Protein Science (1994), *3*:1871–1882. Cambridge University Press. Printed in the USA.

A quantitative methodology for the de novo design of proteins

STEVEN E. BRENNER[1,2] AND ALAN BERRY[1,3]

[1] Cambridge Centre for Molecular Recognition, Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1QW, United Kingdom
[2] Medical Research Council Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom

(RECEIVED May 4, 1994; ACCEPTED July 25, 1994)

Abstract

We have developed a general quantitative methodology for designing proteins de novo, which automatically produces sequences for any given plausible protein structure. The method incorporates statistical information, a theoretical description of protein structure, and motifs described in the literature. A model system embodying a portion of the quantitative methodology has been used to design many protein sequences for the phage 434 Cro and fibronectin type III domain folds, as well as several other structures. Residue sequences selected by this prototype share no significant identity with any natural protein. Nonetheless, 3-dimensional models of the designed sequences appear generally plausible. When examined using secondary structure prediction methods and profile analysis, the designed sequences generally score considerably better than the natural ones. The designed sequences are also in reasonable agreement with a sequence template. This quantitative methodology is likely to be capable of successfully designing new proteins and yielding fundamental insights about the determinants of protein structure.

Keywords: de novo protein design; fibronectin type III domain; phage 434 Cro; quantitative; secondary structure prediction; sequence profiles; statistics

Over the last few decades, advances in X-ray crystallography and high-field NMR spectroscopy have permitted the elucidation of many protein structures, which has led to a much increased understanding of some of the basic principles governing protein folds. At the same time, the advent of modern molecular biology has provided the methods to manipulate the amino acid sequences of proteins and even to build new enzyme activities into known, existing enzymes (Fersht & Winter, 1992; A. Berry, N.S. Scrutton, R.N. Perham, in prep.). However, one of the ultimate aims of structural molecular biology, the construction of tailor-made enzymes for desired functions, is still some way off. Nevertheless, we are now in a position to attempt the first stage in this process: designing new amino acid sequences to fold into desired structures.

By selecting sequences inaccessible to nature but theoretically plausible for a given fold, we can ask precise questions about the general determinants of protein structure (Pastore & Lesk, 1991). In addition to the increase in fundamental scientific understanding that this will bring (Pabo, 1983), there are tremendous potential medicinal and biotechnological uses for custom-designed proteins (Drexler, 1981).

Several de novo protein design experiments have already sought alternative amino acid sequences for well-studied folds (Sander et al., 1992b; for a review, see DeGrado & Matthews, 1993), such as the 4-helix bundle (Regan & DeGrado, 1988; Hecht et al., 1990; Hill et al., 1990; Sander et al., 1992a; Schafmeister et al., 1993), β-bell structures (Richardson & Richardson, 1987), and α/β-barrels (Hubbard & Blundell, 1989; Goraj et al., 1990). In a similar manner, folds not found naturally, such as an "open sandwich," have been designed (Fedorov et al., 1992). Additionally, nonprotein templates such as circularized peptides (Mutter et al., 1992, 1993), porphyrin macrocycles (Akerfeldt et al., 1992), or metal atoms (Ghadiri et al., 1992; Sasaki & Lieberman, 1993) have been employed to bind individual elements of secondary structure together and thus have expedited some attempts to build proteinlike structures. For all of their diversity, however, virtually all these design attempts have shared a similar methodology. The most important step in the protocol, choosing an appropriate amino acid sequence, was performed manually using "physical, statistical, and intuitive criteria" (Sander, 1991).

However, protein design is a delicate task requiring a panoply of subtle criteria to be examined and carefully balanced against each other. Most protein design experiments have failed to conclusively demonstrate that a correct fold was formed. Al-

Reprint requests to: Alan Berry, Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK; e-mail: ab117@biovax.leeds.ac.uk. Correspondence to: Steven E. Brenner, MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK; e-mail: s.e.brenner@bioc.cam.ac.uk.

[3] Present address: Department of Biochemistry and Molecular Biology, University of Leeds, Leeds LS2 9JT, UK.

though very recent experiments show that some new proteins have actually adopted their desired structures (Schafmeister et al., 1993; Kuroda et al., 1994; Tanaka et al., 1994), these have required intimate knowledge of the desired fold that cannot be easily applied to other structures. In short, previous reports have demonstrated that protein design is too complex for manual and intuitive approaches to have significant general success on a large scale. Additionally, although manual methods may succeed in building small proteins, to extract fundamental information about the nature of protein structure from such studies, a generalized approach to protein design is necessary. Consequently, to design proteins successfully and concomitantly learn about the determinants of protein structure, we must create a quantitative method that uses a set of discrete rules for transforming a desired structure into a sequence.

A quantitative methodology

We have developed a conservative methodology for the de novo design of proteins that is principally a quantitative version of the intuitive methods used by other investigators. The system is based upon statistics derived from the native structures of naturally occurring proteins as deposited in the Brookhaven National Laboratory's Protein Data Bank (PDB) (Bernstein et al., 1977; Abola et al., 1987), as well as theoretical models of protein structure. We have chosen to consider virtually all variables that could have effects on structure and that can be measured and fruitfully applied to protein design. It is probable that some of these features are of only marginal importance. However, our methodology provides the ability to explicitly specify the importance of one feature relative to another, and rules derived from less important parameters can be assigned lower weightings.

The method operates by scoring a sequence on several criteria and then computing a weighted sum that is designated the sequence's "quality" for the desired structure. The fold-determining factors in the design methodology can be divided into 4 broad categories: *position*, *neighbor*, *uniqueness*, and *hints*. These are described briefly below. A function of these various parameters is optimized to find the best sequence (i.e., the one with the highest quality score). Thus, as shown in Figure 1, the methodology acts as a machine accepting general information about proteins and a particular fold to design, and produces a sequence.

The core of the design system is in the first 2 categories. These are comprised of "statistical residue preference parameters," whose analysis and use have been described by several groups (Sander et al., 1992a; Singh & Thornton, 1992). These parameters are derived by scanning the protein structure database and ascertaining various statistical features of individual residues.

Position preferences are those that relate a position in a protein structure with an amino acid at that point and are independent of the types of other residues in the protein. A canonical example of this class is secondary structure preference, which is derived from the frequency (in the PDB) of each type of residue in each of helices, sheets, and turns. Because 47% of alanines occur in helices, it would be preferred in a helix over proline, which occurs in helices only 17% of the time. More precise data could be collected by considering the preferred position of various residues within structures, to reflect, for example, the 3.5:1 preference for asparagine at the N-cap position of helices versus other positions (Richardson & Richardson, 1988). Another very important member of this category is solvent accessibility, reflecting the preference of different residues to be found buried within proteins or exposed on the surface. Less important contributions would be made from statistical preferences for different torsion angles and side-chain bonding characteristics.

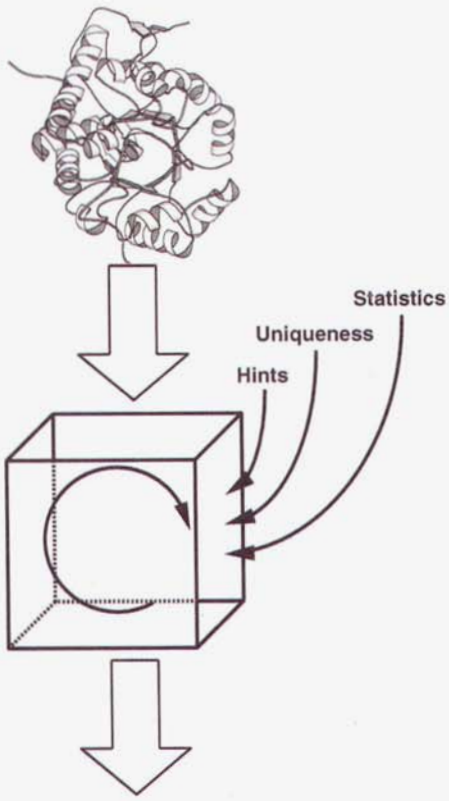

Fig. 1. Schematic diagram of the quantitative protein design system. The design system incorporates statistical parameters derived from known protein structures, various qualitative observations about protein structure and other hints necessary to design satisfactory sequences, and a theoretical model of protein structure to help ensure prohibition of the formation of alternative structures. It then applies this information to any protein structure to automatically compute a sequence that should fold into that structure.

