Function driven protein evolution. A possible proto-protein for the RNA-binding proteins.

Jacquelyn S. Fetrow[1] and Adam Godzik[2]

[1]University at Albany, SUNY, Albany, NY
[2]The Scripps Research Institute, La Jolla, CA

Summary. We introduce a hypothesis that present day proteins evolved from "proto-proteins," small 15-20 residue peptides with some elements of secondary structure and primitive function. Increasingly stable and functional proteins arose by adding structural elements to produce the small domains or protein modules that we would recognize today. From this point of view, the surprising similarities between small structural fragments of large proteins, that are usually taken as examples of convergent, function-driven evolution, are interpreted in exactly the opposite way—as traces of common evolutionary origin. As an example, a hypothetical evolutionary tree for two families of RNA binding proteins, the OB fold, a family of all β proteins, and RBD fold, an α/β protein family is presented. We argue that both protein families could have evolved from the same RNA-binding proto-protein, which had a form of β-loop-β RNA binding motif.

1. Introduction

Proteins are very efficient, highly optimized molecular machines that are products of an extremely long evolution-driven design process. We can learn about how they evolve from studies of sequence divergence between proteins from homologous organisms. However, since species differentiation was a relatively recent process as compared to the earliest evolution of proteins and their functions, most of this information doesn't tell us much about how the folds themselves evolved or about how function and structure influenced each other during evolution. Even the simplest organisms have much of their molecular machinery essentially complete, with almost all protein functional groups and fold families already in place [1].

Almost all existing studies of molecular evolution are based on comparison of protein or DNA sequences. *De facto* in the field of protein analysis, the term "homology" has become synonymous with above-random sequence similarity; whereas, homology truly means a evolutionary divergence from a common ancestor. Unfortunately, sequence based methods are not able to follow evolutionary relations back in time beyond the point where the accumulated number of mutations make similarities between sequences of truly homologous proteins indistinguishable from random similarities between sequences of unrelated proteins. In most cases this point is reached rather quickly, when the structure and function of a protein are still very close to what we observe today.

The inability to follow sequence similarity in very distantly related homologous proteins has led to many claims in the literature of convergent evolution, that proteins have a different origin, but converge to a similar structure or function [2]. As has been pointed out, claims of convergent evolution are difficult to verify and such claims should be made quite carefully [3]. A recent example of such a claim is in the convergent evolution of nucleic acid binding domains [4]. These researchers presented a case for convergent evolution of two protein families: the RBD RNA-binding protein family and the cold shock domains, nucleic acid binding proteins which fold into the OB protein fold motif. Members of both of these families bind nucleic acids and both exhibit the RNP1 and RNP2 sequence signature on adjacent β-strands for RNA binding. Despite the functional and sequence similarity, Graumann and Marahiel ruled out divergent evolution because of the different protein structure, different β-strand order in the β-sheet and different protein topologies found in these two structural families [4].

In this paper we present an alternative hypothesis: that these two families of proteins arose from divergent evolution. We show that Graumann's basic argument holds true only if stable, functioning entities, protein domains or modules, simply "appeared" in the primordial soup. We suggest that this is not the case, but rather that proteins evolved from "proto-proteins," small peptides of minimal structure and function. Increasingly stable and functional proteins arose by adding structural elements to produce the small domains or protein modules that we would recognize

today. From this point of view, the surprising similarities between small structural fragments of these proteins, noticed by many others [4, 5] and taken as examples of convergent, function-driven evolution, are interpreted in exactly the opposite way—as traces of common evolutionary origin. We are fully aware that talking about scenarios of early protein evolution is speculation. But such speculations can offer important insights into relationships between proteins and might enable more accurate prediction of functions and structures of newly sequenced proteins. These in turn can be verified experimentally, thus turning "speculations" into verifiable scientific hypotheses.

