Protein Science (1994), *3*:2358-2365. Cambridge University Press. Printed in the USA.
Copyright © 1994 The Protein Society

Determinants of protein side-chain packing

RYUJI TANIMURA, AKINORI KIDERA, AND HARUKI NAKAMURA
Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan

(RECEIVED July 20, 1994; ACCEPTED September 16, 1994)

Abstract

The problem of protein side-chain packing for a given backbone trace is investigated using 3 different prediction models. The first requires an exhaustive search of all possible combinations of side-chain conformers, using the dead-end elimination theorem. The second considers only side-chain–backbone interactions, whereas the third neglects side-chain–backbone interactions and instead keeps side-chain–side-chain interactions. Predictions of side-chain conformations for 11 proteins using all 3 models show that removal of side-chain–side-chain interactions does not cause a large decrease in the prediction accuracy, whereas the model having only side-chain–side-chain interactions still retains a significant level of accuracy. These results suggest that the 2 classes of interactions, side-chain–backbone and side-chain–side-chain, are consistent with each other and work concurrently to stabilize the native conformations. This is confirmed by analyses of energy spectra of the side-chain conformations derived from the fourth prediction model, the Independent model, which gives almost the same quality of the prediction as the dead-end elimination. The analyses indicate that the 2 classes of interactions simultaneously increase the energy difference between the native and nonnative conformations.

Keywords: consistency in side-chain packing; dead-end elimination; energy spectra; Independent model; prediction of side-chain conformation; side-chain packing; side-chain rotamer

Individual proteins are characterized by their native backbone folds and side-chain arrangements. Because more than half of the degrees of freedom are required in defining the latter, the problem of side-chain packing is an important subproblem in protein folding. To tackle this problem of side-chain packing, it has often been assumed that the main-chain atoms are fixed at their native coordinates (Lee & Subbiah, 1991; Tufféry et al., 1991, 1993; Desmet et al., 1992; Dunbrack & Karplus, 1993). This assumption makes the problem far more tractable. When the main-chain conformation is considered as variable, the conformational energy, E, is given by,

$$E(\mathbf{c},\mathbf{\Phi}) = E_{\text{template}}(\mathbf{\Phi}) + \sum_i E_1(c_i) + \sum_i E_2(c_i,\psi_i) + \cdots + \sum_i E_4(c_i,\psi_i,\phi_{i+1},c_{i+1}) + \cdots \quad (1)$$

where the side-chain χ angles for residue i are collectively represented as c_i [$\mathbf{c} = (c_1, c_2, \ldots)$]; $\mathbf{c}$ corresponds to a set of side-chain rotamers (Ponder & Richards, 1987). The main-chain (ϕ,ψ) angles [$\mathbf{\Phi} = (\phi_1, \psi_1, \ldots)$] are explicitly written in the equation. E_{template} is the energy term for interactions within the main-chain atoms, and E_m is the interaction energy described by m variables. The number m is up to $2N$, where N is the number of residues in a protein. This coupling of a large number of variables hinders a direct enumeration of all possible conformations and represents one of the major complications faced in the folding problem. However, when the main-chain atoms are considered as fixed, Equation 1 becomes much simpler, as has been described by Desmet et al. (1992),

$$E(\mathbf{c}) = E_{\text{template}} + \sum_i E_1(c_i) + \sum_{i<j} E_2(c_i,c_j). \quad (2)$$

E_{template} is now constant, and the coupling of more than 2 variables disappears in Equation 2.

The number of possible combinations of side-chain conformations (c_i and c_j in E_2) is still too large to be enumerated exhaustively, but due to the simplicity of Equation 2, various heuristic approaches have successfully predicted side-chain conformations (Reid & Thornton, 1989; Summers & Karplus, 1989; Holm & Sander, 1991; Lee & Subbiah, 1991; Tufféry et al., 1991, 1993; Dunbrack & Karplus, 1993; Wilson et al., 1993; Koehl & Delarue, 1994). Whereas all these algorithms are based on approximations, Desmet et al. (1992) used Equation 2 to develop an exact method called dead-end elimination (DEE), which is able to determine the global minimum energy structure for all possible combinations of side-chain conformations.

Reprint requests to: Akinori Kidera, Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka 565, Japan; e-mail: kidera@pes4.peri.co.jp.

Besides predicting the side-chain conformations, this simple expression of the side-chain packing (Equation 2) can also give valuable information about the factor determining the native side-chain conformations. Eisenmenger et al. (1993) presented a prediction model which suggested that the combinatorial problem of side-chain conformations hardly exists because the position of each side-chain is essentially determined by the environment provided by the backbone atoms. Because this simplistic approach gave good predictions, it was concluded that side-chain–side-chain interactions (E_2 of Equation 2) can be neglected from the calculation of side-chain conformations.

Do these prediction results really mean that side-chain–side-chain interactions are unimportant in determining the side-chain conformations? The purpose of this paper is to answer this question by interpreting the role of the energy term $E_2(c_i, c_j)$ of Equation 2. For this purpose, we first compare the DEE model with implementations of Equation 2 that set either $E_1 = 0$ or $E_2 = 0$. Differences in the prediction results should help to explain the importance of E_2 in prediction.

