Protein Science (1995), *4*:2006–2018. Cambridge University Press. Printed in the USA.

De novo design of the hydrophobic cores of proteins

JOHN R. DESJARLAIS AND TRACY M. HANDEL
Department of Molecular and Cell Biology, University of California at Berkeley, Berkeley, California 94720

(RECEIVED May 4, 1995; ACCEPTED July 26, 1995)

Abstract

We have developed and experimentally tested a novel computational approach for the de novo design of hydrophobic cores. A pair of computer programs has been written, the first of which creates a "custom" rotamer library for potential hydrophobic residues, based on the backbone structure of the protein of interest. The second program uses a genetic algorithm to globally optimize for a low energy core sequence and structure, using the custom rotamer library as input. Success of the programs in predicting the sequences of native proteins indicates that they should be effective tools for protein design.

Using these programs, we have designed and engineered several variants of the phage 434 cro protein, containing five, seven, or eight sequence changes in the hydrophobic core. As controls, we have produced a variant consisting of a randomly generated core with six sequence changes but equal volume relative to the native core and a variant with a "minimalist" core containing predominantly leucine residues. Two of the designs, including one with eight core sequence changes, have thermal stabilities comparable to the native protein, whereas the third design and the minimalist protein are significantly destabilized. The randomly designed control is completely unfolded under equivalent conditions. These results suggest that rational de novo design of hydrophobic cores is feasible, and stress the importance of specific packing interactions for the stability of proteins. A surprising aspect of the results is that all of the variants display highly cooperative thermal denaturation curves and reasonably dispersed NMR spectra. This suggests that the non-core residues of a protein play a significant role in determining the uniqueness of the folded structure.

Keywords: computational; 434 cro; genetic algorithm; protein design; uniqueness

Considerable effort has been directed toward the de novo design of protein sequences that will fold into a predetermined 3D topology (DeGrado et al., 1991; Betz et al., 1993). In general, the results of these studies suggest that the desired protein architectures can be achieved with relatively simple design rules, as demonstrated by the minimalist design approach and the more recently reported binary patterning strategy (DeGrado et al., 1989; Hecht et al., 1990; Kamtekar et al., 1993). These approaches rely heavily on the idea of a "hydrophobic inside–hydrophilic outside" placement of amino acid types along the target structure. Additional design features often include the introduction of potential electrostatic interactions to promote secondary structure formation, secondary structural propensities, or modest consideration of specific packing interactions within the hydrophobic core.

The common result of most design attempts has been a protein that contains significant amounts of the desired secondary structure and that appears to fold into an approximately correct topology (Raleigh & DeGrado, 1992; Handel et al., 1993; Munson et al., 1994). However, a protein has yet to be designed that possesses all of the characteristics of a natural protein. Unlike most natural proteins, designed proteins generally lack a well-defined and uniquely structured folded state. These proteins usually display weak cooperativity in their unfolding transitions and poorly dispersed NMR spectra. Because of the poorly structured nature of these proteins, determination of high-resolution structures for these molecules has also been hampered. The consensus emerging from these studies is that the designed proteins are lacking the specific interactions necessary to stabilize the folded protein into a single, structurally unique state. Further design attempts have incorporated this thinking via the introduction of disulfide bonds (Quinn et al., 1994; Yan & Erickson, 1994), metal binding sites (Regan & Clarke, 1990; Handel et al., 1993), or specific packing interactions (Raleigh & DeGrado, 1992). These designs often show an improvement toward native-like behavior, but complete success remains elusive.

A common suspicion is that the specific packing interactions observed in structures of natural proteins (Richards, 1977) are not sufficiently present in designed proteins, and that the addition of these interactions, if possible, might lead to a higher degree of specificity for the folded state. The development of several computer programs for the design of well-packed hydro-

Reprint requests to: John R. Desjarlais, 206 Stanley Hall, University of California at Berkeley, Berkeley, California 94720; e-mail: johnd@paradise1.berkeley.edu.

phobic cores (Ponder & Richards, 1987; Hellinga & Richards, 1994; Kono & Doi, 1994) and their success in predicting the core sequences of natural proteins suggest that rational design of hydrophobic cores is indeed possible.

We wish to investigate the impact of the design of the hydrophobic core sequences on the uniqueness and stability of designed proteins. To this end, we have developed a computational approach that overcomes some of the limitations of earlier methods. Importantly, we test the approach experimentally by making and characterizing several predicted variants of a natural protein. As a stringent test of our algorithms, we have focused on some variants that bear little resemblance to the native hydrophobic core sequence. Our results demonstrate that the rational de novo design of hydrophobic cores is possible, and emphasize the importance of packing interactions in stabilizing the folded conformation of a protein. An assessment of the agreement between the 3D structures of these variants and their predicted structures awaits further study.

Results

Description of programs

The first and simpler of the two programs executes a search for low-energy rotamers of a subset of hydrophobic amino acids (V, I, L, F, A, and sometimes W) at each core position of the protein of interest. The energies (see the Materials and methods) are evaluated against the complete input structure, which includes the backbone structure as well as non-core side chains. We refer to the output of this search as a "custom-made" rotamer library, where a separate library is created for each position of interest. The advantage of this approach is that the rotamer set chosen for each position reflects the particular atomic details of the template structure. This represents a major departure from existing methods that use a library of rotamers derived statistically from structures in the Protein Data Bank (Ponder & Richards, 1987). Importantly, it can be shown that the rotamer distributions derived from the program for model secondary structures (α-helix and β-strand) closely resemble those derived statistically (McGregor et al., 1987). Given this consistency, we assume that the computer search also accurately reflects the energetics of side-chain conformers in regions of nonstandard secondary structure, which often show significantly different profiles of energetically favorable rotamer distributions.

A custom rotamer set for a single amino acid side chain can consist of up to 18 members (see the Materials and methods for more details). Given this set size for five or six hydrophobic side-chain types, and a core consisting of 10–20 positions, finding an optimal core sequence and structure is a problem of significant combinatorial complexity: for a small protein with 10 core positions, more than 10^{18} structural solutions exist with roughly 10^{10} sequence combinations. To deal with this, we have developed a second program, called ROC (repacking of cores), which uses a genetic algorithm to search efficiently through sequence-rotamer space.

Genetic algorithms are a class of optimization techniques, based on concepts derived from biological evolution, which have shown great promise in dealing with problems of significant complexity (Holland, 1992). This type of optimization is different from the more common approach of Monte Carlo sampling, where typically a single test solution is derived in each new cycle by random mutation of an existing solution and evaluated by some criterion for acceptance as a new solution. In contrast, a genetic algorithm involves an "evolving" population of solutions where new populations are generated as a result of mutation and recombination of the solutions within the existing population. In our case, a solution consists of a single residue/rotamer combination for each core position. Taken together with the fixed input structure of the backbone and non-core side chains, a solution contains the information necessary to construct a model structure for the whole test protein, the energy of which can then be evaluated. The evolution begins with a collection of model structures whose core sequence and structure are randomly chosen. According to the calculated energy of each structure, recombination probabilities are assigned to each solution such that solutions encoding lower energy structures will recombine more frequently. Recombinations are carried out at randomly chosen crossover points according to these probabilities. Because a solution encodes a complete model structure, a recombination is equivalent to swapping segments of two model structures upstream or downstream from a chosen crossover point within the backbone (leading to two new solutions). After a subsequent round of random mutation of side-chain conformation or identity, the new population is evaluated and the cycle begins anew. This process is typically repeated several hundred times in order to assure complete convergence of the population to a single species.

Prediction of core sequence

The first test of such a design program should be its success in predicting the sequences of natural proteins. Experimental studies have shown that, for a given protein, many core sequences may exist that yield stably folded functionally active proteins (Lim & Sauer, 1989; Richards & Lim, 1993). Thus, it is not expected that the native sequence be predicted exactly. However, sequences resembling the native to some extent are expected if the program is to be useful for protein design. Ideally, the program would report several different sequences consistent with the target structure. These sequences could then be compared to each other using a number of criteria, including calculated energetic components, the sizes and distribution of cavities, and visual inspection. The original algorithm, which was designed to converge to the same sequence and structure for each run, was modified in order to encourage convergence to a different sequence for each run. This was accomplished using small perturbations to the energetics used within the genetic algorithm. Typically, for 100 runs of the program, between 20 and 80 different sequences are generated, depending on the particular test protein.

A set of four proteins was chosen to demonstrate the validity of the program. Included in the set are two proteins with a large number of core positions. Two points should be noted regarding prediction of larger core sequences. First, the statistical difficulty of correctly predicting the native sequence increases exponentially with the number of core positions. Second, cores of this size have been difficult to deal with using existing algorithms, because of the inability of exhaustive search procedures to examine all possibilities in a reasonable amount of time and the potential difficulty of determining a global minimum using nonexhaustive search methods.

