Protein Science (1998), 7:1772–1780. Cambridge University Press. Printed in the USA.

Selecting near-native conformations in homology modeling: The role of molecular mechanics and solvation terms

AJIT JANARDHAN AND SANDOR VAJDA

Department of Biomedical Engineering, Boston University, 44 Cummington St., Boston, Massachusetts 02215

(RECEIVED August 6, 1997; ACCEPTED April 17, 1998)

Abstract

A free energy function, combining molecular mechanics energy with empirical solvation and entropic terms, is used for ranking near-native conformations that occur in the conformational search steps of homology modeling, i.e., side-chain search and loop closure calculations. Correlations between the free energy and RMS deviation from the X-ray structure are established. It is shown that generally both molecular mechanics and solvation/entropic terms should be included in the potential. The identification of near-native backbone conformations is accomplished primarily by the molecular mechanics term that becomes the dominant contribution to the free energy if the backbone is even slightly strained, as frequently occurs in loop closure calculations. Both terms become equally important if a sufficiently accurate backbone conformation is found. Finally, the selection of the best side-chain positions for a fixed backbone is almost completely governed by the solvation term. The discriminatory power of the combined potential is demonstrated by evaluating the free energies of protein models submitted to the first meeting on Critical Assessment of techniques for protein Structure Prediction (CASP1), and comparing them to the free energies of the native conformations.

Keywords: free energy; loop closure; protein conformation; side-chain search

Identifying native and near-native folds among a set of conformations is an important step in a variety of applications involving conformational search. Here we focus on homology modeling that aims to build the structure of a target protein beginning with the coordinates of a homologue serving as the template. Homology modeling employs search algorithms to determine the structure of certain loop regions and to place nonconserved side chains. Accordingly, this paper deals with the problems of ranking near-native conformations that are generated by side-chain search and loop closure algorithms, or have been obtained from a template by various homology modeling procedures.

As shown by Novotny and co-workers (Novotny et al., 1988), molecular mechanics energy functions may be unable to distinguish between correct and misfolded conformations. While molecular mechanics is a useful tool for studying the effects of covalent bonding, excluded volumes, and coulombic electrostatics, it is inadequate for a thermodynamical description of stable, compact protein folds that may be heavily influenced by the nature of their solvent exposed surfaces (Vajda et al., 1997). A further disadvantage of molecular mechanics is its tendency to yield a rugged energy surface with countless local minima, resulting in an extreme sensitivity to small perturbations in the atomic coordinates. Due to these difficulties, molecular mechanics has been increasingly replaced by simplified, structure-based potentials (Vajda et al., 1997; Sippl, 1995). The main applications have been fold recognition (Godzik et al., 1992; Maiorov & Crippen, 1992; Bryant & Lawrence, 1993; Sippl, 1993), and ab initio folding of polypeptides or small proteins (Wilson & Doniach, 1989; Srinivasan & Rose, 1995; Jernigan & Bahar, 1996; Yue & Dill, 1996). However, there is no guarantee that any of these empirical potentials will be able to distinguish reasonably well between native and near-native protein folds (Huang et al., 1996; Park & Levitt, 1996).

The goal of this paper is twofold. First, we describe a free energy potential that expands a molecular mechanics energy function by empirical solvation and entropic terms, and is computationally efficient to be used in a variety of applications involving conformational search. The same potential has been extensively used for docking and binding free energy calculation (Vajda et al., 1994; Gulukota et al., 1996; King et al., 1996; Weng et al., 1996). Although the various terms of the free energy function are based on very different models, we have shown that they are consistent with each other and with thermodynamic data (Vajda et al., 1995; Weng et al., 1997). However, in docking and binding free energy calculation, it is frequently assumed that either both molecules are rigid, or the energy change due to flexible deformations is small relative to other contributions to the binding free energy (Novotny

Reprint requests to: Sandor Vajda, Department of Biomedical Engineering, Boston University, 44 Cummington St., Boston, Massachusetts 02215; e-mail: vajda@enga.bu.edu.

et al., 1989; Vajda et al., 1994; Jackson & Sternberg, 1995; Nauchitel et al., 1995; Verkhivker et al., 1995; Wallqvist et al., 1995). These assumptions clearly do not apply to homology modeling where the search is performed over a set of different conformations that may be heavily strained. In fact, as we will show, the internal energy terms that occur due to the deformations of the polypeptide geometry are generally very important, and can even dominate the free energy expression.

Our second goal is to demonstrate the usefulness of the empirical free energy potential in distinguishing native or near-native protein conformations from others that are less native-like. Tests involve ensembles of decoys generated by search algorithms that are part of homology modeling, i.e., side-chain search and loop closure. The discriminatory power of the potential is further demonstrated by evaluating the free energies of protein models submitted to the first meeting on Critical Assessment of techniques for protein Structure Prediction (CASP1), and comparing them to the free energies of the native conformations.

Empirical free energy functions

Here we describe the basic principles of free energy evaluation by empirical approaches, with more details given in Methods. The free energy difference, $\Delta G = G - G_o$, where G_o is the free energy in a reference conformation, is calculated by the expression

$$\Delta G = \Delta E + \Delta G_d - T\Delta S_c \quad (1)$$

where E, G_d, and S_c denote the molecular mechanics energy, the desolvation free energy, and the conformational entropy, respectively. The energy E is calculated by a molecular mechanics potential for the conditions of a reference medium, which can be either vacuum or an organic liquid. In the most general case, E includes van der Waals, electrostatic, and internal energy terms, $E = E_{vdW} + E_{elec} + E_{int}$, where the internal (bonded) energy E_{int} is the sum of bond stretching, angle bending, torsional, and improper terms, $E_{int} = E_{bond} + E_{angle} + E_{dihedral} + E_{improper}$. The desolvation free energy G_d is defined as the free energy of transferring the protein from water into the reference medium, and is based on the classical atomic solvation parameter (ASP) model (Eisenberg & McLachlan, 1986; Wesson & Eisenberg, 1992). The reference state is a folded conformation, and hence the difference in entropy, ΔS_c is restricted to side chains (Pickett & Sternberg, 1993).

Notice that using Equation 1 we calculate the free energy difference between two states rather than just the energy difference. Indeed, both E and G_d represent an entire ensemble of equienergetic structures, such as side-chain rotamers, rather than a single conformation. Each ensemble has some conformational entropy, resulting in the entropy change term $T\Delta S_c$. Furthermore, both E and G_d are implicitly averaged over an ensemble of water configurations. In particular, G_d includes both the energy and the entropy of desolvation.

