The proportion of polypeptide chains which generate native folds—part 3: designed secondary structures

Royal Truman

In this six-part series, we are evaluating those reports which claim that a high proportion of random polypeptide chains would produce protein native-like folds. Often claims are made by others who cite these papers without explaining the designed elements which went into the experimental protocol. A thorough understanding of the work actually performed, or neglected, is necessary before general statements can be made for truly random protein chains of biologically relevant length. Here, we discuss a paper published by a research group at the University of Zurich, which performed one of the most sophisticated attempts to create new but unspecified folds through intelligent design. The results show that, even though all information known about the requirements to generate secondary structures (binary patterning; length; N-caps; C-caps; optimal choice of amino acids) went into the design, under natural aqueous conditions no examples of native folded proteins were found.

Several research groups[1–3] have attempted to create novel proteins with new polypeptide sequences. Usually *a specific target fold is selected* and they attempt to reproduce its shape and characteristics using other combinations of amino acids. Although such artificial proteins can be intelligently designed, creating novel ones which fold properly has been far more challenging than expected. None of these studies provided the quantitative data we wished to find to help estimate the proportion of *random* sequences between 50 and 350 AA (amino acids) which would produce native-like folds under biologically relevant conditions. Therefore, those studies which targeted a specific fold, almost always unsuccessfully, will not be reviewed in this series of papers.

Our purpose here is to evaluate the evolution-inspired premise that native-like folds are easy to generate naturally. To address this, we will discuss research involving a series of expressed cassettes containing code for semi-random short protein structures. The design of these structures was based on statistical properties of protein secondary structures, to provide optimal chances of generating artificial proteins with native-live folds. As we will see, about 10^9 sequences were generated, but none of them were reported to possess properties typical of true native-like states. Only under abnormal conditions, such as high salt concentration, were some poorly-defined globular-like properties identified, lacking a discrete tertiary structure.

To determine whether many or few polypeptide sequences would produce native-like folds, the Zürich group[4] constructed artificial genes based on polar and non-polar residue patterns which favour secondary structure[5] (α-helices, β-strands and turns, i.e. the components of protein folds). These genes produced several libraries of synthetic proteins which should possess the following characteristics: all α-helix; all β-strand; and α-helix plus β-strand, having an average length of 100 amino acid residues. As a rule, about 90% of a biological protein consists of secondary structure.[6] Since random sequences will rarely form secondary structures consistently under biologically relevant temperatures and aqueous conditions, the authors decided to restrict the space of possibilities to semi-random chains, most likely to form such structures. Otherwise very few, if any, artificial proteins with properly folded conformations would be identified.

The experiment

The experiment conducted by the Zürich group was carefully designed and was based upon the upfront knowledge we have on naturally occurring functional proteins. In order to optimize the chance to find 'random' new combinations, they started with short amino acid patterns known to favour α-helix and β-strand formation. The designed details include 1) kinds of residues in tailored environments with other specific amino acids, 2) that α-helices in nature average about ten amino acids and 3) that β-strands typically use about five amino acids. In addition, tri-nucleotides were used to code for the amino acids most likely to generate the secondary structures desired. Short chains were produced and these linked together to express proteins with a high proportion of stable secondary structures.

Construction of α-helices

To construct these common protein structures, blocks of five and ten amino-acids were designed. To define one end (the N-cap side) of the helix, the researchers benefited[7] from statistical studies[8] that show amino acids Asn, Ser, Asp, Thr, or Gly followed by Pro or Glu occur with highest frequency. They used Ser followed by either Pro or Glu, mixed in equal proportions. At the other end (the C-cap), they used the statistically preferred residue Gly.[8]

The choice of amino acids used between the bounding N-cap and C-cap was guided by the studies of West and Hecht.[9] They had created a database of penta-peptide polar (*p*) and non-polar (*n*) patterns found empirically in helices (the '*binary patterning*'). The most frequently occurring

patterns, *pnnpp*, *nnppn* and *nppnp*, were used.[4] to design the five-residue α-helix containing cassettes. For the ten-residue cassette, the most frequent pattern, *ppnnppnnpp*, was selected.