Figure 1 shows partial outputs for the four test proteins. As a simplification, we show only the five lowest energy sequences out of a total of 100 runs. The prediction success is variable, but the results on some of the proteins are quite promising. Also note that, although buried volume (Richards, 1977) was not included as an explicit constraint, the volumes of the sequences predicted are close to the native volume. The high degree of similarity in the total energies indicates the subtlety involved in determining an optimal core sequence. Underneath each of the partial sequence sets we show a consensus for the entire set: the two most frequent residues in each position are included, but only if the second residue is found in more than 20% of the runs. With this constraint, all positions can be narrowed down to one or two possible residue types. Even after this reduction, the wild-type residue is frequently included in the consensus. In general, although perfect sequence prediction is not achieved, the sequences predicted are very similar to the native, especially when compared to the low similarity seen with a set of sequences generated randomly with a native-like volume constraint (not shown).

```
(A) Major Cold Shock Protein (1MJC; 2.0 Å)                  Etot   ΔVol

N: V F I V V I L V F V

1  V F I I I I L I F V                                      -62.6  +81
2  I F I I I I L V F V                                      -62.2  +81
3  V F V I I I L I F V                                      -61.9  +54
4  V F V I I I L L F V                                      -61.8  +53
5  V F I I I I L V F V                                      -61.7  +54

C: V F I I I I L V F V
   I   V V       I   L

(B) Thioredoxin (2TRX; 1.68 Å)

N: I L I I L I Y L V L L L L F L L

1  I V I L I L F I F I L V L F L I                          -73.8  +20
2  I L V L L L F I F F F V L F L I                          -73.7  +87
3  I V I L L L F I F I L V L F L I                          -73.1  +19
4  I L V L I L F I F F F V L F L I                          -73.0  +88
5  L L V L L L F I F F F V L F L I                          -72.7  +16

C: I V I L L L F I F I L V L F L I
     L V I   V   L I L F L V

(C) Basic Fibroblast Growth Factor (2FGF; 1.77 Å)

N: L C L I V L L V I L M L F Y V L F

1  L V L I I L L V L L L L F F V I F                        -62.8  +59
2  L V L I I L L V L L L L F L V I F                        -62.2  +24
3  L V L I I L V F I L L L F F V I F                        -61.3  +95
4  L V L I I I V F L L L L F F V L F                        -61.1  +94
5  L V L I I L V F I L L L F F V L F                        -61.0  +94

C: L V L I I L L V L L L L F F V I L
     I   V     V F I V       L L L F

(D) Interleukin 4 (1RCB; 2.25 Å)

N: I L L V I F F A V L F L L L L L F L L M

1  L I L L F I L V V I F F L L L L I I I L  -76.5  +42
2  L I L L I F L V V I F F L L L L I I V L  -75.5  +15
3  L I L V V I L I V V F F L L L V F L I L  -75.3  -38
4  L L L L I F L V V F F F L L L L I I L L  -75.1  +74
5  L I L V V I L I V V F F L L L L F I I L  -74.9  -11

C: L I L V I I L V V I F F L L L V I L I L
     L   L V F   I   F           L F I V
```

Fig. 1. Prediction of hydrophobic core sequences using crystal structures of native proteins. **A:** Major cold shock protein (Schindelin et al., 1994), positions 9, 12, 21, 30, 32, 37, 45, 51, 53, and 67. **B:** Thioredoxin (Katti et al., 1990), positions 4, 24, 38, 41, 42, 45, 49, 53, 55, 78, 80, 94, 99, 102, 103, and 107. **C:** Basic fibroblast growth factor (Zhang et al., 1991), positions 23, 25, 32, 34, 40, 53, 55, 63, 65, 74, 76, 82, 94, 106, 116, 118, and 139. **D:** Interleukin 4 (Wlodaver et al., 1992), positions 10, 14, 17, 29, 32, 33, 45, 48, 51, 52, 55, 79, 83, 86, 90, 109, 112, 113, 116, and 120. Using the crystal structures including non-core side chains as input, the hydrophobic core sequence is predicted using the program ROC. Each protein was subjected to 100 runs of the program. For simplicity, we show only the five different lowest energy sequences found for each protein. Shown above is the native core sequence (N) and below each list is a consensus sequence (C) for all 100 runs. Total calculated energy (arbitrary units) and difference in core residue volume (Å^3) relative to the native are also reported for each sequence. Core residue volumes are calculated as the sum of the individual residue volumes (Chothia, 1975).

As an additional test of sequence prediction, we have simulated genetic experiments that were performed on the proteins λ repressor and T4 lysozyme (Lim & Sauer, 1989; Baldwin et al., 1993). In both studies, five out of a larger set of core residues were subject to genetic mutation and functional sequences were collected by selection. Our simulations allow the same five residues to be mutated in each case, but also allow the conformation of the surrounding core side chains to vary, because this may affect the range of permitted sequences. The results of the simulations are shown in Figure 2. The sequences found in our simulation for λ repressor bear a strong resemblance to several of those found to be fully or partially functional by genetic selection. The best scoring sequence for λ repressor is one residue different from the native sequence and is identical to one of the fully functional mutations. In the case of T4 lysozyme, our

```
(A) Lambda Repressor (1LMB; 1.8 Å)          Etot   ΔVol

N: L I Y L V M V L F L L L V F

1  L       I M V   F L L                    -44.3  +26
2  L       I M V   L F L                    -44.2  +26
3  M       L M V   F L I                    -44.0  +29
4  L       I L V   F L L                    -43.8  +23
5  L       M V L   F L L                    -43.8  +25

F: L       I M V   F L L
F: L       I M V   L L L
F: L       I L I   F L L
P: L       M M L   F L L
P: L       I V L   F L L

(B) T4 Lysozyme (2LZM; 1.7 Å)

N: V L L M V F L L A L W V I F

1                M V L   I   F          -61.9  +80
2                M V M   I   F          -62.1  +83
3                M V L   I   I          -59.3  +46
4                M V L   L   I          -59.2  +45
5                V V F   I   F          -58.4  +86
```

Fig. 2. Computer simulation of genetic selection experiments (Lim & Sauer, 1989; Baldwin et al., 1993) using the crystal structures of λ repressor (Beamer & Pabo, 1992) and T4 lysozyme (Weaver & Matthews, 1987) as input. Prediction experiments are similar to those in Figure 1, except here the identities of only those core residues that were subject to mutation in the actual genetic experiments were allowed to vary, whereas only the conformation of the surrounding core residues is allowed to vary. **A:** The five lowest energy sequences predicted for the core of λ repressor. Listed below these are several genetically selected fully (F) or partially (P) functional sequences (Lim & Sauer, 1989) that are identical or highly similar to the computer-selected sequences. **B:** The five lowest energy sequences predicted for the core of T4 lysozyme. None of the sequences match the six reported selected sequences (Baldwin et al., 1993), although there are some similarities.

predicted sequences bear only modest resemblance to the six reported functional variants, although this is perhaps not unexpected for such a small comparison set. One limitation of our program is that it does not allow for changes in backbone conformation to accommodate mutations. Because this does occur in the genetically selected proteins, it may limit the agreement between the genetic and computationally derived sequences (Baldwin et al., 1993; Lim et al., 1994).

Although reasonable sequence prediction is obviously important, correct prediction of the spatial orientation of the core residues is also expected if our methodology is to be confirmed. In Figure 3 we show the predicted structure of the lowest energy core sequence for the major cold shock protein compared to the crystal structure of the native protein. Even though this particular model has three sequence differences from the native, it is difficult to distinguish the structures. Interestingly, experimental studies of core variants have shown that in many cases the orientations of mutated side chains are preserved even when the sequence is altered (Baldwin et al., 1993).

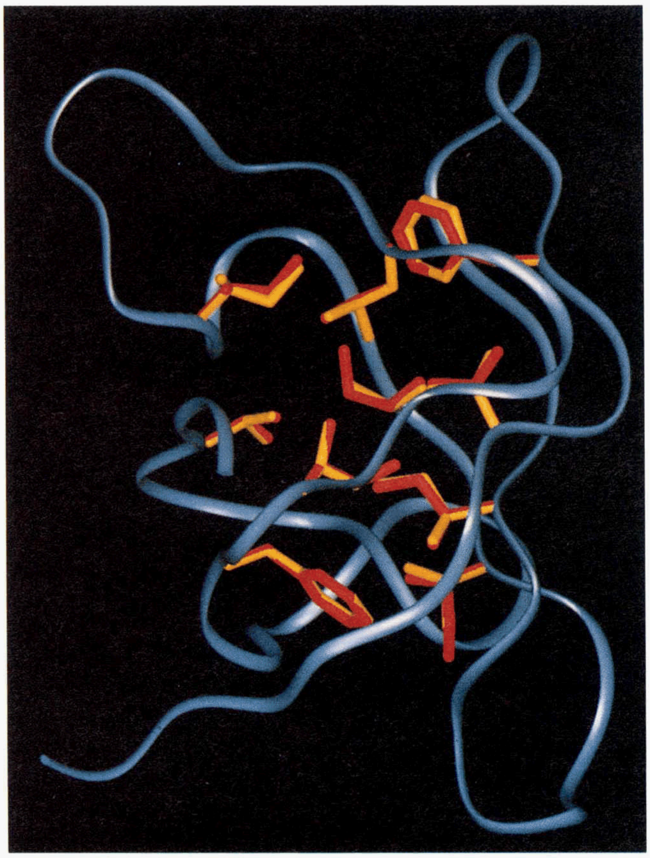

Fig. 3. Model structure for the lowest energy predicted sequence of the major cold shock protein. Core structures of the model (red) and the crystal structure (blue) are shown with a ribbon diagram of the backbone structure. Models were displayed using the program INSIGHT II (Biosym Technologies, San Diego, California).

