Eur Biophys J (2001) 30: 1–10
DOI 10.1007/s002490000111

ARTICLE

**Akira R. Kinjo · Akinori Kidera · Haruki Nakamura**
**Ken Nishikawa**

# Physicochemical evaluation of protein folds predicted by threading

Received: 8 May 2000 / Revised version: 1 September 2000 / Accepted: 1 September 2000 / Published online: 9 December 2000


**Abstract** Protein structure prediction remains an unsolved problem. Since prediction of the native structure seems very difficult, one usually tries to predict the correct fold of a protein. Here the "fold" is defined by the approximate backbone structure of the protein. However, physicochemical factors that determine the correct fold are not well understood. It has recently been reported that molecular mechanics energy functions combined with effective solvent terms can discriminate the native structures from misfolded ones. Using such a physicochemical energy function, we studied the factors necessary for discrimination of correct and incorrect folds. We first selected correct and incorrect folds by a conventional threading method. Then, all-atom models of those folds were constructed by simply minimizing the atomic overlaps. The constructed correct model representing the native fold has almost the same backbone structure as the native structure but differs in side-chain packing. Finally, the energy values of the constructed models were compared with that of the experimentally determined native structure. The correct model as well as the native structure showed lower energy than misfolded models. However, a large energy gap was found between the native structure and the correct model. By decomposing the energy values into their components, it was found that solvent effects such as the hydrophobic interaction or solvent shielding and the Born energy stabilized the correct model rather than the native structure. The large energetic stabilization of the native structure was attained by specific side-chain packing. The stabilization by solvent effects is small compared to that by side-chain packing. Therefore, it is suggested that in order to confidently predict the correct fold of a protein, it is also necessary to predict correct side-chain packing.

**Key words** Protein structure prediction · Molecular mechanics force field · Solvent effects · Side-chain packing

A.R. Kinjo · K. Nishikawa (✉)
Center for Information Biology,
National Institute of Genetics,
Mishima, Shizuoka 411–8540, Japan
E-mail: knishika@genes.nig.ac.jp
Tel.: +81-559-816859
Fax: +81-559-816889

A. Kidera
Department of Chemistry, Graduate School of Science,
Kyoto University, Sakyo-ku, Kyoto 606-8502, Japan

H. Nakamura
Institute for Protein Research, Osaka University,
3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

A.R. Kinjo
K. Nishikawa, Department of Genetics,
The Graduate University for Advanced Studies,
Mishima, Shizuoka 411-8540, Japan

## Introduction

Phrases such as "protein structure prediction is one of the most important but unsolved problems in the field of molecular biology" are already classical, but they still hold true. These statements refer to the cases when there is no apparent homology between the sequence of a protein of unknown structure and sequences of known structure. The ultimate goal of the protein structure prediction problem is the prediction of the native structure which is defined by both the backbone and side-chain conformations. It is believed that the native structure of a protein is of the lowest energy conformation. Therefore, given the correct energy function, predicting the native structure is equivalent to finding the conformation with the lowest energy. Since finding the global minimum conformation seems very difficult, one often tries to predict the native "fold" of a protein. Although the definition of a "fold" is in general artificial, it is usually interpreted as a set of similar backbone structures of proteins. A problem arises when one tries to find the native fold because a structure with the native fold is not necessarily the same as the native structure, and therefore there are no known physical factors that

determine the native fold. For the prediction of the native fold, coarse-grained protein models, in which residue-wise interactions are taken into account, are often used. Then the fold at the minimum of such residue-wise interaction potentials is assumed to be the native fold. Threading in general and some ab initio methods utilize such a strategy, but their success is currently limited. Typical residue-wise interaction potentials are derived from structural databases (Sippl 1995). Not only because the structural databases contain the sources of errors and noises, but also because the functional form of the potential depends on one's intuition, it is often difficult to identify what is wrong with the potentials, which in turn makes it difficult to improve them. It is therefore preferable to use more physically well-grounded potential functions. Although the "true" potential function is difficult to know, the conventional molecular mechanics force fields should be the best possible choice.

Since the pioneering work by Novotny et al. (1984), it has widely been believed that the molecular mechanics energy function is unable to discriminate the native structure from the misfolded ones. However, some authors recently have begun to claim that a molecular mechanics energy function combined with a hydration term can discriminate the native structure from the misfolded structure. Janardhan and Vajda (1998) investigated the use of a molecular mechanics energy function combined with solvation and entropic terms for selecting near-native structures among homology-based models. They showed that the solvation and molecular mechanics energy terms are useful for selecting models with good side-chain packing and well-built loops, respectively. Vorobjev et al. (1998) developed an elaborate method to calculate the conformational free energy of proteins, combining molecular dynamics simulation with explicit and implicit solvent models. Their method incorporates conformational entropy as well as the solvation free energy. They applied the method to the native and misfolded structures of nine small proteins and found that the native structures always gave lower conformational free energies than the misfolded ones. Lazaridis and Karplus (1999b) have developed an effective energy function that combines the conformational potential function with a simple solvent model. Their effective energy function was successful in discriminating the native structures from misfolded structures and hundreds of decoys (Lazaridis and Karplus 1999a).

In the present study, we investigate the physicochemical factors that are necessary for discriminating between the native folds and misfolds. To do so, we construct "near-native" models as representatives of the native (i.e., correct) fold. The backbone structure of a near-native model is almost the same as that of the native structure, but its side-chain conformation is different from that of the native structure. These near-native models are constructed by the same procedure as misfolded models. From comparison of the energy components of the native structure and near-native and misfolded models, it became possible to identify the factors that are necessary for the prediction of the native fold.

## Materials and methods

### Force field

For the energy minimization of all-atom models, we used the AMBER all-atom force field (Weiner et al. 1986) with a distance-dependent dielectric constant ($\epsilon = 2r$, where $r$ is the distance between two interacting atoms) with or without implicit solvent. We employed the implicit solvent model of Ooi et al. (1987), which is based on the solvent-accessible surface area. We call this hydration free energy "OONS." In order to incorporate the surface area-based implicit solvent into energy minimization, we used the analytical method of Richmond (1984) with minor modifications (Wesson and Eisenberg 1992). For the final evaluation of the total energy, the electrostatic energy based on the continuum dielectric model of the protein-water system was used instead of the Coulomb energy with the distance-dependent dielectric constant (see below).

