Biochem. Soc. Symp. **57**, 25–33
Printed in Great Britain

Inverting the Protein-Folding Problem

CHRIS SANDER

E.M.B.L., Postfach 10.2209, Meyerhofstrasse 1, D-6900 Heidelberg, F.R.G.

The Protein-Folding Problem

In the protein-folding problem, one would like to understand how a globular protein reaches its particular three-dimensional shape, starting from the linear sequence of amino acids coded on its gene. More precisely, given a protein sequence, one would like to be able to calculate (predict) the (average) atomic coordinates of the biologically functional protein. Methods from theoretical molecular physics, statistics, information and decision theory, artificial intelligence, etc. have over the last 20 years not solved this central problem of structural biology. Physical simulations at present take too much computer time or fail to correctly simulate the essential interactions. Statistical or empirical rules for structure prediction are generally weak. Three principal difficulties lie in the co-operativity of the folding process, in the delicate energetic balance between enthalpy and entropy and in the fact that protein sequences have evolved to satisfy diverse functional constraints and are not optimized merely to fold up correctly. Today, the protein-folding problem as a structure prediction problem remains fundamentally unsolved [1].

Partial solution to the problem: model building by homology

A partial solution of the problem is based on a remarkable fact of biological evolution: protein molecules retain their basic shape and some aspects of their sequence over millions of years, and across a wide spectrum of species. Whenever we can make the evolutionary connection between a protein of unknown and one of known three-dimensional structure (e.g. by comparing the information in their sequences), we can predict with fair accuracy the unknown protein structure. The art of model building by homology has developed into a powerful tool in the last 2 or 3 years and is the basis of most predictions of protein tertiary structure today (e.g. [2]). Exercise of the art requires skills in the alignment of multiple sequences and of three-dimensional structures, in computer graphics, in intelligent database access and in conformational molecular dynamics. In this context, the prediction of the effects of point mutations on protein structure and function is an area of intense current activity. However, the evolutionary approach to protein folding neither enables us to predict the structure of a new class of proteins nor to construct novel proteins which are unlike those found in nature.

Inverting the problem: protein design

For a more comprehensive solution of the problem, new approaches are needed. Several years ago, in an E.M.B.O. course in September 1986 [3], we attempted to break new ground by turning the question upside down 'Ask not to calculate the structure from the sequence, but ask instead to calculate a sequence from the structure'. In protein design, the problem to be solved is not structure prediction, but sequence prediction.

Inverting the problem leads to new ways of thinking about protein folding. Consider these new questions: 'Enumerate all protein sequences compatible with a given fold!'; or 'Design a sequence that will fold up like lysozyme!'; or 'How much sequence variation at the surface of a globular protein is compatible with maintaining the chain fold?'. The inverted problem may actually be easier to solve than the forward problem, for several reasons. First, the forward problem has only one essentially correct solution, namely the native structure. The inverse problem, if one is satisfied with one sequence that folds up as planned, has many solutions: we know that in nature a protein family typically has many different sequences of essentially identical three-dimensional structure. Secondly, sequence prediction needs pay no heed to functional requirements; the sequence can be optimized for the sole purpose of yielding the correct structure *in vitro.*

To illustrate the possible simplicity of a designed sequence compared with a natural sequence, consider the following example. Suppose you want to construct an all-α-helical protein with a certain given topology. You have at your disposal tables of single residue preferences for certain positions in secondary structures, in loops, in interfaces between secondary structures, and on the surface; also, you have tools to optimize packing by removing steric clashes. Suppose you then simple-mindedly choose for each position residues optimal merely from the point of view of single residue preferences, naively ignoring one of the principal rules of protein folding 'local sequence information is insufficient to fully determine the native fold' (e.g. the structure of oligopeptides as long as five or six residues can be dominated by the surrounding tertiary structure interactions [4]). Is it not conceivable that such a protein sequence, although clearly less complicated than natural sequences, may well do the job? Is it not conceivable that natural sequences are more complicated because they have been subjected to additional functional selective pressure and randomizing mutational events? We shall see in due course.

Protein design as sequence prediction is a new approach to the old problem of protein folding. To the extent that designed proteins are not subject to the same evolutionary constraints as are natural proteins, the protein-design problem transcends the classical protein-folding problem.

