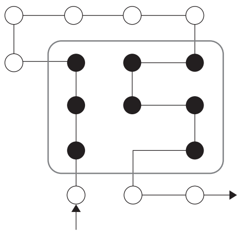

Figure 9. In a two-dimensional lattice permitting only two kinds of elements, neighbouring dark ones not adjacent along the chain are assumed to provide favourable interactions. Once a conformation with a favourable local minimum energy is formed, it will be difficult to unfold to permit searching for an absolute best conformation.

NR5: binary assumption

The authors assume that only hydrophobic ('H-H') interactions need to be considered to model protein folding. Therefore, their amino acids are either hydrophobic, 'H', or not hydrophobic, 'P'. They refer to experimental work done elsewhere which used optimal proportions of amino acid chains with these characteristics, for example, "… the experiments of Rao *et al.*[26] on random polymer sequences of lysine, alanine, and glutamic acid". Lysine is charged, alanine is hydrophobic and glutamic acid is polar (hydrophilic). These are optimized choices known to generate secondary structure and stable folds. No attempt was made to calibrate these other kinds of experiments with biologically relevant amino acid proportions.

Similar work has been done in professor Sauer's lab at MIT.[27] In their case leucine (hydrophobic), glutamine (polar, hydrophilic), and a small amount of Arginine were used. Their choice was based on the high propensity of leucine and glutamine to form alpha helices, which should facilitate formation of stable, folded proteins.

But none of these others studies, which tested large numbers of binary sequences, were able to identify a native-like folded protein, as also reported in this journal.[28] Therefore, it is not justified to restrict the parameters which cause folding to only hydrophobicity considerations.

Levinthal Paradox revisited

The polypeptide backbone chain, figure 5, is represented by sequences –N-C_a-C-N-C_a-C-, where C refers to the carbonyl carbon. The next carbon along the side chain attached to C_a is called C_b, and so forth. Atoms C-N-C_a-C define the torsion angle **f,** whereas N-C_a-C-N- defines angle y. Conformations about these two angles create unfavourable contacts with neighbouring atoms which limit the conformation space used by native folds. Ramachandran diagrams[29] plot y angles on the y-axis and the f on the x-axis, both from –180 to 180 degrees. Experimentally observed angles for residues from biological proteins have been used to generate contours of the regions used in nature, one contour representing b-strands, the other a-helices and a small region of left- handed helices, which are very rare. These contours show that only about 30% of the space of backbone conformations are used by biological proteins per residue.[30] Since most proteins consist of hundreds of residues, a random sequence is extremely unlikely to find itself in a biologically relevant conformation.

Glycine, with a hydrogen as the side chain, is not subject to this restriction. Turns are also not taken into account. Since the average protein size is greater than 300 residues[31–33] (a conservative figure often used by creation scientists), we can simplify our calculations to obtain a reasonable estimate for the portion of backbone conformational space used by native-like folds:

$$(0.3)^{300} = 10^{-157}$$

Residues in an acceptable contour region may need to unfold if the true native fold is to be found. For example, a portion of the backbone which initially would possess angles consistent with a-coils may need to rearrange entirely to participate in a b-strand with distant regions of the folding protein. To illustrate, translation occurs at rates of about 20 residues per second,[34] yet many soluble proteins fold entirely within 0.1 seconds. This suggests that in the cell restrictions exist for the folding process, or, alternatively, there might be a way to reopen the chain to permit the correct final states to be formed.

The lattice model oversimplifies and neglects this reality. Finding the lowest energy conformation would not occur, even in billions[13] of years, for most random polypeptides of average protein length. Therefore, specific amino acid sequences must be present which funnel the folding into a limited number of paths. And this implies that a very small number of random sequences would provide the necessary physical details to ensure proper folding in a biologically relevant time scale.

Furthermore, we suggest that some well-defined amino acids must be present at various positions to *prevent* incorrect possibilities from being explored. The folded proteins are generally only between 15 and 40 kJ/mole stabler [35] than the open chain,[36] a remarkably low value, given how often and how quickly the native state is found. One way to prevent incorrect conformations may be through the use of well-designed *turns*. Turns are combinations of amino acids leading to specific shapes and features, and connect alpha helices and beta coils in proteins. We suggest that careful analysis and designed mutations of turns may reveal their importance not only in generating the correct final layout but also in preventing incorrect alternatives during the process of searching for the native state. Turns are good candidate portions of proteins to test this hypothesis: are undesired conformations often hindered by the sequences present in turns.

It is also possible that undiscovered cellular equipment besides chaperons help guide the folding process. These nano-machines could be located as part of ribosomes or near them. Formation of individual secondary structure in an open chain is not such an easy matter.

