The proportion of polypeptide chains which generate native folds—part 4: reusing existing secondary sequences

Royal Truman

Various experimental approaches to generate native-like protein folds from scratch have failed. This has prompted origin-of-life researchers to develop a new strategy. A large research team randomly combined gene sections able to generate 605 helices, 328 strands and 246 loops, which were found in stably folded biological proteins. Not all 108 clones were examined in detail, but evaluation of the most promising candidates failed to identify any with native-like folds. These experiments demonstrate the difficulty of generating proteins with all the necessary structural parts present harmoniously to produce a stable native-like fold, and illustrates the immense difficulty for this to occur by chance.

In Part 3 of this series, we mentioned[1] that several research groups have attempted to design novel artificial proteins based on a specific target fold[2–4] or binary patterns with remarkably little success.[5] We concluded that native-like folds were not formed, an opinion shared by others:[6]

> "A more recent report details the use of cassettes of binary patterned residues around known protein motifs that are ultimately randomly assembled without fitting any designed folding constraint. However, these latter methods have not yet yielded stable, novel folds under physiological conditions."[7]

This prompted a research group[6] to run an experiment in which it semi-randomly combined sequences from real biological genes that are responsible for the secondary structures found in the resulting proteins. In this manner portions of proteins known to be involved in secondary and tertiary structure would be present in the novel proteins and likely provide the scaffold for novel folds.

The outcome

First, all sequences leading to alpha-helices, beta-strands, and loops found in 190 non-redundant protein present in *E. coli* were identified. From these a library consisting of 605 helices, 328 strands and 246 loops was assembled,[8] using optimal stoichiometries. Carefully designed *polymerase chain reactions* linked these strands and loops together to produce new genes, to ensure that the same orientation of the original helices and sheets found in *E. coli* would result.

The PCR fragments formed were submitted to agarose gel electrophoresis to remove lengths <300 bp[10] so that proteins of 100 or more amino acids would result when expressed in *E. coli*. On average about seven fragments for each clone were used in the artificial genes created. These were fused to a sequence which would produce a GFP[9] protein at one end of the product translated from the gene.[10] The new genes were inserted into a plasmid (EGFP fusion vector JG1)[10] and these transferred into *E. coli* hosts.

The modified *E. coli* were sorted using fluorescence activated cell sorting (FACS), retaining those showing a high fluorescence (GFPuv fluorescence > 80 RFU). This fluorescence comes from the small *GFP folding reporter* mentioned above, deliberately generated as part of each member in the library of the artificial genes. In the resulting protein, high fluorescence may indicate this GFP portion is soluble and stable to proteolysis,[10] and therefore potentially pointing to it being located in a folded environment.

Of 1 x 10^8 cells sorted by fluorescence, 1.6 x 10^6 (1.6% of the total) were identified[10] as high fluorescence, meaning they might have folded. This was too many for detailed analysis, so 1,149 of the most promising clones were chosen for further analysis.

Of the 1,149 clones, 70 were selected on the basis of various criteria: 44 with very high fluorescence (>1,000 RFU); 22 with significant solubility scores,[11] of which fourteen were already part of the high fluorescence set;[12] and 18 clones lacking high fluorescence and significant solubility scores but displaying[12] high Nickel-HRT assay scores (associated with protein solubility).

It was possible to obtain sufficient protein from seven of the 70 clones for further study. One was a duplicate and another was omitted for further investigation since a nonsense mutation (which leads to a premature 'stop' codon) had occurred.

The dye BisANS[13] was added to each of the five clones, since its fluorescence increases upon binding to unfolded and molten globule proteins. Fluorescence was measured at pH values varying[14] between 3.6 and 9.6. The changes seem to indicate that the five proteins are undergoing pH induced structural changes, one trait of folded proteins.

Next, ultraviolet circular dichroism spectra were obtained[14] for the five clones at wavelengths between 250 nm and 200 nm. This is a method used to identify the presence of secondary structure, especially alpha helices.