Since the solvent is not modeled explicitly, the calculation of solvent–solute van der Waals (vdW) interactions requires approximations. The most straightforward strategy is to account for these interactions in the desolvation term ΔG_d, which can be accomplished by using vacuum as the reference medium (Wesson & Eisenberg, 1992; Abagyan & Totrov, 1994; Smith & Honig, 1994; Pellequer & Chen, 1997). The solute–solute van der Waals interactions are obtained by the usual 6–12 formula. However, as we will further discuss, this approach has substantial shortcomings in free energy calculations. In fact, since the solute–solvent and solute–solute vdW terms are based on very different models, the free energy function is very sensitive to small perturbations in the atomic coordinates, leading to a rugged free energy surface as in the case of molecular mechanics.

An alternative approach, frequently used in binding free energy calculations, is based on the approximation that the solute–solute and solute–solvent interfaces are equally well packed, and hence the van der Waals contacts lost between solvent and solute are balanced by new solute–solute contacts formed upon protein folding (Adamson, 1982; Novotny et al., 1989; Nicholls et al., 1991). Due to this cancellation both solute–solvent and solute–solute van der Waals terms can be excluded from the model. Within the framework of an atomic solvation parameter model, this can be easily accomplished by considering an organic liquid as the reference medium, because the free energy of transfer between two liquids includes only relatively small, differential van der Waals effects (Weng et al., 1997). The removal of the solute–solute van der Waals term reduces the molecular mechanics energy to

$$\Delta E = \Delta E_{elec} + \Delta E_{int} \quad (2)$$

resulting in a relatively "smooth" free energy function.

The disadvantage of the above approach is its insensitivity to steric clashes or cavities. To some degree, this problem can be overcome by van der Waals normalization prior to the free energy calculations. Van der Waals normalization means that all conformations are minimized for some number of steps, the structure with the lowest van der Waals energy is selected, and all the other structures are further minimized to attain the same van der Waals energy. The additional minimization removes steric clashes or cavities, and thereby assures that the conditions for assuming van der Waals cancellation are better satisfied. Notice that the removal of major packing errors generally increases the internal energy and hence the free energy.

Results

Analysis of decoys obtained by side-chain search

Side chains were removed from different surface regions of HPR, a phosphocarrier protein which was one of the homology modeling targets of the CASP1 meeting (Mosimann et al., 1995). Using a fixed backbone, a variety of new conformations were generated by a side-chain search algorithm (Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992). The resulting structures were subjected to van der Waals normalization, and the free energies were calculated by Equation 1. Since the side-chain conformational entropy depends only on the backbone conformation which is held invariant, we set $T\Delta S_c = 0$. The X-ray structure of the protein was used as the reference state. For the generated decoys, Figures 1–3 show the free energy difference ΔG, the conformational or molecular mechanics energy change ΔE, and the desolvation free energy difference ΔG_d, respectively, all as functions of the RMSD from the X-ray structure of the HPR protein.

The decoys used in this test have been generated in three different runs involving residues 1–12, 33–45, and 55–66, respectively, of the HPR protein. The side-chain search has been accomplished by using the ALL option of the CONGEN program.

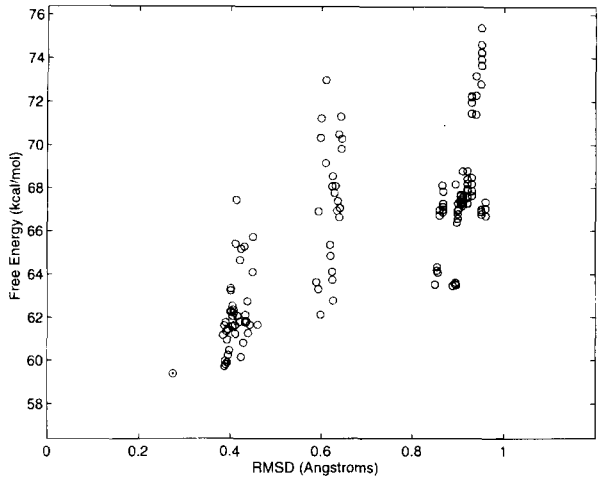

Fig. 1. Free energies (ΔG) of decoys generated by side-chain search.

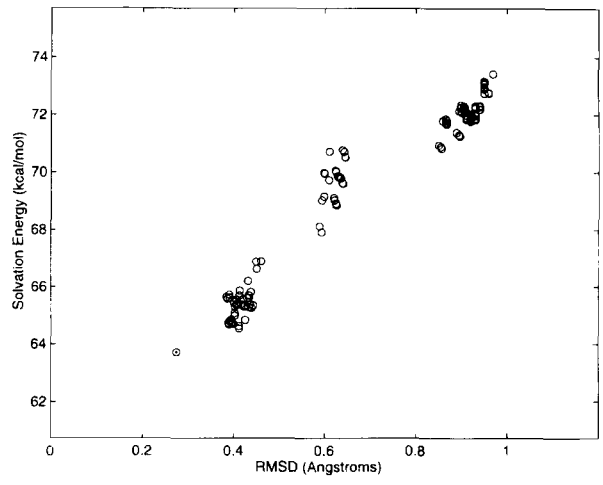

Fig. 3. Desolvation free energies (ΔG_d) of decoys generated by side-chain search.

Each CONGEN search yields a long list of conformations, ranked according to their CONGEN energies, accounting for the internal energy of the side chains and their interactions with the rest of the protein (Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992). From each list the 100 lowest energy structures were selected for van der Waals normalization. Free energies are shown for fewer decoys, however, because the van der Waals energies of some structures could not be reduced to the common value in the van der Waals normalization. Figure 1 also shows the free energy of the X-ray structure of the HPR protein after 200 steps of minimization (circle with a dot). Notice that the RMSD of this conformation slightly differs from zero due to the minimization.

Analysis of decoys obtained by loop closure

The analysis exploits the library of decoys that has been generated by J. Moult and co-workers for a number of short loops by extensive conformational searches (Moult & James, 1986; Fidelis et al., 1994). Results are shown for loop 132–136 from the serine proteinase 2sga and loop 120–124 from the dihydrofolate reductase 3dfr. Prior to free energy evaluation, all conformations taken from the decoy library were subjected to 200 steps of minimization. As will be described in Discussion, in evaluating loop decoys the van der Waals normalization can be replaced by simple minimization with little effect on the results. Figures 4–6 show the free energy difference ΔG, the molecular mechanics energy change ΔE, and the desolvation free energy difference ΔG_d, respectively, for loop 132–136 from 2sga. The same quantities for loop 120–124 from 3dfr are shown in Figures 7–9.