The paper is built as follows. In the first section of the paper, the family of proteins—cold shock and RBD domain proteins—are introduced and the folds and functions of each class are briefly surveyed. This analysis leads to the identification of the minimally functional unit that can be used as a possible candidate for a functional proto-protein peptide. We then construct an "evolutionary tree" based not on sequence comparison, but on comparison of protein morphology. We argue that, in the specific case of these two nucleic acid binding proteins, existing structural data enable us to construct an ancestral proto-protein, the minimal functional unit, which consists of a nucleic acid-binding β-loop-β structure.

2. Results and Discussion

2.1. The set of nucleic acid-binding proteins and analysis of the binding motif in each family

The set of proteins used in this study is listed in Table 1 and models of representative structures are shown in Figure 1. Structures analyzed below come from organisms from different domains, both procariota and eucariota, suggesting that these two families of nucleic acid binding proteins were already well formed before the split between procaryotes and eucaryotes occurred. This fact was previously pointed out for the ribosomal proteins [6].

As can be seen in Figure 1, the cold shock (OB/CSD) family is composed of all-β proteins. The three stranded sheet is at right angles to a second three stranded β-sheet, resulting in a "two-layer sandwich" structure [7]. The RBD proteins also fall into the category of two-layer sandwich proteins, where the first layer is a four stranded β-sheet, but the second layer is composed of α-helices. Thus, the OB/CSD proteins are all-β proteins, while the RBD proteins are α/β proteins. At the sequence level, both families contain the recognizable RNP sequence signature for binding RNA [8], as illustrated in Figure 2 for CspB and U1A proteins. The RNP sequences are on adjacent β-strands in the 3-D structure, as marked in Figure 1.

Table 1. List of RNA/DNA binding protein structures used in this study.

PDB	Protein	Organism	Reference
Family:	**OB/CSD domain (ssDNA binding domain, β–β–β–β–β)**		
1nmg	Cold shock protein B	*B. subtilis*	9
1mjc	Cold shock protein A	*E. coli*	10
1sro	S1 RNA binding domain (PNPase)	*E. coli*	11
1pfs	ssDNA binding protein	pf3 bacteriophage	12
Family:	**RBD domain (RNA binding or RRM motif; β–α–β–β–α–β)**		
1ris	Ribosomal protein S6	*B. thermophilus*	5
1sxl	sex lethal protein	*D. melanogaster*	13
1urn	U1A complex with RNA	*H.sapiens*	14

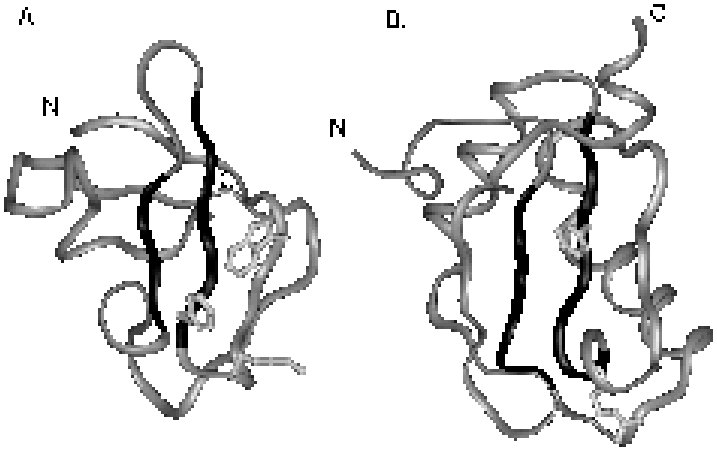

Figure 1. Ribbon models of the protein structure of the topology of cold shock protein B (1nmg [9]) as a representative of the OB/CSD family (A) and U1A (1urn [14]) as a representative of the RBD family (B). The protein structure is shown as a gray ribbon and the location of the RNP1 and RNP2 sequence signatures are in black. Some residues which have been shown by mutagenesis to reduce or eliminate nucleic acid binding are shown as stick figures (Trp8, Lys13 and Phe15 in CspB, Arg52 and Phe56 in U1a). The N- and C- temini of each protein are marked. The cartoon drawings of both topologies are shown in Figure 5.

