Methods and Algorithms for the Design of Proteins

STEVEN E. BRENNER

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, U.K.
*Department of Biochemistry, University of Cambridge, Cambridge CB2 1QW, U.K.*

Abstract. Protein design promises immense rewards for science, medicine, and technology. However, to date, virtually all approaches to the problem have been qualitative and largely intuitive. This paper presents a quantitative methodology for the de novo design of proteins as well as supporting algorithms, including a parallel simulated annealing optimization protocol. With these algorithms it is possible to choose sequences that should adopt any plausible structure. A prototype design system has succeeded in selecting sequences that appear correct to a variety of analyses.

1. Introduction

Proteins are linear pseudo-polymers of amino acids that fold into unique three-dimensional structures whose shapes depend entirely upon the linear sequence. Though researchers have been attempting to understand the forces governing this folding process for decades, the problem remains largely unsolved: we can not determine the conformation a protein will adopt from knowledge of its sequence. Within the past few years, however, it has been recognized that the reverse problem—finding a sequence which will fold into a desired structure—may be considerably simpler.

The recent explosion in the techniques of molecular biology has provided the practical tools needed to construct novel proteins, and studies of the three-dimensional structures of proteins over the last few decades have provided many of the basic principles which are important in maintaining these structures. The combination of these two areas of research with the goal of designing new amino acid sequences that will adopt a desired structure is now one of the outstanding aims of modern biochemistry.

The benefits of an ability to design "custom-made" enzymes are incredible. For example, while most conclusions about the fundamental elements of protein structure have been from the observation of natural proteins, evolution has probably constrained natural sequences and structures to a subset of the full potential range [1, 2]. Separating chance from necessity has been one of the principal challenges of modern protein structure studies. Without manipulating protein sequences directly, it is impossible to ask questions such as whether alternative packings between secondary structure elements are forbidden by nature or simply inaccessible to evolution, and what general constraints exist on protein sequences in a given fold [3].
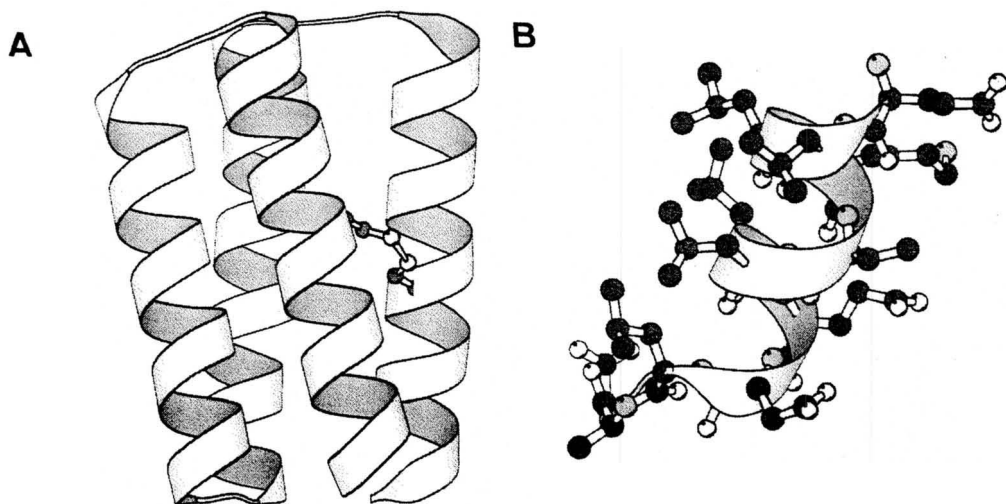
Figure 1. Models of two designed four-helix proteins: Felix and Alpha-1. *A.* Felix was designed to adopt a four-helix bundle with a disulfide linkage. The structure was drawn from the model coordinates in PDB entry 1flx. *B.* The structure of Alpha-1 was confirmed by crystallography (PDB entry 1al1), but these helices aggregated to form a hexamer instead of the desired tetramer. These figures and others in the text were drawn with molscript [67].

In addition to fundamental scientific understanding, there are several other reasons for designing proteins. Perhaps the most obvious utility of this work is the construction of new catalysts, since proteins may be capable of carrying out reactions presently requiring dangerous or expensive reagents at high concentrations, temperatures, or pressures. Of similar economic importance is the ability to create "molecular toolkits" of custom-made proteins, designed to fulfill a variety of complementary tasks. Visionaries imagine using these individually in nanotechnology or in coordinated groups for "molecular computing" [4]. Protein design also has implications for medicine in roles such as antigen display [5, 6] and degradation-resistant protein mimetics. Most importantly, a better understanding of protein structure will aid our ability to comprehend and subvert the molecular mechanisms of illness and disease [7].

2. Protein Design

2.1. Design by Modeling

It is unsurprising, therefore, that many groups have attempted to build new proteins. One of the earliest de novo protein design attempts was to build a beta-bell protein with a structure similar to catalase domain 2 [8]. After selecting a suitable structure and making a "template" upon which to place a sequence, the sequence itself was selected. The first consideration in the beta-sheets was the hydrophobic-hydrophilic character of single positions. At the turns, statistics about residue frequencies in specific positions [9] were manually, qualitatively applied. Similar criteria [10-13] were then applied to find residues which most commonly form beta-sheets. Compromises were made to

provide sequence diversity and to facilitate packing within the sheet, and further trade-offs accommodated three dimensional "pairing preferences." With a complete sequence in hand, the structure was visually examined for inter-sheet packing, and some further modifications made. The final structure was then subjected to various tests ranging from energy minimization to manually shaking a CPK model. While initial attempts to build the protein failed, iterative modification of the sequence has led to the construction of proteins with some characteristics of the desired structure.

This design approach has been concisely summarized [14], and Figure 2 diagrammatically shows the procedure. Note that the most important step, choosing an appropriate amino acid sequence, is performed manually using "physical, statistical, and intuitive criteria" [15]. This clearly described and relatively straightforward, iterative method is what I term *design by modeling*, because the principal task is ensuring that the designed sequence could reasonably fit into the desired structure.

This general protocol has been employed by a large number of different groups to seek alternative amino acid sequences for a variety of well-studied folds, Perhaps the most commonly attempted structure is the four-helix bundle [16-20], two example of which are shown in Figure 2. Another popular structure is the alpha-beta barrel [21, 22]. In addition, a variety of other natural-like structures have also been designed [14, 23]. In a similar manner, sequences have also been selected for folds not found naturally, such as a miniature antibody-type structure [6], and an "open sandwich," consisting of a four-stranded antiparallel beta-sheet with one side screened by two alpha-helices [24].

Unfortunately, with one notable exception [22], none of these medium-size proteins has been shown to adopt the structure desired. Perhaps because of this, much current research on de novo design uses simpler systems to address more direct questions. One such approach is to build only single units of secondary structure rather than linking them to form a full protein.

For example, one group has attempted to build a four-helix bundle out of four aggregating alpha-helices called Alpha-1 [18]. This approach permitted the inclusion of sequences with extremely high helix-forming tendencies because there was no need to "break" the helix to form turns. A crystal structure showed a curious result: while the sequence had formed helices and dimerized, the dimers trimerized to form a hexamer with novel structure. A 29 residue peptide designed to dimerize into a double-stranded parallel coiled-coil found a similar fate: x-ray crystallography revealed that a triple-stranded antiparallel coiled coil was formed instead [25].
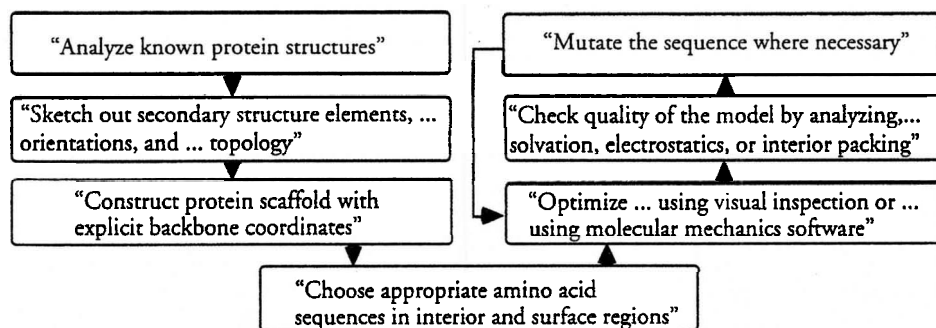

Figure 2. Design by modeling: A flow chart of the protein design methodology described by Sander [14]. Variants of this procedure have been used in nearly all protein design projects.

Nonetheless, these simpler approaches hold great promise, for the partial failures described above were instructive. A four-helix bundle based on Alpha-1 now shows many characteristics of native proteins [26], and a similar bundle made of four helical peptides has been shown to adopt the correct structure [20]. However, perhaps the greatest success of design by modeling has been the construction of a small protein which forms two alpha-helices linked with a disulfide bond (Figure 3). NMR studies have revealed that while the ends of the helices are somewhat frayed, the protein structure is much as desired [27]. This experiment clearly aside doubt about the possibility of protein design: in a small way, it has been successfully accomplished.

