

De Novo Protein Design. I. In Search of Stability and Specificity

Patrice Koehl* and Michael Levitt*

Department of Structural
Biology, Fairchild Building
Stanford University, Stanford
CA, 94305, USA

We have developed a fully automated protein design strategy that works on the entire sequence of the protein and uses a full atom representation. At each step of the procedure, an all-atom model of the protein is built using the template protein structure and the current designed sequence. The energy of the model is used to drive a Monte Carlo optimization in sequence space: random moves are either accepted or rejected based on the Metropolis criterion. We rely on the physical forces that stabilize native protein structures to choose the optimum sequence. Our energy function includes van der Waals interactions, electrostatics and an environment free energy. Successful protein design should be specific and generate a sequence compatible with the template fold and incompatible with competing folds. We impose specificity by maintaining the amino acid composition constant, based on the random energy model. The specificity of the optimized sequence is tested by fold recognition techniques. Successful sequence designs for the B1 domain of protein G, for the lambda repressor and for sperm whale myoglobin are presented. We show that each additional term of the energy function improves the performance of our design procedure: the van der Waals term ensures correct packing, the electrostatics term increases the specificity for the correct native fold, and the environment solvation term ensures a correct pattern of buried hydrophobic and exposed hydrophilic residues. For the globin family, we show that we can design a protein sequence that is stable in the myoglobin fold, yet incompatible with the very similar hemoglobin fold.

© 1999 Academic Press

Keywords: protein structures; protein sequence; design; specificity; random energy model

*Corresponding author

Introduction

Solving the protein folding problem remains the “holy grail” for computational structural biology; the goal is to predict the three-dimensional structure of a protein from its sequence (Dill *et al.*, 1995; Dill & Chan, 1997; Levitt *et al.*, 1997; Nath & Udgaonkar, 1997; Shakhnovich, 1997; Dobson *et al.*, 1998; Skolnick *et al.*, 1998). Most efforts in the field are directed towards the understanding of the basic physical and chemical laws that define a protein structure, as well as towards unraveling the steps of the folding process itself. Though substantial exper-

imental and theoretical progresses have been made in both directions, the “grail” is still out of reach. An alternative route that could lead indirectly to this goal is to invert the problem: reformulate the quest as searching for protein sequences that fold into a given stable conformation. This is the “inverse folding problem” (Drexler, 1981; Pabo, 1983), which has attracted considerable interest as it is fundamental for protein design and engineering (Mutter & Tuchscherer, 1997; Smith & Regan, 1997; Cao *et al.*, 1998; Giver & Arnold, 1998; Regan & Wells, 1998; Schafmeister & Stroud, 1998; Shakhnovich, 1998). Since the function of a protein is directly related to its three-dimensional structure, manipulation of the structure *via* the sequence changes will provide functional diversity. Protein molecules can be engineered to optimize their activities as well as to alter their pharmacokinetic properties, which is of interest for therapeutically important molecules. Another aim of protein engin-

Abbreviations used: GB1, B1 unimmunoglobulin-binding domain of protein G; vdW, van der Waals; SD, design sequences; SN, native sequences.

E-mail addresses of the corresponding authors:
koehl@hyper.stanford.edu and
michael.levitt@stanford.edu

engineering is to synthesize proteins that exhibit novel activities. An example is the chemical addition of a toxin to antibodies specific for cancer cells so as to enable more efficient, targeted treatment of tumors (Pastan *et al.*, 1995; Chowdhury *et al.*, 1998; Kreitman & Pastan, 1998).

It is clear that the folding problem and the inverse folding problem are related: the physical laws that govern folding also stipulate the protein sequence. There are major differences in the approaches used to solve these problems. A natural protein can be assumed to adopt a unique conformation under given environmental conditions; this is the assumption that computational structural biology relies upon (Anfinsen, 1973). For a given protein, this unique structure can be found experimentally (both X-ray crystallography and NMR spectroscopy have matured into reliable techniques for high-resolution structure determination).

Protein design is different in that the inverse folding problem usually does not provide a unique sequence as an answer. The key question to be answered is: how many, and which sequences can fold into a given conformation? This involves a search in sequence space for sequences that make the native structure both stable and unique. Though chemical synthesis of protein is becoming standard, an experimental exhaustive search of sequence space is totally unrealistic, and the protein engineer is faced with the problem of defining which sequence or family of sequences to synthesize. One approach is to rely solely on evolution (Arnold, 1998a,b). Alternatively, rational protein design usually proceeds in a hierarchical fashion following the hierarchy of forces that stabilize protein structures (Bryson *et al.*, 1995). These approaches have led to impressive successes for local protein design, such as the introduction of a metal binding site for affinity purification (Arnold & Haymore, 1991), changes in substrate specificity (Wilks & Holbrook, 1991; Elhawrani *et al.*, 1994), and transfer of active sites to small natural peptide scaffolds (Vita *et al.*, 1995; Vita, 1997; Mer *et al.*, 1998). They have been successfully extended to the design of small compact proteins (Lazar *et al.*, 1997; Dahiyat & Mayo, 1997a; Walsh *et al.*, 1999). It is worth noting that these recent successes in protein design are not the result of a simple hierarchic procedure; they rely more on exhaustive, fully automatic computer-based strategies, combined with feedback from experiments.

A successful protein design calculation should generate a sequence compatible with the template fold (the "design in" procedure), and incompatible with competing folds (the "design out" procedure, or specificity problem). A rigorous solution to this problem requires simultaneous exploration of both the sequence space and the conformation space. While this may be feasible for a short peptide chain with a simplified representation (Seno *et al.*, 1996), it cannot be applied to a longer protein chain with a detailed all-atom representation. In some studies, this problem has been ignored:

Malakauskas & Mayo (1998) successfully redesigned the core of the B1 domain of protein G, using a variant of the dead-end elimination algorithm, without explicit consideration of specificity. Ignoring specificity, however, has not always been successful. In the case of the HP model, for example, super-stable sequences have been designed with all-H inside and all-P at the surface (Shakhnovich & Gutin, 1993a,b). These sequences, however, are not specific, and can fold into many "native" conformations (Yue & Dill, 1992). The design of metal-ion binding sites with specified spatial arrangement sites was only possible by considering specificity (Coldren *et al.*, 1997; Pinto *et al.*, 1997; Hellinga, 1998b,c). Shakhnovich & Gutin (1993a,b) proposed a simple, approximate solution to the problem of specificity, based on the random energy model (for a review, see Pande *et al.*, 1997). In their approach, sequence design proceeds by selecting sequences that have low energy in the template conformation at fixed amino acid composition. This procedure has been applied to protein design simulation on lattice, and is fast enough to be used for off-lattice sequence optimization.

The design of a protein sequence, S , for a given template conformation, C , cannot be considered complete before it is shown that S folds, and that its structure is C . The obvious solution to this problem is to synthesize S , fold it experimentally, and study its conformation using either NMR spectroscopy or X-ray crystallography. Failure at this level can then be used as corrective feedback for sequence optimization. This is the design cycle proposed by Dahiyat & Mayo (1996, 1997a,b); it was shown to be successful for the design of the core of protein G (Malakauskas & Mayo, 1998), of surface residues in GCN4 (Dahiyat *et al.*, 1997), and for the nearly complete design of a small zinc finger protein (Dahiyat & Mayo, 1997a). While experimental validation must be the ultimate choice for testing any theoretical approach, it is of great practical value to have powerful design tools that consider specificity.

A true computing equivalent to the experimental procedure involves solving the protein folding problem. This has been carried out for simple problems, with a simplified HP alphabet on a lattice (for a review, see Dill *et al.*, 1995); extension to more complex systems, such as the full atom protein model considered here, is still a very difficult, as yet unsolved problem. An alternative approach is to use a fold recognition technique (for reviews, see Jones *et al.*, 1995; Fischer & Eisenberg, 1996; Fischer *et al.*, 1996; Miller *et al.*, 1996; Mirny & Shakhnovich, 1998), which usually rely on "hide-and-seek" computer experiments. In this procedure, also referred to as "threading", the template structure, C , for the designed sequence, S , is hidden among a large number of non-native folds, N ; the design is said to be successful if C can be separated from all N based on an energy criteria, i.e. $E(S,C) < E(S,N)$ for all N .

Here, we present a simple protein design strategy that incorporates all the elements described

above. We use a full atom representation of the protein and a physical energy function. Optimization is based on a search in sequence space, where random moves are either accepted or rejected using the Metropolis Monte Carlo criterion. The amino acid composition is maintained constant, following the "canonical" method (Shakhnovich & Gutin, 1993a,b). At each step of the calculation, a hybrid protein based on the known template backbone structure and the current designed sequence is modeled using our own very fast method for side-chain placement (Koehl & Delarue, 1994a). The energy of the model is then used to drive the Monte Carlo optimization. In previous work (Koehl & Delarue, 1997) an initial version of this procedure, using only the van der Waals energy to measure the fitness of a sequence on a structure, proved unsuccessful. Here we use a more general energy function including van der Waals (vdW), electrostatics and solvent interactions. This energy functions is deliberately chosen to only include physical terms, so as to allow better insight into the sequence optimization process. All designed sequences are tested for specificity, using a fold recognition technique.

In this report, the first in a series of two, the method is fully described, emphasizing the importance of each term of the energy function considered. The success of the procedure is assessed with respect to its ability to design in and design out sequences for different template backbones, including the B1 domain of protein G, lambda repressor and myoglobin. Comparisons with experimental mutation data are provided for protein G and lambda repressor. In the accompanying paper, we show in greater detail how much sequence information can be retrieved from the backbone template of a protein using our physical energy function (Koehl & Levitt, 1999).

Results

Reaching stability by "design in"

A test case: GB1

We chose the B1 immunoglobulin-binding domain of streptococcal protein G (GB1) as the first test molecule for our sequence design procedure. Protein G is small (56 residues), highly stable and very regular, with 80% of the residues participating in secondary structure. It has been used as a model for studying β -sheet propensities of amino acid residues (Minor & Kim, 1994a,b; Smith *et al.*, 1994; Smith & Regan, 1995), as a scaffold on which a metal binding site was engineered (Farinas & Regan, 1998), and as a template in the Paracelsus challenge, where it was converted to a fully helical protein by changing only 50% of its sequence (Dalal *et al.*, 1997). Its structure has been solved by NMR (Gronenborn *et al.*, 1991) and by X-ray crystallography (Gallagher *et al.*, 1994). We chose the X-ray structure as the target native con-

formation. Coordinates for the backbone atoms were extracted from the PDB file 1PGB, and all side-chain atoms were discarded. We only used the amino acid composition of the native sequence as input for all the sequence optimizations described below.

As described above, the complete energy function contains three terms: a Lennard-Jones or van der Waals term (E_{vdW}) for steric interactions, a Coulomb term for electrostatics (E_{elec}), and a surface dependent, semi empirical free energy of environment (E_{env}). In order to assess the relative importance of each of three components, sequence design calculations were performed for GB1 according to the procedure described in Methods, using sequence selection based on vdW interactions, on vdW and electrostatics, and on the full energy function, respectively. In all three optimizations, the initial sequence was chosen to be a complete random reshuffling of the native sequence of GB1. Variation of each individual energy term along the course of the three simulations is shown in Figure 1.

Optimization based on E_{vdW} alone improves packing as defined by Lennard-Jones interactions, yielding protein models with vdW energies much lower than the corresponding energy of the native protein. This had already been observed in the earlier study of chymotrypsin inhibitor (Koehl & Delarue, 1997). Interestingly, optimization of E_{vdW} does not reduce the electrostatics and environment free energies. Similar behavior is observed in the second simulation, where both the vdW and electrostatic energies of the final model are much lower than the corresponding energies of the native structure. Again, little reduction of the environment free energy is observed. When the complete energy function is used, all three components are reduced, but to a lesser degree than before. In every case, the total energy of the model protein built from the optimized sequence is lower than that of the native protein.

Repeating the sequence design procedure by setting the temperature of the Monte Carlo to infinity results in no optimization as all random moves are accepted. From such a simulation, we determine that the average level of sequence identity for random sequences with the GB1 amino acid composition to be around 10%. Sequence optimizations based on each of the three different terms mentioned give higher levels of sequence identity (Figure 2). A detailed analysis of S_{opt} , the sequence obtained by optimization with all three energy components, is given in Figure 3. S_{opt} is 28.7% identical with the native sequence of protein G. A FASTA search against the non-redundant SWISS-PROT sequence database does identify S_{opt} as being similar to the native sequence, S_{nat} but with marginal significance (E value of 0.4). No other sequence scores better. PHD secondary structure prediction (Rost & Sander, 1993) on S_{opt} correctly identifies three out of the four strands, as well as the central alpha helix.