However, prediction by energy minimization treats only the minimum energy structure. To elucidate the mechanism by which side-chain conformations are determined, it is important to consider the characteristics of the native state relative to other conformations or, more specifically, how other conformations are eliminated in the search for a global minimum. Such information can be obtained from the energy spectra for all possible side-chain conformations. Unfortunately, the DEE method does not give such energy spectra because it determines only the global minimum energy structure. The A* algorithm of combinatorial optimization gives all structures within a certain energy level (Leach, 1994), but this algorithm is computationally too intensive for a large protein. Therefore, we have developed an approximate method that gives good agreement with the DEE model while enabling us to evaluate the energy spectra.

Using the DEE model and our new model, we analyze the interaction energy in detail and discuss what determines the unique and stable side-chain conformations found in native proteins.

Prediction models

Here we explain the various prediction models used in this paper. In all models, discrete side-chain conformations, i.e., rotamers, are assumed for c_i in Equation 2 (see Methods for the definition). Minimization of Equation 2 in terms of $\mathbf{c}$, i.e.,

$$\min_{\mathbf{c}}[E(\mathbf{c})] = \min_{\mathbf{c}}\left[\sum_i E_1(c_i) + \sum_{i<j} E_2(c_i, c_j)\right], \qquad (3)$$

is solved by the DEE method (Desmet et al., 1992) and by 3 kinds of approximation: the Independent, Template, and Phantom-Template models.

The global minimum energy conformation is found by the DEE theorem. A rotamer r_i for residue i can be eliminated from the conformational search when there is another rotamer s_i that satisfies

$$E_1(r_i) + \sum_{j\neq 1} \min_{c_j}[E_2(r_i, c_j)] > E_1(s_i) + \sum_{j\neq 1} \max_{c_j}[E_2(s_i, c_j)]. \qquad (4)$$

Equation 4 is applied successively to all possible rotamers to eventually give a set of rotamers defining the global minimum energy conformation.

Now we introduce an approximation to the problem of Equation 3. This approximation reduces the degrees of freedom searched to those of a subset of M ($<N$) side chains $\mathbf{a}$ [$=(a_i, a_k, \ldots)$; $\mathbf{a} \subset \mathbf{c}$] whose conformations are predicted. Each of the remaining side chains $\mathbf{b}$ ($=\mathbf{c} - \mathbf{a}$; a subset of $N - M$ side chains) is frozen in the conformation that gives the minimum interaction energy with the set $\mathbf{a}$. We call this the minimum-field approximation. Because the subset $\mathbf{b}$ is uniquely determined for a given set of $\mathbf{a}$, we do not have to consider the combinations of the 2 subsets $\mathbf{a}$ and $\mathbf{b}$. The same operation is repeated for subset $\mathbf{b}$ to determine the conformations of the remaining side chains. Instead of Equation 3, we have an approximation,

$$\min_{\mathbf{c}}[E(\mathbf{c})] \sim \min_{\mathbf{a}}[E_{mf}(\mathbf{a})] + \min_{\mathbf{b}}[E_{mf}(\mathbf{b})], \qquad (5)$$

where the minimum-field energy, E_{mf}, is written by,

$$E_{mf}(\mathbf{a}) = \sum_{i\in\mathbf{a}} E_1(a_i) + \sum_{i<k\in\mathbf{a}} E_2(a_i, a_k) + \frac{1}{2}\sum_{j\in\mathbf{b}} \min_{b_j}\left[\sum_{i\in\mathbf{a}} E_2(a_i, b_j)\right]. \qquad (6)$$

$E_{mf}(\mathbf{b})$ is obtained by exchanging $\mathbf{a}$ with $\mathbf{b}$ in Equation 6. Because the 2 sets of conformations for $\mathbf{a}$ and $\mathbf{b}$ are independently determined, these subsets $\mathbf{a}$ and $\mathbf{b}$ are not necessarily consistent with each other. Thus, we call this the Independent model. In this study, we use the simplest approximation for N subsets with $M = 1$, i.e.,

$$\min_{\mathbf{c}}[E(\mathbf{c})] \sim \sum_i \min_{a_i}[E_{mf}(a_i)] \qquad (7)$$

where E_{mf} is now,

$$E_{mf}(a_i) = E_1(a_i) + \frac{1}{2}\sum_{j\neq 1} \min_{b_j}[E_2(a_i, b_j)]. \qquad (8)$$

To determine the side-chain conformation of residue i, we search the rotamers of all the other residues j ($=1, \ldots, i-1, i+1, \ldots, N$) to find which conformations give the lowest interaction energy with the rotamer a_i. This operation is repeated for all possible rotamers in residue i, and the rotamer, a_i, giving the lowest energy is chosen as the prediction. Equation 8 is applied to all side chains, $i = 1, \ldots, N$. Because minimization in this model simply involves sorting E_1 and E_2, the energy spectra of all side-chain conformations are easily evaluated.

A more crude approximation is given by ignoring all terms of E_2 from Equation 3 or 8, i.e.,

$$\min_{\mathbf{c}}[E(\mathbf{c})] \sim \sum_i \min_{c_i}[E_1(c_i)]. \qquad (9)$$

We call this the Template model, and it essentially corresponds to Eisenmenger's prediction model (1993) except we consider side-chain rotamers rather than continuous χ angles.

