Inter-residue and solvent-residue interactions in proteins: A statistical study on experimental structures

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · APRIL 2004

Impact Factor: 2.63 · DOI: 10.1002/prot.20030 · Source: PubMed

Riccardo Chelli
University of Florence
87 PUBLICATIONS 1,747 CITATIONS

Francesco Luigi Gervasio
University College London
94 PUBLICATIONS 3,040 CITATIONS

Piero Procacci
University of Florence
117 PUBLICATIONS 2,073 CITATIONS

Vincenzo Schettino
University of Florence
195 PUBLICATIONS 3,427 CITATIONS

Inter-residue and Solvent-residue Interactions in Proteins: A Statistical Study on Experimental Structures

Riccardo Chelli,[1,2]* Francesco Luigi Gervasio,[3,4] Piero Procacci,[1,2] and Vincenzo Schettino[1,2]
[1]*Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy*
[2]*European Laboratory for Nonlinear Spectroscopy (LENS) Via Nello Carrara 1, 50019 Sesto Fiorentino, Italy*
[3]*Centro Svizzero di Calcolo Scientifico, Via Cantonale, CH-6928 Manno, Switzerland*
[4]*Physical Chemistry, ETH Zurich, Hönggerberg, CH-8093 Zurich, Switzerland*

ABSTRACT **A large set of protein structures resolved by X-ray or NMR techniques has been extracted from the Protein Data Bank and analyzed using statistical methods. In particular, we investigate the interactions between side chains and the interactions between solvent and side chains, pointing out on the possibility of including the solvent as part of a knowledge-based potential. The solvent-residue contacts are accounted for on the basis of the Voronoi's polyhedron analysis. Our investigation confirms the importance of hydrophobic residues in determining the protein stability. We observe that in general hydrophobic-hydrophobic interactions and, more specifically, aromatic-aromatic contacts tend to be increasingly distally separated in the primary sequence of proteins, thus connecting distinct secondary structure elements. A simple relation expressing the dependence of the protein free energy by the number of residues is proposed. Such a relation includes both the residue-residue and the solvent-residue contributions. The former is dominant for large size proteins, whereas for small sizes (number of residues less than 100) the two terms are comparable. Gapless threading experiments show that the solvent-residue knowledge-based potential yields a significant contribution with respect to discriminating the native structure of proteins. Such contribution is important especially for proteins of small size and is similar to that given by the most favorable residue-residue knowledge-based potential referring to hydrophobic-hydrophobic interactions such as isoleucine-leucine. In general, the inclusion of the solvent-residue interaction produces a relevant increase of the free energy gap between the native structures and decoys. Proteins 2004;55:139–151.**

INTRODUCTION

With the growing number of protein structures made available on the Protein Data Bank (PDB)[1] and similar databases, statistical studies on their ensemble are increasing both in number and significance.[2–11] Most of these studies are centered on the relative abundance and geometry of specific residue-residue contacts.[5–11] The general idea[12] in such approaches is that the probability of finding any two residues at a given distance is related to the Boltzmann factor of an effective interresidue potential of mean force (PMF) in physiological conditions. Studies based on this principle have been used mainly i) to gain insights on the importance of specific interactions between pairs of residues in stabilizing the protein structures and ii) to devise knowledge-based potentials for predicting the tertiary structure of a protein given the primary one. As to the first issue, several articles[9,10,13] have highlighted the role of the interactions between hydrophobic residues in the core and charged residues on the protein surface. In particular it has been shown that aromatic residues assume sandwich conformations more often than predicted by a random distribution,[14–19] suggesting that their interaction might be important in the stabilization of the tertiary structure of proteins. As to the second issue, i.e., tertiary structure prediction from knowledge-based PMFs, one of the simplest and most followed approaches is based on pairwise residue-residue contact potentials.[20] From the contact maps averaged over experimental structures, a contact energy parameter set, representing residue-residue interaction free energy (or PMF), can be built. Hence a minimalistic square well potential can be defined as follows: zero beyond the threshold distance ($R_{\max}$), infinity for distances below the sum of hard core radii of the particle residues ($R_{\min}$), and equals to the PMF in the [$R_{\min}$, $R_{\max}$] interval. Whether such simple interresidue pair potentials, based on contact frequencies and excluded volumes, correctly reflect the basic physical forces stabilizing the native structure of proteins remains a subject of debate.[21–24] Several studies indicate that a simple residue-residue PMF cannot stabilize the correct fold of native proteins.[20,25] In an effort to go beyond the simple pairwise contact potential, secondary structure specific effects have been included by devising expanded 60-residue alphabet[26]

The Supplementary Materials Referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/index.html

Grant sponsors: Italian Ministero dell'Istruzione, dell'Università e della Ricerca (MIUR) and the European Union; Grant numbers: ERB-FMGE-CT950017 and HPRI-CT-1999-00111.

*Correspondence to: Riccardo Chelli, Dipartimento di Chimica, Università di Firenze, Via della Lastruccia 3, 50019 Sesto Fiorentino, Italy. E-mail: chelli@chim.unifi.it.

Received 13 May 2003; Revised 1 August 2003; Accepted 10 September 2003

(20 amino acids combined with three secondary structure states), chain connectivity and solvent effects have been taken into account using a more careful definition of the residue-residue PMF,[13] short range multibody terms have been explicitly included,[27] three-body "rope thickness" related potential has been introduced.[28,29]

In the present work we build an updated knowledge-based residue-residue and solvent-residue PMF based on a database of 1671 nonhomologous protein structures.[30,31] The residue-residue PMF for a given residue pair is calculated on the basis of the number of times those residues come in contact in the protein data set. Solvent-residue interactions are treated on the same footing of the residue-residue ones. The solvent is assumed to be a delocalized residue in contact with exposed side chains. The side chains are classified as solvent-exposed by using the Voronoi's polyhedron[32,33] analysis. In this way we establish an energy parameter set that explicitly takes into account the contribution of the residues on the protein surface to the overall free energy. As expected, the solvent-residue PMF is favorable for charged and polar residues and highly unfavorable for the hydrophobic ones. The resulting free energy function confirms that not only the hydrophobic-hydrophobic interactions, but also the interactions between the solvent and the charged (or polar) residues, are important in determining the structural properties of the proteins. We also find that the interactions between aromatic residues occur more frequently for distal than for proximal pairs in the primary structure, hence connecting distinct secondary structure elements with enhanced probability. The protein free energy, calculated from the contact maps and from the PMFs, is found to depend only on the interresidue interactions for large size proteins, whereas solvent-residue contribution becomes important for small chains.

A test of the proposed knowledge-based potential is performed with standard gapless threading experiments.[34,35] The free energy of the threaded folds is calculated using the residue-residue PMF with and without the solvent-residue contribution. Results for threading experiments show that the inclusion of the solvent-residue PMF produces an increase of the average free energy gap between the native fold and the decoys. Also in discriminating the native structure of a protein with respect to different possible folds, the solvent seems to play an important role, especially for small size proteins.

MATERIALS AND METHODS

Definition of the Residue-Residue Potential of Mean Force

Many researchers[9,13,36] assume that any two residues are "in contact" when their "distance" is below a certain threshold. In this simplified approach, therefore, two residues may be either in contact or not. The value of the distance, in turn, depends on how it is defined. Different residue-residue contact definitions can be based on i) side chain centroid-centroid distance, ii) nearest atom-atom distance, and iii) C_α–C_α distance. These contact definitions have threshold consensus distances of 7–8 Å,[26,37] 4–5 Å,[36,38] and 6–7 Å,[36] respectively. In the present work we have used the centroid-centroid or Miyazawa-Jernigan convention.[37] From the computational standpoint this method is much cheaper than the all-atom approach, while being physically more accurate than the "C_α" one, as it takes explicitly into account the side chain-side chain interactions. It has been recently shown[13] that pairing frequencies beyond a distance of about 7.5 Å using the centroid-centroid criterion can be completely explained by the naturally inhomogeneous distribution of residues, by the finite size of the proteins and by the chain connectivity. In the present study, the used threshold distance varies from 7.0 to 8.0 Å.