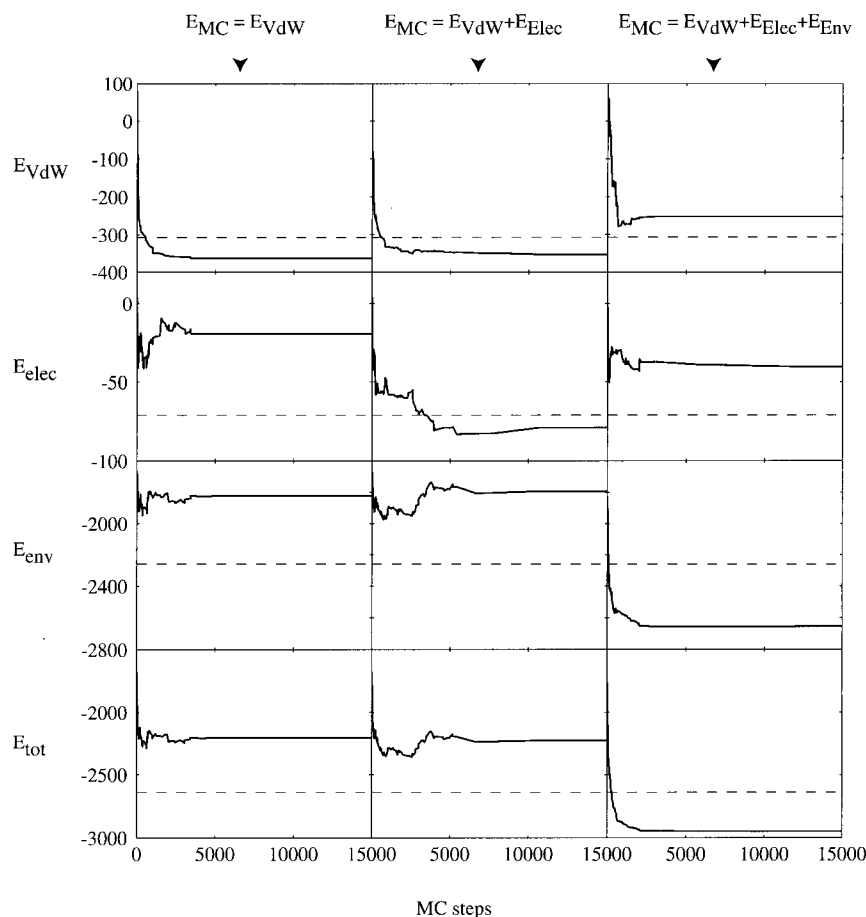


Figure 1. Evolution of the different components of the energy function upon design of the B1 domain of protein G. Three different optimizations are described (down the columns), in which the energy function used to drive the sequence selection contains a vdW term only (E_{vdW}), vdW + electrostatics ($E_{vdW} + E_{Elec}$), and the complete energy function (E_{tot}), respectively. All energy terms are monitored for each optimization, even if they are not optimized (along the row). In each case, the energy of the native protein is shown as a broken line.

Contributions of each energy term

It is well known that the same fold can accommodate a large variety of sequences and that proteins are quite tolerant to mutations. It is, therefore, important that a sequence design procedure be able to generate a family of sequences. This is possible with the procedure described here, as it is relatively fast (complete sequence design for GB1 takes five hours and 30 minutes on a DEC processor alpha at 533 Mhz) and can be repeated from different initial shuffled sequences.

Sequence design for protein G was repeated ten times, for each of five variants of the energy function: vdW only, the environment free energy only (env), the two combinations vdW + elec and vdW + env, and the total energy (Tot). Results showing the sequence variations for six selected residues of GB1 are given in Table 1. These six residues were chosen to illustrate the importance of each term of the energy function: Leu5, Ala26, Phe52 and Val54 are fully buried core residues, whereas Gly14 and Val21 are exposed. Residue

Gly14 in GB1 has a positive ϕ angle and negative ψ angle, and is therefore expected to be a glycine residue. Though residue 21 is highly accessible, it is a valine residue in the native sequence.

Sequence design based on vdW only. The Lennard-Jones term is a measure of steric constraints in the protein in that any "forbidden" contact is strongly penalized. Thus, the glycine residue is selected in all ten E_{vdW} designed sequences for residue Gly14. Similarly, while residue Ala26 is part of the central helix of GB1, it is completely buried, and there is not much room to replace it with a larger residue: as a consequence, alanine is generally selected at position 26. The Lennard-Jones term does not account for electrostatics or environment. As a consequence, a buried hydrophobic residue can be replaced by a similarly shaped charged or polar residue at little cost in energy. This is observed for residue Val54, which is fully buried in the native structure, and for which optimization generally selects lysine or threonine residues. In general, E_{vdW} does not restrict the choice of amino

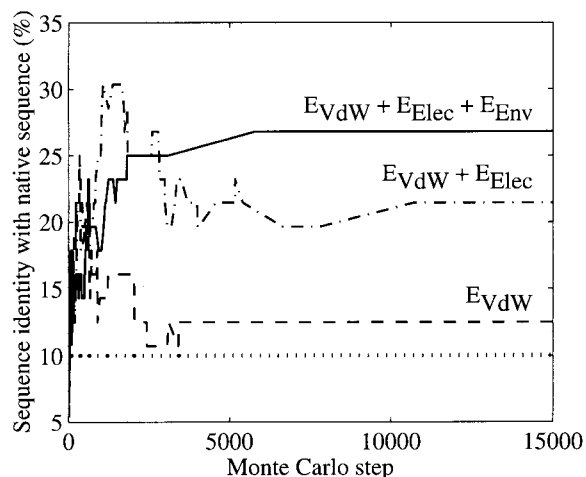


Figure 2. Variation of percentage identity to the native sequence of GB1 of sequences designed for the B1 domain of protein G with respect to the Monte Carlo cycle number, during three different sequence designs based on E_{vdW} only (---), $E_{\text{vdW}} + E_{\text{elec}}$ (-·-·-) and $E_{\text{tot}} = E_{\text{vdW}} + E_{\text{elec}} + E_{\text{env}}$ (—). The broken line shows the average sequence identity with the GB1 amino acid composition, respectively.

acid selection for residues that are not sterically constrained. Threonine, which is the most abundant amino acid residue in GB1 (11 out of 56 residues, or nearly 20% in composition), is the most likely choice for most of the positions considered here.

Sequence design based on environment only. Solvent interactions are considered to be the major forces for protein folding, and consequently also for sequence specificity. The environment energy E_{env} provides the hydrophobic/hydrophilic partitioning of the sequence in our procedure. This is observed for all core positions in GB1 that are generally occupied by hydrophobic residues in the designed sequences. The tendency may be too strong: residue Val21, which is fairly exposed, is

always replaced by an aspartic acid residue in the designed sequences. Similarly, residue Val14, which has an accessibility of 38% in the native protein, is mainly replaced with a lysine residue. While this makes sense in terms of accessibility, it does not correctly illustrate the fact that residue 14 is structurally constrained by close contacts (see above). The same is true for residue Ala26, which is buried in the native sequence (Ala), but often replaced by large non-polar residues in the designed sequences.

Combining the energy terms. Combination of E_{vdW} and E_{elec} cannot correctly partition hydrophobic residues in the core, and exposed residues at the surface (Table 1). It should be noted that E_{elec} includes a Coulomb term for intra-protein interactions, and does not take in account the solvent. Similarly, the combination $E_{\text{elec}} + E_{\text{env}}$ cannot ensure proper packing interactions, such as those observed on Gly14 (result not shown). The combined energy function based on E_{vdW} and E_{env} corrects for the limitation of each of these components used alone, while maintaining the advantages of both. Further improvement is observed when all three terms are included in the energy function, mainly in terms of recovering the native sequence. For example, position Val54 is a valine residue three times out of ten when the design selection is based on E_{vdW} and E_{env} ; when E_{elec} is added it is a valine residue eight times out of ten.

Reaching specificity by design out

A simple test case: GB1

We have seen above in the case of GB1 that each additional term of our energy function provides information that helps in the process of extracting sequence information from a protein backbone. The complete energy function, however, is the only one that provides specificity to the designed sequence. This was shown by performing protein fold recognition for the designed sequence, and monitoring the scores for the target fold GB1

	1	10	20	30	40	50
GB1	MTYKLIILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWYDDATKFTVTVE					
Design	TTFWLVIAAATDETGTQTEDKEETDKAATYFKTYLNKKGVGGNVNLDTDAYKMTVEA					
PHD Design	EEEEEEE		HHHHHHHHHHHHHHHHHH		EEEEEE	EEEE
DSSP GB1	EEEEEEE	SS	EEEEEEE	SSHHHHHHHHHHHHHHTT		EEEEETTTEEEEE
PHD GB1	EEEEEE		EEEEHHHHHHHHHHHHHHHH		EEEE	EEEE

Figure 3. Characteristics of a designed sequence for the B1 domain of protein G: alignment with the native sequence (sequence identity: 26.8%) and comparison of the secondary structure elements of the designed sequence, with those of the actual native sequence for GB1. Secondary structure prediction is performed using PHD (Rost & Sander, 1993). The raw score of the sequence alignment based on the Blosum50 scoring matrix is 71. The designed sequence was tested against the non-redundant SWISSPROT database using FASTA, and the native sequence of protein G scored best, with an E -value of 0.4

Table 1. Structure dependent substitution matrix for GB1

Ene	Res.	Struct ^a	Access (%)	Access																			
				G	A	V	I	L	F	P	M	W	C	S	T	N	Q	Y	H	D	E	K	R
vdW	L5	β	0		1									1	6							2	
	G14	γ	38	10																			
	V21	α_r	75			2								7	1								
	A26	α_r	0		7	1								1							1		
	F52	β	2		1									3	1					2		3	
Env	V54	β	0			1					1			4								4	
	L5	β	0		2	1	1	2	1		2							1					
	G14	γ	38		1					1									1			7	
	V21	α_r	75																		10		
	A26	α_r	0		3	1	1	1			4												
VdW + Elec	F52	β	2			1			5	2											2		
	V54	β	0							8		1									1		
	L5	β	0		1												7					2	
	G14	γ	38		10																		
	V21	α_r	75				2										7						
vdW + Env	A26	α_r	0		1	7																1	
	F52	β	2			1																7	
	V54	β	0				5															3	
	L5	β	0		4	3																	
	G14	γ	38		9																	1	
vdW + Elec + Env	V21	α_r	75																		10		
	A26	α_r	0		9	1																	
	F52	β	2			4				4		1									1		
	V54	β	0		2	3				1	1			1							2		
	L5	β	0		4	2				4												2	
vdW + Elec + Env	G14	γ	38		9																	1	
	V21	α_r	75																			9	
	A26	α_r	0		10																	1	
	F52	β	2			3				4		1	1								1		
	V54	β	0			8				1											1		

^a Geometry of the residue defined from its position on the Ramachandran plot: β stands for $\Phi < 0$, $\Psi > 0$, α_r for $\Phi < 0$, $\Psi < 0$, and γ stands for $\Phi > 0$, $\Psi < 0$ (region mainly populated by glycine residues).

during the Monte Carlo optimization, using PROSA (Hendlich *et al.*, 1990; Sippl & Weitkus, 1992; Sippl, 1993; see also <http://lore.came.sbg.ac.at/>) (Figure 4(a)) and THREADER (Jones *et al.*, 1992; see also <http://globin.bio.warwick.ac.uk/jones/threader.html>) (Figure 4(b)). Since these two well-respected programs both use scoring functions based on statistics of interactions in known native protein structures rather than physical forces, they can be considered as reasonable independent assessors. It is worth noticing that the sequence designed for GB1 based on the total energy function achieves a specificity towards the GB1 fold very similar to that of the native sequence based on PROSA, while assessment of the same sequence using THREADER yields a lower specificity of approximately 0.5. The optimized sequence, S_{opt} for GB1 (see Figure 3) was recognized to adopt the protein G fold by PROSA, THREADER, and two profile methods (Gribskov *et al.*, 1987; Fischer & Eisenberg, 1996; Rice & Eisenberg, 1997) (Table 2). The fold of GB1 is not unique, and other proteins have been found to adopt very similar folds. It is interesting that for all protein fold recognition methods but one (PROFILESEARCH), GB1 itself came out as the best scoring fold. In the case of PROFILESEARCH, GB1 only ranked 6; all sequences that did score better correspond to protein with similar folds (PDB files

codes were 1FCC, 2IGG, 1IGCA, 1IGD and 1PGX, in descending order of Z-scores).