Instead of ignoring E_2, it is possible to remove E_1 from Equations 7 and 8 as,

$$\min_{\mathbf{c}}[E(\mathbf{c})] \sim \sum_i \min_{a_i}\left\{\frac{1}{2}\sum_{j\neq i}\min_{b_j}[E_2(a_i,b_j)]\right\}, \tag{10}$$

which is called the Phantom-Template model. In this situation, the DEE theorem (Equation 3 with Equation 4) cannot be applied because of the conformational degeneracies found for residues exposed to solvent. The Template and Phantom-Template models provide a complementary set to assess the importance of the E_2 term.

A schematic example of the side-chain prediction is given in Figure 1.

Table 1. *Proteins considered in the prediction of the side-chain structure*

Name of protein	PDB code	Resolution (Å)	Number of residues[a]
Crambin	1CRN	1.5	46 (26)
Ovomucoid third domain	2OVO	1.5	56 (39)
Trypsin inhibitor	5PTI	1.0	58 (36)
L7/L12 C-terminal domain	1CTF	1.7	68 (46)
Ubiquitin	1UBQ	1.8	76 (65)
Plastocyanin	2PCY	1.8	99 (77)
Insulin dimer	3INS	1.5	102 (76)
Thioredoxin	2TRX	1.68	108 (80)
Lysozyme[b]		1.5	130 (95)
Interleukin-1β	1I1B	2.0	151 (132)
Papain	9PAP	1.65	212 (154)

[a] Numbers in parentheses are the numbers of residues other than Ala, Cys (S-S form), Gly, or Pro.

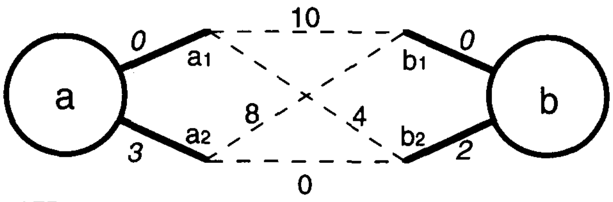
[b] Coordinates of lysozyme are those determined by Kidera et al. (1994), to be submitted to PDB.

Results and discussion

Predictions of side-chain conformations

We predicted the side-chain conformations of 11 proteins, which have a variety of sizes and folding patterns (Table 1). Four different prediction models, DEE, Independent, Template, and Phantom-Template models, were implemented, each with 3 different sets of rotamer libraries, small size, medium size, and large size (Table 2). The prediction results are summarized in Table 3, together with the upper and lower limits of prediction accuracy (the prediction of lysozyme is illustrated in Fig. 2). Details for each protein are summarized in Tables S1, S2, and S3, available as supplementary materials on the Diskette Appendix.

When either the small-size or medium-size rotamer library was used, the DEE successfully converged to give the global minimum energy conformation (the DEE models). In papain, with the medium-size library, DEE was able to determine the global



DEE

Combination	Energy
(a1,b1)	*0* + *0* + 10 = 10
(a1,b2)	*0* + *2* + 4 = 6
(a2,b1)	*3* + *0* + 8 = 11
(a2,b2)	*3* + *2* + 0 = 5*

Approximation

	Energy		
Rotamer	Independent (Eq. 8)	Template (Eq. 9)	Phantom-Template (Eq. 10)
a1	*0* + 4/2 = 2*	*0**	4/2 = 2
a2	*3* + 0/2 = 3	*3*	0/2 = 0*
b1	*0* + 8/2 = 4	*0**	8/2 = 4
b2	*2* + 0/2 = 2*	*2*	0/2 = 0*

Fig. 1. A schematic example after Lasters and Desmet (1993) explains how the 4 models are determined. This example is composed of 2 residues *a* and *b*, each of which has 2 possible rotamers 1 and 2. The values of E_1 (Equation 2) are given in italics. The sc-sc interaction energy, E_2, is represented by dashed lines with their values. The DEE model is the true minimum energy combination, (a_2,b_2), among the 4 possible combinations. Asterisks indicate the rotamers chosen by the model. The Independent model chooses the rotamers (a_1,b_2), each of which yields the lowest value of E_{mf} defined by Equation 8. The Template model, (a_1,b_1), neglects E_2, whereas the Phantom-Template model, (a_2,b_2), considers only E_2.

Table 2. *Number of rotamers in the 3 sets of rotamer libraries*[a]

	Small size	Medium size	Large size
Arg	11	15	76
Asn	3	9	21
Asp	3	6	18
Cys	3	3	27
Gln	6	6	33
Glu	7	7	29
His	7	54	67
Ile	5	6	39
Leu	8	12	28
Lys	13	17	41
Met	10	13	34
Phe	5	32	32
Ser	3	3	27
Thr	3	3	27
Trp	7	40	48
Tyr	4	34	68
Val	3	3	9
Total	101	263	624

[a] The rotamer libraries are determined from the following 49 crystal structures whose PDB codes are: 1CSE(I), 1CSE(E), 7RSA, 1UTG, 4PTP, 1ECA, 1MBD, 256B(A), 8ABP, 2SGA, 3B5C, 4CPV, 3GRS, 5CPA, 2ER7(E), 2RHE, 351C, 3TLN, 4BP2, 2CPP, 3LZM, 2WRP(R), 6XIA, 2CCY(A), 1FKF, 2CYP, 2LTN(A), 2LTN(B), 3DFR, 3CLA, 1GD1(O), 1TGS(I), 2AZA(B), 2CDV, 2RNT, 4FXN, 2TIM(B), 1FD2, 2FB4(H), 2TSC(A), 1BBP(A), 1GOX, 1GP1(B), 1HOE, 1LH1, 1R69, 2CAB, 2MHR, 2RSP(B). The codes in parentheses are the chain names.