> "Although helices have regular repeating hydrogen bonds coupled with a uniformity of bond lengths and angles this periodicity masks their marginal stability. In an isolated state most helices will unfold and only synthetic propyl-alanine helices are reasonably described as stable. Initiation of helix formation is a slow and unfavourable process. This arises because five residues must be precisely positioned to define the first hydrogen bond between residues 1 and 5."[37]

Summary

This paper showed that the program developed by Lau and Dill is severely hampered. As a computer simulation, it has little, if any, relevance to the subject modelled. More severe, still, is that it has never been tested nor calibrated against real biological data. This seems to be a general trend in Evoland. Apparently, unproven models and unrealistic simulations are taken as substitutes for real biology.

Although the high proportion claimed for random polypeptides able to produce native-like folds might be attractive to those favouring a naturalistic origin for globular proteins, modern investigators know that this widely cited paper does not produce quantitatively meaningful information. Referring specifically to Lau and Dill's work, Backofen and Will comment:

> "The main motivation for studying simplified protein models is to be able to predict model structures much more quickly and more accurately than is possible for real proteins. However, up to now there was a dilemma: the algorithmically tractable, simple protein models can not model real protein structures with good quality and introduce strong artefacts."[38]

Conversely, the computer expert may confirm that the algorithm to find all conformations for any length *n* of residues would be correct for the lattice model, but have no idea whether the model represents anything biologically relevant.

It is imperative that any computer simulation must be tested against real data before making quantitative claims. This has not been done in this report,[1] nor any follow-up work, but then surprisingly, the results were extrapolated to larger lattices with no knowledge of how real proteins react to increasing chain length. In addition, the authors ignored the time necessary to find proper folds and just assumed that if they existed, nature would easily find them.

The quantitative claims are little more than speculation and could easily be wrong by a hundred or more orders of magnitude for reasonably sized proteins. Quoting the quantitative conclusions is irresponsible and to be avoided even if in harmony with one's preconceptions. Some creation scientists might feel that the lower bound of 10^{-10} for a hundred residues is acceptable, since 3 or 4 such domains in a protein, or large domains, would imply for an average size protein a proportion of about 10^{-35}, but this temptation is to be avoided.[39] Inevitably, others would argue that the smaller domains arose somehow and then nature 'only' had to link them together.

It is regrettable that we don't have a simplifying algorithm, properly calibrated, to calculate the correct proportion[43] of random protein sequences able to generate native-like folds. A computer program able to perform the calculations in a physically realistic manner to test random sequences exceeds by far the world's current computing resources.

What is the correct proportion? The data in this six-part series do not provide the answer, except to confirm that it is lower than the estimates each author suggested. It would be sensible to examine the facts *first* and *then* to reason out an interpretation within our worldview. For example, if large proportions of random sequences were to fold in energetically almost degenerate states, then so be it. The creationist would then point out the consequence, that very few would be found at a given time in the biologically relevant conformation and would not remain folded corrected for very long in that shape.

Papers modelling reality should be much more critically scrutinized for their scientific merit. A reason why they are not receiving adequate critique is the need for readers to cover multiple disciplines. In the paper mentioned above,[1] the person knowledgeable in cell biology might be intimidated by unfamiliar techniques or phrases, such as “a recently developed lower-resolution lattice statistical mechanics model of protein folding”.[40] Papers such as these often also include some very intimidating and unfamiliar mathematics[4] which a biologist may feel unqualified to challenge. Therefore, creation scientists should not only be specialists in their own fields, but should also be encouraged to become multidisciplinary.