Evaluation of predicted structures from CASP1

Prior to the CASP1 meeting, a number of research groups predicted structures for seven proteins provided as homology modeling targets (Mosimann et al., 1995). Here we calculate the free energy of each prediction for the four proteins with the highest

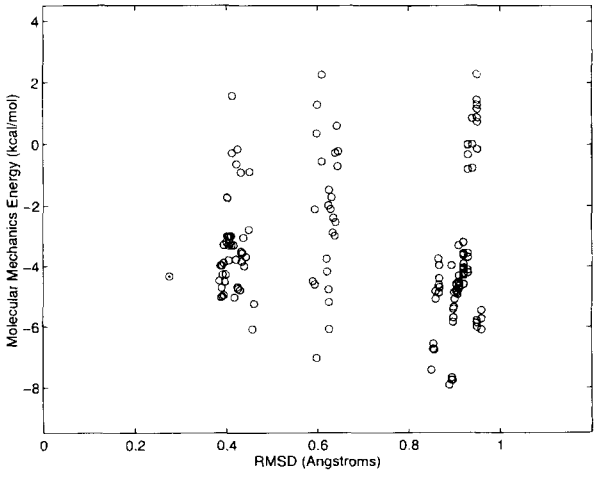

Fig. 2. Molecular mechanics energies (ΔE) of decoys generated by side-chain search.

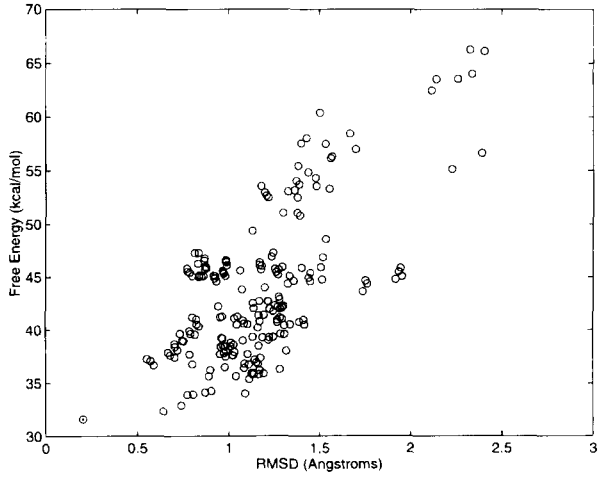

Fig. 4. Free energies (ΔG) of decoys generated for loop 132^{Ala}–136^{Asp} of 2sga.

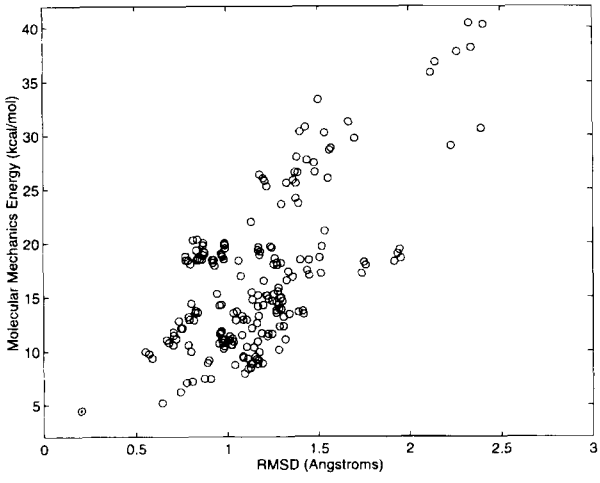

Fig. 5. Molecular mechanics energies (ΔE) of decoys generated for loop 132^{Ala}–136^{Asp} of 2sga.

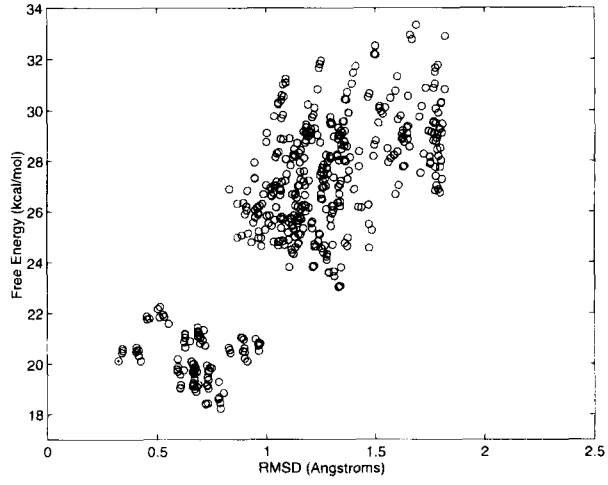

Fig. 7. Free energies (ΔG) of decoys generated for loop 120^{Gly}–124^{Gly} of 3dfr.

number of predictions submitted, and compare it to the free energy of the native conformation. Table 1 shows the four targets and the best templates available in the Protein Data Bank at the time of the prediction contest, along with relevant data illustrating the degree of difficulty associated with the modeling of each protein. The first and second columns give the target name and template pdb code, respectively, with the resolution of the crystal structure in parenthesis. The third and fourth columns show the sequence identity and number of gaps, respectively, between the target and template proteins. The remaining columns show the spread of RMSD values attained by the predictors, and the mean RMSD, calculated for both backbone and all heavy atoms.

Tables 2–5 show the results of free energy calculations performed on the predictions submitted. Some predictions have been left out due to errors in sequence or missing coordinates. For each protein we van der Waals normalized the submissions, allowing us to remove the van der Waals terms from the potential. In all tables we list the electrostatic energy ΔE_{elec}, the desolvation free energy ΔG_d, the internal energy ΔE_{int} defined as the sum of bond length, bond angle, dihedral, and improper energy terms, and the molecular mechanics energy $\Delta E = \Delta E_{int} + \Delta E_{elec}$. The next column, $\Delta \tilde{G}$, is defined as the sum of nonbonded terms only, $\Delta \tilde{G} = \Delta E_{elec} + \Delta G_d$, and is included to demonstrate the failure of a free energy function that lacks the internal energy term. The last columns in each table, ΔG, is the free energy function that includes both bonded and nonbonded energy terms, $\Delta G = \Delta E_{int} + \Delta E_{elec} + \Delta G_d$. The reasons for omitting the side-chain conformational entropy change term, $T\Delta S_c$ will be discussed in Methods.

