Protein Science (1998), 7:1898–1914. Cambridge University Press. Printed in the USA.

Effects of salt bridges on protein structure and design

CHARLES V. SINDELAR, ZACHARY S. HENDSCH, AND BRUCE TIDOR
Department of Chemistry, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307

(RECEIVED March 2, 1998; ACCEPTED May 18, 1998)

Abstract

Theoretical calculations (Hendsch ZS & Tidor B, 1994, *Protein Sci 3*:211–226) and experiments (Waldburger CD et al., 1995, *Nat Struct Biol 2*:122–128; Wimley WC et al., 1996, *Proc Natl Acad Sci USA 93*:2985–2990) suggest that hydrophobic interactions are more stabilizing than salt bridges in protein folding. The lack of apparent stability benefit for many salt bridges requires an alternative explanation for their occurrence within proteins. To examine the effect of salt bridges on protein structure and stability in more detail, we have developed an energy function for simple cubic lattice polymers based on continuum electrostatic calculations of a representative selection of salt bridges found in known protein crystal structures. There are only three types of residues in the model, with charges of −1, 0, or +1. We have exhaustively enumerated conformational space and significant regions of sequence space for three-dimensional cubic lattice polymers of length 16. The results demonstrate that, while the more highly charged sequences are less stable, the loss of stability is accompanied by a substantial reduction in the degeneracy of the lowest-energy state. Moreover, the reduction in degeneracy is greater due to charges that pair than for lone charges that remain relatively exposed to solvent. We have also explored and illustrated the use of ion-pairing strategies for rational structural design using model lattice studies.

Keywords: electrostatics; ion pairs; lattice models; protein design; protein electrostatics; protein stability; salt bridge; specificity

Of fundamental interest to protein scientists is not only the relative strengths of interactions in macromolecules, but also the interplay among individual interactions and the relationship between structural interactions and overall function. Studies of "compound" and "cooperative" interactions have expanded our understanding beyond simply excluded volume, dispersion interactions, the hydrophobic effect (Kauzmann, 1959), hydrogen bonds (Baker & Hubbard, 1984; Fersht et al., 1985; Yang et al., 1992), salt bridges (Fersht, 1971, 1972; Marqusee & Baldwin, 1987; Anderson et al., 1990; Horovitz et al., 1990; Serrano et al., 1990; Dao-pin et al., 1991; Merutka & Stellwagen, 1991; Lyu et al., 1992; Huyghues-Despointes et al., 1993; Hendsch & Tidor, 1994; Waldburger et al., 1995; Wimley et al., 1996) and local backbone and side-chain entropy (Page & Jencks, 1971; Matthews et al., 1987; Straatsma & McCammon, 1989; Tobias et al., 1989; Pickett & Sternberg, 1993) to include more aggregate and delocalized effects. Examples range from the medium-ranged, such as helix-capping interactions (Hol et al., 1978; Sheridan et al., 1982; Hol, 1985; Shoemaker et al., 1987; Nicholson et al., 1988, 1991; Presta & Rose, 1988; Richardson & Richardson, 1988; Šali et al., 1988; Fairman et al., 1989; Serrano & Fersht, 1989; Lyu et al., 1993; Hendsch & Tidor, 1994; Tidor, 1994; Prévost, 1996) to the global, such as overall vibrational entropy (Page & Jencks, 1971; Irikura et al., 1985; Finkelstein & Janin, 1989; Tidor & Karplus, 1993, 1994; Gilson et al., 1997).

Reprint requests to: Bruce Tidor, Department of Chemistry, Room 6-135, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139-4307; e-mail: tidor@mit.edu.

Here we focus on the role of salt bridge interactions on the folding equilibrium of proteins. Recent theoretical and experimental evidence has suggested that an individual salt bridge may be less stabilizing than a structurally equivalent hydrophobic interaction (Hendsch & Tidor, 1994; Waldburger et al., 1995; Wimley et al., 1996). That is, salt bridges in proteins are generally destabilizing relative to replacement with hydrophobic groups of the same size and shape. It has been suggested that these essentially destabilizing electrostatic interactions may play a role in defining specificity, the uniqueness of the lowest free-energy state (the ground state), by two mechanisms: (1) the unfavorability of burying charged groups could reduce the degeneracy of the ground state by disfavoring charge burial altogether (Paul, 1982) and (2) the requirement to compensate buried charges through salt bridges and hydrogen bonds could also reduce ground-state degeneracy (Tanford et al., 1960; Hendsch & Tidor, 1994). Indeed, some experimental evidence for an increase in degeneracy upon replacement of charged or polar interactions with hydrophobic ones has been seen in the coiled coils (Harbury et al., 1993; Lumb & Kim, 1995). In experiments using the Arc repressor, no such increase in structural degeneracy was seen (Waldburger et al., 1995). This could be due to a larger network of cooperative interactions defining the ground-state structure in the Arc repressor as compared with the coiled coils.

The purpose of the current work is to explore the consequences of the notion that salt bridges destabilize the folding of proteins and to investigate principles of protein design using patterns of electrostatic and hydrophobic interactions. For this type of study, thorough searches of conformational and sequence spaces are es-

sential; it is necessary to determine the global free energy minimum structure (or structures, if the ground state is degenerate) for a given sequence and it is desirable to find families of sequences that fold to the same structure. Complete conformational searching using a full atomic model with independent Cartesian (or even just torsional) degrees of freedom for peptides or proteins is currently computationally prohibitive (Clarage et al., 1995) and searching sequence space using the full alphabet of the 20 common amino acids can be, likewise, daunting. In this study we used a simplified lattice representation of the protein, which significantly reduced the number of chain conformations available, as well as a reduced sequence alphabet. The advantage of these simplifications is that they allowed a complete search of conformational space and a large region of sequence space. Complete searching was accomplished using enumeration to examine every conformation and sequence that might be of interest. Enumeration, while generally considered primitive when compared to more sophisticated global optimization tools, such as simulated annealing (Kirkpatrick et al., 1983) and genetic algorithms (Unger & Moult, 1993), has a number of advantages. Principal among these is that there is no question of convergence; if the search is done in a discrete space, such as the lattice used in the current work, when the search is finished it is certain that the global minimum was located. Moreover, it is trivial to tabulate more information about the space than is generally available by other methods, such as to keep track of all conformations that share the ground-state energy or are within a certain tolerance (such as k_BT) of the ground-state energy.

Lattice simulation has proven to be a useful and robust method for studying protein polymers (Skolnick & Kolinski, 1989; Karplus & Shakhnovich, 1992; Dill et al., 1995). Work has been carried out using a number of different lattice types (ranging from simple two-dimensional square and three-dimensional cubic to diamond, knight's walk, and even more sophisticated spaces), chemical representations (Hinds & Levitt, 1992) and amino-acid alphabets [from the simple HP (for hydrophobic and polar) alphabet (Dill et al., 1995) to the 20-amino-acid alphabets (Miyazawa & Jernigan, 1985)]. Here we use a three-dimensional simple cubic lattice with one node representing a single residue and with three amino-acid types (hydrophobic, positively charged, and negatively charged). The proteins studied were of length 16 (i.e., they were peptides), which allowed enumeration of all conformations and many sequences. The alternative strategy of examining only maximally compact conformations allows study of longer polymers for similar computational effort (such as 27-mers confined to a $3 \times 3 \times 3$ cube), but was not used here because submaximally compact structures do populate ground states often in our model and would be missed. One disadvantage of this model is that such short polymers do not possess a hydrophobic core on three-dimensional cubic lattices but do possess such a core on two-dimensional square lattices (which are also computationally less demanding). However, the greater realism and conformational freedom of three-dimensional space caused us to choose cubic lattices here.

Results and discussion

Effects of charge on structure, energy, and degeneracy

The results of continuum electrostatic calculations on 21 protein salt bridges (Hendsch & Tidor, 1994) were fit to three similar free energy functions described in Methods. Briefly, each energy function included a hydrophobic term, which represents a favorable contribution for burying a residue in the folded state. This term was applied to all residues, but was the only term applied to hydrophobes. Charged residues also paid an electrostatic desolvation penalty and made electrostatic interactions with adjacent residues in the lattice. The hydrophobic and desolvation terms were of larger magnitude for more buried residues. The interaction term was stronger for more buried residues (representing a lower effective dielectric), and residues making multiple charge–charge interactions had their interactions scaled down to represent the limited ability of a side chain to make strong interactions in many directions simultaneously, as well as the expected repulsion between multiple salt bridge partners of an individual residue.

These energy functions were used to determine the ground-state energy and structure(s) for a set of 16-residue sequences using structural enumeration (systematic search). The sequences were all based on the parent sequence + 0000000000000 − (where 0 indicates a hydrophobe and + or − indicates a unit positive or negative charge, respectively); all sequences with up to four additional charges in place of hydrophobes in the parent sequence were studied. This set of sequences allowed us to examine the effect of adding charged groups to a mostly hydrophobic polymer; the positively and negatively charged termini of the parent sequence were retained to model the N- and C-termini of proteins. Statistics for ground-state energy and degeneracy are listed in Table 1 for all 19,321 sequences with up to four added charges studied (i.e., not including the charged termini) using each of three energy functions, which differ in the relative strength (low, middle, or high) of the hydrophobic term. Statistics were averaged over sequences with the same number of added charges.

The average folding energy became less favorable as charged residues were substituted in the parent sequence. For the middle-weighted energy function, the average energies for sequences with zero, one, two, three, and four additional charges were −17.5, −17.0, −16.4, −15.7, and −15.0 kcal/mol, respectively. This illustrates a feature of our energy functions (see Methods) that was a direct result of continuum electrostatic calculations (Hendsch & Tidor, 1994): The electrostatic penalty for charge burial outweighed favorable charge–charge interaction. Note that, because changing sequence can change the lowest energy structure(s), adding a charge did not necessarily increase the folding energy for every sequence.