Table 3. *Results of predictions*[a]

	Random prediction		Phantom-Template model		Template model		Independent model		DEE model		Best rotamer	
Rotamer library	RMSD (Å)	C (%)	RMSD (Å)	C (%)	RMSD (Å)	C (%)	RMSD (Å)	C (%)	RMSD (Å)	C (%)	RMSD (Å)	C (%)
Small size	3.33	22	2.62	41	2.13	64	1.91	69	2.01	67	0.88	89
	3.35	23	2.21	52	1.76	72	1.68	77	1.69	74	0.70	94
Medium size	3.39	21	2.77	38	2.03	63	1.76	68	1.75	68	0.76	94
	3.40	22	2.42	51	1.46	73	1.10	80	1.21	79	0.54	97
Large size	3.31	26	2.76	42	2.04	65	1.78	70			0.58	95
	3.36	27	2.25	53	1.58	77	0.95	85			0.42	99

[a] The average values of the predictions for the 11 proteins listed in Table 1, weighted by each residue number. RMSD is the RMS deviation of side-chain coordinate from the corresponding X-ray structure and C is percentage of correctly predicted χ angles; upper for all residues and lower for core residues. C_β atoms are excluded from the RMSD calculation. The correct χ angles are defined by the criterion that both χ_1 and χ_2 have deviations within 40° from the corresponding X-ray values. There are 6 models using 3 sets of rotamer libraries. Because the large-size library did not give DEE models for 8 proteins, the average values are not given. Random prediction is the result of averaging all possible rotamers with equal probability, which provides the lowest limit of the prediction accuracy. Best rotamer gives the upper limit of the accuracy and is defined by the set of rotamers having the smallest RMSD value with the X-ray structure.

minimum energy structure from 2.9×10^{146} possible rotamer combinations. However, with the large-size library, the algorithm only reached convergence for 3 small proteins (crambin, ovomucoid third domain, and trypsin inhibitor). This is because the large-size library yields many rotamer combinations with similar energy levels, which cannot therefore be eliminated by the DEE theorem (Equation 4).

Effects of rotamer library size on prediction

Improvements of the prediction are attained by increasing the size of the rotamer library, particularly for buried residues in the Independent model (Table 3). There are 2 major sources of error in the global minimum energy structures. First, the energy term used here does not consider solvent effects properly. Even a simple surface area model of solvation is not compatible with Equation 2 because the calculation of surface area requires 4-body interactions (Kratky, 1981). This is one reason the predictions of exposed residues are worse than those of buried residues. A second problem is the restricted search of rotamer space over discrete orientations. Because the discrete rotamers do not perfectly match with the X-ray structure, the best rotamer could have a bad contact with the other atoms. Such a rotamer of high energy would not be chosen as the prediction and therefore could lead to erroneous results. As shown in Table 3, increasing the size of the library alleviates this problem by providing rotamers that are closer to the X-ray structure and therefore improving the prediction accuracy.

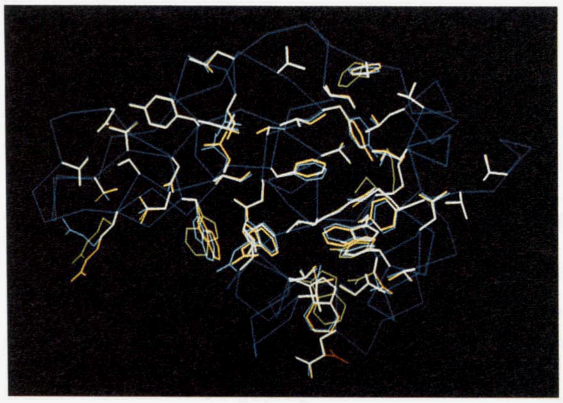


Fig. 2. Comparison between the predicted models and the X-ray structure of human lysozyme. As the main-chain trace, Cα atoms are shown (dark blue). Only side-chains in the core regions are shown for X-ray structure (light blue), DEE (yellow), Independent (red), and Template models (green). At most residues, the Independent model agrees well with the DEE model. Side chains in white indicate overlap of these models.

Template and Phantom-Template models

The difference between the Template and DEE models is in the neglect of the side-chain–side-chain (abbreviated as sc-sc) interactions, the E_2 term of Equation 2. In Table 3, it is seen that the E_2 term gives rather small improvements in the accuracy (about 5%). This finding is consistent with the prediction results of Eisenmenger et al. (1993).

The results of the Template model suggest that sc-sc interactions play only a minor role in determining side-chain conformations. Nevertheless, the Phantom-Template model, containing only the sc-sc interactions, retains 42% (core 53%) accuracy, which is much better than a random prediction. Because the Phantom-Template model does not contain either main-chain (mc) atoms nor C_β atoms, the number of interacting atoms considered is much smaller than in the Template model. However, the Phantom-Template model still keeps a significant level of the accuracy. These results suggest that the sc-sc and the sc-mc interactions work concurrently to stabilize the native state. As a first approximation, the native side-chain conformation, being the minimum energy conformation, should correspond not only to the minimum sc-mc interaction energy, but also to the minimum sc-sc interaction energy. The 2 classes of interactions are consistent with each other and work simultaneously to stabilize the same side-chain conformations. This consistency may account for the good level of prediction quality of both methods, in spite of their different treatments of the side-chain interactions.