	RNP1	pos	RNP2	pos
CspB	**KGFGFIEV**	13-20	**VFVH**	26-29
U1A	**RGQAFVIF**	56-63	**IYIN**	12-15

Figure 2. RNP1 and RNP2 RNA binding sequences from CspB and U1A. Note that in CspB the RNP2 motif is located after the RNP1along the sequence, while in˝U1A the situation is reversed

Closer observation of the proteins shows that both families bind nucleic acids in a similar fashion: the nucleic acids bind across the face of a β-sheet that is rich in aromatic amino acids, the side chains of which form stacking interactions with the bases in the nucleic acids. Positively charged residues, which interact with the negatively charged phosphates on the nucleic acid backbone, are found in loops around the β-sheet. The structure of the human U1A protein bound to stem-loop II of the U1 snRNA (small, nuclear RNA) provides a good example (Figure 3, [15]). In this structure, the top of the RNA loop lays across the plane of the β-sheet. Tyr13 and Phe56 on the two central strands, β1 and β3, of the β-sheet form stacking interactions with a cytosine and adenine in the RNA loop. An omega loop or hairpin between the strands 2 and 3 is inserted into the center of the RNA loop and positively charged residues, Arg 47, Lys50, and Arg52, can form electrostatic interactions with the phosphates on the nucleic acid backbone.

A similar motif is repeated in OB/CSD family. In the cold shock protein, CspB, from *B. subtilis*, mutational analysis shows ssDNA binding in vitro is reduced or eliminated when residues Trp8, Phe15, Phe27, Phe30, Lys7, Lys13, or Arg56 are mutated [16]. As in the U1A protein, the aromatic residues are located on the β-sheet and the positively charged residues surround the aromatic residues (Figure 3B). In particular, Lys13 is found in the loop between β1 and β2.

2.2. An evolutionary tree based on protein morphology: divergent evolution of the nucleic acid binding domains

Thus, a very similar motif, RNA binding on the surface of a β-sheet, is found in two families, RBD and CSD, which exhibit no observable sequence similarity, except in the RNP sequences themselves.

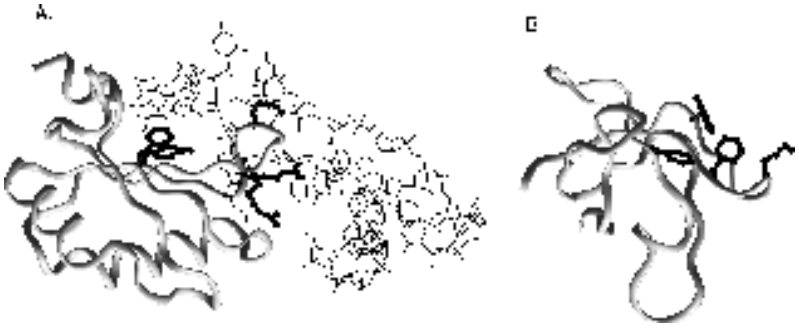

Figure 3. **A**. A close-up view of the loop in U1 RNA interacting with the U1A protein, 1urn [15]. The protein backbone is a ribbon and side chains of Tyr13, Phe56, Arg47, Lys50, and Arg53 are shown as ball and sticks. **B**. A close up view of the proposed ssDNA binding site in cold shock protein CspB from *B. subtilis*, 1nmg [17]. Trp8, Phe15, Phe27, and Lys13 are shown as ball and sticks. The protein backbone is a light gray ribbon.