2.2. Abiotic Protein Design

To surmount the relatively small preferences of natural amino acids for forming one type of secondary structure over another, many groups have begun experimenting with non-protein amino acids. To this end, alpha-aminoisobutyric acid (Aib) and alpha-beta-dehydrophenylalanine have been demonstrated to be powerful 3_{10}- and alpha-helix initiators [28, 29], while 4-(2-aminoethyl)-6-dibenzofuranpropionic acid assists in the nucleation of beta-sheet structure [30], and the planar structure of alpha, beta-unsaturated amino acids has been found to help introduce type two beta-turns [31]. Another non-protein amino acid, (S)-a-amino-(2,2'-bipyridine)-6-propanoic acid has been used as a highly ion-specific metal ligand [32]. One of the interesting results of this work with non-protein amino acids was the discovery that even strong helix-breaking sequences such as Gly-Pro cannot terminate a helix nucleated by Aib [28]. This suggests
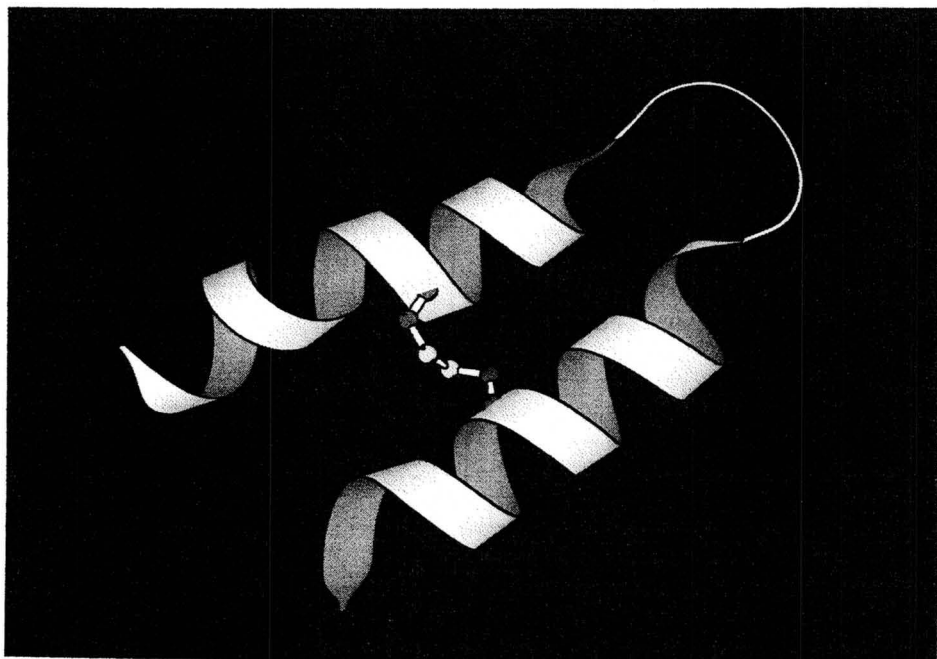
Figure 3. NMR structure of ALIN. The structure of this antiparallel two-helix designed protein with a disulfide bridge was confirmed by 2D-NMR spectroscopy [27]. Coordinates for the structure were provided by Y. Kuroda.

that an even balance needs to be kept between secondary structure initiating and terminating sequences. This, in turn, provides a rationale for the relative instability of proteins [33].

Using different abiotic means, several groups have taken a more expedient approach to building new medium-size proteins. For example, a membrane-channel four-helix bundle called Tetraphilin is held together using an o-amidophenyl tetraphenylporphyrin ring [34]. In order to be assured of forming helices, this new protein also incorporates an Aib residue in each helix. "Template-Assisted Synthetic Proteins" employ a slightly different tack to force the association of otherwise independent elements of secondary structure: the peptide chains are connected to the lysines of a circularized peptide [35]. When forced to associate with other helices on a template, peptides with no defined structure when free in solution gain high helix content [36].

Metal ions provide another way to bind independent elements of secondary structure together. For example, a Ru(II) ion was coordinated to four 15-residue alpha-helices via a linker to form a remarkably stable metalloprotein. A similar system making use of a Fe(II) ion to assemble a three-helix bundle linked by a bipyridine moiety [37] is almost certainly the most elegant experiment in de novo design to date. Because of the asymmetry of the linker, four stereoisomers of the molecule (due to different coordination of the linker to the Fe(II)) are possible. Reverse phase HPLC separates the isomers, making it possible to see if any forms were favored. For some sample helices, there was no preference between different forms, suggesting a "molten globule" interior. Most design experiments require sophisticated and difficult methods to elucidate the protein's structure—and these succeed only when the protein fold is somewhat stable and unique. However, this approach provides easy access to information about the protein's structure, allowing researchers to quickly determine whether they have succeeded in producing specific inter-helix interactions.

2.3. Generalized Approaches

While the various design efforts described above have been scientifically valuable, 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. In particular, design by modeling is fundamentally irreproducible. Therefore, some investigators have created more generalized procedures to explore protein design.

Experimentalists have produced fascinating results by making large numbers of random sequences. For example, when proteins were made of 80 to 100 glutamine, leucine, and arginine residues in random sequence, five percent are soluble and resistant to intracellular degradation [38]. By contrast, when a four-helix bundle is made with designed turns, but random hydrophobic sequence on the interior and hydrophilic on the exterior, some 60 percent of the proteins survive in the cell [39]. While this result does not show that the desired structure was formed frequently, if at all, it does clearly demonstrate the importance of the hydrophobic effect in protein structure.

One of the earliest theoretical studies in protein design worked on a different principle, assuming that residue packing is the dominant force in protein folding. To this end, Ponder and Richards developed an algorithm to determine which residues could pack to form a particular structure [40]. Thus, it designs sequences whose sidechain rotamers are sterically compatible with a given fold. However, this criterion
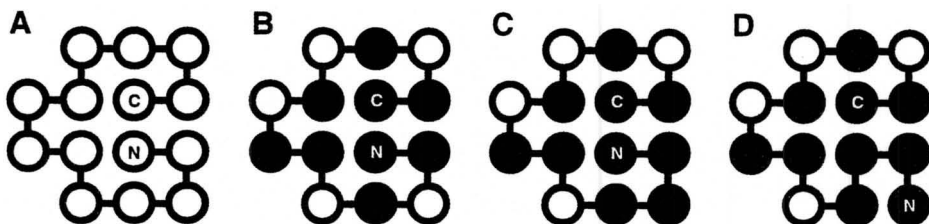
Figure 4. HP (Hydrophobic/Polar) model of protein structure. In this model, the structure quality is simply the number of hydrophobic (dark) residues neighboring other hydrophobics; polar (white) residues have no interactions. If the structure to be designed is *A*, then *B* and *C* are two optimal-quality sequences for the fold. However, the variation at position three means that the sequence shown in *C* can adopt sequence *D* equally well. The heuristics developed in [42] would select *B* over *C*, even though the two sequences are of equal quality, because *B* incorporates "negative design" for structure *D*. (Adapted from [42], with permission.)

seems to become weak and the algorithm too computationally difficult when normal variation [41] is permitted within the core of the protein.

Consequently, other groups have attacked the protein design problem by making different assumptions and considerable simplifications. One design system describes proteins as strings of hydrophobic and polar residues that lie on a lattice [42]. In this model, hydrophobic interactions are the only force driving protein folding; thus, every hydrophobic residue touching another hydrophobic residue stabilizes the protein, while all other interactions are ignored. The protein design problem is then defined not as finding the most stable sequence for a given fold, but finding the sequence that, when folded into the desired structure, has a lower energy than when folded into any other structure (Figure 4).

Because this model is uncomplicated, it elegantly contains none of the arbitrary parameters and compromises that plague all the other design methods. More important, it is defined rigorously enough to be amenable to mathematical analysis. To this end, a number of heuristics have been developed for rapidly distinguishing structures likely to fold uniquely into the desired structure.

This approach suggests another reason why protein structures have such poor stabilities: rather than having the lowest-energy sequence, a given structure has a sequence that has destabilized itself in ways which destabilize other structures even more. One can imagine evolution blindly playing a game of brinkmanship, making a protein sequence increasingly unstable for the desired structure, in the hope of making undesirable conformations have a very high free energy.

If this is indeed the case, and if secondary structure lies in a delicate balance between nucleation and termination, then protein design is a task requiring even more subtle weighting and compromise than would be suggested by the experience of design by modeling. To attempt to optimize a function of such a huge number of parameters with any degree of sophistication, a quantitative method is necessary. In addition, reproducibility of the procedure when applied to different structures is necessary to ensure that it encapsulates genuine general and fundamental information about protein structure. For this reason, it is necessary to make use of a fully qualified methodology to select new protein sequences.

3. A Quantitative Methodology for Design of Proteins

3.1. Methodology

In order to explore the field of protein structure, I have developed a design methodology that attempts to combine the comprehensibility and relative success of design-by-modeling experiments with the scientific rigor of the generalized methodologies [43, 44]. At its core, this extensible protocol uses quantitative statistics about all known protein structures. In addition, the methodology employs a theoretical model to incorporate anti-design (as in Figure 4) and various "hints" such as specific features of particular folds and requirements for functionality. These parameters are combined as a weighted sum which measures the quality of a particular sequence for a particular structure. To design a protein, this quality evaluation function is optimized for a given structure.

As the evaluation function incorporates information about all natural protein structures, the procedure is general, automated, and reproducible. Any fold to be designed can be fed into the design system, and optimization of the evaluation function will generate a sequence that should adopt that structure.