Contribution of packing, electrostatics, and solvent to specificity

A more complete analysis of the ten sequences designed for GB1 was performed (see Figure 5 and Table 3). With PROSA, the GB1 sequences designed using the environment energy (E_{env}) only, appear to be as specific as sequences designed with the complete energy function (Tot) (see Figure 5(a)). This suggests that PROSA is mainly dominated by the contribution of its environment term, which measures the hydrophobic/hydrophilic pattern in the protein. With THREADER, the results (Figure 5(b)) are compatible with what we deduced based on the sequence only: E_{vdW} alone or E_{env} alone does not provide specificity, which is obtained only by combining the two terms. Addition of the electrostatics term increases the retention of wild-type sequence features, but does not improve specificity.

Our assessment of these two fold recognition techniques is based on information obtained by comparing the designed sequences with the native sequence for GB1. To rule out the possibility that PROSA or THREADER only detects sequence similarities, ten independent Monte Carlo sequence

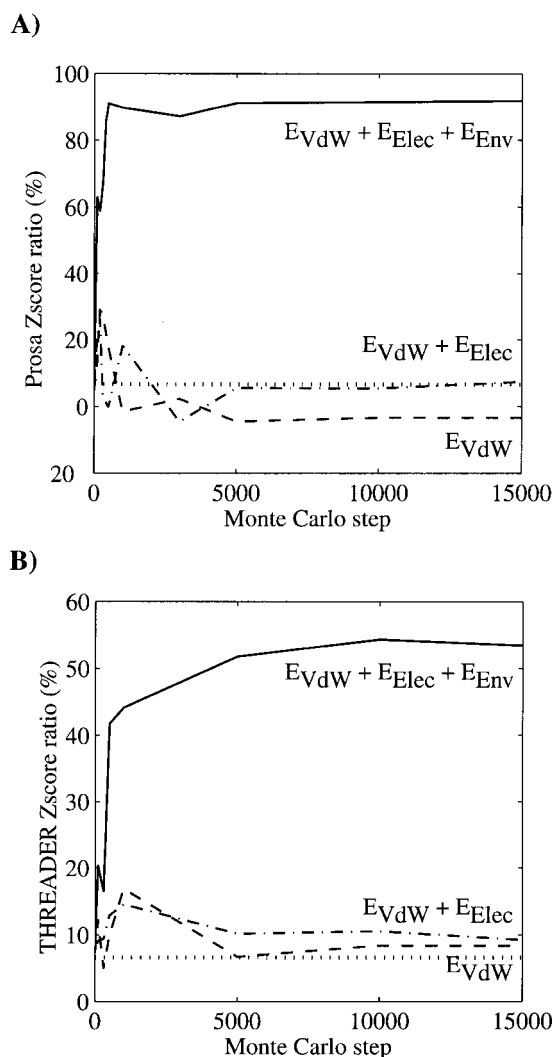


Figure 4. (a) Variation of specificity of sequences designed for the B1 domain of protein G with respect to the Monte Carlo cycle number, during three different sequence designs based on E_{vdW} only (---), $E_{vdW} + E_{elec}$ (- · - ·) and $E_{tot} = E_{vdW} + E_{elec} + E_{env}$ (—). The specificity of the sequence for the structure of protein G is evaluated by (a) PROSA and (b) THREADER. The variation of the ratio of the Z-score of the designed sequence and the Z-score of the native sequence is shown with respect to the number of MC steps: a ratio of 0 indicates poor specificity, while a ratio of 100% correspond to a specificity equal to that of the native sequence. The broken lines show the average specificity found for ten random sequences with the GB1 amino acid composition, respectively.

design were performed at infinite temperature over 100,000 cycles. (i.e. without any selection criteria). The sequence with the highest identity to the native GB1 is selected in each run. In each case this sequence shows a level of identity close to 30%; none however shows a significant “specificity” to the native fold for GB1, as determined by PROSA and THREADER (point “Bias” in Figure 5(a) and

Table 2. Specificity of the sequence designed for protein G

Method	S_{nat}	S_{opt}
PROSA	-7.08 (1)	-6.50 (1)
THREADER	18.23 (1)	9.74 (1)
PROFILESEARCH	40.23 (1)	39.21 (6)
H3P2	3.00 (1)	2.70 (1)

Ability to recognize the correct fold (given in the PDB file 1PGB), both for the native sequence (S_{nat}) and for the designed sequence (S_{opt}). Four different methods for fold recognition are considered: PROSA, THREADER, PROFILESEARCH and H3P2 (see Methods for a presentation of each technique). Negative Z-scores indicate stability (except for PROSA, which defines the Z-score with an opposite sign). The rank of the native fold is given in parenthesis.

(b)). Clearly the level of specificity observed in the optimized sequences for GB1 cannot be due to the level of sequence similarity.

Based on these results, THREADER was chosen to be the better tool for assessing specificity. We also tested the 3D-1D profile methods available on the fold recognition server by Fisher and Eisenberg (see Methods), which compares well with THREADER (results not shown). The choice for THREADER was also influenced by the fact that we can run the program locally, with full control on its input.

Specificity within fold families: the globin case

The first two high-resolution protein structures determined by the pioneers of protein X-ray crystallography were myoglobin (Kendrew *et al.*, 1960), and hemoglobin (Perutz *et al.*, 1960). These two proteins turn out to have very similar folds: this could not have been predicted from their sequences alone, which show very low level of similarity. Globins which constitute a large family of similar folds (Bashford *et al.*, 1987) remain one of the most studied protein families, which includes myoglobins, hemoglobins, erythrocourins, leghemoglobins and plant phycoyanins. As such, globins represent an interesting challenge for protein design. For example, can we optimize a sequence for myoglobin, which has poor specificity for hemoglobin? To study this specificity issue, we designed the whole sequence of sperm whale myoglobin, based on the structure of its backbone (extracted from the PDB file 5MBN), and its amino acid composition. The optimization required 40,000 MC step, using the complete energy function for sequence selection. The designed sequence, SD, and the native sequence, SN, of 5MBN are 22.9% identical. The THREADER Z-scores for SD and SN are 8.5 and 13, respectively; both values are significantly higher than random. This indicates that our designed sequence is specific for myoglobin, but is it specific enough, however, that it does not recognize hemoglobin?

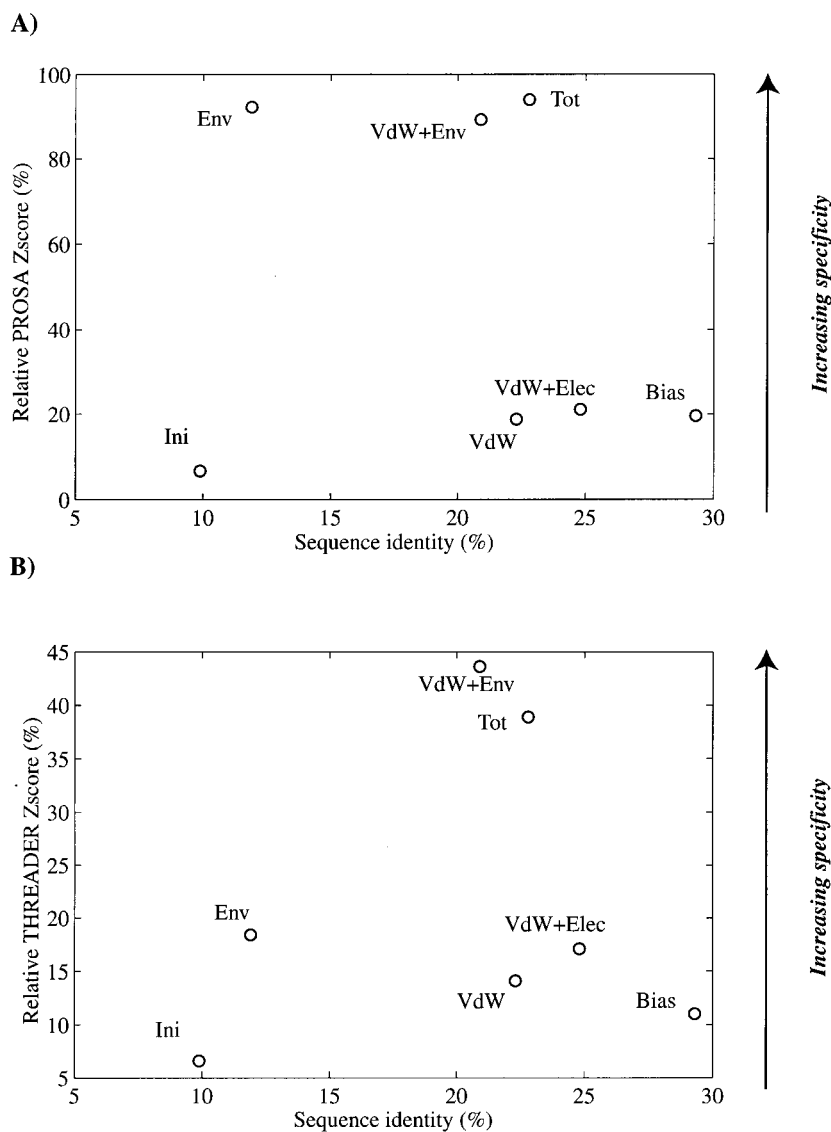


Figure 5. Testing of fold recognition techniques: ten independent sequence design experiments were carried out for protein GB1, using five different energy functions: vdW, env, vdW + env, vdW + elec, and Tot = vdW + elec + env. For reference, Ini corresponds to random sequences with the GB1 amino acid composition. A sixth set of ten optimization was carried out at infinite temperature over 100,000 cycles, in which the sequence with highest identity to the native sequence of protein G is selected; this set is defined as Bias, since it is based on the native sequence. The abilities of the designed sequences to recognize the native fold of protein G are plotted *versus* their similarities to the native sequence, for (a) PROSA and (b) THREADER. In all cases, the average value of the ratio of the Z-scores for the designed and native sequence is reported as a measure of specificity, while the average sequence identity to the native sequence is used to measure sequence similarity.

To answer this question, we extracted all 349 globin chains from the entire PDB database. The sequences of each of these globins were aligned with both the native and designed sequence, and the corresponding identity scores are plotted *versus* the structural similarities between the globin and 5MBN shown in Figure 6(a) and (b), respectively. The pattern observed with respect to the native sequence is clearly repeated for the designed sequence. Interestingly, while the native MBN sequence is between 80% and 100% identical with the myoglobins in the PDB and on average 25% identical with the other globins, these values become 25% and 15%, respectively, in the case of the designed sequence. Hence discrimination of myoglobins is observed at the sequence level, even though the scores are below the limits usually considered to be reliable. To analyze further the specificity issue, the native sequence (SN) was tested with THREADER, including all 349 globins in the library of folds. In Figure 6(c), Z-scores obtained

for each globin-SN pair are plotted against the coordinate RMS deviation between the corresponding globin and 5MBN (structure superposition was performed by STRUCTAL (Subbiah *et al.*, 1993), which allows gaps in both structures). The THREADER Z-score clearly separates myoglobins from the other globins, both for the native sequence and the designed sequence (Figure 6(d)). This is a good indication that the physical energy function we use in conjunction with the Monte Carlo procedure does succeed in designing a sequence stable in the myoglobin fold but with poor specificity to hemoglobin.

Comparison with experiments

The amino acids forming the hydrophobic core of a protein encode for its stability. Consequently, there has been considerable experimental work aimed at the stability contribution of individual residues within the core. These data provide a sen-

Table 3. Fold recognition as a measure of specificity

Energy type	(Seq. Ident.)	(PROSA-all)	(PROSA-pair) ^a	(PROSA-surf) ^b	(THREADER)
vdW	22.3	18.8	22.6	16.2	14.1
Env	11.9	92.2	58.1	108.3	18.4
vdW + Elec	24.8	21.2	34.6	13.1	17.1
vdW + Env	20.9	89.2	67.9	99.5	43.6
Tot	22.8	94	76.9	101.9	38.9
Random ^c	29.3	19.6	6.9	25.3	11

All values are averaged over ten independent computer designed sequences for GB1.

^a PROSA Z-score based on its pair potential only.

^b PROSA Z-score based on its mean force potential energy.

^c The "random" sequences have been derived from ten simulation of 100,000 Monte Carlo steps at infinite temperature, selecting the sequence with highest identity compared to the native sequence of GB1 for each simulation.

sitive test for our design procedures. Here we analyze our designed sequences and compare them with experimental results on protein core residues, both for protein G, our test protein, and lambda d repressor.