As described in the introduction, this observation led to the hypothesis of convergent evolution between these two families of RNA binding proteins [4]. We propose that this is not, in fact, a case of convergent evolution, but rather a case of divergent evolution. Our hypothesis is based on several lines of evidence. First, as discussed above, because RNA is proposed to be the original organic polymer upon which life was built [18], the RNA-binding proteins are likely to be the most ancient family of proteins. For instance, ribosomal proteins are among few protein groups which orthologous evolution can be easily followed among organisms from different domains, such as procariota and eucariota [6]. Second, even when sequence similarity is non-existent, we argue that proteins with similar topologies and similar functions could have diverged from a similar ancestor. Time has simply erased any trace of sequence similarity. This already has been suggested by several researchers in relation to some of RNA-binding families [11, 19-21]. Third, and most important, the RNA binding motifs themselves are located on similar super-secondary structures. In particular, one of the structures is a β-strand and the adjacent loop (Figures 3 and 4). We propose that a β-loop-β structure was the original nucleic acid proto-protein from which both of these protein families diverged.

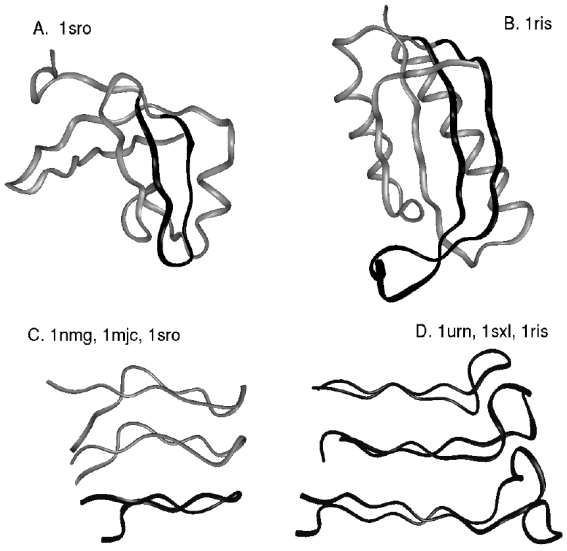

Figure 4. Ribbon structures of 1sro from the OB/CSD family (A) and 1ris from the RBD family (B) with the proposed proto-protein β-loop-β shown in black. Comparison of the β-loop-β structure from other members of both families in shown in parts C and D of the figure. See Table 1 for protein names and references.

Assuming that the β-hairpin was the original proto-protein for the OB/CSD and RBD proteins, can we propose an evolutionary tree for these two families based on their protein morphology? This tree is based on the idea that evolution can occur by addition or deletion of structural elements and strong evidence exists to support this mode of evolution. For instance, in the example of mandelate racemase and enolase, the first two αβ units of the regular (αβ)8 barrel of mandelate racemase are changed to a unusual α–β–β–α structure, thus introducing one anti-parallel strand to the usually all parallel TIM-like β barrel. The family of copper binding proteins can serve as another example. An archetypal protein from this family, plastocyanin, is a Greek-key eight strand β barrel. Other members of this family retain most of the fold, specially around the copper binding site, but introduce additional structural elements, either at one of the termini, such as an α helix at the C-terminus in pseudoazurin (1paz, [22]) or an additional β/α hairpin at the N-terminus in the cucumber basic protein (1cbp, [23]) or in the middle of the sequence (but at the edge of the structure), such as an additional hairpin in azurin (2aza, [24]). These, and other similar examples found in other protein groups, suggest that protein fold is flexible and adding or removing secondary structure elements or even larger domains happened in many protein families even in relatively recent times.

Given this mode of evolution, a proposal for an evolutionary tree for the OB/CSD and RBD proteins built on the basis of protein morphology is presented in Figure 5. The first two levels of the tree are hypothetical. We suggest that proteins with such structures, in fact, do not exist today, but were the original proto-proteins, proteins not as stable and as functionally specific as the structures we currently recognize as proteins. Synthesizing sequences of such short, functional peptides could be an attractive target for experiment. From observation of the protein structures, we can also propose that once an RNA-binding proto-protein developed, the next step was to add another layer of structural elements to stabilize that proto-protein. The stabilizing layer in the OB/CSD family was an additional β-sheet, while the stabilizing layer in the RDB family was a layer of helices. Thus, the two-layer sandwich motif found in both of these families is a result of a divergent, rather than convergent, evolution.