The decision to embrace statistics for the evaluation function relies upon the following assumptions: 1) It is possible to derive statistics about residues that are capable of incorporating sufficient thermodynamic information to form correctly folded proteins and 2) Thermodynamics will be completely dominant over kinetics in the formation of the protein structures; i.e., folding intermediates need not be considered.

Unfortunately, there is no conclusive proof that either of these assertions is true, and a formidable amount of evidence mitigates their veracity. Indeed, the success or failure of the design system is a partial test of these claims. However, the success of the statistically-based quantitative design system may not be predicated on these assumptions being entirely true, and it seems that both assumptions have some merit. In particular, there is much evidence that statistics can reveal and describe significant features of protein structure. Not only are many statistics intuitively comprehensible (e.g., the high fraction of prolines in turns), but statistical methods seem to be capable of distinguishing some correct structures from incorrect ones and even matching sequences with folds [45-48]. Additionally, while it is clear that, in general, the folding process cannot be entirely ignored (as shown by the existence of proteins that can not

Category	Parameters	Category	Parameters
Position	Secondary Structure Solvent Accessibility "Cap" Positions Torsion Angles	Neighbor	Primary Structure Neighbors Secondary Structure Neighbors Tertiary Structure Neighbors Total Surface Area Neighbors Size
Uniqueness	Theoretical Models	Hints	Motifs from the Literature Functionality Ease of Synthesis Suitability for Characterization

Table 1. Several parameters for the quantitative design methodology.

be reversibly denatured), it seems unlikely that any particular folding intermediates are absolutely necessary for the small proteins currently being designed.

The statistics are obtained by scanning the protein databank (PDB) [49, 50] and ascertaining various characteristics of individual residues. The method includes virtually all variables which could have effects on structure and which can be measured and fruitfully applied to protein design. It is probable that some of these features are of marginal importance. However, because the methodology provides the ability to specify explicitly the importance of one feature relative to another, rules derived from less important parameters can be assigned lower weightings.

All of these statistical parameters are divided into two categories for the computational reasons described in §3.3. *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. The complementary category, *neighbor preferences*, describes the likelihood of a given residue having specific atoms, residues, or structures nearby. Examples of these parameters are shown in Table 1.

However, 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" [51], it is necessary to incorporate "anti-design" for other potential structures. For this reason, the heuristics developed by Yue & Dill [42] (see §2.3 and Figure 4) are included to provide *uniqueness* of structure.

Since it is possible that particular structures have features which are not general, it also necessary to provide various *hints* to the design system. In addition to including these large motifs, this category can also provide features related to functionality or information to aid synthesis and characterization of the designed protein.

3.2. Protein Structure Database Weighting

For a statistically-based design system to succeed, the selection of parameters measured is of paramount importance. However, in order for these parameters to be meaningful, it is also imperative that the statistics be collected in a carefully selected manner, taking account of the data-set from which they derive. This is particularly important when dealing with protein structure information, because the PDB is immensely biased towards certain structures and contains data of variable quality. To surmount this problem, many researchers simply select a small number of "model" structures [52]. However, recently a new, quantitative method for selecting a unique subset of the PDB of high quality has been developed; in effect, it selects a single structure of high quality from each cluster of homologous proteins in the database [53].

Even this, though a vast improvement, is not entirely satisfactory for statistical use of the PDB. By selecting only a single representative of each fold, the overwhelming majority of the entries in the database—and the data they contain—is excluded from consideration. For this reason, I have developed a protein structure weighting algorithm that removes bias while ensuring arbitrarily high average structure quality and retaining nearly all of the information in the database [43, 54]. Under this scheme, better quality structures or those that are unique in the database garner high weightings, while poor or over-represented ones receive low weightings. If natural proteins are assumed to be evenly distributed and the center of gravity of known protein structure classes is assumed to correspond with that of natural proteins, then these weightings can then be applied

to statistical data from the PDB to deduce "global" features of proteins.

Figure 5 diagrammatically shows how the algorithm would operate on two proteins with some overlapping content. If the two proteins shown in Figure 5A were both to be used in some statistical analysis, a bias would result from the overlapping character of the two proteins. To compensate for this by weighting, the similar portions of the two proteins are first notionally separated from those regions which are unique, as shown in Figure 5B. (The weighting protocol does not actually separate portions of sequence or structure and weight them differently. Rather, it assigns the whole protein the weighted average of the scores computed for overlapping and unique sections of the protein.) The similar regions of the two proteins are then individually considered (5C) and assigned a weighting according to their quality, such that the sum of the two weightings is unity (the weighting of a unique protein). Quality is determined by considering 1) how fine the experimental data are (the resolution), 2) how well the protein structure model fits the data (the R-value), and 3) how well the structure resembles "reality" (the bond angles and other stereochemical checks). In Figure 5D, the protein on the left is of about half the quality of the darker one on the right. If the weighted portions of protein are reassembled, as in Figure 5E, it is clear that there is now a uniform weighting of different structure types, where redundant information has been weighted according to quality.
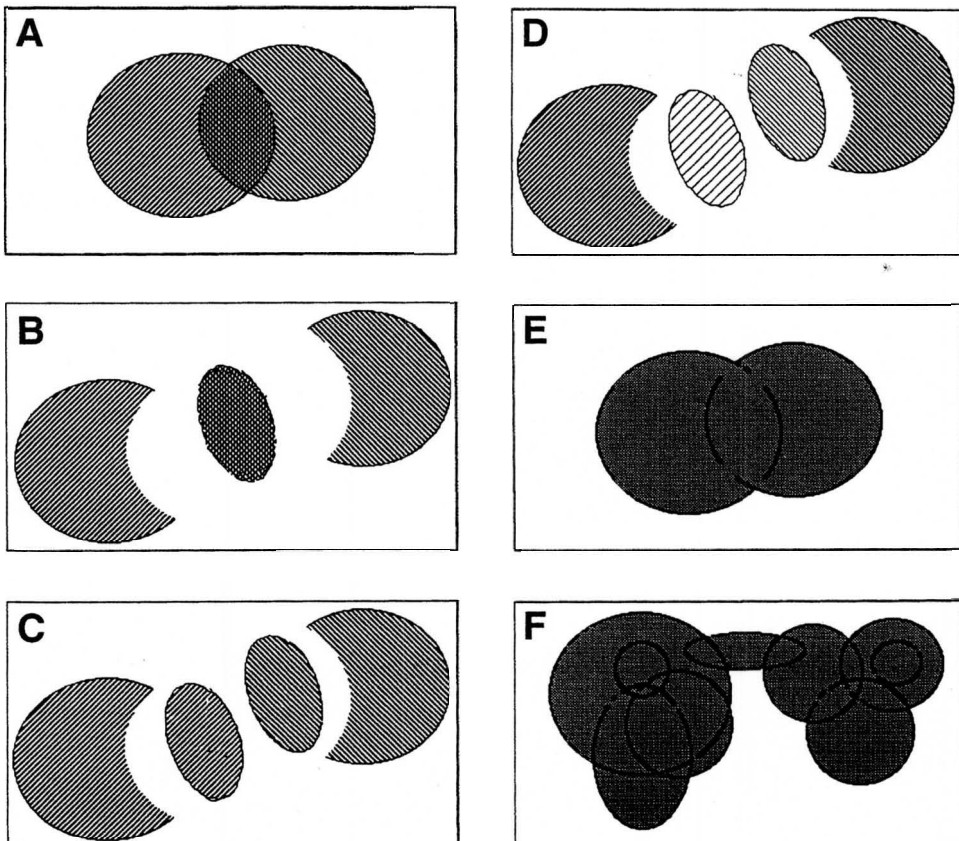

Figure 5. A protein structure weighting scheme. See §3.2 for explanation; reproduced from [43, 44].

Figure 5F shows how a database might "appear" after being weighted according to this method.

3.3. Ternary Division Sum Optimization

Once parameters have been selected for inclusion in the design system and the relevant statistics collated, they can be used for automatically designing protein sequences. This is accomplished by optimizing a function over sequences which assigns qualities for a given structure.

It is trivial to optimize a function of the *position preferences*; at each position, simply select the residue that has the highest statistical quality. Because every site is independent, the complexity of such an algorithm would be merely $O(n)$. However, the *neighbor preferences* mean that each residue is not independent. If it were necessary to consider all potential interactions between different positions, the optimization problem would be of exponential complexity. However, if only immediate neighbors are considered, the problem becomes markedly simpler: a high-order polynomial algorithm exists to optimize position and neighbor preferences in one dimension (Steven P. Ketchpel, personal communication). Ternary division sum optimization (TDSO), a variant on Ketchpel's algorithm which is considerably more efficient, can be applied to protein design.