The average degeneracy for sequences with zero, one, two, three, and four additional charges (middle-weighted energy function) were 267.0, 88.5, 27.9, 8.8, and 4.4, respectively. Clearly, added charge dramatically reduced ground-state degeneracy. This result is especially significant because it provides a possible explanation for the existence of salt bridges in proteins; namely, geometric restrictions required for the compensation of (partially) buried charged groups contribute significantly to the uniqueness of the folded state (Hendsch & Tidor, 1994). Furthermore, it suggests that there may be a trade-off between stability and uniqueness. That is, one class of very stable proteins might be extremely hydrophobic (e.g., the parent sequence) but highly degenerate. One strategy for designing sequences with less degenerate structures would be to include salt bridges (and other complementary polar interactions), which would generally require sacrificing folding energy. An alternative explanation for these data is that added charges reduce structural degeneracy not by forming buried salt bridges but by remaining unpaired on the surface. The requirement to not bury lone charges would be a reasonable mechanism for reducing structural degeneracy (Paul, 1982). This idea leads to the design principle of plac-

Table 1. *Average properties per sequence group by number of added charges*

Number of added charges[a]	Number of sequences[b]	Nondegenerate ground states[c] (%)	Average energy[d]	Average degeneracy[e]	Average number of contacts[f]
		Low-weighted energy function			
0	1	0	−11.0	267.0	13.0
1	28	0	−10.7	160.1	12.4
2	364	5	−10.0	64.5	12.5
3	2,912	18	−9.3	53.9	12.1
4	16,016	22	−8.8	20.9	11.7
		Middle-weighted energy function			
0	1	0	−17.5	267.0	13.0
1	28	0	−17.0	88.5	12.9
2	364	6	−16.3	27.8	12.9
3	2,912	19	−15.7	8.8	12.8
4	16,016	36	−15.0	4.4	12.7
		High-weighted energy function			
0	1	0	−24.0	267.0	13.0
1	28	0	−23.5	81.7	13.0
2	364	6	−22.8	23.7	13.0
3	2,912	17	−22.2	8.0	13.0
4	16,016	37	−21.4	3.8	12.9

[a]Each line in the table represents a class of sequences with the given number of added charges beyond the charged N- and C-termini of each 16-mer.
[b]The number of sequences in each class.
[c]The percentage of sequences in each class with a single structure in the ground state.
[d]The energy of the ground state, in kcal/mol, averaged over the sequences in each class.
[e]The degeneracy of the ground state averaged over the sequences in each class.
[f]The average number of contacts in the ground state averaged over the sequences in each class.

ing hydrophobic groups at buried positions and hydrophilic ones at exposed positions (Kametekar et al., 1993; Munson et al., 1994; Dahiyat & Mayo, 1997).

An analysis was performed to examine the question of whether lone charges or salt bridging charges played a larger role in reducing structural degeneracy. Figure 1 shows the average ground-state energy (A) and degeneracy (B) plotted as a function of the number of added total and lone charges (middle-weighted energy function). Each group of bars in the figure refers to the total number of added charges (zero, one, two, etc.); within each group of bars, the left-most bar indicates zero lone charges, the next bar one lone charge, etc. Figure 1A demonstrates that, for a fixed total number of added charges, more lone charges led to higher average energies and degeneracies, or, restated, adding charges in a manner that produced salt bridges in the ground state resulted in more stable and less degenerate ground states than resulted from an equal number of lone charges. This supports the notion that the reduction in degeneracy was due to specific salt bridge effects rather than simply to greater sequence heterogeneity with additional charges.

A related effect observed in Table 2 is that, for a given number of added charges, sequences in which the number of positive charges was balanced by the number of negative charges resulted in more unique structures than others with the same total number of added charges. For example, among sequences with four added charges, the highest fraction with nondegenerate ground states was in the group with two positive and two negative added charges (abbreviated [++−−]). Of these sequences, 39% uniquely selected the ground-state structure, compared to 36% for [+++−] and [−−−+], and just 24% for [++++] and [−−−−] (middle-weighted energy function). The average degeneracy, 3.5, was also lowest for the [++−−] group, followed by 4.2 for [+++−] and [−−−+] and 8.0 for [++++] and [−−−−]. These data show that adding charges with the potential to form salt bridges (i.e., matched pairs of oppositely charged residues) was generally more effective at reducing the degeneracy than adding other charge combinations (such as like-charged residues).

This difference between salt bridged and lone charges can be rationalized from the potential: Both types of charge increase the desolvation penalty for folding, but salt bridging interactions recover a partial compensation. In every case in Figure 1, adding a salt bridging charge increased the average energy less, and reduced the average degeneracy more, than adding a lone charge in a sequence group. This suggests that, in general, a protein engineer should introduce a salt bridge, rather than a pair of lone charges, to reduce degeneracy most efficiently (i.e., with minimal stability loss). An exception might be lone charges that extend out from the protein surface (not considered in this simple lattice model) that could reduce degeneracy without incurring a significant desolvation penalty upon folding. Also not addressed here is the relative stability effect of compensating a charge with polar groups, which might be qualitatively similar to compensation with ionic groups.

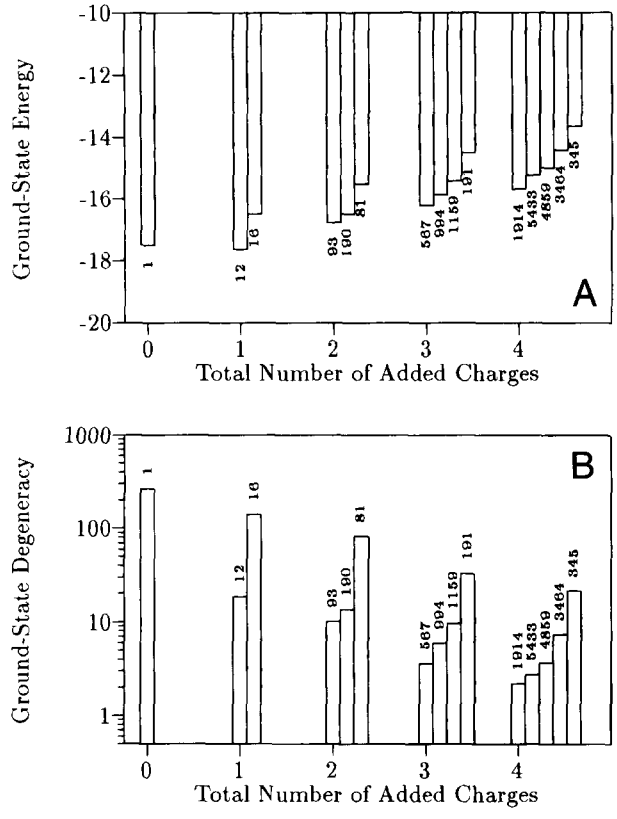

Fig. 1. The (**A**) ground-state energy and (**B**) degeneracy for all sequences of length 16 with up to four added charges using the middle-weighted energy function. Each group of bars refers to the total number of added charges (zero, one, two, etc.); within each group of bars, the leftmost bar indicates zero lone charges, the next bar indicates one lone charge, etc. The number above each bar is the number of sequences contributing, rounded to the nearest integer.

A significant fraction of our sequence space codes for nondegenerate structures. For the low-weighted energy function, 21% of the sequences investigated produced unique structures, while for the middle- and high-weighted energy functions the value was 33%. This result is surprising, especially given the relatively simple amino-acid alphabet (three types of residues: +1, 0, and −1 charge) and the relatively short length (nominally 16 residues) of proteins (peptides) modeled. Certainly the restrictions imposed by the lattice are a strong factor leading to low degeneracy (Honig & Cohen, 1996). The fraction of sequences producing a unique structure increased with added charge, in parallel with the decrease in average degeneracy per sequence. For the middle-weighted energy function, this fraction was 0% with one added charge, 6% with two added charges, 19% with three added charges, and 36% with four added charges (Table 1).

The average compactness (the number of inter-residue contacts, with 13 being "maximally compact" for a 16-node conformer on a simple cubic lattice) decreased somewhat with added charge [13.0, 12.9, 12.9, 12.8, and 12.7, respectively, for zero, one, two, three, and four additional charges (middle-weighted energy function)], although the hydrophobic term of the energy function strongly modulated the magnitude of this effect (Fig. 2). For the high-weighted energy function, all sequence groups remained quite compact, with average compactness of 13.0, 13.0, 13.0, 13.0, and 12.9, respectively, for zero, one, two, three, and four additional charges. The same sequence groups were much less compact for the low-weighted energy function, 13.0, 12.4, 12.5, 12.1, and 11.7, respectively. Interestingly, the decrease in average compactness with added charge was largely due to lone charges, and structures were fully compact if all charges were compensated by other charges. Indeed, in general, a weak hydrophobic component caused sequences with lone charges to lose compactness, while a strong hydrophobic component drove sequences to maximal compactness even if this required burying lone charges.

Designable regions of structure space

A structure (a trace on the simple cubic lattice) is defined here as "designable" if there was at least one sequence for which this was the nondegenerate lowest-energy state (Li et al., 1996). The number of 16-mer conformations is given as a function of number of contacts in Table 3. There were 2,134 maximally compact (13-contact) traces, 4,654 twelve-contact traces, 72,056 eleven-contact traces, and the number rapidly increased with even fewer contacts (interestingly, there were more 1-contact than 0-contact traces). Many of these traces could be designed from our 19,321-sequence set (with up to four added charges). For the middle-weighted energy function, 25% of the maximally compact traces could be designed, 10% of the 12-contact traces, just 22 of 72,056 (0.03%) 11-contact traces and no traces with fewer than 11 contacts. Larger numbers of additional charges allowed more structures to be designable (see below). One reason contributing to the lower designability of less compact structures was the occurrence of loops extending into solvent that could bend without altering the inter-residue contact pattern (Fig. 3A). The 22 eleven-contact designable structures were relatively noncompact, yet did not possess such loops. They fit into either of two shapes (i.e., spatial arrangement of monomer units, regardless of connectivity), which are shown in Figure 3B.

The first entry in Table 1 is for the 267 structures that were the lowest-energy degenerate conformations for the parent sequence, regardless of whether the low-, middle-, or high-weighted energy function was used. These structures were a subset of the 2,134 maximally compact structures for a 16-mer; all 267 contained an exposed salt bridge contact between the N- and C-termini. Nearly all of these (260) could be designed from at least one sequence with up to four added charges using any of the three energy functions. The remaining seven could be designed using sequences with more than four charges for the middle- and high-weighted energy functions (results not shown).