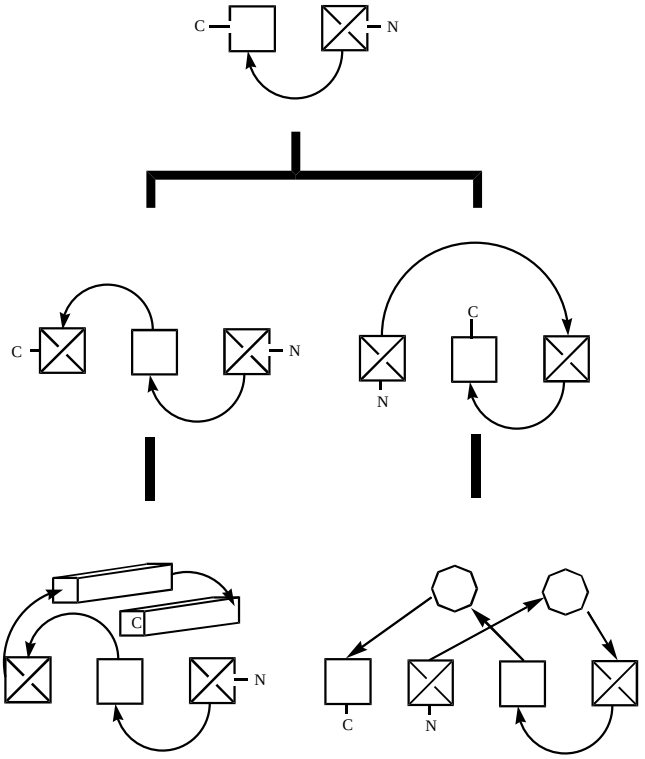

Figure 5. A proposed evolutionary tree constructed using the morphology of the RNA binding proteins listed in Table 1. The tree is based on the hypothesis that the β-hairpin shown on the first level of the tree was the proto-RNA-binding protein. The first two levels of the tree are hypothetical. No complete RNA binding domains of that construction have been found, but constructing such a peptide could be an attractive target for experiment. In the last level of the tree, the cartoon representation of 1urn and 1nmg, shown earlier in Fig. 1 are shown. The β-strands containing RNP1 and RNP2 are highlighted.

2.3. Sequence alignments provide further support of the proposed evolutionary tree

A test of our model would be to show structures or sequences of related proteins that fit into the tree or that fill the obvious “holes” in the tree. For example, our model suggests that in sequence alignments of the proteins in these families, the β-strands involved in RNA binding should be most conserved, while the structural elements involved in the second, stabilizing layer should be much more variable. Analysis of the sequence variability in all protein families supports this hypothesis. As an example, alignments of the cold shock protein 1mjc with two of its homologues, the cold shock protein C from *E.coli* (CSPC_ECOLI) and the glycine