In one clone, no helical content was determined. For the other four clones, a considerable amount of alpha-helical structure was identified, although the values deviated dramatically from that expected based on the characteristics of the segments they were built from (which were designed from natural proteins found in *E. coli*). Also disconcerting, the AGADIR program that predicts alpha-helical structure predicts these four clones should have little to no alpha-helical character.

Clone Name	CD spectra	Predicted from library content	AGADIR
5.1	40	22	1.2
5.6	29	6	7.8
5.26	15	23	4.8
5.31	47	17	0.3

Table 1. Percent alpha-helical content predicted using three methods.[14]

Thermal denaturation of the four proteins was determined by monitoring circular dichroism at 222 nm. The temperature was raised slowly and then lowered again between 20°C and 90°C[11] leading to near super-imposable folding and unfolding profiles for three of the clones (not clone 5.26).

Finally, the researchers subjected the proteins of three clones (5.6, 5.26 and 5.31) to NMR analysis, a key method to determine protein structures.

> "The chemical shift dispersion of the 1-D NMR spectra of proteins is a qualitative measure of protein folding and tertiary packing of protein side chains."[15]

The NMR spectra resemble the results from a reference, unfolded protein (FKBP12), leading the authors to conclude that

> "The selected clones should therefore not be considered viewed as 'native-like' proteins but rather 'molten globule like'."[15]

Discussion

In this study, a large number of gene sequences known to produce secondary structures of properly folded biological proteins were combined to generate novel artificial genes. Surprisingly, none of the carefully constructed and tested artificial proteins showed the expected protein folds. As usual in science, though, when unexpected results occur, one can often postulate a good reason, which may or may not be correct. In this case the authors suggest,

> "The lack of correlation between the fragment origins and the structure they appear to assume in the selected polypeptides is not too surprising since it is well known that primary sequence is not the sole determinant of secondary structure formation."[15]

Nevertheless, the authors stated[15] that they have no explanation for the positive CD spectra that indicated the presence of alpha-helices.

The proteins generated by the synthetic/artificially constructed (i.e. designed) genes was only examined for some of the 10^8 *E. coli* isolated. The results reported permit some estimates for the proportion of sequences which would fold for the carefully optimized original library.

1) We can assume, as the authors do, that of all the modified *E. coli* created, those displaying GFPuv fluorescence of less than 80 RFU did not contain a synthetic gene which leads to a native-like protein. Then, of the original pool, 0.16 are candidates for having folded properly.

2) Of the 1.6 x 10^6 cells comprising promising clones, 1,149 clones were selected. We will assume that the selection was random, and that no criteria, such as strength of fluorescence to pick the best candidates was used. This is not clear from the paper, and would seem like a foolish decision. Of this set, the 70 which displayed evidence they might contain folded proteins were selected: 70/1,149 = 0.061.

3) Of the seventy best candidates so far, seven were picked. It is not clear how to interpret this: perhaps some of the remaining sixty three were unsuitable. Possibly abnormal behaviour on the separation ("*Comassie-stained bands by SDS PAGE after a single nickel chelating FPLC purification step*"[12]) column was observed. Of the seventy potential candidates, none were native-like folded proteins, leading to a factor ranging between 1/70 = 0.014 and 7/70 = 0.1. In favour of the naturalist view, we'll use the more generous value of 0.1.

These three terms allow us to state that a library of proteins carefully designed as described contained a proportion less than 0.16 x 0.061 x 0.1 = 1 x 10^{-3} folded proteins. The only quantitative conclusion possible is that the proportion of native-like folds among random polypeptides must be many orders of magnitude lower than 10^{-3}.