Before calculating the PMF between two side chains of type i and j (by meaning as "type of a side chain" the corresponding amino acid), one must evaluate the ij pair correlation function. This last can be conveniently defined as follows:

$$g_{ij} = \frac{\sum_{p=1}^{N} \Delta n_{ij}^{(p)}(R_0)\, n_p}{\sum_{p=1}^{N} n_{ij}^{(p)}} \tag{1}$$

where the sums run over all the N proteins of the database, $\Delta n_{ij}^{(p)}(R_0)$ is the number of centroid-centroid contacts (below the threshold distance R_0) for the ij residue pair in the protein p, n_p is the number of residues in the protein p, and $n_{ij}^{(p)}$ is the number of all the possible contacts between residues of type i and j in the protein p. In the present case, the PMF for a ij residue pair is calculated from the normalized form of g_{ij}, namely[9,39]:

$$G(i, j) = g_{ij}\left(\frac{1}{400}\sum_{k=1}^{20}\sum_{l=1}^{20} g_{kl}\right)^{-1} \tag{2}$$

where the sums run over all the residue types. The normalization factor for g_{ij} corresponds to a well-defined choice of the reference state. This state is assumed to be an ideal globular protein where the radial distribution function of any pair of residues (independent on their types) is given by the average of all the pair distribution functions obtained by applying Eq. (1) to the protein database. The n_p factor in Eq. (1) makes the pair correlation function independent on the protein size for fixed residue composition. In fact, because the dependence of $n_{ij}^{(p)}$ and $\Delta n_{ij}^{(p)}(R_0)$ is quadratic and linear in n_p, respectively, the global trend of g_{ij} is constant as the protein size increases. The definition of g_{ij} of Eq. (1) is based on the following assumption: for a protein p, the correct ij pair distribution function is defined as

$$g_{ij}^{(1)} = \frac{\Delta n_{ij}^{(p)}(R_0)/V_0}{n_{ij}^{(p)}/V_p} \tag{3}$$

where V_0 is the volume of a sphere of radius R_0 and V_p is the volume of the protein p. The $g_{ij}^{(1)}$ function measures the fractional deviation of the two-residues contact correlation

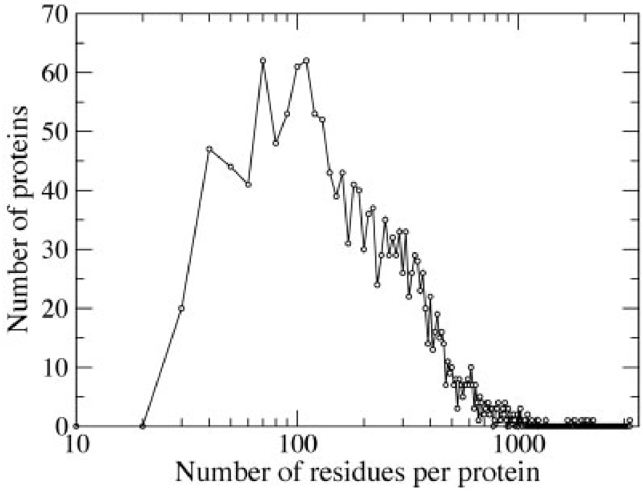

Fig. 1. Distribution of the number of proteins in the database as a function of the number of residues per protein (logarithmic scale). The line is reported as a guide for eyes.

function from the value obtained as if the residues i and j were homogeneously distributed in the protein p according to their natural abundance. The volume V_p can be assumed to be approximately proportional to the total number of residues in the protein p (namely, $V_p = Kn_p$). Hence using this approximation, the ij pair distribution function is given by

$$g_{ij}^{(1)} = \frac{\Delta n_{ij}^{(p)}(R_0)\; n_p\; K}{n_{ij}^{(p)}\; V_0} \qquad (4)$$

Extending Eq. (4) to all protein domains in the database, we recover Eq. (1), apart from the constant factors, K and V_0. Because such factors can be simplified in Eq. (2), g_{ij} instead of $g_{ij}^{(1)}$ can be used to calculate $G(i,j)$.

The pair correlation function g_{ij} of Eq. (1) could be defined in a different way, namely by averaging $g_{ij}^{(1)}$ of Eq. (4) over all the protein database:

$$g'_{ij} = \frac{1}{N} \sum_{p=1}^{N} \frac{\Delta n_{ij}^{(p)}(R_0)\; n_p}{n_{ij}^{(p)}} \qquad (5)$$

where N is the number of proteins in the database. Eqs. (1) and (5) are equivalent had all the proteins in the database the same size and composition. According to Eq. (5), small and large proteins have the same statistical weight. Therefore, given the distribution of the number of proteins (in the database) as a function of the number of residues reported in Figure 1 (see below for the selection criteria of the protein database), Eq. (5) yields a strong small proteins bias. On the other hand one should consider that energetic stabilization factors (e.g., presence of external cofactors, contribution from the secondary structure, etc.), which are different from the normal "affinity" between side chains considered here, are more important for small proteins. So, in order to avoid overestimation of these features on the PMF, $G(i,j)$ [Eq. (2)] in combination with g_{ij} [Eq. (1)] has been used.

Once $G(i, j)$ is calculated, the residue-residue PMF (in RT units) between the residues of type i and j is given by[40]

$$w(i, j) = -\ln[G(i,j)] \qquad (6)$$

Starting from the knowledge-based contact potential, the residue-residue contribution to the free energy of any given protein can be easily computed via its contact map.[41] The contact map $\mathbf{S}$ of a protein with n_p residues can be associated to an $n_p \times n_p$ symmetric matrix S, whose elements are as follows:

$$S_{lk} = S_{kl} = \begin{cases} 1, & \text{if the residues } k \text{ and } l \text{ are in contact} \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

Of course, the diagonal elements of S have no physical meaning, because they represent residue self-interactions. However, for suitability of notation, the condition $S_{ii} = 0$ will be assumed in the following. Using the previous definition for $\mathbf{S}$, the residue-residue contribution to the protein free energy is given by

$$E_{rr}(\mathbf{A}, \mathbf{S}) = \frac{1}{2} \sum_{k=1}^{n_p} \sum_{l=1}^{n_p} w[t(l), t(k)]\, S_{kl} \qquad (8)$$

where $\mathbf{A}$ denotes the primary structure of the protein chain [$\mathbf{A} = (a_1, a_2, \ldots, a_{n_p})$], $w(i, j)$ is the PMF of the ij residue pair [Eq. (6)], and $t(k)$ is a function that associates the amino acid type (TYR, ALA, etc.) to the kth residue.

Definition of Exposed Residues and Solvent-Residue Potential of Mean Force

Solvent-residue interaction is indirectly taken into account in the overall protein free energy [Eq. (8)]. In fact, because of the presence of the solvent, we expect hydrophobic-hydrophobic contact free energies to be lower than those corresponding to contacts between charged or polar residues, because these last occur mostly on the protein surface. On the other hand, a residue on the protein surface is thermodynamically stabilized not only by the interactions engaged with the neighboring residues, but also by the interactions engaged with the solvent molecules. For example, when charged residues on the protein surface strongly interact with the solvent (generally water) through hydrogen bonds, their contribution cannot certainly be neglected in computing the effective free energy of the protein in solution. In past studies there have been several attempts to explicitly include solvent-protein contribution in knowledge-based free energies, including, for example, free energy terms based on the mean coordination number[41] or on one-residue potentials accounting for the non-polar-in–charged-out character.[13] In the present case, the solvent is treated as an *additional* and *delocalized* residue, i.e., a matrix where all elements (proteins) of the statistical ensemble are embedded. We must hence establish a criterion for counting the "contacts" between the matrix "solvent" and standard residues. To this end, we consider the protein (whose side chains are represented by the corresponding centroids) to be embedded in a large (compared with the protein dimensions) cubic box with eight dummy particles at its vertices. We then compute the volume of the Voronoi's polyhedra[32] for each side chain, using, when needed, the dummy particles for closing the polyhedra. In this manner exposed residues are characterized by large volumes. Given that side chains are highly packed in proteins, Voronoi volumes that are fairly above the mean bulk side chain volume are indicative of solvent exposure. Hence, in analogy with the assumption of contact radius, we define a residue to be in contact with the solvent if its Voronoi volume exceeds a certain arbitrary threshold V_{ex}. The optimum value of V_{ex} is chosen so as to minimize the number of proteins (in the protein database) that yield a *positive* folding free energy (see next section for the folding free energy definition).

