STRATEGIES FOR SELECTION FROM PROTEIN LIBRARIES COMPOSED OF DE NOVO DESIGNED SECONDARY STRUCTURE MODULES

TOMOAKI MATSUURA* and ANDREAS PLÜCKTHUN*
Biochemisches Institut, Universität Zürich, Winterthurerstr. 190, CH 8057, Zürich, Switzerland
(author for correspondence, e-mail: plueckthun@bioc.unizh.ch, phone: +41-1 635 5571, fax: +41-1 635 5712*

(Received 1 November 2002; accepted in revised form 3 April 2003)

Abstract. As more and more protein structures are determined, it has become clear that there is only a limited number of protein folds in nature. To explore whether the protein folds found in nature are the only solutions to the protein folding problem, or that a lack of evolutionary pressure causes the paucity of different protein folds found, we set out to construct protein libraries without any restriction on topology. We generated different libraries (all α-helix, all β-strand and α-helix plus β-strand) with an average length of 100 amino acid residues, composed of designed secondary structure modules (α-helix, β-strand and β-turn) in various proportions, based primarily on the patterning of polar and non-polar residues. From the analysis of proteins chosen randomly from the libraries, we found that a substantial portion of pure α-helical proteins show properties similar to native proteins. Using these libraries as a starting point, we aim to establish a selection system which allows us to enrich proteins with favorable folding properties (non-aggregating, compactly folded) from the libraries. We have developed such a method based on ribosome display. This selection is based on two concepts: (1) misfolded proteins are more sensitive to proteolysis, (2) misfolded and/or aggregated proteins are more hydrophobic. We show that by applying each of these selection criteria proteins that are compactly folded and soluble can be enriched over insoluble and random coil proteins.

Keywords: binary patterning, combinatorial approach, ribosome display, protein fold

1. Introduction

The number of all possible sequences of a protein with a length of 100 amino acids is $20^{100} \approx 10^{130}$. If we take into account proteins with different lengths, this number increases even more, and thus it is impossible that nature has explored all of the sequence space available to proteins. When the issue comes to protein structures, it is not clear how many possible folds can exist, although the number might eventually be estimated when structure predictions from the primary sequence of a protein become possible. Nevertheless, it seems to be clear from the available sequences and structures that the number of protein folds in nature are limited

* Present address: PRESTO, Japan Science and Technology Corporation, Symbiotic Engineering Laboratory, Department of Bioinformatic Engineering, Graduate School of Information Science and Technology, 2-1 Yamadaoka, Suita, Osaka 565-0871, Japan.

Origins of Life and Evolution of the Biosphere **34:** 151–157, 2004.
© 2004 *Kluwer Academic Publishers. Printed in the Netherlands.*

(Thornton *et al*., 1999). The question thus becomes: have all possible stable folds been realized by nature or has only a small subset – through diversification and selection – given rise to the repertoire of modern proteins? A number of theoretical studies have addressed these questions (Govindarajan and Goldstein, 1996; Helling *et al*., 2001; Yue and Dill, 1995). Our desire, in contrast, is to attack this problem experimentally by generating novel proteins and subsequently investigating their structures.

We describe here the concept for this approach and the experimental tools currently available. While some parts of this work have already been published elsewhere in greater detail (Matsuura *et al*., 2002; Matsuura and Plückthun, 2003), we want to consider the strategies in context.

As most folds consist mainly of secondary structure modules, we began by generating a protein library comprised of designed secondary structure elements (Matsuura *et al*., 2002). Furthermore, using this library as a starting point, we aim to mimic part of the process of protein evolution by using ribosome display (Hanes and Plückthun, 1997). Ribosome display is an *in vitro* selection system which has been shown to be a powerful tool to select antibodies, and more recently, other natural scaffolds (P. Forrer *et al*., unpublished) against many different ligands with high affinity and specificity. Using ribosome display, we have now developed a method to perform selections based on the folding properties of displayed proteins, such as solubility and protease resistance (Matsuura and Plückthun, 2003).