Designing Proteins *De Novo*

The number of proteins designed *de novo* is gradually increasing. An excellent recent review [5] presents important examples. Here, I merely mention proteins designed in the 1986 E.M.B.O. course [3]: two triose phosphate isomerase (TIM) barrels (4-fold symmetric $\alpha/\beta/\alpha/\beta$ barrels), two alternating α/β proteins (flavodoxin topology) and two α-helical bundles. The approach was to start from an idealized form of a well-known protein fold and to design one sequence that would produce that fold. This was done in two steps. (i) Choose a typical protein fold (say, a $\beta\alpha\beta$ nucleotide-binding fold) and idealize (simplify) its architecture. Incorporate, if possible, a binding site for a ligand the binding of which can be detected experimentally. The result is a backbone coordinate set that clearly belongs to a well-known structural class, yet is unlike any of the natural members of the class in detail. (ii) Invent, calculate and predict an appropriate amino acid sequence. The result is a full coordinate set (all atoms) of the designed protein and a list of its amino acid sequence. An attempt at experimental verification of one of these designs, the TIM barrel babarellin, is in progress, by expression of a synthetic gene (H. J. Fritz, personal communication).

Redesigning Protein Topology

A hybrid approach to the protein-design problem involves the redesign of a natural protein, keeping part of its original amino acid sequence and changing other parts, in the pursuit of particular design goals. In collaboration with S. Emery, W. Klaus & H. Blöcker at G.B.F. in Braunschweig, as well as M. Sagermann and D. Tsernoglou at E.M.B.L. in Heidelberg, and W. Eberle and P. Roesch at M.P.I.M.F. in Heidelberg, we have re-engineered the topology of loop connections in a bundle of four α-helices, the rop (repressor of primer) protein, and assayed the re-engineered structure by n.m.r. spectroscopy.

The particular hypothesis tested by the rop redesign was motivated by a number of fundamental observations about the changes in protein structures and sequences in natural evolution (e.g. [6]). The most strongly conserved aspects of the tertiary structure of globular proteins are a structural core and the topology of loop connections between elements of secondary structure in that core. In contrast, the detailed structure of loops undergoes large variations. The evidence comes from superposition of related natural protein structures. Corresponding multiple alignments of amino acid sequences show that evolutionary constraints on protein sequences are strongest in the interior of the conserved structural core, where residues make numerous intraprotein contacts, and in functionally important sequence positions. Sequence variation is much greater where residues make few contacts with other residues, e.g. in solvent-exposed surface or loop regions. The successful transfer, by cut-and-paste engineering, of a β-turn structure between two proteins, preserving the original loop structure, directly illustrates the possible variability and independence of loop conformation [7].

The fact that a structurally conserved core can be delineated in space and that there are stronger variations outside of that core lead to the 'core hypothesis': *the dominant factor determining the native protein fold is the interaction of residues in the structural core*; and the corollaries: (i) to first approximation, the effect on protein structure of sequence changes in the core and outside of the core, especially loops, can be separated, i.e. treated independently in prediction or construction of protein structure; (ii) radical changes in the topology of loop connections, within certain weak constraints, are consistent with the preservation of the structural core, although such changes are (almost) never observed in natural protein families. Proof or disproof of these statements about protein folding is very difficult to achieve in a single set of experiments. The rop redesign was intended as a first step in the verification (or eventual modification) of the 'core hypothesis'.

The natural protein on which the redesign was based is a small bacterial protein involved in regulating DNA replication via interacting with an RNA primer [8], called rop. This protein of 63 amino acids forms a homodimer, as shown by crystallography [9] and two-dimensional n.m.r [10]. Each subunit has an antiparallel pair of two helices and a flexible *C*-terminal tail of seven residues (57-GDDGENL-63). The two pairs of helices associate in antiparallel orientation to form a four-helix bundle (Fig. 1). Each of the four helices is oriented antiparallel to its two nearest neighbours and parallel to its diagonal, more distant, neighbour. Residues on the interior helix faces form a hydrophobic core. The loop forming the chain reversal of the helical hairpin involves only three residues on each subunit, L29, D30 and A31. At each end of the bundle there are four helix termini, two connected by a short loop, and two as polypeptide *N*- and *C*-termini.