### Decoys

In order to test the general ability of our energy function, we first apply it to the problem of discriminating the native structure out of hundreds of decoys. The four-state reduced decoy sets created by Park and Levitt (1996) were obtained from a web site (http://dd.stanford.edu/). Each decoy set consists of more than 600 decoy structures. The structures of the Park and Levitt decoy set were first energy-minimized for 300 steps using the AMBER potential function, excluding electrostatic and hydrogen bond terms, with the positional restraints on the backbone heavy atoms. Another 100 steps of conjugate gradient energy minimization were carried out using AMBER with the implicit solvent term. We used only three sets (1ctf, 1r69, 3icb) out of the seven sets by eliminating those having disulfide bonds or iron-sulfur clusters in order to make the comparison of different structures possible. Also the decoy set for 2cro (434 Cro protein) was not used because it is similar to 1r69 (434 repressor).

### Generation of models for correct and incorrect folds

We arbitrarily selected seven proteins of various structural classes (all $\alpha$, all $\beta$, $\alpha/\beta$, and $\alpha+\beta$) which are composed of approximately 100 amino acid residues. These proteins are called "target" proteins (Table 1) and we construct correctly and incorrectly folded models of these target proteins. All the protein structures were obtained from the Protein Data Bank (url: http://www.rcsb.org/pdb/) and the names of all the proteins in this paper are referred to by their Protein Data Bank (PDB) codes. The target proteins do not have disulfide bonds in the structural core. For those having a disulfide bond (i.e., 1rgeA and 1thx), we deleted the disulfide bonds so that all the cysteine residues are treated in their reduced form. The backbone structures of correctly or incorrectly folded models of a target protein were selected using a conventional threading method in the following manner. Each target sequence was threaded through structures whose number of residues was larger than that of the target in a fold library without gaps using the threading program S3 (Ota et al. 1999). The program S3 employs a sequence-structure compatibility function which is statistically derived from a structural database. The compatibility function consists of four terms: side-chain packing, hydration, local conformation, and hydrogen bonds. These terms are similar to those of Matsuo et al. (1995), but local structures are treated in more detail (Ota et al. 1999). The candidates for the predicted structures were selected

**Table 1** Target proteins

| Target[a] | Nres[b] | Class[c] | Name | Ref |
|---|---|---|---|---|
| 1lmb4 | 92 | All $\alpha$ | Lambda repressor | Clarke et al. (1991) |
| 1molA | 94 | $\alpha+\beta$ | Monellin | Tomic et al. (1992) |
| 1plc | 99 | All $\beta$ | Plastocyanin | Guss et al. (1992) |
| 1rgeA | 96 | $\alpha+\beta$ | Ribonuclease Sa | Sevcik et al. (1993) |
| 1thx | 108 | $\alpha/\beta$ | Thioredoxin | Saarinen et al. (1995) |
| 2hmzA | 113 | All $\alpha$ | Hemerythrin | Holmes and Stenkamp (1991) |
| 3chy | 128 | $\alpha/\beta$ | CheY protein | Volz and Matsumura (1991) |

[a] The Protein Data Bank (PDB) codes of the target proteins. The fifth letter, if present, indicates the chain identifier
[b] Number of residues of the target protein
[c] Structural classification according to the SCOP database (Murzin et al. 1995)

from those giving the top 10 scores. Those candidates are called "templates" and are listed in Table 2. In order to avoid non-compact structures, only those proteins whose sequences were longer than the target by at most five residues were used; the obtained template structures are indeed compact (Table 2).

Since only the backbone atoms are represented explicitly in the threading, we next have to build the side-chain conformations in order to apply the molecular mechanics energy function. The model building was done as shown in Table 3 using the program EMBOSS (Nakai et al. 1993), which was originally developed for structure determination by NMR spectroscopy. EMBOSS can perform efficient conformational sampling by simulated annealing mass-weighted molecular dynamics in four-dimensional space (Havel 1991; Nakai et al. 1993). A random coil was given as the initial conformation. After the simulated annealing in four-dimensional space, the weight $k_{4D}$ of the fourth-dimensional energy is increased to compress the fourth coordinate and to obtain the three-dimensional structure. From stages 1 to 4 in Table 3, the distance geometry force field was used, which consists of only local geometry terms and long-range soft repulsion terms, but with no attractive term. Positional restraints on all the backbone atoms were imposed throughout the optimization steps so that the backbone structure of the model becomes the same as the coarse-grained model used in the threading. Note that side-chain conformations are determined only by the repulsive term, that is, they are determined simply by minimizing the atomic overlaps with each other and with backbone atoms. A residue-based cutoff scheme was applied. A cutoff length of 6 Å was used through the stages 1–4 in Table 3, and 12 Å was used when a minimization involved the AMBER force field. The interaction tables were updated every 100 steps through stages 1–4, and every 20 steps for all other cases. For comparison, the native structure of each target determined experimentally was also minimized for 500 steps of the conjugate gradient method with positional restraints on the backbone atoms. By the minimization, the native structures deviated from the experimental structures by a root mean square deviation (RMSD; all heavy atoms) of at most 0.3 Å.

Computations were done on a VPP500 (Fujitsu) with vector processors and AP3000 (Fujitsu) workstations with Ultra SPARC-II (296 MHz) CPUs. A typical modeling procedure from stage 1 to stage 4 took about 1 h for each model on the VPP500. One hundred steps of conjugate gradient minimization with AMBER and the implicit solvent term took about 3 min for each model of 1thx (108 residues) on the AP3000.

## Continuum electrostatics calculation

In order to evaluate the electrostatic energy, especially the contribution from the reaction field of the protein-solvent system, continuum electrostatics calculations (Nakamura 1996) were carried out for the models constructed by the procedure described above. The electrostatic potential is obtained by numerically solving the Poisson equation:

$$\nabla \cdot \epsilon(\mathbf{r})\nabla\psi(\mathbf{r}) = -4\pi \sum_i q_i \delta(\mathbf{r}-\mathbf{r}_i) \quad (1)$$

where $\epsilon(\boldsymbol{r})$ and $\psi(\boldsymbol{r})$ are the dielectric constant and the electrostatic potential at the position $\boldsymbol{r}$, respectively, $q_i$ and $\boldsymbol{r}_i$ are the charge and the position of the $i$th protein atom; $\epsilon(\boldsymbol{r})$ is set to $\epsilon_p$ in the protein region, and to $\epsilon_s$ in the solvent region. The protein region is defined by the excluded volume, which in turn is defined by the van der Waals radii of the protein atoms. In order to avoid the divergence of the electrostatic potential due to the self-Coulomb energy of the point charges of the protein atoms, we solve the above equation twice, once for the protein-solvent system ($\epsilon_s = 80$) and again for the protein in a vacuum ($\epsilon_s = \epsilon_p$). Thus we obtain two solutions $\psi^{sol}$ and $\psi^{vac}$, respectively. By taking their difference, we obtain the reaction