Discussion

Analysis of decoys obtained by side-chain search

The three distinguishable clusters of points in Figure 1 correspond to three separate runs involving the side chains of residues 1–12, 33–45, and 55–66. The separate circle with a dot represents the

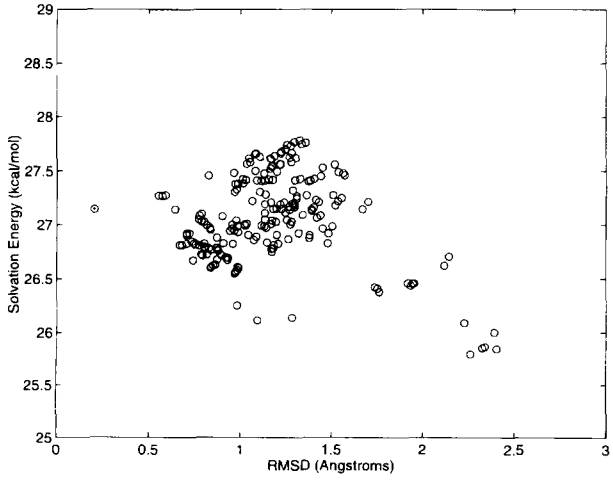

Fig. 6. Desolvation free energies (ΔG_d) of decoys generated for loop 132^{Ala}–136^{Asp} of 2sga.

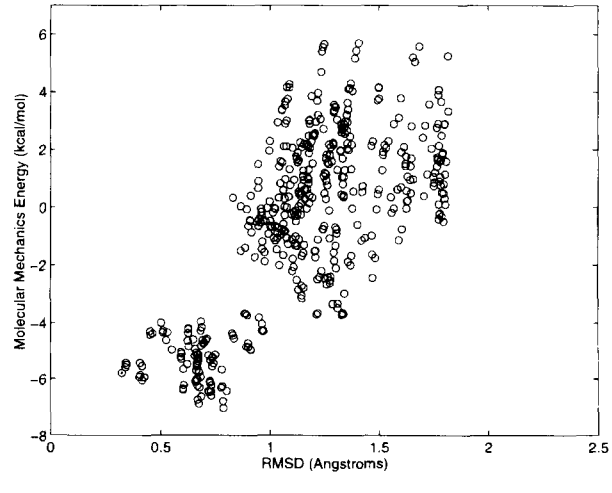

Fig. 8. Molecular mechanics energies (ΔE) of decoys generated for loop 120^{Gly}–124^{Gly} of 3dfr.

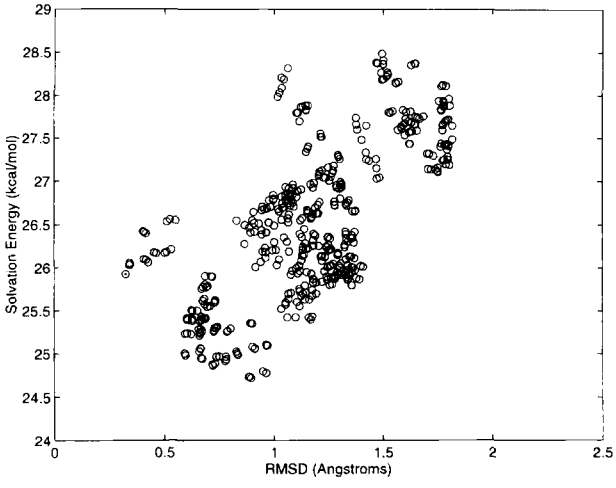

Fig. 9. Desolvation free energies (ΔG_d) of decoys generated for loop 120^{Gly}–124^{Gly} of 3dfr.

minimized X-ray structure of the HPR protein. In each run, all conformations generated are in a narrow RMSD range, suggesting that CONGEN is not an ideal device for side-chain search. Notice, however, that only the 100 lowest energy structures have been retained from each CONGEN calculation, and there exist many further structures that have higher RMSD. Although retaining only 100 structures gives a relatively poor sampling within each run, the three runs together show that selecting lower free energy structures generally yields lower RMSD. Furthermore, the minimized X-ray structure has the lowest free energy.

It is well known that solvation effects are important to guide side-chain placement (Schiffer et al., 1993; Johnson et al., 1994). According to the relationship between the molecular mechanics energy and RMSD, shown in Figure 2, we can even conclude that molecular mechanics on its own is practically useless in this problem. Although the variation in RMSD is limited among the 100 lowest energy structures selected from each run, the value of the molecular mechanics energy varies by as much as 10 kcal/mol. What is even worse, energies substantially lower than that of the native structure can be attained. By contrast, the solvation term (ΔG_d) shown in Figure 3 ranks the generated conformations almost ideally, both among and even within the clusters.

As shown in Figure 1, the well-known shortcoming of purely molecular mechanics potentials (Novotny et al., 1988) can be alleviated by adding solvation. However, it appears that with a fixed backbone, after major clashes have been eliminated by energy minimization, the best criterion for selecting native and near-native side-chain conformations is the solvation free energy, because the molecular mechanics energy does not provide any structural information (see Figure 2).

Analysis of decoys obtained by loop closure

In loop calculations both backbone and side chains vary within a short fragment. As shown in Figure 4 for loop 132–136 of 2sga, there exists a correlation between free energy and RMSD, and hence selecting low free energy structures one can identify near-native conformations among such decoys. As shown in Figure 5, the contribution of the molecular mechanics term is much more important than in the case of side-chain search. In fact, the overall free energy function closely follows the molecular mechanics energy. Compared to the variation of almost 40 kcal/mol in the molecular mechanics energy, the variation in the solvation energy, shown in Figure 6 is negligible (2.5 kcal/mol). The entropic contribution, $T\Delta S_c$, is even smaller (not shown separately). Thus, the discriminant is the molecular mechanics energy. The two components of this term, i.e., internal and electrostatic energies, behave very similarly to each other, with larger variation in the internal energy term (30 kcal/mol vs. 10 kcal/mol).

Region 120–124 of 3dfr includes a loop that is exposed to solvent to a greater extent than loop 132–136 of 2sga, which is largely buried. Therefore we expect the solvation effect, shown in Figure 9, to contribute more to the variation in the free energy than in the previous case. Indeed, for this loop the molecular mechanics and solvation energies are on about the same scale, although the molecular mechanics term shown in Figure 8 remains the more important of the two contributions. The free energy function is a relatively good predictor of the RMSD, with the exception of a low energy cluster centered at 0.7 Å. The same low energy cluster seen in both Figures 8 and 9, and thus both components of the free energy function identify this cluster as energetically more favorable than the native. The conformations in this cluster primarily differ from the native in the orientation of the Glu-123 side chain, which, with an average *B*-factor of 94.2, is essentially undefined in the X-ray structure. As in the case of the 2sga loop, the internal energy alone would be a good predictor of the RMSD, although it has somewhat smaller variation (9 kcal/mol) than the total molecular mechanics term.