Independent model

When comparing the Template and DEE models in detail, we notice that the 2 models show agreement of 74% (core 79%) (Table 4), whereas the Independent model resembles the DEE model much more closely, about 10% better agreement, and gives almost the same prediction accuracy (Table 3). This suggests that the minimum-field approximation of the Independent model correctly estimates the sc-sc interactions. The advantage of the Independent model is in its computational simplicity. The DEE method requires an iteration procedure until all rotamers except one are eliminated. On the other hand, the Independent model can be obtained by a simple sorting procedure using any size of rotamer library. Therefore, the Independent model with the large-size library gives significantly better predictions than the DEE model with the medium-size library (Table 3). In addition, it gives information on the rotamer energies other than the global minimum, thus enabling us to evaluate energy spectra for the side-chain conformations. The following discussion is based on the Independent model with the large-size rotamer library.

Table 4. *Percentage of agreement between the 3 prediction models*[a]

	Template	Independent
Independent	80 (85) 77 (82) 75 (84)	
DEE	76 (80) 74 (79) –	88 (90) 84 (89) –

[a] The 3 values of percentage agreement between 2 models, top, middle, and bottom, are those calculated with the small-size, medium-size, and large-size rotamer libraries, respectively. The values in parentheses are of core residues.

Energy spectra of side-chain conformations

The minimum energy conformation gives only limited information about the energetics of the side-chain structure. For a better understanding, we should investigate the mechanism by which incorrect rotamers are eliminated. This can be done with the energy spectra of all possible side-chain conformations. The Independent model provides the energy spectra for individual residues in the form of Figure 3. The spectra of crambin provide an example showing how the native conformations are selected. For buried residues, nonnative conformations give very high energies, whereas exposed residues can give low energies even for nonnative conformations. Figure 3 also demonstrates the close similarity between the DEE and Independent models.

Here we define the discriminating power of the native conformation by the energy difference between 2 conformations at a correctly predicted residue; one is of the lowest energy (the correct conformation) and the other is the next lowest energy with an incorrect conformation. When this difference is large, the side-chain conformation should be unique and stable, and the prediction should be reliable. Figure 4 shows 2 conformations of papain defining the discriminating power, the lowest energy (Independent model), and the next lowest energy conformations.

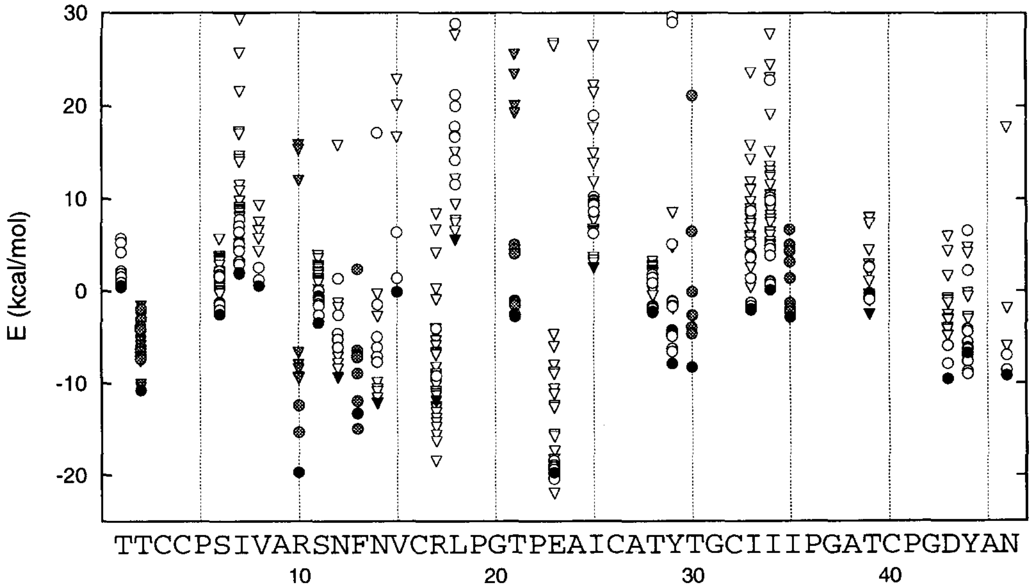


Fig. 3. Energy-spectra of side-chain rotamers in the Independent model of crambin (1CRN) calculated with the large-size rotamer library. The conformational energy for each side-chain rotamer is represented by either of 2 symbols, ○ for correctly predicted and ▽ for incorrect. Correctness is defined by the same criterion as the correct χ angles in Table 3. Shaded symbols are those of buried residues. The lowest energy rotamers correspond to the Independent model. Filled symbols indicate rotamers of the DEE model, which coincide well with the Independent model. Spectra for Ala, Cys (S-S form), Gly, and Pro are not given here because they are considered part of the template.