Redesign of a natural protein core: 434 cro

The result that the core sequences and structures are well predicted for a number of proteins implies that the program could indeed be applied to the de novo design of a protein core. To test this, we chose to redesign the hydrophobic core of a natural protein as a first step toward computationally driven total de novo design. An important aspect of this approach is that it allows us to compare critically the structure, dynamics, and stability of the designed cores to the native core.

We chose the cro protein from bacteriophage 434 as an initial experimental system for several reasons. (1) 434 cro is a small protein with a single contiguous hydrophobic core. (2) It does not rely on disulfide bonds or metal binding for stability. (3) The small size of this protein makes it feasible to construct whole genes out of a small number of synthetic oligonucleotides. (4) Because the native protein itself is not substantially stable, this protein serves as a stringent test of the success of designed core sequences. (5) The small size should facilitate eventual structure determination of variants by NMR. (6) As a DNA-binding protein, it can be tested for preservation of specific binding to its natural DNA sequence (future work).

The program was run using the crystal structure of 434 cro protein (2CRO) as input (Mondragon et al., 1989). All non-core side chains were present, being an important constraint for predicting core sequences compatible with a natural molecule. Figure 4 shows the complete output of core sequences for 100 runs of the program on 434 cro. For this protein, the native sequence is not predicted. There are, however, a number of sequences very similar to the native. The most consistent differences in sequence from the native core are the replacement of the first two leucines with isoleucines and the replacement of cysteine in position 54 with valine. Cysteine was not included in the original run of the program, but even when included as a possible residue in this position, valine was found to be preferred.

Several sequences also bear a strong resemblance to the core sequence of 434 repressor protein. This is not surprising given the strong homology between 434 cro and repressor proteins. A comparison of the crystal structure of 434 repressor (1R69) (Mondragon et al., 1989) with that of 434 cro indicates that the structures themselves are indeed extremely similar, with a backbone RMSD of less than 1 Å. Particularly comforting is the fact that the equivalent of cro position 54 in 434 repressor contains a valine instead of a cysteine. Also note that the first two core positions in repressor are isoleucine and valine, more closely resembling sequences found in the output generated for 434 cro.

A useful feature of the multiple sequence output is the recurrence of a number of sequences. This implies that these particular sequences are less sensitive to small changes in the energetic parameters used within the program. Along with the energy output for each sequence, the number of occurrences serves as an additional criterion for choosing a particular sequence. Using these parameters and visual inspection as a guide, we chose to make the proteins corresponding to three of the sequence variants from the list in Figure 4.

A sequence with high occurrence and low energy should serve as the best test of the potential application of the program for de novo protein design. This is particularly important given that the native sequence is not predicted. For this reason, we chose sequence number 2, which has a low total energy, a low pack-

	2	6	13	20	26	31	34	45	58	52	54	58	59					
N:	L	L	L	L	V	I	I	L	I	L	C	W	L	Etot	Es-b	Es-s	ΔVol	n
1	L	I	L	L	I	I	I	L	L	L	V	W	L	-68.6	-85.9	-11.9	+63	1
2	I	I	L	L	I	I	L	L	I	L	V	W	L	-68.4	-86.7	-15.8	+64	15
3	L	I	L	L	I	I	I	L	I	L	V	W	L	-68.4	-85.4	-13.4	+64	1
4	I	I	L	L	V	I	L	L	I	L	V	W	L	-68.3	-87.0	-15.7	+37	16
5	I	I	L	L	I	V	L	L	L	L	V	W	L	-68.0	-87.3	-14.7	+36	10
6	I	I	L	L	I	V	L	L	I	L	V	W	L	-67.8	-87.2	-15.1	+37	14
7	I	I	L	L	I	I	I	L	I	L	V	W	L	-67.6	-85.3	-11.7	+65	1
8	L	I	L	L	I	V	I	L	I	L	V	W	L	-67.5	-86.1	-12.0	+37	1
9	L	I	L	L	I	V	I	L	L	L	V	W	L	-67.5	-88.4	-13.1	+36	1
10	I	I	L	L	L	I	L	L	I	L	V	W	L	-67.4	-83.8	-14.9	+63	3
11	L	V	L	L	I	I	I	L	I	L	V	W	L	-67.4	-87.0	-12.0	+37	1
12	I	I	L	L	V	I	L	L	L	L	V	W	L	-67.3	-85.6	-13.2	+36	1
13	I	I	L	L	I	V	I	L	L	L	V	W	L	-67.3	-88.2	-11.1	+37	1
14	L	I	L	L	V	I	I	L	I	L	V	W	L	-66.9	-86.1	-14.3	+37	1
15	I	I	L	L	L	V	L	L	I	L	V	W	L	-66.9	-84.1	-15.4	+36	11
16	I	I	L	L	L	V	L	L	L	L	V	W	L	-66.5	-84.6	-15.8	+35	6
17	L	F	L	V	L	V	I	L	L	L	V	W	L	-65.8	-84.6	-16.3	+43	2
18	I	F	L	V	L	V	I	L	L	L	V	W	L	-65.6	-83.6	-16.9	+44	2
19	F	I	L	L	L	V	L	I	L	L	V	W	L	-65.3	-79.6	-17.2	+70	1
20	I	I	L	L	L	V	L	L	L	V	V	W	L	-65.1	-86.4	-13.5	+ 9	1
21	F	I	L	L	I	I	L	I	I	L	V	W	L	-64.6	-78.2	-16.2	+99	1
22	I	F	L	V	L	V	L	L	L	L	V	W	L	-64.6	-83.5	-17.3	+43	2
23	F	F	L	V	L	V	L	L	L	L	V	W	L	-63.6	-77.9	-16.8	+77	1
24	I	F	L	V	I	V	L	L	L	L	V	W	L	-63.2	-81.0	-15.7	+44	1
25	I	F	L	V	I	V	I	L	I	L	V	W	L	-63.1	-80.3	-14.7	+46	1
26	I	I	L	A	L	L	L	L	L	F	V	W	L	-62.4	-83.5	-15.3	+20	1
27	I	I	L	A	L	I	L	L	L	F	V	W	L	-62.3	-84.3	-14.8	+20	1
R:	I	V	L	L	T	I	L	L	L	L	V	W	L					

Fig. 4. Complete core sequence output from 100 runs of the program ROC using the crystal structure of 434 cro (2CRO) (Mondragon et al., 1989) including non-core side chains as input. The core sequences of native 434 cro (N) and 434 repressor (R) with core sequence positions are shown for comparison. Sequences chosen as design candidates are underlined (see also Fig. 5). The total energy of each sequence (E_{tot}) and the contributions of side-chain–backbone ($E_{s\text{-}b}$) and side-chain–side-chain ($E_{s\text{-}s}$) energies (arbitrary units) are reported, as well as the volume difference (Å^3) from the native. The integer n is the number of occurrences of a particular sequence in the total of 100 runs. Not shown are the energies of interaction of core side chains with non-core side chains, which can be derived from the other terms.

ing energy (indicated by the side-chain–side-chain energy term), and a high frequency of occurrence. Sequence number 4 might also have been chosen; it differs by one residue from sequence 2, but the difference represents a slight loss in volume, as well as one less change in sequence relative to the native protein. These two sequences together represent one-third of all occurrences. We will refer to the protein corresponding to sequence number 2 as D-5, meaning a designed sequence with five differences from the native sequence. Figure 5B shows the core structure of the variant D-5. Also shown is a comparison of the model structure for the variant C-1 versus the crystal structure (Fig. 5A), demonstrating the accuracy of the structure prediction for the native core.

We have also made a D-7 and a D-8 protein that were chosen for different reasons. First, they both bear little resemblance to the native core sequence. D-7 was chosen because it represents one of many similar sequences, all of which contain a phenylalanine in position 8, and complementary valines in positions 22 and 33. This class of sequences all have low side-chain–side-chain energies as well as low side-chain–backbone energies. D-8 was made because it had a very low side-chain–side-chain energy but a much higher side-chain–backbone energy. The idea was to test the possibility that if good packing is possible, the backbone will relax slightly to accommodate any side-chain–backbone strain. D-7 and D-8 also incorporate an aromatic residue into a core that originally contained only aliphatic side chains.

Three control proteins were also produced. The first, R-6 (random with six differences from native), has a hydrophobic core that was chosen randomly using the single constraint that the core maintain a volume close to that of the native sequence. This particular sequence is a good control because it has fewer differences from the native protein than two of our designs, has a very high energy, and is conserved in volume. It thus serves as a strong test of the importance of packing, as well as our ability to distinguish good from bad sequences. As a positive control and native sequence standard, the protein C-1 was made with a valine substituted for cysteine in position 54 (see the Materials and methods). Finally, it has been hypothesized that the increased rotameric freedom of a collection of leucine side chains (McGregor et al., 1987) leads to a more dynamic folded state (Handel et al., 1993). Thus, with the intent of creating a dynamic model system, the "minimalist" protein M-5 was made with leucines in most core positions (DeGrado et al., 1989). However, in order to preserve native-like volume, the conserved valine and tryptophan residues were retained.