Protein G

Gronenborn *et al.* (1996) have created a library of core mutants of protein G (GB1), and tested by

NMR and fluorescence the stability and structure of a selected subset of this library. All mutants involve five residues, four of which are both in the core and in the β -sheets (Leu5, Leu7, Phe52 and Val54), while the fifth is a solvent-exposed residue used as a control (Ile6). The library of mutants consists of the native GB1 sequence in which one or more of these five residues have been replaced either by Leu, Val, Ile, Phe or Met. We performed ten independent sequence optimizations of GB1,

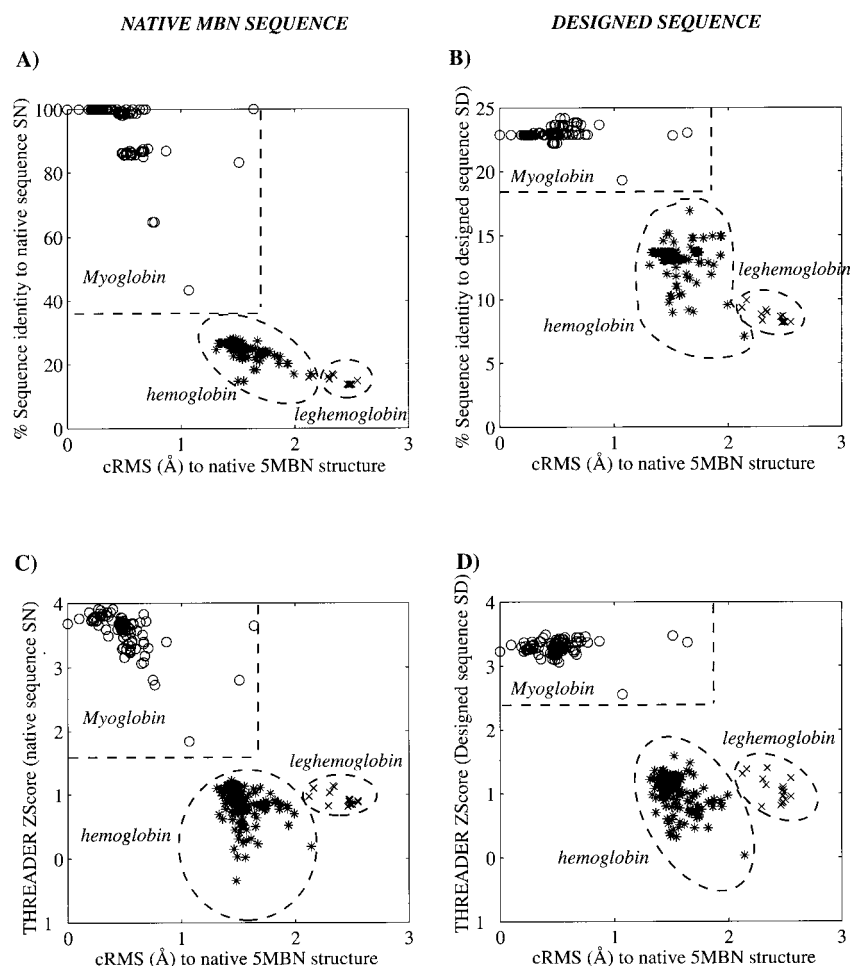


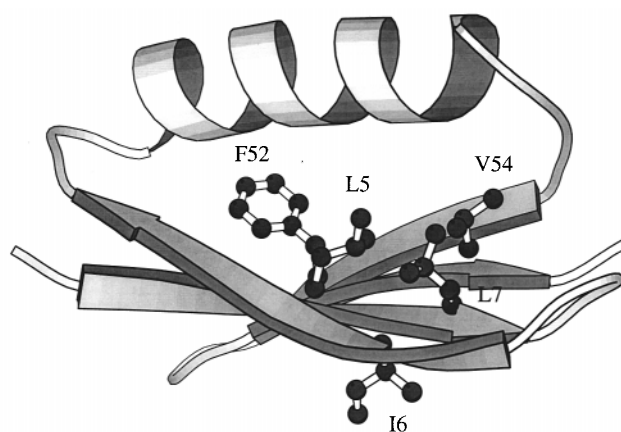
Figure 6. Each sequence of all 349 globins in the PDB database is aligned to the native sequence of (a) 5MBN, SN, as well as to a sequence designed for (b) 5MBN, SD, and the corresponding identity scores are plotted versus the cRMS deviation for the optimal alignment between the globin and the native structure 5MBN. Note the difference in the scale of the y -axis. The native sequence, SN, and the designed sequence, SD, for 5MBN were also threaded on the 349 different globin folds F contained in the PDB using THREADER. The fitness of SN and SD on each fold F are defined by Z-score, and these Z-scores are plotted versus the cRMS deviation for the optimal alignment between the fold and the native structure 5MBN in (c) and (d), respectively. Myoglobin folds are shown as circles (O), while hemoglobin and leghemoglobin folds are shown as stars (*) and crosses (X), respectively. Structural alignments were generated using STRUCTAL (Subbiah *et al.*, 1993).

starting from ten random shuffled sequences and using our complete energy function. In Figure 7, we compare the amino acids observed in the designed sequences at the five positions 5, 6, 7, 52 and 54, with the amino acids observed in the most stable mutants reported by Gronenborn *et al.* (1996). Positions 5, 7, 52 and 54 of GB1 are buried; they should consequently accommodate non-polar amino acids. Indeed, in our design sequences, Val, Leu, Ile, Phe and Met represent more than 60% of the amino acid residues observed at these positions (90% for 7 and 54, 80% for 52 and 60% for 5; see Table 4), which is in good agreement with the experimental mutation studies (Figure 7). Position 6 is exposed to solvent, and should therefore be polar. As expected, in our ten designed sequences for GB1, the five non-polar amino acids Leu, Val, Ile, Phe and Met represent only a small fraction of the amino acids observed at this position (30%; the other residues observed there are tyrosine and lysine, both hydrophilic residues). Among these five amino acid residues, only the presence of valine at position 6 is statistically significant (20%, compared to 7% for a random shuffling; Table 4), and it is interesting to note that this corresponds to the substitution observed when comparing the B1 and B2 domain of protein G (Achari *et al.*, 1992).

Lambda repressor

Sauer and co-workers (Lim & Sauer, 1989, 1991; Bowie *et al.*, 1990; Reidhaarolson *et al.*, 1991) have

used cassette mutagenesis to alter randomly sets of residues within the core of the N-terminal domain of phage lambda repressor. By selecting the resultant mutants for repressor function, they have identified residues compatible with each core position in the lambda repressor (Lim & Sauer, 1989). The comparison of these experimentally determined residues with the list of residues derived from our design procedure is given in Figure 8. Sequence design was performed on the backbone template taken from the PDB file 1LMB (chain 4). Ten independent optimizations were carried out, starting from ten different initial shuffled sequences. All 92 residues of lambda repressor were allowed to change using the complete energy function. For all seven core positions, the amino acid from the native sequence was detected in at least one of the designed sequences. In all but two cases (the tyrosine residue at position 47 and the alanine residue at position 51), the residues found in the optimized sequences were experimentally found to be compatible with the function of the protein. While there is no experimental evidence that residue 47 could not accommodate a tyrosine residue, it was shown that the single mutation F51 → A yields an inactive protein (Lim & Sauer, 1989). Note that our designed sequence containing an alanine residue at position 51 does not correspond to single point mutation. Experimentally, cysteine was found compatible with all seven core positions of the lambda repressor. This could not have been found in our designed sequences, since



Protein G Core Design

Position	5	6	7	52	54
Native	L	I	L	F	V
Experiment	L, M, V	L, V	F, I, L, V	F, L, V	L, V
Design	L, V	L, V	F, I, M	F, M, V	L, V

Figure 7. Protein design of the B1 domain of protein G: comparison of the predicted amino acid found at four core positions (5, 7, 52 and 54) and one exposed position (6) with those found in stable mutants of GB1. Results are given for five types of residues (Val, Ile, Leu, Phe and Met). The drawing of the protein was generated using MOLSCRIPT (Kraulis, 1991).

Table 4. Protein G core design: comparison with experiments

Position		5	6	7	52	54
Native		L	I	L	F	V
Amino acid	F	0 (4)	0 (5)	40 (4)	40 (4)	0 (4)
	I	0 (2)	0 (2)	20 (2)	0 (7)	0 (7)
	L	40 (5)	10 (5)	0 (5)	0 (7)	10 (5)
	M	0 (2)	0 (2)	30 (2)	10 (2)	0 (2)
	V	20 (7)	20 (7)	0 (7)	30 (7)	80 (7)
	Other	40 (80)	70 (79)	10 (80)	20 (73)	10 (75)

The occurrences (in %) at each of the five positions of each five amino acid type derived from ten independent design optimizations are compared with those expected from random permutation of the native sequence of GB1 (given in parenthesis).

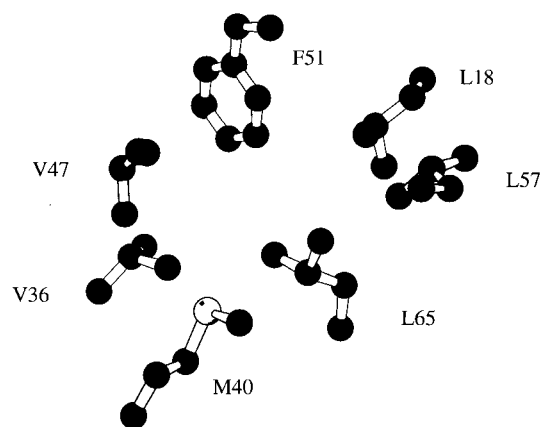
the native sequence for 1LMB4 does not contain a cysteine residue.

Discussion

Defining an energy function for protein design

Finding an energy function for inverse protein folding remains an active field of research, whose goal is to define a function for which the native structure of a protein is at the global minimum for the native sequence. An attractive idea for such an energy function is to use the increasing number of high-resolution protein structures available. This has led to the definition of potentials of mean

forces (Tanaka & Scheraga, 1976; Sippl, 1990), or knowledge-based scoring functions, which have attracted considerable interests in the recent past (see, for example, Sippl & Jaritz, 1994; Sippl, 1995; Moult, 1997; Skolnick *et al.*, 1997). Besides legitimate concerns about the physical basis of these function (Thomas & Dill, 1996; Ben-Naim, 1997), their applications to the specific problem of sequence design have been questioned (Rooman & Wodak, 1995). These energy terms are usually based on two-body contacts, and it was shown that a simple pairwise potential approximation is too crude to describe real proteins (Vendruscolo & Domany, 1998; Vendruscolo *et al.*, 1999). For all these reasons, we have decided to use an energy function based on physical terms only.



Lambda Repressor Core Design

Position	18	36	40	47	51	57	65
Native	L	V	M	V	F	L	L
Experiment	A, C, I, L, M, V	A, C, F, I, L, M, T, V	A, C, F, I, L, M, V	A, C, F, I, L, M, T, V	C, F, I, L, M, V	C, F, I, L, P, V	A, C, F, I, L, M, S, T, V
Design	A, L, M	I, L, M, V	I, L, M	L, M, V, Y	A, F, I, L, M	F, L, P	I, L, M

Figure 8. Tolerance for substitution observed at seven positions in the core of the N-terminal domain of the lambda repressor protein experimentally, and in protein design simulations. The schematic drawing of the relative positions of the seven residues was generated using MOLSCRIPT (Kraulis, 1991).

Our procedure for sequence design contains two steps: side-chain modeling using a self-consistent mean field approach, followed by a Monte Carlo protein sequence selection. Side-chain conformation prediction based on a known backbone template is mainly a geometrical problem, in that it aims to define the optimal side-chain packing with fewest close contacts. For this purpose, a van der Waals potential has been proven effective (Koehl & Delarue, 1994a). The sequence design problem is quite different. In earlier work (Koehl & Delarue, 1997), we have applied the sequence optimization scheme described above to the design of a chymotrypsin inhibitor (PDB code 2CI2), in which van der Waals only is used to drive the Monte Carlo computation in sequence space (Koehl & Delarue, 1997). The model with the final optimized protein sequence has better van der Waals packing interactions than the native structure, and its packing density is also higher than for native structure. The algorithm also positions glycine residues where they should be favored due to steric hindrance. However, the optimized sequence was shown to have very little specificity to the target 2CI2 fold (Koehl & Delarue, 1997). This was considered to be a consequence of the inadequacy of using only the van der Waals interactions to drive sequence design. This is in agreement with the conclusion reached by Behe *et al.* (1991), i.e. that packing is not the principal cause of conformational specificity. For these reasons, we used a more complete energy function that includes van der Waals for packing, electrostatics and solvent interactions.

Ultimately, the same energy function should be used for generating the model structure using SCMF, and for sequence selection using Monte Carlo, and this is under study. It should be noted, however, that side-chain modeling using SCMF based on vdW only has the advantage to be fast, and that it was shown to be as accurate as other methods based on more general energy functions (Koehl & Delarue, 1994a). Inclusion of electrostatics and a term for solvent interaction is therefore not expected to change the results.