The dimeric rop bundle has a definite handedness (left handed in our convention as well as in that of Presnell & Cohen [11]). To see the non-equivalence of the two chiral alternatives, follow the U shape of one subunit (helix–loop–helix) from *N*- to *C*-terminus; as you arrive at the *C*-terminus, the hydrophobic dimer interface (and hence the other subunit) lies either on one side or the other of the plane formed by the two helices (Fig. 1). To remember the left–right convention (valid for both monomeric and dimeric bundles), place the first two helices (counting from the *N*-terminus) in the plane of the paper, the other two behind that plane. The chain trace of the first two helices, from *N*- to *C*-terminus, runs clockwise from right-handed bundles and anti-clockwise for left-handed bundles. The existence of both right-handed and left-handed helix bundles in nature [11], suggests the possibility that the topology of loop connections can be arranged independently of helix packing.

Genetic alteration [12] has shown that the wild-type rop (Wtrop) structure is very resistant to single non-conservative point mutations on the helical surfaces and near the loop region. Evidence comes from functional integrity (interaction with RNA primer) or, in cases of functional change, successful overproduction and/or covalent modification of buried Cys residues. Deletion of five and insertion of two amino acids in the loop region also leaves the structure intact; conservative point mutations in the interior are tolerated, e.g. L48 to V and L41 to V, but at a loss of thermodynamic stability ([12]; G. Cesareni and M. Kokkinidis, personal communications). Wtrop is thus an excellent candidate for the type of re-engineering performed here. The analysis of known helix bundles and mutation experiments with rop suggested that the core hypothesis is valid for this class of proteins. The design of the left-handed monomer of rop (Lmrop) represents one extreme case consistent with this hypothesis: complete conservation of the protein sequence in the helical

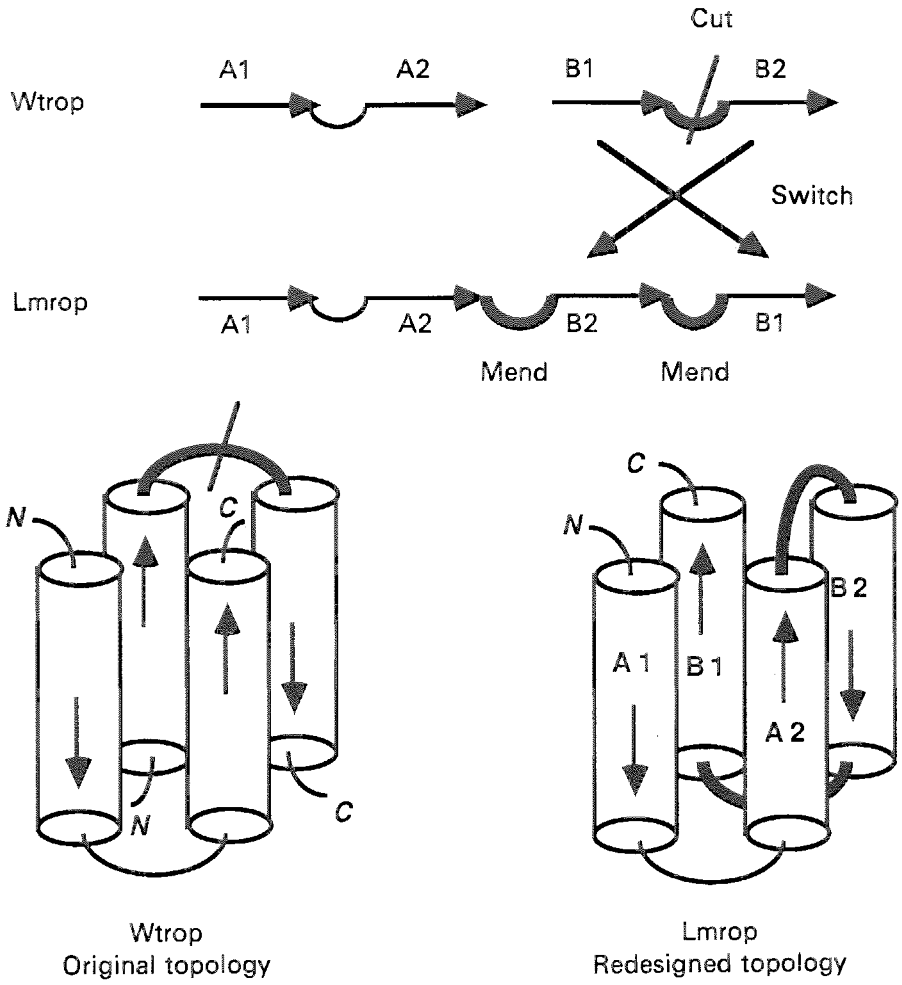

Fig. 1. *Cut-and-mend protein engineering: design steps from Wtrop dimer to Lmrop monomer involve deletion of one loop, addition of two new loops and a C-terminal extension*