Following the treatment done for the residue-residue pair distribution function (see previous section), one can define the solvent-residue pair distribution function as

$$g_s(i) = \frac{\sum_{p=1}^{N} \Delta s_i^{(p)}\, n_p^{2/3}}{\sum_{p=1}^{N} n_i^{(p)}} \qquad (9)$$

where $\Delta s_i^{(p)}$ is the number of exposed residues of type i in the protein p, $n_i^{(p)}$ is the number of residues of type i in the protein p, and the sums run over all the proteins of the database. The solvent-residue [Eq. (9)] and residue-residue [Eq. (1)] pair correlation functions are fully consistent. The $n_p^{2/3}$ factor is approximately proportional to the area of the protein surface and plays the same role of the n_p factor in Eq. (1), i.e., it makes the potential independent on the protein size at fixed residue composition. For the solvent-residue PMF evaluation, we consider the normalized form of $g_s(i)$:

$$G_s(i) = g_s(i)\left(\frac{1}{20}\sum_{k=1}^{20} g_s(k)\right)^{-1} \qquad (10)$$

In Eq. (10) we implicitly assume that the reference state for the solvent-residue interaction is the same of the residue-residue one [see Eq. (2)]. The solvent-residue PMF (in RT units) for the residue of type i can be then calculated from $G_s(i)$:

$$w_s(i) = -\ln[G_s(i)] \qquad (11)$$

On the basis of the solvent-residue PMF of Eq. (11), one can calculate the contribution given by the solvent-protein interaction to the total free energy of a protein. Following Eq. (7), a contact map $\mathbf{T}$ for the solvent-residue interactions can be defined and associated to a vector T of n_p elements:

$$T_k = \begin{cases} 1, & \text{if the residue } k \text{ is solvent-exposed} \\ 0, & \text{otherwise} \end{cases} \qquad (12)$$

Analogously to Eq. (8), the solvent-residue contribution to the protein free energy is

$$E_{sr}(\mathbf{A}, \mathbf{T}) = \sum_{k=1}^{n_p} w_s[t(k)]\, T_k \qquad (13)$$

where T_k is defined by Eq. (12), $w_s(i)$ is defined by Eq. (11), and $t(k)$ is the function used in Eq. (8).

Free Energy of a Protein

The knowledge-based pair potentials $w(i, j)$ and $w_s(i)$ account for residue-residue and solvent-residue interactions, respectively. In principle, these terms alone should account for most of the protein folding free energy, which is the reversible work needed to fold a protein in its native state starting from a disordered unfolded state. The role of the primary structure connectivity is simply that of preemptively defining a kinetic route to the supposedly unique native fold. In fact, during the folding process, the chain connectivity is preserved, so that the thermodynamically accessible conformational states are severely limited by the distributions of charged or polar (i.e., preferentially exposed) and hydrophobic (i.e., buried) residues along the protein chain. When evaluating free energy difference between folded states (characterized by different contact maps **S**, but identical primary structure **A**), the probability bias in the residue-residue pairing frequencies, due to the presence of the distance constraint imposed by the primary structure topology, can be reasonably assumed to depend only on the invariant sequence **A**. Therefore the associated free energy is the same for any folded state. The free energy of a given folded state, characterized by an **S** residue-residue map and a **T** solvent-residue map, can be therefore calculated as

$$E(\mathbf{A}, \mathbf{S}, \mathbf{T}) = E_{rr}(\mathbf{A}, \mathbf{S}) + E_{sr}(\mathbf{A}, \mathbf{T}) \qquad (14)$$

where $E_{rr}(\mathbf{A}, \mathbf{S})$ and $E_{sr}(\mathbf{A}, \mathbf{T})$ are defined by Eqs. (8) and (13), respectively. In Eq. (14) the reference state corresponds to that assumed for $w(i, j)$ [Eq. (6)] and $w_s(i)$ [Eq. (11)]. However, a different reference state can be considered, namely the state where the protein is completely unfolded. From the residue-residue and solvent-residue contact map standpoint such an unfolded reference state corresponds to an "ideal" state where each residue is exposed to the solvent [all the elements of the T vector of Eq. (12) are one] and none of the residues is involved in residue-residue interactions [all the elements of the S matrix of Eq. (7) are zero]. Hence, the free energy [Eq. (14)] of this unfolded state is

$$E_{\mathrm{u}}(\mathbf{A}) = \sum_{k=1}^{n_p} w_s[t(k)] \qquad (15)$$

and the total protein free energy with respect to the above reference state is

$$E'(\mathbf{A}, \mathbf{S}, \mathbf{T}) = E_{rr}(\mathbf{A}, \mathbf{S}) + E_{sr}(\mathbf{A}, \mathbf{T}) - E_{\mathrm{u}}(\mathbf{A}) \qquad (16)$$

Selection of the Protein Database

A set of 2516 protein structures was extracted from the PDB[1] (ensemble FSSP[30,31]). Before performing the statistical analysis, the structures were selected according to the following criteria: i) PDB files with only C_α-traces have been discarded; ii) PDB files containing defective or nonstandard residues have been discarded; iii) PDB files containing only α-helix or β-sheet have been discarded (see discussion below); iv) when several model structures are proposed in a PDB file, only one of them has been randomly retained; v) only nonhomologous primary structures have been considered, two sequences being defined homologous when they share more than 25% residues.[30,31] For homologous sequences, only one has been randomly retained. However, some PDB files referring to complexes may contain protein chains also included in other PDB files either as monomer or as chain involved in different complexes. In this occasional case the chains have been retained. After the screening procedure, we were left with 1671 PDB files (the list of the processed PDB files is available as supplementary material),[47] corresponding to 1671 distinct proteins domains (with the few and statistically irrelevant exceptions due to chains present in distinct complexes). In the overall, the processed database contained a total number of 4.71×10^5 residues. The distribution of the number of proteins as a function of the number of residues per protein is reported in Figure 1. The maximum of the distribution occurs between 70 and 100 residues.

RESULTS AND DISCUSSION

Residue-Residue and Solvent-Residue Potential of Mean Force

The residue-residue and solvent-residue PMFs [Eqs. (6) and (11), respectively] are reported in Figure 2 (top) in a 3D plot. In the figure, the residues are ordered on the basis of their hydrophobicity, whereas the unlabeled topmost and rightmost positions refer to the solvent-residue PMF. This order has been established, not only on the basis of the chemical intuition, but also considering the fraction of solvent-exposed residues calculated for each residue type in the PDB. (In the Voronoi's polyhedron analysis, the threshold volume V_{ex} of 350 Å^3 has been assumed for the solvent-exposure criterion). The fraction of residues exposed to the solvent is also reported in Figure 2 (bottom). It can be observed that there exists a correspondence between solvent-exposure and solvent-residue PMF: high solvent-exposure corresponds to low solvent-residue PMF and *vice versa*. This is trivially due to the fact that solvent-residue PMF is proportional to the number of solvent-exposed residues [see Eq. (9)]. As anticipated before, PMFs show that the hydrophobic residues tend to be closely packed in the protein core. In fact, as found in other studies,[41] the least energy contact PMFs are those corresponding to the triad VAL, ILE, LEU, and to the diad TRP, PHE. Correspondingly, the probability of being solvent-exposed for these residues is around 0.1, only the probability of CYS solvation being smaller. This is reflected in the associated solvent-residue PMFs, that are by far the largest for VAL, ILE, LEU, TRP, and PHE (apart for the solvent-CYS interaction). By inspection of the solvent-exposure of CYS (<5%) and of the residue-CYS PMFs, one can deduce that, when CYS is present in a protein, it is engaged almost exclusively in sulfur bridges, without any preferential correlation with the remaining residue types. In fact, the PMF between CYS and all other hydrophobic residues is uniform from ASN to VAL, while being greater in the case of charged residues. Hydrophilic residues weakly interact with other residues and tend to be distributed on the protein surface so as to maximize their interaction with the solvent. Note that the PMF relative to residue pairs with opposite charges (e.g., LYS-GLU) is much lower than that corresponding to pairs with charges of the same signs (e.g., LYS-LYS).