Packing interactions

“Knowledge-based” protein design experiments have suggested that it is relatively easy to generate proteins which have approximately the correct target fold, but hard to design the specificity which ensures a stable, unique fold (for a review, see Hellinga, 1998a). This indicates that correct side-chain packing plays an important role in defining the specific native fold of a protein. Detailed information on packing can only be obtained by using a full atom representation of the protein, in which case the energy function must include a term for measuring geometric overlap. This is usually handled by the repulsive part of the Lennard-Jones potential for van der Waals interactions. Potential energy functions based on vdW interactions only have proven useful for predicting the conformation

of all side-chains of a protein, based on a given backbone. This term was consequently used here for the generation of a model structure based on the target backbone and a given sequence. The results described in Table 1 show, however, that sequence optimization protocol cannot rely solely on vdW interaction. This is in agreement with the experimental finding that local packing in the vicinity of β -sheets are not optimal in order to preserve the backbone-to-backbone hydrogen bond patterns within secondary structures (Schultz-Beardsley & Kauzmann, 1996).

All protein models considered here are based on a discrete representation of the side-chains as rotamers. While this approximation greatly reduces computing time by artificially reducing the search in conformational space, it does sometimes introduce steric clashes. Rigidly defined rotamers have been shown to mislead the analysis of allowed residues in a given protein site if a classical vdW potential is used (Koehl & Delarue, 1994a). One approach to this problem is to soften the Lennard-Jones potential (Levitt, 1976) by introducing an upper limit for its value. Such an approximation has proven valuable for side-chain prediction (Koehl & Delarue, 1994a); here we use it and show that it is also useful for protein design. It is worth mentioning that all sequence design procedures using rotamers are based on a single backbone. We are currently working on that problem.

Solvent interactions

Hydrophobic interaction are generally recognized as the major forces of protein stability (Dill, 1990); their quantification, however, remain an open question. The most popular model for the hydrophobic interactions is the so-called solvation free energy model introduced by Eisenberg & McLachlan (1986). Although the underlying assumptions to this model (such as the linear relationship relating atomic surface area and solvation free energy) have been questioned (Wood & Thompson, 1990; Ben-Naim, 1994; Simonson & Brunger, 1994), it has proven useful for testing protein models (Chiche *et al.*, 1990; Holm & Sander, 1992) as well as in molecular dynamics simulations (Wesson & Eisenberg, 1992; Schiffer *et al.*, 1992, 1993). The solvation free energy is designed to implicitly account for interaction with the solvent and, as such, does not include any information on internal contacts in the protein. It can, however, be easily modified to an environment energy E_{env} that considers the complete environment of each atom (Koehl & Delarue, 1994b). We have shown here that E_{env} is able to generate an hydrophobic/hydrophilic pattern when applied to the problem of protein design. It is worth noting that it provides a poor amino acid specificity by itself (see Table 3), and does not account for packing interactions.

Models for hydrophobic interaction based on atomic surface areas or volumes have the draw-

back that they cannot be represented by a sum of pair-wise energy terms. Most of the efficient methods for discrete conformational searches however require pair-wise additive energy functions. Recognition of this problem has led to the development of pair-wise approximations both for surface area calculations (Street & Mayo, 1998) and volume calculations (Augspurger & Scheraga, 1996). The Monte Carlo procedure we have chosen does not require the pair-wise approximation, since it can directly consider a global energy term such as the environment energy considered here.

Electrostatics

We believe that electrostatics is an important part of any potential energy function used for protein sequence design that involves both core and exposed residues. Electrostatics should account for interaction within the protein, as well as interaction with the solvent. The energy function used in this work is a first approximation to this problem: the solvent is considered implicitly by setting a screening factor in the Coulomb equation, i.e. defining the dielectric "constant" to be a linear function of the distance r . There is no physical reason for this modification and the screening effect it provides is quite crude. This approximation has been used before in several force fields and our partial charges for the Coulomb interactions are derived from the standard polar hydrogen parameters (param19) of CHARMM (Brooks *et al.*, 1993; Neria *et al.*, 1996). From a practical point of view, the r -dielectric has the advantage of being computationally efficient as it does not require square-root evaluation. It should be mentioned that E_{env} does include an electrostatics component, since it takes in account the nature of inter-atomic contacts in the protein so there may be some double counting. We show here that even this crude energy function improves the results of our design strategy and are currently testing other forms for the electrostatics potential.

Stability and specificity of the designed sequences

While the energy function is used to provide stability, it is crucial that the optimization process include a foldability criterion: the designed sequence should fold to the specific native structure. The straightforward solution to this problem is to evaluate the effect of each "mutation" process on the folding capacity of the sequence. This approach requires complete exploration of possible conformations and cannot be used for anything more than simple lattice models. Here we use an approximate method to obtain this folding specificity, based on the random energy model (REM), which can be implemented by keeping the amino acid composition constant (Shakhnovich & Gutin, 1993a,b). A major feature of this approach is that it is computationally feasible, even in the case of full

atom representations. The dimension of the sequence space is considerably reduced (from 20^N to $N!/\prod_{j=1}^{20} n(j)!$, where N is the length of the sequence, and $n(j)$ the number of amino acid of type j in the native sequence for the target structure).

The REM is characterized by a statistical independence of states (for a review, see Pande *et al.*, 1997). As a consequence, the REM energy spectrum consists of a continuous part that is independent of disorder, or sequence, and a few discrete energy levels that are placed very individually, and are sequence specific. In that context, design based on energy optimization for the target conformation should "pull down" a single energy level (the continuous part is not affected, since it is independent of sequence), resulting in the apparition of a large energy gap, required for stability (Shakhnovich, 1998). This selective pull down provides specificity. REM is, however, not universally valid, and in fact cannot be exact for heteropolymers (Pande *et al.*, 1997). For example, it can be violated in cases of long-range interactions, such as Coulomb interactions between charges (Pande *et al.*, 1996). In the case of proteins, however, electrostatics interactions can be considered, as a first crude approximation, more as a short range than a long-range interaction, due to solvent screening. This was taken in account in the Coulomb term included in our energy function, which considers a r -dependent dielectric constant in order to significantly decrease its range. Application of the REM to protein sequence design will also not work if it is not possible to pull down one structure. This is the case for example in the HP model, because the "native" structure in that case is not unique, as shown by the collaboration of the groups of Dill and Shakhnovich (Yue *et al.*, 1995). In this challenge, the Shakhnovich group designed a series of HP sequences for a given lattice configurations using Monte Carlo in sequence space with fixed amino acid composition. These sequences were then folded by the group of Ken Dill, and it was found that their ground state generally differ from their targets that were imposed. This failure of fixed composition was assigned to the simplified alphabet they used.

Although REM cannot be exact for proteins, it is a very good approximation. The problem of the simplified alphabet of HP model was solved by Shakhnovich (1994), by using a greater number of types of amino acid residues. In this work, we have applied REM on full atom protein design, using a physical potential in which long-range interactions have been screened, and our results suggest that it did succeed in designing specificity (see, for example, Figure 6). At this stage, however, specificity has only been evidenced by computer simulation, and a direct experimental test of our designed sequences is planned.

The amino acid composition of a protein is highly correlated with its folding class (see, for example, Chou, 1995). It is therefore intuitive to

maintain this composition constant: we know the structure of the template backbone, hence we know the folding class. This procedure, however, introduces a limitation: a given protein sequence may not include all 20 amino acid residues, and the missing types will never be considered in our sequence design procedure. We are also working on that problem.

Concluding Remarks

Three conclusions can be drawn from this study. Firstly, protein design using all-atom models is computationally feasible. Secondly, sequences can be designed to be stable and specific to their target conformation using a simple physical energy function. Thirdly, optimization in sequence and conformational spaces can be performed sequentially. For a given sequence, a full atom model is built based on the target conformation, using mean field theory and a simple pair-wise energy function, and the energy (non-necessarily pair-wise) of the optimized model is then used to drive the design in sequence space.

Methods

Off lattice Monte Carlo protein sequence optimization

The procedure for protein sequence optimization is outlined in Figure 9. Each step is detailed below.

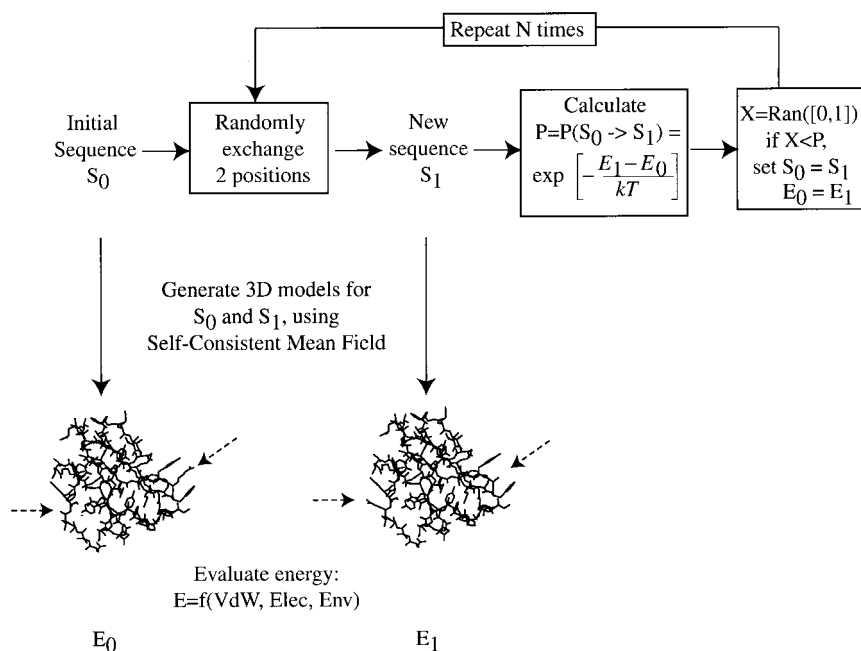


Figure 9. Schematic description of the method. In the illustration of the amino acid swap, the two residues involved in the swap are highlighted by broken-line arrows. Ran corresponds to a uniform distribution.

Sequence optimization

The designed sequence must be stable in the known native conformation, and also be specific to that structure, i.e. incompatible with competing folds; these two conditions are dealt with in the so-called design in and design out procedures, respectively. The problem can be reformulated as finding the sequence, S , such that it has a high probability, P , to be in the template conformation, C_{nat} at room temperature. P is given by:

$$P = \frac{\exp\left[-\frac{E(C_{\text{nat}}, S)}{kT}\right]}{\sum_C \exp\left[-\frac{E(C, S)}{kT}\right]} \quad (1)$$

$E(C, S)$ is the energy of sequence S in conformation C , T is the temperature and k is the Boltzmann constant. The denominator in equation (1) corresponds to a partition function, Z . A rigorous approach to the problem of maximizing the probability P would require simultaneous and complete explorations of all of sequence space and conformation spaces.

The random energy model as described in the canonical ensemble optimization procedure, makes the problem computationally more tractable (Shakhnovich & Gutin, 1993a,b). In this approach, the partition function Z (denominator of equation (1)) is assumed to depend only on the amino acid composition and not on the ordered sequence itself. Given this approximation, specificity can be achieved by optimization in sequence space alone, provided that the amino acid composition of the sequence is held constant.

Starting from a random sequence, S_0 , of the required composition, a model structure is built, and its energy is evaluated and stored as E_0 . Two positions are then chosen at random, and the corresponding amino acid types in S_0 are exchanged, yielding a new sequence S_1 . A new

model structure is built based on S_1 , and its energy is stored as E_1 . The sequence move from S_0 to S_1 is accepted or rejected according to the Metropolis Monte Carlo probability (Metropolis *et al.*, 1953) given by:

$$P\{S_0 \rightarrow S_1\} = \begin{cases} e^{-E_1 - E_0/T_{MC}} & \text{if } E_1 - E_0 > gT_0 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where T_{MC} is a parameter usually referred to as the temperature of the Monte Carlo simulation. The procedure is repeated until the system has equilibrated and the energy remains steady.

Full atomic model for the test sequences

The current sequence in the Monte Carlo optimization is threaded on the backbone template of the known native structure, and side-chains are positioned using an iterative self-consistent mean field approach (Koehl & Delarue, 1994a). This method is based on a rotamer library of side-chain conformations. We use a modified version of the averaged library by Tuffery *et al.* (1991), which has been corrected for duplicate rotamers. Three additional rotamers obtained from the library by Ponder & Richards (1987) are added for proline residues. The backbone-independent rotamer library by Dunbrack & Cohen (1997) was used for leucine residues. The procedure iteratively refines a conformational matrix, \mathbf{CM} , of the side-chains of the protein such that its current element at each cycle $\mathbf{CM}(i,j)$ is the probability that side-chain i of the protein adopts the conformation of its possible rotamer j . Use of the mean field implies that each residue feels the average energy of all possible environments, weighted by their respective probabilities. Interactions and hence probabilities depend solely on a Lennard-Jones function for vdW interactions (electrostatics and solvent interactions are ignored). The procedure converges in a few cycles. The rotamer with the highest probability in the optimized conformational matrix is used to define the conformation of the side-chain in the final model.