2. Results and Discussions

Discrete secondary structure elements can be encoded by specific patterning of polar and non-polar residues. Indeed, the occurrence of such patterns is highly correlated with the existence of α-helices and β-strands in natural proteins (West and Hecht, 1995). For the construction of the protein library, we first designed three secondary structure motifs (α-helix, β-strand and turn) mainly based on the patterning of polar and non-polar residues (binary patterning) (for details see Matsuura *et al*., 2002). These motifs were prepared using trinucleotide building blocks (Virnekäs *et al*., 1994), which allowed us to tailor the amino acid mixtures to favor the formation of the desired secondary structure element. These elements were then polymerized in different combinations to produce three different protein libraries (all α-helix, all β-strand and both α-helix and β-strand) with an average length of about 100 amino acid residues (Figure 1).

Amino acid sequences of randomly chosen members of each library showed low homology to any known sequence, indicating that the library members are distant in sequence space from known natural proteins. Nevertheless, three randomly chosen proteins from the all α-helix libraries were found to be soluble, monomeric and helical in the presence of high concentrations of NaCl. Moreover, these proteins showed cooperative urea equilibrium unfolding behavior in the pres-

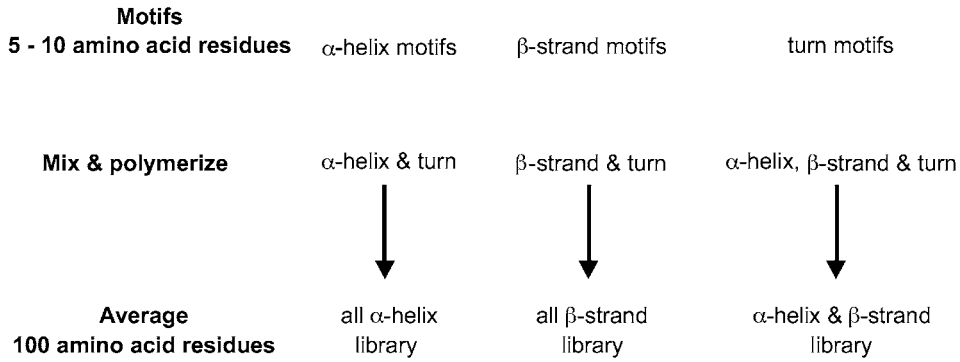

Figure 1. Schematic overview of the protein library construction based on the *de novo* designed secondary structure modules. For details, see Matsuura *et al.* (2002).

ence of high concentrations of $(NH_4)_2SO_4$. However, these proteins also bound to 1,8-anilinonaphtalene sulfonate (ANS), which indicates that these proteins have molten-globule like properties. Despite the fact that the members of our libraries are distant from natural sequences in protein sequence space, we estimated that about 1 in 6 members of the all-α-helix libraries exhibit properties of the molten globule state. In contrast, the libraries incorporating β-strand motifs produced proteins that are more aggregation-prone. Therefore, while with α-helix modules it is relatively easy to obtain proteins with many basic characteristics of natural globular proteins, the presence of β-strands seems to require more precise topological arrangements to prevent aggregation (for details, see Matsuura *et al.*, 2002).

Natural globular proteins are not only soluble but are also resistant to proteolytic digestion. These properties were acquired as a consequence of natural evolution, which occur by many consecutive cycles of mutation and selection. Our desire is to establish an evolutionary system, which can perform rounds of selection and diversification rapidly, and is capable of applying three different selection pressures: for function, solubility and protease resistance. We have developed such a system using ribosome display (for details, see Matsuura and Plückthun, 2003).