The reliability of the PMF shown in Figure 2 critically depends on the statistical significance of the underlying database. In order to roughly assessing the statistical error, we randomly divided the database in two parts of approximately equal number of proteins. For the two sets, the residue-residue and solvent-residue PMFs have been calculated. From these two independent sets of PMFs, we have then calculated the average value of the difference (in absolute value) of corresponding residue-residue PMFs. The estimated error is only 0.027 RT, well below the differences observed for the PMFs of different residue pairs (see the chromatic scale in Fig. 2).

By inspection of Figure 2, we can loosely subdivide the 20 residue types in three somehow overlapping categories: a) hydrophobic or core residues: CYS, VAL, ILE, LEU, PHE, TRP, TYR, and MET; b) intercalating residues: GLY, ALA, THR, HIS, SER, and PRO; c) hydrophilic or surface residues: ARG, LYS, GLU, ASP, GLN, and ASN. The intercalating residues have roughly the same probability of being on the surface or in the protein core and, in general, show quite weak PMF ($-0.25\,RT < w(i,j) < 0.25\,RT$) with any partner, including solvent. The topological distribution of the hydrophobic and hydrophilic residues along the protein chain very likely codes for the native state. In fact, given the constraint of the invariance of $\mathbf{A}$ for any folded state, they restrain the accessible conformational space by strongly interacting with each other (hydrophobic) or with the solvent (hydrophilic). Because the intercalating residues have no specific interactions with other residue types and with the solvent, they play probably a minor role in the folding process and in stabilizing the native state. This picture of the folding process of a copolymer chain, i.e., a chain made of residues that set the path to the ground state and of indifferent residues as filling intercalating neutral units, do not take into account the existence of the secondary structure. We might infer that secondary structure formation is a *structural adjustment* of a proto native fold determined mainly by the elementary residue-residue and solvent-residue interactions and by the topological distribution of the residues along the primary structure. In other terms, the secondary structure is formed, in a second stage or concurrently, in those regions where it is possible to strain or linearize the structure without strong variation of the residue-residue and solvent-residue stabilization energy. In absence of specific interactions, such sequence independent "linearization" potentials have been recently postulated in the determination of the optimal shape of closely

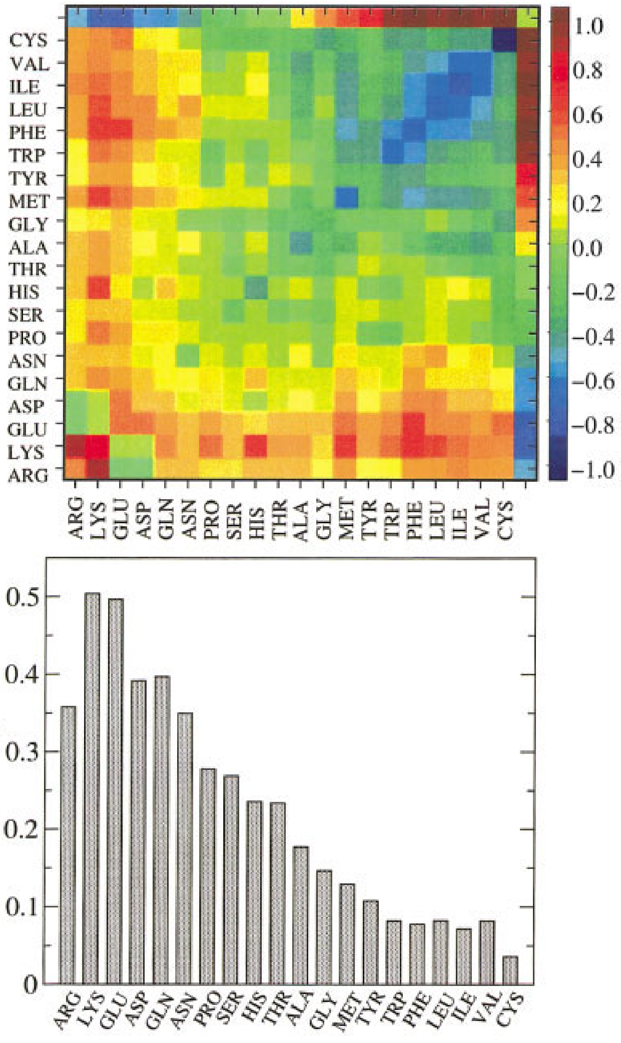

Fig. 2.

packed compact strings[28] and of secondary structures in proteins.[29]

Role of the Aromatic-Aromatic Interaction in Protein Folding

In Figure 3 we show the residue-residue PMFs computed using Eq. (1) with $\Delta n_{ij}^{(p)}(R_0)$ and $n_{ij}^{(p)}$ referring to contact pairs separated by more than 40 and 100 residues in the primary structure. Because of the elimination of the negative bias yielded by the chain connectivity that forces repelling residues to be in contact, the wide minimum involving hydrophobic residues deepens with respect to the PMF of Figure 2. As previously suggested,[16,18,19] it is indeed noticeable the emerging role of the interaction between aromatic residues (PHE and TRP in particular) in connecting different secondary structure elements. The role of the aromatic residues can be better appreciated in Figure 4, where the PMFs for aromatic-aromatic and aliphatic-aliphatic pairs are reported as a function of the number of separating residues (N_r in figure). In particular

Fig. 2. Top: residue-residue and solvent-residue PMFs. The residue-residue PMF has been calculated using a threshold distance R_0 = 8 Å. The topmost row and rightmost column refer to the solvent-residue PMF. The chromatic scale (right of the picture) is in *RT* units. The numerical values of the PMF are available as supplementary material. Bottom: fraction of solvent-exposed residues for each amino acid type. The results are from the Voronoi's polyhedron analysis using a threshold volume of 350 Å^3.

Fig. 3. Left (PMF-40): residue-residue PMF calculated considering only the contribution from the pairs of residues far more than 40 residues in the primary sequence. Right (PMF-100): residue-residue PMF calculated considering only the contribution from the pairs of residues far more than 100 residues in the primary sequence. The chromatic scale (right of the pictures) is in *RT* units for both the PMF-40 and PMF-100 cases. The residue-residue PMFs have been calculated using a threshold distance R_0 = 8 Å. The solvent-residue contribution (also shown in Fig. 2) is reported for comparison. The numerical values of the PMFs are available as supplementary material.