Sequence-structure fitness

The thermodynamic stability of a sequence S is measured by the difference in free energies between its native state, N , and an unfolded state, U :

$$\Delta G_{U \rightarrow N}(S) = G_N(S) - G_U(S) \quad (3)$$

where the total free energy G can be partitioned into:

$$G(S) = G^{\text{bon}}(S) + G^{\text{nb}}(S) + G^{\text{env}}(S) + G^{\text{ent}}(S) \quad (4)$$

$G^{\text{bon}}(S)$, $G^{\text{nb}}(S)$, $G^{\text{env}}(S)$ and $G^{\text{ent}}(S)$ are the covalently bonded, the non-bonded, the environment interactions and entropy, respectively. The energy difference between two sequences S_0 and S_1 is given by:

$$\Delta \Delta G_{U \rightarrow N}(S_0 \rightarrow S_1) = \Delta G_{U \rightarrow N}(S_1) - \Delta G_{U \rightarrow N}(S_0) \quad (5)$$

where N is the template structure used for design.

Using equation (2), the same difference can be rewritten as:

$$\Delta \Delta G(S_0 \rightarrow S_1) = (G_N(S_1) - G_N(S_0)) - (G_U(S_1) - G_U(S_0)) \quad (6)$$

Equation (5) is the natural free energy difference, while equation (6) is a mathematical expression not related to physical chemistry, but directly applicable to a computer calculation. Each term in equation (6) can be partitioned into the four types of interaction described by equation (4), yielding an exact expression for the difference of energy between the two models built from sequences S_0 and S_1 .

Use of equation (6) requires that the free energy of the sequence be computed in the "unfolded" state, which is usually taken to be the fully extended structure. The denatured "state" of a protein is known to be a distribution of different molecular conformations, and the extended conformation is certainly a poor model of this state (for a review, see Dill & Shortle, 1991). Though inter-residue contacts do exist in the denatured states, it is fair to assume that most of the energy is dominated by local interactions. This energy mainly depends on the amino acid composition of the sequence, rather than the sequence itself. In the case of a canonical sequence optimization with fixed amino acid composition, this leads to:

$$G_U(S_1) \approx G_U(S_0) \quad (7)$$

for all test sequences S_0 and S_1 . As a consequence, the denatured states have little influence in the design process if the sequence composition is kept fixed. Macro-canonical optimization (i.e. optimization with unconstrained amino acid composition) requires more care (Sun *et al.*, 1995; Deutsch & Kurosky, 1996; Seno *et al.*, 1996; Vendruscolo *et al.*, 1997) and is not considered here.

Based on this definition of the denatured state, let us review the various components of equation (6).

Bonded interactions. The contribution of the bonded interactions is given by:

$$\Delta \Delta G^{\text{bon}}(S_0 \rightarrow S_1) = (G_N^{\text{bon}}(S_1) - G_N^{\text{bon}}(S_0)) - (G_U^{\text{bon}}(S_1) - G_U^{\text{bon}}(S_0)) \quad (8)$$

The bonded interactions are local interactions, and to a first approximation, only depend on the amino acid composition of the sequence of the protein of interest. In the framework of the canonical optimization with fixed amino acid composition, bonded interactions do not change on folding:

$$\Delta \Delta G^{\text{bon}}(S_0 \rightarrow S_1) = 0 \quad (9)$$

Non-bonded interactions. The non-bonded interactions are described by the sum of a Lennard-Jones potential and a Coulomb potential for vdW and electrostatics interactions, respectively. The Lennard Jones potential of a protein of sequence S and conformation C is given by:

$$E_{vdW}(S, C) = \sum_i \sum_{j < i} \varepsilon_{ij} \left[\left(\frac{r_{ij}^o}{r_{ij}} \right)^{12} - 2 \left(\frac{r_{ij}^o}{r_{ij}} \right)^6 \right] \quad (10)$$

where the summation extends over all pairs of atom (i,j) , r_{ij} is the inter-atomic distance between i and j , and ε_{ij} and r_{ij}^o are constants that depend on the chemical nature of i and j . No scaling of the 1-4 interactions was performed. E_{vdW} as given by equation (10) is also used in

the SCMF procedure for calculating the model side-chain conformations (see above).

The electrostatics potential is given by:

$$E_{\text{elec}}(S, C) = \sum_i \sum_{j < i} \frac{q_i q_j}{D_r r_{ij}} \quad (11)$$

where the summation extends over all pairs of atom (i, j) . r_{ij} is the inter-atomic distance, and q_i and q_j are the partial charges of i and j , respectively. The solvent plays a significant role in determining the electrostatic energy of a protein, most notably through a screening of the electrostatic interactions. As a first approximation, this screening is included in the calculation by damping E_{elec} with a distance dependent dielectric constant, so that:

$$D_r = 4r_{ij} \quad (12)$$

The total non-bonded free energy difference between sequences S_0 and S_1 adopting the same conformation N is given by:

$$\Delta \Delta G^{\text{nb}}(S_0 - S_1) = E_{\text{vdW}}(S_1, N) + E_{\text{elec}}(S_1, N) - E_{\text{vdW}}(S_0, N) - E_{\text{elec}}(S_0, N) \quad (13)$$

as the non-bonded energy of the denatured states are assumed to cancel.

The environment free energy. One of the most appealing quantification of the solvation interactions empirically relates the solvation free energy to the accessible surface area (Eisenberg & McLachlan, 1986). In this scheme, the solvation free energy is defined as:

$$G_S = \sum_i \text{ASP}_i \text{ASA}_i \quad (14)$$

where the sum extends over all atoms i of the protein. ASA_i is the accessible surface area of atom i , and ASP_i is the corresponding atomic solvation parameters that converts surface area into energy. The ASP parameters are derived from experimental data on free energy of transfer of amino acid analogs from octanol to water (Fauchere & Pliska, 1983; Sharp *et al.*, 1991). This model was recently generalized in order to take into account contributions not only from protein/solvent interactions, but also from internal protein/protein interactions (Koehl & Delarue, 1994b). In the original Eisenberg & McLachlan (1986) formalism the surface area of polar atoms buried upon folding yields an unfavorable contribution to the free energy, irrespective of the environment of the polar atoms in the core of the folded protein. However, in the core of a protein, polar atoms may be in contact with other polar atoms. Intuitively, this is not as unfavorable as having the polar atoms in contact with non-polar atoms. Similarly, in the original formalism non-polar atoms buried upon folding always contribute favorably to the free energy even if they are in contact with polar atoms inside the protein and this also seems counter intuitive. For that reason, Koehl & Delarue (1994b) introduced a free energy of environment, G^{env} , in order to take account for the full environment for each atom of the protein (solvent and other protein atoms):

$$G^{\text{env}} = \sum_i [A_i(\text{ASA}_i + \text{PCA}_i) + B_i \text{NPCA}_i] \quad (15)$$

where the summation extends over all atoms i , PCA_i and

NPCA_i are the surface areas of atom i occluded by polar and non-polar atoms, respectively (also described as the polar and non-polar contact area of i). A_i and B_i are surface tension factors similar to the atomic solvation parameters, ASP. Let us define TASA_i the total accessible surface area of atom i in the presence of local interaction only:

$$\text{TASA}_i = \text{ASA}_i + \text{PCA}_i + \text{NPCA}_i \quad (16)$$

The environment free energies of sequence S in the target structure N and in the denatured state are therefore given by:

$$G_N^{\text{env}}(S) = \sum_i (B_i - A_i) \text{NPCA}_i + \sum_i A_i \text{TASA}_i \quad (17)$$

and:

$$G_U^{\text{env}}(S) = \sum_i A_i \text{TASA}_i \quad (18)$$

For two sequences S_0 and S_1 having the same amino acid composition:

$$\begin{aligned} \Delta \Delta G^{\text{env}}(S_0 \rightarrow S_1) &= \sum_{i \in S_1} (B_i - A_i) \text{NPCA}_i \\ &\quad - \sum_{i \in S_0} (B_i - A_i) \text{NPCA}_i \end{aligned} \quad (19)$$

Entropy. The loss of entropy upon formation of the folded protein is an important contribution to the stability of a protein. To a first approximation, the entropy in the highly ordered folded state is assumed to be small compared to the entropy in the unfolded state. Consequently:

$$\Delta \Delta G^{\text{ent}}(S_0 \rightarrow S_1) = G_U^{\text{ent}}(S_1) - G_U^{\text{ent}}(S_0) \quad (20)$$

Within the framework of canonical sequence design, this leads to:

$$\Delta \Delta G^{\text{ent}}(S_0 \rightarrow S_1) = 0 \quad (21)$$

A computer program for protein sequence optimization

Protein backbone

The coordinates of the backbone atoms (including C^β) of the desired target structure are given as input. These coordinates are derived from the PDB file describing the native structure of the protein of interest. For residue positions occupied by a glycine residue in the native sequence, a C^β atom is built using standard geometry. All other side-chain atoms are discarded. All main-chain atoms are fixed in space in all subsequent calculation.

Computer implementation

The self consistent mean field approach for side-chain conformation prediction is based on a fixed set of rotamers, whose positions within the framework of the target structure can be pre-computed once, for all positions and all types of amino acid. In our initial implementation

for sequence design (Delarue & Koehl, 1996) all interactions between side-chains were also pre-computed, yielding a large but efficient energy matrix which was then used at each step of the Monte Carlo sequence design procedure for side-chain placement. While this method works for small chains (<100 residues), it becomes impractical for larger chains because of the size of the energy matrix (for a protein length of 200 residues, with a rotamer library including 100 conformations for all 20 amino acids, the full interaction matrix would require 1.6 GBytes of memory). In the current implementation, the energy matrix is built based on the initial sequence in the Monte Carlo procedure, and updated each time an amino acid residue is changed. Since only pair interactions are included for side-chain placement, the update only requires that the rows and columns of the matrix corresponding to the position considered be modified.

The same update procedure is used for the non-covalent part of the sequence-structure fitness function. The latter, however, also includes an environment term involving surface area calculations, which needs to be computed completely for all subsequent models, since it is not a pair-wise calculation. We use the very efficient algorithm by Legrand & Merz (1993) to calculate surface area.

Our procedure requires two seconds for building a full atomic model of a 150-residue sequence threaded into the target structure and evaluating the total energy of this model (on a Compaq Alpha processor at 533 MHz). The computing time grows linearly with the size of the protein.

Parameters

Parameters for the vdW and electrostatics calculation were taken from CHARMM19 (Brooks *et al.*, 1983). The Lennard-Jones potential described by equation (9) becomes infinite when the inter-atomic distance r tends to 0. This could be a problem in the formalism used here, since side-chains are positioned according to fixed rotamer orientations and clashes can occur. To remove the consequent infinite energy barriers, E_{vdW} and E_{elec} were truncated to a maximum value of 10 kcal/mol, as described by Levitt (1983). The atomic solvation parameters ($A_i - B_i$) in equation (19) were taken from Koehl & Delarue (1994b), and multiplied by 20 for proper scaling with the other terms in the energy function.

Unless specified otherwise, the Monte Carlo "temperature" T was set to 2, and the design was performed over 40,000 cycles for all proteins considered. In all cases described here, this led to a rejection rate of 95% over the 40,000 cycles.

Probing sequence-structure specificity

The issue of specificity can be addressed by protein fold recognition techniques, which usually rely on "hide-and-see" computer experiments (Hendlich *et al.*, 1990; Sippl & Weitckus, 1992). In this procedure, the target structure X for a given sequence, S , is hidden among a large number of non-native folds, C , and the task is to retrieve X using an energy criteria. Success is achieved if the energy of S threaded on X , $E(S,X)$, is lower than any $E(S,C)$. A measure of this success is provided by a Z -score (Bowie *et al.*, 1991):

$$Z = \frac{E(S, X) - \langle E(S, C) \rangle}{\sigma} \quad (22)$$

where $\langle \rangle$ stands for the average over all conformations C , and σ is the corresponding standard deviation. If the discrimination is significant, Z is expected to be negative and large.