The second group of parameters considered can be termed *neighbor preferences*. These parameters describe the likelihood of a given residue having specific atoms, residues, or structures nearby. The most intuitive parameter in this category is spatial neighbor, which incorporates the contact surface area between one type of residue and its neighbors; for example, leucine touches other leucine residues far more often than it does glutamic acid (Sander et al., 1992a). In addition, one can look specifically at those interactions that are found between and within elements of secondary structure. Furthermore, primary sequence neighbor preferences also appear to be significant (Richardson & Richardson, 1987; Neher, 1994). Although the spatial neighbor preferences embody some information about the size of different residues, this information should also be explicitly included to ensure correct uniform packing density (i.e., to avoid "holes" and steric clashes) using methods such as that of Gregoret and Cohen (1990).

The third type of parameter to be considered by the protein design system is *uniqueness*. This reflects the need for what Richardson has called "negative design," which is needed because "showing that a sequence fits well with one particular structure does nothing to prove that there is not another structure it fits even better" (Richardson et al., 1992). Yue and Dill have developed heuristics that can help ensure that not only will the designed sequence be of low energy in the desired conformation, but that it will have high energy in all other folds (Yue & Dill, 1992). As shown in Figure 2, this method operates on the simple assumption that there are only 2 types of residues, hydrophobic and polar, that lie on a rectilinear lattice, and that the sole determinant of protein folding is the maximization of hydrophobic–hydrophobic interactions. Although this model is quite removed from real proteins, it can provide useful guidelines about how to avoid making sequences that appear suitable for the desired structure but actually fold into other conformations of even lower energy.

Finally, "*hints*" is a broad category of parameters that contains any data about the desired protein needed by the design system that is not included in the previous statistical parameters. For example, large features documented in the literature (e.g., the hydrophobic diamond at helix–sheet interfaces [Cohen et al., 1982]) can be incorporated here, as well as more generalized features, such as folding initiation sites as proposed by Moult and Unger (1991). Additionally, *hints* contains information about any functionality the protein should have, details that will make the protein easier to synthesize, and patterns that will facilitate inferring and solving the protein's structure. Finally, successful design should not be sacrificed for complete generality. Therefore, if the design system consistently selects an implausible segment of sequence and no general rules seem capable of rectifying the problem, the system will be explicitly told what to do. Although this would reflect a partial failure of the general quantitative approach, it would not be surprising to discover that certain folds have some unique properties that the purely statistical parameters fail to incorporate.

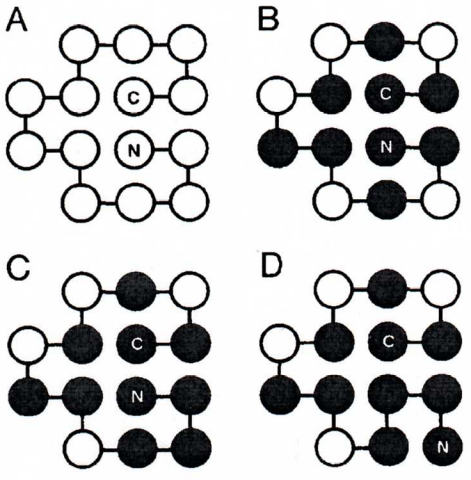

Fig. 2. Model of the binary lattice design system developed by Yue and Dill (1992). In this model, the quality of the structure is the sum of hydrophobic (dark) residues neighboring other hydrophobics; polar (white) residues have no interactions, and the structure to be designed is shown in **A**. Two optimal-quality sequences for this fold are shown in **B** and **C**. However, the variation at position 3 means that the sequence shown in C can also adopt the structure in **D** equally well. The heuristics developed by Yue and Dill would select sequence B over C, even though they are of equal quality, because B incorporates "negative design" for structure D. (Adapted from Yue & Dill [1992], with permission.)

A model system

In order to assess the feasibility of such a general quantitative design method and to examine the effects of varying the weighting of the different design parameters described above, a simplified design system was created using a subset of rules extracted from the overall methodology. We have used the model system to design various protein structures and have subjected the resulting new amino acid sequences to a battery of subjective and objective analyses.

Rather than taking a generalized protein structure as input, the prototype system requires data about a real protein, namely the secondary structure into which each residue is to fit, and its solvent accessibility. The rules in the model system are based only upon secondary structure, solvent accessibility, primary sequence neighbors, and diversity (see below, and Methods). As described in the Methods, measures of these parameters were summed to form a pseudo-energy function that was optimized by simulated annealing to find the best sequence.

As an example of the results generated by the prototype system, we describe here some sequences selected to fold into the same structure as phage 434 Cro protein—PDB structure 2CRO (Mondragón et al., 1989) and tenascin's third fibronectin type III domain—1TEN (Leahy et al., 1992). Sequences for several other protein folds, including lactate dehydrogenase—5LDH (Grau et al., 1981), hemoglobin—4HHB (Fermi et al., 1984), and ubiquitin—1UBQ (Vijaykumar et al., 1987), were also designed. Because the design system is general, it is capable of choosing sequences for any other plausible structure as well.

Design of protein sequences to adopt the Cro fold

The Cro protein from phage 434 contains 5 α-helices and is a member of the large class of helix-turn-helix DNA-binding proteins (Dodd & Egan, 1990). Figure 3 shows the amino acid sequence of the natural 434 Cro protein (line a), its crystallographically assigned secondary structure (line b), and the sequences of designed proteins produced by the prototype system using different weightings for the constituent parameters described above (lines 1–50). A comparison of the sequences designed using the same weighting criteria shows that the system was annealed too rapidly to find the global minima for a given weighting. However, the variations in sequence are very small, and variations in quality (as measured by the design system) were usually far less than 0.1% (data not shown).

One of the most obvious, yet most intriguing, observations about the designed sequences is that none show significant identity to the natural Cro protein. Indeed, a database search found that none of the designed sequences had significant identity to any known natural protein. Nevertheless, all the designed sequences show considerable similarity to each other, but no positions are strictly conserved. With the exception of those sequences generated by extremely high diversity requirements, all proteins designed using the same rule weightings are nearly identical, and the substitutions fall into certain patterns, isoleucine for valine, for example. Because multiple runs of selection

Designed Sequence
Derivation
Prediction
Profile
Phage 434 Cro
434 Cro structure

procedure with the same parameters all begin with different random sequences but converge to nearly the same final sequence as a result of random mutations (Fig. 4), the method is clearly specific.

Some interesting points to note about the designed proteins are that buried polar residues in the real protein, such as the Glu-35 that contacts helices 1 and 2, are almost always replaced with nonpolar residues. Similarly, because hydrogen bonding is not explicitly described in the methodology, it was not surprising to find that Gln-28 and Gln-32, which interact with each other, were replaced by a variety of different residues in the designed sequences.

It was noted from the structural determination of 434 Cro protein that residues 27–30, 32, and 33 are important in making contact with the DNA (Mondragón et al., 1989). However, the model system is intended only to produce amino acid sequences that will adopt the same fold as naturally occurring proteins, and it specifically does not attempt to incorporate function into the designed proteins. Thus, with the exception of the highly solvated lysine at position 27, the design system usually selected different residues at DNA-binding positions because it knew nothing of the protein's function.