The 6,307 sequences (with up to four added charges) that produced unique structures (middle-weighted energy function) actually produced only 1,007 distinct structures, giving an average of six sequences that design each structure. Some structures were much more "popular" than average, with as many as 122 sequences defining them uniquely, while 330 structures could be designed by only one sequence. Statistics for eight structures are given in Table 4. All but one of these was significantly more popular than average, having 22 to 122 design sequences, and there was one structure of about average popularity (six sequences). These structures were examined in more detail to seek out guidelines that might be useful in intentionally choosing a sequence that will lead to a given target structure as its nondegenerate ground state. That is, an attempt was made to use reverse engineering to extract design principles from these "suc-

Table 2. *Average properties per sequence group by types of added charges*[a]

Added charges	Number of sequences	Nondegenerate ground states (%)	Average energy	Average degeneracy	Average number of contacts
		Low-weighted energy function			
	1	0	−11.0	267.0	13.0
+/−	28	0	−10.7	160.1	12.4
+−	182	4	−10.1	75.6	12.6
++/−−	182	5	−9.9	53.5	12.4
++−/−−+	2,184	18	−9.4	55.7	12.2
+++/−−−	728	16	−9.2	48.6	11.7
++−−	6,006	26	−8.9	15.9	11.9
+++−/−−−+	8,008	22	−8.8	19.9	11.7
+++−/−−−−	2,002	10	−8.5	39.9	11.1
		Middle-weighted energy function			
	1	0	−17.5	267.0	13.0
+/−	28	0	−17.0	88.5	12.9
+−	182	4	−16.5	26.7	12.9
++/−−	182	7	−16.2	29.0	12.8
++−/−−+	2,184	20	−15.8	7.9	12.9
+++/−−−	728	15	−15.3	11.3	12.6
++−−	6,006	39	−15.2	3.5	12.8
+++−/−−−+	8,008	36	−15.0	4.2	12.7
+++−/−−−−	2,002	23	−14.4	8.0	12.2
		High-weighted energy function			
	1	0	−24.0	267.0	13.0
+/−	28	0	−23.5	81.7	13.0
+−	182	4	−23.0	20.6	13.0
++/−−	182	7	−22.7	26.7	12.9
++−/−−+	2,184	20	−22.3	7.2	13.0
+++/−−−	728	9	−21.7	10.3	12.9
++−−	6,006	40	−21.7	3.2	13.0
+++−/−−−+	8,008	38	−21.4	3.7	12.9
+++−/−−−−	2,002	22	−20.7	5.6	12.8

[a]See Table 1 for definitions of column headings.

cesses" that were chosen largely because they were more "popular," and apparently easier to design, than the average structure.

In 104 of the 122 sequences that uniquely selected structure I (see Fig. 4A), an identical salt bridge contact pattern was found. It consisted of a cyclic, four-way salt bridge network among both ends and residues 6 and 11. These sequences differed only in the number and placement of lone charges on the surface of the structure. Of the remaining 18 sequences, six had another salt bridge in addition to the four-way network. The other 12 sequences formed an end salt bridge network with residue 6 or 11 (but not both), and additional salt bridges or lone charges were required to successfully design the structure.

The second most popular structure was selected uniquely by 92 sequences. It is shown in Figure 4B and listed as structure II in Table 4. Because our enumeration scheme distinguished chain direction, there were actually two versions of the second most popular structure (differing in chain direction). Due to symmetry, the complete set of design sequences for one can be converted to that for the other by reversing the charge polarity and order (to account for the opposite sign of N- and C-terminal residues). Only one version of structure II is discussed here. Most sequences that uniquely designed this structure (78 of 92 sequences) had salt bridge contacts consisting solely of a four-way cyclic network involving both ends, three of the remaining sequences included both this network and other salt bridge contacts, and the remaining 11 sequences retained one or the other (but not both) of the added charges that together would make the four-way network (but also included other salt bridges and/or lone charges).

A key design feature of structures I and II was the four-way cyclic salt bridge network to the ends. Significantly, this network alone was sufficient to uniquely select each of these structures without the addition of other charges and was the most stable of any of their design sequences, yielding a folding energy of −18.0 kcal/mol for both structures. It seems that the placement of specific electrostatic restraints at one end of the molecule combined with the hydrophobic effect and connectivity can be enough to define a unique structure. Thus, there is a straightforward design principle for structures I and II: form a cyclic network to the end–end salt bridge and follow it by additional lone charges and

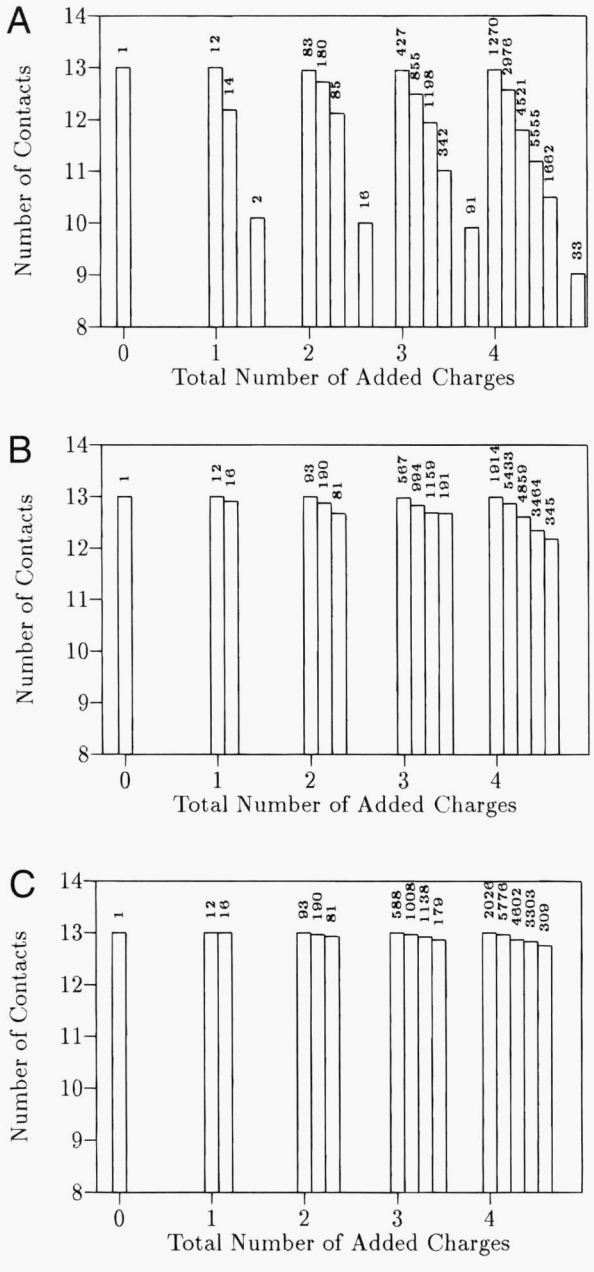

Fig. 2. The average compactness for all sequences of length 16 with up to 4 added charges, for each of the (**A**) low, (**B**) mid, and (**C**) high energy functions. As in Figure 1, the bars represent sequence groups sorted by number of added charges, and numbers above each bar is the number of sequences contributing, rounded to the nearest integer.

salt bridges. We explore the general utility of this strategy in the section "Modeling the design process."

Structure III (Fig. 4C) is interesting because it does not permit a cyclic four-way salt bridge network to the ends; the two residues contacting the N- and C-termini (residues 6 and 7, respectively) cannot directly interact with each other because they are covalently linked. Every sequence that uniquely selected this structure (with up to four additional charges) was charged at one or the other, but not both, of these residues; however, these charges alone did not allow a successful design. At least three additional charges were required at specific locations on the surface of the structure. Examples (shown in Fig. 4C) include sequences with three lone charges (five sequences), one lone charge and one salt bridge (16 sequences), and a two salt bridge network (one sequence). The latter sequence is the only design (with up to four additional charges) of structure III without lone charges but is nearly the least stable

Table 3. *Accessible backbone traces*

Number of contacts	Number of traces
0	64,430,202
1	104,301,551
2	99,758,684
3	75,349,797
4	47,612,959
5	26,552,533
6	13,325,102
7	6,394,581
8	2,593,219
9	904,680
10	247,762
11	72,056
12	4,654
13	2,134
Total	441,549,914

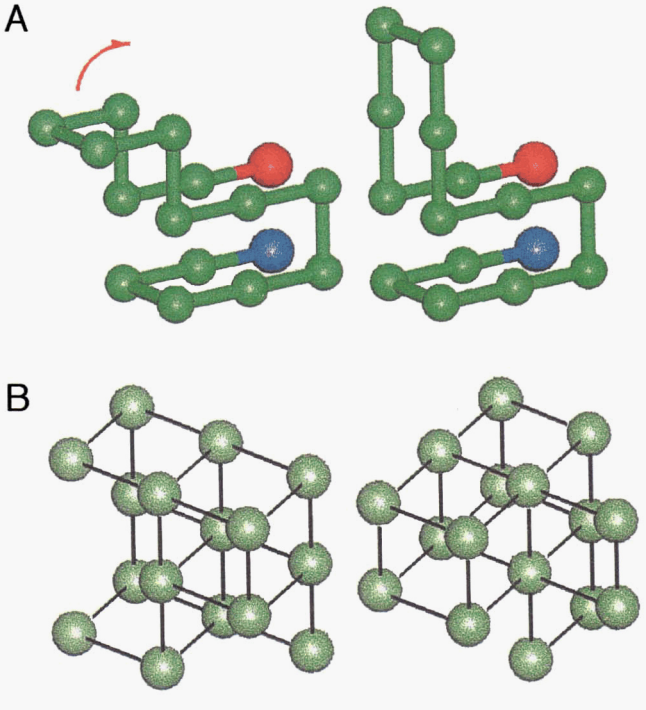

Fig. 3. A: A structure having a flexible "loop" with two conformations, indistinguishable by the type of contact potential used in the current work. **B:** The two arrangements of monomer units found for sequences (up to four added charges) with nondegenerate 11-contact ground-state structures. All 22 such sequences gave structures, which conformed to these shapes, within rotational and mirror-image symmetry transformations (see Methods). Note the absence of flexible loops of the type seen in **A**.

Table 4. *All designs with up to four added charges for structures I–VIII*

Target structure	Contacts	Energy function	Number of sequences that select target structure in the ground state (uniquely/uniquely or nonuniquely)					
			+0...0−	1 Charge	2 Charges	3 Charges	4 Charges	Total
I	13	Low	0/1	0/0	1/4	16/22	60/91	77/118
		Mid	0/1	0/8	1/23	18/58	103/187	122/277
		High	0/1	0/8	1/25	18/72	115/255	134/361
II	13	Low	0/1	0/0	1/4	10/20	28/72	39/97
		Mid	0/1	0/8	1/24	16/49	75/143	92/225
		High	0/1	0/8	1/26	16/63	83/196	100/294
III	13	Low	0/1	0/2	0/19	0/47	0/64	0/133
		Mid	0/1	0/8	0/29	0/68	22/133	22/239
		High	0/1	0/8	0/29	0/73	23/150	23/261
IV	13	Low	0/1	0/4	0/28	0/56	0/89	0/178
		Mid	0/1	0/10	0/39	0/99	6/192	6/341
		High	0/1	0/10	0/41	0/112	7/244	7/408
V	12	Low	0/0	0/0	0/1	3/5	23/38	26/44
		Mid	0/0	0/0	0/0	3/3	17/27	20/30
		High	0/0	0/0	0/0	0/0	4/9	4/9
VI	13	Low	0/1	0/2	0/11	0/24	5/48	5/86
		Mid	0/1	0/6	0/15	4/32	29/65	33/119
		High	0/1	0/6	0/15	4/36	29/77	33/135
VII	13	Low	0/1	0/3	0/20	0/41	7/60	7/125
		Mid	0/1	0/7	0/25	0/61	10/129	10/223
		High	0/1	0/7	0/25	1/66	12/145	13/244
VIII	13	Low	0/1	0/3	0/20	2/41	9/62	11/127
		Mid	0/1	0/7	0/26	5/65	27/143	32/242
		High	0/1	0/7	0/26	5/73	29/163	34/270
Total sequences			1	28	364	2,912	16,016	19,321

(folding energy of −14.8 kcal/mol). The one-lone-charge, one-salt bridge group was most stable, with each member having a −15.6 kcal/mol folding energy, while the three-lone-charge group was least stable (−14.6 kcal/mol for each sequence). While there does not appear to be a foolproof design principle for structure III, our results suggest a basic strategy similar to that found for structures I and II: place at least one salt bridge networked to the ends and try different combinations of surface salt bridges and lone charges until a successful design is reached (i.e., semi-rational design).