Using the tri-nucleotide technology developed by Virnekäs,[10] the five amino acids most frequently occurring in helices[7] were used: Gln, Glu, Lys, Arg and Ala for the polar ones (*p*); and Ile, Phe, Leu, Met and Ala for the non-polar residues (*n*). Ala was allowed in the experiments at the *n* and *p* location, since it is known to have the highest propensity to form helices (figure 1).

α-helix design

α_5 cassette

-[S-(P/E)-*nppnn*-G]-
-[S-(P/E)-*ppnnp*-G]-
-[S-(P/E)-*nppnp*-G]-

α_{10} cassette

-[S-(P/E)-*ppnnppnnpp*-G]-

p: Q, E, K, R or A
n: I, F, L, M, or A

β-strand design

β cassette

-[S-R-*pnpnp*-G]-
-[S-R-*npnpn*-G]-

p: V, Y, T or S
n: I, F, L or V

Amino acid abbreviations

G = glycine	T = threonine
A = alanine	C = cysteine
V = valine	Y = tyrosine
L = leucine	N = asparagine
I = isoleucine	Q = glutamine
M = methionine	D = aspartic acid
F = phenylalanine	E = glutamic acid
W = tryptophan	K = lysine
P = proline	R = arginine
S = serine	H = histidine

turn design

turn cassette

-[S-(D/S/N/T)-(D/S/T/P)-(D/S/N/A)-G]-

Figure 1. Design of the secondary structure (helix, strand and turn) cassettes later linked to form new artificial genes. Polar and non-polar amino acids are shown as *p* and *n*, respectively.[14]

Construction of β-strands

Binary patterns *npnpn* and *pnpnp* are the most commonly found folding patterns in beta-strands, especially in domains exposed to the exterior of the folded protein.[11] To construct these patterns, Arg was also designed into the cassette, because it is known to be a good N-capping amino acid for β-strands.

In addition, Val, Tyr, Thr and Ser were chosen[7] to encode *p* in the binary pattern. The reason for this is that these amino acids are known to have the highest frequency on the water-exposed side of the alternating β-strand.[12] In the buried hydrophobic interior of proteins, the order of preference is, in contrast, Val, Cys, Phe, Ile and Leu. Therefore, these AA were used to encode the non-polar (*n*) residues (figure 1). Cys was excluded[7] to avoid the complications which would result from the formation of random disulfide bridges.

Construction of β-turns

β-turns are another important structure component of proteins. They reverse the peptide chains at the surface of the protein and permit them to become globular.[13] The most common, *type I turn*,[13] was used in the experiment. It consists of a hydrogen bond formed between the main-chain position *i* and *i+3*. Amino acids between *i* and *i+3* were included based on their known frequency of occurrence[13] in the turn cassette, which was later bonded to other cassettes. Asp, Ser, Asn, and Thr were chosen for position *i*; Asp, Ser, Thr, and Pro for position *i+1*; Asp, Ser, Asn, and Ala for position *i+2*; and Gly for position *i+3* (figure 1).

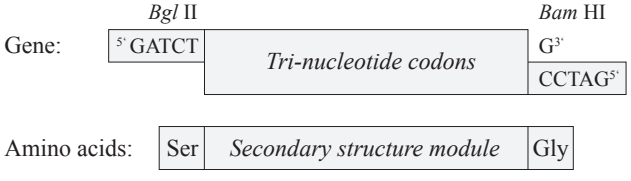

Figure 2. The structural cassettes carry the enzymatic restriction sites *Bg* III and *Bam* HI. This permits them to be ligated together. The amino acid pair Gly-Ser connects the cassettes. A, C, G and T are the DNA four nucleotides used in the genetic code.

The cassettes prepared were amplified using PCR and digested with restriction enzymes *Bam*HI and *Bgl*II, leading to the overhangs shown in figure 2. Since DNA base-pairing is between A-T and C-G, the ends of the structural cassettes can be ligated together to form longer oligonucleotides. Multiple rounds of ligation and digestion with the enzymes were carried out to generate four libraries of around 100 amino acids in length.