Ribosome display is based on the translation of peptides or proteins from mRNA to generate ternary complexes consisting of mRNA, the encoded protein and the ribosome, thereby coupling phenotype and genotype. Ribosome display is shown schematically in Figure 2a. In ribosome display, ternary complexes are generated using an *in vitro* translation system with mRNAs lacking stop codons. Transferring the complexes into a high Mg^{2+} concentration in an ice-cold buffer immediately after *in vitro* translation also increases the stability of the ternary complexes, presumably by stabilizing the ribosome. Once stable ribosomal complexes are generated, these can be used in affinity selection. Those complexes, which display proteins with sufficiently high affinity to a ligand, will remain bound after the washing steps. Ultimately, mRNAs that encode such proteins are isolated either by eluting with EDTA, which leads to dissociation of the ternary complexes, or by

Figure 2. (a) Principle of ribosome display. For details, see text. (b) Selection based on protease resistance. (c) Selection based on protein hydrophobicity. For details, see Matsuura and Plückthun, 2003.

adding a high concentration of free ligand as a competitor. The recovered mRNAs are reverse transcribed and amplified by PCR to generate the library for the next round of affinity selection. As mutations are automatically introduced during the PCR amplification step, diversification steps are an inherent part of the ribosome display procedure (Figure 2a). Since ribosome display works entirely *in vitro*, a larger library can be screened than with *in vivo* based methods. Moreover, because of the defined nature of the system, possible problems of genetic instability or toxicity of the proteins which might occur *in vivo* can be circumvented.

We have now further adapted ribosome display to select for proteins based on their folding properties. This is based on two concepts: (1) misfolded proteins are more sensitive to proteolysis and (2) misfolded and/or aggregated proteins are more hydrophobic.

Misfolded and unstructured proteins are not sufficiently compact to be protected from proteolytic digestion. Based on this idea, we attempted to establish a system capable of enriching folded proteins over random coil and aggregated proteins using ribosome display (Figure 2b). We used three different proteins in the model selection, (i) proteins that exhibit random coil formation (RC) as evidenced by CD spectroscopy; (ii) proteins that form amyloid-like fibrils (FB); and a compact model protein, (iii) protein D (pD), a capsid protein from the lambda phage (Yang *et al*., 2000). When starting from a mixture of the mRNA encoding the three proteins, we saw a clear enrichment of pD in increasing concentrations of trypsin. We could also see that pD is the most stable protein among them. These results indicate that a selection based on the protease stability of the displayed protein can be achieved with ribosome display (for details see Matsuura and Plückthun, 2003).

Misfolded or aggregated proteins, in general, have a greater exposure of hydrophobic residues than those which are folded properly. Based on this idea, we attempted to establish a system capable of enriching less hydrophobic proteins over those with higher hydrophobicity (Figure 2c). Using same three proteins (RC, FB and pD), protein-mRNA-ribosome ternary complexes were incubated with butyl-agarose in the presence of 2.7 M KCl. Those which did not bind to the butyl ligands were rescued. We again could see the enrichment of pD, which indicates that a selection based on the hydrophobicity of the displayed protein can be achieved by ribosome display (for details see Matsuura and Plückthun, 2003). When both strategies were combined, a very significant enrichment of a native protein both over random coil proteins and fibril forming proteins could be achieved (Figure 3).

The increasing number of whole genome sequences and the experimentally determined 3D-structures leads to a number of fundamental questions regarding the origin of life. Why are the number of folds limited in nature? Why are some folds frequently used and others not? Are these properties inherently limited to the fact that proteins are made of a standard set of 20 amino acids?

Our approach is to tackle these questions by repeating the course of evolution from a synthetic basis. The basic concept is to develop both polypeptide libraries and selection systems for this purpose. Our libraries are very different from natural proteins, but still contain the fundamental building blocks of secondary structure elements. We feel that this is a useful compromise, as total random sequence of the library will "dilute" sequences with biophysical properties even remotely similar to natural proteins. The selection system, ribosome display, can because of its in vitro nature, use libraries with high diversity, and we have shown a variety of strategies how one can select for the biophysical properties of the protein.