We have used PROSA (Sippl, 1993) and THREADER (Jones *et al.*, 1992) for these tests. PROSA performs threading without gaps in the sequence or in the model structure. The sequence of interest is threaded on a large ensemble of conformations (the so-called "polyprotein"; Sippl & Jaritz, 1994), and each sequence-structure match is given a score based on a combined potential of mean force, including a pair-wise, distance-dependent contact potential, as well as a surface potential. All scores are transformed into Z -score based on equation (22). THREADER aligns the sequence of interest on a library of folds, based on a double dynamic programming algorithm, which allows for insertions as well as deletions in the sequence. The default library of fold provided with THREADER is CATH. Z -scores defined by THREADER have opposite signs compared to equation (22).

In both PROSA and THREADER, the compatibility score between a sequence and a given fold is computed directly from a given structural model. Alternatively, the protein structure information can be reduced into one dimension, yielding the so-called 3D-1D profile. In this approach, a position and structure dependent scoring table is built, which contains as many rows as residues in the structure, and a column for each of the 20 amino acid residues. Aligning a sequence on such a scoring table, or profile, resorts to dynamic programming methods developed for pair-wise sequence comparisons. We have used two variants of this technique, based on two different scoring tables: (i) the score function is derived from the GONNET (Gonnet *et al.*, 1992) substitution matrix, plus secondary structure information; this was originally proposed by Fisher & Eisenberg (1996); (ii) the scoring matrix is based on sequence information, secondary structure information, and solvent accessibility information; it was specifically computed from a database of structural pairs with low sequence similarity, in order to improve recognition of distantly related sequence-structure pairs; this is the H3P2 matrix by Rice & Eisenberg (1997). This procedure makes use of the secondary structure of the query sequence, predicted using the PHD program by Rost & Sander (1993).

Both methods are available online at the site <http://fold.doe-mbi.ucla.edu>.

Protein sequence database search

Protein fold recognition can be performed solely at the sequence level by comparing the test sequence to a database of known protein sequences. For that purpose, we have used FASTA (Lipman & Pearson, 1985) on the non-redundant SWISSPROT database (Bairoch & Boeckmann, 1991; Bairoch & Apweiler, 1999), release of May 1998. Individual alignments are performed with an opening and extension penalties for gaps of -12 and -2 , respectively. The Blosum50 substitution matrix (Henikoff & Henikoff, 1992) is used for scoring.

Structural alignments

We have used STRUCTAL (Subbiah *et al.*, 1993) for protein structure superposition. This method starts with an arbitrary equivalence of the residues of the two pro-

teins. This equivalence is used to perform a classical superposition of the two structures, from which a structural alignment matrix SA is calculated. The best structural alignment is then obtained by standard global dynamic programming on SA. Since the alignment may depend on the initial residue equivalence, the procedure is repeated for five different initial sets of correspondence, and the optimal alignment is taken as that with the highest score.

Acknowledgments

This work was supported by National Institutes of Health grant GM45415 to M.L. The research was carried out while P.K. was on leave of absence from the CNRS institute, Strasbourg, France, partially funded by a long term fellowship from the Union Internationale Contre le Cancer, Geneva, Switzerland.

References

- Achari, A., Hale, S. P., Howard, A. J., Clore, G. M., Gronenborn, A. M., Hardman, K. D. & Whitlow, M. (1992). 1.67-angstrom X-ray structure of the B2 immunoglobulin-binding domain of streptococcal protein-G and comparison to the NMR structure of the B1 domain. *Biochemistry*, **31**, 10449-10457.
- Anfinsen, C. (1973). Principles that govern the folding of protein chains. *Science*, **181**, 223-230.
- Arnold, F. H. (1998a). Design by directed evolution. *Acc. Chem. Res.* **31**, 125-131.
- Arnold, F. H. (1998b). When blind is better: protein design by evolution. *Nature Biotechnol.* **16**, 617-618.
- Arnold, F. H. & Haymore, B. L. (1991). Engineered metal-binding proteins: purification to protein folding. *Science*, **252**, 1796-1797.
- Augsburger, J. D. & Scheraga, H. A. (1996). An efficient, differentiable hydration potential for peptides and proteins. *J. Comp. Chem.* **17**, 1549-1558.
- Bairoch, A. & Apweiler, R. (1999). The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucl. Acids Res.* **27**, 49-54.
- Bairoch, A. & Boeckmann, B. (1991). The Swiss-Prot protein-sequence data-bank. *Nucl. Acids Res.* **19**, 2247-2248.
- Bashford, D., Chothia, C. & Lesk, A. M. (1987). Determinants of a protein fold: unique features of the globin amino-acid-sequences. *J. Mol. Biol.* **196**, 199-216.
- Behe, M. J., Lattman, E. E. & Rose, G. D. (1991). The protein-folding problem: the native fold determines packing, but does packing determine the native fold?. *Proc. Natl Acad. Sci. USA*, **88**, 4195-4199.
- Ben-Naim, A. (1994). Solvation of large molecules: some exact results on the dependence on volume and surface-area of the solute. *Biophys. Chem.* **51**, 203-216.
- Ben-Naim, A. (1997). Statistical potentials extracted from protein structures: are these meaningful potentials? *J. Chem. Phys.* **107**, 3698-3706.
- Bowie, J. U., Reidhaarolson, J. F., Lim, W. A. & Sauer, R. T. (1990). Deciphering the message in protein sequences: tolerance to amino-acid substitutions. *Science*, **247**, 1306-1310.
- Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164-170.
- Brooks, B., Brucoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). CHARMM: a program for macromolecular energy minimization and dynamics calculations. *J. Comput. Chem.* **4**, 187-217.
- Bryson, J. W., Betz, S. F., Lu, H. S., Suich, D. J., Zhou, H. X., O'Neil, K. T. & DeGrado, W. F. (1995). Protein design: a hierarchic approach. *Science*, **270**, 935-941.
- Cao, A. N., Lai, L. H. & Tang, Y. Q. (1998). The current state and prospect of *de-novo* protein design. *Prog. Biochem. Biophys.* **25**, 197-201.
- Chiche, L., Gregoret, L. M., Cohen, F. E. & Kollman, P. A. (1990). Protein model structure evaluation using the solvation free-energy of folding. *Proc. Natl Acad. Sci. USA*, **87**, 3240-3243.
- Chou, K. C. (1995). A novel-approach to predicting protein structural classes in a (20-1)-D amino-acid-composition space. *Proteins: Struct. Funct. Genet.* **21**, 319-344.
- Chowdhury, P. S., Vasmatzis, G., Lee, B. & Pastan, I. (1998). Improved stability and yield of a Fv-toxin fusion protein by computer design and protein engineering of the Fv. *J. Mol. Biol.* **281**, 917-928.
- Coldren, C. D., Hellinga, H. W. & Caradonna, J. P. (1997). The rational design and construction of a cuboidal iron-sulfur protein. *Proc. Natl Acad. Sci. USA*, **94**, 6635-6640.
- Dahiyat, B. I. & Mayo, S. L. (1996). Protein design automation. *Protein Sci.* **5**, 895-903.
- Dahiyat, B. I. & Mayo, S. L. (1997a). *De-novo* protein design: fully automated sequence selection. *Science*, **278**, 82-87.
- Dahiyat, B. I. & Mayo, S. L. (1997b). Probing the role of packing specificity in protein design. *Proc. Natl Acad. Sci. USA*, **94**, 10172-10177.
- Dahiyat, B. I., Gordon, D. B. & Mayo, S. L. (1997). Automated design of the surface positions of protein helices. *Protein Sci.* **6**, 1333-1337.
- Dalal, S., Balasubramanian, S. & Regan, L. (1997). Protein alchemy: changing beta-sheet into alpha-helix. *Nature Struct. Biol.* **4**, 548-552.
- Delarue, M. & Koehl, P. (1997). The inverse protein folding problem: self consistent mean field optimisation of a structure specific mutation matrix. In *Proceedings of the Pacific Symposium on Biocomputing*, 1997 (Altman, R., Dunker, A., Hunter, L. & Klein, T., eds), pp. 109-121, World Scientific, Singapore.
- Deutsch, J. M. & Kurosky, T. (1996). New algorithm for protein design. *Phys. Rev. Letters*, **76**, 323-326.
- Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, **29**, 7133-7155.
- Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nature Struct. Biol.* **4**, 1-10.
- Dill, K. A. & Shortle, D. (1991). Denatured states of proteins. *Annu. Rev. Biochem.* **60**, 795-825.
- Dill, K. A., Bromberg, S., Yue, K. Z., Fiebig, K. M., Yee, D. P., Thomas, P. D. & Chan, H. S. (1995). Principles of protein folding - a perspective from simple exact models. *Protein Sci.* **4**, 561-602.
- Dobson, C. M., Sali, A. & Karplus, M. (1998). Protein-folding: a perspective from theory and experiment. *Angew. Chem. Internat. Edit.* **37**, 868-893.
- Drexler, K. E. (1981). Molecular engineering: an approach to the development of general capabilities for molecular manipulation. *Proc. Natl Acad. Sci. USA*, **78**, 5275-5278.

- Dunbrack, R. L. & Cohen, F. E. (1997). Bayesian statistical-analysis of protein side-chain rotamer preferences. *Protein Sci.* **6**, 1661-1681.
- Eisenberg, D. & McLachlan, A. (1986). Solvation energy in protein folding and binding. *Nature*, **319**, 199-203.
- Elhawrani, A. S., Moreton, K. M., Sessions, R. B., Clarke, A. R. & Holbrook, J. J. (1994). Engineering surface loops of proteins: a preferred strategy for obtaining new enzyme function. *Trends Biotechnol.* **12**, 207-211.
- Farinas, E. & Regan, L. (1998). The *de-novo* design of a rubredoxin-like Fe site. *Protein Sci.* **7**, 1939-1946.
- Fauchere, J.-L. & Pliska, V. (1983). Hydrophobic parameters π of amino acid side-chains from the partitioning of N-acetyl-amino-acid amides. *Eur. J. Med. Chem-Chim Therap.* **18**, 369-375.
- Fischer, D. & Eisenberg, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947-955.
- Fischer, D., Rice, D., Bowie, J. U. & Eisenberg, D. (1996). Assigning amino-acid-sequences to 3-dimensional protein folds. *FASEB J.* **10**, 126-136.
- Gallagher, T., Alexander, P., Bryan, P. & Gilliland, G. L. (1994). Two crystal structures of the B1 immunoglobulin-binding domain of streptococcal protein G and comparison with NMR. *Biochemistry*, **33**, 4721-4729.
- Giver, L. & Arnold, F. H. (1998). Combinatorial protein design by *in-vitro* recombination. *Curr. Opin. Chem. Biol.* **2**, 335-338.
- Gonnet, G. H., Cohen, M. A. & Benner, S. A. (1992). Exhaustive matching of the entire protein-sequence database. *Science*, **256**, 1443-1445.
- Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355-4358.
- Gronenborn, A. M., Filpula, D. R., Essig, N. Z., Achari, A., Whitlow, M., Wingfield, P. T. & Clore, G. M. (1991). A novel highly stable fold of the immunoglobulin binding domain of streptococcal protein-G. *Science*, **253**, 657-661.
- Gronenborn, A. M., Frank, M. K. & Clore, G. M. (1996). Core mutants of the immunoglobulin binding domain of streptococcal protein-G: stability and structural integrity. *FEBS Letters*, **398**, 312-316.
- Hellinga, H. W. (1998a). Computational protein engineering. *Nature Struct. Biol.* **5**, 525-527.
- Hellinga, H. W. (1998b). Construction of a blue copper analog through iterative rational protein design cycles demonstrates principles of molecular recognition in metal center formation. *J. Am. Chem. Soc.* **120**, 10055-10066.
- Hellinga, H. W. (1998c). The construction of metal centers in proteins by rational design. *Fold. Design*, **3**, R1-R8.
- Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990). Identification of native protein folds amongst a large number of incorrect models. *J. Mol. Biol.* **216**, 167-180.
- Henikoff, S. & Henikoff, J. G. (1992). Amino-acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915-10919.
- Holm, L. & Sander, C. (1992). Evaluation of protein models by atomic solvation preference. *J. Mol. Biol.* **225**, 93-105.
- Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature*, **358**, 86-89.
- Jones, D. T., Miller, R. T. & Thornton, J. M. (1995). Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins: Struct. Funct. Genet.* **23**, 387-397.
- Kendrew, J., Dickerson, R., Strandberg, B., Hart, R., Davies, D. & Philips, D. (1960). Structure of myoglobin: a three-dimensional Fourier synthesis a 2 Å resolution. *Nature*, **185**, 422-427.
- Koehl, P. & Delarue, M. (1994a). Application of a self consistent mean field theory to predict protein side-chains conformation and estimate their conformational entropy. *J. Mol. Biol.* **239**, 249-275.
- Koehl, P. & Delarue, M. (1994b). Polar and non-polar atomic environment in the protein core: implications for folding and binding. *Proteins: Struct. Funct. Genet.* **20**, 264-278.
- Koehl, P. & Delarue, M. (1997). The native sequence determines sidechain packing in a protein, but does optimal sidechain packing determine the native sequence? In *Proceedings of the Pacific Symposium on Biocomputing* (Altman, R. B., Dunker, A. K., Hunter, L. & Klein, T., eds), pp. 198-209, World Scientific, Singapore.
- Koehl, P. & Levitt, M. (1999). *De novo* protein design. II. Plasticity in sequence space. *J. Mol. Biol.* **293**, 1183-1193.
- Kraulis, P. J. (1991). Molscrip: a program to produce both detailed and schematic plots of protein structures. *J. Appl. Crystallog.* **24**, 946-950.
- Kreitman, R. J. & Pastan, I. (1998). Immunotoxins for targeted cancer-therapy. *Advan. Drug Deliv. Rev.* **31**, 53-88.
- Lazar, G. A., Desjarlais, J. R. & Handel, T. M. (1997). *De novo* design of the hydrophobic core of ubiquitin. *Protein Sci.* **6**, 1167-1178.
- Legrand, S. M. & Merz, K. M. (1993). Rapid approximation to molecular-surface area *via* the use of Boolean logic and look-up tables. *J. Comput. Chem.* **14**, 349-352.
- Levitt, M. (1976). Simplified representation of protein conformations for rapid simulation of protein folding. *J. Mol. Biol.* **104**, 59-107.
- Levitt, M. (1983). Protein folding by constrained energy minimization and molecular dynamics. *J. Mol. Biol.* **170**, 723-764.
- Levitt, M., Gerstein, M., Huang, E., Subbiah, S. & Tsai, J. (1997). Protein-folding: the endgame. *Annu. Rev. Biochem.* **66**, 549-579.
- Lim, W. A. & Sauer, R. T. (1989). Alternative packing arrangements in the hydrophobic core of lambda-repressor. *Nature*, **339**, 31-36.
- Lim, W. A. & Sauer, R. T. (1991). The role of internal packing interactions in determining the structure and stability of a protein. *J. Mol. Biol.* **219**, 359-376.
- Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. *Science*, **227**, 1435-1441.
- Malakauskas, S. M. & Mayo, S. L. (1998). Design, structure and stability of a hyperthermophilic protein variant. *Nature Struct. Biol.* **5**, 470-475.
- Mer, G., Kellenberger, E. & Lefevre, J. F. (1998). Alpha-helix mimicry of a beta-turn. *J. Mol. Biol.* **281**, 235-240.
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