**Table 2** Templates obtained by ungapped threading

| Target | Correct | | | Incorrect misfolded[d] | $R_g$ ratio[e] |
|---|---|---|---|---|---|
| | NN[a] | HM[b] | MA[c] | | |
| 1lmb4 | 1lmb4 | – | – | 1frrA 1molA 1nsgB 1pdr 1rgeA 1ris 1sxl 2hgf 2rgf | 0.93 (0.05) |
| 1molA | 1molA | – | – | 1hsbB 1nct 1plc 1prtF 1ris 1stfI 1sxl 2acy 2rgf | 0.95 (0.03) |
| 1plc | 1plc | – | – | 1ag2 1audA 1bftA 1dcpC 1hsbB 1lktA 1lt5D 1tlk 1tul | 1.06 (0.04) |
| 1rgeA | 1rgeA | – | – | 1aps 1audA 1hsbB 1iuz 1nct 1pcs 1stfI 1tiiD 2ncm | 1.04 (0.04) |
| 1thx | 1thx | 2trxA | 1tof | 1bcpD 1cewI 1csyA 1kpeA 1picA 1rblM 1rtu | 1.06 (0.06) |
| 2hmzA | 2hmzA | | 2mhr | 1cd8 1dutA 1hcd 1kb5A 1pbk 1rot 1tvdA 2rspA | 1.02 (0.04) |
| 3chy | 3chy | – | – | 135l 1a25A 1adl 1aizA 1bbhA 1bff 1cpq 1ftpA 1kuh | 1.06 (0.07) |

[a] The templates for "near-native" models
[b] The template for a "homologous" model
[c] The template for "misaligned" models
[d] The templates for "misfolded" models
[e] Average ratio (and its standard deviation in parentheses) for the radius of gyration of constructed models to that of the native structure

**Table 3** Protocol of simulated annealing optimization

| | A random coil is generated as the initial structure[a] |
|---|---|
| Stage 1 | 500 steps conjugate gradient minimization ($k_{4D} = 0.05$)[b] |
| | Distance geometry force-field |
| Stage 2 | 5000 steps molecular dynamics at 1000 K ($k_{4D} = 0.05$) |
| | Distance geometry force field |
| | Atomic mass 1000 Da, step size 50 fs, coupling constant 40 ps |
| Stage 3 | 100,000 steps molecular dynamics to 1 K ($k_{4D} = 0.05$) |
| | Distance geometry force field |
| | Atomic mass 1000 Da, step size 50 fs, coupling constant 40 ps |
| | Cooling rate 1 K/100 steps |
| Stage 4 | 3000 steps conjugate gradient minimization ($k_{4D} = 10.0$) |
| | Distance geometry force field |
| Stage 5 | 500 steps conjugate gradient minimization |
| | AMBER force field with OONS[c] |

[a] Positional restraints on all the backbone atoms are imposed throughout the stages
[b] The weight of the fourth dimension
[c] For the calculations shown in Fig. 3B, OONS was not included

field, $\psi^{\text{react}} = \psi^{\text{sol}} - \psi^{\text{vac}}$, from which the self-energies and Coulomb interaction energies are eliminated. Finally, the electrostatic energy of the protein-solvent system ($E_{\text{p-s}}$) is given by:

$$E_{\text{p-s}} = \frac{1}{2}\sum_i q_i \psi^{\text{react}}(\mathbf{r}_i) + \sum_{i<j}{}' \frac{q_i q_j}{\epsilon_{\text{p}} |\mathbf{r}_i - \mathbf{r}_j|} \quad (2)$$

The second sum on the right-hand side of this equation is the Coulomb interaction energy, which is calculated only for 1–4 and 1–5 interacting pairs of atoms, as is done for usual calculations with the AMBER force field. In this case, no distance cut-off is applied for the Coulomb interaction term. During the development of this study we tested a few different values for the dielectric constant of the protein region, $\epsilon_{\text{p}}$. Using low values for $\epsilon_{\text{p}}$ (such as 2 or 4), the native structures did not always yield a lower electrostatic energy than misfolded models (data not shown). The best result was obtained when we set $\epsilon_{\text{p}} = 10$. Therefore, $\epsilon_{\text{p}}$ is set to 10 in the continuum electrostatic calculation shown later in this paper.

The charges and van der Waals radii of the protein atoms were taken from the values of the AMBER all-atom force field. Note that the continuum electrostatic energy is not used in minimization of the models, but it is simply applied to the models constructed by minimization with the AMBER force field with a distance-dependent dielectric constant.

The calculations of the continuum electrostatics were done on the VPP500. It took 5–10 min before a set of calculations was complete for each model.

## Results and discussion

### Discriminating the native structure from decoys

In order to test the energy function, we first calculated the energy values of a large number of decoys created by Park and Levitt (1996) for three proteins (1ctf, 1r69, and 3icb). All the decoys were energy minimized according to the procedure described above. Figure 1 shows the minimized energy values (AMBER + OONS) for the decoy sets. For all three cases, the native structure has an energy lower than any decoys. For 3icb, there is one decoy which has only 0.7 kcal/mol higher energy than the native. However, this structure is actually very close to the native one with a $C_\alpha$ RMSD of less than 2 Å. Therefore, the energy function in the present study is proved to be sufficiently good for use in discrimination of the native structure from misfolded structures.