The algorithm which works on the principle that the best sequence is the one with the best beginning plus middle plus end. As shown in Figure 6, TDSO first appends left and right "end-caps" to the sequence. Then it computes the position preference for each of the 20 types of residue at the central location. It then adds, to each of these, the recursively computed qualities of the left and right sides using the current central residue as the right and left end-cap, respectively, for the sides. Recursion is ceased when no
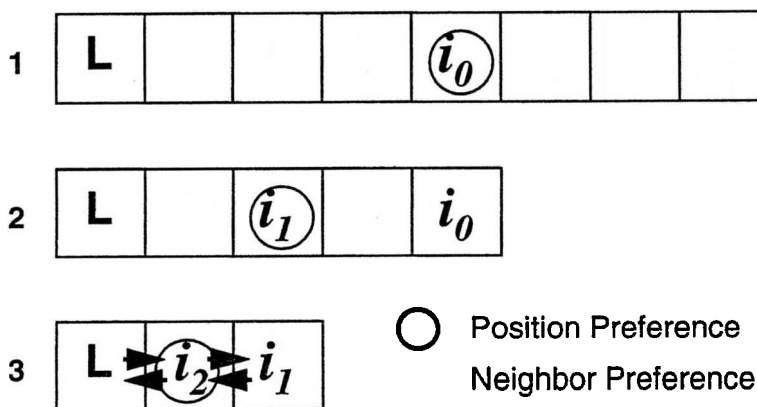
Figure 6. Schematic representation of TDSO algorithm. Line *1* shows a model sequence of length seven, with left and right end-caps appended. The optimal sequence is that which has the best left, middle, and right portions. The middle is simply the position preference of a particular residue i_0 in the central position. The optimal left and right halves for that residue are computed recursively using the current middle residue as an end-cap. Line *2* shows the left side, and line *3* shows the left side of that sub-sequence. Recursion ceases when, as in line *3*, the left and right sides consist only of end-caps; at this point all half-neighbor preferences are summed with the position preference.

central position can be selected, and then several half-neighbor preferences are computed. (A half-neighbor interaction is the preference of the residue at one position for its neighbor on one side. In total, there are four half-neighbor interactions at each position: the preferences of the residue at that position for each neighbor, and the preferences of both neighbors for it.)

When TDSO is extended beyond one dimension to an arbitrary graph, it is impossible to pick a center pivot position which will divide the graph into two disjoint ones. Instead, it is necessary to find a set of nodes which together partition the graph into two subgraphs. However, instead of needing to test 20 residues at the center position at each recursive level, it is necessary to test all combinations of residues at all the positions. If there are m border nodes, then 20^m different partitions must be tested and must then summed with the best "left" and "right" subgraphs to find the best set of residues for the entire graph. As expected, the complexity of the algorithm increases as the connectivity increases.

Although the running time of this algorithm grows dramatically with sequence length, it may still be tractable if various approximations are applied, including alpha/beta cutoffs [55, 56] and heuristics that reduce the number of choices at each position from 20 residues to just a few classes (e.g., hydrophobic, polar, and neutral). The algorithm is simple to parallelize, and it should be able to achieve near-ideal speedup on massively parallel computers because it has virtually no communications overhead.

3.4. Delayed-Update Parallel Simulated Annealing

While (nearly) every natural protein adopts only a single structure, any given natural structure can be formed by several different sequences. Consequently, it is unnecessary to select the best possible sequence as computed by the design methodology; near-optimal sequences will provide satisfactory results. Indeed, since the margins of error in the various rules which contribute to the overall methodology are considerable, the best sequence computed by the statistics is unlikely to optimize the underlying parameters that the statistics attempt to describe. Moreover, as sequences similar to the optimal one will be close to it in energy, stochastic sampling and optimization methods suggest themselves as time-efficient methods of designing new proteins.

It is possible to apply a straightforward simulated annealing protocol [57, 58] to the protein design problem by setting the elements as follows:

Configuration: The protein is a sequence of residues 1...n, each of which can be any of the 20 principal biological amino acid residues.

Rearrangements: A residue at a particular position may be mutated to any other randomly selected residue.

Objective function: This is the quality measure described in §3.1.

Annealing schedule: While the precise schedule must vary with the particular parameters included in the objective function, it has been found useful to make rearrangements in an ordered way, by iterating over the sequence until specified number of mutations in the sequences occur, at which point the temperature is dropped.

While this method can produce results very quickly when using easily computable quality measures on small proteins, it becomes inconvenient when working with large proteins and complex objective functions. Even in these complex instances, the
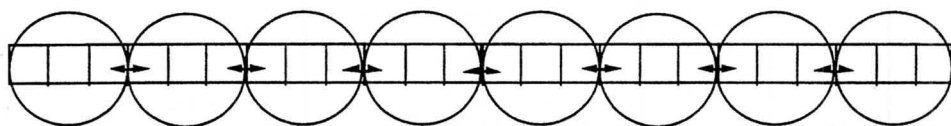
Figure 7. Schematic representation of DUPSA data distribution. The boxes represent positions in a sequence to be selected, and each circle represents a compute node containing a small number of these positions. Because the optimization requires neighbor information, communication (arrows) is required between nodes holding neighbor residues.

computational design procedure is much faster than actually synthesizing the designed protein. However, to develop, test, and improve the design methodology it is necessary to generate large numbers of sample sequences. Therefore, methods to increase the speed of designing new sequences are of considerable practical utility.

Because the objective function is principally the sum of the qualities of each of the individual positions (the *uniqueness* and *hints* parameters may abrogate this rule and need to be dealt with specially), it is not difficult to parallelize the simulated annealing algorithm. To do so, the sequence is divided evenly across the compute nodes of the parallel computer and each node finds the best amino acids for the region of sequence assigned to it. Because of the neighbor preferences, every node needs to keep a record of the current neighbors of all of the positions which it is trying to optimize. Inter-node communication is used to provide this information as the simulated annealing progresses. Figure 7 diagrammatically represents the division of sequence over the compute nodes.

Synchronous communication must be used if this algorithm is to be isomorphic with the serial optimization algorithm. That is, each node must wait until it has received messages containing information about the current neighbors before proceeding with the annealing. Coded this way, the parallel algorithm does provide significant speed enhancement over the serial in many cases. However, as the number of neighbors increases, the communication overhead grows dramatically.

Asynchronous message passing provides a partial solution to this problem: rather than waiting to receive information about changed neighbors, each compute node optimizes its sequence with the most recently received information. This means that there is a potentially problematic delay between when a neighbor residue is changed and when this change is registered. However, for the design systems tested, this gap during which out-of-date neighbor information is used had no detrimental effect on the final sequences. Thus, this Delayed-Update Parallel Simulated Annealing (DUPSA) protocol selects sequences as well as the serial simulated annealing protocol in a fraction of the time.

4. A Model System

4.1. Construction

A complex quantitative design system contains a multitude of parameters, making it difficult to understand directly how particular characteristics have interacted to produce the resultant sequence. Therefore, a comparatively simple model protein design system was created to gain some understanding about the overall feasibility and promise of

quantitative protein design. In addition, one of the most difficult tasks in building the design system is determining appropriate weightings for each of the various parameters (e.g., how important is secondary structure preference relative to solvent accessibility?). A basic system which contained only the most important parameters facilitated setting the weightings on these parameters to approximately correct values, which will be altered only slightly when more abstruse criteria from the methodology are added to fine-tune the output.

The central components of the model system were secondary structure preference and solvent accessibility, both of which are statistical position preferences (see §3.1). Primary structure neighbor preference was also incorporated mainly as a test of the optimization procedures' ability to operate on neighbor parameters. In addition, it readily became apparent that a diversity hint, which ensures that the composition of the sequences models that of natural sequences, was necessary because of the particular way in which solvent accessibility is measured. (For a detailed description of the measurement of these parameters, and the details of both optimization procedure and analysis, see [43, 44].)
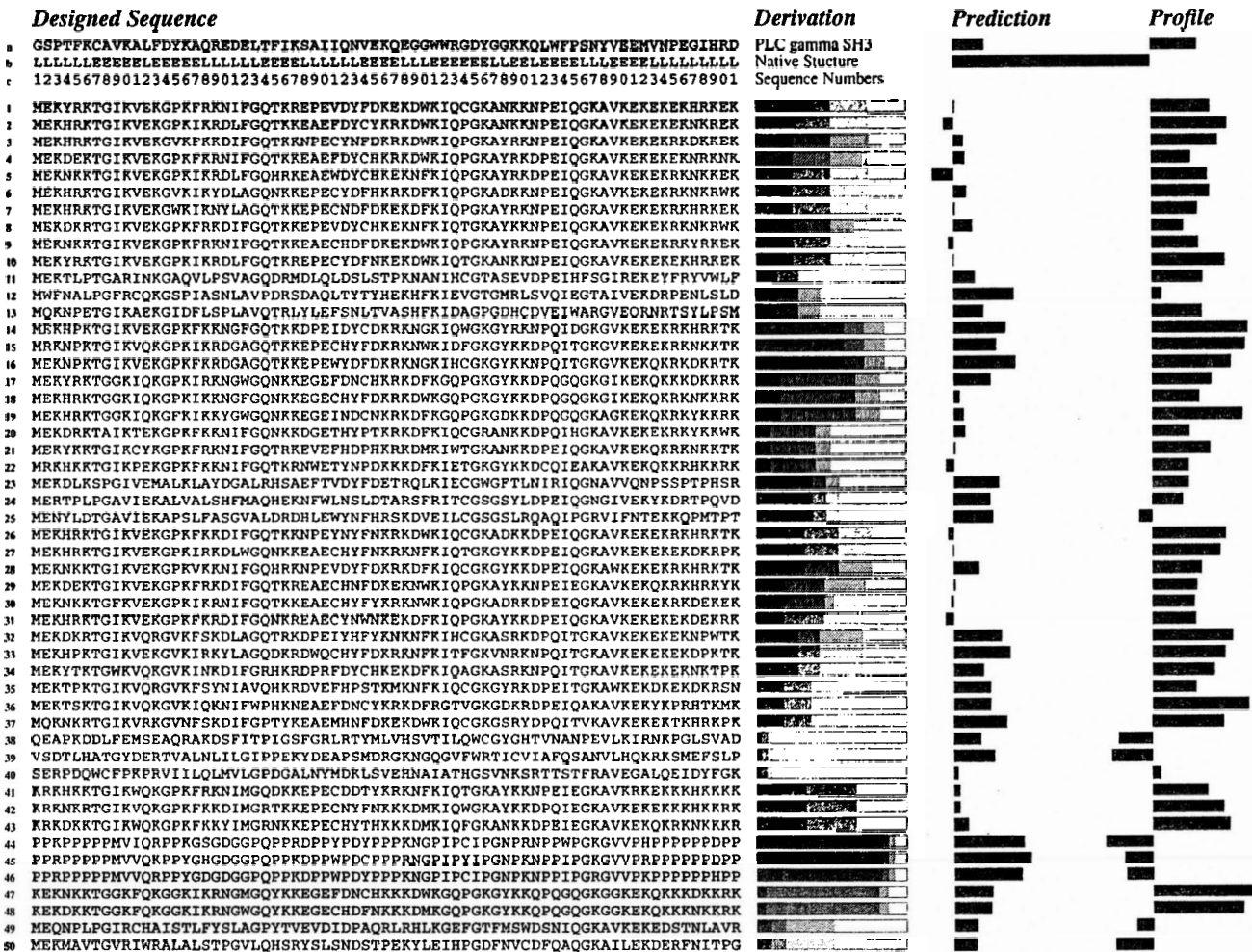
The design system accepts detailed information about the protein to be designed, and then applies the parameters listed above to find the optimal sequence for that structure. For example, alanine would receive a good position preference score at a hydrophobic position in an alpha-helix, while lysine would score well in a hydrophilic helical position, and proline in a turn. Optimizing all the parameters over the entire structure results in the selection of a single protein sequence.