Top: the orientation in space and packing of the four helices remains intact. Bottom: two helical segments are switched at the gene level. In detail: Wtrop has two identical subunits, say A and B, and a total of four helices, say A1, A2 and B1, B2 (cylinders/arrows). To achieve the desired topology of a left-handed four-helix bundle which starts with helix A1, delete the loop between helices B1 and B2; switch the order (in the chain sense) of these helices from B1, B2 to B2, B1); connect by a new loop helices A2, B2 and by another loop helices B2, B1; complete helix B1 by addition of a few residues. Loops involved in these changes are highlighted (thick lines).

regions, both in the interior (deemed essential) and on the helical surface (kept for convenience), and complete change in topology and amino acid sequence of loop regions (except on loop).

The redesign steps in going from the dimer Wtrop to the monomer Lmrop are these (Fig. 1): delete the loop on one two-helix subunit, say the second one; switch the order of the two helices in that subunit; and introduce two new loops, one between the subunits and one between the two helices of the second subunit, at the helix end opposite to where the original loop was deleted. The wild-type topology, A1–loop–A2–break–B1–loop–B2, becomes the Lmrop topology, A1–loop–A2–loop–B2–loop–B1–tail, where A1, A2, B1, B2 are α-helices.

Sequence design (Figs. 2 and 3), production of Lmrop proteins and determination of its three-dimensional structure were undertaken to verify the design and to provide a well-characterized vehicle for future protein folding studies [13]. Milligram amounts of pure protein allowed analysis by two-dimensional n.m.r. spectroscopy. Although we have not yet solved the structure of Lmrop in full atomic detail, the analysis proves that Lmrop does have four α-helical segments and that these helices come together in the three-dimensional structure in the same relative orientation, with the same interhelix contacts and the same bundle handedness at Wtrop.

The successful redesign of Lmrop is consistent with the core hypothesis, i.e. that the native fold is determined by interaction between the helices and that loops merely provide weak constraints which rule out certain arrangements. It remains an open question as to what precisely the constraints on loop connections are. We think that the principal restrictions are minimum loop length consistent with the arrangement of secondary structure elements in space and

helices A1+A2		loop L						helix B2	loop M			helix B1	tail N			
A1....A55	A56	L1	L2	L3	L4	L5	L6	B32...B56	M1	M2	M3	B3...B29	N1	N2	N3	name
S		K	K	P	G	Q	I		G	G	S		A	K	G	res

num	name	res	
	B1	M	deleted
	B2	T	deleted
	B30	D	deleted
	B31	A	deleted
55	A55	S	to avoid clash, replaces R A55
56	A56	F	kept
57	L1	K	continue helix after F A56
58	L2	K	continue helix
59	L3	P	turn PG as in loop PGQI of trypsin, contact with F A56
60	L4	G	as in trypsin res 174
61	L5	Q	as in trypsin res 175
62	L6	I	as in trypsin res 176
88	M1	G	flexible
89	M2	G	flexible
90	M3	S	possible H-bond to E B5
118	N1	A	for helix-helix packing - nothing bigger here !
119	N2	K	for charge interaction with COO of E's and/or C-term
120	N3	G	allows flexible adjustment of COO location

Fig. 2. *Protein sequence design of Lmrop and details of sequence changes relative to Wtrop [13]*