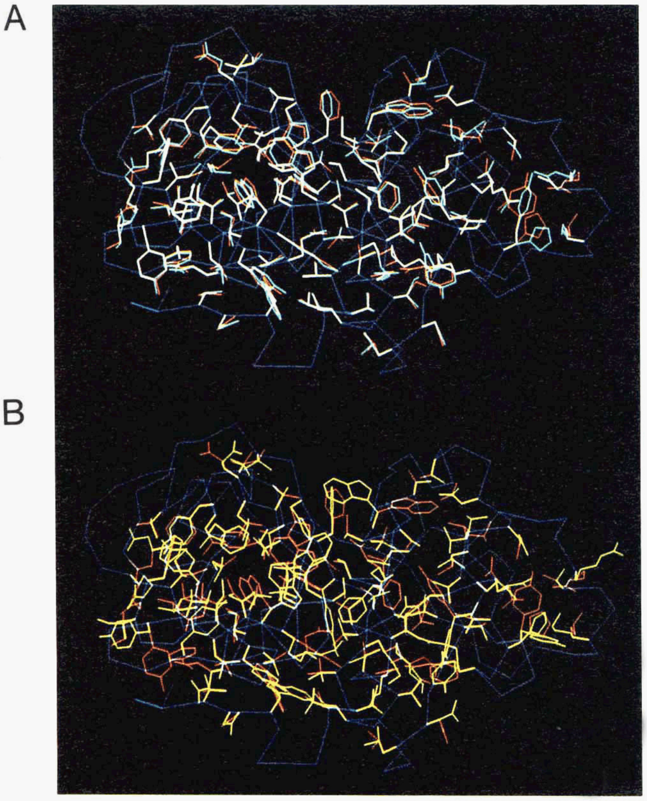


Fig. 4. Two conformations of papain defining the discriminating power. **A:** The lowest energy conformation corresponding to the Independent model (red) is compared with the X-ray structure (light blue). **B:** The second lowest energy conformation (yellow) with different χ angles (χ_1 and/or χ_2 are different from the corresponding angles in A by more than 40°) is shown with the Independent model (red). As in Figure 1, Cα atoms are shown as the main-chain trace (dark blue) and side chains in the core regions are shown.

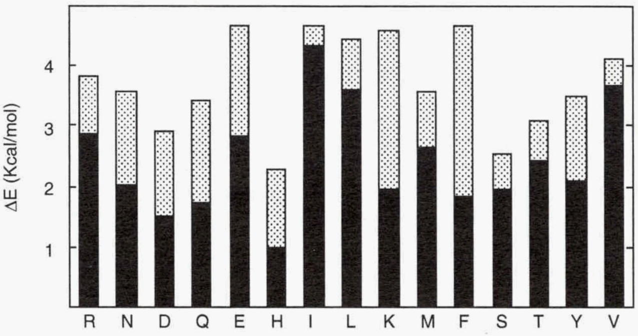


Fig. 5. Average discriminating power ΔE of the native side-chain conformation for each amino acid type. ΔE consists of 2 contributions, the sc-mc interactions (black) and the sc-sc interactions (shaded). The values for Cys (SH form) and Trp are not given because there are too few cases. These values have been calculated by the Independent model for the 11 proteins using the large-size rotamer library. Discriminating power is defined by the energy difference between 2 rotamers for correctly predicted sites; one is of the lowest energy (thus correct conformation) and the other is of the lowest energy in the incorrectly predicted rotamers. To avoid large statistical errors, statistics are taken only for energy differences less than 10 kcal/mol. Some of the incorrectly predicted rotamers have bad contacts with an extremely high interaction energy.

The discriminating power can be decomposed into sc-mc and sc-sc interaction contributions. Figure 5 shows the average discriminating power for each amino acid type. Both sc-mc and sc-sc interactions increase the discriminating power. This result is consistent with the Phantom-Template model, which also shows that these interactions simultaneously favor the native side-chain conformation. It is noted that the Independent model always underestimates the sc-sc interactions because of its minimum-field approximation, and therefore the real sc-sc interactions should be larger than the values shown in Figure 5.

Figure 6 shows which interactions penalize incorrect conformations, sc-mc or sc-sc. As in Figure 5, the energy difference is calculated for each of the correctly predicted residues between the lowest energy rotamer and all other rotamers having an energy difference of less than 30 kcal/mol. This threshold is used because the Independent model is chosen from the rotamers whose backbone interaction energies are less than 30 kcal/mol (see Methods for detail). When the energy difference is decomposed into the sc-mc and sc-sc contributions, we regard the interaction giving a larger positive contribution as the one penalizing the rotamer. In Figure 6, the average fraction of rotamers penalized by the sc-sc interactions is summarized for each amino acid type. This is another view of the discriminating power of the native conformation. We observe a good agreement with the results of Figure 5; the correlation coefficient between 2 values, the sc-sc contributions in Figure 5 and the fractions in Figure 6, is 0.69. Both of these interactions not only stabilize the native conformation but also destabilize nonnative conformations.

The number of the rotamers with energies close to the minimum shows a picture consistent with the discriminating power. Figure 7 shows the average number of the rotamers in the buried residues whose energies are within 3 kcal/mol of the lowest energy. Only 62 (62/100)% of the rotamers in the Template model have correct conformations. The sc-sc interactions in the Independent model divide the rotamers of the Template model into 2 parts; about half (55 = 42 correct + 13 incorrect) with 76 (42/55)% accuracy are kept within 3 kcal/mol, and the rest (45 = 20 correct + 25 incorrect) with only 44 (20/45)% accuracy

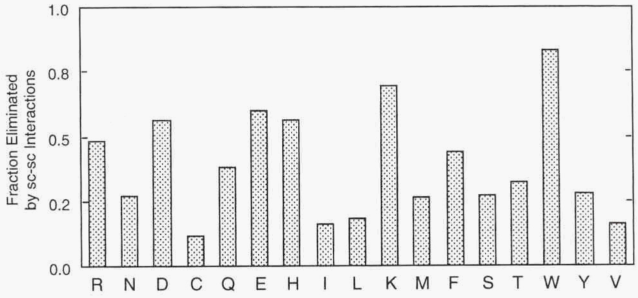


Fig. 6. Fraction of rotamers penalized by sc-sc interactions. These values were calculated from the Independent model of the 11 proteins with the large-size library. This is based on the energy difference between 2 rotamers at a correctly predicted residue; one is of the lowest energy and the other is any rotamer of incorrect conformation. When the energy difference is decomposed into the sc-mc and sc-sc contributions, the rotamer is regarded as being penalized by the interaction giving the larger positive contribution. Statistics were taken only for rotamers having an energy difference less than 30 kcal/mol. Because of the large number of rotamers for each residue, the values of Cys (SH form) and Trp are statistically significant.