The core sequences of all experimental 434 cro variants are shown in Figure 6 along with calculated energies of each variant. Although the variants C-1, R-6, and M-5 were not derived originally from the program, calculated energies for these sequences can be determined by restricting the program to each of the individual sequences for a number of runs. In order to properly compare all of the test sequences, the designed sequences were also subjected to the same number of runs with

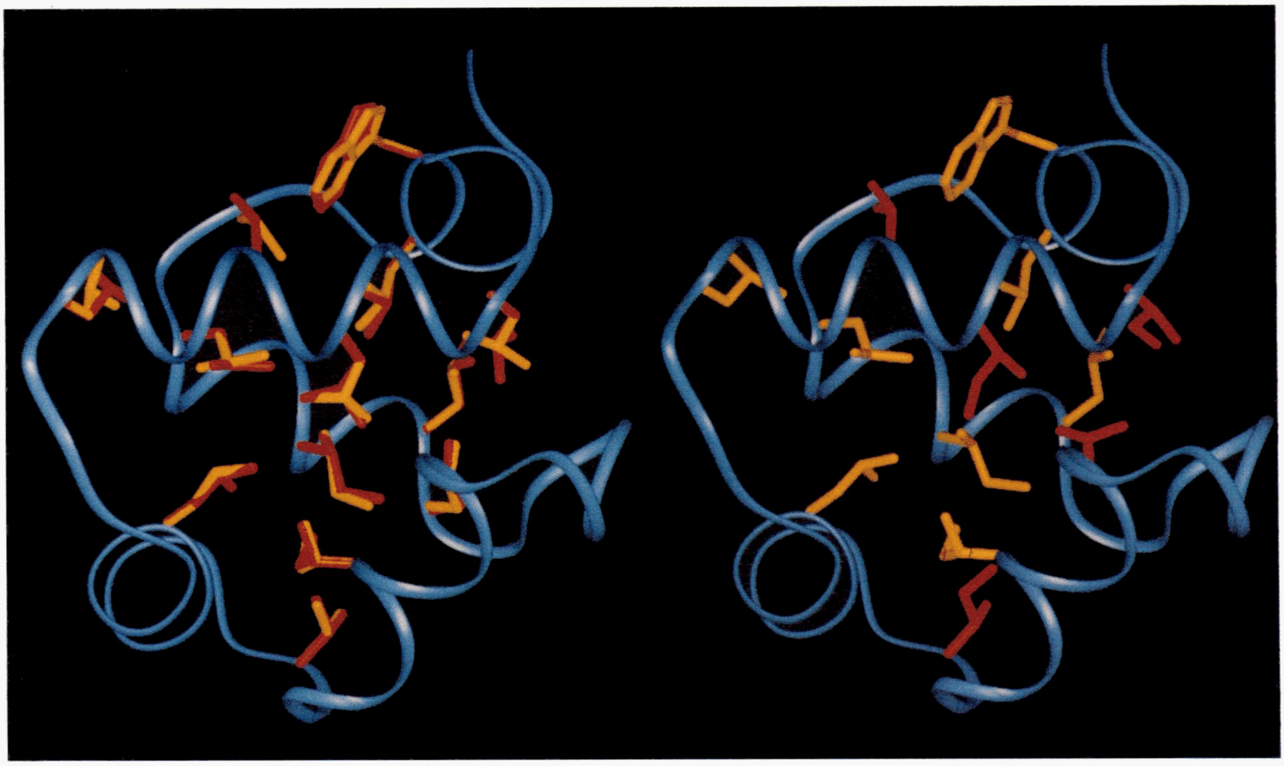

Fig. 5. Model structures for 434 cro variants compared to the native. **A:** Predicted core structure of the native-like C-1 variant (red) compared to the crystal structure of 434 cro (blue). **B:** Model core structure for the D-5 variant. Residues different from the native sequence are highlighted in red.

their sequences fixed. As seen in Figure 6, the randomly generated core sequence has a high energy compared to the other variants. To ensure that this was not an unusual occurrence, several random sequences were run through the program and were all determined to have comparatively high energies (not shown).

The absence of the native sequence in the predictions for 434 cro is slightly disconcerting. However, the energies shown in Figure 6, where the C-1 variant can be taken as representative of the native sequence, indicate that the native sequence is judged to be of comparatively higher energy using our current set of pa-

	2	6	13	20	26	31	34	45	48	52	54	58	59	E_{tot}	E_{s-b}	E_{s-s}	ΔVol	T_m
C-1	L	L	L	L	V	I	I	L	I	L	$\boxed{V}$	W	L	-64.3	-87.6	-15.0	+36	56
D-5	$\boxed{I}$	$\boxed{I}$	L	L	$\boxed{I}$	I	$\boxed{L}$	L	I	L	$\boxed{V}$	W	L	-68.8	-87.0	-15.8	+64	60
D-7	$\boxed{I}$	$\boxed{F}$	L	$\boxed{V}$	$\boxed{L}$	$\boxed{V}$	I	L	$\boxed{L}$	L	$\boxed{V}$	W	L	-66.2	-85.6	-17.0	+44	17
D-8	$\boxed{F}$	$\boxed{I}$	L	L	$\boxed{L}$	$\boxed{V}$	$\boxed{L}$	$\boxed{I}$	$\boxed{L}$	L	$\boxed{V}$	W	L	-67.0	-81.3	-17.3	+70	50
M-5	L	L	L	L	$\boxed{L}$	$\boxed{L}$	$\boxed{L}$	L	$\boxed{L}$	L	$\boxed{V}$	W	L	-59.8	-88.1	-13.3	+59	33
R-6	$\boxed{I}$	L	$\boxed{I}$	L	V	$\boxed{L}$	I	$\boxed{I}$	$\boxed{L}$	L	$\boxed{V}$	W	L	-52.8	-80.6	-11.7	+37	-

Fig. 6. Designed and control variants of 434 cro. Sequence names are derived from their origin as designed (D), positive control (C), minimalist (M), or random (R), and the number of differences between a given sequence and the native (e.g., D-5 = designed with five differences from native). Residues different from the native sequence are boxed. In order to compare the different sequences accurately and to evaluate sequences that were not part of the original output, the program was run 40 times while restricted to each individual sequence, resulting in a range of energies for each. For simplicity, we show only the lowest value of each energy term found for all runs. Different terms are therefore not mutually consistent. Experimentally determined melting temperatures (°C) are also reported for each sequence (see Fig. 7).

rameters. This is also true if the side-chain torsion angles are derived directly from the crystal structure and input into our program. These observations and a similar trend with other proteins suggest that the current potential could be improved. However, the similarity of many of the designed sequences to the native, as well as the experimental results that follow, suggest that the current potential is reasonably accurate.

Thermal stability and cooperativity of 434 cro variants

Thermal denaturation of each of the variants was monitored using the change in CD signal at 222 nm as a function of temperature. These data, corrected for sloping baselines, are shown in Figure 7 as apparent fractions of unfolded protein. As the data clearly demonstrate, two of the designed proteins (D-5 and D-8) are of comparable thermal stability to the C-1 native control. In fact, the high occurrence, low energy protein D-5 is actually more thermally stable than the native control. The proteins D-7 and M-5, however, are both significantly destabilized relative to the native and are fighting a balance between hot and cold denaturation. This behavior, surprising at first, is completely predictable assuming two-state unfolding characteristics and a heat capacity change consistent with the small size of this protein. The most dramatic result is that the R-6 protein, which has a randomly chosen hydrophobic core, is completely unfolded under these conditions.

Most designed proteins do not exhibit cooperative thermal unfolding transitions, presumably due to a significant population of intermediate states or because of unsubstantial enthalpies of denaturation. The level of cooperativity observed in the thermal denaturation experiments with the 434 cro variants is therefore key to evaluating the impact of packing interactions on these properties. For the three most stable proteins (D-5, C-1, and D-8), it can be seen that the level of cooperativity of each of these proteins is approximately the same, although it does appear from the unfolding profiles that the C-1 native control protein is slightly more cooperative than the others. This is confirmed by curve-fitting analysis of the thermal denaturation data, which reveals that the C-1 protein has the most substantial van't Hoff enthalpy value (Table 1).