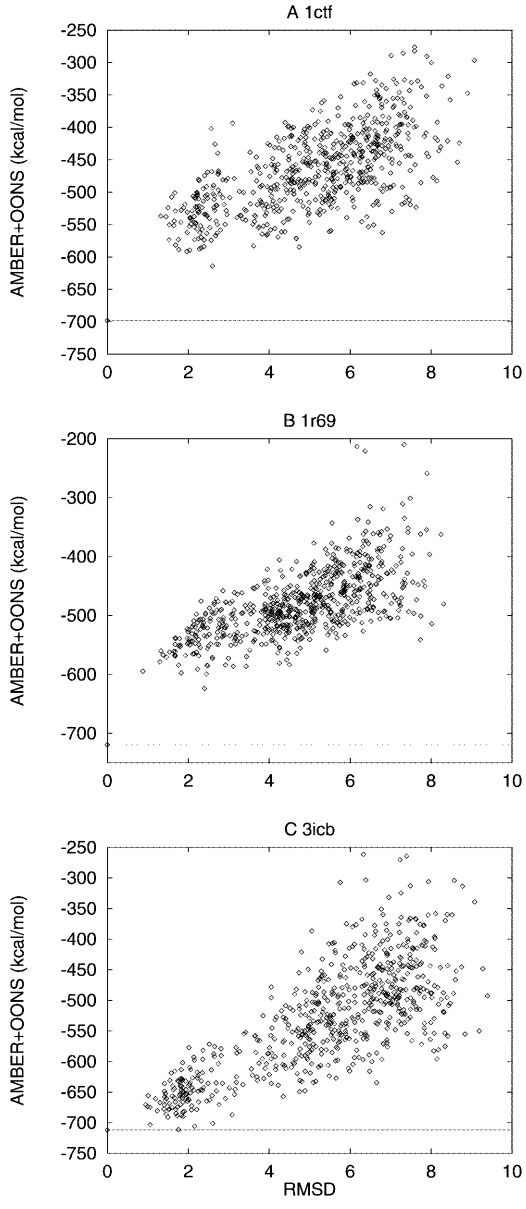


**Fig. 1A–C** Results for the Park and Levitt (1996) decoy set. The *horizontal axis* is the RMSD (Å) of the decoy from the native structure. The *vertical axis* is the minimized energy of the decoys. The *dashed line* indicates the energy of the native structure

### Evaluation of the structures selected by threading

The conventional ungapped threading search was performed for each target in Table 1 using the program S3 (Ota et al. 1999). The 10 structures giving the best 10 scores of S3 were selected as the template structures for

the target, and were subject to all-atom modeling and energy minimization by the method described above. In all the cases we tried, the native conformations were ranked at the top as expected (Jones and Thornton 1996). The selected structures are summarized in Table 2. All the templates are of protein-like structures in that they are compact and contain significant amounts of secondary structures. A model based on the native backbone structure itself is called the "near-native" model. Note that the side-chain conformations of the near-native models are completely reconstructed in the same manner as other models. Structural differences between the native and reconstructed near-native structures are shown in Table 4. For the target 1thx, two homologous structures (2trxA and 1tof) were incidentally found (Table 2). The alignment of 1thx and 2trxA is correct. We call the model "1thx-2trxA" (the model of the target "1thx" based on the template "2trxA") the "homologous" model. The alignment of 1thx and 1tof is partly incorrect (Fig. 2A); thus we call the model 1thx-1tof the "misaligned" model (Table 2). Based on the alignments by the threading, the sequence identity of the templates 2trxA and 1tof with the native sequence of 1thx is 42.6% and 17.6%, respectively. The backbone RMSD of the models 1thx-2trxA and 1thx-1tof from the native structure (1thx) is 1.2 Å and 3.7 Å, respectively. For the target 2hmzA, one homologous structure (2mhr) was found by threading and their alignment is partly incorrect (Fig. 2B) with the sequence identity of 33.6%; thus the model 2hmzA-2mhr is another misaligned model. Its backbone RMSD from the native (2hmzA) is 3.4 Å. The native, near-native, homologous, and misaligned models are defined to be "correct" models. Other models are of totally different fold from the native structure with RMSDs of more than 8 Å; thus they are called "misfolded" models and accordingly they are defined to be "incorrect" models (Table 2).

**Table 4** The difference between reconstructed near-native models and the native structures

| Model | RMSD (Å) | | $\chi$ correct (%)[c] | |
|---|---|---|---|---|
| | Backbone[a] | All[b] | $\chi^1$ | $\chi^2$ |
| 1lmb4-1lmb4 | 0.136 | 1.62 | 78.6 | 80.0 |
| 1molA-1molA | 0.116 | 1.90 | 76.5 | 72.7 |
| 1plc-1plc | 0.121 | 1.45 | 70.8 | 22.2 |
| 1rgeA-1rgeA | 0.104 | 1.51 | 82.4 | 83.3 |
| 1thx-1thx | 0.130 | 1.57 | 64.3 | 88.9 |
| 2hmzA-2hmzA | 0.117 | 2.01 | 65.5 | 79.2 |
| 3chy-3chy | 0.123 | 1.55 | 75.0 | 85.7 |
| Average | 0.121 | 1.66 | 73.3 | 73.1 |

[a] RMSD for backbone heavy atoms between the native and near-native model
[b] RMSD for all the heavy atoms between the native and near-native model
[c] The fraction of correctly predicted $\chi$ angles of buried residues. Residues are defined to be buried if the solvent-accessible surface area is less than 10% of the extended Gly-X-Gly conformation. $\chi$ angles with deviation from the native of less than 40° are considered to be correct

After the all-atom modeling and the energy minimization, the electrostatic energy based on the continuum dielectric model was calculated for each model. The results of the energy calculations are shown in Fig. 3A. In this figure, the total energy is defined as the sum of the AMBER energy without the Coulomb interaction terms, the OONS hydration free energy, and the electrostatic energy $E_{p-s}$ obtained by Eq. (2). Large energy gaps of more than 100 kcal/mol were found between the native structures and the near-native or any misfolded models. Although the near-native structures have much higher energies than the native, their energies are nevertheless lower than those of the misfolded ones (Fig. 3A). Even the misaligned models have fairly low energies compared to the other misfolded models. It is because their backbone topologies are similar to those of the native as mentioned above and in Fig. 2. The homologous model

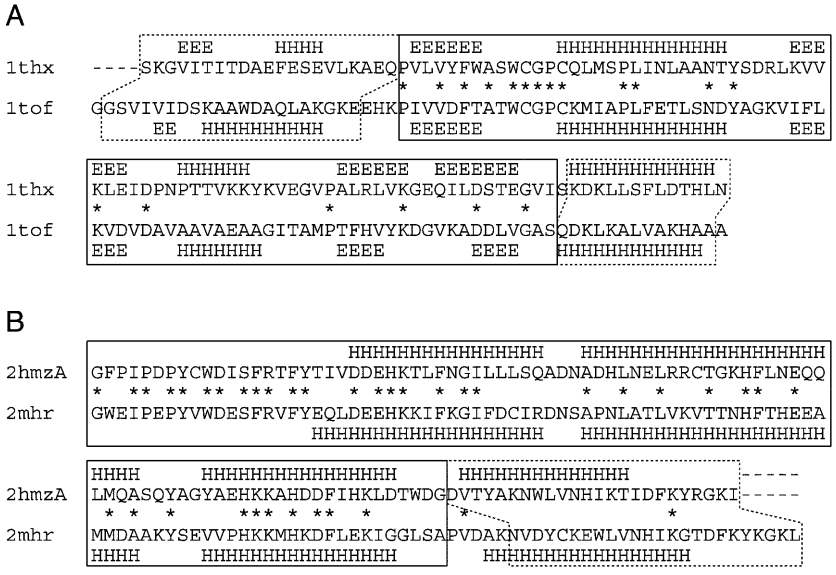


**Fig. 2A, B** The alignments of misaligned models obtained by ungapped threading. **A** The alignment for the model 1thx-1tof. **B** The alignment for the model 2hmzA-2mhr. The region bound by a *solid box* coincides with the correct alignment. The region bound by a *dotted box* indicates incorrectly aligned sites. The secondary structures ("H" for $\alpha$ helices, "E" for $\beta$ strands) are also shown. The alignment sites with identical residues are marked with *asterisks*