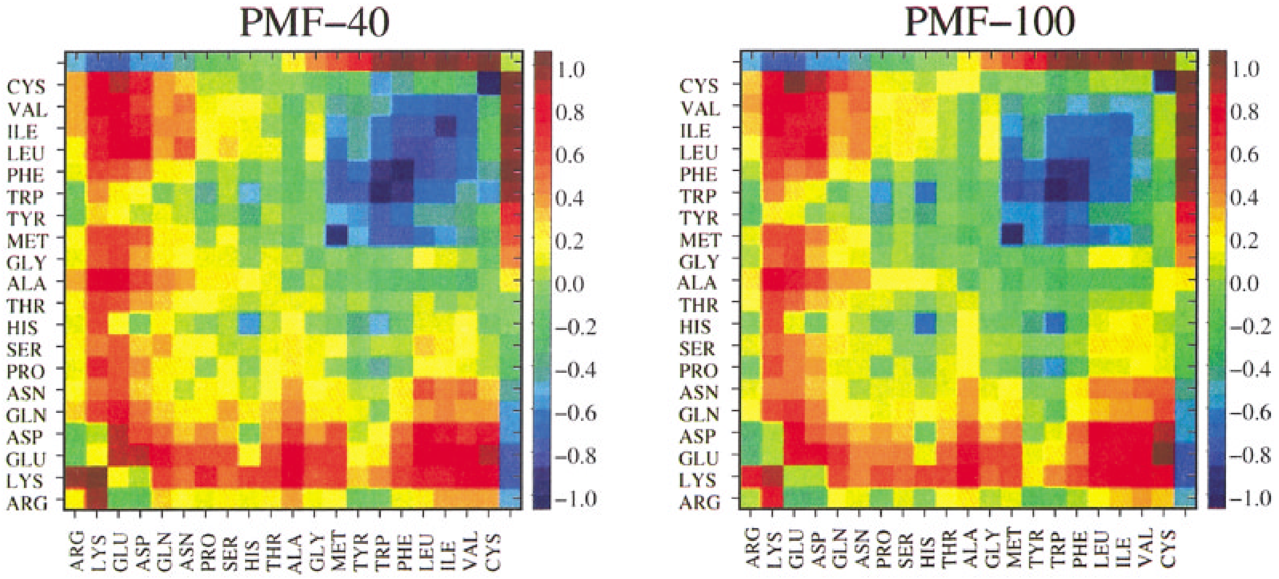

Fig. 3.

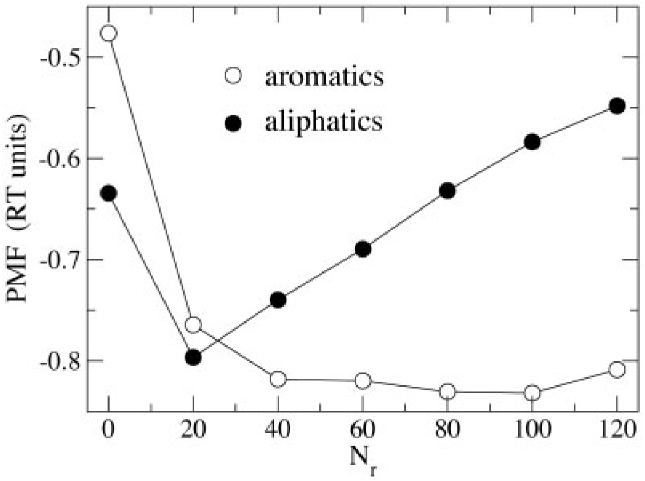

Fig. 4. Average residue-residue PMF for aromatic-aromatic pairs (PHE-PHE, PHE-TYR, PHE-TRP, TYR-TYR, TYR-TRP, and TRP-TRP) and aliphatic-aliphatic pairs (LEU-LEU, LEU-ILE, LEU-VAL, ILE-ILE, ILE-VAL, and VAL-VAL) as a function of N_r, namely the residue-residue distance (in "number of residues" units) in the primary sequence. The lines are reported as a guide for eyes.

we report the PMFs obtained by averaging over the pairs made of only TRP, TYR, and PHE (i.e., PHE-PHE, PHE-TRP, etc.) and of only ILE, VAL, and LEU. When all interactions are accounted for in the calculation of PMF ($N_r = 0$ in Fig. 4), aliphatic residues have a PMF lower than the aromatic ones. This last behavior is observed for $N_r < 20$. The relative large PMF values for $N_r = 0$ are probably due to two main factors: i) neighboring residues along the primary structure are partially constrained to interact independently of their affinity and ii) the distribution of the residues along the primary structure is almost random. For $N_r > 20$ the PMF of aromatic residues is constantly below -0.75 RT, while for the aliphatic residues the PMF versus N_r shows a monotonous increasing trend. For $N_r = 120$ the difference in PMF between aromatic and aliphatic residues is > 0.25 RT. This enhanced contribution of aromatic with respect to other hydrophobic residues in the folding stabilization of globular proteins can be ascribed to the stabilization due to the stacking conformation of the aromatic rings.[19,42]

From Figure 3, we can also observe that an important contribution in connecting different parts of the protein chain is given by the interaction of MET with other hydrophobic residues and in particular by MET self-interactions. This behavior is actually not clear, but probably it can be attributed to the high flexibility of MET, which allows a close packing of the side chains in the protein core.

Free Energy of Native States

In Figure 5 (top) we show the residue-residue and solvent-residue contributions to the free energy of the proteins as a function of the number of residues. The solvent-residue free energy is computed using a volume threshold for the solvent-exposure (V_{ex}) of 350 Å^3 (see Materials and Methods). By fitting the residue-residue free energy with the $y = ax$ function, we found $a = -0.14$ RT (a corresponds to the average residue-residue free energy per residue) and a correlation coefficient of 0.64. With respect to the previous work of Bahar and Jernigan,[9] the residue-residue stabilization free energy is lower (-0.14 versus -0.43 RT), but in that case a different way to calculate the PMF of the residue-residue interactions was used. For the solvent-residue free energy, the dependence on the number of residues of a protein is less simple. Hypothesizing the trend of the solvent-residue free energy (E_{fe}) to be linear with the number of the solvent-exposed residues (N_s) in a protein, one can write $E_{\mathrm{fe}} = E_{\mathrm{fe}}^{(1)} N_s$, where $E_{\mathrm{fe}}^{(1)} = -0.20$ RT is the calculated average solvent-residue free energy per solvent-exposed residue. By assuming a simple model of protein, made of spherical closely packed residues forming a globular (nearly spherical) protein, one can find a simple and approximate formula for correlating E_{fe} to the number of residues in the protein, n_p:

$$E_{\mathrm{fe}} = 3E_{\mathrm{fe}}^{(1)} n_p^{2/3} \tag{17}$$

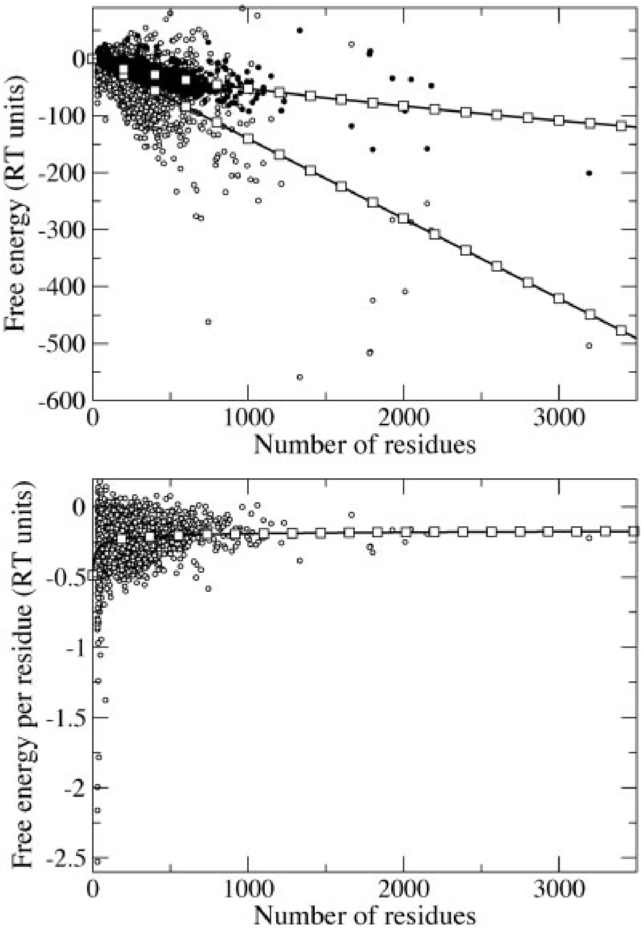

Fig. 5. Top: residue-residue (○) and solvent-residue (●) free energy as a function of the number of residues. The full lines with open squares are fits of the data with a model behavior (see text for details). Bottom: total free energy per residue as a function of the number of residues. The full line with open squares has been obtained from Eq. (18).