rich protein from *A. thaliana* (JQ1061) are shown in Figure 6. These proteins, with 70% and 43% overall sequence identity to the 1mjc sequence are clearly homologous. But for both, the variability is markedly different between the front and the back β sheet (see Figure 6). For CSPC_ECOLI the two halves of the alignment show 83% and 53% sequence identity, while for JQ1061 the respective conservation is 63% for the front sheet versus 21% for the second layer. All of the proteins homologous to 1mjc, as well as for other proteins studied here, show the similar pattern of mutations. Another way of representing this effect is shown on Figure 7 for the ribosomal protein S6 (PDB code 1ris). The sequences of this protein was used as input for a Blastp search [25] of the OWL data base [26]. Sequences were then aligned using an automated profile editor adapted from [27]. To show the degree of variability at each amino acid position, the variability was calculated as a mean of the mutation probability averaged for all pairs at a given position [28] and plotted versus residue number (see Figure 7). Data, such as shown on Figures 6 and 7, clearly indicate that the sequence variability in these proteins is dramatically larger in the second, stabilizing layer of structural elements, while the most conserved region in all proteins in the end of the β-strand and the beginning of the β-turn (Figure 7). This effect, seen so clearly even for closely homologous proteins, could lead to significant changes in the overall fold over longer periods of time.

```
EEEEEEEEETTTTEEEEEETTS   EEEEGGGB  TTT    TT EEEEEEE  SSS EEEEEEE
A            *******       *****
KMTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAKGPAAGNVTSL
|  | |||||  |||||||| ||||||||||||||  | | | ||| | | |  | ||||| |||
KIKGQVKWFNESKGFGFITPADGSKDVFVHFSAIQGNGFKTLAEGQNVEFEIQDGQKGPAAVNVTAI
B
KMTGIVKWFNADKGFGFITPDDGSKDVFVHFSAIQNDGYKSLDEGQKVSFTIESGAKGPAAGNVTSL
:..| ||||:.:||||||||||||:.|:||| |.|..:|::||..:: |.|.:| :....: :  .|
RRKGSVKWFDTQKGFGFITPDDGGDDLFVHQSSIRSEGFRSLAAEEAVEFEVEIDNNNRPKAIDVSG
```

Figure 6. Sequence alignments for two proteins from the OB/CSD protein family to the 1mjc sequence (shown in bold). **A**. Cold shock protein C from *E.coli* (CSPC_ECOLI), **B**. Glycine rich protein from *A. thaliana* (JQ1061). The secondary structure assignment and positions involved in RNA binding for 1mjc are shown in the top part of the figure.

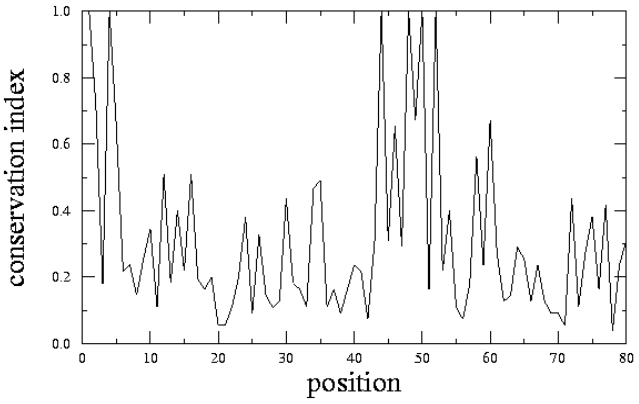

Figure 7. Sequence variability index for ribosomal protein S6 family (the RDB structural family). The index was calculated as a mean of the mutation probability averaged for all pairs at a given position, using a variant of the approach used in [28], and plotted versus residue number. In this protein, the helices forming the second layer (see text) are located approximately between the positions 16-32 and 66-80.

3. Conclusions

We propose that the first proteins were peptides with some minimal, probably a binding and stabilizing, function. These proteins evolved to more stable, more functionally specialized proteins, by adding structural elements at the peptide ends and, later as the proteins grew, by insertions and additions of larger domains. In this paper, we show that that this scenario suggests that the OB/CSD proteins and the RBD proteins arose by divergent evolution, rather than the convergent evolution previously proposed [4]. We show an evolutionary tree based on the protein structural morphology of these two families. Sequence data is consistent with the proposed mechanism, showing a much larger sequence variation in the second structural layer. Further work will involve including more RNA binding protein families in the tree and pursuing further sequence and structural analysis to expand the proposed tree.

The question of the origin of present day proteins, which is addressed here from the viewpoint of function of small structural units has been asked many times in various contexts. According to "the exon theory of genes" [29], present day proteins were formed by reshuffling short DNA fragments. In number of cases it was shown that exons indeed code for compact, semi-autonomous peptide fragments [30]. Unfortunately, there are other cases where it is not true [31]. Small, autonomous folding units were identified in structures of several proteins, and even though their extent only weakly coincides with exon boundaries, it was suggested that they may play a role in the evolution of protein structures [32]. A similar role was proposed for "structural compact modules" [33]. Understanding early protein evolution could help us understand today's protein; however, at present both problems remain far from being solved.