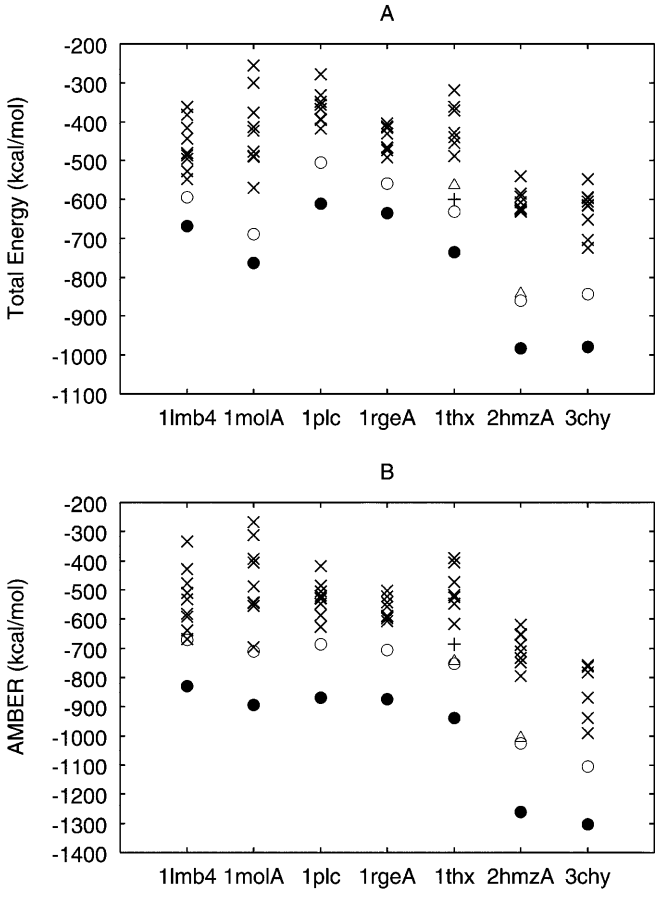


**Fig. 3A, B** Total energies of the native structures, correct and incorrect models. The *horizontal axis* indicates target proteins. The symbols are defined as follows: ●, native structures; ○, near-native models; +, homologous model; △, misaligned models; ×, misfolded models. See Table 2 for the nomenclature of the models. **A** The sum of the bond length, bond angle, torsion angle, improper torsion angle, 1–4 and 1–5 van der Waals, and hydrogen bond terms of the AMBER force field, the OONS hydration free energy, and the continuum electrostatic energy. **B** The AMBER force-field energy with the dielectric constant $\epsilon = 2r$

1thx-2trxA has higher energy than the near-native model 1thx-1thx, and lower energy than the misaligned model 1thx-1tof. This trend is reasonable, considering the structural similarity of these three models to the native structure of 1thx. This result shows that the present energy function is also able to discriminate the correct folds from incorrect ones. In the following section, we examine combinations of energy terms in search of physicochemical determinants of the native structure and the correct fold.

## Solvent effect

In order to see the effect of the solvent, we carried out the minimization without the hydration term (Fig. 3B). In this case, the compared energy is the AMBER force field with a distance-dependent dielectric constant ($\epsilon = 2r$). The native structures always have lower energies than any other near-native or misfolded structures. Although the near-native structures have lower energies than the misfolded structures, the energy difference is small (less than 15 kcal/mol) for targets such as 1lmb4 and 1molA. Also, the homologous model 1thx-2trxA has higher energy than the misaligned model 1thx-1tof, in contrast to the result for the total energy discussed above (Fig. 3A). It seems that the solvent effect is important for stabilizing near-native structures rather than the native structure. The solvent effects are further discussed in the following paragraphs.

The "hydrophobic interaction" is believed to be a dominant factor in stabilizing protein structures (Dill 1990). In the present study, the hydration effect is taken into account in terms of the implicit solvent model of Ooi et al. (1987). Figure 4A shows that the correct structures, including the native, do not necessarily have lower hydration free energies than incorrect structures. This trend was also observed by others (Wako 1989; Vorobjev et al. 1998; Lazaridis and Karplus 1999a). Makhatadze and Privalov (1995) showed that the main contributors to the hydrophobic interactions are van der Waals interactions in addition to the hydration effect. Therefore, we compared the sum of the van der Waals

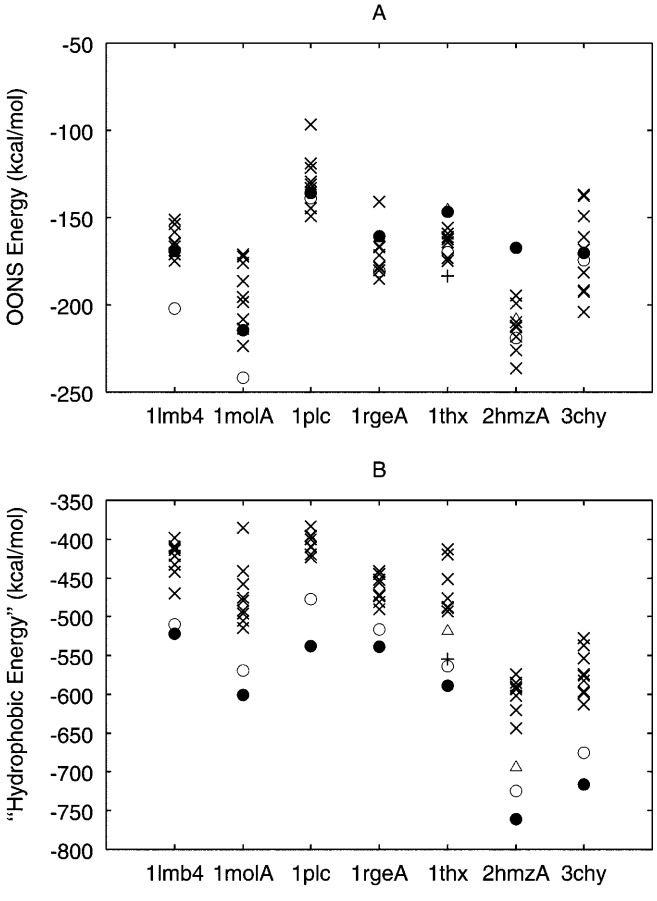


**Fig. 4** **A** The hydration free energy. **B** The "hydrophobic energy," which is the sum of the van der Waals energy and the hydration free energy. The symbols are defined in Fig. 3

