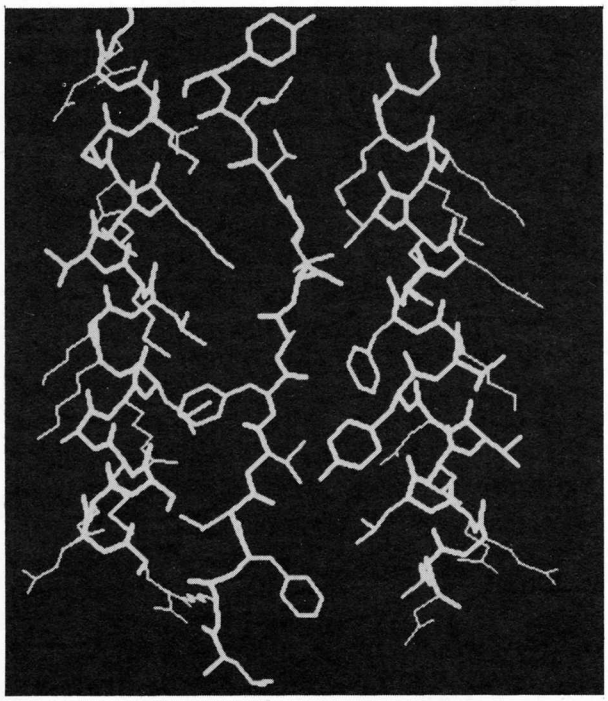
Figure 4

Molecular model of the predicted signal-binding site

Models were constructed on New England Biographics Promodeler I software, by residue replacement of the SRP sequences of the h2 and h3 helices on a classical α-helix, and by residue replacement of a β-strand found in the crystal structure of β-lactoglobulin (residues 47–58) with residues 1–12 of the rat pre-elastase signal sequence.



across the bilayer could then take place, possibly through a channel comprising, at least in part, the sec61 protein [12] and/or a 43 kDa integral membrane protein previously shown as a component of rough microsomes that cross-links to a photoreactive signal peptide [13]. The signal sequence would finally be proteolytically removed by signal peptidase in the lumen of the ER.

1 Austen, B. M. and Westwood, O. M. R. (1991) Protein Targeting and Secretion. 'In Focus' series (Rickwood, D., ed.), IRL Press, Oxford, New York, Tokyo
2 Walter, P. and Blobel, G. (1980) Proc. Natl. Acad. Sci. U.S.A. **77**, 7112–7116
3 High, S. and Dobberstein, B. (1991) J. Cell Biol. **113**, 229–233
4 Bernstein, H. D., Poritz, M. A., Strub, K., Hoben, P. J., Brenner, S. and Walter, P. (1989) Nature (London) **340**, 482–487
5 Robinson, A., Westwood, O. M. R. and Austen, B. M. (1990)
6 Zopf, D., Bernstein, H. D., Johnson, A. E. and Walter, P. (1990) EMBO J. **9**, 4511–4517
7 Kaderbhai, M. A. and Austen, B. M. (1985) Eur. J. Biochem. **153**, 167–178
8 Robinson, A., Meredith, C. and Austen, B. M. (1986) FEBS Lett. **203**, 243–246
9 Oliver, D. (1985) Annu. Rev. Microbiol. **39**, 615–648
10 Reddy, G. L. and Nagaraj, D. (1989) J. Biol. Chem. **264**, 16591–16597
11 Segrest, J. P., Jackson, R. L., Morrisett, J. D. and Gotto, A. M. (1974) FEBS Lett. **38**, 247–253
12 Gorlich, D. and Rapoport, T. A. (1993) Cell **75**, 615–630
13 Robinson, A., Kaderbhai, M. A. and Austen, B. M. (1987) Biochem. J. **242**, 767–777
14 Provencher, S. W. and Glockner, J. (1981) Biochemistry **20**, 33–37

Received 7 July 1994

A prototype computer system for *de novo* protein design

A. Berry*‡ and S. E. Brenner*†

*Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, U.K. and †M.R.C. Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.