TABLE I. Percentage of Proteins with Positive Folding Free Energy[†]

	No solvent[a]	V_{ex}[b] = 150 Å^3	V_{ex}[b] = 350 Å^3
PMF8[c]	14.5	2.3	1.2
PMF7[d]	20.4	5.4	3.5

[†] The percentage is calculated with respect to the number of proteins in the database (1671). See Materials and Methods for computational details.

[a] Results obtained without solvent-residue contribution.

[b] V_{ex} is the threshold volume used to calculate the solvent exposure (see text for details). PMF7 and PMF8 share the calculation method of the solvent-residue contribution.

[c] The residue-residue PMF is calculated using $R_0 = 8$ Å.

[d] The residue-residue PMF is calculated using $R_0 = 7$ Å.

It is easy to demonstrate that, for a model where a larger compactness is assumed, a prefactor greater than three can be obtained in Eq. (17). By fitting the solvent-residue free energy of Figure 5 with a function $y = bE_{\rm fe}^{(1)}x^{2/3}$, a value of 2.5 for b has been found. Excluding from the fit the proteins with <500 residues, i.e., considering only large proteins that more probably can have a globular form, we found $b = 2.7$. This value, slightly smaller than that of the compact-sphere model, is compatible with the minimalist view that a (large) protein can be represented as a globular cluster made of shapeless residues.

In Figure 5 (bottom) we report the free energy per residue (residue-residue + solvent-residue) as a function of the number of residues n_p. In the figure we also report the free energy per residue function as obtained from the model discussed before, namely,

$$y = -0.14 - 0.5\, n_p^{-1/3} \tag{18}$$

The deviation of the protein free energy per residue from Eq. (18) is greater for small proteins for which various factors not considered in the calculation (e.g., secondary structure) might be important. The value of the stabilization free energy per residue obtained in the present study agrees almost qualitatively with recent investigations on protein databases,[11] where the knowledge-based potential was obtained by maximizing the thermodynamic average of the overlap between protein native structure and a Boltzmann ensemble of alternative structures. In addition, the rather small value ($-0.21\ RT$ in average) appears to be plausible given that the observed denaturation free energies are of the order of tens of kcal/mol for most proteins.[43]

The folding free energy can be calculated with respect to a completely unfolded state as explained in Materials and Methods [$E'(\mathbf{A}, \mathbf{S}, \mathbf{T})$ of Eq. (16)]. When plotted as a function of the number of residues of the protein, $E'(\mathbf{A}, \mathbf{S}, \mathbf{T})$ shows an almost linear trend. A fit with a function $y = ax$ gives $a = -0.35\ RT$ with a correlation coefficient of 0.87. Similar behavior is obtained using the PMF evaluated by Bastolla et al.[11] In this last case the fit gives $a = -0.41\,RT$, with a correlation coefficient of 0.97. In Table I we report the number of proteins (in percentage with respect to the number of proteins in the database) having an unstable folded state with respect to the hypothetical unfolded one. In the table, different entries refer to different ways used to calculate the residue-residue and solvent-residue PMFs. The better result, i.e., the smallest number of unstable native folds, is obtained using a cut-off radius of 8 Å for the centroid-centroid distance and a volume V_{ex} of 350 Å^3 for the solvent-exposure calculation. However, no relevant differences are observed for changes of the cut-off radius or when different V_{ex} is used. On the other hand, differences even of one order of magnitude are obtained when the solvent-residue contribution to the free energy is neglected. The role of the solvent in stabilizing the folded states is increasing in importance for medium or small proteins. In fact, for proteins having >320 residues, the residue-residue PMF is able alone to give stable folds.

The residue-residue PMF proposed by Bastolla et al.,[11] supplemented with our solvent-residue PMF, yields for all 1671 proteins negative folding free energies. As by definition the residue-residue part is absent in the arbitrary unfolded state, this result may be simply due to a hidden or arbitrary negative offset existing between the residue-residue PMFs. A more consistent comparison between the two knowledge-based potentials will be given in the next section where results of threading experiments are reported.

We would now discuss the results on the folding free energies obtained using $V_{ex} = 350$ Å^3 and $R_0 = 8$ Å. In this case the number of folds with positive free energy is 20. Among these seemingly "unstable" proteins, only four have >200 residues. It should be stressed that, because of the statistical nature of the knowledge-based potential and given that the secondary structure stabilization energy is not included in the knowledge-based potential, positive folding free energies can be found for small proteins because in these systems specific interactions are more important than in large domains. The four unstable large size folds are 1fgj, 1fxk, 1quu, and 2nsy. 1fgj refers to a monomer of hydroxylamine oxidoreductase. Stabilization in this case may be due to oligomerization (hydroxylamine oxidoreductase is indeed stable as a trimer) and to the presence of the heme groups.[44] As to the others, 1fxk, 1quu, and 2nsy, 76%, 83%, and 60% of the residues are in α-helix, well above the average percentage of α-helix of ~33%.[43] Hence in these cases stabilization due to secondary structure must certainly play an important role.

Threading Experiments

Knowledge-based potentials are generally tested using so-called gapless threading experiments.[34,35] For a given primary sequence with n residues, this very simple methodology provides a fast mean of producing plausible tertiary structures by "threading" the n residues putative fold from proteins with N residues ($N \geq n$). So, for example, a protein of N residues produces $N - n + 1$ folds (or decoys) for a primary sequence made of n residues. Decoys are guaranteed to represent physical folds as they have been "threaded" from an existing structure. It has been often pointed out that gapless threading is not an efficient way of producing low energy folds.[25,36] In fact, low energy folds can be more effectively produced by stochastic or

TABLE II. Percentage of Misclassified Proteins[†]

	No solvent[a]	V_{ex}[b] = 150 Å^3	V_{ex}[b] = 350 Å^3
PMF8[c]	9.5	7.8	7.8
PMF7[d]	10.4	8.0	8.1
PMF4.5[e]	17.4	8.5	8.3
BAST[f]	13.6	9.7	9.5

[†] A protein is considered as misclassified if there exists at least one decoy with energy below that of the native structure. The percentage is calculated with respect to the number of proteins in the database (1671). The total number of threaded folds for the proteins of the database is about 3.18×10^8.

[a] Results obtained without solvent-residue contribution.

[b] V_{ex} is the threshold volume used to calculate the solvent exposure (see text for details). All the potentials (PMF7, PMF8, PMF4.5, and BAST) share the calculation method of the solvent-residue contribution.

[c] The residue-residue PMF is calculated using $R_0 = 8$ Å.

[d] The residue-residue PMF is calculated using $R_0 = 7$ Å.

[e] The residue-residue PMF is calculated using the all atom method with a threshold atom-atom cut-off distance of 4.5 Å.

[f] The residue-residue PMF is taken from Ref. 11.

deterministic conformational search.[45] However, in the case of protein folding, these methodologies are computationally demanding and gapless threading for conformational sampling is a useful practical and simple zero-order test for validation of knowledge-based potentials. In the present case, the number of decoys generated by gapless threading is $\sim 3.18 \times 10^8$. The number of threaded structures per protein as a function of the residue number clearly depends on the distribution function reported in Figure 1. In our case, this number obeys to the following phenomenological law: $N_{\text{dec}}(n) = N_1 \exp[(n_1 - n)/n_{av}] + N_2 \exp[-(n_2 - n)^2/\sigma^2]$ with $N_1 = 406398$, $n_1 = 31$, $n_{av} = 220$, $N_2 = 7130$, $n_2 = 1000$, $\sigma = 850$. The evaluation of the solvent-residue map **T** (see Materials and Methods) for the threaded structures requires in principle a Voronoi's analysis, one for each decoy. In practice the solvent exposure has been estimated approximately with a more (computationally) efficient but almost equally accurate procedure. Let us consider a protein J with N residues. Starting from the Voronoi polyhedron partitioning of J, for each residue k belonging to a jth decoy of J, we computed the total area A_{kj} defined by neighbors not belonging to the jth decoy. The kth residue of the jth decoy is assumed to be solvent-exposed when A_{kj} is greater than a given threshold A_T (depending in turn on the volume V_{ex} used for the Voronoi's analysis of J). We find that for $V_{ex} = 350$ Å^3, the optimum value of A_T is 40 Å^2, yielding in 90% of the cases the correct solvent exposure calculated by means of genuine Voronoi's analysis.