The design protocol has been applied to a variety of different proteins including phage 434 cro [43, 44], myohemerythrin [59], myoglobin, hemoglobin, lactate dehydrogenase, and ubiquitin. Here, I report sequences designed by the model system to adopt the structure of two different SH3 domains.

4.2 Designed Sequences for the SH3 Fold

SH3 domains derive their name from "src homology," and characteristically bind small proline-rich peptides (for a review, see [60]). Typically, they consist of two beta-sheets which lie against each other to form a barrel-like structure. Generally, the sheets are composed of five to seven strands, and frequently the strands are not linear, due either to bulges or turns. In addition, some SH3 domains contain one or more small helices. In the case study here, sequences were designed to fold into the structure of the SH3 domains of human phospholipase C gamma (PDB entry 1hsp) [61] and chicken brain alpha spectrin (PDB entry 1shg) [62]. These sequences are 71 and 57 residues long, respectively, although the termini of 1hsp are from the expression vector and are without experimental restraints.

A wide variety of different weightings of the four constituent parameters were used in designing the sequences, which are shown in Figures 8 and 9. The designed sequences looked very similar to each other, though they have no significant identity ($p < 0.001$ for blastp [63] with the seg filter and match matrix on the May 1994 nr database) to either the natural sequences nor to any other protein. Intriguingly, even residues at positions strongly conserved between different SH3 domains [62] were not retained by the designed sequences. Notwithstanding this, most sequences appeared fairly normal, although there was an apparent excess of certain residues (such as lysine) and a strange

S.E. Brenner/Design of Proteins

Designed Sequence   Derivation   Prediction   Profile

```
a  GSPTFKCAVRALFDYKAQREDELTFIKSAIIQNVEKQEGGWWRGDYGGKKQLWFPSNYVEEMVNPEGIHRD   PLC gamma SH3
b  LLLLLLEEEEELEEEEEELLLLLLEEEELLLLLLLEEEELLLEEEEEELLEELEEEELLLEEEELLLLLLLLL  Native Stucture
c  12345678901234567890123456789012345678901234567890123456789012345678901   Sequence Numbers

 1  MEKYRKTGIKVEKGPKFRKNIFGQTRREPEVDYFDKEKDWKIQPGKANKRNPEIQGKAVKEKEKEHRKEK
 2  MEKHRKTGIKVEKGPKIKRDLFGQTKKEAEFDYCYKRKDWKIQPGKANKRNPEIQGKAVKEKEKERNKREK
 3  MEKHRKTGIKVEKGVKFKRDIFGQTKKNPECYNFDKRKDWKIQPGKAYRRNPEIQGKAVKEKEKRRKDKKEK
 4  MEKDERTGIKVEKGPKFRKNIFGQTKKEAEFDYCHKRKDWKIQPGKAYRKDPEIQGKAVKEKEKEKNRKNK
 5  MEKNRKTGIKVEKGPKIKRDLFGQHRKEAEWDYCHKEKNFKIQPGKAYRKDPEIQGKAVKEKEKERNKKEK
 6  MEKHRRTGIKVEKGVKIKYDLAGQNKKEPECYDFHKRKDFKIQPGKADKRNPEIQGKAVKEKEKRRKNKRWK
 7  MEKHRTGIKVEKGWKIKNYLAGQTKKEPECNDFDKEKDFKIQPGKAYRKNPEIQGKAVKEKEKRRHKRNK
 8  MEKDKRTGIKVEKGPKFRKDIFGQTKREPEVDYCHKEKNFKIQTGKAYKRNPEIQGKAVKEKEKRNKRWK
 9  MERNKRTGIKVEKGPKFRKNIFGQTKKEAECHDFDKEKDWKIQPGKAYRRNPEIQGKAVKEKERRRYRKEK
10  MEKYRKTGIKVEKGPKIKRDLFGQTKREPECYDFNKEKDWKIQTGKANKRNPEIQGKAVKEKEKEKHRKEK
11  MEKTLPTGARINKGAQVLPSVAGQDRMDLQLDSLSTPKNANIHCGTASEVDPEIHFSGIREKEYFRYVWLF
12  MWFNALPGFRCQKGSPIASNLAVPDRSDAQLTYTYHEKHFKIEVGTGMRLSVQIEGTAIVEKDRPENLSLD
13  MQKNPETGIKAEKGIDFLSPLAVQTRLYLEFSNLTVASHFKIDAGPGDHCDVEIWARGVEORNRTSYLPSM
14  MEKHPKTGIKVEKGPKFKKNGFGQTKKDPEIDYCDKRRKNGKIQWGKGYRRNPQIDGKGVKEKEKRKHRKTK
15  MRRNPKTGIKVQKGPKIKRDGAGQTKKEPECHYFDKRKNWKIDFGKYRKDPQITGKGVKEKEKRKHKKTK
16  MEKNPKTGIKVEKGPKFKRRDGAGQTKKEPEWYDFDKRKNGKIHCGKGYKKNPQITGKGVKEKQKRKDKRTK
17  MEKYRKTGGKIQKGPKIRRNGWGQNKKEGEFDNCHKRRKDFKGQPGKGYRKDPQGQGKGIKEKQKKKDKKRK
18  MEKHRKTGGKIQKGPKIKRKNGFGQNKKEGECHYFDKRKDWKGQPGKGYRKDPQGQGKGIKEKQKRKNKKRK
19  MEKHRRTGGKIQKGFKIKKYGWGQNKKEGEINDCNKRRKDFKGQPGKGDKKDPQGQGKAGKEKQKRYKKRR
20  MEKDRKTAIKTEKGPKFRKNIFGQNKKDGETHYPTKRRKDFKIQCGWGFTLNIRIQGNAVVQNPSSPTPHSR
21  MEKYRKTGIKCYKGPKFRKNIFGQTKREVEFHDPHKRRDMKIWTGKANKRDPEIQGKAVKEKQKRNKKTK
22  MRRHKKTGIKPEKGPKFRKNIFGQTKRNWETYNPDKKKDFKIETGKGYKKDCQIEAKAVKEKQKRHKKRK
23  MEKDLKSPGIVEMALKLAYDGALRHSAEFTVDYFDETRQLRIECGWGFTLNIRIQGNAVVQNPSSPTPHSR
24  MERTPLPGAVIEKALVALSHFMAQHEKNFWLNSLDTARSFRITCGSGSYLDPEIQGNGIVERYKDRTPQVD
25  MENYLDTGAVIEKAPSLFASGVALDRDHLEWYNFHRSKDVEILCGSGSLRQAQIPGRVIFNTEKKQPMTPT
26  MEKHRKTGIKVEKGPKFKKNPEYNYFNKRRKDWKIQPGKADKRDPEIQGKAVKEKERRHKRTK
27  MEKHRRTGIKVEKGPKIKRDLWGQNKKEAECHYFNKRKNFKIQTGKGYKKDPEIQGKAVKEKEKEKDKRPK
28  MEKNKKTGIKVEKGPKVKKNIFGQHRKNPEVDYFDKRKDFKIQCGKGYKKDPQGKAWKEKEKRHRKTK
29  MEKDERTGIKVEKGPKFRKDIFGQTKREAECHNFDKEKNWKIQPGKAYKKNPEIEGKAVKEKQKRKHRRYK
30  MEKNKKTGFKVEKGPKFRKNIFGQTKKEAECHYFYKRKNWKIQPGKADKRDPEIQGKAVKEKEKRRKDEKER
31  MEKHRRTGIKVEKGPKFRKNREAECYNWNKEKDFKIQPGKAYKKDPEIQGKAVKEKEKEKDEKRK
32  MEKDKRTGIKVQRGVKFSKDLAGQTKRDPEIYHFYKNKNFKIHCGKASRKDPQITGKAVKEKEKEKNPWTK
33  MEKHPKTGIKVEKGVKIKRKYLAGQDKRDWQCHYFDKRRNFKITFGKVNRKNPQITGKAVKEKEKEKDPKTK
34  MEKYTKTGWKVQKGVKINKDIFGRHKKDPRFDYCHKEKDFKIQGKASRKNPQITGKAVKEKEKERKTPKR
35  MEKTPKTGIKVQRGVKFSYNIAVQHKRDVEFHPSTKMKNFKIQCGKGYRKDPEITGKAWKEKDKEKDRSN
36  MEKTSKTGIKVQKGVKIQKNIFWPHKNEAEFDNCYKRKDFRGTVGKGDKRDPEIQAKAVKEKYKPRHTKMK
37  MQRNKRTGIKVRKGVNFSKDIFGPTYKEAEMHNFDKEKDWKIQCGKGSRYDPQITVKAKERTKHRRYK
38  QEAPKDDLFEMSEAQRAKDSFITPIGSFGRLRTYMLVHSVTILQWCGYGHTVNANPEVLKIRNKPGLSVAD
39  VSDTLHATGYDERTVALNLILGIPPEKYDEAPSMDRGKNGQGVFWRTICVIAFQSANVLHQKRKSMEFSLP
40  SERPDQWCFPRPRVIILQLMVLGPDGALNYMDRLSVEHNAIATHGSVNKSRTTSTFRAVEGALQEIDYFGK
41  KRRHKKTGIKWQKGPKFRKNIMGQDKKEPECDDTYKRRKNFKIQTGKAYKRNPEIEGKAVRRKEKKKHKKKK
42  KRRNKRTGIKVQKGPKFRKDIMGRTKKEPECNYFNKRKDMKIQKGKATRKDPQIEGKAVKEKQKRKNKKKK
43  KRRDKKTGIKWQRGPKFKKYIMGRNKKEPECHYTHKKKDMKIQFGKANKRKDPEIEGKAVKEKQKRKNKKKR
44  PPRPPPPPPMVIQRPPKGSGDGGPQPPRDPPYPDYPPPKNGPIPCIPGNPRNPPWPGKGVVPHPPPPPPDPP
45  PPRPPPPPMVVQKPPYGHGDGGPQPPRPPRNGPIPYIPGNPKNPPIPGKGVVPRPPPPPDPP
46  PPRPPPPPPMVVQRPPYGDGDGGPQPPKDPPWPDYPPPKNGPIPCIPGNPKNPPIPGRGVVPKPPPPPPHPP
47  KEKNKKTGGKFQKGGKIKRNGMGQYKKEGEFDNCHKKKDWKGQPGKGYKKQPQGQGKGGKEKQKKKDKRRK
48  KEKDKKTGGKFQKGGKIKRNGWGQYKKEGECHDFNKKKDMKGQPGKGYKKQPQGQGKGGKEKQKRKEKTKHKK
49  MEQNPLPGIRCHAISTLFYSLAGPYTVEVDIDPAQRLRHLKGEFGTFMSWDSNIQGKAVKEKEDSTNLAVR
50  MEKMAVTGVRIWRALALSTPGVLQHSRYSLSNDSTPEKYLEIHPGDFNVCDFQAQGKAILEKDERFNITPG
```