energy and the hydration free energy as the "hydrophobic energy." Figure 4B shows that all the correct structures can be readily discriminated from their misfolded counterparts. The difference in the hydrophobic energy between the native structure and near-native model is in most cases less than 50 kcal/mol, which is small compared to the difference in the total energy (compare Fig. 4B with Fig. 3A). This fact suggests that the hydrophobic interaction stabilizes near-native structures more than the native structure itself.

The electrostatic energies $E_{p-s}$ based on Eq. (2) are shown in Fig. 5A. Except for the target 2hmzA, the native structure has the lowest electrostatic energy, and for most cases the near-native model has the next lowest electrostatic energy. For the target 2hmzA, there are buried glutamates which bind to irons in the experimental structure. Since these irons are ignored in the calculation, the electrostatic energy of the native structure of 2hmzA is not the lowest. This exception demonstrates the importance of the Born energy, which penalizes the buried charges. To see the importance of the solvent shielding and the Born energy, we also examined the Coulomb energy as calculated with a distance-dependent dielectric constant, $\epsilon = 2r$ (Fig. 5B). In this case, although all the native structures have the lowest Coulomb energy, the near-native models show the energy value close to, or even higher than, their misfolded counterparts. The stabilization by the solvent shielding and the Born energy, in other words, the contribution of the solvent to the electrostatic energy, seems more important for the near-native models than for the native structures.

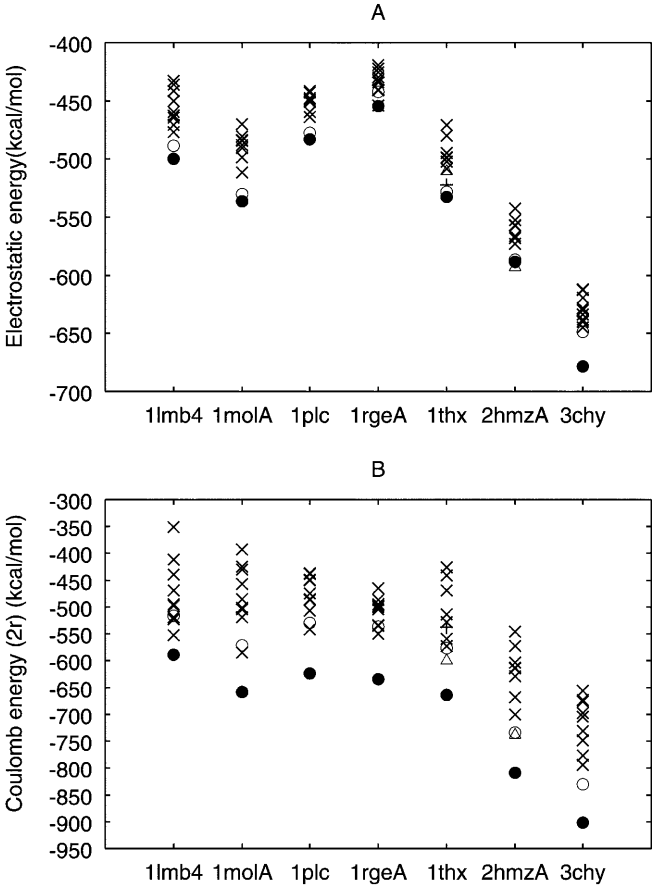


**Fig. 5** **A** The electrostatic energy calculated by solving the Poisson equation for the protein-solvent system (see the section Continuum electrostatics calculation). **B** The electrostatic energy calculated with a distance-dependent dielectric constant ($\epsilon = 2r$). The symbols are defined in Fig. 3

## Side-chain packing

Vorobjev et al. (1998) reported that the "packing energy," that is, the sum of the local geometry terms (bond lengths, bond angles, and torsion angles) and the van der Waals energy, was in favor of the native structure. Our results are consistent with their observation (Fig. 6A). The native structures have about 100 kcal/mol more stable packing energies than other models. This amount of stabilization is significant compared to the hydrophobic energy (Fig. 4B) and electrostatic energy (Fig. 5A). We also find that the near-native structures have lower packing energy than any other misfolded structures, but the energy difference is marginal in a few cases. While the native structures always have significantly low van der Waals energy, the van der Waals energy alone cannot necessarily discriminate near-native from misfolded structures (Fig. 6B). Therefore, local geometry and van der Waals energies are consistent only for the correct models, but this consistency alone is not enough to give the correct models significantly lower packing energy than the misfolded ones. In other words, the packing energy is a good index for discriminating the native structure, but it is not so for discriminating near-native structures. The native structure of a globular protein shows specific and close packing of the side chains (Richards and Lim 1993). The constructed near-native models have the side-chain conformations similar to the native structures (Table 4), but Figs. 3 and 6 show that the differences are significant from the energetics point of view. Note that the differences in van der Waals energy between the native structure and near-native models are about 50 kcal/mol, which is small compared to the difference in the packing energy. This shows that not only the close packing but also the less distorted local geometry contributes significantly to the stabilization of the native structure.