The amino acid sequence of Lmrop is derived from that of Wtrop by gene duplication (residue with names B1–B56 are a duplicate of A1–A56), deletion of four residues (B1, B2, B30, B31), mutation of one residue (A55), addition of nine residues in new loops (L1–L6 and M1–M3) and addition of a three-residue tail (N1–N3) for helix extension. 'num' is the sequential residue number in Lmrop (1–120); 'name' is the invariant residue name (chain and number, same in Wtrop and Lmrop when possible); 'res' is the amino acid type in one-letter code (A, Ala; G, Gly; etc.). The loop in trypsin was identified in Protein Data Bank data set 2PTN.

```
    M  T  K  Q  E  K  T  A  L  N  M  A  R  F  I  R  S  Q  T  L    A1
5' ATGACCAAACAGGAAAAAACCGCCCTTAACATGGCCCGCTTTATCAGATCTCAGACATTA

    T  L  L  E  K  L  N  E  L  D  A  D  E  Q  A  D  I  C  E  S    A1-loop-A2
   ACGCTTCTGGAGAAACTCAACGAGCTGGACGCGGATGAACAGGCAGACATATGTGAATCG

    L  H  D  H  A  D  E  L  Y  R  S  C  L  A  S* F  K* K* P* G*   A2-loop
   CTTCACGACCACGCTGATGAGCTTTACCGCAGCTGCCTGGCTAGCTTCAAAAAGCCGGGT

    Q* I* D  E  Q  A  D  I  C  E  S  L  H  D  H  A  D  E  L  Y    loop-B2
   CAGATCGATGAACAGGCAGACATCTGTGAATCGCTTCATGATCACGCTGATGAGCTTTAC

    R  S  C  L  A  R  F  G* G* S* K  Q  E  K  T  A  L  N  M  A    B2-loop-B1
   CGCAGCTGCCTGGCCAGGTTCGGTGGATCCAAACAGGAAAAAACCGCCCTTAACATGGCC

    R  F  I  R  S  Q  T  L  T  L  L  E  K  L  N  E  L  A* K* G*   B1-end
   CGCTTTATCAGAAGCCAGACATTAACGCTTCTAGAGAAACTCAACGAGCTCGCGAAAGGT 3'
```

Fig. 3. *DNA sequence design of the coding strand of the synthetic gene encoding Lmrop protein* [*13*] Residues in helical segments kept from Wtrop are underlined. Residues mutated or inserted (13 out of 120) are denoted by a star (*)

sufficient polar groups in side-chains to assure solvation. The stronger the loop constraints are, the easier the combinatorial problem in structure prediction; the weaker they are, the wider the options in protein engineering.

Support for the core hypothesis can also be derived from known cases of cyclically permuted protein sequences with preserved native structure. This simple form of topological rearrangement was observed *in vivo* comparing the structures of two leguminous lectins, favin and concanavalin A (Con A) [14,15]. Favin protein is similar to Con A in sequence (38% identical residues) and homologous in structure (both are the same type of antiparallel β-sheet sandwich), except for a cyclic permutation of the protein chain. The permutation is thought to occur not at the DNA level but after production of a precursor protein chain and without significant refolding of the initial translation product [16,17].

Pioneering experiments by Goldenberg & Creighton [18] and Kirschner and co-workers [19,20] produced two examples of proteins permuted cyclically by artificial means. The former made both a cyclic [21,22] and a cyclically permuted [18] form of pancreatic trypsin inhibitor by chemical modification and proved that the rearranged forms maintain the native fold. Luger *et al.* [19] achieved cyclic permutations of the yeast enzyme phosphoribosyl anthranilate isomerase (4-fold $\beta\alpha\beta\alpha$ barrel structure as deduced by sequence homology), by a gene construction permuting the order of the four $\beta\alpha\beta\alpha$ units from ABCD to DABC (plus shift by an extra helix in a second construct). Enzymic activity of the permuted forms as well as spectroscopic results (e.g. circular dichroism) indicated an essentially unperturbed structure compared with the wild type.

Preservation of globular structure in the face of a cut-and-mend operation appears consistent with the exon shuffling hypothesis [23–25]. However, we have no evidence that the individul helices of the bundle are independent folding units transferable to another three-dimensional structural context.

Rather, it seems likely that precise complementary packing of all four is required for proper folding. So, while sequential segment swap within one protein was shown here to be feasible, the transfer of folding units between proteins would be much more difficult, both for the protein engineer and for nature. Whether such transfer events are frequent in the natural evolution of globular proteins depends on the likelihood of mutations in the interface (in three-dimensions) between exchanged units, i.e. whether good core packing is achieved before the rearranged gene is eliminated.