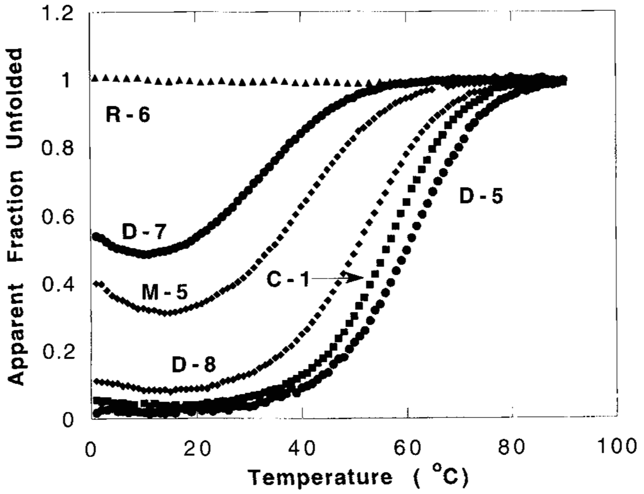

Fig. 7. Thermal denaturation of 434 cro variants. Unfolding of each variant was detected by monitoring the CD signal at 222 nm. Data are plotted here as the apparent fraction of unfolded protein versus temperature, after curve fitting of the original data. Data for the variants M-5 and D-7 reveal that these proteins are in a balance between hot and cold denaturation (not unexpected for proteins of this size and stability).

In attempts to observe full thermal denaturation profiles for all of the variants, we repeated the thermal denaturation experiments in the presence of 40% glycerol, a strong renaturant. These data are shown in Figure 8, again as apparent fractions of unfolded protein. All of the proteins, including the R-6 variant, can be fully folded under these conditions. Surprisingly, all of the proteins seem to possess an approximately equivalent amount of cooperativity in these transitions. This includes the minimalist protein M-5 and the random control protein R-6, neither of which was designed with respect to specific packing interactions. Again, the C-1 protein exhibits a slightly more cooperative unfolding transition than the designs. In general, the glycerol data are completely consistent with the water data, both qualitatively and quantitatively: van't Hoff enthalpy values and melting temperatures are both well correlated (not shown). This indicates that the presence of glycerol leads to the desired stabilization without significantly affecting the relative order of stabilities or the relative levels of cooperativity observed.

Under conditions at which all of the variants are close to being fully folded, there are no significant differences in the CD spectra taken in the range of 200–300 nm (data not shown). This is consistent with all of the variants having a folded structure similar to the native protein.

1D NMR of selected 434 cro variants

Designed proteins typically have very poorly dispersed and broadened NMR spectra due to a combination of exchange broadening and chemical shift averaging caused by a dynamic folded state. 1D NMR provides a qualitative characterization of the behavior of the 434 cro variants that is complementary to that provided by thermal unfolding cooperativity. We collected 1D proton NMR spectra for the three representative 434 cro variants C-1, D-8, and M-5 at 7 °C, where the variant M-5 is maximally folded. These data are shown in Figure 9.

The NMR spectra are surprising in that all of the spectra are reasonably well dispersed relative to the spectrum obtained for

Table 1. *Thermodynamic parameters*[a]

Protein	T_m (°C)	ΔH_{vH} (kcal mol^{-1})	ΔC_p (kcal mol^{-1} K^{-1})
C-1	56	31.3	0.70
D-5	60	29.1	0.59
D-8	50	24.0	0.66
D-7	17	5.2	0.73
M-5	33	13.9	0.70

[a] Thermodynamic parameters for unfolding of 434 cro variants. Van't Hoff enthalpies, melting temperatures, and ΔC_p values were derived from curve-fitting analysis of the thermal denaturation data. Because values for ΔC_p cannot be determined with high accuracy from these data alone, we report the values here for completeness only. However, a plot of the enthalpy values versus melting temperatures r͜veals that the values of ΔC_p are approximately correct.

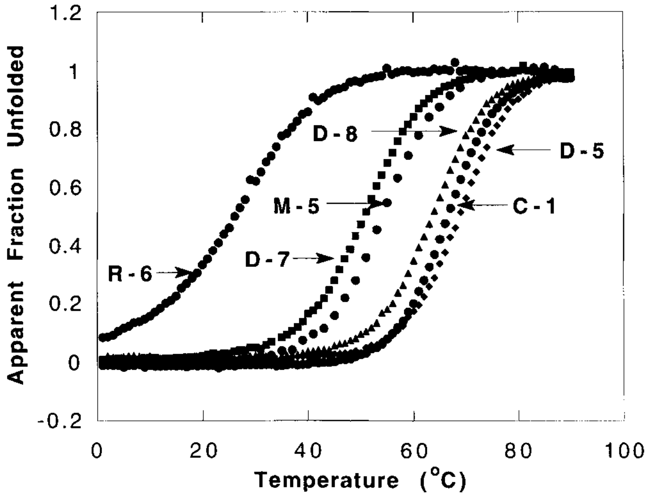

Fig. 8. Thermal denaturation of 434 cro variants in the presence of 40% glycerol. Data are plotted as described in Figure 6. Note the observable folding of the variant R-6 and the same relative order of thermal stabilities as in Figure 6.

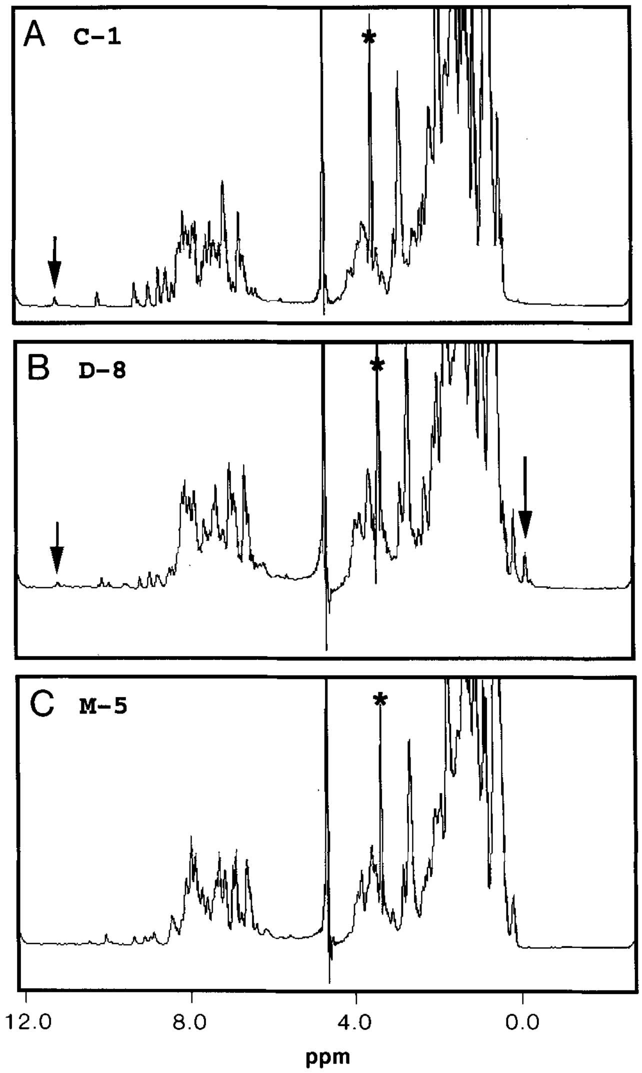

Fig. 9. 1D NMR of selected 434 cro variants. Spectra are shown for the variants (**A**) C-1, (**B**) D-8, and (**C**) M-5, representing native, designed, and minimalist proteins, respectively. The asterisk marks a small-molecule impurity seen in all of the spectra. The large arrow in B marks a potentially ring current-shifted methyl group that is absent in A and C. There is also a smaller peak upfield of this one that is difficult to see in the figure. Small arrows in A and B mark an unusually downfield-shifted amide resonance (also difficult to observe).

the native C-1. This implies that for these representative variants, the folded state is well ordered. Although this interpretation is not quantitative, the level of dispersion seen for all three variants is much higher than that observed for a typical de novo-designed protein. This is potentially surprising for the variant M-5, which might have been expected to fold into a highly dynamic structure. These observations are consistent with the interpretation of the cooperativity of thermal denaturation determined for all of the variants.

Of the three variants examined by NMR, only D-8 has an aromatic residue included in the hydrophobic core. An intriguing feature of the NMR spectrum of D-8 are two apparent ring current shifted methyl groups below 0 ppm. These resonances are absent in either of the other spectra, suggesting that the phenylalanine in position 4 of D-8 is uniquely placed in the folded structure. Based on the model structure of D-8, we suggest that the shifted resonances are from protons within the methyl groups of leucine 36. The variant D-8 and the native C-1 protein also share an upfield-shifted amide proton resonance found beyond 11 ppm, whereas M-5 contains no resonance at this frequency. This suggests the possibility that the local structure of that amide proton is not preserved in M-5.

Discussion

This study was initiated to examine the impact of hydrophobic core design on the thermodynamic and structural behavior of a protein. An integral component of such an examination is, of course, the assessment of our ability to design a hydrophobic core that compares favorably to a hydrophobic core selected by natural evolution. Although several computer programs designed for this purpose have been reported, we are aware of only one experimental application of such a program, in which a fraction of the total core was redesigned (Hurley et al., 1992). Here we have presented a newly developed set of programs that are appropriate for the design of hydrophobic cores consisting of a large number of residues. Because the intended use of the programs is in application to de novo protein design, we apply the program simultaneously to all core residues. To assess this potential, several designed and control sequences for the core of 434 cro have been characterized experimentally.