Janardhan and Vajda (1998) reported that the solvation term was important, but the molecular mechanics energy was useless for selecting near-native models with good packing. However, their molecular mechanics energy did not include the van der Waals term, and their electrostatic energy did not incorporate the solvent shielding and the Born energy. Therefore, their result for the molecular mechanics energy is more or less similar to Fig. 3B in our case. However, they used the atomic solvation parameters whose reference state was an organic solvent to complement the absence of the van der

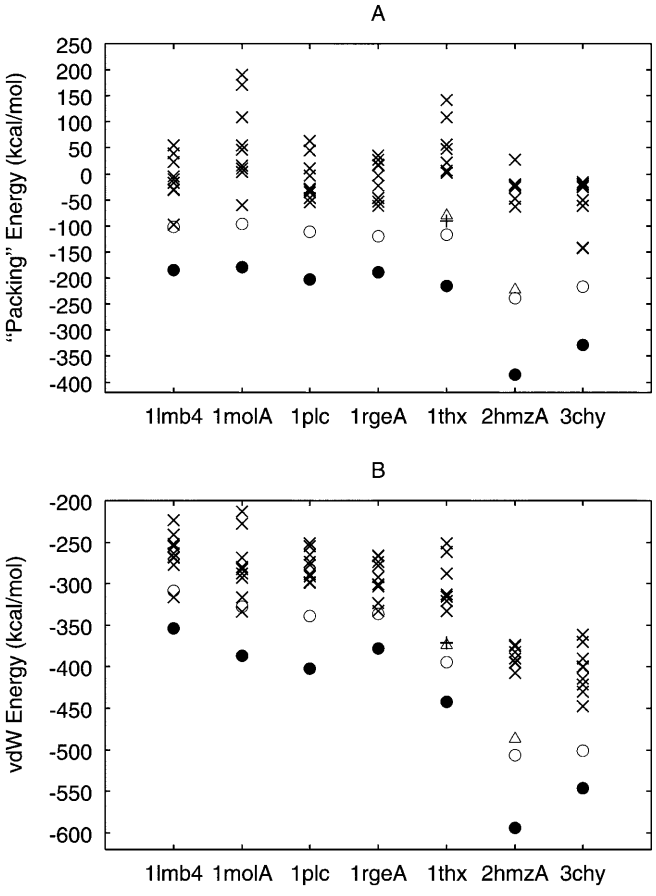


**Fig. 6** **A** The "packing energy." In our case, the packing energy is defined as the sum of the terms for bond lengths, bond angles, proper and improper torsion angles, and 1–4 and 1–5 van der Waals interactions. **B** The sum of 1–4 and 1–5 van der Waals energy terms. The symbols are defined in Fig. 3

Waals term, whereas the reference state of the parameters by Ooi et al. (1987) is the vacuum. Hence, their solvation free energy corresponds to the hydrophobic energy in our case (Fig. 4B). The results of Janardhan and Vajda (1998) are actually consistent with our observations.

Sahasrabudhe et al. (1998) applied a homology modeling method (Li et al. 1997) to model the structure of an RNA-binding protein using two kinds of template proteins, a ferredoxin-like fold as the correct fold and cold shock protein A as an incorrect fold. They found that the former had a plausible negative value of the energy, while the latter showed an unrealistically high value (Sahasrabudhe et al. 1998). In addition to the molecular mechanics energy in vacuum, they used structural restraints derived from both the backbone and side-chain conformations of the template. Consequently, the side chains of their incorrect model were forced into unrealistic conformations, hence the unrealistically high energy. In our case, no restraints were imposed on the side chains. Therefore, the side chains of the incorrect folds can adopt relatively relaxed conformations compared to the case of Sahasrabudhe et al. (1998). Nevertheless, our result (Fig. 6A) indicates that structures with good packing (i.e., the native structures) can be readily discriminated even in the absence of any artificial restraints for side-chain conformations.

## Implications for structure prediction

We have shown that stabilization by solvent effects such as "hydrophobic energy," solvent shielding, and the Born energy are good indexes for the discrimination of near-native models from misfolded ones. Also the homologous model 1thx-2trxA was found more stable than the misaligned model 1thx-1tof when solvent effects were included (Figs. 3A, 4B, and 5A). These results indicate that the solvent effects do not depend on structural details, but depend on more global features such as the pattern of hydrophobic and hydrophilic residues in the three-dimensional space. Hydrophobicity and the Born energy are effectively taken into account in the compatibility functions for threading (Nakamura 1996). This is one of the reasons for the success of threading in predicting approximately correct folds.

Although the hydrophobic energy can well discriminate the global fold of the native structure (Fig. 4B), its contribution is rather small compared to that from the packing energy (Fig. 6A). In fact, the large energy gap between the native structure and near-native model comes mainly from specific side-chain packing (Figs. 3A and 6A). Consequently, the lack of detailed treatment of side-chain packing may be the reason for the false positives often found by threading. In fact, some attempts have been made that incorporate more or less detailed representations of side-chain packing into statistical potentials and their results show significant improvements in recognizing the native folds (Matsuo et al. 1995; Samudrala and Moult 1998). However, since threading involves alignment of the target sequence to structures, exact modeling of side-chain conformations is in principle impossible. Also, it is difficult to model the side-chain conformations based on a distantly related template structure whose backbone structure differs to some extent from that of the native structure of the target protein (Chung and Subbiah 1996). Therefore, we cannot adopt any conventional algorithms for side-chain packing prediction which require the rigorously fixed backbone conformation (Levitt et al. 1997). Instead, we have to allow the backbone to move to an extent so that correct side-chain packing can be achieved. The present study treats the backbone conformation restrained to a given template, but allows it more or less to move. In order to solve the more general side-chain packing problem allowing backbone movement, and to reach the true native from a near-native structure, it will be necessary to employ powerful conformational sampling techniques such as generalized ensemble methods (e.g., Nakajima et al. 1997; Sugita and Okamoto 1999). The computation might be accomplished by sufficient and adequate conformational sampling in the optimization

process, as far as the native structure is located at the global minimum of the energy surface of the protein molecule. Attempts along this line are currently under way.

## Concluding remarks

For all of the seven protein sequences examined in the present study, the native conformation, minimized from the X-ray structure, was always the lowest in total energy among those conformations selected by threading. This fact, together with the test runs performed for decoy models (Fig. 1), validates the energy function employed. Misfolded conformations always have relatively higher energies, and correct models including the near-native and homologous models, were situated in the middle between the native and misfolded conformations (Fig. 3A). The near-native model, reconstructed according to the native backbone as the template, has an almost identical backbone conformation to the native (Table 4), but different side-chain conformations which were attained by minimizing atomic overlaps. Therefore, a distinct energy difference between the native and the near-native conformations (Fig. 3A) is mainly attributed to the difference in the side-chain conformations between them. However, the difference is little (Fig. 5A) or small (Fig. 4B) in electrostatic and hydrophobic energies, respectively, indicating that these energy terms are relatively insensitive to the side-chain conformation. On the other hand, the packing energy alone seems to yield the net change between the native and near-native structures (Fig. 6A), although the distinction between the near-native and misfolded models becomes somewhat less clear in the packing energy than in the total energy. Taking all these results into account, the energetic contributions to a native protein are summarized into two categories: one mainly depending on the topology or backbone conformation of a protein molecule (i.e., electrostatic and hydrophobic terms), and the other depending on the detailed side-chain conformation (i.e., packing energy). The importance of both contributions to the protein stability delineates the limit of the threading treatment in which the side chain is in principle simplified as a norm and therefore detailed side-chain packing must be totally neglected. Given an approximately correct topology (backbone conformation) alone, the next step for us is to realize the true native conformation for whole protein atoms. This would be one of the necessary steps toward ab initio predictions.