The *de novo* design of proteins, i.e. the selection of amino acid sequences that fold into a desired predetermined structure, is one of the outstanding challenges in modern biochemistry. There is little wonder, therefore, that various attempts have already been made to design novel amino acid sequences [1] for naturally occurring protein folds including four-helix bundles [2–4], β-bell structures [5] and α/β barrels [6,7]. Similarly, folds not found naturally, such as an 'open sandwich', have also been designed [8]. In all these cases, the approach has been essentially the same, to manually choose appropriate sequences of amino acids using "physi-

Abbreviation used: PDB, Brookhaven Protein Data Bank.

‡To whom correspondence should be addressed. Present address: Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, U.K.

cal, statistical and intuitive criteria" [9]. Although most of the designed proteins have not been subjected to a full structural characterization, many seem to have elements of the correct secondary structure, and some recent reports indicate that new sequences actually adopt the desired structure [10,11]. However, in all these experiments, an intimate knowledge of the particular protein fold was an essential prerequisite, and it is not at all clear how this approach might be extended to have significant general success.

A widely applicable approach to the design problem is to use a computer, initially to derive an approximation to the 'rules' of structure in the form of statistical parameters, and then to find a sequence which optimizes these rules for a particular target fold. This is the approach which we have taken (S. E. Brenner and A. Berry, unpublished work) [12]. The method operates by evaluating a number of criteria and then computing a weighted sum which is designated the sequence's 'quality' for the desired structure. These criteria can be divided into four broad categories: position, neighbour, uniqueness and hints.

The heart of the design system lies in the first two categories, *position* and *neighbour. Position* preferences are statistical measures such as secondary-structure preference or solvent accessibility, reflecting the preference of a particular type of residue to be found within a given secondary structure, or to be found buried within proteins or exposed on the surface. *Neighbour* preferences are measures of the likelihood of finding one residue type near another, either in the primary sequence or in space [5,13]. The statistical information which is used to design optimized sequences is derived from the native structures of known proteins deposited in the Brookhaven Protein Data Bank (PDB) [14,15]. The analysis of the database is particularly problematic because there is severe bias in the data set resulting from large numbers of structures of the same or very closely related proteins. Furthermore, the quality of individual PDB co-ordinate files varies widely and it is undesirable to give considerable weight to data which is inaccurate. A weighting algorithm has therefore been created which takes account of these factors without needlessly throwing away information (S. E. Brenner and A. Berry, unpublished work).

The third type of parameter is *uniqueness* and is incorporated to reflect the need for 'negative design', i.e. the need to ensure that a designed sequence not only fits well in the desired structure but does not fit even better in another structure [16]. Finally, *hints* is a broad category which contains data about the desired protein which is needed by the design system but which is not included in the previous parameters. In addition, this section could have any information about the functionality of the protein, details that will make it easier to synthesize, and patterns that will facilitate structural studies on the protein.

In order to assess the feasibility of such a computerized design method, we have created a simplified system using a subset of the parameters outlined above. We have used this prototype to design various protein structures and have subjected the resulting amino acid sequences to a range of analyses. The prototype system requires the secondary structure and solvent accessibility of each amino acid in the protein structure we are trying to build. The rules which the system then uses to determine the novel sequence are based only on secondary structure, solvent accessibility, primary-sequence neighbour and diversity (a parameter incorporated to ensure that the residue composition of designed sequences mirrors that of natural sequences). A weighted sum of these parameters is optimized using a simulated annealing protocol to find the best sequences for the structure.

As an example of the results generated by the design system, we describe here some sequences designed to fold into a four α-helix bundle based on the structure of myohemerythrin (Figure 1), PDB

Figure 1

Molecular graphics representation of the structure of myohemerythrin (PDB structure 2mhr [17])

The figure was produced using the program MOLSCRIPT [20]

Figure 2

Amino acid sequences (*a*) and secondary structures (*b*) of the natural myohemerythrin [17] and four examples of designed sequences based on this structure

Figure (*b*) shows representations of, from top to bottom, the structure derived from the crystallography study [17], the secondary-structure prediction for the natural sequence [18,19], and secondary-structure predictions for the designed sequences.