The libraries were constructed as follows:

Library 1 ($\alpha_{10}\alpha_5 t$): Using structural modules from α_{10}, α_5 and turn cassettes

Library 2 ($\alpha_{10} t$): α_{10} and turn cassettes

Library 3 ($\alpha_{10}\alpha_5\beta t$): α_{10}, α_5, β and turn cassettes

Library 4 (βt): β and turn cassettes.

All libraries were cloned in a plasmid[15] expression vector[16] (derivative of pQE16[17]) obtained from Qiagen[18] with a T5 promoter (which is recognized by *E. coli* RNA polymerase) and a strong Shine-Dalgarno sequence[19] (which included a histidine tag at the C-terminal to promote transcription). The vectors were introduced into *E. coli* and exposed to ampicillin and chloramphenicol[20] to kill the bacteria lacking the modified plasmid. The expressed proteins of the seven most promising colonies from the four libraries were examined using various standard tests.

Results

Library 1 and 2 (α10α5t and α10t)

Three proteins were examined, but the authors did not explain the basis for their selection nor why only three were chosen. Under *normal, biological aqueous* conditions the circular dichroism (CD) spectra indicated a substantial proportion of random coil formation (i.e. absorption having a minimum at 200 nm) and little or no evidence for α-helices.[21]. Only under very high concentrations of NaCl did the CD spectra suggest formation of helices. Salts affect properties of

proteins such as stability and solubility. Therefore, the helix formation and the ensuing tertiary structure formation may be at least partially due to the high concentration of salt added.[23]

Gel filtration experiments indicated that the three proteins were larger than expected for native globular proteins, which is typical for molten globule, but atypical for properly folded proteins. Sedimentation equilibration experiments in $(NH_4)_2SO_4$ indicated that one of the proteins formed as a trimer and the other two as an ill-defined mixture of probably monomer, dimers and trimers. The fact that they could be isolated, and had not been proteolyzed in *E. coli*, is likely due to their aggregate states, which protect them. Had they been degraded by proteases in the bacteria, this could be interpreted as lack of stable folding. But the opposite, lack of degradation, need not demonstrate stable folding, as suggested in the paper.[22]

1,8-Anilinonaphtalene sulfonate (ANS) is a fluorescent probe that can detect the accessible hydrophobic core area. Detectable binding is consistent with molten-globule states of a protein, whereas it generally shows only weak binding to native or totally unfolded structures.

All three proteins tested showed no detectable binding to ANS in the absence of salts, but significant binding in 3.5 M NaCl.[23] The lack of binding in the absence of NaCl indicates lack of secondary or tertiary structure, in agreement with the CD tests. In *high concentrations* of NaCl, these three proteins do form helices, based on the CD results and gel filtration experiments, and binding to ANS is consistent with formation of a molten-globule.

Urea equilibrium unfolding experiments in the presence of both NaCl showed no cooperative unfolding behaviour.

Taking all the results together, the experiments imply that under aqueous conditions as typically found in nature, these three proteins remain in an unfolded state, but in the presence of high concentrations of NaCl, they form a molten-globule state.[24] The results provide no evidence that the sequences in this library led to a native folded state. That only three candidates were selected, possibly on the basis of their greater quantities, implies to us that the authors found this library to be unpromising.

Library 3 and 4 (α10α5βt and βt)

Two proteins from the $\alpha_{105}\beta t$ library were chosen for further characterization.[24] Why only two were selected, after so much effort in creating the library, and why these particular two, was not discussed by the authors. Neither bound to Ni-NTA agarose beads, except in the presence of 6 M GdnHCl. This indicates that the histidine tag is inaccessible and implies aggregation of the protein. Furthermore, one of the proteins precipitated above 0.5 M NaCl and the other at 1.5 M NaCl. These facts imply that these proteins have exposed hydrophobic patches due to the β sheet.[24]

CD spectra of both proteins showed the presence of a random coil conformation and no α-helical character, except for a very weak signal in the presence of high concentration of NaCl. The signals in the absorption spectra were much too weak to determine whether caused by α-helix or β-sheet formation.