**Acknowledgements** We thank Drs. Takeshi Kawabata and Motonori Ota for extensive discussions, and Drs. Shoji Takada, Yuji Sugita, Steven E. Brenner, and Gaetano T. Montelione and Mr. Motoki Susa for helpful comments. We also thank Yuichi Kawanishi for programming assistance and Biomolecular Engineering Research Institute (Osaka, Japan) for providing the program EMBOSS. A.R.K. is a predoctoral research fellow of the Japan Society for the Promotion of Science. This work was partly supported by a grant-in-aid from the Ministry of Education, Science, Sports and Culture, Japan.

## References

Chung SY and Subbiah S (1996) A structural explanation for the twilight zone of protein sequence homology. Structure 4: 1123–1127

Clarke ND, Beamer LJ, Goldberg HR, Berkower C, Pabo CO (1991) The DNA binding arm of lambda repressor: critical contacts from a flexible region. Science 254: 267–270

Dill KA (1990) Dominant forces in protein folding. Biochemistry 29: 7133–7155

Guss JM, Bartunik HD, Freeman HC (1992) Accuracy and precision in protein crystal structure analysis: restrained least-squares refinement of the crystal structure of poplar plastocyanin at 1.33 angstroms resolution. Acta Crystallogr Sect B 48: 790–811

Havel TF (1991) An evaluation of computational strategies for use in the determination of protein structure from distance constraints obtained by nuclear magnetic resonance. Prog Biophys Mol Biol 56: 43–78

Holmes MA, Stenkamp RE (1991) Structures of met and azidomet hemerythrin at 1.66 Å resolution. J Mol Biol 220: 723–737

Janardhan A, Vajda S (1998) Selecting near-native conformations in homology modeling: the role of molecular mechanics and solvation terms. Protein Sci 7: 1772–1780

Jones DT, Thornton JM (1996) Potential energy functions for threading. Curr Opin Struct Biol 6: 210–216

Lazaridis T, Karplus M (1999a) Discrimination of the native from misfolded protein models with an energy function including implicit solvation. J Mol Biol 288: 477–487

Lazaridis T, Karplus M (1999b) Effective energy function for proteins in solution. Proteins 35: 133–152

Levitt M, Gerstein M, Huang E, Subbiah S, Tsai J (1997) Protein folding: the endgame. Annu Rev Biochem 66: 549–579

Li H, Tejero R, Monleon D, Bassolino Klimas D, Abate Shen C, Bruccoleri RE, Montelione GT (1997) Homology modeling using simulated annealing of restrained molecular dynamics and conformational search calculations with CONGEN: application in predicting the three-dimensional structure of murine homeodomain Msx-1. Protein Sci 6: 956–970

Makhatadze GI, Privalov PL (1995) Energetics of protein structure. Adv Protein Chem 47: 307–425

Matsuo Y, Nakamura H, Nishikawa K (1995) Detection of protein 3D-1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions. J Biochem (Tokyo) 118: 137–148

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540

Nakai T, Kidera A, Nakamura H (1993) Intrinsic nature of the three-dimensional structure of proteins as determined by distance geometry with good sampling properties. J Biomol NMR 3: 19–40

Nakajima N, Nakamura H, Kidera A (1997) Multicanonical ensemble generated by molecular dynamics simulation for enhanced conformational sampling of peptides. J Phys Chem B 101: 817–824

Nakamura H (1996) Roles of electrostatic interaction in proteins. Q Rev Biophys 29: 1–90

Novotny J, Bruccoleri R, Karplus M (1984) An analysis of incorrectly folded protein models – implications for structure predictions. J Mol Biol 177: 787–818

Ooi T, Oobatake M, Nemethy G, Scheraga HA (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. Proc Natl Acad Sci USA 84: 3086–3090

Ota M, Kawabata T, Kinjo AR, Nishikawa K (1999) Cooperative approach for the protein fold recognition. Proteins Suppl 3: 126–132

Park BH, Levitt M (1996) Energy functions that discriminate X-ray and near-native fold from well-constructed decoys. J Mol Biol 258: 367–392

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

Richmond TJ (1984) Solvent accessible surface area and excluded volume in proteins. Analytical equations for overlapping spheres and implications for the hydrophobic effect. J Mol Biol 178: 63–89

Saarinen M, Gleason FK, Eklund H (1995) Crystal structure of thioredoxin-2 from anabaena. Structure 3: 1097–1108

Sahasrabudhe PV, Tejero R, Kitao S, Furuichi Y, Montelione GT (1998) Homology modeling of an RNP domain from a human RNA-binding protein: homology-constrained energy optimization provides a criterion for distinguishing potential sequence alignments. Proteins 33: 558–566

Samudrala R, Moult J (1998) An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275: 895–916

Sevcik J, Zegers I, Wyns L, Dauter Z, Wilson KS (1993) Complex of ribonuclease Sa with a cyclic nucleotide and a proposed model for the reaction intermediate. Eur J Biochem 216: 301–305

Sippl MJ (1995) Knowledge-based potential for proteins. Curr Opin Struct Biol 5: 229–235

Sugita Y, Okamoto Y (1999) Replica-exchange molecular dynamics method for protein folding. Chem Phys Lett 314: 141–151

Tomic MT, Somoza JR, Wemmer DE, Park YW, Cho JM, Kim SH (1992) $^{1}$H resonance assignments, secondary structure and general topology of single-chain monellin in solution as determined by $^{1}$H-2D-NMR. J Biomol NMR 2: 557–572

Volz K, Matsumura P (1991) Crystal structure of *Escherichia coli* CheY refined at 1.7-Å resolution. J Biol Chem 266: 15511–15519

Vorobjev YN, Almagro JC, Hermans J (1998) Discrimination between native and intentionally misfolded conformations of proteins: ES/IS, a new method for calculating conformational free energy that uses both dynamics simulations with an explicit solvent and an implicit solvent continuum model. Proteins 32: 399–413

Wako H (1989) Monte Carlo simulations of a protein molecule with and without hydration energy calculated by the hydration-shell model. J Protein Chem 8: 733–747

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

Wesson L, Eisenberg D (1992) Atomic solvation parameters applied to molecular dynamics of proteins in solution. Protein Sci 1: 227–235