We have previously used a free energy potential that included all terms of Equation 1 but the internal energy. In particular, we cal-

Table 1. *Comparative modeling cases analyzed*[a,b]

Target (resolution)	Template (resolution)	Sequence identity (%)	No. of gaps[c]	Backbone RMSD		All atom RMSD	
				Range	Mean	Range	Mean
EDN (2.2)	6rsa (2.0)	39.7	4	3.7–5.3	4.6	4.9–6.4	5.8
CRABP I (2.7)	2hmb (2.1)	43.1	3	2.0–3.7	2.7	2.6–4.4	3.4
HPR (2.0)	1pch (1.8)	41.4	0	1.0–4.1	1.5	1.7–4.3	2.1
NM23 (2.0)	1ndl (2.4)	77.5	0	0.4–2.0	1.2	1.3–3.1	2.3

[a]From the CASP1 meeting.
[b]Resolution and RMSD values are given in Å.
[c]Based on results of sequence alignment using GCG software.

Table 2. *Predictions for CRABP I*[a]

Group	ΔE_{elec}[b]	ΔG_d[c]	ΔE_{int}[d]	ΔE[e]	$\Delta \tilde{G}$[f]	ΔG[g]
Abagyan	26.8	3.3	32.3	59.1	30.0	62.4
Moult 1	39.4	*−0.2*	32.3	71.7	39.3	71.6
Moult 2	27.5	2.1	22.8	50.4	29.6	52.5
Sali	45.1	*−3.2*	10.5	55.5	41.9	52.4
Vinals 1	*−12.4*	10.9	53.4	41.0	*−1.4*	52.0
Vinals 2	66.9	*−3.8*	78.5	145.4	63.1	141.6
Vinals 3	64.5	*−0.3*	89.1	153.6	64.2	153.3
Vriend	52.7	*−9.1*	12.8	65.5	43.6	56.3
Weber 1	*−8.6*	11.0	74.7	66.1	2.4	77.0
Weber 2	*−3.8*	8.8	74.7	70.9	5.0	79.7

[a]Energy values relative to that of the native structure.
[b]Electrostatic energy.
[c]Desolvation free energy.
[d]$\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper}$.
[e]$\Delta E = \Delta E_{int} + \Delta E_{elec}$.
[f]$\Delta \tilde{G} = \Delta E_{elec} + \Delta G_d$.
[g]$\Delta G = \Delta E_{elec} + \Delta E_{int} + \Delta G_d$.

culated the free energies of protein unfolding and showed that the results are in good agreement with the experimentally determined values (Weng et al., 1997), suggesting that both unfolded and native folded protein conformations are free of significant strains that would affect the folding free energy. By contrast, most loop conformations in the decoy library are heavily strained even after minimization. This should not come as a surprise since the local minimization methods applied to each loop conformation from the decoy library are expected to find only local energy minima, and thus are generally unable to proceed to a fully relaxed conformation of the molecule. The fact that we find very interesting and useful is that the molecular mechanics energy values at these local minima correlate with the RMSD from the native structure as shown in Figures 5 and 8, and hence can be used to select the least distorted states.

Table 3. *Predictions for EDN*[a]

Group	ΔE_{elec}[b]	ΔG_d[c]	ΔE_{int}[d]	ΔE[e]	$\Delta \tilde{G}$[f]	ΔG[g]
Biosym	48.8	13.7	50.4	99.2	62.4	112.9
Koehl	54.5	16.1	24.1	78.7	70.6	94.7
Moult	45.3	2.1	46.2	91.5	47.4	93.6
Sali 1	45.6	8.2	27.5	73.1	53.8	81.3
Sali 2	51.6	2.5	29.4	81.0	54.1	83.5
Saqi 1	31.9	8.2	52.3	84.2	40.1	92.4
Vinals 1	35.8	2.4	102.3	138.2	38.2	140.5
Vinals 2	14.6	6.4	131.1	145.7	20.9	152.0
Vinals 3	6.2	6.6	61.1	67.3	12.9	73.9
Weber	32.7	0.4	112.8	145.5	33.0	145.9

[a]Energy values relative to that of the native structure.
[b]Electrostatic energy.
[c]Desolvation free energy.
[d]$\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper}$.
[e]$\Delta E = \Delta E_{int} + \Delta E_{elec}$.
[f]$\Delta \tilde{G} = \Delta E_{elec} + \Delta G_d$.
[g]$\Delta G = \Delta E_{elec} + \Delta E_{int} + \Delta G_d$.

Table 4. *Predictions for HPR*[a]

Group	ΔE_{elec}[b]	ΔG_d[c]	ΔE_{int}[d]	ΔE[e]	$\Delta \tilde{G}$[f]	ΔG[g]
Abagyan	23.8	*−3.1*	*−2.0*	21.8	20.7	18.7
Biosym	8.0	*−2.9*	4.7	12.7	5.1	9.8
Koehl 1	33.1	*−3.1*	1.8	34.9	30.0	31.8
Koehl 2	27.8	2.6	3.1	30.9	30.4	33.5
Mosenkis	35.1	*−5.7*	1.8	36.9	29.4	31.2
Moult	20.1	*−3.3*	1.8	21.9	16.8	18.6
Vriend	35.3	3.2	4.0	39.3	38.5	42.4
Weber	24.2	4.5	13.0	37.2	28.8	41.7

[a]Energy values relative to that of the native structure.
[b]Electrostatic energy.
[c]Desolvation free energy.
[d]$\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper}$.
[e]$\Delta E = \Delta E_{int} + \Delta E_{elec}$.
[f]$\Delta \tilde{G} = \Delta E_{elec} + \Delta G_d$.
[g]$\Delta G = \Delta E_{elec} + \Delta E_{int} + \Delta G_d$.

The above observation may also explain why the van der Waals normalization can be replaced by simple minimization in loop closure problems. Recall that we introduced van der Waals normalization because the free energy function does not include vdW terms, and hence packing errors (e.g., atomic overlaps) could go unnoticed. However, if there is an overlap of backbone atoms or any distortion in the backbone geometry, the minimization will substantially increase the molecular mechanics energy term and thereby the free energy.