Three-dimensional models of designed Cro

Several models of each designed sequence's structure were computed (see Methods) and, although the resultant structures should not be considered real, they do provide some insight into the prototype design system's strengths and weaknesses. A structural model of one of the designed proteins is shown in Figure 5 and Kinemage 1.

General inspection of the structure shows that it is not implausible, although there may be too few interhelix contacts, with a tryptophan filling what would otherwise be a gap. As Figure 6 shows, there are a considerable number of bad van der Waals contacts. However, most of these result from 2 prolines whose position had been overconstrained by holding the backbone constant while building the model. Therefore, these can probably be dismissed as an artifact of the model and not the sequence. In any case, the lack of optimal internal structure is not remarkable, for the prototype design system's rules only tangentially address this issue through solvent accessibility and primary sequence preference. The large number of residues with long, charged side chains also seems unusual and is probably due to both the method of considering solvent accessibility (see Discussion) and the large number of charged side chains found in the natural Cro protein, for binding DNA.

Similarly, although most structural components of the protein seem appropriate to their position, some detailed information is missing. For example, Mondragón et al. (1989) note that "the glycine at position 25 seems to be important to accommodate a tight bend. . . . the presence of an alanine at position 21, almost invariant in all related molecules, is also forced by the sharpness of the bend which reduces the space available." Only a few of the designed structures (notably, those with very high weightings on secondary structure preference) chose residues appropriate to the tight turn.

Objective analysis of designed Cro sequences

The designed proteins were subjected to 2 types of analysis to assess quality. A secondary structure prediction algorithm was used to determine whether correct secondary structure elements were likely to form, and a profile method was employed to evaluate tertiary structure.

The PHD secondary structure prediction method (Rost & Sander, 1992, 1993) was usually able to detect correctly the positions of helices and turns in designed sequences, although it was often incorrect in determining their boundaries (Fig. 3). Indeed, the predicted structures of most designed proteins were closer to the correct structure than that predicted from the natural Cro sequence—even though Cro was part of the data set used to design the prediction system. The PHD algorithm works, in part, by aligning query sequences with others of known structure. As expected, most sequences could not be aligned with any other known sequence. However, PHD did occasionally claim to find homologues to the designed sequences, even though other methods did not find any. Most sequences with PHD-assigned homologues (indicated by asterisks in Fig. 3) had seriously flawed secondary structure predictions.

A score was computed by comparing PHD's predictions for each sequence with the desired structure (Fig. 3, line b). Those sequences that were designed with a heavy emphasis on secondary structure rules (e.g., sequence 44) generally scored very well. Noticeably, sequences constructed with solvent accessibility rules weighted heavily (e.g., sequence 47) scored very poorly.

The potential of each designed sequence to fold into the desired tertiary structure was measured using a 3-dimensional profile method (Lüthy et al., 1992), which determines whether a sequence is compatible with a given structure. Many of the designed sequences score better with its criteria than the natural Cro sequence (Fig. 3), and all but one (a high diversity sequence) are deemed acceptable. Because secondary structure preference is also part of the 3-dimensional profile, those sequences generated with a heavy weighting on secondary structure do particularly well. However, this is probably because the criteria used to design the sequences are not independent of those used to assess them by the profile. The diversity hint appears to have

Fig. 3 (*facing page*). Designed sequences and analysis. Line a shows the sequence of phage 434 Cro, and line b numbers the residues from −1 to 63 in accordance with Mondragón et al. (1989). Lines 1–50 list designed sequences. The colors of the sequences (a and 1–50) show the secondary structure predictions: yellow for helix, red for sheet, and blue for turn, whereas line b shows the actual secondary structure of 434 Cro. The orange secondary structure column indicates how well the predicted structure agrees with the desired structure, with the bar for line b indicating what score a perfect prediction would receive. An asterisk next to the line number indicates that the secondary structure algorithm inappropriately aligned the sequence with others. In the magenta profile column, the scores indicating how appropriate the sequence is for the 434 Cro structure are shown. Any length greater than 0 indicates that the sequence is suitable for the structure; only sequence 38 is inappropriate. The derivation indicates the relative weighting of the parameters in the design system (see Methods), where secondary structure is red, solvent accessibility is yellow, primary sequence neighbor is green, and diversity is blue.

```
ZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZZ
DTFCGEWIBZAOKPGUJPLGQRGJHBAVBLSKESNVBCTSPUGBOJKJBKAPWNKXKVHAFHALV
NKWSIMLFIRHIQNBHEFILGTVRSDSBQQPMXGMZQKQYHLQTJOMTKQYRHWDPZLIHVTSSB
BQBPLKEHDPYNICMTAXTHEJDMKQFFWVGLFSHRHIMYOIEHTLFPDFSTMQPQRVRPUWAYG
MVPGIHFFPKHLSAWEVAMRMHYQIARGHKDAPRTGVPSQLPPFVYLEEIVKTCKPANRRSHMNR
MQVFGWSWDSQAERLRMARNYGIMKUGFNNRMMHDFWYVNEOKGCLSXSMKEIEDOUTTNQIDAZ
MQLFIEVVTHNMDYFYAPTRQQEWDFIGGSTLERRLISGKSTMSUSAPEVADGYQGFFNHLEKLK
KEWJIEGANQEFEPLDNRLKPQYEERHPRHKAGRSZLPGYCITGVSVDREAEAESGFMQTVSKVD
MFLLFEGRRKDLDSGKGDHEEHAQILARNPKLWRNVAIWQGRESRYTEQHKFVLPANKDNYHKPA
MTNFIQIKEKHIRHWSLRIDKYLEDVTSQDDHHRQTGRSQGKSALNLDELZEINYSETKNVQTGR
MKTIVRGVRKHAKNFKITYNECAEAGYGRDKSLEDMVPSPGQRVNSFKEATRFPKGWTNDFQDAK
MRFLVEIVEKEAKHFYGNYQRCMKELPGKDKSAKDIGSGKAEKGKPLKRILRLPQAEAENWNKIK
MDTVIRLYEKEAKNIKGHYDEIVKKGTGKNKSAKQFVSGQAQRGKPLKEAPRVPKGRLEDWNKFK
MKTIVEVARKEAKDFDGNHQELIKKGPGKNKYCRQILSVKAQKGNPLKEALELPKGRFRDWYKAK
MKTVIEVARKEAKDFNGNHQEALKKGPGKDKYCKQIVPLRAQKGKPLKEVTRINKGRFEDWYKAK
MKSVIRVLEKEAKDFKGNHQEALKKGTGKDKYCKQAVPIRAQKGNPLKEIVRLPKGKFRDWYKAK
MKTVIEVLEKEAKDFKGNHQRAVKKGPGKDKYCKQAVPLRAQKGKPLKEINKINKGRFRDWYKAK
MKSVIEVLEKEAKDFKGNHQRAVKKGPGKDKYCKQAVPLRAQKGNPLKEILKINKGRFRDWYKAK
MKTVIEVLEKEAKDFKGNHQRAVKKGPGKDKYCKQAVPLRAQKGKPLKEINKINKGRFRDWYKAK
MKTVIEVLEKEAKDFKGNHQRAVKKGPGKDKYCKQAVPLRAQKGNPLKEILKINKGRFRDWYKAK
MKTVIEVLEKEAKDFKGNHQRAVKKGPGKDKYCKQAVPLRAQKGKPLKEINKINKGRFRDWYKAK
```

Fig. 4. Annealing of designed Cro sequence. A few of the sequences along the annealing pathway for line 1 of Figure 3 are shown. The initial nonsense sequence is converted to a random sequence, which is repeatedly mutated. Slowly the sequence anneals to the final selection. Note how some positions (e.g., the N-terminal methionine) become fixed very early in the annealing process, whereas others remain variable until the sequence is frozen. Additional intermediates are shown on the Diskette Appendix.