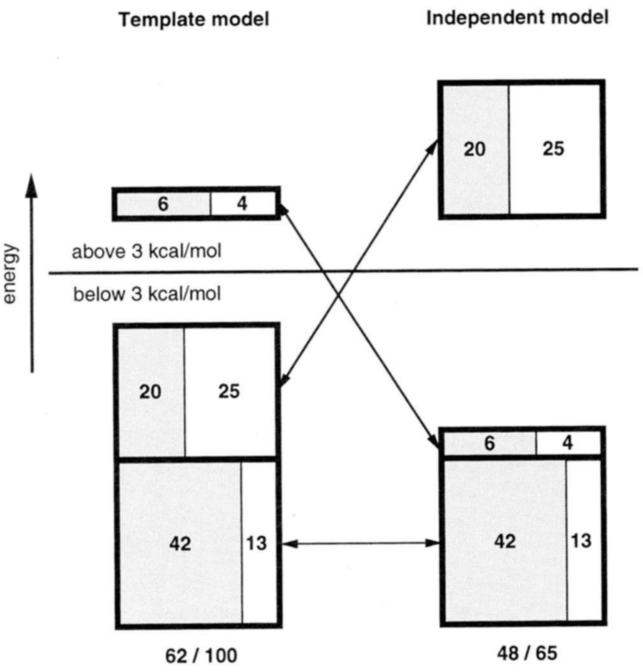


Fig. 7. Number of rotamers having low interaction energies. The number of rotamers for buried residues are counted for the Template and Independent models if a rotamer has an interaction energy within 3 kcal/mol of the lowest energy rotamer. These quantities were calculated with the large-size library. The size of each block corresponds to the number of rotamers and is normalized against the number of rotamers of the Template model within 3 kcal/mol. The shaded area is for the correctly predicted rotamers and the white area is for incorrectly predicted rotamers. Arrows indicate the correspondence between 2 models.

are eliminated from the low energy region. In this manner, the sc-sc interactions enhance the discriminating power, and increase the prediction accuracy from 62% to 74 (48/65)%.

Conclusion

Using the 4 prediction models, the Phantom-Template, Template, Independent, and DEE models, we investigated the role of sc-sc interactions in determining side-chain conformations. The sc-mc and sc-sc interactions work concurrently to favor the native conformations. This complementary nature would discriminate in favor of the unique and stable side-chain conformations found in native proteins.

This conclusion corresponds to the consistency principle of Gō (1983) or the principle of minimum frustration of Bryngelson and Wolynes (1987). These concepts state that the short-range interactions governing secondary structure propensities are consistent (or not in conflict) with the long-range interactions. Such a consistency has been found here in partitioning the interactions into sc-sc and sc-mc interactions.

Methods

We predicted side-chain conformations of 11 proteins (Table 1) by 4 different models using 3 different sets of rotamer libraries (Table 2). Here we explain the computational details.

For the DEE, we adopted the original algorithm of Desmet et al. (1992) and Lasters and Desmet (1993). First, E_1 and E_2 of Equation 2 are calculated for all rotamers, and rotamers having high interaction energies (>30 kcal/mol) with the backbone atoms are eliminated. Then, Equation 4 is successively applied to each rotamer and their pairwise combinations. Further, the generalized dead-end theorem for multiple residues is used with the "add on" procedure. After the DEE operation, we did not perform any further energy minimization.

Calculations of the Independent, Template, or Phantom-Template model are simply done by sorting E_1 and E_2 of Equation 2. The elimination of rotamers by the 30-kcal/mol threshold is also applied to the Independent model. This process ensures that rotamers having bad contacts do not contribute to the minimum field of the Independent model. Therefore, these calculations correspond to the first stage of the DEE algorithm.

The conformational energy in the computation is the sum of the interaction energies given by the all-atom parameters of the AMBER forcefield (Weiner et al., 1986), where the bond and angle energies are not included because of the rotamer approximation. The dielectric constant, ϵ, for the electrostatic potential was the distance-dependent $\epsilon = 2r$ where r is in Å. No cut-off operation was applied. Ala, Cys with a disulfide bridge, Gly, and Pro were all regarded as part of the template, as were C_β atoms. The coordinates of hydrogen atoms were generated from the standard AMBER geometries. The hydrogen coordinates given in the Protein Data Bank (PDB; Bernstein et al., 1977) entries were not used. Where the PDB entry has disorders with alternate locations, the coordinates of the larger occupancy or of the identifier A (when occupancy = 0.5) were used.