Three aspects of the design of hydrophobic cores are important in the de novo design process. First, the core sequence must be compatible with the target structure of the protein. Second, the core sequence must confer optimal stability to the folded state of the protein. Finally, the core should be consistent with a uniquely folded state by being inconsistent with other folded states. Here "other folded states" does not refer to gross changes in global topology but rather significant excursions of atomic coordinates from the average structure.

The effect of hydrophobic core mutations on the stability of proteins has been a subject of intensive study (Sandberg & Terwilliger, 1989; Shortle et al., 1990; Lim & Sauer, 1991; Eriksson et al., 1992, 1993; Lim et al., 1992; Jackson et al., 1993) and is well covered in a recent review by Richards and Lim (1993). Several general classes of mutations are possible. Using the terminology of Richards and Lim, substitutions that conserve hydrophobicity can be classified as disruptive and nondisruptive. Nondisruptive mutations involve a loss in side-chain volume by the removal of atomic groups while the remaining side-chain atoms retain a similar geometry. The best example of this type is a substitution of valine for isoleucine. Disruptive mutations involve a change in the covalent connectivity of the side-chain atoms, such as in an isoleucine to leucine substitution or an increase in volume. Numerous studies have shown, through analysis of the effects of these different types of substitutions, that specific packing interactions are important for the stability of the protein. This is most clearly demonstrated by the existence of significantly destabilizing pairs of mutations that preserve composition (Sandberg & Terwilliger, 1989). Because most previous studies have focused on, at most, five mutations in a hydrophobic core (Lim & Sauer, 1989; Baldwin et al., 1993), the data presented herein serve to extend these observations to include substitutions in a larger number of core positions. A recent report describes the use of existing patterns in the protein Rop to guide replacements in the hydrophobic core, resulting in a few proteins with a very large number of substitutions (Munson et al., 1994). These guided substitutions generally lead to an impressive increase in the thermal stability of the protein but a slight decrease in the stability determined using chemical denaturation.

The variants characterized in this report contain different numbers of potentially disruptive mutations as well as moderately different volumes. The important differences between variants are the extent to which packing requirements were used in their design. The randomly designed variant R-6 consists solely of conservative volume mutations but contains four disruptive mutations. This protein is shown to be completely unfolded in water; further experiments in 40% glycerol imply that this variant is severely destabilized relative to the native protein. Different effects are seen for the minimalist design M-5, which consists predominantly of leucine core residues, but also contains four potentially disruptive mutations. This protein is significantly destabilized but not nearly to the same extent as R-6. It is possible that the ability of leucines to adopt a large number of side-chain conformations reduces the impact of the four disruptive mutations. At another extreme, it may have been expected that the presence of many leucines in a hydrophobic core would lead to a more dynamic folded state and a high stability due to an increase in conformational entropy. This influence, if present, is apparently not sufficient to overcome the loss of stability due to the lack of good packing interactions.

All of the designed variants contain at least four potentially disruptive mutations. The variant D-5, which contains four disruptive mutations, is more thermally stable than the native C-1. The variant D-8 has eight mutations relative to the native, six of which are potentially disruptive mutations. This protein is only slightly destabilized relative to the native, which is surprising given the number of substitutions. D-8, as well as D-5, was designed with respect to packing interactions, suggesting that packing geometries significantly different from that of the native can be successfully engineered using our computational approach.

The remaining variant D-7 contains six potentially disruptive mutations and was also designed computationally. Indeed, it scores extremely well in our program as being both well packed and having low energy side-chain orientations relative to the backbone. This protein, however, is significantly destabilized relative to the native and is even slightly less stable than the minimalist variant M-5. It is unclear why this is so. It is possible that factors such as secondary structural propensities influence the stability of the variant D-7, especially considering the presence of two extra valines in α-helical positions.

Our program, written with van der Waals' potentials alone, was created as a method for selecting potentially permissive sequences from a very large sequence space but only as a crude predictive tool for the energetics of variants. Given the current level of understanding of effects such as secondary structure propensities in buried positions, side-chain entropy loss upon burial, effects of mutations on the denatured state, and other effects, it is impossible to account for more detailed differences in the experimentally determined stabilities of these proteins. This view was confirmed in earlier studies (Hurley et al., 1992), where energy minimization, helical propensity, buried surface areas, and side-chain entropy loss were all included in the calculated energies. The authors noted that no correlation is seen between observed and calculated stabilities. In a different study, empirical weights for energetic parameters were derived from a large database of experimental energies (Wilson et al., 1991). The derived weights were then used to successfully design mutants of α-lytic protease with altered substrate specificity. Similar attempts on 434 cro variants would require a much larger set of variants than that presented herein. Finally, the lack of detailed knowledge about the unfolded state may be particularly relevant in the case of 434 cro, as the homologous 434 repressor has been shown to contain residual structure in the urea-induced unfolded state (Neri et al., 1992).

A potentially major contribution to the inaccuracy of the energetic calculations is the fixed backbone assumption used in the current version of our program. The ability of the backbone to relax to accommodate hydrophobic core mutations is now well documented (Eriksson et al., 1992, 1993; Baldwin et al., 1993; Lim et al., 1994). The effects of this relaxation on the energetics are not accounted for in the calculated energies of the different variants. The fixed backbone assumption is an acknowledged limitation of any existing hydrophobic core design methodology. In the current study, we are designing novel cores for an existing backbone structure, using a simple set of potentials for selecting permissive sequences. It is doubtful that overcoming fixed backbone limitations would significantly increase the success of our predictions. For de novo design efforts, however, where an exact backbone structure may be unknown, a methodology that allows backbone variation may be extremely important. In light of this concern, it is promising that the variant D-8, which was selected on the basis of good packing but unimpressive side-chain–backbone energies, leads to a stable folded structure. This suggests that selecting sequences based on good packing with only reasonable side-chain–backbone energies will be a useful trick for de novo design efforts, especially if combined iteratively with energy minimization methods.

The effect of hydrophobic core design on the uniqueness of the folded structure is difficult to determine at this point. The surprising result of our studies is that all of the variants exhibit approximately equal amounts of cooperativity in their thermal unfolding transitions. Consistent with this, high amounts of dispersion are seen in proton NMR spectra collected for variants representing designed and nondesigned sequences. This is in stark contrast to the low levels of dispersion present in a typical de novo-designed protein (Handel et al., 1993). In the absence of any other data, one might conclude from these observations that rational hydrophobic core design is unnecessary for specifying uniqueness of the folded state. We suggest the alternative conclusion that uniqueness is overdetermined in natural proteins. Hence, there may be enough information in the non-core sequence to enforce uniqueness of structure at the current level of observation. If this is the case, de novo design efforts to incorporate non-core tertiary interactions involving hydrogen bonds, electrostatic interactions, and other specific interactions are well placed. On the other hand, in a total de novo design scenario, hydrophobic core design may in fact contribute significantly to reducing the number of folded conformations.

A reasonable conclusion from the data discussed above is that the expectations of the program have been fulfilled. In general, sequences selected by our program as permissive lead to stable folded structures, whereas sequences designed with very crude criteria are not generally successful. Indeed, the variant D-5, based on a strongly selected sequence, appears to have a stability greater than that of the native protein. In addition, a large number of substitutions are shown to be tolerated in the case of the variant D-8, where specific packing interactions are modeled explicitly.

Although stability may be a loose correlate of structural integrity, structural studies are necessary to define the true extent to which the variants differ in structure from the native protein and the extent to which they match our computer predictions. The present study raises the following questions. Do the designed variants have structures more similar to the native structure? Are the designed variants functionally active as DNA binding proteins? And how do the dynamics of the variants compare? These questions will be addressed in future structure and function studies of these partially designed proteins. The computational efforts presented herein are intended as a first test of packing predictability; structural characterization of the variants described will hopefully lead to an improvement of the current methodology.

Materials and methods

Program descriptions

The first of the programs, called NBSEARCH, is given an input structure in Protein Data Bank format, a list of core positions chosen by visual inspection of the structure, and a list of potential residue types (typically V, I, L, F, and W) to search through for each position. The search is done exhaustively through all reasonable values of the side-chain torsion angles $\chi 1$ and $\chi 2$ at 5° increments. The allowed ranges for $\chi 1$ in degrees are 50–90, 125–225, and 250–325. Those for $\chi 2$ are 25–120, 125–220, and 240–285. For selection of side-chain conformers out of these searches, only the six lowest energy orientations within each $\chi 1$ range are retained. This selective retention represents the major decrease in the number of allowed conformers. For completeness, we have typically supplemented the custom library with a statistically derived rotamer library (Tuffery et al., 1991). In general, however, this does not significantly affect the performance of the program: most of the final rotamer values chosen using the program ROC (see below) are directly from our custom library. This is presumably a result of the custom fit of our rotamer values to a given backbone position together with a finer sampling level. Because methionine residues have three dihedral angles, custom searches are impractical. Thus, for the genetic simulations described in Figure 2, the methionine rotamer values were derived solely from the statistical library.