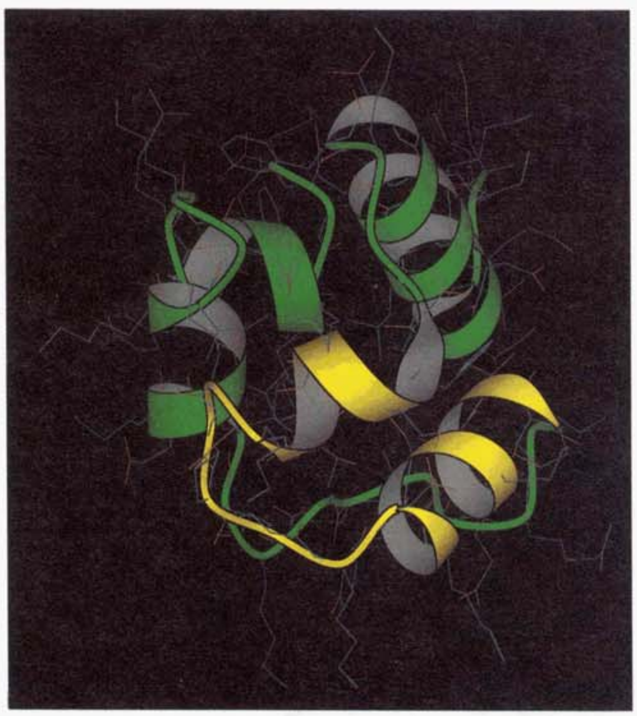

Fig. 5. Model of designed protein. The backbone structure is that of phage 434 Cro, and the model was created as described in the Methods. The N-terminus is on the right, and helices 2 and 3, which interact with DNA in the natural protein, are shown in yellow. The image was generated with MOLSCRIPT (Kraulis, 1991). The coordinates are on the Diskette Appendix.

a detrimental effect when it dominates the design criteria, but other trends are hard to find, particularly because the variation in the profile quality of sequences generated by the same rules is considerable.

Sequences designed to adopt the fibronectin type III fold

That the Cro designed sequences showed no significant sequence identity to the natural sequence was expected because the natural protein represents only a single point in an enormous potential sequence space. Indeed, in many large protein families, some members have virtually no pairwise sequence identity with each other. However, thus far it has always been possible to create a template for such families that specifies a few positions at which there is considerable conservation of either an individual residue or a class of residues.

Successfully designed proteins need not even fit the template for a widely diverged family because the path taken by evolution to create the family's members may have greatly restricted the sequence space available. In addition, functionally important residues generally appear prominently in templates but are irrelevant for the design of structure. However, one would still expect that designed sequences would be somewhat compatible with templates for the same fold because the templates incorporate some general information necessary for determining structure.

One protein family with a large number of highly diverged members and for which a template has been derived is the fibronectin type III (Fn3) domain. Fn3 structures are about 90 residues long and consist of 7 β-strands. These domains are used

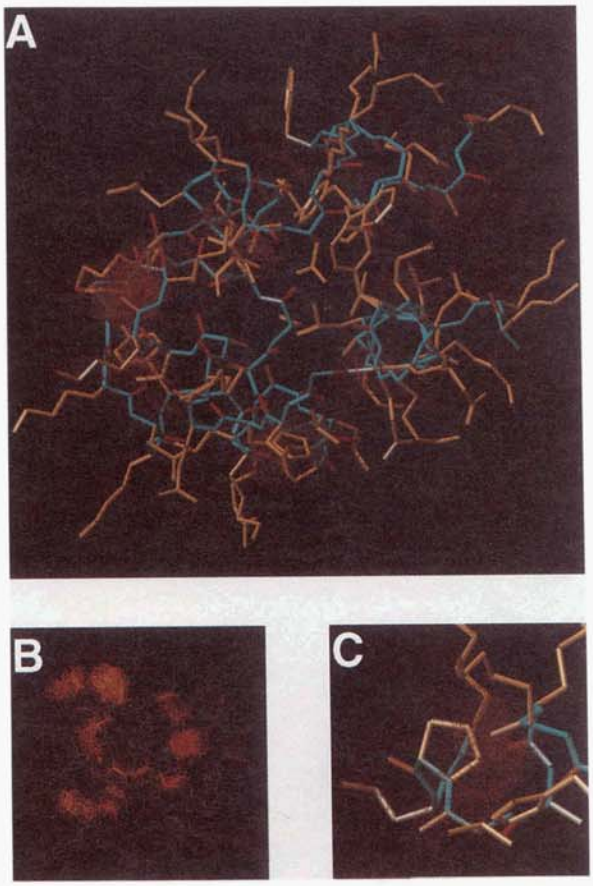

Fig. 6. Steric visualization of designed protein. **A** and **B** show 2 different views of the same designed protein with red shells indicating bad van der Waals contacts. **C**, which portrays 1 of the 2 dense regions in detail, shows that a proline in tight turn is the cause of most steric clashes, which are probably an artifact of the modeling procedure. The figures were generated by the program Sculpt (Surles, 1992).

for protein–protein interaction (Pierschbacher & Ruoslahti, 1984) and frequently occur in tandem chains, presumably as "spacers" (Campbell & Spitzfaden, 1994; Huber et al., 1994). The Bork and Doolittle (1993) template for this structure specifies particular residues at 6 positions and residue classes (hydrophobic or turn) at 17 positions. These positions are graphically represented on the structure in Figure 7.

The third Fn3 domain in tenascin (Leahy et al., 1992) was used as input to the model design system. Although the designed sequences have virtually none of the consensus identities in the template, they are otherwise quite compatible with it (Fig. 8). In particular, although precise matches of designed sequence with the template might not be expected for the reasons noted above, about 70% of designed sequence positions are in a residue class that matches the template. For example, KRP is a good sequence for the turn at positions 824–826, and F is almost certainly a reasonable substitution for the conserved Y at position 879. The designed sequences usually have hydrophobic residues at hydrophobic positions in the consensus. Only a few positions are incompatible with the Fn3 template, and careful inspection of the structure suggests that these substitutions are generally reasonable. Perhaps the only worrying discrepancy is the frequent selection of D at buried position 837, but this appears to be an artifact of the way solvent accessibility is measured in the model system (as noted below).

These results do not conclusively prove that the designed sequences would adopt the correct structure, but they demonstrate that even the simple model design system selects generally reasonable sequences that usually fit the sequence consensus. These results also highlight areas where the natural sequences may have been evolutionarily limited or where the model system requires additional parameters.

Discussion

Methodologies

Design by modeling has now been successfully employed to design new sequences for particular protein structures. Moreover, these methods have proven capable of asking specific questions about the particular folds in a way that exploration of natural sequences cannot. Design may therefore also provide insight into the general nature of protein stability. However, the qualitative and intuitive nature of manual design makes it impossible to fully describe all of the criteria and knowledge that enter into the method; design by modeling is fundamentally irreproducible.

For this reason, it is necessary to make use of a fully qualified methodology to select new protein sequences. The reproducibility of the procedure when applied to different structures ensures that it does indeed encapsulate genuine and general information about protein structure. Furthermore, a quantitative method can be analytically dissected to reveal this information.

In addition to the fundamentally statistical approach described, several other potential quantitative methodologies exist for the design of proteins. Instinctively, one wishes to use the most detailed description of atomic systems available; however, quantum-wave theory is intractable for macromolecular design, and even molecular dynamics, although useful for exploring the conformation space of a few small peptides (Floegel & Mutter, 1992), is computationally too expensive and provides no clear path from structure to sequence.