Two proteins from library βt were expressed and studied in more detail. Electrospray mass spectrometry showed they had the expected mass, but they could not be solubilized in high concentration of denaturants like 8 M urea or in 6 M GdnHCl. Electron micrographs showed clear evidence they were clumping together, forming amyloid-like fibrils. This effect illustrates the difficulty of generating β-sheets which could in principle participate in native folds built using amino acid patterns of *pnpnp* and *npnpn* (figure 1).

The failure to identify properly folded artificial proteins in this library is not entirely surprising. Matsuura *et al.* point out: "West *et al.* found that a combinatorial library of β-strands, in which each module has alternating polar and non-polar residues, tends to form soluble amyloid fibrils."[25]

Evaluation of results

The work described in this paper is one of the most carefully planned experiments designed to produce folded proteins using semi-random sequences.

It is of note that introducing β-strand cassettes into the protein made them very aggregation-prone. This experimental insight is useful and led the authors to conclude that "the presence of β-strands seems to require precise topological arrangement to prevent aggregation".[26] Furthermore, "It follows that in all-β proteins a number of mechanisms must be operative to prevent this fibril formation, including topology enforcing turns (underrepresented in our libraries) and deviations from the regular spacing of the β-strands in the sequence."

Also, native-like folded new proteins were not found, in spite of the astuteness which went into the design. Remarkably, here the authors offer the correct insight: "Because of the restricted nature of our cassettes (only 4–5 amino acids allowed at most positions) and the absence of any selection applied, we cannot reasonably expect to obtain the finely tuned packing required for natively folded proteins."

In other words, evolution could not have started with far fewer than the twenty natural amino acids, and built proteins from them.

From figure 1 we see that seventeen different amino acids were used in the polypeptides, placed at the residue position most likely to be useful, but this was not sufficient. The specific residues which should be used at each position[27] depend on the overall context and design of the whole protein in its final folded state.

Quantitative conclusions

As scientists, we are naturally disappointed that the authors were unable to identify new kinds of folds. This could have provided insights into the puzzle of how and why naturally occurrring proteins fold so quickly and reliably in the face of so many alternative conformations and such small difference in energy between the most stable native-fold and the random state. New folds might also offer possibilities for medical or industrial applications. But our purpose here is to

evaluate the evolution-inspired premise that native-like folds are easy to generate naturally.

The library contained about 10^9 different gene sequences, but the number of plasmids actually introduced into *E. coli* was not estimated in the paper. Presumably a correspondingly large number was used. The implication is that a pool of several million different sequences were available, from which the most promising ones were selected for characterization. A high proportion of the bacteria generated measurable amounts of protein: "If only clones with correct reading frames are considered, 37, 72, 90, and 10% of libraries $\alpha_{10}t$, $\alpha_{10}\alpha_5\beta t$, βt, *and* $\alpha_{10}\alpha_5$, respectively, have detectable protein expression."[20]

Conclusion

The data reported are important, since they show that even sequences only 100 amino acids long, optimally designed, have a chance under 10^{-9} of producing true protein folds. Even though all information known about the requirements to generate secondary structures (binary patterning; length; N-caps; C-caps; optimal choice of amino acids) went into the design, under natural aqueous conditions no examples of native folded proteins were found.

The authors believe that a few of the proteins generated have some kind of molten globule structure. However, they admit,

> "Although natural proteins, in general, have a distinct global free energy minimum that allows them to fold into one unique structure, molten-globule–like properties probably lack a distinct global minimum, and thus do not have specific tertiary structure."[26]

And a non-specific tertiary structure cannot be used for biological purposes, as an evolutionary starting point, nor be of value which natural selection could fine-tune.

Even a biological protein possesses a multitude of false folded states in a local minimum energy state. However, a sequence which folds close to the native state is needed as an evolutionary starting point.