The rotamer libraries used here were prepared from the structure data of 49 nonhomologous proteins whose PDB codes are listed in the caption of Table 2. These nonhomologous proteins were chosen by the following criteria: the 11 test proteins listed in Table 1 were excluded; the resolution must be better than 2.0 Å; the stereochemical quality must satisfy the 3 criteria defined by Morris et al. (1992); class 1 or 2 for the ϕ, ψ distribution, the χ_1 standard deviation, and the hydrogen bond energy. Any side-chain having an atom with B-factor > 30 Å^2 was ignored in the analysis. We prepared 3 sets of libraries, small size, medium size, and large size, to examine the effects of the library size on the prediction accuracy. The number of rotamers in each library is summarized in Table 2. The differences among these 3 sets are in the number of rotamers assigned to each cluster in the χ angle distribution and in whether the rotamers are assigned for minor clusters. The small-size library corresponds to the Ponder and Richards library (1987). By increasing the number of rotamers for aromatic amino acids according to Desmet et al. (1992), we obtained the medium-size library. Finally, the number of rotamers for flexible side chains was increased in the large-size library.

The prediction models were assessed mainly from the RMS deviation (RMSD) of side-chain atoms from the corresponding X-ray coordinates and the percentage of correctly predicted χ angles. C_β atoms were not counted in the RMSD. The correct χ angles were defined by the criterion that both χ_1 and χ_2 deviate by less than 40° from the corresponding X-ray values. Only χ_1 angles were assessed for amino acids having no χ_2 angle. For both the RMSD values and the correct χ angles, the following pairs of atoms were treated as equivalent: OD1 and ND2 of Asn, OE1 and NE2 of Gln, ND1 and CD2 for His, and NE2 and CE1 for His. The symmetry of the side-chain structure was also considered for Arg, Asp, Glu, Phe, and Tyr. Statistics are given either for all side chains or for buried side chains. The buried side chains are defined by the criterion that the accessible surface area (Shrake & Ruply, 1973) of the side-chain atoms is less than 30% of the value of a tripeptide Gly-X-Gly.

All computations were performed on a DEC 3000 500X workstation.

Supplementary materials on the Diskette Appendix

Detailed results of the predictions, RMSD values against the X-ray structures, and percentages of correctly predicted χ_1 and χ_2 angles are summarized in 3 tables as supplementary materials. Tables S1, S2, and S3 are for the small-size, medium-size, and large-size rotamer libraries, respectively. The files are located in the Tanimura.SUP subdirectory of the SUPLEMNT directory.

Acknowledgments

We are grateful to Drs. M.B. Swindells and P. Jewsbury for stimulating discussions.

References

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.

Bryngelson JD, Wolynes PG. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci USA 84*:7524–7528.

Desmet J, De Maeyer M, Hazes B, Lasters I. 1992. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature 356*:539–542.

Dunbrack RL Jr, Karplus M. 1993. Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *J Mol Biol 230*:543–574.

Eisenmenger F, Argos P, Abagyan R. 1993. A method to configure protein side-chains from the main-chain trace in homology modelling. *J Mol Biol 231*:849–860.

Gō N. 1983. Theoretical studies of protein folding. *Annu Rev Biophys Bioeng 12*:183–210.

Holm L, Sander C. 1991. Database algorithm for generating protein backbone and side-chain co-ordinates from a C^{α} trace. Application to model building and detection of co-ordinate errors. *J Mol Biol 218*:183–194.

Kidera A, Inaka K, Matsushima M, Gō N. 1994. Response of dynamic structure to removal of a disulfide bond: Normal mode refinement of C77A/C95A mutant of human lysozyme. *Protein Sci 3*:92–102.

Koehl P, Delarue M. 1994. Application of a self-consistent mean field theory to predict protein side-chains' conformation and estimate their conformational entropy. *J Mol Biol 239*:249–275.

Kratky KW. 1981. Intersecting disks (and spheres) and statistical mechanics. I. Mathematical basis. *J Statist Phys 25*:619–634.

Lasters I, Desmet J. 1993. The fuzzy-end elimination theorem: Correctly implementing the side chain placement algorithm based on the dead-end elimination theorem. *Protein Eng 6*:717–722.

Leach A. 1994. Ligand docking to proteins with discrete side-chain flexibility. *J Mol Biol 235*:345–356.

Lee C, Subbiah S. 1991. Prediction of protein side-chain conformations by packing optimization. *J Mol Biol 217*:373–388.

Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. 1992. Stereochemical quality of protein coordinates. *Proteins Struct Funct Genet 12*:345–364.

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

Reid LS, Thornton JM. 1989. Rebuilding flavodoxin from Cα co-ordinates: A test study. *Proteins Struct Funct Genet 5*:170–182.

Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol 79*:351–371.

Summers NL, Karplus M. 1989. Construction of side-chains in homology modelling. Application to the C-terminal lobe of rhizopuspepsin. *J Mol Biol 210*:785–812.

Tufféry 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.

Tufféry P, Etchebest C, Hazout S, Lavery R. 1993. A critical comparison of search algorithm applied to the optimization of protein side-chain conformations. *J Comput Chem 14*:790–798.

Weiner SJ, Kollman PA, Nguyen DT, Case DA. 1986. An all atom force field for simulation for proteins and nucleic acids. *J Comput Chem 7*:230–252.

Wilson C, Gregoret LM, Agard DA. 1993. Modeling side-chain conformation for homologous proteins using an energy-based rotamer search. *J Mol Biol 229*:996–1006.