(*a*)

Natural myohemerythrin
GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFDCYSIRDNSAPNLATLVKVTTNHFTHEEAMMDAAKYSEVVPHKKMHKDFLEKIGGLSAPVDAKNVDYCYSKEWLVNHIKGTDFKYK

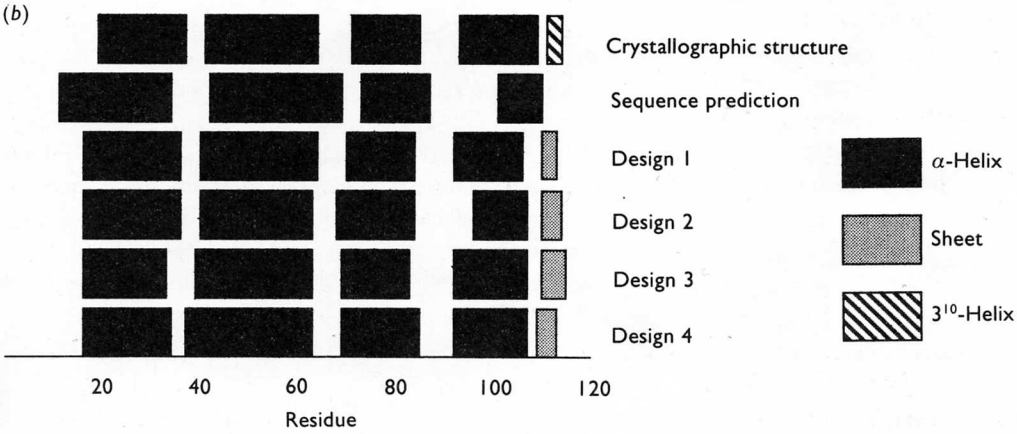
Design 1
MKRYGKRPQGDKDFSTKTKEALDIAEKALKALNILQKNTVEKAVDNFYKGAHRALKNIFNWLHRGKGRPVKEIDKQLKDFSKEIYPGKGKPEKEIGRQLWQMALSVAMGVTYKCHPKTP

Design 2
MKRDGKRPQGNKDFSPKTKRALDILEKALKALNIAQKNTVEKAGQHFYKGINEALKHLFNWLHRGKGRSIKEAQKDIKDLSKEIYPGKGKPEKRVGEDFWQMALSVAMGVTYKCYPKTP

Design 3
MKRNGKRPQGYKDVSPKTKEALDILEKILKALHGFDKNTVEKAVDNIYKIAHEALKNIFHWMNRGKGRPVKEAQRQLKDLSKDFYPGKGKPEKEIGRQAWQMALSVAFGVCTKCYSKTP

Design 4
MKRYGKRPQGHKDVSPKTKEALDILEKAMKALNIFDKETVEKAVDNIYKGINEALKHLFNWINRGKGRPVKEAQKDLKQLSKDFYSGKGRPEKEIGRQAWQMALSVAFGVTYKCHPKTP

(*b*)



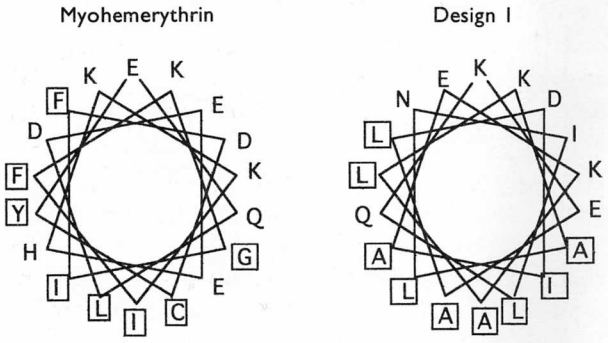
structure 2mhr [17]. Figure 2 shows the sequences of the natural myohemerythrin together with the sequences of four proteins designed with the default weights for the various parameters (S. E. Brenner and A. Berry, unpublished work). None of the designed sequences shows any significant identity to the natural sequence; however, the designed sequences are nearly identical. Indeed, a search of the Swissprot protein database found that none of the designed sequences shared any significant identity with any natural protein.

Secondary-structure prediction showed that four α-helices were likely to form in the designed sequences (Figure 2), although the boundaries were often not positioned exactly as desired [18,19]. It is interesting to note that around residues 110–113, where a 3_{10}-helix is found in the natural protein, the designed sequences were always predicted to have an extended conformation.