Figure 8. Sequences designed to adopt the phospholipase C gamma SH3 domain fold [61], PDB entry 1hsp. Line *a* shows the natural protein sequence (residues 1, 2, and 65–71 are from the expression vector) and the secondary structure is shown on line *b*. Designed sequences are on lines *1–50*, and the profile cut-off for these sequences is 14.3. See the caption to Figure 9 for more details.

Figure 9. Sequences designed to adopt the alpha spectrin SH3 domain fold [62], PDB entry 1shg. Line *a* shows the natural protein sequence and *b* lists the crystallographically assigned secondary structure (H: helix; E: sheet; L: loop). The numbering on *c* is in accordance with [62]. Sequences designed by the prototype system (§4) to adopt this fold are given on line *1-50*. A variety of different weightings of the constituent parameters were used, as shown in the *Derivation* column; from darkest to lightest, the regions represent the relative weighting on an arbitrary scale of secondary structure preference, solvent accessibility preference, primary structure neighbor preference, and diversity. The *Prediction* column (lines *a, 1-50*) shows how well the sequences' secondary structure prediction agreed with the desired secondary structure on line *b*. For scale, line *b* shows the score that would be received by a perfect prediction. The *Profile* column shows how well the structural models of the designed sequences rate in a profile analysis. The bars show by how much the assigned score exceeds the minimum quality cutoff of 11.5. Sequences *38-40* received zero scores because they contained incorrect residues.

repetitiveness was found in some unstructured regions (e.g., positions 63-71 of Figure 8). In addition, sequences whose design criteria were dominated by one parameter often looked peculiar. For example, when secondary structure had a very large weighting, the sequences degenerated into poly-proline at turns and poly-valine in sheets (Figures 8 and 9, lines 44-46). Surprisingly, when diversity was the overwhelming criterion for making alpha spectrin's SH3 domain, the optimization procedure was apparently unable to select residues that satisfy all the criteria and inserted the nonsense residues B and Z into the sequences.

The sequences were subjected to secondary structure analysis by the PHD algorithm [64, 65]. The predicted structure was then compared against the desired structure, and a score was computed. Almost all of the designed sequences did relatively poorly on this test, as shown by the small size of the prediction bands in Figures 8 and 9. However, since the natural sequences also failed to produce accurate predictions, it is unclear how to interpret the result. As the prediction algorithm is honed to work on families of homologous sequences, it is perhaps unsurprising that it did not produce good results when asked to evaluate the sequences individually. The problem was probably exacerbated by the mostly beta-sheet structure, as beta-strands are short relative to most alpha-helices.

The designed sequences were also evaluated at the tertiary structure level. Several three dimensional models (Figures 10 and 11) were constructed for each of the designed sequences and visual inspection did not reveal any major flaws. (Some aromatic residues were innocuously misshapen because of an error in the force field used to create the models from the designed sequences.) For example, the core of the barrel of a 1hsp analog seemed suitably filled with hydrophobic residues, and the surface was mostly covered with largely polar and charged residues. It is interesting that even non-polar groups on the surface of the protein thought to be functionally important [61] were generally replaced with hydrophilic residues in the designed sequences because the design system was not provided with any parameters relating to function. There is also no reason to believe that the beta-bulges (such as that at positions 51-52 in 1hsp) and
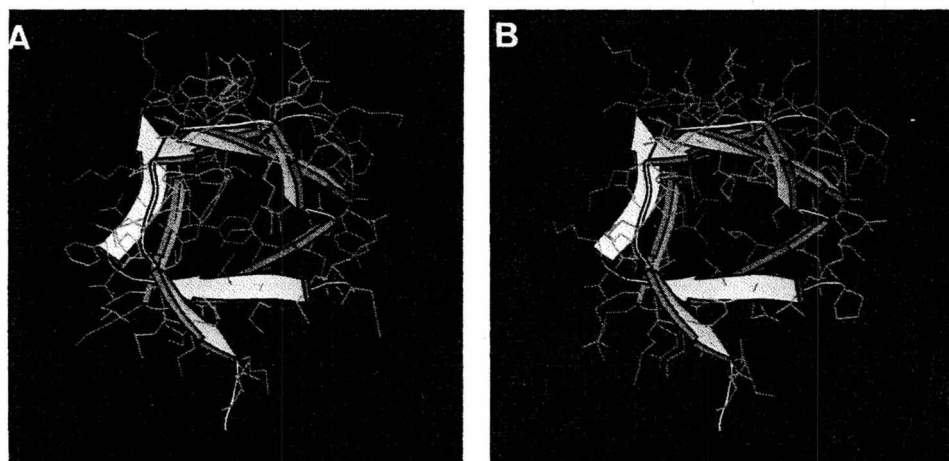
Figure 10. Human phospholipase C gamma SH3 domain structures threaded with natural and designed sequences. The coordinates and secondary structure assignments of PDB entry 1hsp are viewed in two orientations to highlight the barrel (*A, B*) and beta-sandwich (*C, D*) characteristics of the

other deformities in the native structure would exist in the designed proteins, as the model design system did not include statistical information at that level of detail.

The structural models were tested using a profile methodology [45], and most of the designed sequences received scores well above the minimum acceptable [66]. Indeed, as shown in Figure 8 and 9, many scored better than the natural sequence. However, as the criteria used to design the sequences are similar to those used by the profile method, these results must be considered suspect.

The results for sequences designed to fold into these two SH3 domains form an interesting contrast to those designed to adopt the phage 434 cro fold [44]. Sequences for this five-helix protein received very good secondary structure prediction scores and astounding profiles. Because the analyses were so laudatory over a wide variety of different weightings, it was difficult to discern the relative importance each of the constituent parameters of the optimization function. However, since the apparent quality of the sequences designed for the SH3 domains varied more widely, it was possible to glean additional insight into the interactions of the different parameters.