A characteristic of the structure III designs is that residues 6 and 7 were never simultaneously charged (although, as noted above, a charge on one or the other was found in all successful designs). The generality of this feature was tested by modifying the three-lone-charge designs found, for structure III, to place a charge on the hydrophobic member of the the pair of residues 6 or 7 (thus, forming a salt bridge network between residues 6–1–16–7). In fact, these new designs (which were not previously considered because they contained more than four additional charges) were also successful. Thus, design features that appeared in the sequence space searched here were not always necessary in the global sequence space, due to the restricted number of charges allowed.

Structure IV (Fig. 4D) could be designed by only six sequences with up to four added charges, but there was a clear similarity between target interactions from these sequences and those specifying structure III uniquely. One or both of the possible end-networked salt bridge contacts were always formed, plus specific combinations of lone and salt bridged charges were placed at residues 2, 4, 7, and 10.

One example of a less compact, yet still relatively popular, structure is V (Fig. 4E), for which 20 design sequences were possible with up to four charges. A cyclic four-way salt bridge network including the ends was present in each sequence uniquely specifying the structure. In this case the cyclic network was not sufficient and had to be supplemented by lone charges placed on the surface. Two residues in structure V (12 and 15) make no inter-residue contacts and are thus fully solvent-exposed in our energy model. Interestingly, a charge on either or both of these residues (12 or 15), in conjunction with the cyclic end network, produced successful designs that were the most stable of any designs for structure V (five sequences, −15.9 kcal/mol each; note that not all eight sequences fitting this description were successful because for three either the target structure was not in the ground state or was in a nonunique ground state). Moreover, every successful design sequence found carried a charge on residue 12, residue 15, or both. This is plausible in that charges placed at these positions incur no desolvation penalty in the target structure, but may help to select against other structures in which one or both of these positions are partially or fully buried.

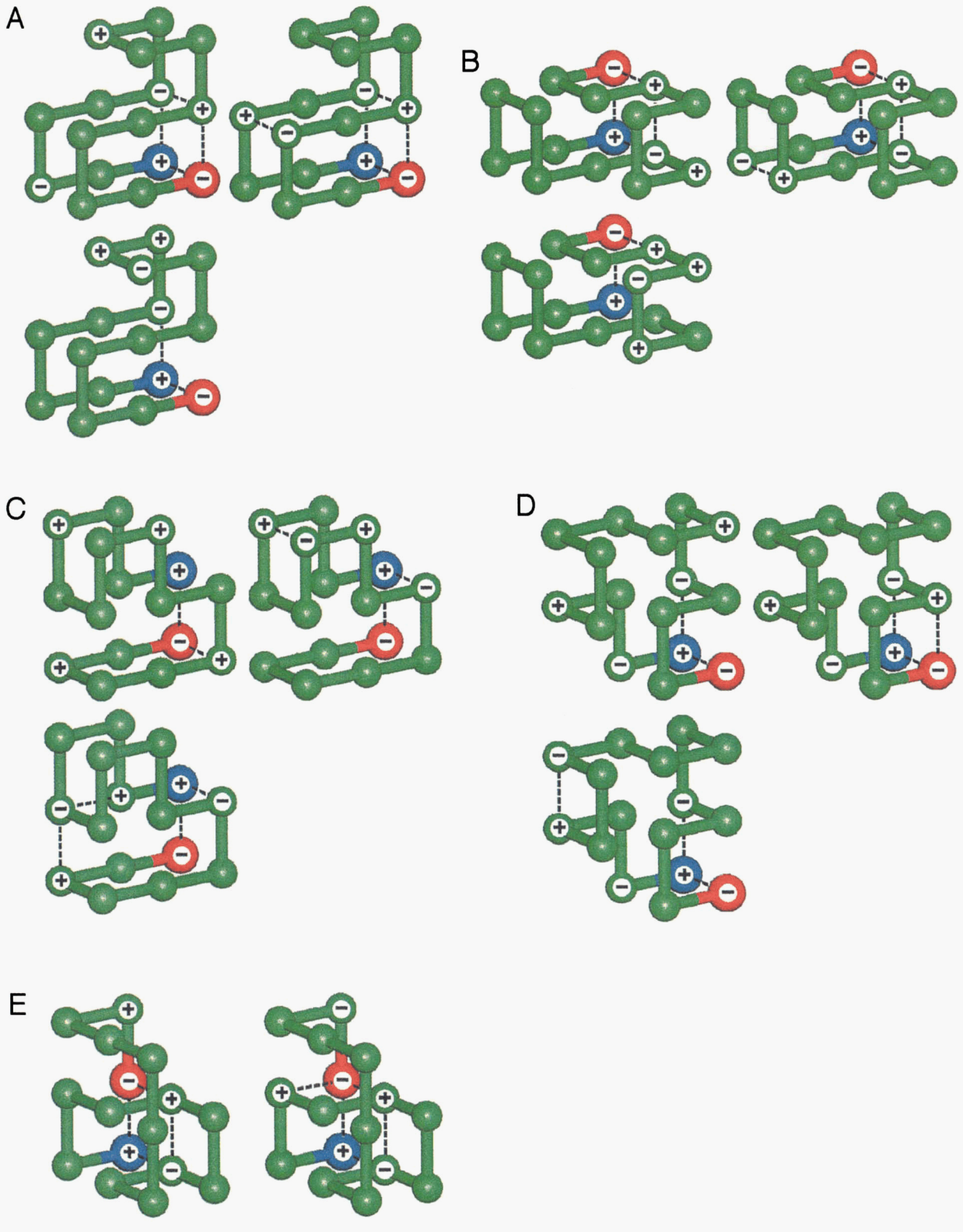

Fig. 4. Examples of sequences that produced a unique ground state, for structures I–V. The dashed lines indicate salt bridge contacts recognized by the energy function. **A:** Three sequences that uniquely select structure I—a design with two lone charges and a cyclic four-way salt bridge involving both ends; a design with the cyclic end network and a second, independent salt bridge; and a design with three lone charges but lacking the cyclic end network. **B–E:** Sample designs for structures II–V, respectively.

For structures I–V, the sequences that solved the design problem had certain similarities. Many sequences chosen from one of these groups were related by just one or two single-site mutations to another sequence that uniquely selected the same structure. This suggested a certain continuity of this sequence space, which was investigated further.

Mutational effects on structure and stability

Given a wild-type sequence that uniquely selected some structure, how many nearby mutations are there that preserve the structure as a nondegenerate or degenerate ground state? In our model a 16-mer has 14 potential sites of mutation (maintaining the charge on the N- and C-termini), leading to $2^1 \cdot \binom{14}{1} = 28$ single or $2^2 \cdot \binom{14}{2} = 364$ double mutants. We randomly chose 22 wild-type sequence/structure pairs from the full set of 6,307 sequences with unique ground-state structures and examined all single- and double-mutant sequences (a total of 392 in each set of "homologous" sequences). As a control, for every wild-type sequence this analysis was repeated five times using 392 random sequences derived from the homologues by shuffling the residues of the central 14 positions. Results of analysis on homologous and control sequence sets are presented in Table 5.

Randomly chosen control sequences were exceedingly unlikely to uniquely specify the wild-type structures of any of the 22 wild-type sequences: The average rate was less than 0.1%. In contrast, an average of 4.8% of homologous (single- or double-mutant) sequences uniquely selected wild-type structures, a rate 50 times greater on average. Mutants nearby to the wild-type sequence thus had a small but significant preference for uniquely specifying the original structure. This rate did vary among both homologous and control sets for different wild-type sequences, but with just one exception the rate was always higher for homologous than control sequences.

Many more homologous sequences selected the wild-type structure as one member of a degenerate ground state. Homologous sequences with this property again outnumbered control sequences, on average, by a factor of 19 and comprised 8.9% of all homologous sequences. Thus, a significant number of homologous sequences defined the wild-type fold, but nonuniquely.

Finally, we considered sequences with a unique ground-state structure that was closely related to, but distinct from, the wild-