Side-chain geometries, atomic parameters, and energetic potentials are equivalent between the two programs. Side chains with defined torsion angles are built by subroutines within each program, using bond length and bond angle geometries as defined within the molecular modeling program INSIGHT II (Biosym Technologies, San Diego, California). Hydrogen atoms are included explicitly, except for methyl group hydrogens. In this case, a united atom centered at the carbon atom was found to give better results. Hydrogen atoms were added to the crystal structure using the program INSIGHT II. Atomic parameters for all atom types are listed in Table 2. Parameters were originally taken from Hagler (Dauber & Hagler, 1980). Two empirical adjustments were made to the parameters. First, a uniform reduction in atomic radius was applied. The choice of this reduction parameter was the one that gave the best overall qualitative agreement of rotamer probabilities within a model α-helix and β-strand to those derived statistically from a set of protein structures (McGregor et al., 1987). Second, a standard energetic adjustment for the placement of a given residue type into a position was defined. A search within the standard torsion angle ranges (defined above) was conducted for each residue within a model α-helix; a standard adjustment was defined as the amount required to bring the energy of the lowest energy orientation to a zero value. Further minor adjustments were made empirically, using subjective evaluation of prediction success on a set of protein structures as criteria (different from the test proteins described in Fig. 1). Additional adjustments to phenylalanine atomic parameters were judged by the same criteria.

Nonbonded energies between atoms i and j are defined by the following standard functional forms:

$$R = 2\sqrt{(R_i \cdot R_j)}$$

$$E = [(R/d)^{12} - 2(R/d)^6] \cdot \sqrt{(e_i \cdot e_j)},$$

where R_i, R_j and e_i, e_j are the van der Waals' radii and well depth, respectively, for atoms i and j, and d is the distance be-

Table 2. *Atomic parameters*[a]

	Radius (Å)	ϵ (a.u.)
Hydrogen	1.23	0.038
Methyl hydrogen	0.00	–
Carbon	1.94	0.039
Backbone carbonyl carbon	1.81	0.148
Methyl carbon	1.95	0.160
Aromatic carbon	1.78	0.110
Nitrogen	1.75	0.167
Oxygen	1.43	0.228
Sulfur	1.95	0.160

[a] Atomic parameters used in the programs NBSEARCH and ROC. The ϵ values are in arbitrary units.

tween atoms i and j. The E term is capped at 100, and interactions outside of 8 Å are neglected.

The second program, called ROC, reads an input structure in PDB format, a list of core positions, residues to be included in the search for each core position, and a rotamer library customized for each position (output from the first program). The program then first evaluates all potential side-chain–backbone energies and all potential side-chain–side-chain energies. Optimization for a low energy core sequence/structure is then achieved using a genetic algorithm (Holland, 1992). Briefly, a population of strings of information, each string encoding a complete model structure, is created at each round of the genetic algorithm. Although the original population is created randomly, each subsequent population is generated by recombination between strings in the current round. This is performed such that strings encoding model structures with lower energies recombine most frequently with other strings with low energies. This weighted recombination is the selection pressure for "evolution" of strings encoding the best model structures. The force field dictating the "fitness" of each structure can be thought of as the "environment" in which the strings are selected to survive. The encoding of model structures into strings is trivial for the current problem: each bit in the string encodes a given residue type with defined torsion angles. The location of the bit within the string is related to the core position. The input rotamer library for each core position is thus a list of residue/torsion possibilities for the string location corresponding to the core position. An inversion operator within the genetic algorithm encourages different linear orderings of the bits, allowing the potential for genetic "linkage" between pairs of bits.

A typical run of the genetic algorithm is for 500 rounds of recombination, inversion, and mutation, resulting in convergence to a single sequence/structure. Usually 100 supercycles (one supercycle = one run of 500 rounds) are done in order to generate multiple outputs. To encourage different output for each new supercycle, two perturbations to the calculated energies are made. First, the weight of the side-chain–side-chain energies relative to the side-chain–backbone energies is made incrementally larger as the supercycles progress, up to a factor of two for the final supercycle. Second, for each new supercycle, a random energy perturbation of up to 0.5 energy units is made to each possible rotamer throughout all rotamer libraries. After completion of all supercycles, these energetic perturbations are filtered out to yield the "true" calculated energies of all final sequences/structures, where all energy terms are equally weighted and no random perturbations are included. A typical run of 100 supercycles of 500 rounds each takes 1–3 h to run on a 150-MHz R4400 processor of a Silicon Graphics Indigo2 computer. Output consists of a report of the sequence, rotamer values, and the "perturbed" and "true" energies for the final sequence/structure of each supercycle. Additional output includes a table of all unique sequences, the number of occurrences of each sequence, and the lowest energy found for each energy term throughout all occurrences of a particular sequence. A final program has been written that creates a model structure directly from any of the output sequence/rotamer lists and the input structure.

Construction of 434 cro variant genes

Synthetic genes encoding each of the 434 cro variant proteins were constructed individually from a set of 6 synthetic oligonucleotides, based predominantly on the wild-type 434 cro gene sequence, with the addition of a cysteine codon (for eventual chemical cleavage) preceding the methionine start codon. The ligated synthetic genes were subcloned into a system for expression of a fusion protein between the 434 cro variant and the insoluble protein LE′. This system was derived from the vector pTrpLE′ (Kleid et al., 1981), which contains the LE′ gene expressed from a Trp promoter. We adapted the LE′ fusion into a T7 RNA polymerase system by subcloning the LE′ gene into the vector pG5 (Alexander et al., 1992).

Expression and purification of 434 cro variants

LE′–434 cro fusion proteins were produced in BL21-pLysS (Studier et al., 1990) *Escherichia coli* cells, using 0.5 mM IPTG for induction of protein expression. Cells were collected by centrifugation typically after 3–4 h of growth postinduction and frozen at −80 °C. Cells were lysed by thawing and the addition of 0.1% Triton-X 100. The insoluble fusion protein was purified as described (Kleid et al., 1981).

434 cro or variants were then separated from the fusion protein using a chemical cleavage at cysteine residues (Jacobson et al., 1973). This cleavage method necessitated the replacement of the wild-type cysteine residue at position 54 with a valine. Fusion protein was solubilized in a solution of 6 M urea (USB) and 0.1 M Tris-acetate, pH 8.0. Approximately 0.1 g of 2-nitro-5-thiocyanobenzoic acid (Sigma) were added to the solution, which was then incubated at room temperature for 30 min to 1 h. The pH of this solution was then raised to 9.5 by the addition of NaOH, and the solution was incubated overnight at 37 °C to enhance cleavage.

The resulting solution was then loaded onto a column containing approximately 20 mL of SP-Sephadex C-25 resin (Pharmacia) equilibrated in 6 M urea, 50 mM Tris, pH 8.0, 50 mM NaCl. The column was washed with 60 mL of the same buffer. The column containing 434 cro protein was then washed with 50 mM Tris, pH 8.0, 50 mM NaCl to remove the urea. Finally, 434 cro protein was eluted with 50 mM Tris, pH 8.0, 0.8 M NaCl. The eluted 434 cro protein was further purified by reverse-phase HPLC on a Vydac C-18 semipreparative column.

CD and thermal denaturation

All CD experiments were performed on an Aviv 62DS CD spectrometer. Samples were prepared in 10 mM Tris, pH 7.0, or in the case of the glycerol experiments, in 10 mM Tris, pH 7.0, 40% glycerol. Spectra were collected at 1- or 2-nm increments across the range 200–300 nm, with a typical signal averaging time of 10–20 s.

Thermal denaturation experiments were performed at 1 or 2 °C temperature increments, with an equilibration time of 1 min, and a typical signal averaging time of 20 s. Unfolding of protein was detected by monitoring the CD signal at 222 nm.

Thermal denaturation profiles were fitted, assuming a two-state equilibrium, to the following equation using the program Kaleidagraph (Synergy Software, Reading, Pennsylvania):

$$\text{Signal} = [U(T) - L(T)]/[1 + \exp(-\Delta G/RT)] + L(T).$$

ΔG is given by the Gibbs–Helmholtz equation:

$$\Delta G = \Delta H(T_m) + \Delta C_p[T - T_m] - T\{\Delta H(T_m)/T_m + \Delta C_p \cdot \ln[T/T_m]\},$$

$U(T)$ and $L(T)$ (upper and lower) are functions of the form $y = mT + b$ representing temperature-dependent baselines for unfolded and folded protein, respectively, T_m is melting temperature, ΔC_p is the heat capacity increment upon unfolding, and $\Delta H(T_m)$ is the van't Hoff enthalpy of unfolding at the melting temperature.

Melting temperatures as well as folded and unfolded temperature-dependent baselines were determined in the curve-fitting analysis and confirmed by visual inspection. Apparent fractions of unfolded protein (as shown in Figs. 7, 8) were determined by applying the relation $F_{app} = [\text{signal}(T) - L(T)]/[U(T) - L(T)]$ to the actual data using the upper and lower sloped baselines determine by the fits. For the two variants M-5 and D-7, folded baselines are not observed directly in the raw data. However, the data fit well to a two-state equilibrium with values that are consistent with the values derived for the more stable proteins and are also consistent with values determined from the glycerol experiments.