Natural evolution of protein structures appears to have occurred along continuous pathways of point mutations and recombination of domains. A natural pathway from dimeric Wtrop to monomeric Lmrop would have to involve gene duplication, elimination of a stop codon, segment swap (invert the order of two helices) and mutations in loop regions, and all this without losing continuity of reading frame or viability of the organism. While there appear to be in nature examples of gene duplication resulting in a single protein of double size, it appears unlikely that natural evolution would be capable of evolving a molecule like Lmrop from a protein like Wtrop — the pathway from one to the other is very discontinuous. In this sense, protein engineering is capable of evolutionary jumps that transcend natural evolution.

The notion that solvent-exposed loop regions are the most strongly varying parts of protein structures in natural evolution is well established by the observation of numerous mutations in natural protein families. Our own and related [19, 20] results suggest that not only point mutations, and insertions or deletions, but also topological reconnections can be tolerated without affecting the core structure of a globular protein. The recognition that topological restrictions are much weaker than anticipated from the observation of naturally evolved protein structures may open up a new degree of freedom in the design of useful proteins.

It will be interesting to see if future repeat cycles of protein redesign, production, structure determination, analysis and redesign as performed once for Lmrop, will lend further support to the 'core hypothesis'. Meanwhile, it is already clear that the protein engineer has considerable latitude in the design of new structures.

References

1. Blundell, T. L., Sibanda, B. L., Sternberg, M. J. & Thornton, J. M. (1987) *Nature (London)* **326**, 347–352
2. Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., Colman, P. M., Spinelli, S., Alzari, P. M. & Poljak, R. J. (1989) *Nature (London)* **342**, 877–883
3. Sander, C. (ed.) (1987) *Protein Design Exercises*, EMBL BIOcomputing Technical Document 1, available from EMBL, D-6900, Heidelberg, RAULFS@EMBL.bitnet
4. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. U.S.A.* **81**, 1075–1078
5. Richardson, J. S. & Richardson, D. C. (1989) *Trends Biochem. Sci.* **14**, 304–309
6. Chothia, C. & Lesk, A. M. (1986) *EMBO J.* **5**, 823–826
7. Hynes, T. R., Kautz, R. A., Goodman, M. A., Gill, J. F. & Fox, R. O. (1989) *Nature (London)* **339**, 73–76
8. Cesareni, G. & Banner, D. W. (1985) *Trends Biochem. Sci.* **10**, 303–306
9. Banner, D. W., Kokkinidis, M. & Tsernoglou, D. (1987) *J. Mol. Biol.* **196**, 657–675

10. Eberle, W., Klaus, W., Cesareni, G., Sander, C. & Rösch, P. (1990) *Biochemistry* **29**, 7402–7407
11. Presnell, S. R. & Cohen, F. E. (1989) *Proc. Natl. Acad. Sci. U.S.A.* **86**, 6592–6596
12. Castagnoli, L., Scarpa, M., Kokkinidis, M., Banner, D. W., Tsernoglou, D. & Cesareni, G. (1989) *EMBO J.* **8**, 621–629
13. Emergy, S. C., Sagermann, M., Eberle, W., Klaus, W., Blöcker, H., Roesch, P., Tsernoglou, D. & Sander, C. (1990) submitted
14. Reeke, G. N. Jr, Becker, J. W. & Edelman, G. M. (1975) *J. Biol. Chem.* **250**, 1525–1547
15. Reeke, G. N. Jr & Becker, J. W. (1986) *Science* **234**, 1108–1111
16. Hemperly, J. J., Mostov, K. E. & Cunningham, B. A. (1982) *J. Biol. Chem.* **257**, 7903–7909
17. Bowles, D. J. *et al.* (1986) *J. Cell Biol.* **102**, 1284–1297
18. Goldenberg, D. P. & Creighton, T. E. (1983) *J. Mol. Biol.* **165**, 407–413
19. Luger, K., Hommel, U., Herold, M., Hofsteenge, J. & Kirschner, K. (1989) *Science* **243**, 206–210
20. Goldenberg, D. P. (1989) *Protein Eng.* **2**, 493–495
21. Chazin, W. J., Goldenberg, D. P., Creighton, T. E. & Wuethrich, K. (1985) *Eur. J. Biochem.* **152**, 429–437
22. Goldenberg, D. P. & Creighton, T. E. (1984) *J. Mol. Biol.* **179**, 527–545
23. Gilbert, W. (1978) *Nature (London)* **271**, 501
24. Doolittle, W. F. (1978) *Nature (London)* **272**, 581–582
25. Blake, C. C. F. (1979) *Nature (London)* **277**, 598