One of the problems frequently encountered by protein designers has been unsatisfactory packing of residues within the protein. This problem is only tangentially addressed in our methodology by

Figure 3

Helical-wheel diagrams for helix 1 from the natural sequence and designed sequence 1, showing the hydrophobic faces of both helices



spatial neighbour preferences; nevertheless, it was interesting to look at the nature of the faces of the designed helices as an indication of how they might interact. An example of this analysis is shown in Figure 3, where the residues of helix 1 in both the natural and designed sequences are displayed on a

helical-wheel diagram. In both cases, the helices show clearly defined hydrophobic and hydrophilic faces. This result is also repeated for helices 2, 3 and 4 (results not shown). It appears, therefore, that not only is the designed sequence predicted to form four helices but that they come together in some way to bury the hydrophobic surfaces of the helices. It remains to be determined experimentally whether or not this is the case.

In conclusion, we have developed a general quantitative methodology for designing proteins that relies upon statistical information derived from the structures of all known proteins, a theoretical model of protein structure, and motifs described in the literature. A prototype version of the design system has been used to produce new sequences for a wide variety of protein folds, and these sequences appear generally plausible for the desired structure. In addition to designing proteins with potential medicinal and biotechnological uses, the quantitative nature of the design system will provide us with a fuller understanding of protein structure.

The authors thank the Science and Engineering Research Council, the Royal Society and the National Science Foundation for financial support. A.B. was a Royal Society 1983 University Research Fellow and S.E.B. is a Herchel Smith Harvard Scholar.

1 DeGrado, W. F. and Matthews, B. W. (1993) Curr. Opin. Struct. Biol. **3**, 547–548
2 Hecht, M. H., Richardson, J. S., Richardson, D. C. and Ogden, R. C. (1990) Science **249**, 884–891
3 Hill, C. P., Anderson, D. H., Wesson, L., DeGrado, W. F. and Eisenberg, D. (1990) Science **249**, 543–546
4 Regan, L. and DeGrado, W. F. (1988) Science **241**, 976–978
5 Richardson, J. S. and Richardson, D. C. (1987) in Protein Engineering (Oxender, D. L. and Fox, C. F., eds.), pp. 149–163, Alan R. Liss Inc., New York
6 Goraj, K., Renard, A. and Martial, J. A. (1990) Protein Eng. **3**, 259–266
7 Hubbard, T. J. and Blundell, T. L. (1989) in Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications (van Gunsteren, W. F. and Weiner, P. K., eds.), pp. 168–182, ESCOM, Leiden
8 Federov, A. N., Dolgikh, D. A., Chemeris, V. V., Chernov, B. K., Finkelstein, A. V., Schulga, A. A., Alakhov, Y. B., Kirpichnikov, M. P. and Ptitsyn, O. B. (1992) J. Mol. Biol. **225**, 927–931
9 Sander, C. (1991) Curr. Opin. Struct. Biol. **1**, 630–637
10 Kuroda, Y., Nakai, T. and Ohkubo, T. (1994) J. Mol. Biol. **236**, 862–868
11 Tanaka, T., Kimura, H., Hayashi, M., Fujiyoshi, Y., Fukuhara, K. and Nakamura, H. (1994) Protein Sci. **3**, 419–427
12 Brenner, S. E. (1993) M. Phil. thesis, University of Cambridge, Cambridge
13 Sander, C., Scharf, M. and Schneider, R. (1992) in Protein Engineering: A Practical Approach (Rees, A. R., Sternberg, M. J. E. and Wetzel,. R., eds.), pp. 88–115, Oxford University Press, Oxford
14 Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. and Weng, J. (1987) in Crystallographic Databases: Information Content, Software Systems, Scientific Applications (Allen, F. H., Bergerhoff, G. and Sievers, R., eds.), pp. 107–132, Data Commission of the International Union of Crystallography, Cambridge
15 Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. **112**, 535–542
16 Richardson, J. S., Richardson, D. C., Tweedy, N. B., Gernert, K. M., Quinn, T. P., Hecht, M. H., Erickson, B. W., Yan, Y. B., McClain, R. D., Donlan, M. E. and Surles, M. C. (1992) Biophys. J. **63**, 1186–1209
17 Sheriff, S., Hendrickson, W. A. and Smith, J. L. (1987) J. Mol. Biol. **197**, 273
18 Rost, B. and Sander, C. (1992) Nature (London) **360**, 540
19 Rost, B. and Sander, C. (1993) J. Mol. Biol. **232**, 584–599
20 Kraulis, P. J. (1991) J. Appl. Crystallogr. **24**, 946–950

Received 7 July 1994