Evaluation of predicted structures from CASP1

CRABP, EDN, and HPR constitute moderately difficult homology modeling problems, with sequence identities around 40% between template and target, while the task of modeling NM23 is trivial, with 77.5% sequence identity. Notice, however, that the average backbone RMS deviation is nearly identical for HPR (with 41.4% identity) and NM23 (with 77.5% identity), while CRABP I and EDN predictions have deviations that are nearly twice as large. The reason for this is that there are no gaps present in the alignments of HPR and NM23 with their templates, while those of

Table 5. *Predictions for NM23*[a]

Group	ΔE_{elec}[b]	ΔG_d[c]	ΔE_{int}[d]	ΔE[e]	$\Delta \tilde{G}$[f]	ΔG[g]
Koehl	*−16.7*	*−2.4*	26.9	10.2	*−19.1*	7.8
Sali	*−0.9*	3.1	0.9	*0.0*	2.2	3.1
Vihinen	40.6	*−1.5*	11.5	52.1	39.0	50.6
Vriend	11.2	*−6.4*	39.4	50.6	4.8	44.3
Weber 1	*−1.2*	7.3	18.7	17.5	6.1	24.8
Weber 2	0.1	6.9	27.5	27.6	7.0	34.5

[a]Energy values relative to that of the native structure.
[b]Electrostatic energy.
[c]Desolvation free energy.
[d]$\Delta E_{int} = \Delta E_{bond} + \Delta E_{angle} + \Delta E_{dihedral} + \Delta E_{improper}$.
[e]$\Delta E = \Delta E_{int} + \Delta E_{elec}$.
[f]$\Delta \tilde{G} = \Delta E_{elec} + \Delta G_d$
[g]$\Delta G = \Delta E_{elec} + \Delta E_{int} + \Delta G_d$.

CRABP and EDN contain several. Thus, the backbones of HPR and NM23 are fully defined by their templates. Furthermore, the *all atom* RMSD between HPR and its template is noticeably smaller than those of CRABP and EDN, even though all three share similar sequence identity with their respective templates (around 40%). Clearly then, the most important step in homology modeling is backbone coordinate determination, as the accuracy of the backbone coordinates limits the accuracy of side-chain placement.

Results in Tables 2–5 show that the empirical free energy function ΔG defined by Equation 1 discriminates the X-ray structure from the conformations predicted by homology modeling in all four problems, often by a considerable margin. As in the case of loop closure, the molecular mechanics energy is generally much larger than the desolvation term and, apart from a single case, discriminates the native structures on its own. The exception is the structure by Sali in Table 5, which has the same molecular mechanics energy as the X-ray structure of the target. However, we can accept that this prediction with the α-carbon RMSD of 0.43 Å is indistinguishable from the native structure, since such RMSD can be seen between two X-ray structures of the same protein.

Notice that there is a strong correlation between the backbone RMSD and the internal energy component of the molecular mechanics energy. As we pointed out, the alignments of CRABP and EDN (Tables 2 and 3, respectively) contained gaps, resulting in high deviations between predicted and actual backbone coordinates. For these two proteins the internal energy, ΔE_{int}, ranks the native fold as the lowest energy structure by far. On the other hand, the alignments of HPR and NM23 were gap free, and thus the backbone coordinates of the target are nearly identical to those of the template, leading to relatively small backbone RMSD. Tables 4 and 5, of HPR and NM23, respectively, reveal that for these proteins, the internal energy of the predicted structures is about the same and sometimes even lower than that of the native. Since the internal energy depends mainly on the backbone, the necessary similarity of the backbone structures in target and template proteins suggests a simple test. If the internal energy of a predicted target conformation is much higher than the internal energy of the template, then the prediction is likely to have a distorted backbone. Since such distortions may occur due to erroneous alignment, the simple test can be very useful.

According to our results, the molecular mechanics energy discriminates the native structure among the predictions, and generally dominates the free energy expression unless the backbone conformation is very close to the native, such as in the prediction by Sali in Table 5. It is interesting that the desolvation term attempts to compensate for large changes in the molecular mechanics term. For example, three predictions in Table 2 (Vinals 1, Weber 1, and Weber 2) have lower electrostatic energies than the native. It is very likely that these models have been heavily minimized using a molecular mechanics energy function, and such minimizations frequently yield very low electrostatic and van der Waals energies. However, as seen in Table 2, the same three models have the highest values of the desolvation free energy ΔG_d. However, for the Vinal 1 prediction the compensation is not strong enough, and the conformation is distinguished from the X-ray due to its much higher internal energy.

During the last few years a large variety of structure-based and hydrophobic potentials have been developed that do not include internal or molecular mechanics energy terms (Vajda et al., 1997). The primary application of these potentials is threading and ab initio simulation of small proteins. Since threading uses "strain-free" backbone structures observed in proteins, the potentials can perform reasonably well. Similarly, in folding studies it is meaningful to assume that local strains are removed on a much shorter time scale than that of the folding itself, and hence one can restrain consideration to terms representing desolvation and hydrogen bonding. By contrast, our results show that in homology modeling the selection of native and near-native conformations generally requires including molecular mechanics energy terms in the potential. To emphasize this observation, in Tables 2–5 we list the values of a nonbonded free energy potential, $\Delta\tilde{G}$, that does not include the internal energy. While $\Delta\tilde{G}$ was shown to provide an adequate tool for calculating the binding free energy in receptor-ligand complexes (Vajda et al., 1994; Gulukota et al., 1996; King et al., 1996; Weng et al., 1996), in homology modeling it fails to discriminate the native structure in two of the four cases (see Tables 2, 4).

We have found that for each homology modeling target, the free energies of the CASP1 predictions were higher than those of the X-ray structure. However, we were unable to find a correlation between free energy and RMSD, although such correlations were easily identifiable in side-chain search and loop closure. We think that comparing entire models built by homology modeling is more difficult than comparing simple decoys, all generated by the same method. In fact, homology modeling requires not only loop closure and side-chain search, but also the alignment of target and template sequences, the selection of loop regions to be built, and the refinement of the derived structures by some type of energy minimization. The various groups used very different assumptions in these steps when working on their submissions to CASP1, resulting in predictions that differ from each other in a variety of ways. Since the differences affect many variables, we can regard these predictions as points defined in a very high dimensional conformational space. For each target, the set of submitted predictions can be regarded as a sample of ten or fewer points, which is clearly too small for the analysis of possible interactions.

Conclusions

In the eighties Novotny and coworkers (Novotny et al., 1988) made the protein community accept that molecular mechanics alone is unable to distinguish between correct and misfolded protein conformations, and one has to add measures of solvation and possibly entropic effects. It is interesting that now we may have moved too far away from molecular mechanics. During the last few years, simplified potentials have been increasingly used in protein modeling, primarily for threading and folding simulations. These potentials attempt to represent the main factors that are known to contribute to the stability of folded proteins, i.e., hydrophobic interactions and hydrogen bonding, and generally do not include internal energy terms (Vajda et al., 1997).