In their seminal work exploring the sequences compatible with particular structures, Ponder and Richards (1987) developed an algorithm to determine which residues could pack to form a particular structure; as such, it designs sequences whose side-chain rotamers are sterically compatible with a given fold. However, this criterion is almost certainly not sufficient to guarantee proper folding of a structure. Moreover, we are attempting to select sequences that adopt a particular fold without requiring precise distances between elements of the protein, because even highly homologous proteins have small variations of structure within the core (Chothia & Lesk, 1986). However, permitting this flexibility renders the Ponder and Richard's computation extremely difficult and allows far too many sequences as possibilities.

Consequently, other groups have attacked the protein design problem by making considerable simplifications. A simple lattice model (Yue & Dill, 1992) has already been discussed. A method operating on the same hydrophobic/polar model has been developed by Shakhnovich and Gutin (1993) that principally attempts to optimize interresidue contacts in a given structure, much like the neighbor preferences described above. Yue et al. (1994) have found that this method fails the uniqueness criterion; however, it appears that when expanded to systems

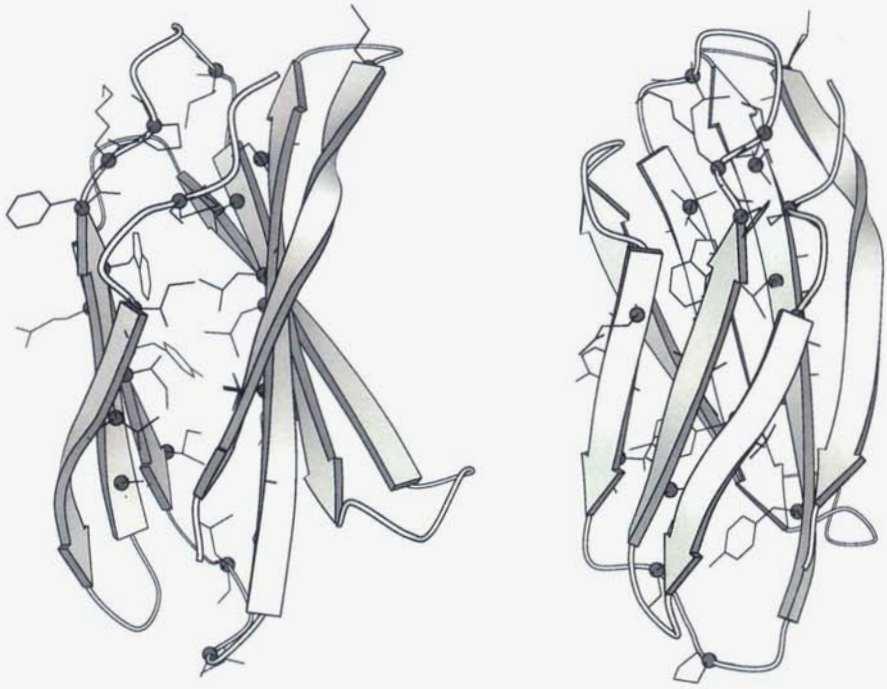

Fig. 7. Cartoons of tenascin third Fn3 repeat, highlighting template residues. The Fn3 domain has 7 strands: A, B, C, C′, E, F, and G. The residue positions that form the template defined by Bork and Doolittle (1993) are drawn, with large spheres at the α-carbon positions. The twist at the beginning of strand G has been described as a single turn of a nonideal polyproline II helix, which is longer in other Fn3 domains (Huber et al., 1994). The coordinates are taken from PDB entry 1TEN (Leahy et al., 1992) and the image was generated with MOLSCRIPT (Kraulis, 1991).

with real residues, the method may produce better results. Elegant experimental work exploring this simplified domain of protein design was done by Kamtekar and coworkers (1993), who constructed 4-helix bundles where the helices contained random hydrophobic residues on the interior and hydrophilic on the exterior. They found that 60% of the sequences were soluble and resistant to intracellular degradation, whereas only 5% of random sequences of 80–100 glutamine, leucine, and arginine residues meet similar criteria (Davidson & Sauer, 1994).

We have taken an approach that lies closer to the methods used by most investigators, using a variety of statistical measures that are supplemented by a uniqueness criterion. However, the decision to embrace statistics either explicitly, as described here, or implicitly, as in most design projects, entails 2 fundamental assumptions. First, statistics about residues must be capable of incorporating sufficient thermodynamic information to form correctly folded proteins. Although this claim is tenuous, the principal rationales for believing it are: (1) many statistics appear meaningful, e.g., proline occurs mainly at turns, (2) proteins designed using these criteria have usually approximated the desired structure in some way (Richardson & Richardson, 1989; Lovejoy et al., 1993), and (3) methods based on statistical data, such as templates, profiles, and threading, have had remarkable success in assessing the correctness of protein models and in predicting protein structure (Overington et al., 1990; Bowie et al., 1991; Lesk & Boswell, 1992; Lüthy et al., 1992). In short, statistics seem to work, but protein structure is not well enough understood to fully realize why.

The second assumption, that thermodynamics will be completely dominant over kinetics in the formation of the protein

```
a    P  h  h       h  h Wttt   h   h h h                    th h   L P   Y  h h A    t
b ldapsqie vkdvtdttal itwfkplaei dgieltygik dvpgdrttid ltedenqysi gnlkpdteye vslisrrgdm ssnpaketft t
c       AA AAAAA  BBB BBBB        CCCCCCCC      C'C'C' C'   EEEEE E     FFFF FFFFFFF  G GGGGGGGGGG G
d 34567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1234567890 1
1 MDVAKRMK NKRNHQYSVF TIVKRPKGKP QGVDISDINP KDENPKQKYK ARTEDEKEKV AKGKTKGKFI TFTYVKRGKE KAKHWKQEIK G
2 MDIGKRMK NKRHNQTPVF YVIKRPKGKP EGVDVSDINS KNQSPKQKYK ARTEDEKEKV AKGKTKGKFI TFTYCRKGKE KAKHWKQEIK A
3 MNIAKRMK NKRNHQPSTF TVIKRPKGKP QGVDVYDINS KDEYPKQKSK ARTEDEKEKV AKGKTKGKFI VFTYVKRGKE KAKHWKQEIK G
4 MNVAKRMK NKRNDQYSCF TVIKRPKGKP EGIDVSDITS KDENPKQKYK ARTQHEKEKV AKGKTKGKFI TFVYIKRGKE KPKHWKQEVK G
5 MNVAKRMK TKRDYQYSHF TVIKKPKGKP EGIDVSDINS KDENPKQKSK ARTEDEKEKV AKGKTKGKFV TFHYIKRGKR KPKNWKQEIK G
6 MDIPKRMK NKRNYQYSVF TVIKRAKGKP EGVDISDINS KDENPKQKSK ARTQHEKEKV ARGKTKGKFV TFTYVKKGKE KPKHWKQEIK G
7 MHIAKRMK NKKDNQHPVF TVIKRPKGKP EGVDISDITS KDENPKQKYK ARTEDEKEKV AKGKTKGKFV YFTYCRKGKR KAKNWKQEIK G
8 MDIAKRMK NKRDNQTSVF TIVKRPKGKP EGVDYSDVNP KHQYPKQKSK ARTEDEKEKV AKGKTKGKFI TFHYIKKGKR KAKNWKQEIK G
9 MHVAKRMK NKRDYQTSVF TAVKKPKGKP EGVDVSDINS KNENPKQKYK ARTQHEKEKI AKGKTKGKFI TFIYVKRGKR KPKDWKQEIK G
0 MNIGKRMK NKRNHQYSVF TVAKRPKGKP EGIDISDVTS KDEHPKQKYK ARTEDEKEKV AKGKTKGKFV TFIYVKKGKR KPKNWKQEIK A
```

Fig. 8. Fibronectin type III template, tenascin third Fn3 repeat sequence, and several designed sequences. The template, a, is that of Bork and Doolittle (1993), where capital letters indicate conserved residues, and lowercase letters represent conserved classes: h is hydrophobic and t is turn. The numbering, d, of the tenascin sequence, b, starts at 803 and follows that of Leahy et al. (1992). The positions of the crystallographically determined strands are shown on line c. The 10 designed sequences (1–0) were all generated using the same parameter mixture as the first 10 designed sequences in Figure 3. Additional Fn3 designed sequences, with the same variation in parameters as the Cro sequences in Figure 3, are on the Diskette Appendix.

structures, means folding intermediates need not be considered. This claim is slightly easier to accept, if only for the very small structures now being built. For proteins of this size, it is difficult to imagine what absolutely necessary folding intermediates might be present. The only considerable worry is that a given secondary structure element may not be initiated or terminated properly, and that other structure elements may then induce an incorrect fold. This issue will gain greater importance as larger proteins are designed and built.