References

1. Lau, K.F. and Dill, K.A., Theory for protein mutability and biogenesis, *Proc. Natl. Acad. Sci. USA* **87**:638–642, 1990.
2. Graziano, J.J., Wenshe, L., Perera, R., Geierstanger, B.H., Lesley, S.A. and Schultz, P.G., Selecting Folded Proteins from a Library of Secondary Structural Elements, *J. Am. Chem. Soc.* **130**(1):176–185, 2008. doi:10.1021/ja074405w. See p. 186.
3. Backofen, R., The protein structure prediction problem: a constraint optimization approach using a new lower bound, *Constraints* **6**(2/3):223–255, 2001; ISSN:1383-7133.
4. Lau and Dill, ref. 1, p. 641.
5. Lau and Dill, ref. 1, p. 642.
6. Lau and Dill, ref. 1, p. 640.
7. Lau and Dill, ref. 1, p. 641; see their fig. 5.
8. Lau and Dill, ref. 1, p. 639.
9. Lau, K.F. and Dill, K.A., A lattice statistical mechanics model of the conformational and sequence spaces of proteins, *Macromolecules* **22**:3986–3997, 1989.
10. Lau and Dill, ref. 1, p. 639: “Singly degenerate sequences occur largely in the composition range of 30–70% H.”
11. en.wikipedia.org/wiki/Pac_man.
12. Beginning with a two-residue chain, we add a third residue to the end, noting that there are three possible positions on the lattice. Adding a fourth residue at the end of the preceding three-member chain produces a total of three times more chains. For *n* amino acids and *C* chains we easily see the relationships: {n:C}: {2:1} {3:3} {4:9} {5:27} {6:81} and so on, from which the relationship for C = $3^{(n-2)}$ was derived.
13. en.wikipedia.org/wiki/Levinthal_paradox.
14. Lodish, H., Berk, A., Zipursky, S.L., Matsudaira, P., Baltimore, D. and Darnell, J., *Molecular Cell Biology*, 4th ed., W.H. Freeman and Company, p. 62, 2000: “Any polypeptide chain containing *n* residues could, in principle, fold into 8^n conformations. This value is based on the fact that only eight bond angles are stereochemically allowed in the polypeptide backbone.”
15. For example, for *n* =100: 3^{2n-2} gives 3.0 x 10^{94} and 8^n leads to 2.0 x 10^{90}; for n=150: 3^{2n-2} gives 1.5 x 10^{142} and 8^n leads to 2.9 x 10^{135}.
16. The average size of a globular domain, according to the CATH database is 153 residues: Shen, M-y., Davis, F.P. and Sali, A., The optimal size of a globular protein domain: A simple sphere-packing model, *Chemical Physics Letters* **405**:224–228, 2005.
17. For a high-quality image see: gibk26.bse.kyutech.ac.jp/jouhou/image/protein/2ndst/alpha_st.gif.
18. For a high-quality image see: gibk26.bse.kyutech.ac.jp/jouhou/image/protein/2ndst/beta_anti_st.gif.
19. Whitford, D., *Proteins: Structure and Function*, John Wiley & Sons, p. 41, July 2008, ISBN-13: 978-0471-49893-3.
20. Whitford, ref. 19, p. 44.
21. Whitford, ref. 19, p. 46.
22. Whitford, ref. 19, pp. 47–48.
23. www.bio.davidson.edu/biology/aatable.html.
24. en.wikipedia.org/wiki/Salt_bridge_%28protein_and_supramolecular%29.
25. en.wikipedia.org/wiki/Disulfide_bond.
26. Rao, S.P., Carlstrom, D.E. and Miller, W.G., Collapsed structure polymers. A scattergun approach to amino acid copolymers, *Biochemistry* **13**:943–952, 1974.
27. Davidson, A.R. and Sauer, R.T., Folded proteins occur frequently in libraries of random amino acid sequences, *Proc. Natl Acad. Sci. USA* **91**:2146, 1994.
28. Truman, R., The proportion of polypeptide chains which generate native folds—part 1: analysis of reduced codon set experiments, *J. Creation* **25**(1):77–85, 2011.
29. en.wikipedia.org/wiki/Ramachandran_plot.
30. Whitford, ref. 19, p. 49.
31. The thousands of proteins found in yeast have, on average, 466 residues. en.wikipedia.org/wiki/Protein.
32. Drummond, D.A., Bloom, J.D., Adami C., Wilke, C.O. and Arnold, F.H., Why highly expressed proteins evolve slowly, *Proc. Natl Acad. Sci. USA* **102**:14338–4343, 2005. Here an average-length of 415 AA was reported for yeast proteins.
33. As a test, I arbitrarily examined the first Proteobacteria listed in the Microbial Genome Database, mbgd.genome.ad.jp/. This was *Caulobacter crescentus*, the genes of which are listed at mbgd.genome.ad.jp/htbin/MBGD_gene_list.pl?spec=ccr. After removing t-RNAs, hypothetical proteins and 5S ribosomal RNA, I found 2,210 identified proteins with an average size of 376 residues.
34. Whitford, ref. 19, p. 415.
35. Protein stabilities can be reported in calories or joules; 1 kcal = 4.186 kJ.
36. Whitford, ref. 19, p. 403.
37. Whitford, ref. 19, p. 409.
38. Backofen, R. and Will, S., A constraint-based approach to fast and exact structure prediction in three-dimensional protein models, *Constraints* **11**(1):5–30, January 2006; ISSN:1383-7133.
39. Domains are believed to fold independently. If the probability of proper folding for a single domain in a protein is 10^{-10}, then three independent domains of that size on a protein would have a probability of 10^{-10x3}.
40. Lau and Dill, ref. 1, p. 638.

Royal Truman has bachelor’s degrees in chemistry and in computer science from SUNY Buffalo, on M.B.A. from the University of Michigan, a Ph.D. in organic chemistry from Michigan State University and post-graduate studies in bioinformatics from the universities of Heidelberg and Mannheim. He works for a large multinational in Europe.