Table 5. *Homologous sequence statistics for sequences A–V*

	Wild-type sequence																				
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	Fraction of sequences (%)																					
Uniquely select wild-type structure																						
Single mutants (38)	17.9	21.4	0.0	14.3	7.1	17.9	17.9	10.7	14.3	17.9	10.7	7.1	17.9	17.9	21.4	14.3	39.3	14.3	14.3	14.3	14.3	3.6
Single or double mutants (392)	5.1	9.2	0.3	3.1	2.6	6.1	5.6	4.1	4.1	6.6	3.1	2.0	6.4	4.1	6.4	3.6	15.3	5.9	4.3	2.8	3.3	1.8
Sets of control sequences (392)	0.0	0.8	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.0	0.3	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0
	0.3	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.0	0.0
	0.3	0.0	0.0	0.3	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.3	0.0	0.3	0.0	0.0	0.0
	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.3	0.0	0.0	0.0
	0.3	0.5	0.3	0.3	0.0	0.3	0.3	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.3	0.0	0.0	0.0	0.0	0.0	0.0
Uniquely or nonuniquely select wild-type structure																						
Single mutants (28)	32.1	32.1	10.7	14.3	14.3	42.9	25.0	17.9	17.9	17.9	25.0	21.4	21.4	32.1	25.0	32.1	46.4	25.0	35.7	25.0	14.3	10.7
Single or double mutants (392)	11.5	14.3	3.3	3.8	5.6	12.0	8.7	8.4	6.1	7.1	7.9	6.9	7.4	12.5	8.2	12.5	19.1	8.7	14.3	8.4	4.3	4.8
Sets of control sequences (392)	0.0	0.8	0.3	0.3	0.0	0.3	1.8	1.0	0.0	0.0	0.0	0.3	0.3	0.8	1.3	0.8	0.0	0.5	0.5	0.8	0.3	0.0
	0.5	1.0	0.0	0.5	0.0	0.5	1.8	0.5	0.5	0.0	0.3	0.8	0.3	1.3	0.5	1.0	0.0	0.0	0.5	0.3	0.3	0.0
	1.0	0.5	0.3	0.3	0.0	0.5	0.5	1.3	0.3	0.3	0.5	1.0	0.3	0.5	0.3	1.3	0.3	0.0	1.8	0.5	0.0	0.0
	0.8	0.8	0.0	0.8	0.0	0.5	1.0	0.0	0.0	0.0	0.5	0.3	0.3	0.5	0.5	1.0	0.0	0.0	0.5	0.3	0.3	0.0
	1.0	1.3	0.3	1.8	0.0	0.5	1.3	0.3	0.0	0.0	0.8	0.0	0.0	0.8	0.8	1.8	0.0	0.0	1.5	0.3	0.0	0.0
Uniquely select "nearby" structure																						
Single mutants (28)	53.6	32.1	3.6	28.6	7.1	35.7	28.6	21.4	21.4	42.9	46.4	35.7	57.1	42.9	46.4	25.0	57.1	25.0	21.4	53.6	32.1	14.3
Single or double mutants mutants (392)	32.9	23.2	3.1	12.2	4.8	21.2	23.2	17.3	9.9	21.9	29.3	26.3	36.7	27.0	28.8	19.4	35.5	23.0	17.3	26.5	24.2	13.8
Sets of control sequences (392)	3.6	3.6	3.8	4.8	0.3	5.6	7.1	2.8	4.1	0.5	5.1	3.8	4.1	5.6	5.1	5.9	2.3	2.6	4.1	5.4	1.5	3.1
	5.1	3.8	2.6	3.8	0.5	4.1	7.1	1.5	2.6	0.5	7.1	5.6	3.1	2.6	4.1	3.8	2.6	2.8	2.8	5.1	2.8	3.6
	5.4	4.8	3.1	4.3	0.8	2.6	4.8	4.1	2.3	1.0	5.6	2.6	3.1	4.1	3.6	5.6	2.8	4.1	4.1	2.6	2.0	3.3
	4.3	3.1	3.8	6.9	0.8	5.1	7.9	3.8	2.6	2.3	4.1	6.1	5.4	2.8	5.9	3.8	3.1	4.3	3.6	3.3	1.5	4.6
	6.1	1.5	4.6	3.1	0.8	4.3	7.4	2.8	2.3	1.3	5.4	4.1	4.6	4.1	3.3	4.1	3.3	2.8	2.6	3.1	3.8	1.8

type structure. Contact homology was chosen as a measure of structural similarity. Contact homology was defined as the fraction of inter-residue contacts shared by two conformations. (The number of contacts in the wild-type structure was chosen to be the denominator of this fraction.) An average of 21.7% of homologous sequences were found to have unique native folds related to the wild-type structure (more than 50% contact homology), a rate much higher than that found in the control sequences (3.7%, on average).

One might imagine constructing a path through sequence space along which the ground-state structure remains unchanged or partially conserved (defined here as 50% contact homology) through the accumulation of single-substitution mutations. Here we have mapped out a small portion of that space (all single and double mutants about a small number of starting sequences). A mutation series was found for sequence A that nicely illustrates some of the above characteristics of homologous sequence sets. The wild-type structure for sequence A is shown in Figure 5A. The elimination of a lone charge on residue 6 (+ → 0) was fully conservative in that the mutant uniquely selected the same ground-state structure as the wild-type (Fig. 5B) and stabilized the fold by an additional 1.0 kcal/mol. When this was followed by the mutation of residue 12 (0 → +), which introduced a buried lone charge in the original structure, a conformational change was induced (Fig. 5C). The new positive charge on residue 12 moved to a fully solvent-exposed position at the same time as residue 8, which was hydrophobic, shifted to become more buried. Although this rearrangement avoided the costly burial of the newly charged residue, the overall loss of compactness caused a significant stability decrease, +1.5 kcal/mol (middle-weighted energy function) over the previous mutant in Figure 5B. The resulting structure preserved all salt bridge contacts and had a high contact homology (69%) with the wild-type ground state. If applied in order, the above two mutations provide a single-mutation-at-a-time pathway from the original structure of Figure 5A to the structure of Figure 5C that maintains a unique native fold at every step. As such, they illustrate one possible partially conservative, protein evolution pathway.

Modeling the design process

Developing design strategies, even heuristic ones, provides a useful test of our understanding of protein structure and function; protein design and rational engineering is also an important goal. In the current study we have an ideal system for formulating and testing design strategies within the confines of our model, because the entire conformational space of any given 16-mer sequence can be searched to locate the global minimum. Given that we have a "complete" understanding of the energetics (i.e., we know the energy function), it would seem that design should be simple. However, while the energy function helps us to choose a sequence that will have a low energy of folding to the target structure, it does not directly indicate whether the target structure is the global minimum and whether that minimum is nondegenerate. One especially difficult aspect inherent in this, as well as many real-world design problems, is the need to design out alternatives as well as to design in the target structure. For our model, two simple design principles were explored to determine whether they might provide sufficiently successful trial sequences to eventually reduce the need for exhaustive enumeration of conformational and sequence space.

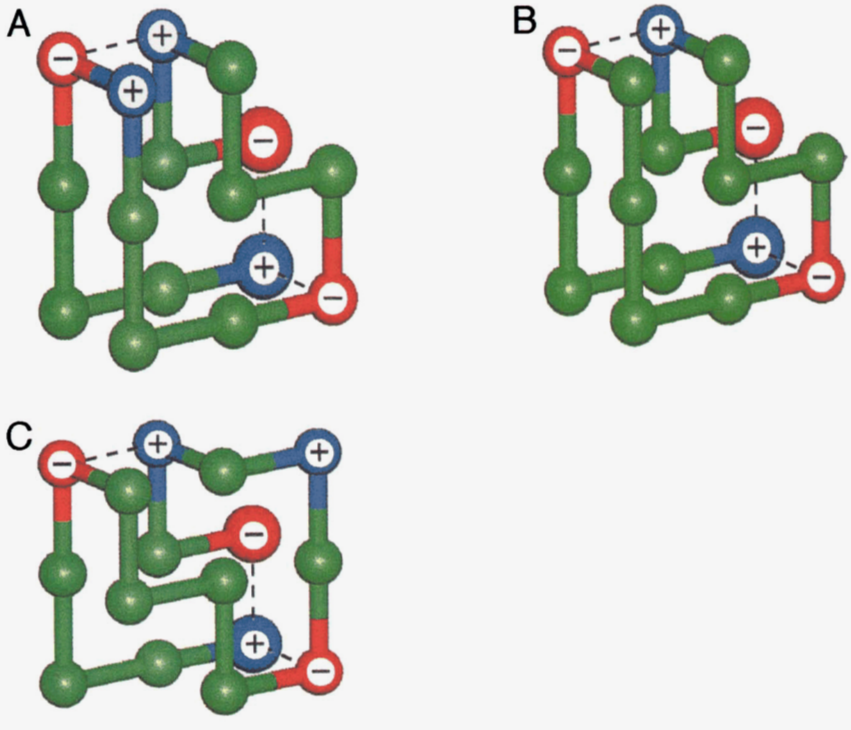

Fig. 5. A mutation series beginning with sequence **A**. (**A**) The unique ground-state structure of sequence **A**; this unique ground-state structure is conserved following elimination of the lone charge on residue 6 (**B**); upon a second mutation (0 → +, residue 12) residues 8 and 12 shift to expose the newly added charge, in the resulting unique ground-state structure (**C**).

Our results suggest that salt bridges may have design advantages over lone charges (at least in the crude models explored here) for two reasons. First, the average effect of adding a salt bridge to a sequence reduced structural degeneracy more than adding a pair of lone charges (Fig. 1A). Second, an added salt bridge was less destabilizing than a pair of added lone charges (Fig. 1B). Irrespective of these generalities, it should be noted that some structures were assisted by lone charges for a nondegenerate design, such as structures IV and V. Moreover, other structures, such as structure III, were more stable when specified uniquely by sequences resulting in a lone charge than with only salt bridges. It is an important goal to be able to design sequences that fold to desired, nondegenerate structures without enumerating all other structures, particularly using more detailed and realistic models for which such searches are beyond the reach of current methodology. With this in mind, we adopted and tested the following two design strategies. In the first, we started from the parent sequence in the desired target structure and added two salt bridges (the two-bridge method); these initial designs were refined by single mutations. In the second, we continued adding salt bridges until as many as possible had been placed in the structure (the all-charge method). These methods were evaluated for their ability to construct sequences that select the desired structure as a degenerate and as a nondegenerate ground state.

The two-bridge method

Any specific design strategy might include protocols for choosing where and how many salt bridges and lone charges should be introduced into a largely hydrophobic parent background sequence. Here we test a "two-bridge" method in which two salt bridge contacts were introduced. The first charge was added to form a networked salt bridge with the end-to-end salt bridge contained in all target structures. Then a salt bridge was added (two charged residues, or just one if the bridge was being formed with an existing charge). This strategy was applied combinatorially to produce sequences yielding all pairs of salt bridges in the target structure, with the exception that highly buried salt bridges (in which each residue formed three or more contacts) were not considered. The two-bridge strategy was applied to structures I–VIII. Enumeration of the two-bridge designs for these structures yielded 17–43 sequences per structure, each of which was analyzed for ground-state energy, structure, and degeneracy.

While this strategy was quite successful in identifying sequences that specified the target structure nonuniquely, it did not result in many sequences that specified the target structure uniquely (Table 6). For the middle-weighted energy function, only three of the structures were uniquely specified by any of their two-bridge design sequences. For these structures, fewer than 25% of trial

Table 6. *Two-bridge design attempts for structures I–VIII*

			Uniquely select target structure		Uniquely or nonuniquely select target structure	
Structure	Energy function	Number of attempts	Number of sequences	Range of energies (kcal/mol)	Number of sequences	Range of energies (kcal/mol)
I	Low	17	1	−11.5	5	−11.5 to −9.6
	Mid	17	3	−18.0 to −16.1	5	−18.0 to −16.1
	High	17	3	−24.5 to −22.6	5	−24.5 to −22.6
II	Low	17	1	−11.5	3	−11.5 to −9.6
	Mid	17	3	−18.0 to −16.1	3	−18.0 to −16.1
	High	17	3	−24.5 to −22.6	3	−24.5 to −22.6
III	Low	21	0	—	8	−11.3 to −9.1
	Mid	21	0	—	9	−17.8 to −15.6
	High	21	0	—	10	−24.3 to −22.1
IV	Low	19	0	—	5	−10.8 to −9.6
	Mid	19	0	—	6	−17.3 to −16.1
	High	19	0	—	6	−23.8 to −22.6
V	Low	43	0	—	0	—
	Mid	43	0	—	0	—
	High	43	0	—	0	—
VI	Low	41	0	—	11	−11.3 to −9.1
	Mid	41	0	—	11	−17.8 to −15.6
	High	41	0	—	13	−24.3 to −22.1
VII	Low	38	0	—	10	−10.8 to −9.1
	Mid	38	0	—	11	−17.3 to −15.6
	High	38	1	−22.1	12	−23.8 to −22.1
VIII	Low	38	1	−9.9	10	−10.8 to −9.1
	Mid	38	1	−16.4	11	−17.3 to −15.6
	High	38	1	−22.9	12	−23.8 to −22.1

sequences were successful. Those most successfully designed were structures I and II, for each of which 3 out of 17 designs gave the correct nondegenerate ground-state fold. Structures I and II were the two most popular design structures found by the earlier sequence enumeration for up to four added charges (Table 4). The only other structure uniquely specified by a two-bridge design was structure VIII, which was the unique native fold of just 1 of 38 design attempts.