A credulous interpretation of the results suggests that these assumptions do hold. Even with the very simple criteria used by the model design system, it was able to choose sequences that appeared plausible to humans and excellent to a variety of analysis methods. It is perhaps even more fascinating that, even when the relative importance of these different criteria were varied widely, the resulting sequences, although quite different, nearly always received similarly high scores from the analysis methods. The scores' resilience to different parameter weightings may be a genuine consequence of enormous flexibility in natural sequence–structure relationships. More likely, however, it reflects our ignorance of their complexities.

If the analysis systems are indeed very imprecise and are falsely suggesting that the designed protein sequences are appropriate, then one of the largest problems to be surmounted for successful protein design by the generalized quantitative system is the determination of the relative importance of the various parameters. For example, if only a small set of sequences received high scores, this could be taken as an indication that the parameter weightings used to design those sequences were somehow better than other combinations; as it is, we are able to draw no such conclusions.

Because the analyses reveal little about the relative importance of different parameters, they usually provide even less information about the particular nuances of measuring a particular parameter. For example, although primary sequence neighbor preference appears easy—if not trivial—to describe, there are a wide number of different means of collecting this information, all of which produce different results. For example, the most straightforward approach is to simply search a protein sequence database and count how many times 2 particular residues occur in apposition, perhaps making the order significant. However, this incorporates bias in residue frequency, which may or may not be desired. The termini presented another problem because it is unclear what weighting they should receive. Finally, it must be decided if the parameter should vary linearly with the number of occurrences of a given pair or if it should be a more complex function, perhaps representing the information content (Garnier et al., 1978; Sander et al., 1992a). Without useful feedback from the design system, these decisions become disappointingly arbitrary and are guided principally by scientific intuition.

Despite the difficulty in using the objective analyses of the designed sequences to refine the quantitative design methodology, it is still possible to learn from the sequences themselves and the models of their structure. Early in the design process, even before full sequences had been generated, some problems with particular parameters appeared. For instance, alanine occurs adjacent to alanine more often than next to any other residue in sequence databases. However, this means that the primary structure neighbor preference parameter will give extremely high scores to long chains of polyalanine. Although these long stretches may be reasonable to provide flexibility in loops and linker regions, they certainly appear odd in the core of a globular protein structure. The conclusion is that statistical neighbor preferences generated from all protein sequences do not provide an excellent model for individual domains. In the model system, this particular problem was partially compensated for by reducing the primary structure neighbor preference of each residue for itself.

A similar problem arose with the measurement of solvent accessibility. One of the simplest possible models for optimizing solvent accessibility is to place residues in positions where the exposure to water is approximately equal to their own "preferred" average solvent accessibility. However, in phage 434 Cro, the area exposed to solvent by individual residues ranges from 0 to 170 $Å^2$, whereas the average solvent accessibilities for residues in α-helices in the PDB ranges from 10.8 $Å^2$ for cysteine to 94.6 $Å^2$ for lysine. It is not difficult to see why the averages are more limited in distribution than the actual values for exposure found in Cro. However, the solvent accessibility parameter consequently scores lysine as the optimal residue in 20% of the helix positions in Cro. In the prototype system, this problem was largely circumvented by introducing a "diversity" parameter that places a penalty on residues occurring as a greater fraction of the designed sequence than of average sequence in the protein databases. The need for this correction suggests that average solvent-accessible area is not a robust parameter for describing characteristics of residues.

By looking in detail at the designed sequences, it is also possible to learn about some features of natural proteins that are omitted from the prototype system. For example, as mentioned above, the prototype system will never incorporate internal hydrogen bonds that may be essential for the stability of a small protein. We can also see from the prototype system the importance of steric information in a design system if the core is to pack in a manner similar to that of natural proteins. To summarize, the model design system has provided some insight into important statistical features of proteins. Ironically, however, its very success at designing sequences that appear suitable to analysis has hindered efforts to abstract more precise data.

Conclusions

Protein design is a daunting task because of the vast number of subtle parameters and the nearly infinite number of possible conformations available to a polypeptide chain. From subjective and objective analyses of the sequences selected by the model protein design system, the quantitative methodology's potential to attack these problems is clear. With even a small set of rules and varied weightings, the system designed proteins that have reasonable sequences, and examination of computational models shows that they could fold into plausible structures. Objective analyses show that some of the designed sequences are very likely to form the correct secondary structure elements and could reasonably adopt the desired tertiary structure.

A successful protein design methodology—or even one that is only partially effective—will have many uses beyond its intended task of selecting sequences for new proteins. The protein design rules can be applied to a preexisting sequence and structure to verify that the two are compatible, similar to the profiles methods used above. The protein design system can also be used as a testbed to evaluate new parameters that might be

important in protein folding. Whereas the only straightforward method of assessing a general hypothesis about protein structure is to consider the same data (the PDB) used to generate it, this system could provide an "experimental" method of testing new ideas.

Moreover, the weightings of the parameters used to design new proteins will embody a relationship between easily measurable characteristics of protein sequences and structure. Consequently, in addition to building new proteins, the quantitative methodology holds the potential to provide fundamental insights into the general determinants of protein structure.

Methods

Model system

Parameters

In the model system, 4 parameters were used. Two were position preference parameters: secondary structure and solvent accessibility. Primary structure neighbor preference was the neighbor preference used, and diversity was a hint. No uniqueness parameters were incorporated.

The secondary structure term was computed by assigning every residue in a weighted representation of the complete PDB (S.E. Brenner & A. Berry, in prep.) a secondary structure with the computer program DSSP (Kabsch & Sander, 1983). For statistics quoted in this paper, outputs of G, H, and I were conflated as helices, but in the design system, all different identified structure types (B, E, G, H, I, S, T, and none) were treated independently. The secondary structure parameter $struc_{r,s}$ for a given residue r is equal to the fraction of occurrences of that residue in a particular secondary structure type s.

The solvent accessibility parameter $solv_{r,s}$ for each residue is computed independently for each of the DSSP-defined secondary structure types and is equal to the average solvent accessibility (in Å^2) of each residue r in a particular secondary structure s. It is important to separate the averages for different types of secondary structures because these exposures vary considerably.

We have collected the occurrence of sequentially neighboring residues in the SwissProt 25 database of protein sequences (Bairoch & Boeckmann, 1991) as preference information for this characteristic. A score $neigh_{i,j,l}$ was computed counting the fraction of times each residue (or terminus), j, occurs adjacent to each residue, i, on the C or N side, l, and then normalizing by the frequency of the neighbor, j. If the neighbors i and j were identical, the score was divided by 3.