- Miller, R. T., Jones, D. T. & Thornton, J. M. (1996). Protein fold recognition by sequence threading: tools and assessment techniques. *FASEB J.* **10**, 171-178.
- Minor, D. L. & Kim, P. S. (1994a). Context is a major determinant of beta-sheet propensity. *Nature*, **371**, 264-267.
- Minor, D. L. & Kim, P. S. (1994b). Measurement of the beta-sheet-forming propensities of amino-acids. *Nature*, **367**, 660-663.
- Mirny, L. A. & Shakhnovich, E. I. (1998). Protein-structure prediction by threading: why it works and why it does not. *J. Mol. Biol.* **283**, 507-526.
- Moult, J. (1997). Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* **7**, 194-199.
- Mutter, M. & Tuchscherer, G. (1997). Nonnative architectures in protein design and mimicry. *Cell. Mol. Life Sci.* **53**, 851-863.
- Nath, U. & Udgaonkar, J. B. (1997). How do proteins fold. *Curr. Sci.* **72**, 180-191.
- Neria, E., Fischer, S. & Karplus, M. (1996). Simulation of activation free energies in molecular systems. *J. Chem. Phys.* **105**, 1902-1921.
- Pabo, C. (1983). Designing proteins and peptides. *Nature*, **301**, 200.
- Pande, V. S., Grosberg, A. Y., Joerg, C., Kardar, M. & Tanaka, T. (1996). Freezing transition of compact polyampholytes. *Phys. Rev. Letters*, **77**, 3565-3568.
- Pande, V. S., Grosberg, A. Y. & Tanaka, T. (1997). Statistical-mechanics of simple-models of protein-folding and design. *Biophys. J.* **73**, 3192-3210.
- Pastan, I. H., Pai, L. H., Brinkmann, U. & Fitzgerald, D. J. (1995). Recombinant toxins: new therapeutic agents for cancer. *Ann. NY Acad. Sci.* **758**, 345-354.
- Perutz, M., Rossmann, M., Cullis, A., Muirhead, G., Will, G. & North, A. (1960). Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5 Å resolution, obtained by X-ray analysis. *Nature*, **185**, 416-422.
- Pinto, A. L., Hellinga, H. W. & Caradonna, J. P. (1997). Construction of a catalytically active iron superoxide-dismutase by rational protein design. *Proc. Natl Acad. Sci. USA*, **94**, 5562-5567.
- Ponder, J. W. & Richards, F. M. (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.
- Regan, L. & Wells, J. (1998). Engineering and design. Recent adventures in molecular design. *Curr. Opin. Struct. Biol.* **8**, 441-442.
- Reidhaarolson, J. F., Bowie, J. U., Breyer, R. M., Hu, J. C., Knight, K. L., Lim, W. A., Mossing, M. C., Parsell, D. A., Shoemaker, K. R. & Sauer, R. T. (1991). Random mutagenesis of protein sequences using oligonucleotide cassettes. *Methods Enzymol.* **208**, 564-586.
- Rice, D. W. & Eisenberg, D. (1997). A 3d-1d substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **267**, 1026-1038.
- Rooman, M. J. & Wodak, S. J. (1995). Are database-derived potentials valid for scoring both forward and inverted protein folding? *Protein Eng.* **8**, 849-858.
- Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584-599.
- Schafmeister, C. E. & Stroud, R. M. (1998). Helical protein design. *Curr. Opin. Biotechnol.* **9**, 350-353.
- Schiffer, C. A., Caldwell, J. W., Stroud, R. M. & Kollman, P. A. (1992). Inclusion of solvation free-energy with molecular mechanics energy: alanyl dipeptide as a test case. *Protein Sci.* **1**, 396-400.
- Schiffer, C. A., Caldwell, J. W., Kollman, P. A. & Stroud, R. M. (1993). Protein-structure prediction with a combined solvation free energy-molecular mechanics force-field. *Mol. Sim.* **10**, 121-134.
- Schultz-Beardsley, D. & Kauzmann, W. (1996). Local densities orthogonal to β -sheet amide planes: patterns of packing in globular proteins. *Proc. Natl Acad. Sci. USA*, **93**, 4448-4453.
- Seno, F., Vendruscolo, M., Maritan, A. & Banavar, J. R. (1996). Optimal protein design procedure. *Phys. Rev. Letters*, **77**, 1901-1904.
- Shakhnovich, E. I. (1994). Proteins with selected sequences fold to their unique native conformation. *Phys. Rev. Letters*, **72**, 3907-3910.
- Shakhnovich, E. I. (1997). Theoretical-studies of protein-folding thermodynamics and kinetics. *Curr. Opin. Struct. Biol.* **7**, 29-40.
- Shakhnovich, E. I. (1998). Protein design: a perspective from simple tractable models. *Fold. Design*, **3**, R45-R58.
- Shakhnovich, E. I. & Gutin, A. M. (1993a). Engineering of stable and fast-folding sequences of model proteins. *Proc. Natl Acad. Sci. USA*, **90**, 7195-7199.
- Shakhnovich, E. I. & Gutin, A. M. (1993b). A new approach to the design of stable proteins. *Protein Eng.* **6**, 793-800.
- Sharp, K. A., Nicholls, A., Friedman, R. & Honig, B. (1991). Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models. *Biochemistry*, **30**, 9686-9687.
- Simonson, T. & Brunger, A. T. (1994). Solvation free-energies estimated from macroscopic continuum theory: an accuracy assessment. *J. Phys. Chem.* **98**, 4683-4694.
- Sippl, M. (1990). Calculation of conformational ensembles from potentials of mean force: an approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **190**, 859-883.
- Sippl, M. (1993). Boltzmann's principle, knowledge-based mean fields and protein folding. An approach to the computational determination of protein structures. *J. Comput. Aided. Mol. Des.* **7**, 473-501.
- Sippl, M. (1995). Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* **5**, 229-235.
- Sippl, M. J. & Jaritz, M. (1994). Predictive power of mean force pair potentials. In *Proteins Structure by Distance Analysis* (Bohr, H. & Brunak, S., eds), pp. 113-134, IOS Press, Amsterdam.
- Sippl, M. & Weitckus, S. (1992). Detection of native-like models for amino-acid sequences of unknown three dimensional structure in a database of known protein conformation. *Proteins: Struct. Funct. Genet.* **13**, 258-271.
- Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997). Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? *Protein Sci.* **6**, 676-688.
- Skolnick, J., Kolinski, A. & Ortiz, A. R. (1998). Reduced protein models and their application to the protein-folding problem. *J. Biomol. Struct. Dynam.* **16**, 381-396.
- Smith, C. K. & Regan, L. (1995). Guidelines for protein design: the energetics of beta-sheet side-chain interactions. *Science*, **270**, 980-982.

- Smith, C. K. & Regan, L. (1997). Construction and design of beta-sheets. *Acc. Chem. Res.* **30**, 153-161.
- Smith, C. K., Withka, J. M. & Regan, L. (1994). A thermodynamic scale for the beta-sheet forming tendencies of the amino-acids. *Biochemistry*, **33**, 5510-5517.
- Street, A. G. & Mayo, S. L. (1998). Pairwise calculation of protein solvent-accessible surface-areas. *Fold. Design*, **3**, 253-258.
- Subbiah, S., Laurents, D. V. & Levitt, M. (1993). Structural similarity of DNA-binding domains of bacteriophage repressors and the globin core. *Curr. Biol.* **3**, 141-148.
- Sun, S., Brem, R., Chan, H. S. & Dill, K. A. (1995). Designing amino acid sequences to fold with good hydrophobic cores. *Protein Eng.* **8**, 1205-1213.
- Tanaka, S. & Scheraga, H. A. (1976). Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules*, **9**, 945-950.
- Thomas, P. D. & Dill, K. A. (1996). Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* **257**, 457-469.
- Tuffery, P., Etchebest, C., Hazout, S. & Lavery, R. (1991). A new approach to the rapid determination of protein side-chain conformations. *J. Biomol. Struct. Dynam.* **8**, 1267-1289.
- Vendruscolo, M. & Domany, E. (1998). Pairwise contact potentials are unsuitable for protein-folding. *J. Chem. Phys.* **109**, 11101-11108.
- Vendruscolo, M., Maritan, A. & Banavar, J. R. (1997). Stability threshold as a selection principle for protein design. *Phys. Rev. Letters*, **78**, 3967-3970.
- Vendruscolo, M., Najmanovich, R. & Domany, E. (1999). Protein folding in contact map space. *Phys. Rev. Letters*, **82**, 656-659.
- Vita, C. (1997). Engineering novel proteins by transfer of active-sites to natural scaffolds. *Curr. Opin. Biotechnol.* **8**, 429-434.
- Vita, C., Roumestand, C., Toma, F. & Menez, A. (1995). Scorpion toxins as natural scaffolds for protein engineering. *Proc. Natl Acad. Sci. USA*, **92**, 6404-6408.
- Walsh, S. T., Cheng, H., Bryson, J. W., Roder, H. & deGrado, W. F. (1999). Solution structure and dynamics of a *de novo* designed three helix bundle. *Proc. Natl Acad. Sci. USA*, **96**, 5486-5491.
- Wesson, L. & Eisenberg, D. (1992). Atomic solvation parameters applied to molecular-dynamics of proteins in solution. *Protein Sci.* **1**, 227-235.
- Wilks, H. M. & Holbrook, J. J. (1991). Alteration of enzyme specificity and catalysis by protein engineering. *Curr. Opin. Struct. Biol.* **2**, 561-567.
- Wood, R. H. & Thompson, P. T. (1990). Differences between pair and bulk hydrophobic interactions. *Proc. Natl Acad. Sci. USA*, **87**, 946-949.
- Yue, K. & Dill, K. A. (1992). Inverse protein folding problem: designing polymer sequences. *Proc. Natl Acad. Sci. USA*, **89**, 4163-4167.
- Yue, K., Fiebig, K. M., Thomas, P. D., Chan, H. S., Shakhnovich, E. I. & Dill, K. A. (1995). A test of lattice protein folding algorithms. *Proc. Natl Acad. Sci. USA*, **92**, 325-329.

Edited by F. E. Cohen

(Received 15 July 1999; received in revised form 8 September 1999; accepted 9 September 1999)