The successful designs that gave a unique native state for structures I, II, and VIII, while few in number, were exceptionally stable. The most stable two-bridge designs for structures I and II, each −18.0 kcal/mol (middle-weighted energy function), were in fact the most stable sequences found for these structures by the earlier sequence enumeration (this was true for all three variants of our energy function: low, middle, and high). The single nondegenerate design for structure VIII had a stability of −16.4 kcal/mol, which was surpassed by only one other uniquely specifying sequence (−16.5 kcal/mol) for structure VIII found with up to four added charges (Table 4). By adding a small number of charges compensated by salt bridge interactions, the two-bridge strategy tended to produce very stable designs in the few cases where it produced the desired target uniquely.

If degenerate ground states were included, a significant fraction of the two-bridge design sequences for structures I–VIII selected the target structure. For six of the structures, more than 25% of the designs specified the target structure, and such designs were found for all but one of the eight structures (Table 6). We attempted to define properties of the two-bridge designs themselves that predicted whether they specified the target structure with low degeneracy. Such properties could allow trimming of the two-bridge designs for a particular structure to a smaller number of design candidates more likely to be successful. While such sequences may not be completely successful as designs due to the degeneracy of the ground state, they might be useful intermediates in the design process. To test whether these sequences could be easily refined to uniquely select the target structure, the design attempts for structure III were examined in more detail.

Nine out of 21 sequences (43%) specified structure III nonuniquely with the middle-weighted energy function (none specified the structure uniquely). Two properties appeared to correlate with a design's success in specifying structure III. First, the likelihood of specifying structure III improved for designs in which the second salt bridge of the strategy was maximally exposed (the only contact of either residue was the salt bridge contact). Of the eight designs with such salt bridges, six selected structure III (75% success rate), accounting for two-thirds of the target-specifying designs. Designs that exploited the odd–even rule (see Methods) were also more likely than average to select structure III. One such design is shown in Figure 6A, for which the odd–even rule prevented either residues 3 or 8 from forming favorable salt bridge contacts except with each other. This effect enhanced the design by eliminating folds with unintended salt bridges involving these residues (we term salt bridges with this feature "snookered"). While real proteins do not have exactly this type of constraint, the regularity of secondary structure (in which every other residue is on the opposite face of a regular β-strand and heptad repeats define packing interactions for certain coiled coils, for example) may impose analogous constraints. Of the 10 designs whose second salt bridge was snookered, six specified structure III, for a success rate of 60%.

The snookered salt bridge contributed to a very low degeneracy for the sequence in Figure 6A; there were only two ground-state structures (middle-weighted energy function), including the target structure, and they shared an identical salt bridge contact pattern. When the polarity of the 3-to-8 salt bridge in Figure 6A was reversed (permitting residues 3 and 8 to make favorable salt bridge contacts with other residues), the design continued to nonuniquely select the target structure, but had higher degeneracy. In addition to those in Figure 6A, ground-state conformations for this latter design included the three structures of Figure 6B. These structures were distinctive in that they had lost their end-to-end contact. However, they maintained the general salt bridge contact motif of the other native conformations with one three-way network of salt bridges and one other maximally exposed ion pair.

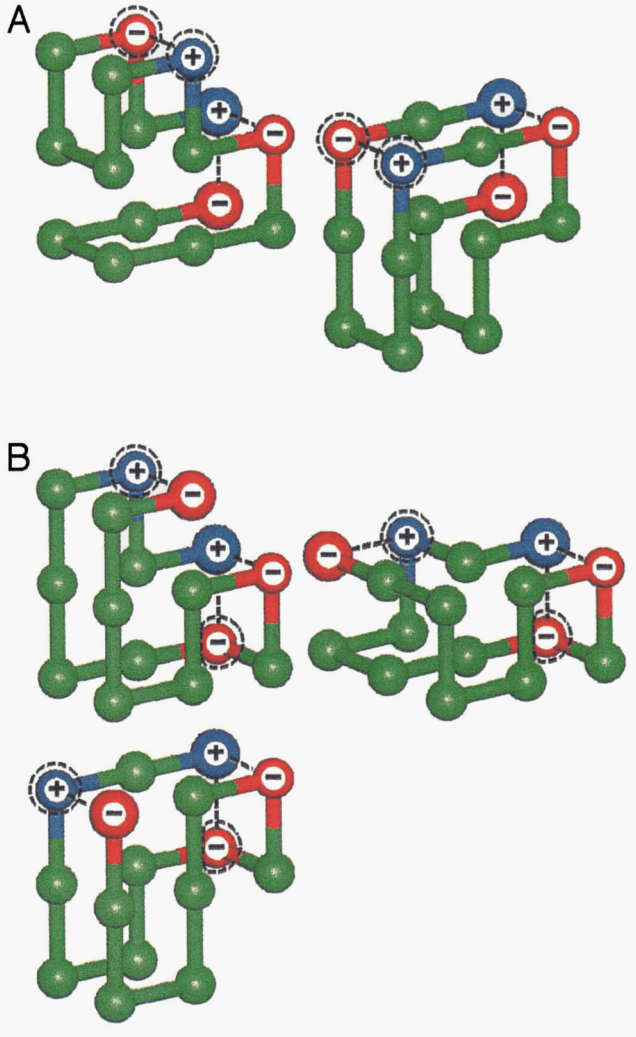

Fig. 6. A: Ground-state structures for a design of structure III with a "snookered" (see text), and maximally exposed, salt bridge between residues 3 and 8 (both surrounded by dashed circles). The negative charge on residue 3 was of the proper sign but incorrect spacing to pair with residue 1 on the lattice. Likewise, the positive charge at position 8 was unable to pair with negatively charged residues at positions 10 and 16 due to lattice constraints (the odd–even rule). The ground state included both the target (structure III) and one other conformation. The nontarget conformation maintained the salt bridge contacts of the target; hydrophobic segments were shifted. **B:** If the polarity of the 3–8 salt bridge contact was reversed relative to the sequence in **A**, the ground state degeneracy of the design attempt increased from two to five. While both conformations shown in **A** were still included, the three additional conformations shown here were also found in the ground state for the sequence of **B**.

The two-bridge designs whose second salt bridge was both snookered and maximally exposed (of which the design in Figure 6A is an example) were remarkably successful. All four such sequences specified structure III and had the lowest degeneracy of all target-specifying design attempts, with two sequences of ground-state degeneracy two and two of degeneracy three. Thus, for structure III the two-bridge strategy could have been enhanced by considering only the designs with both of these two particular salt bridge properties. This would have resulted in a small group of sequences that all contained the target structure in a ground state of very low degeneracy.

The target-specifying designs for structure III were also tested for single mutations that refined the design to the correct, unique native fold. There were six two-bridge designs capable of such a mutation, and four designs in particular had multiple possibilities, that is the same four that were singled out above for having especially low degeneracy. This demonstrated that a small population of the two-bridge designs with certain definable characteristics (namely, their use of the odd–even rule and high solvent exposure for the second salt bridge) was easiest to refine to the correct unique native fold. Moreover, the low degeneracy of these designs proved to be a strong indication of their potential for successful refinement. For structure III a successful design strategy could thus be formulated. First, determine the two-bridge designs that select the target structure, narrowing the search to designs only with maximally exposed and snookered second salt bridges. Second, refine those sequences by making single mutations until the correct unique fold is found. With this improved procedure, it was possible to find at least one uniquely specifying design for six of the eight structures in Table 6 (results not shown).

The all-charge method

The second strategy examined was to place charged residues to produce as many salt bridges simultaneously in the target structure as possible without introducing repulsive charge–charge contacts [the effects of repulsive contacts have been discussed (Chan & Dill, 1996)]. This effort was greatly assisted by the odd–even rule, which guaranteed (for this lattice) that at least one fully-charged sequence could be designed for every structure in which only favorable salt bridge contacts occurred. Thus we termed this strategy the "all-charge" method, because the resulting designs contained charges at every sequence location. The basis for choosing this strategy was that such designs would have the maximum number of specific interactions; however, the large number of partially buried charges might lead to destabilization of the target structure (indeed, of many compact structures).

This strategy was tested on the same set of eight target structures examined above. For each structure, every all-charge sequence was enumerated, which resulted in a total of 4–16 designs per target structure (fewer than for the two-bridge strategy). Analysis of these sequences is presented in Table 7. All structures in Table 7 included the all-charge design sequence + − + − + − + − + − + − + − + −. Due to the odd–even rule, this sequence resulted in only complementary charge–charge contacts for any cubic-lattice conformation. Thus, it is an all-charge design sequence for every structure. This sequence produced a highly degenerate ground state with 154 (middle- and high-weighted energy functions) or 555 (low-weighted energy function) conformations. Only three of the seven maximally compact structures in Table 7 were included in the ground state for this alternating sequence (all energy functions).

For the middle- and high-weighted energy functions, the all-charge strategy was more successful than the initial two-bridge approach at defining structures uniquely. Six target structures could be defined uniquely by at least one all-charge sequence, and both of the remaining structures were members of a degenerate ground state. In general the all-charge sequences produced structures that were less energetically stable than those produced by other sequences. The two unique designs for structure III, for example, had a folding energy of −12.8 kcal/mol (middle-weighted energy function). A design of nearly 3 kcal/mol greater stability (−15.6 kcal/mol) was found by the two-bridge strategy for the same structure. The all-charge strategy produced a large fraction of sequences that uniquely defined structure VI, in which the overall topology abstractly resembled a four-helix bundle—9 of 16 attempts using the middle- and high-weighted energy functions. A similar success rate was found for structures VII (2/4) and VIII (2/4) using the same energy function, although there were many fewer all-charge sequences in total for these structures.