The popular hypothesis that random gene shuffling produced countless new functional proteins is not compatible with these results. Based on detailed bio-informatic analysis from an evolutionary perspective, Conant and Wagner reached a similar conclusion:

> "We have studied shuffling in genes that are conserved between distantly related species. Specifically, we estimated the incidence of gene shuffling in ten organisms from the three domains of life: eukaryotes, eubacteria, and archaea, considering only genes showing significant sequence similarity in pairwise genome comparisons. We found that successful gene shuffling is very rare among such conserved genes."[16]

Conclusion

Random shuffling of pre-existing protein-domains appears not to be a mechanism to account for novel functional proteins. Rather, protein-building elements need to be carefully organized together for a new protein to produce a native-like fold. This all points in the direction of design, hence a designer.

References

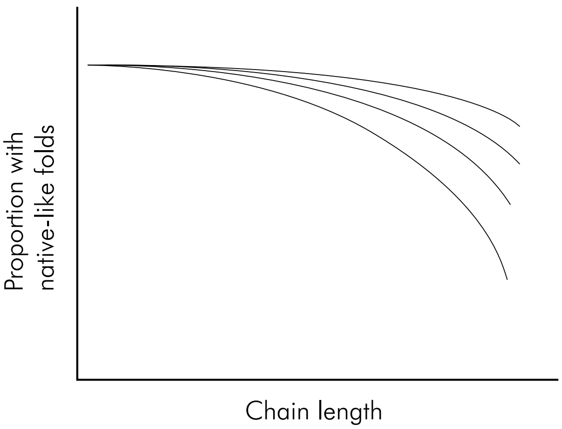
1. Truman, R., part 3. The proportion of polypeptide chains which generate native folds: Based on designed secondary structures, *J. Creation* **25**(3):102–105, 2011.
2. Kamtekar, S., Schiffer, J.M., Xiong, H., Babik, J.M. and Hecht, M.H., Protein design by binary patterning of polar and nonpolar amino acids, *Science* **262**:1680–1685, 1993.
3. Hecht, M.H., De novo proteins from designed combinatorial libraries, *Protein Science* **13**:1711–1723, 2004.
4. Kuhlman, B., Dantas G., Ireton, G.C., Varani, G., Stoddard B.L. *et al.*, Design of a novel globular protein fold with atomic-level acuracy, *Science* **302**:1364–1368, 2003.
5. Matsuura, T., Ernst, A. and Plückthun, A., Construction and characterization of protein libraries composed of secondary structures modules, *Protein Science* **11**:2631–2643, 2002.
6. 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.
7. Graziano *et al*, ref. 6, p. 177.
8. Graziano *et al*, ref. 6, p. 181.
9. GFP: Green Fluorescent Protein, often used for laboratory purposes, en.wikipedia.org/wiki/Green_fluorescent_protein.
10. Graziano, ref. 6, p. 182.
11. Graziano, ref. 6, p. 180. Screening of target clones for soluble protein was performed in parallel. Solubility scores were calculated by an automated procedure (Ni-HRP A_{420nm}).
12. Graziano, ref. 6, p. 183.
13. 5,5'-bis(8-anilino-1-naphthalenesulfonate)
14. Graziano, ref. 6, p. 184.
15. Graziano, ref. 6, p. 185.
16. Conant, G.C. and Wagner, A., The rarity of gene shuffling in conserved genes, *Genome Biology* **6**:R50.1– R50.14, 2005; doi:10.1186/gb-2005-6-6-r50.

Royal Truman has bachelor's degrees in chemistry and in computer science from SUNY Buffalo, an 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.

Errata

Journal of Creation **25**(2)

- Truman, R., The proportion of polypeptide chains which generate native folds—part 2: theoretical studies.
  On p. 100, figure 2 should be:



- Tomkins, J. and Bergman, J., The chromosome 2 fusion model of human evolution II: Re-analysis of the genomic data.
  On p. 114, in the caption to figure 2, 'TTAGGn' should read 'TTAGGGn'.

- Hartnett, J., Does the Bible really describe expansion of the universe?
  On p. 126, second column,
  – line 16, should read '… the heavens are spread out (Hebrew, *mathach*)'.
  – line 35, should read '… Hebrew words, *natah* and *mathach*'.