1D NMR

NMR spectra were collected on a 600-MHz DMX spectrometer equipped with 3-axis pulsed field gradients on samples that were approximately 200 μM in 90% H_2O/10% D_2O, pH 6.0. One-dimensional spectra were recorded at 7 °C as the first slice of a NOESY sequence with essentially no mixing time (5 μs) and a WATERGATE (Sklenar et al., 1993) pulse train as the read pulse to enhance water suppression. Data were collected with 1K complex points, apodized with a 90°-shifted sine bell, and zero-filled to 2K points prior to Fourier transformation.

Acknowledgments

T.M.H. is an NSF Young Investigator. J.R.D. is a fellow of the Jane Coffin Childs Memorial Fund for Medical Research, fellowship number 61-948. This work was supported in part by an NSF grant, MCB-9458201. We thank the Dupont Merck Company for supporting J.R.D. for his first year and a half as a postdoctoral fellow, and Susan Marqusee for use of her CD spectrometer.

References

Alexander P, Fahnestock S, Lee T, Orban J, Bryan P. 1992. Thermodynamic analysis of the folding of the streptococcal protein G IgG-binding domains B1 and B2: Why small proteins tend to have high denaturation temperatures. *Biochemistry 31*:3597-3603.

Baldwin EP, Hajiseyedjavadi O, Baase WA, Matthews BW. 1993. The role of backbone flexibility in the accommodation of variants that repack the core of T4 lysozyme. *Science 262*:1715-1718.

Beamer LJ, Pabo CO. 1992. Refined 1.8 angstroms crystal structure of the lambda repressor-operator complex. *J Mol Biol 227*:177-196.

Betz SF, Raleigh DP, DeGrado WF. 1993. De novo protein design: From molten globules to native-like states. *Curr Opin Struct Biol 3*:601-610.

Chothia C. 1975. Structural invariants in protein folding. *Nature 254*:304-308.

Dauber P, Hagler AT. 1980. Crystal packing, hydrogen bonding, and the effect of crystal forces on molecular conformation. *Acc Chem Res 13*: 105-112.

DeGrado W, Raleigh D, Handel T. 1991. Protein design, what are we learning? *Curr Opin Struct Biol 1*:984.

DeGrado WF, Wasserman ZR, Lear JD. 1989. Protein design, a minimalist approach. *Science 243*:622-628.

Eriksson AE, Baase WA, Matthews BW. 1993. Similar hydrophobic replacements of Leu99 and Phe153 within the core of T4 lysozyme have different structural and thermodynamic consequences. *J Mol Biol 229*:747-769.

Eriksson AE, Baase WA, Zhang XJ, Heinz DW, Blaber M, Baldwin EP, Matthews BW. 1992. Response of a protein structure to cavity-creating mutations and its relation to the hydrophobic effect. *Science 255*:178-183.

Handel TM, Williams SA, DeGrado WF. 1993. Metal ion-dependent modulation of the dynamics of a designed protein. *Science 261*:879-885.

Hecht MH, Richardson JS, Richardson DC, Ogden RC. 1990. De novo design, expression, and characterization of Felix: A four-helix bundle protein of native-like sequence. *Science 249*:884-891. [Published erratum appears in *Science 249*(4972):973.]

Hellinga HW, Richards FM. 1994. Optimal sequence selection in proteins of known structure by simulated evolution. *Proc Natl Acad Sci USA 91*:5803-5807.

Holland JH 1992. *Adaptation in natural and artificial systems.* Cambridge, Massachusetts: The MIT Press.

Hurley JH, Baase WA, Matthews BW. 1992. Design and structural analysis of alternative hydrophobic core packing arrangements in bacteriophage T4 lysozyme. *J Mol Biol 224*:1143-1159.

Jackson SE, Moracci M, elMasry N, Johnson CM, Fersht AR. 1993. Effect of cavity-creating mutations in the hydrophobic core of chymotrypsin inhibitor 2. *Biochemistry 32*:11259-11269.

Jacobson GR, Schaffer MH, Stark GR, Vanaman TC. 1973. Specific chemical cleavage in high yield at the amino peptide bonds of cysteine and cystine residues. *J Biol Chem 248*:6583-6591.

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

Katti SK, LeMaster DM, Eklund H. 1990. Crystal structure of thioredoxin from *Escherichia coli* at 1.68 Å resolution. *J Mol Biol 212*:167-184.

Kleid DG, Yansura D, Small B, Dowbenko D, Moore DM, Grubman MJ, McKercher PD, Morgan DO, Robertson BH, Bachrach HL. 1981. Cloned viral protein vaccine for foot-and-mouth disease: Responses in cattle and swine. *Science 214*:1125-1229.

Kono H, Doi J. 1994. Energy minimization method using automata network for sequence and side-chain conformation prediction from given backbone geometry. *Proteins Struct Funct Genet 19*:244-255.

Lim WA, Farruggio DC, Sauer RT. 1992. Structural and energetic consequences of disruptive mutations in a protein core. *Biochemistry 31*:4324-4333.

Lim WA, Hodel A, Sauer RT, Richards FM. 1994. The crystal structure of a mutant protein with altered but improved hydrophobic core packing. *Proc Natl Acad Sci USA 91*:423-427.

Lim WA, Sauer RT. 1989. Alternative packing arrangements in the hydrophobic core of lambda repressor. *Nature 339*:31-36.

Lim WA, Sauer RT. 1991. The role of internal packing interactions in determining the structure and stability of a protein. *J Mol Biol 219*:359-376.

McGregor MJ, Islam SA, Sternberg MJ. 1987. Analysis of the relationship between side-chain conformation and secondary structure in globular proteins. *J Mol Biol 198*:295-310.

Mondragon A, Subbiah S, Almo SC, Drottar M, Harrison SC. 1989. Structure of the amino-terminal domain of phage 434 repressor at 2.0 Å resolution. *J Mol Biol 205*:189-200.

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

Munson M, O'Brien R, Sturtevant JM, Regan LR. 1994. Redesigning the hydrophobic core of a four-helix-bundle protein. *Protein Sci 3*:2015-2022.

Neri D, Billeter M, Wider G, Wüthrich K. 1992. NMR determination of residual structure in a urea-denatured protein, the 434-repressor. *Science 257*:1559-1563.

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

Quinn TP, Tweedy NB, Williams RW, Richardson JS, Richardson DC. 1994. Beta-doublet: De novo design, synthesis, and characterization of a beta-sandwich protein. *Proc Natl Acad Sci USA 91*:8747-8751.

Raleigh DP, DeGrado WR. 1992. A de novo designed protein shows a thermally induced transition from a native to a molten globule-like state. *J Am Chem Soc 114*:10079-10081.

Regan L, Clarke ND. 1990. A tetrahedral zinc(II)-binding site introduced into a designed protein. *Biochemistry 29*:10878-10883.

Richards FM. 1977. Area, volumes, packing, and protein structure. *Annu Rev Biophys Bioeng 6*:151-176.

Richards FM, Lim WA. 1993. An analysis of packing in the protein folding problem. *Q Rev Biophys 26*:423–498.

Sandberg WS, Terwilliger TC. 1989. Influence of interior packing and hydrophobicity on the stability of a protein. *Science 245*:54–57.

Schindelin H, Jiang W, Inouye M, Heinemann U. 1994. Crystal structure of CspA, the major cold shock protein of *Escherichia coli*. *Proc Natl Acad Sci USA 91*:5119–5123.

Shortle D, Stites WE, Meeker AK. 1990. Contributions of the large hydrophobic amino acids to the stability of staphylococcal nuclease. *Biochemistry 29*:8033–8041.

Sklenar V, Piotto M, Leppik R, Saudek V. 1993. Gradient-tailored water suppression For H-1-N-15 HSQC experiments optimized to retain full sensitivity. *J Magn Reson 102*:241–245.

Studier FW, Rosenberg AH, Dunn JJ, Dubendorff JW. 1990. Use of T7 RNA polymerase to direct expression of cloned genes. *Methods Enzymol 185*: 60–89.

Tuffery P, Etchebest C, Hazout S, Lavery R. 1991. A new approach to the rapid determination of protein side chain conformations. *J Biomol Struct & Dyn 8*:1267–1289.

Weaver LH, Matthews BW. 1987. Structure of bacteriophage T4 lysozyme refined at 1.7 angstroms resolution. *J Mol Biol 193*:189–199.

Wilson C, Mace JE, Agard DA. 1991. Computational method for the design of enzymes with altered substrate specificity. *J Mol Biol 220*:495–506.

Wlodaver A, Pavlovsky A, Gustchina A. 1992. Crystal structure of human recombinant interleukin-4 at 2.25 Å resolution. *FEBS Lett 309*:59–64.

Yan Y, Erickson B. 1994. Engineering of betabellin 14D: Disulfide-induced folding of a β-sheet protein. *Protein Sci 3*:1069–1073.

Zhang JD, Cousens LS, Barr PJ, Sprang SR. 1991. Three-dimensional structure of human basic fibroblast growth factor, a structural homolog of interleukin 1 beta. *Proc Natl Acad Sci USA 88*:3446–3450. [Published erratum appears in *Proc Natl Acad Sci USA 88*(12):5477.]