In Table II we report the results obtained by using several combinations of contact threshold distances (R_0) and of exposure volumes (V_{ex}). We notice that, when the solvent-residue interactions are included, the impact of R_0 becomes less important. In fact, PMF7 and PMF8 produce essentially the same results on the number of misclassified proteins, i.e., the number of proteins for which there exists at least one decoy with energy below that of the native structure. Hence, the inclusion of the solvent-residue PMF appear to diminish (or modulate) the effect of R_0 in the detection of misclassified proteins. Threading experiments appear to be rather insensitive to V_{ex} as well. The importance of the solvent-residue PMF in stabilizing the protein structure is well evident from Table II: when the solvent-residue PMFs are not included in the calculation, the overall percentage of misclassified proteins increases by at least 1.7% and as much as 9.1% (for PMF8 and PMF4.5, respectively). These increases represent a 1.2- to 2-fold greater occurrence of misclassified proteins without solvent-residue interactions with respect to the cases where they are considered. The adoption of an all-atom criterion for the definition of the PMF with respect to threading can be assessed by inspecting the entries corresponding to PMF4.5 in Table II. PMF4.5 has been calculated using the nearest atom-atom threshold distance set to 4.5 Å (see Ref. 11 for a detailed description of the model). PMF4.5 differ from PMF7 and PMF8 only for the residue-residue part, the solvent-residue part being the same. In spite of the more physically accurate criterion for defining the chemical contact, the residue-residue knowledge-based PMF4.5 produces surprisingly a much larger number of misclassified proteins with respect to the traditional Miyazawa-Jernigan criterion (PMF7 and PMF8). Remarkably, when solvent-residue free energy is included, the number of misclassified proteins obtained with PMF4.5 is strongly reduced (from 17.4% to 8.3% with $V_{ex} = 350$ Å^3), a result comparable to that obtained with PMF7 or PMF8. In Table II we also report, for comparison, the results obtained with the residue-residue PMF proposed by Bastolla et al.[11]

In Table III we report the results of Table II but relating to three different subsets of proteins. These subsets have been selected on the basis of the protein residue number, n_p: for the first subset $n_p \leq 140$, for the second one $140 < n_p \leq 340$, for the third one $n_p > 340$. The first subset contains the largest number of misclassified folds. For example, in the PMF8/350 case (i.e., $R_0 = 8$ Å and $V_{ex} = 350$ Å^3), ~21% of the proteins (namely 116 over 559) have at least one threaded structure with energy below the native state energy. This is hardly surprising as in small proteins the weight of secondary structure and of specific effects may well out-weight that of the contact thermodynamic potential computed over a statistical ensemble. In the intermediate protein subset, for the PMF8/350 potential, we find at most 10 misclassified structures over 647, corresponding to 1.5%: 1ctq, 1c3w, 1ees, 1fqy, 1cf4, 1eai, 3prn, 1ppb, 2por, and 1cip. BAST/350, besides these structures, yields also 1qu0 as a misclassified fold; PMF4.5/350 does not yield 1c3w, 1fqy, 1cip, while misclassifying 1cdy. Actually for all the structures except 1c3w, 1fqy, and 3prn, misclassification is due to only few decoys or even, in most cases, to a single decoy having the primary sequence (up to few point mutations) and fold (up to minor changes in the contact map) equal to that of the misclassified protein. The truly misclassified folds in the intermediate subset, 1c3w, 1fqy, and 3prn, have a percentage of secondary structure well above 90%. We may therefore infer that free energy stabilization in these three misclassified proteins stems from the secondary structure free energy gain. In fact, the large percentage of secondary structure with respect to the

TABLE III. Percentage of Misclassified Proteins for Different Protein Subsets[†]

$n_p \leq 140$; Proteins[a]: 559; Decoys[b]: 1.81×10^8			
	No solvent[c]	$V_{ex}{}^d = 150$ Å^3	$V_{ex}{}^d = 350$ Å^3
PMF8[e]	26.3	20.8	20.8
PMF7[f]	29.0	22.2	22.2
PMF4.5[g]	44.7	23.8	23.1
BAST[h]	37.4	26.7	26.1

$140 < n_p \leq 340$; Proteins[a]: 647; Decoys[b]: 1.11×10^8			
	No solvent[c]	$V_{ex}{}^d = 150$ Å^3	$V_{ex}{}^d = 350$ Å^3
PMF8[e]	1.4	1.3	1.5
PMF7[f]	1.5	1.2	1.4
PMF4.5[g]	6.0	1.2	1.2
BAST[h]	2.8	2.0	1.7

$n_p > 340$; Proteins[a]: 465; Decoys[b]: 0.26×10^8			
	No solvent[c]	$V_{ex}{}^d = 150$ Å^3	$V_{ex}{}^d = 350$ Å^3
PMF8[e]	0	0	0
PMF7[f]	0	0	0
PMF4.5[g]	0	0	0
BAST[h]	0	0	0

[†] A protein is considered as misclassified if there exists at least one decoy with energy below that of the native structure. The (three) protein subsets ($n_p \leq 140$; $140 < n_p \leq 340$; and $n_p > 340$) are chosen on the basis of the number of residues per protein, n_p. The percentage values reported for a given subset are calculated with respect to the number of proteins of that subset.

[a] Number of proteins in the subset.

[b] Number of decoys in the subset.

[c] Results obtained without solvent-residue contribution.

[d] V_{ex} is the threshold volume used to calculate the solvent exposure (see text for details). All the potentials (PMF7, PMF8, PMF4.5, and BAST) share the calculation method of the solvent-residue contribution.

[e] The residue-residue PMF is calculated using $R_0 = 8$ Å.

[f] The residue-residue PMF is calculated using $R_0 = 7$ Å.

[g] The residue-residue PMF is calculated using the all atom method with a threshold atom-atom cut-off distance of 4.5 Å.

[h] The residue-residue PMF is taken from Ref. 11.

mean percentage (40%) results in increasing the complexity of the free energy landscape. This complexity leads to frustration of the system so that the knowledge-based energy function has difficulties in recognizing the native structure.

In the last protein subset (n_p >340) we do not find misclassified folds, probably because of the limited sampling obtained by threading. For example, according to the experimental structure,[44] 1fgj (499 residues) has eight heme groups whose contacts with the surrounding residues must heavily contribute to the stabilization of the fold. In fact, as discussed above in Free Energy of Native States, the free energy for 1fgj is positive, indicating that this fold is unstable as far as residue-residue and residue-solvent are concerned. We hence do expect 1fgj to be misclassified in threading experiment using a simple residue-residue and residue-solvent knowledge-based potential. However, none of the 51253 threaded structures has a free energy value lower than that (positive) of the native state, indicating clearly that the extent of sampling provided by gapless threading is not sufficient in this case.

In order to further quantitatively assess the solvent-residue contribution with respect to discriminating native structures, we have performed threading experiments by neglecting in turn the ILE-LEU and the PHE-PHE PMFs using the PMF8 method without solvent-residue contribution. ILE and LEU are hydrophobic residues and their interaction is unanimously considered very important for the stabilization of the core in globular proteins. PHE-PHE interaction is very common again to connect distinct secondary structure elements in the hydrophobic core (see also discussion in Role of the Aromatic-Aromatic Interaction in Protein Folding). The data are shown in Table IV. For reference, the entries of Tables II and III referring to PMF8/*no solvent* and PMF8/150 are also reported in Table IV. On the overall (see row labeled *All proteins* in Table IV), the effect of neglecting ILE-LEU or PHE-PHE PMF on the percentage of misclassified folds is of the same order of magnitude of that found by including the solvent-residue PMF. As expected, in the intermediate subset ($140 < n_p \leq 340$) the solvent-residue contribution is less important than that given by the ILE-LEU and PHE-PHE interactions because the ratio between the solvent-residue and the residue-residue contacts decreases with increasing protein size (see Materials and Methods).