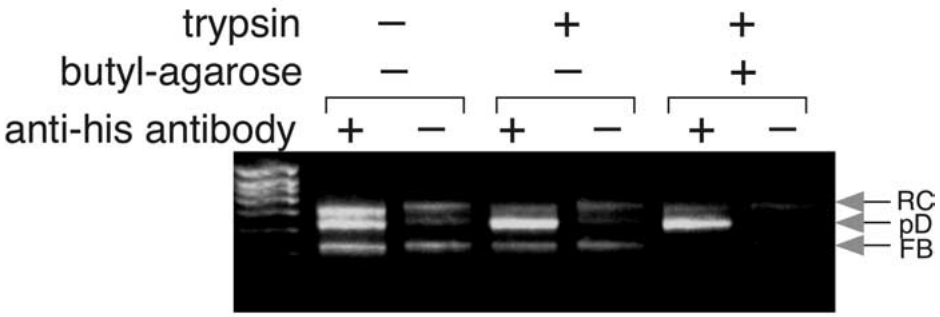

Figure 3. When the selections based on surface hydrophobicity and protease resistance of the displayed protein are combined, the native protein pD can be enriched over random coil proteins (RC) and a fibril forming one (FB). Ternary ribosomal complexes of the model proteins were incubated either with (+) or without (–) 70 nM trypsin, followed by an incubation with (+) or without (–) butyl-agarose beads. Affinity selection of his-tagged protein was performed in the presence (+) or absence (–) of anti-his-tag antibody. RT-PCR products were analyzed by agarose gel electrophoresis. (Reproduced with permission from Matsuura and Plückthun, 2003).

We are now able to mimic important steps of evolution by combining the secondary structure based protein libraries reported here and ribosome display which allows rapid cycles of diversification and selection.

Acknowledgement

We would like to thank Andreas Ernst for helpful discussions.

References

Govindarajan, S. and Goldstein, R. A.: 1996, Why Are Some Proteins Structures So Common?, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 3341–3345.

Hanes, J. and Plückthun, A.: 1997, In vitro Selection and Evolution of Functional Proteins by Using Ribosome Display, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 4937–4942.

Helling, R., Li, H., Melin, R., Miller, J., Wingreen, N., Zeng, C. and Tang, C.: 2001, The Designability of Protein Structures, *J. Mol. Graph. Model.* **19**, 157–167.

Matsuura, T., Ernst, A. and Plückthun, A.: 2002, Construction and Characterization of Protein Libraries Composed of Secondary Structure Modules, *Protein Sci.* **11**, 2631–2643.

Matsuura, T. and Plückthun, A.: 2003, Selection Based on the Folding Properties of Proteins with Ribosome Display, *FEBS Lett.* **539**, 24–28.

Pearl, F. M., Lee, D., Bray, J. E., Buchan, D. W., Shepherd, A. J. and Orengo, C. A.: 2002, The CATH Extended Protein-Family Database: Providing Structural Annotations for Genome Sequences, *Protein Sci.* **11**, 233–244.

Thornton, J. M., Orengo, C. A., Todd, A. E. and Pearl, F. M.: 1999, Protein Folds, Functions and Evolution, *J. Mol. Biol.* **293**, 333–342.

Virnekäs, B., Ge, L., Plückthun, A., Schneider, K. C., Wellnhofer, G. and Moroney, S. E.: 1994, Trinucleotide Phosphoramidites: Ideal Reagents for the Synthesis of Mixed Oligonucleotides for Random Mutagenesis, *Nucleic Acids Res.* **22**, 5600–5607.

West, M. W. and Hecht, M. H.: 1995, Binary Patterning of Polar and Nonpolar Amino Acids in the Sequences and Structures of Native Proteins, *Protein Sci.* **4**, 2032–2039.

Yang, F., Forrer, P., Dauter, Z., Conway, J. F., Cheng, N., Cerritelli, M. E., Steven, A. C., Plückthun, A. and Wlodawer, A.: 2000, Novel Fold and Capsid-Binding Properties of the Lambda-Phage Display Platform Protein gpD, *Nat. Struct. Biol.* **7**, 230–237.

Yue, K. and Dill, K. A.: 1995, Forces of Tertiary Structural Organization in Globular Proteins, *Proc. Natl. Acad. Sci. U.S.A.* **92**, 146–150.