Acknowledgments

This research was supported by NSF grant MCB-9506278. JSF thanks Jeffrey Skolnick for support during her sabbatical.

References

1. Riley, M. and Labedan, B. J. Mol. Biol. **268:**857-868, 1997
2. Branden, C. and Tooze, J., *Introduction to protein structure*. 1991, New York and London: Garland Publishing, Inc. 302.
3. Doolittle, R.F. TIBS **19:**15-18, 1994
4. Graumann, P. and Marahiel, M.A. BioEssays **18:**309-315, 1996
5. Lindahl, M., *et al.* EMBO J. **13:**1249-1254, 1994
6. Wool, I.G., Chan, Y.-L. and Glueck, A. Biochem. Cell Biol. **73:**933-947, 1995
7. SCOP Structural classification of proteins, MRC Cambridge, Oxford, http://www.bio.cam.ac.uk/scop, 1995
8. Bairoch, A. 1994
9. Schnuchel, A., *et al.* Nature **364:**169-171, 1993
10. Schindelin, H., Jiang, W., Inouye, M. and Heinemann, U. Proc. Natl. Acad. Sci. USA **91:**5119-5123, 1994
11. Bycroft, M., Hubbard, T.J., Proctor, M., Freund, S.M. and Murzin, A.G. Cell **88:**235-242, 1997
12. Folmer, R.H.A., Nilges, M., Konings, R.N.H. and Hilbers, C.W. EMBO J. **14:**4132-4142, 1995
13. Lee, A.L., Kanaar, R., Rio, D.C. and Wemmer, D.E. Biochemistry **33:**13775-13786, 1994
14. Oubridge, C., Ito, N., Evans, P.R., Teo, C.-H. and Nagai, K. Nature **372:**432-438, 1994
15. Gubser, C.C. and Varani, G. Biochemistry **35:**2253-2267, 1996
16. Schroder, K., Graumann, P., Schnuchel, A., Holak, T.A. and Marahiel, M.A. Mol. Microbiol. **16:**699-708, 1995
17. Schindelin, H., Marahiel, M.A. and Heinemann, U. Nature **364:**164-168, 1993

18. Joyce, G.F. Nature **338:**217-224, 1989
19. Biamonti, G. and Riva, S. FEBS Lett. **340:**1-8, 1994
20. Golden, B.L., Ramakrishnan, V. and White, S.W. EMBO J. **12:**4901-4908, 1993
21. Hoffman, D.W., *et al.* EMBO J. **13:**205-212, 1994
22. Petratos, K., Dauter, Z. and Wilson, K.S. Acta Crystallogr., B. **44:**628, 1988
23. Guss, J.M., Merritt, E.A., Phizackerley, R.P. and Freeman, H.C. J. Mol. Biol. **262:**686, 1996
24. Baker, E.N. J. Mol. Biol **203:**1071-1095, 1988
25. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. J. Mol. Biol. **215:**403-410, 1990
26. Bleasby, A.J., Akrigg, D. and Attwood, T.K. Nucl. Acid Res. **22:**3574-3577, 1994
27. Rychlewski, L. and Godzik, A. Prot. Eng. in press, 1997
28. Godzik, A. and Sander, C. Prot. Engineer. **2:**589-596, 1989
29. Gilbert, W., de Souza, S. and Long, M. Proc. Natl. Acad. Sci. USA **94:**7698-7703, 1997
30. Gilbert, W. and Glynias, M. Gene **135:**137-144, 1993
31. Dietmaier, W. and Fabry, S. Curr. Gen. **26:**497-505, 1994
32. Panchenko, A.R., Z., L.-S. and Wolynes, P.G. Proc. Natl. Acad. Sci. USA **93:**2008-213, 1996
33. Takahashi, K., OOhashi, M., Noguti, T. and Go, M. FEBS Lett. **405:**47-54, 1997