The all-charge design strategy was less successful using the low-weighted energy function. Only one sequence uniquely determined any of the structures (VIII) and only five of the structures were selected either uniquely or nonuniquely by any of the designs. This was a result of the lesser weight given to favorable energetics of protein collapse, which produced less compact ground states. One example was a design sequence for structure VIII that actually did select the target structure, but had four other ground-state conformations that were not maximally compact and contained fully solvent-exposed loops that could deform without changing their inter-residue contact pattern.

Conclusion

A simple cubic lattice model was used to examine the effect of salt bridge interactions on the ground-state energy and degeneracy of proteins (peptides) using three energy functions obtained largely from continuum electrostatic calculations on structures from the Protein Data Bank (Bernstein et al., 1977; Abola et al., 1987). It was found that the substitution of charged for hydrophobic residues substantially reduced the degeneracy of the ground state. Moreover, charges involved in salt bridges in the ground state were generally, but not always, more effective in reducing degeneracy than those that were not. Together with accumulating evidence that buried polar and charged interactions tend to destabilize folded proteins relative to hydrophobic groups of similar size and shape (Hendsch & Tidor, 1994; Yang & Honig, 1995; Waldburger et al., 1995; Wang et al., 1996; Wimley et al., 1996), this suggests there is, in general, a trade-off between stability and specificity [i.e., the uniqueness of the native fold (Lattman & Rose, 1993)]. The discovery of particular constellations of interactions that impart stability and specificity simultaneously would thus be extremely valuable for protein molecular engineering. Alternatively, sets of polar and charged interactions that impart specificity with a minimal loss of stability would be desirable.

A further finding of this study is that complete knowledge of the energy function did not trivially permit "rational" design, in either the forward or reverse direction. That is, given a sequence, the structure(s) corresponding to its ground state were not readily constructable from the energy function; rather, a search procedure (exhaustive enumeration, here) was used to locate the ground state. Likewise, given a target structure, complete knowledge of the energy function did not readily permit the construction of se-

Table 7. *All-charge design attempts for structures I–VIII*

Structure	Energy function	Number of attempts	Uniquely select target structure: Number of sequences	Uniquely select target structure: Range of energies (kcal/mol)	Uniquely or nonuniquely select target structure: Number of sequences	Uniquely or nonuniquely select target structure: Range of energies (kcal/mol)
I	Low	16	0	—	2	−6.1
	Mid	16	2	−12.6	8	−12.6
	High	16	2	−19.1	8	−19.1
II	Low	8	0	—	0	—
	Mid	8	1	−12.6	3	−12.6
	High	8	1	−19.1	3	−19.1
III	Low	8	0	—	0	—
	Mid	8	1	−12.6	1	−12.6
	High	8	1	−19.1	1	−19.1
IV	Low	4	0	—	0	—
	Mid	4	0	—	1	−12.6
	High	4	0	—	1	−19.1
V	Low	16	0	—	3	−6.1
	Mid	16	0	—	4	−12.1
	High	16	0	—	4	−18.1
VI	Low	16	0	—	16	−6.2
	Mid	16	9	−12.8	16	−12.8
	High	16	9	−19.3	16	−19.3
VII	Low	4	0	—	4	−6.2
	Mid	4	2	−12.8	4	−12.8
	High	4	2	−19.3	4	−19.3
VIII	Low	4	1	−6.2	4	−6.2
	Mid	4	2	−12.8	4	−12.8
	High	4	2	−19.3	4	−19.3

quences that would fold uniquely to it; instead, trial sequences could be constructed from heuristics that were only moderately successful and searches were performed in conformational space to determine the successes and in sequence and conformational space to refine them. This can be compared to the case for oligonucleotides. Given a sequence of one polynucleotide, it is trivial to "design" a complementary one. Likewise, given an RNA secondary structure, it is harder, but still relatively simple, to write a sequence that is likely to fold with that structure. The difference is that there is a relatively straightforward "code" for polynucleotide structure, but either there is not one, or we do not yet know it, for protein structure. The nondeterministic design scheme for protein structure is not a result of complexities introduced due to ensemble averaging and entropic considerations entering into the calculation of the free energy of the system; here only a simple effective potential energy function was used. Rather, it is that energetics does not directly translate into a "code." Thus, as our understanding of protein energetics continues to improve, it will be essential to develop tools (and even heuristics) to allow that understanding to be used effectively in molecular engineering and design. For example, will it be possible to "design in" a given structure without explicitly considering, and "designing out," a number of competitors? This was achieved here, to some extent, by using the odd–even rule to place salt bridges that could not shuffle partners.

While no such fully rational design scheme has emerged yet for the simple lattice models presented here, a few useful heuristics have been shown. The encoding of specificity for the target structure by including complementary salt bridging groups, particularly those interacting with "obligate charges" (the chain termini, in the models here), the arrangement of salt bridging groups so that they were unable to pair with alternative partners (through use use of inherent lattice constraints, here, but perhaps possible using other topological constraints in real proteins) and the placement of salt bridges to be partially exposed as opposed to fully buried in the target structure, were all useful strategies that tended to favor the target structure as, at least, a nondegenerate ground state.

Methods

The model

Protein polymers were modeled as self-avoiding, directional flights on a simple cubic (three-dimensional) lattice. Each flight or "trace" corresponded to a conformation or structure for a protein. Each lattice point occupied by a trace corresponded to a protein residue; each protein contained 16 residues. The residues were of three types: hydrophobic, positively charged, and negatively charged. The N-terminal (first) and C-terminal (last) residue was positively and negatively charged, respectively, whereas the 14 other positions in the sequence could be occupied by any of the three residue types.

Structural enumeration

Full enumeration to determine all self-avoiding, directional conformers of this lattice polymer was carried out. A structure was eliminated during the enumeration if it (1) violated excluded volume, (2) was related by overall translation or rotation to a conformer already accepted in the enumeration, or (3) was the mirror image of an accepted conformer. The effect of these conditions was that each configuration (ignoring chain direction) appeared twice (once in each chain direction) except for configurations containing internal symmetry (e.g., a plane of reflection, ignoring chain direction), which appeared just once. The total number of structures in this enumeration was 441,549,914.

Energy functions

The energy of any given conformation for a defined sequence was computed using the following energy function, which contains three terms and is based on residue–residue contacts. The zero of energy for any sequence corresponded to any conformation with no residue–residue contacts (defined here as a pair of residues not adjacent in the sequence that occupy nearest-neighbor sites on the simple cubic lattice). For example, a fully extended conformation for any sequence had zero energy. The first energy term was used to reward burial of surface area (independent of sequence) and was equal to the total number of contacts in the structure times a scale factor. In three separate versions of the energy function (termed low-, middle-, and high-weighted), the scale factor was -1.0, -1.5, and -2.0 kcal/mol/contact, respectively. Because this was the only energy term applied to hydrophobic residues, and because of its role in discharging–folding–charging pathways (Nicholls et al., 1991), we refer to this as the hydrophobic term. In this lattice model a fully buried residue had four contacts (five if it was the N- or C-terminus), so this corresponded to a 2–4 kcal/mol reward for a fully buried residue. Based on the observation that a buried methyl group is worth about 1 kcal/mol (Kellis et al., 1988; Matsumura et al., 1988; Pace, 1992) this range appears reasonable for polypeptides.

The second term in the energy function was a penalty for burying charged residues. Fitting to continuum electrostatic results from a study of buried salt bridges (Hendsch & Tidor, 1994), this term consisted of a contribution from each charged residue in the chain of 1.00, 2.25, 4.00, 6.50, and 6.50 kcal/mol for 1, 2, 3, 4, and 5 contacts, respectively. This term was fit to per-residue $\Delta\Delta G_{solv} + \Delta\Delta G_{protein}$ values relative to per-residue percentage burial (Hendsch & Tidor, 1994) and so includes the penalty for burying a charge in a hydrophobic protein and the mean-field favorable electrostatic interactions with other neighbors that are not salt bridges.

The third term in the energy function was a charge–charge term calculated as a sum over charged residues in the chain and, for each charge, a sum over its charge–charge contacts of ± 0.50, ± 1.25, ± 2.00, ± 2.50, and ± 2.50 kcal/mol for 1, 2, 3, 4, and 5 contacts (for the first charge), respectively. The minus sign was taken for opposite charges, and the plus for like charges. The values were obtained by fitting to previous results (Hendsch & Tidor, 1994). This term was modified for a charge involved in a charge network; the contribution of such a charge was scaled by 0.75, 0.62, 0.50, and 0.40 if the net charge summed over all its contacts was 2, 3, 4, and 5, respectively. These scaling numbers were chosen somewhat arbitrarily but reflect the limited ability of a single side chain to interact optimally in all directions simultaneously as well as the expected repulsion between a pair of salt bridge partners.

Degeneracy and global minimum energy

The global minimum energy for a given sequence was defined as the lowest energy for any conformation of that sequence on the lattice. The degeneracy of a sequence was defined as the number of structures sharing the global minimum energy (ignoring chain direction). Initially sequence enumeration was performed over the set of structures with nine or more contacts to determine the global minimum energy and degeneracy. Sequences found by this method to have ground-state structures with 10 or fewer contacts were re-evaluated by considering all structures with seven or more contacts; sequences from this group having ground-state structures with eight or fewer contacts were again re-evaluated using all structures with five or more contacts. For sequences with up to four added charges, no ground-state structures were found with fewer than eight contacts by this method (21 sequences with the low-weighted energy function had 8-contact ground-state structures; no such sequences were found for the middle- or high-weighted energy functions).

When examining a sequence class (e.g., the set with two additional charges: $++$), the average minimum energy was taken as an unweighted average over the sequences in the class of the ground-state energy for each. The average degeneracy for such a class was an unweighted average of the degeneracy of each sequence in the set. When examining a structural class (e.g., structures with i added charges and j lone charges), the average energy was a weighted average of the ground-state energy of every sequence contributing to the class weighted by the fraction of structures in the ground state for that sequence that belonged to the class. Likewise, the average degeneracy of such a class was calculated as a weighted average of the degeneracy of every sequence contributing to the class similarly weighted by the fraction of structures in the ground state for that sequence that belonged to the class.

The odd–even rule

One feature of simple square and cubic lattice models of polymers is that there are restrictions on which residues can be in contact. First, by definition, adjacent residues (i with $i \pm 1$) are not considered to form a contact. Second, residues i and $i \pm 2$ cannot contact, and neither can i and $i\pm$ any even integer. The only possible contacts are with residues i and $i\pm$ an odd integer greater than one. This restriction is termed the odd–even rule (Chan & Dill, 1990).