Perhaps the clearest conclusion that can be drawn is that there needs to be some balance between different parameters in order to produce reasonable sequences. When a single criterion dominated, as in Figures 8 and 9, lines 37-39, the results were poor. Moreover, it seems that the best sequences were generated when the sum of the diversity and secondary structure preference was about five times that solvent accessibility (using the scale in Figure 8 and 9), as in lines 14-16 and 32-37. However, as shown by lines 11-13, it seems that secondary structure preference should be the larger of the two contributors to the sum. These conclusions garner some support because the results are consistent between the sequences designed for the 1hsp and the 1shg structures and also because they can be logically interpreted. As described elsewhere [44], the solvent accessibility measure used in the model system strongly favors particular residues, especially lysine. In alpha-helices, this complements the statistically strong helical preference of lysine (data not shown); however, in beta-sheets, these statistical preference are at odds, and some compromise must be made. This compromise can be easily seen
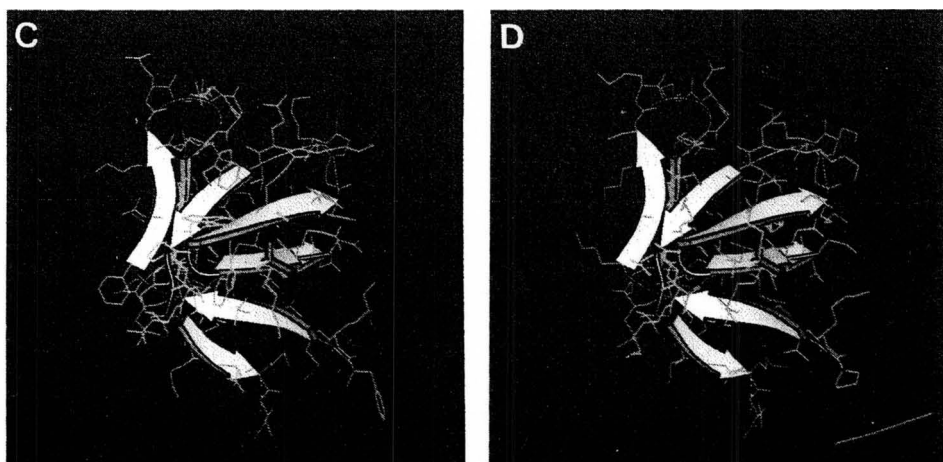
SH3 domain. Only residues 5-64 are shown because the NMR structure was constructed with constraints for residues 8-62 only. The natural protein is shown in A and C, while a model has been constructed by threading a designed sequence (Figure 8, line 14) onto the 1hsp backbone in B and D.

A   **B**
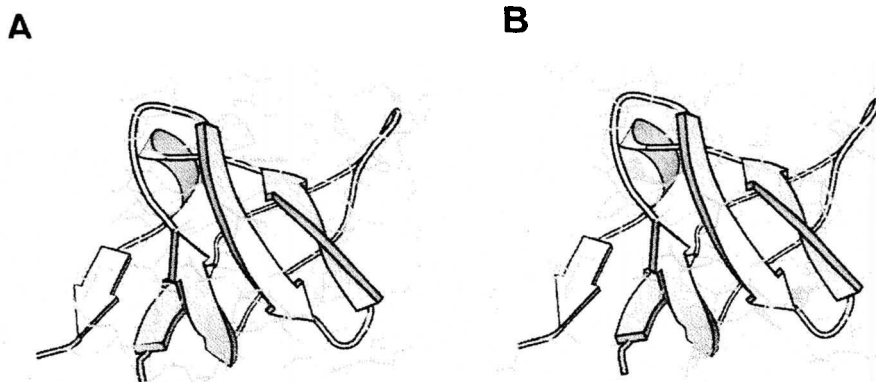
Figure 11. Chicken alpha spectrin SH3 domain structure (PDB entry 1shg) threaded with natural (*A*) and designed (*B*) sequences. The designed sequence is from Figure 9, line 14.

in the designed sequences: lysine occurred with remarkable frequency at the solvent exposed N termini of many strands, but rarely within them. The relative compatibility of solvent accessibility and secondary structure in helices would help to explain why the automated system had apparently more success with all-helix proteins than with the beta-sheet SH3 domains.

In summary, the prototype design system was capable of designing novel sequences for the SH3 domains of both phospholipase C gamma and alpha spectrin, and visual inspection revealed no major flaws in models made with the sequences. While the secondary structure predictions did not agree well with the desired structure, the native sequences generally did equally poorly. In addition, profile analysis suggests that the designed sequences could reasonably adopt the desired structures. Perhaps most intriguingly, distinguishable variation in the qualities of sequences derived from different parameter weighting made it possible to gain some insight into the relative importance of each parameter.

5. Conclusions

Protein design is a stimulating field of structural molecular biology which offers a new way to understand protein structure, and recent experiments have proven that it is possible to make new proteins with a desired fold. However, most work in this field has been irreproducible and unverifiable because of the qualitative and intuitive nature of the design process. To remedy this, I have developed a quantitative methodology for the design of proteins which operates by optimizing a quality function over the structure to be built. To support this methodology, I have created algorithms both to gather statistics for the quality function and to optimize the quality function over a particular structure.

A simplified prototype design system has been used to test the methodology, and the results have been surprisingly positive: designed sequences appear suitable to both manual inspection and profile analysis. Moreover, a tentative understanding about the contributions of different parameters to sequence selection has been gained. However, the designed sequences do reveal flaws in the model system which may be rectified by

a more complete implementation of the design methodology.

This research is now entering its most exciting stage, for more sophisticated versions of the protein design system are being built and experimental work to synthesize some designed proteins and characterize their structure is beginning. When the design methodology systems produces sequences which correctly adopt the desired folds, it will reveal precise new relationships between observed features of proteins and their structural importance.

Acknowledgments

References

[1]   C. Chothia, Proteins: 1000 families for the molecular biologist, *Nature* 357 (1992) 543-544.

[2]   W. Gilbert, The exon theory of genes, *Cold Spring Harbor Symposia on Quantitative Biology* 52 (1987) 901-905.

[3]   A. Pastore and A. M. Lesk, Brave new proteins: What evolution reveals about protein structure, *Current Opinion in Biotechnology* 2 (1991) 592-598.

[4]   M. Conrad, The lure of molecular computing, *IEEE Spectrum* 23 (1986) 55-60.

[5]   K. E. Drexler, Molecular engineering: An approach to the development of general capabilities for molecular manipulation, *Proceedings of the National Academy of Sciences, USA* 78 (1981) 5275-5278.

[6]   A. Pessi, E. Bianchi, A. Crameri, S. Venturini, A. Tramontano and M. Sollazzo, A designed metal-binding protein with a novel fold, *Nature* 362 (1993) 367-369.

[7]   M. Perutz, *Protein Structure: New Approaches to Disease and Therapy.* Freeman, New York, 1992.

[8]   J. S. Richardson and D. C. Richardson, Some design principles: Betabellin. In: D. L. Oxender and C. F. Fox (Eds.), *Protein Engineering.* Alan R. Liss, Inc., New York, 1987, pp. 149-163.

[9]   P. Y. Chou and G. D. Fasman, Beta-turns in proteins, *Journal of Molecular Biology* 115 (1977) 135-175.

[10]   P. Y. Chou and G. D. Fasman, Prediction of protein conformation, *Biochemistry* 13 (1974) 222-245.

[11]   J. Garnier, D. J. Osguthorpe and B. Robson, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *Journal of Molecular Biology* 120 (1978) 97-120.

[12]   O. B. Ptitsyn and A. V. Finkelstein, Theory of protein secondary structure and algorithm of its prediction, *Biopolymers* 22 (1983) 15-25.

[13]   S. Lifson and C. Sander, Specific recognition in the tertiary structure of beta-sheets of proteins, *Journal of Molecular Biology* 139 (1980) 627-639.

[14]   C. Sander, G. Vriend, F. Bazan, A. Horovitz, H. Nakamura, L. Ribas, A. V. Finkelstein, A. Lockhart, R. Merkl, L. J. Perry, S. C. Emery, C. Gaboriaud, C. Marks, J. Moult, C. Verlinde, M. Eberhard, A. Elofsson, T. J. P. Hubbard, L. Regan, J. Banks, R. Jappelli, A. M. Lesk and

A. Tramontano, Protein design on computers: Five new proteins: Shpilka, Grendel, Fingerclasp, Leather, and Aida, *Proteins: Structure, Function, and Genetics* 12 (1992) 105-110.

[15] C. Sander, De novo design of proteins, *Current Opinion in Structural Biology* 1 (1991) 630-637.

[16] M. H. Hecht, J. S. Richardson, D. C. Richardson and R. C. Ogden, De novo design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence, *Science* 249 (1990) 884-891.

[17] L. Regan and W. F. DeGrado, Characterization of a helical protein designed from first principles, *Science* 241 (1988) 976-978.

[18] C. P. Hill, D. H. Anderson, L. Wesson, W. F. DeGrado and D. Eisenberg, Crystal structure of Alpha-1: Implications for protein design, *Science* 249 (1990) 543-546.

[19] C. Sander, M. Scharf and R. Schneider, Design of protein structures. In: A. R. Rees, M. J. E. Sternberg and R. Wetzel (Eds.), *Protein Engineering: A Practical Approach.* Oxford University Press, Oxford, 1992, pp. 89-115.

[20] C. E. Schafmeister, L. J. W. Miercke and R. M. Stroud, Structure at 2.5Å of a designed peptide that maintains solubility of membrane-proteins, *Science* 262 (1993) 734-738.