Sequence selection

The model protein design program used simulated annealing to optimize a function that was the simple sum of qualities assigned by the various rules. The function was

$$F(S) = nD\,div(S) + \sum_{i=1}^{n} A\left(1 - \left|\frac{a_i - solv_{c_i,s_i}}{\max(a_i, solv_{c_i,s_i})}\right|\right) + B(struc_{c_i,s_i}) + C(neigh_{c_i,c_{i-1},C} + neigh_{c_i,c_{i+1},N}),$$

where F is the overall quality of the sequence for the desired structure, n is the sequence length, i is a position in the sequence, c_i is the residue at that position, and a_i and s_i are the desired solvent-exposed area and secondary structure, respectively, of position i. The parameters are *solv*, *struc*, and *neigh*, as described above. The diversity term is defined as

$$div(S) = \sum_{j=residues} 1 - \left|\frac{f_{j,swissprot} - f_{j,S}}{\max(f_{j,swissprot}, f_{j,S})}\right|,$$

where $f_{j,swissprot}$ and $f_{j,S}$ are the frequencies of the residue j in SwissProt 25 and the test sequence, respectively. A, B, C, and D are the parameters' weightings. The derivation column of Figure 3 shows the relative value of the terms after they have been divided by their neutral mix of $A = 2$, $B = 0.7$, $C = 3$, $D = 1$.

Starting with a "poly-Z" sequence (where Z is a dummy residue with 0 quality for all measures), the program applied a simulated annealing protocol (Press et al., 1988) using an annealing schedule of 0.9 (i.e., each successive temperature was 0.9 times the former one). The temperature was dropped after 1,000 iterations over every position without change or if there were the equivalent of 100 changes at each position. (The total number of changes, rather than the number of changes at each position was tracked.) To avoid propagation of major changes and to assist in parallelization, the sequence was iterated over with stride 3.

Computations were carried out on a cluster of Silicon Graphics workstations, a Convex 3800, an Intel Paragon, and a Cray T3D.

Analysis

Sequences

Sequences were compared against SwissProt 25 using Basic Local Alignment Search Tool (BLAST) (Altschul et al., 1990). Results were collected using the match matrix and all other parameters were set to their defaults. Matches were deemed to be possibly significant if $P(N) < 0.10$. Many additional queries made use of the National Center for Biotechnology Information experimental BLAST Network service.

Three-dimensional models

Three-dimensional models of the proteins were made in X-PLOR (Brünger, 1992), by first stripping the side-chain atoms from the structure of 434 Cro (Mondragón et al., 1989) in PDB entry 2CRO. All of the model sequence's heavy side-chain atoms were randomly placed in a Gaussian distribution centered at 1 Å from the α-carbon of each residue, and hydrogens were added to the appropriate heavy atoms. All subsequent steps occurred with the backbone fixed and with a force field including bond lengths and angles, van der Waals interactions, and improper and dihedral bond angles. First, the structure was subjected to 200 rounds of Powell minimization and then was subjected to Verlet dynamics while the temperature was gradually cooled from 1,000 K to 100 K and the van der Waals force increased. Finally, the structure was subjected to another 200 rounds of Powell minimization.

Objective analyses

The potential secondary structures of the sequences were analyzed using the PredictProtein email server at EMBL, which utilizes the PHD algorithm (Rost & Sander, 1992, 1993). A scoring method was used to compare the predicted structure of the designed sequences with the desired structure. For every posi-

tion assigned correctly, 2 points were added, whereas 1 point was deducted for every incorrect structure assignment.

To see whether a structure with the desired backbone coordinates would be compatible with the designed sequences, Verify_3D, a part of a sequence profile package (Bowie et al., 1991), was run on each of 7 structures computed (as above) for each sequence, and only the best score for each sequence was considered. The minimum acceptable score for a protein of Cro's size is 13 (Lüthy et al., 1992), so that was used as a 0 point in considering the data (i.e., 13 is subtracted from each Verify_3D score).

Supplementary material on Diskette Appendix

The Brenner.SUP subdirectory of the SUPLEMNT directory on the Diskette Appendix contains a summary file (Brenner.doc) and 4 files of supplementary information. The file CroTab.seq contains the Cro protein designed sequences listed in this paper in Figure 3. Fn3Align.seq contains sequences designed to adopt the structure of tenascin's third Fn3 repeat with the same parameters as those in Figure 3 (including those in Fig. 8), as well as the Bork and Doolittle (1993) template. The file Anneal.seq contains annealing intermediates (including those in Fig. 4) for several of the Cro designed sequences, and new2cro.xyz contains coordinates of the designed Cro protein structure shown in Figure 5. In the KINEMAGE directory, Brenner.kin shows a comparison of the original 434 Cro structure and the designed model.

Acknowledgments

We thank Dr. Cyrus Chothia for interesting and helpful discussions. S.E.B. was supported by a Herchel Smith Harvard Scholarship. A.B. was a Royal Society 1983 Research Fellow. The authors also thank the members of the CCMR for use of SGI workstations, NSF and SDSC for use of the Paragon, EPSRC and EPCC for use of the T3D, SERC and ULCC for use of the C3800, and the Royal Society for additional support of this work.

References

Abola EE, Bernstein FC, Bryant SH, Koetzle TF, Weng J. 1987. Protein Data Bank. In: Allen FH, Bergerhoff G, Sievers R, eds. *Crystallographic databases: Information content, software systems, scientific applications*. Cambridge: Data Commission of the International Union of Crystallography. pp 107–132.

Akerfeldt KS, Kim RM, Camac D, Groves JT, Lear JD, DeGrado WF. 1992. Tetraphilin: A four-helix proton channel built on a tetraphenylporphyrin framework. *J Am Chem Soc 114*:9656–9657.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol 215*:403–410.

Bairoch A, Boeckmann B. 1991. The Swiss-Prot protein-sequence data-bank. *Nucleic Acids Res 19 S*:2247–2248.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival for macromolecular structures. *J Mol Biol 112*:535–542.

Bork P, Doolittle RF. 1993. Fibronectin type-III modules in receptor phosphatase CD45 and tapeworm antigens. *Protein Sci 2*:1185–1187.

Bowie JU, Lüthy R, Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science 253*:164–170.

Brünger AT. 1992. *X-PLOR: A system for crystallography and NMR. Version 3.1*. New Haven, Connecticut: Yale University Press.

Campbell ID, Spitzfaden C. 1994. Building proteins with fibronectin type III modules. *Structure 2*:333–337.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J 5*:823–826.

Cohen FE, Sternberg MJE, Taylor WR. 1982. Analysis and prediction of the packing of alpha-helices against a beta-sheet in the tertiary structure of globular proteins. *J Mol Biol 156*:821–862.

Davidson AR, Sauer RT. 1994. Folded proteins occur frequently in libraries of random amino-acid sequences. *Proc Natl Acad Sci USA 91*:2146–2150.

DeGrado WF, Matthews BW. 1993. Engineering and design. *Curr Opin Struct Biol 3*:547–548.

Dodd IB, Egan JB. 1990. Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res 18*:5019–5026.

Drexler KE. 1981. Molecular engineering: An approach to the development of general capabilities for molecular manipulation. *Proc Natl Acad Sci USA 78*:5275–5278.

Fedorov AN, Dolgikh DA, Chemeris VV, Chernov BK, Finkelstein AV, Schulga AA, Alakhov YB, Kirpichnikov MP, Ptitsyn OB. 1992. De novo design, synthesis and study of albebetin, a polypeptide with a predetermined three-dimensional structure: Probing the structure at the nanogram level. *J Mol Biol 225*:927–931.

Fermi G, Perutz MF, Shaanan B, Fourme R. 1984. The crystal-structure of human deoxyhemoglobin at 1.74 Å resolution. *J Mol Biol 175*:159–174.

Fersht A, Winter G. 1992. Protein engineering. *Trends Biochem Sci 17*:292–294.

Floegel R, Mutter M. 1992. Molecular-dynamics conformational search of six cyclic-peptides used in the template assembled synthetic protein approach for protein de novo design. *Biopolymers 32*:1283–1310.