The results of the present paper indicate that, apart from side-chain search with a fixed backbone, for ranking near-native conformations that occur in homology modeling it is vital to include the internal energy term, in addition to free energy contributions representing electrostatic, solvation, and entropic effects. While on its own the internal energy may not be able to discriminate between native and non-native folds, its addition to the other terms drastically improves the discriminating power of the free energy function.

Although protein folding is governed by hydrophobic interactions, hydrogen bonding, and the loss of conformational entropy, these quantities are dwarfed in relation to the internal energy over

large fractions of the conformational space. In principle, one can reduce the internal energy of the molecule by relaxing all deformations of the polypeptide geometry, and then use a simplified, empirical potential. However, the methods routinely used in conformational searches are unable to accomplish this, and most trajectories end up in local minima that are still in the region of high internal energy. For example, although the models submitted to CASP1 have been refined by their authors and further minimized by us, the internal energy dominates the free energy in most cases.

Although we emphasized that the molecular mechanics terms cannot be omitted, ranking near-native conformation requires the use of a complete free energy function that also includes solvation and entropic terms. At the beginning of a conformational search, the backbone conformations are usually strained, and the internal energy dwarfs all other free energy terms. However, selecting conformations with low internal energy moves the search into regions of the conformational space where the molecular mechanics and desolvation terms are on the same scale. Finally, after a backbone conformation is accepted, the all-atom RMSD can be further reduced by a side-chain search governed by the desolvation free energy alone. The importance of the molecular mechanics energy has recently been emphasized in a study focusing on loop closure (Pellequer & Chen, 1997), but without noticing the increasing role of solvation as backbone strains are reduced.

Methods

Free energy calculation

The energy change ΔE in Equation 1 was calculated using version 19 of the CHARMm force field (Brooks et al., 1983) with a distance-dependent dielectric coefficient $\epsilon = 4r$, and nonbonded cutoff 17 Å. Only polar hydrogens were used. To refine the proteins before free energy evaluation (i.e., to remove van der Waals clashes or substantial deformations of geometry), we performed either 200 steps of minimization using the CHARMm potential, or applied the *van der Waals normalization* procedure, in which the minimization was carried out until the van der Waals energies of different conformations were within 1 kcal/mol of each other.

The desolvation free energy ΔG_d is based on the atomic solvation parameter (ASP) model

$$G_d = \Sigma A_i \sigma_i$$

where A_i denotes the solvent accessible surface area of the ith atomic group and σ_i is the corresponding atomic parameter (Eisenberg & McLachlan, 1986), obtained from octanol-to-water transfer free energies (Vajda et al., 1994).

The calculation of the side-chain conformational entropy loss ΔS_c is based on an empirical entropy scale (Pickett & Sternberg, 1993) in which the maximum conformational entropy S_c of each side chain was calculated by the classical expression $S_c = -R \times \Sigma_i p_i \ln(p_i)$, where p_i denotes the probability of the ith rotamer. In the free energy calculation we assume that the entire side-chain entropy is lost, i.e., $\Delta S_c = S_c$, if the change ΔA_t in the total solvent accessible surface area of the side chain is more than 60% of its standard side-chain surface area A_t^*(Shrake & Rupley, 1973). Otherwise the entropy loss is scaled according to $\Delta S_c = \alpha S_c$, where $\alpha = \Delta A_t/(0.6A_t^*)$.

We have shown that the above entropy scale agrees very well with side-chain entropies based on calorimetric observation of temperature-induced protein unfolding (Weng et al., 1997). The method has been extensively used for calculating the loss of entropy of the side chains that become part of the receptor–ligand interface upon protein–protein association. However, the approach is less appropriate for calculating the side-chain entropy difference between two folded conformations. In fact, although the side-chain entropy should depend only on the backbone conformation, we use the change in the solvent exposed area of a side chain to assess if it is exposed and hence has its entropy, or becomes buried and hence loses it. Since side-chain positions can vary even with a fixed backbone, the method has an inherent error. In addition, in homology modeling the backbones of the template and the target must be similar, and hence the difference in side-chain entropy is relatively small. This difference may be comparable in magnitude to the inherent error of the method, and hence in some applications it may be more appropriate to ignore the change in side-chain entropy. In particular, we did not include the conformational entropy change term, $T\Delta S_c$, when evaluating the free energy of CASP1 submissions. While side-chain entropy loss could be calculated by a number of more accurate methods requiring simulation or iterative determination of self-consistent rotamer probabilities, the small difference in the side-chain entropy between two similar conformations does not justify the use of these computationally more demanding methods.

Decoy generation

As we described, side-chain decoys were generated by fixing the backbone coordinates, removing all side chains from certain surface regions of the HPR protein, and then searching for acceptable side-chain placement using the ALL option of the CONGEN program (Bruccoleri et al., 1988; Bruccoleri & Novotny, 1992). These searches for side-chain decoys involved fragments 1–12, 33–45, and 55–66 of the HPR protein.

The loop decoys were downloaded from *http://prostar.carb.nist.gov/PDec/PDecInfo.html*, the ProStar website. CASP1 predictions as well as native structures are available for downloading on the web at *http://PredictionCenter.llnl.gov/* from the Protein Structure Prediction Center, Lawrence Livermore National Laboratory, Livermore, California.

Testing an alternative method of free energy calculation

Throughout this paper we used van der Waals normalization in order to avoid the need for estimating vdW interactions between the protein and the solvent. As mentioned in the introduction, an alternative approach to this problem is modeling the solute–solvent vdW interactions implicitly as part of the desolvation free energy ΔG_d. Within the framework of the atomic solvation parameter (ASP) model, this can be accomplished using ASPs based on vapor-to-water transfer free energies (Wesson & Eisenberg, 1992). Then vdW interactions among protein atoms are also included in the free energy, and are calculated based on the Lennard–Jones potential. As we mentioned, the shortcoming of this approach is its extreme sensitivity to atomic positions. Nevertheless, we attempted to use this more straightforward method, and calculated the free energies for the four targets shown in Table 1 and their predictions. Prior to energy evaluation, each conformation was minimized for 200 steps. For EDN and CRABP some of the homology models have lower van der Waals energy than the native (results not shown). This is not surprising considering that mini-

mization in vacuum can lead to structures more compact than the native (Vajda et al., 1993). Although the desolvation energy which includes solute–solvent vdW interactions attempts to counteract the artificially low vdW energies, its magnitude is too small for compensation, and hence we gave up the method in favor of van der Waals normalization.