[21] K. Goraj, A. Renard and J. A. Martial, Synthesis, purification and initial structural characterization of Octarellin, a de novo polypeptide modeled on the alpha/beta-barrel proteins, *Protein Engineering* 3 (1990) 259-266.

[22] T. Tanaka, H. Kimura, M. Hayashi, Y. Fujiyoshi, K. Fukuhara and H. Nakamura, Characteristics of a de novo designed protein, *Protein Science* 3 (1994) 419-427.

[23] T. J. Hubbard and T. L. Blundell, The design of novel proteins using a knowledge-based approach to computer-aided modelling. In: W. F. van Gunsteren and P. K. Weiner (Eds.), *Computer Simulations of Biomolecular Systems: Theoretical and Experimental Applications.* ESCOM, Leiden, Holland, 1989, pp. 168-82.

[24] A. N. Fedorov, D. A. Dolgikh, V. V. Chemeris, B. K. Chernov, A. V. Finkelstein, A. A. Schulga, Y. B. Alakhov, M. P. Kirpichnikov and O. B. Ptitsyn, De novo design, synthesis and study of Albebetin, a polypeptide with a predetermined three-dimensional structure: Probing the structure at the nanogram level, *Journal of Molecular Biology* 225 (1992) 927-931.

[25] B. Lovejoy, S. Choe, D. Cascio, D. K. Mcrorie, W. F. DeGrado and D. Eisenberg, Crystal structure of a synthetic triple-stranded alpha-helical bundle, *Science* 259 (1993) 1288-1293.

[26] W. F. DeGrado and B. W. Matthews, Engineering and design, *Current Opinion in Structural Biology* 3 (1993) 547-548.

[27] Y. Kuroda, T. Nakai and T. Ohkubo, Solution structure of a de novo helical protein by 2D-NMR spectroscopy, *Journal of Molecular Biology* 236 (1994) 862-868.

[28] P. Balaram, The design and construction of synthetic protein mimics, *Pure and Applied Chemistry* 64 (1992) 1061-1066.

[29] K. R. Rajashankar, S. Ramakumar and V. S. Chauhan, Design of a helical motif using alpha,beta-dehydrophenylalanine residues: Crystal-structure of boc-val-delta-phe-phe-ala-phe-delta-phe-v l-delta-phe-gly-och3, a 3_{10}-helical nonapeptide, *Journal of the American Chemical Society* 114 (1992) 9225-9226.

[30] H. Diaz, K. Y. Tsang, D. Choo and J. W. Kelly, The design of water-soluble beta-sheet structure based on a nucleation strategy, *Tetrahedron* 49 (1993) 3533-3545.

[31] T. P. Singh, P. Narula and H. C. Patel, Alpha,beta-dehydro residues in the design of peptide and protein structures, *Acta Crystallographica B* 46 (1990) 539-545.

[32] B. Imperiali, S. L. Fisher, R. A. Moats and T. S. Prins, Unnatural amino-acids in de novo protein design, *Abstracts of Papers of the American Chemical Society* 203 APR (1992) 101.

[33] T. E. Creighton, Protein Folding, *Biochemical Journal* 270 (1990) 1-16.

[34] K. S. Akerfeldt, R. M. Kim, D. Camac, J. T. Groves, J. D. Lear and W. F. DeGrado, Tetraphilin: A four-helix proton channel built on a tetraphenylporphyrin framework, *Journal of the American Chemical Society* 114 (1992) 9656-9657.

[35] M. Mutter, B. Dorner, C. Sigel, R. Floegel, C. Servis and G. Tuchscherer, Topological templates as a tool in molecular recognition and in protein design, *Journal of Cellular Biochemistry* 17C (1993) 211.

[36] G. Tuchscherer, B. Dorner, U. Sila, B. Kamber and M. Mutter, The TASP concept: Mimetics

of peptide ligands, protein surfaces, and folding units, *Tetrahedron* 49 (1993) 3559-3575.

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

[38]   A. R. Davidson and R. T. Sauer, Folded proteins occur frequently in libraries of random amino-acid sequences, *Proceedings of the National Academy of Sciences, USA* 91 (1994) 2146-2150.

[39]   S. Kamtekar, J. M. Schiffer, H. Y. Xiong, J. M. Babik and M. H. Hecht, Protein design by binary patterning of polar and nonpolar amino-acids, *Science* 262 (1993) 1680-1685.

[40]   J. W. Ponder and F. M. Richards, Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes, *Journal of Molecular Biology* 193 (1987) 775-791.

[41]   C. Chothia and A. M. Lesk, The relation between the divergence of sequence and structure in proteins, *EMBO Journal* 5 (1986) 823-826.

[42]   K. Yue and K. A. Dill, Inverse protein folding problem: Designing polymer sequences, *Proceedings of the National Academy of Sciences, USA* 89 (1992) 4163-4167.

[43]   S. E. Brenner, *Development of a Quantitative Methodology for the de novo Design of Proteins.* M. Phil. Thesis, University of Cambridge, 1993.

[44]   S. E. Brenner and A. Berry, A quantitative methodology for the de novo design of proteins [submitted]

[45]   R. Lüthy, J. U. Bowie and D. Eisenberg, Assessment of protein models with three-dimensional profiles, *Nature* 356 (1992) 83-85.

[46]   A. M. Lesk and D. R. Boswell, Does protein structure determine amino acid sequence, *Bioessays* 14 (1992) 407-410.

[47]   J. U. Bowie, R. Lüthy and D. Eisenberg, A method to identify protein sequences that fold into a known three-dimensional structure, *Science* 253 (1991) 164-170.

[48]   D. T. Jones, W. R. Taylor and J. M. Thornton, A new approach to protein fold recognition, *Nature* 358 (1992) 86-89.

[49]   F. C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, The Protein Data Bank: A computer-based archival for macromolecular structures, *Journal of Molecular Biology* 112 (1977) 535-542 .

[50]   E. E. Abola, F. C. Bernstein, S. H. Bryant, T. F. Koetzle and J. Weng, Protein Data Bank. In: F. H. Allen, G. Bergerhoff and R. Sievers (Eds.), *Crystallographic Databases: Information Content, Software Systems, Scientific Applications.* Data Commission of the International Union of Crystallography, Cambridge, 1987, pp. 107-132.

[51]   J. S. Richardson, D. C. Richardson, N. B. Tweedy, K. M. Gernert, T. P. Quinn, M. H. Hecht, B. W. Erickson, Y. B. Yan, R. D. McClain, M. E. Donlan and M. C. Surles, Looking at proteins: Representations, folding, packing, and design, *Biophysical Journal* 63 (1992) 1186-1209.

[52]   R. Lüthy, A. D. Mclachlan and D. Eisenberg, Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein-sequence databases for structural similarities, *Proteins: Structure, Function, and Genetics* 10 (1991) 229-239.

[53]   U. Hobohm, M. Scharf, R. Schneider and C. Sander, Selection of representative protein data sets, *Protein Science* 1 (1992) 409-417.

[54]   S. E. Brenner and A. Berry, A weighting scheme for biomolecular structures and sequences [in preparation].

[55]   E. Horowitz and S. Sahni, *Fundamentals of Computer Algorithms.* Computer Science Press, Rockville, MD, 1978.

[56]   A. V. Aho, J. E. Hopcroft and J. D. Ullman, *Data Structures and Algorithms.* Addison-Wesley, Reading, MA, 1983.

[57]   W. H. Press, B. P. Flannery, S. A. Teukolsky and W. T. Vetterling, *Numerical Recipies in C: The Art of Scientific Computing.* University of Cambridge Press, Cambridge, 1988.

[58]   N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, Equation of state calculations by fast computing machines, *Journal of Chemical Physics* 21 (1953) 1087-1092.

[59]   A. Berry and S. E. Brenner, A prototype computer system for de novo protein design, *Biochemical Society Transactions* [in press].

[60]   J. Kuriyan and D. Cowburn, Structures of SH2 and SH3 domains, *Current Opinion in Structural Biology* 3 (1993) 828-837.

[61]   D. Kohda, H. Hatanaka, M. Odaka, V. Mandiyan, A. Ullrich, J. Schlessinger and F. Inagaki, Solution structure of the SH3 domain of phospholipase C gamma, *Cell* 72 (1993) 953-960.

[62]   A. Musacchio, M. Noble, R. Pauptit, R. Wierenga and M. Saraste, Crystal structure of a src homology 3 (SH3) domain, *Nature* 359 (1992) 851-855.

[63]   S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, Basic Local Alignment Search Tool, *Journal of Molecular Biology* 215 (1990) 403-410.

[64]   B. Rost and C. Sander, Prediction of protein secondary structure at better than 70% accuracy, *Journal of Molecular Biology* 232 (1993) 584-599.

[65]   B. Rost, C. Sander and R. Schneider, PHD: An automatic mail server for protein secondary structure prediction, *Computer Applications in the Biosciences* 10 (1994) 53-60.

[66]   D. Eisenberg, J. U. Bowie, R. Lüthy and S. Choe, Three-dimensional profiles for analyzing protein-sequence structure relationships, *Faraday Discussions of the Chemical Society* 93 (1992) 25-34.

[67]   P. J. Kraulis, Molscript: A program to produce both detailed and schematic plots of protein structures, *Journal of Applied Crystallography* 24 (1991) 946-950.