Garnier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol 120*:97–120.

Ghadiri MR, Soares C, Choi C. 1992. Design of an artificial four-helix bundle metalloprotein via a novel ruthenium (II)-assisted self-assembly process. *J Am Chem Soc 114*:4000–4002.

Goraj K, Renard A, Martial JA. 1990. Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modeled on the α/β-barrel proteins. *Protein Eng 3*:259–266.

Grau UM, Trommer WE, Rossman MG. 1981. Structure of the active ternary complex of pig heart lactate dehydrogenase with LAC-NAD at 2.7 Å resolution. *J Mol Biol 151*:289.

Gregoret LM, Cohen FE. 1990. Novel method for the rapid evaluation of packing in protein structures. *J Mol Biol 211*:959–974.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression, and characterization of felix: A four-helix bundle protein of native-like sequence. *Science 249*:884–891.

Hill CP, Anderson DH, Wesson L, DeGrado WF, Eisenberg D. 1990. Crystal structure of alpha-1: Implications for protein design. *Science 249*:543–546.

Hubbard TJ, Blundell TL. 1989. The design of novel proteins using a knowledge-based approach to computer-aided modelling. In: van Gunsteren WF, Weiner PK, eds. *Computer simulations of biomolecular systems: Theoretical and experimental applications*. Leiden, Holland: ESCOM. pp 168–182.

Huber AH, Wang YE, Bieber AJ, Bjorkman PJ. 1994. Crystal structure of tandem type III fibronectin domains from *Drosophila* neuroglian at 2.0 Å. *Neuron 12*:717–731.

Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Kamtekar S, Schiffer JM, Xiong HY, Babik JM, Hecht MH. 1993. Protein design by binary patterning of polar and nonpolar amino-acids. *Science 262*:1680–1685.

Kraulis PJ. 1991. MOLSCRIPT: A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 24*:946–950.

Kuroda Y, Nakai T, Ohkubo T. 1994. Solution structure of a de novo helical protein by 2D-NMR spectroscopy. *J Mol Biol 236*:862–868.

Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of the selenomethionyl protein. *Science 258*:987–991.

Lesk AM, Boswell DR. 1992. Does protein structure determine amino acid sequence. *Bioessays 14*:407–410.

Lovejoy B, Choe S, Cascio D, Mcrorie DK, DeGrado WF, Eisenberg D. 1993. Crystal structure of a synthetic triple-stranded alpha-helical bundle. *Science 259*:1288–1293.

Lüthy R, Bowie JU, Eisenberg D. 1992. Assessment of protein models with 3-dimensional profiles. *Nature 356*:83–85.

Mondragón A, Wolberger C, Harrison SC. 1989. Structure of phage-434 Cro protein at 2.35 Å resolution. *J Mol Biol 205*:179–188.

Moult J, Unger R. 1991. An analysis of protein folding pathways. *Biochemistry 30*:3816–3824.

Mutter M, Dorner B, Sigel C, Floegel R, Servis C, Tuchscherer G. 1993. Topological templates as a tool in molecular recognition and in protein design. *J Cell Biochem 17C*:211.

Mutter M, Tuchscherer GG, Miller C, Altmann KH, Carey RI, Wyss DF,

Labhardt AM, Rivier JE. 1992. Template-assembled synthetic proteins with 4-helix-bundle topology: Total chemical synthesis and conformational studies. *J Am Chem Soc 114*:1463–1470.

Neher E. 1994. How frequent are correlated changes in families of protein sequences? *Proc Natl Acad Sci USA 91*:98–102.

Overington J, Johnson MS, Sali A, Blundell TL. 1990. Tertiary structural constraints on protein evolutionary diversity: Templates, key residues and structure prediction. *Proc R Soc Lond B 241*:132–145.

Pabo C. 1983. Molecular technology: Designing proteins and peptides. *Nature 301*:200.

Pastore A, Lesk AM. 1991. Brave new proteins: What evolution reveals about protein structure. *Curr Opin Biotech 2*:592–598.

Pierschbacher MD, Ruoslahti E. 1984. Cell attachment activity of fibronectin can be duplicated by small synthetic fragments of the molecule. *Nature 309*:30–33.

Ponder JW, Richards FM. 1987. Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes. *J Mol Biol 193*:775–791.

Press WH, Flannery BP, Teukolsky SA, Vetterling WT. 1988. *Numerical recipes in C: The art of scientific computing*. Cambridge: Cambridge University Press.

Regan L, DeGrado WF. 1988. Characterization of a helical protein designed from first principles. *Science 241*:976–978.

Richardson JS, Richardson DC. 1987. Some design principles: Betabellin. In: Oxender DL, Fox CF, eds. *Protein engineering*. New York: Alan R. Liss, Inc. pp 149–163.

Richardson JS, Richardson DC. 1988. Amino-acid preferences for specific locations at the ends of alpha-helices. *Science 240*:1648–1652.

Richardson JS, Richardson DC. 1989. The de novo design of protein structures. *Trends Biochem Sci 14*:304–309.

Richardson JS, Richardson DC, Tweedy NB, Gernert KM, Quinn TP, Hecht MH, Erickson BW, Yan YB, Mcclain RD, Donlan ME, Surles MC. 1992. Looking at proteins: Representations, folding, packing, and design. *Biophys J 63*:1186–1209.

Rost B, Sander C. 1992. Jury returns on structure prediction. *Nature 360*:540.

Rost B, Sander C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol 232*:584–599.

Sander C. 1991. De novo design of proteins. *Curr Opin Struct Biol 1*:630–637.

Sander C, Scharf M, Schneider R. 1992a. Design of protein structures. In: Rees AR, Sternberg MJE, Wetzel R, eds. *Protein engineering: A practical approach*. Oxford: Oxford University Press. pp 89–115.

Sander C, Vriend G, Bazan F, Horovitz A, Nakamura H, Ribas L, Finkelstein AV, Lockhart A, Merkl R, Perry LJ, Emery SC, Gaboriaud C, Marks C, Moult J, Verlinde C, Eberhard M, Elofsson A, Hubbard TJP, Regan L, Banks J, Jappelli R, Lesk AM, Tramontano A. 1992b. Protein design on computers: Five new proteins: shpilka, grendel, fingerclasp, leather, and aida. *Proteins Struct Funct Genet 12*:105–110.

Sasaki T, Lieberman M. 1993. Between the secondary structure and the tertiary structure falls the globule: A problem in de novo protein design. *Tetrahedron 49*:3677–3689.

Schafmeister CE, Miercke LJW, Stroud RM. 1993. Structure at 2.5 Å of a designed peptide that maintains solubility of membrane-proteins. *Science 262*:734–738.

Shakhnovich EI, Gutin AM. 1993. A new approach to the design of stable proteins. *Protein Eng 6*:793–800.

Singh J, Thornton JM. 1992. *Atlas of protein side-chain interactions*. New York: IRL Press.

Surles MC. 1992. An algorithm with linear complexity for interactive, physically-based modeling of large proteins. *Computer Graphics 2*:221–230.

Tanaka T, Kimura H, Hayashi M, Fujiyoshi Y, Fukuhara K, Nakamura H. 1994. Characteristics of a de novo designed protein. *Protein Sci 3*:419–427.

Vijaykumar S, Bugg CE, Cook WJ. 1987. Structure of ubiquitin refined at 1.8 Å resolution. *J Mol Biol 194*:531–544.

Yue K, Dill KA. 1992. Inverse protein folding problem: Designing polymer sequences. *Proc Natl Acad Sci USA 89*:4163–4167.

Yue K, Fiebig KM, Thomas PD, Chan HS, Shakhnovich EI, Dill KA. 1994. A test of lattice protein folding algorithms. *Proc Natl Acad Sci USA*. Forthcoming.