The results imply that the proportion of sequences leading to a proper fold from among truly random polypeptides must be very small. 10^9 sequences 100 AA long were generated and the most promising ones were analyzed without success. Random sequences would be orders of magnitude even less likely to produce native folds.

References

1. 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.
2. Hecht, M.H., De novo proteins from designed combinatorial libraries, *Protein Science* **13**:1711–1723, 2004.
3. 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.
4. Matsuura, T., Ernst, A. and Plückthun, A., Construction and characterization of protein libraries composed of secondary structures modules, *Protein Science* **11**:2631–2643, 2002.
5. Whitford, D., *Proteins: Structure and Function*, John Wiley & Sons, July 2008, ISBN-13: 978-0471-49893-3, See chapter 3.
6. Chothia, C., Principles that determine the structure of proteins, *Annu. Rev. Biochem.* **53**:537–572, 1984.
7. Matsuura, ref. 4, p. 2632.
8. Blundell, T.L. and Zhu, Z.Y., The α-helix as seen from the protein tertiary structure: A 3-D structural classification, *Biophys. Chem.* **55**:167–184, 1995.
9. West, M.W. and Hecht, M.H., Binary patterning of polar and non-polar amino acids in the sequences and structures of native proteins, *Protein Science* **4**:2032–2039, 1995.
10. Virnekäs, B., Ge, L., Plückthun, A., Schneider, K.C., Wellnhofer, G. and Moroney, S.E., Tri-nucleotide phosphoramidites: Ideal reagents for the synthesis of mixed oligonucleotides for random mutagenesis, *Nucleic Acids Res.* **22**:5600–5607, 1994.
11. Broome, B.M. and Hecht, M.H., Nature disfavors sequences of alternating polar and non-polar amino acids: Implications for amyloidogenesis, *J. Mol. Biol.* **296**:961–968, 2000.
12. Zhu, Z.Y. and Blundell, T.L., The use of amino acid patterns of classified helices and strands in secondary structure prediction, *J. Mol. Biol.* **260:**261–276, 1996.
13. Wilmot, C.M. and Thornton, J.M., Analysis and prediction of the different types of β-turn in proteins, *J. Mol. Biol.* **203**:221–232, 1988.
14. Matsuura, ref. 4, p. 2633.
15. A plasmid is a DNA molecule that is separate from, and can replicate independently of, the chromosomal DNA. Plasmids are often found naturally in bacteria; en.wikipedia.org/wiki/Plasmid.
16. An *expression vector*, otherwise known as an *expression construct*, is generally a plasmid that is used to introduce a specific gene into a target cell; en.wikipedia.org/wiki/Expression_vector.
17. www.embl-hamburg.de/~geerlof/webPP/vectordb/bact_vectors/maps_seqs_mcs/pQE/pQE-16_map.pdf.
18. www.qiagen.com/products/protein/expression/qiaexpressexpression system/c-terminuspqevectorset.aspx.
19. "The Shine-Dalgarno sequence (or Shine-Dalgarno box) is a ribosomal binding site in the mRNA, generally located 8 basepairs upstream of the start codon. The Shine-Dalgarno sequence exists only in prokaryotes. The six-base consensus sequence is AGGAGG; in *E. coli*, for example, the sequence is AGGAGGU. This sequence helps recruit the ribosome to the mRNA to initiate protein synthesis by aligning it with the start codon."; en.wikipedia.org/wiki/Shine-Dalgarno_sequence.
20. Matsuura, ref. 4, p. 2640.
21. Matsuura, ref. 4, p. 2635.
22. Matsuura, ref. 4, p. 2634.
23. Matsuura, ref. 4, p. 2636.
24. Matsuura, ref. 4, p. 2637.
25. Matsuura, ref. 4, p. 2638.
26. Matsuura, ref. 4, p. 2639.
27. In these kinds of studies it is common for the plasmid to include a gene which offers protection against a toxin which is added to the culture, allowing those who have successfully ingested the plasmid to be distinguished from the rest.

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.