Acknowledgments

We would like to thank Zhiping Weng for technical assistance, and Charles DeLisi, Jean Garnier, and Carlos Camacho for insightful discussions. The collaboration between the authors and Dr. Garnier was supported by NATO Collaborative Research Grant CRG 950265. This work was supported by the Donors of the Petroleum Research Fund, administered by the American Chemical Society, and by grant DE-F602-96ER62263 from the Department of Energy.

References

Abagyan R, Totrov M. 1994. Biased probability Monte Carlo conformational searches and electrostatic calculations for peptides and proteins. *J Mol Biol 235*:983–1002.

Adamson AW. 1982. *Physical chemistry of surfaces*. New York: J. Wiley.

Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. 1983. Charmm: A program for macromolecular energy, minimization, and dynamics calculations. *J Comput Chem 4*:187–217.

Bruccoleri RE, Haber E, Novotny J. 1988. Structure of antibody hypervariable loops reproduced by a conformational search algorithm. *Nature (London) 335*:564–568.

Bruccoleri RE, Novotny J. 1992. Antibody modeling using the conformational search program congen. *Immunomethods 1*:96–106.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct Funct Genet 16*:92–112.

Eisenberg D, McLachlan AD. 1986. Solvation energy in protein folding and binding. *Nature (London) 319*:199–203.

Fidelis K, Stern P, Bacon D, Moult J. 1994. Comparison of systematic search and database methods for constructing segments of protein structure. *Protein Eng 7*:953–960.

Godzik A, Kolinski A, Skolnick J. 1992. Topology fingerprinting approach to the inverse folding problem. *J Mol Biol 227*:227–238.

Gulukota K, Vajda S, DeLisi C. 1996. Peptide docking using dynamic programming. *J Comput Chem 17*:418–428.

Huang E, Subbiah S, Tsai J, Levitt M. 1996. Using a hydrophobic contact potential to evaluate native and near-native folds generated by molecular dynamics simulations. *J Mol Biol 257*:716–725.

Jackson RM, Sternberg MJE. 1995. A continuum model for protein–protein interactions: Application to the docking problem. *J Mol Biol 250*:258–275.

Jernigan RL, Bahar I. 1996. Structure-derived potentials and protein simulations. *Curr Opin Struct Biol 6*:195–209.

Johnson MS, Srinivasan N, Sowdhamini R, Blundell TL. 1994. Knowledge-based protein modeling. *Critical Reviews in Biochemistry and Molecular Biology 29*(1):1–68.

King BL, Vajda S, DeLisi C. 1996. Empirical free energy as a target function in docking and design: Application to HIV-1 protease inhibitors. *FEBS Lett 384*:87–91.

Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 227*:876–888.

Mosimann S, Meleshko R, James MNG. 1995. A critical assessment of comparative molecular modeling of tertiary structures of proteins. *Proteins Struct Funct, Genet 23*:301–317.

Moult J, James MNG. 1986. An algorithm for determining the conformation of of polypeptide segments in proteins by systematic search. *Proteins Struct Funct Genet 1*:146–163.

Nauchitel V, Villaverde MC, Sussman F. 1995. Solvent accessibility as a predictive tool for the free energy of inhibitor binding to the HIV-1 protease. *Protein Sci 4*:1356–1364.

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.

Novotny J, Bruccoleri RE, Saul FA. 1989. On the attribution of binding energy in the antigen–antibody complexes McPC 603, D1.3 and HyHEL-5. *Biochemistry 28*:4735–4749.

Novotny J, Rashin AA, Bruccoleri RE. 1988. Criteria that discriminate between native proteins and incorrectly folded models. *Proteins Struct Funct Genet 4*:19–30.

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

Pellequer JL, Chen SW. 1997. Does conformational free energy distinguish loop conformations in proteins? *Biophys J 73*:2359–2375.

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

Schiffer CA, Caldwell JW, Kollman PA, Stroud RM. 1993. Protein structure prediction with a combined solvation free energy-molecular mechanics force-field. *Molec Simul 10*:121–134.

Shrake A, Rupley JA. 1973. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *Biochemistry 79*:353–371.

Sippl MJ. 1993. Boltzmann's principle, knowledge-based mean fields, and protein folding. *J Computer-Aided Molecular Design 7*:473–501.

Sippl MJ. 1995. Knowledge-based potentials for proteins. *Curr Opin Struct Biol 5*:229–235.

Smith KC, Honig B. 1994. Evaluation of the conformational free energies of loops on proteins. *Proteins Struct Funct Genet 18*:119–132.

Srinivasan R, Rose GD. 1995. Linus: A hierarchic procedure to predict the fold of a protein. *Proteins Struct Funct Genet 22*:81–99.

Vajda S, Jafri MS, Sezerman OU, DeLisi C. 1993. Necessary conditions for avoiding incorrect polypeptide folds in conformational search by energy minimization. *Biopolymers 33*:173–192.

Vajda S, Sippl M, Novotny J. 1997. Empirical potentials and functions for protein folding and binding. *Curr Opin Struct Biol 7*:222–228.

Vajda S, Weng Z, DeLisi C. 1995. Extracting hydrophobicity parameters from solute partition and protein mutation/unfolding experiments. *Protein Eng 8*:1081–1092.

Vajda S, Weng Z, Rosenfeld R, DeLisi C. 1994. Effect of conformational flexibility and solvation on receptor–ligand binding free energies. *Biochemistry 33*:13977–13988.

Verkhivker G, Appelt K, Freer ST, Villafranca JE. 1995. Empirical free energy calculations of ligand–protein crystallographic complexes. I. Knowledge-based ligand–protein interaction potentials applied to the prediction of Human Immunodeficiency Virus 1 protease binding affinity. *Protein Eng 8*:677–691.

Wallqvist A, Jernigan RL, Covell DG. 1995. A preference-based free energy parameterization of enzyme-inhibitor binding. Applications to HIV-1-protease inhibitor design. *Protein Sci 4*:1881–1903.

Weng Z, DeLisi C, Vajda S. 1997. Empirical free energy calculation: Comparison to calorimetric data. *Protein Sci 6*:1976–1984.

Weng Z, Vajda S, DeLisi C. 1996. Prediction of complexes using empirical free energy functions. *Protein Sci 5*:614–626.

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

Wilson C, Doniach SA. 1989. A computer model to dynamically simulate protein folding: studies wit crambin. *Proteins Struct Funct Genet 6*:193–209.

Yue K, Dill KA. 1996. Folding proteins with a simple energy function and extensive conformational searching. *Protein Sci 5*:254–261.