Acknowledgments

We thank the Tidor research group and Aaron Dinner, Wayne Hubbell, Martin Karplus, Jamie Williamson, Tau-Mu Yi, and Yaoqi Zhou for helpful discussions and an anonymous reviewer for insightful comments. This work was supported by the National Institutes of Health (GM47678 and GM55758).

References

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

Anderson DE, Becktel WJ, Dahlquist FW. 1990. pH-induced denaturation of proteins: A single salt bridge contributes 3–5 kcal/mol to the free energy of folding of T4 lysozyme. *Biochemistry 29*:2403–2408.

Baker EN, Hubbard RE. 1984. Hydrogen bonding in globular proteins. *Progr Biophys Mol Biol 44*:97–179.

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

Chan HS, Dill KA. 1990. The effects of internal constraints on the configurations of chain molecules. *J Chem Phys 92*:3118–3135.

Chan HS, Dill KA. 1996. Comparing folding codes for proteins and polymers. *Proteins Struct Funct Genet 24*:335–344.

Clarage JB, Romo T, Andrews BK, Pettitt BM, Phillips GN Jr. 1995. A sampling problem in molecular dynamics simulations of macromolecules. *Proc Natl Acad Sci USA 92*:3288–3292.

Dahiyat BI, Mayo SL. 1997. De novo protein design: Fully automated sequence selection. *Science 278*:82–87.

Dao-pin S, Sauer U, Nicholson H, Matthews BW. 1991. Contributions of engineered surface salt bridges to the stability of T4 lysozyme determined by directed mutagenesis. *Biochemistry 30*:7142–7153.

Dill KA, Bromberg S, Yue K, Fiebig KM, Yee DP, Thomas PD, Chan HS. 1995. Principles of protein folding: A perspective from simple exact models. *Protein Sci 4*:561–602.

Fairman R, Shoemaker KR, York EJ, Stewart JM, Baldwin RL. 1989. Further studies of the helix dipole model: Effects of a free α-NH_3^+ or α-COO^- group on helix stability. *Proteins Struct Funct Genet 5*:1–7.

Fersht AR. 1971. Conformational equilibria and the salt bridge in chymotrypsin. *Cold Spring Harbor Symp Quant Biol 36*:71–73.

Fersht AR. 1972. Conformational equilibria in α- and δ-chymotrypsin. The energetics and importance of the salt bridge. *J Mol Biol 64*:497–509.

Fersht AR, Shi J-P, Knill-Jones J, Lowe DM, Wilkinson AJ, Blow DM, Brick P, Carter P, Waye MMY, Winter G. 1985. Hydrogen bonding and biological specificity analysed by protein engineering. *Nature (London) 314*:235–238.

Finkelstein AV, Janin J. 1989. The price of lost freedom: Entropy of bimolecular complex formation. *Protein Eng 3*:1–3.

Gilson MK, Given JA, Bush BL, McCammon JA. 1997. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophys J 72*:1047–1069.

Harbury PB, Zhang T, Kim PS, Alber T. 1993. A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. *Science 262*:1401–1407.

Hendsch ZS, Tidor B. 1994. Do salt bridges stabilize proteins? A continuum electrostatic analysis. *Protein Sci 3*:211–226.

Hinds DA, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.

Hol WGJ. 1985. The role of the α-helix dipole in protein function and structure. *Prog Biophys Molec Biol 45*:149–195.

Hol WGJ, van Duijnen PT, Berendsen HJC. 1978. The α-helix dipole and the properties of proteins. *Nature (London) 273*:443–446.

Honig B, Cohen FE. 1996. Adding backbone to protein folding: Why proteins are polypeptides. *Fold Design 1*:R17–R20.

Horovitz A, Serrano L, Avron B, Bycroft M, Fersht AR. 1990. Strength and co-operativity of contributions of surface salt bridges to protein stability. *J Mol Biol 216*:1031–1044.

Huyghues-Despointes BMP, Scholtz JM, Baldwin RL. 1993. Helical peptides with three pairs of Asp–Arg and Glu–Arg residues in different orientations and spacings. *Protein Sci 2*:80–85.

Irikura KK, Tidor B, Brooks BR, Karplus M. 1985. Transition from B to Z DNA: Contribution of internal fluctuations to the configurational entropy difference. *Science 229*:571–572.

Kametekar 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.

Karplus M, Shakhnovich E. 1992. Protein folding: Theoretical studies of thermodynamics and dynamics. In: Creighton T, ed. *Protein folding*. New York: Freeman. pp 127–195.

Kauzmann W. 1959. Some factors in the interpretation of protein denaturation. *Adv Protein Chem 14*:1–63.

Kellis JT Jr, Nyberg K, Šali D, Fersht AR. 1988. Contribution of hydrophobic interactions to protein stability. *Nature (London) 333*:784–786.

Kirkpatrick S, Gelatt CD Jr, Vecchi MP. 1983. Optimization by simulated annealing. *Science 220*:671–680.

Lattman EE, Rose GD. 1993. Protein folding: What's the question? *Proc Natl Acad Sci USA 90*:439–441.

Li H, Helling R, Tang C, Wingreen N. 1996. Emergence of preferred structures in a simple model of protein folding. *Science 273*:666–669.

Lumb KJ, Kim PS. 1995. A buried polar interaction imparts structural uniqueness in a designed heterodimeric coiled coil. *Biochemistry 34*:8642–8648.

Lyu PC, Gans PJ, Kallenbach NR. 1992. Energetic contribution of solvent-exposed ion pairs to alpha-helix structure. *J Mol Biol 223*:343–350.

Lyu PC, Wemmer DE, Zhou HX, Pinker RJ, Kallenbach NR. 1993. Capping interactions in isolated α helices: Position-dependent substitution effects and structure of a serine-capped peptide helix. *Biochemistry 32*:421–425.

Marqusee S, Baldwin RL. 1987. Helix stabilization by $Glu^- \cdots Lys^+$ salt bridges in short peptides of de novo design. *Proc Natl Acad Sci USA 84*:8898–8902.

Matsumura M, Becktel WJ, Matthews BW. 1988. Hydrophobic stabilization in T4 lysozyme determined directly by multiple substitutions of Ile 3. *Nature (London) 334*:406–410.

Matthews BW, Nicholson H, Becktel WJ. 1987. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc Natl Acad Sci USA 84*:6663–6667.

Merutka G, Stellwagen E. 1991. Effect of amino acid ion pairs on peptide helicity. *Biochemistry 30*:1591–1594.

Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.

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

Nicholls A, Sharp KA, Honig B. 1991. Protein folding and association: Insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins Struct Funct Genet 11*:281–296.

Nicholson H, Anderson DE, Dao-pin S, Matthews BW. 1991. Analysis of the interaction between charged side chains and the α-helix dipole using designed thermostable mutants of phage T4 lysozyme. *Biochemistry 30*:9816–9828.

Nicholson H, Becktel WJ, Matthews BW. 1988. Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. *Nature (London) 336*:651–656.

Pace CN. 1992. Contribution of the hydrophobic effect to globular protein stability. *J Mol Biol 226*:29–35.

Page MI, Jencks WP. 1971. Entropic contributions to rate accelerations in enzymic and intramolecular reactions and the chelate effect. *Proc Natl Acad Sci USA 68*:1678–1683.

Paul CH. 1982. Building models of globular protein molecules from their amino acid sequences. I. Theory. *J Mol Biol 155*:53–62.

Pickett SD, Sternberg MJE. 1993. Empirical scale of side chain conformational entropy in protein folding. *J Mol Biol 231*:825–839.

Presta LG, Rose GD. 1988. Helix signals in proteins. *Science 240*:1632–1641.

Prévost M. 1996. Concurrent interactions contribute to the raised pK_a of His 18 in barnase. *J Mol Biol 260*:99–110.

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

Šali D, Bycroft M, Fersht AR. 1988. Stabilization of protein structure by interaction of α-helix dipole with a charged side chain. *Nature (London) 335*:740–743.

Serrano L, Fersht AR. 1989. Capping and α-helix stability. *Nature (London) 342*:296–299.

Serrano L, Horovitz A, Avron B, Bycroft M, Fersht AR. 1990. Estimating the contribution of engineered surface electrostatic interactions to protein stability by using double-mutant cycles. *Biochemistry 29*:9343–9352.

Sheridan RP, Levy RM, Salemme FR. 1982. α-Helix dipole model and electrostatic stabilization of 4-α-helical proteins. *Proc Natl Acad Sci USA 79*:4545–4549.

Shoemaker KR, Kim PS, York EJ, Stewart JM, Baldwin RL. 1987. Tests of the helix dipole model for stabilization of α-helices. *Nature (London) 326*:563–567.

Skolnick J, Kolinski A. 1989. Computer simulations of globular protein folding and tertiary structure. *Annu Rev Phys Chem 40*:207–235.

Straatsma TP, McCammon JA. 1989. Treatment of rotational isomers in free energy evaluations. Analysis of the evaluation of free energy differences by molecular dynamics simulations of systems with rotational isomeric states. *J Chem Phys 90*:3300–3304.

Tanford C, De PK, Taggart VG. 1960. The role of the α-helix in the structure of proteins. Optical rotatory dispersion of β-lactoglobulin. *J Am Chem Soc 82*:6028–6034.

Tidor B. 1994. Helix-capping interaction in λ Cro protein: A free energy simulation analysis. *Proteins Struct Funct Genet 19*:310–323.

Tidor B, Karplus M. 1993. The contribution of cross-links to protein stability: A normal mode analysis of the configurational entropy of the native state. *Proteins Struct Funct Genet 15*:71–79.

Tidor B, Karplus M. 1994. The contribution of vibrational entropy to molecular association. The dimerization of insulin. *J Mol Biol 238*:405–414.

Tobias DJ, Brooks CL III, Fleischman SH. 1989. Conformational flexibility in free energy simulations. *Chem Phys Lett 156*:256–260.

Unger R, Moult J. 1993. Genetic algorithms for protein folding simulations. *J Mol Biol 231*:75–81.

Waldburger CD, Schildbach JF, Sauer RT. 1995. Are buried salt bridges important for protein stability and conformational specificity? *Nature Struct Biol 2*:122–128.

Wang L, O'Connell T, Tropsha A, Hermans J. 1996. Energetic decomposition of the α-helix–coil equilibrium of a dynamic model system. *Biopolymers 39*:479–489.

Wimley WC, Gawrisch K, Creamer TP, White SH. 1996. Direct measurement of salt bridge solvation energies using a peptide model system: Implications for protein stability. *Proc Natl Acad Sci USA 93*:2985–2990.

Yang A-S, Honig B. 1995. Free energy determinants of secondary structure formation: I. α-Helices. *J Mol Biol 252*:351–365.

Yang A-S, Sharp KA, Honig B. 1992. Analysis of the heat capacity dependence of protein folding. *J Mol Biol 227*:889–900.