In Figure 6 we show the area-normalized distributions of the threaded and native structures as a function of the free energy per residue. The results for the PMF8/350 and PMF8/*no solvent* potentials have been reported. The effect of the solvent in widening the energy gap between threaded folds and native structures is evident from the distributions. The stabilization free energy due to the solvent, i.e., the widening of the gap between the main peak of the two distributions, is about 0.16 *RT* per residue. This value comes from a stabilization of 0.06 *RT* of the native folds and a destabilization of 0.10 *RT* of the decoys.

As discussed previously, threading experiments do not give misclassified folds for proteins with a number of residues greater than 340 (see Table III). At first one could explain this finding by considering that the number of threading generated decoys for large proteins is the lowest (0.26×10^8) in spite of the highest number of accessible folds. However, the absence of misclassified proteins cannot be simply attributed only to the poor sampling of the possible contact maps. It can be also attributed to an inherent narrowing of the distributions of the free energy per residue of native and decoy structures as the number of residues approaches to the thermodynamic limit. For the free energy distribution of native structures, this narrowing can be observed in Figure 5 (bottom), where the spread of the free energy per residue decreases with increasing number of residues. In order to clarify this aspect, we have calculated the distributions of the free energy per residue of native folds and decoys for the three protein subsets considered in Table III. Results are shown in Figure 7 (top). Remarkably the average free energy per residue for native folds and decoys is independent on the number of residues, whereas the width of the distributions decreases in going from small to large proteins. The overlap between the native structure and decoy distribu-

TABLE IV. Percentage of Misclassified Proteins by Deleting Selected Pairs in the PMF[†]

	Full residue-residue PMF with solvent[a]	Full residue-residue PMF without solvent[b]	No ILE-LEU without solvent[c]	No PHE-PHE without solvent[d]
All proteins[e]	7.8	9.5	11.2	9.7
$n_p \leq 140$[f]	20.8	26.3	30.9	26.7
$140 < n_p \leq 340$[g]	1.3	1.4	2.0	1.7

[†] A protein is considered as misclassified if there exists at least one decoy with energy below that of the native structure. The percentage is calculated with respect to the number of proteins in the database (1671).

[a] The free energy is calculated using the model PMF8/150 (entries of Tables II and III).

[b] The free energy is calculated using the model PMF8/*no solvent* (entries of Tables II and III).

[c] The free energy is calculated using the model PMF8/*no solvent* by not considering the ILE-LEU residue-residue contribution.

[d] The free energy is calculated using the model PMF8/*no solvent* by not considering the PHE-PHE residue-residue contribution.

[e] The calculation has been performed for all the proteins of the database.

[f] The calculation has been performed for proteins with a number of residues less or equal to 140.

[g] The calculation has been performed for proteins with a number of residues in the range 140–340.

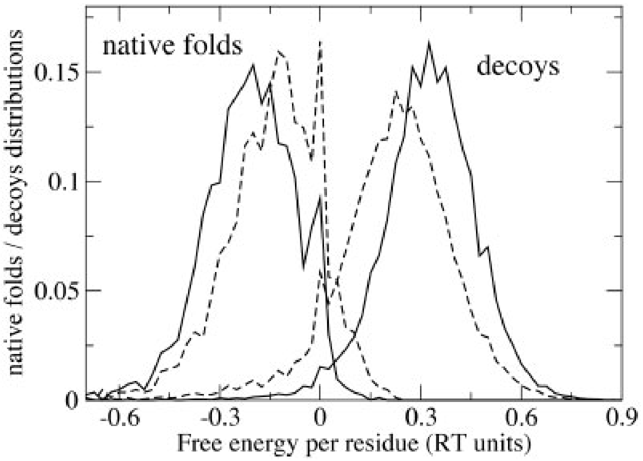

Fig. 6. Area-normalized distributions of the native folds and of decoys as a function of the free energy per residue, obtained from gapless threading experiments. Full line distributions have been obtained using the PMF8/350 potential. Dashed line distributions have been obtained using PMF8/*no solvent* potential.

tions can be highlighted by the product function of the two distributions themselves (bottom part of Fig. 7). The integral of the product function is in turn proportional to the probability of finding misclassified folds by gapless threading (or by any other sampling method) and is the lowest precisely for the large protein subset.

CONCLUSIONS AND PERSPECTIVES

A knowledge-based potential of mean force (PMF) has been evaluated on a large database of protein structures. The PMF has been supplemented with a solvent-residue knowledge-based free energy term calculated by performing Voronoi's polyhedron analysis on the protein database. The obtained PMF shows that amino acid side chains can be classified in three categories: a) hydrophilic residues, i.e., residues that are mostly found surface exposed and bearing large solvent-residue stabilization contribution; b) hydrophobic residues with strong mutual stabilization due to the residue-residue term and having negative solvent-residue stabilization energy; and c) intercalating residues, i.e., residues that may be found with equal probability in the core or on the protein surface. The computed residue-residue PMF shows in general the importance of hydrophobic-hydrophobic interactions in stabilizing protein tertiary structure.

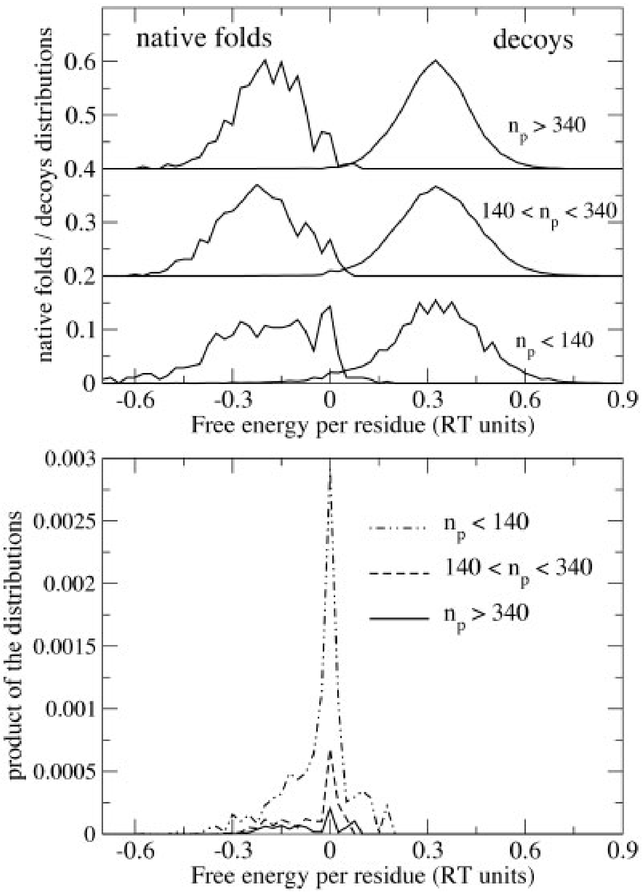

Fig. 7. Top: area-normalized distributions of the native folds and decoys as a function of the free energy per residue, obtained from gapless threading experiments. The distributions relative to the three subsets of proteins of Table III are reported (n_p is the number of residues per protein). The PMF8/350 potential has been used for the calculation. Bottom: product of the distributions shown in the top part. The integral of this product function is proportional to the probability of finding misclassified proteins in the corresponding subset.

We have also observed that aromatic-aromatic contacts occur with greater probability when the interacting residues are distally separated in the primary sequence, thus evidencing the role of the aromatic residues in connecting distinct secondary structure elements. The obtained PMFs predict reasonable values for the free energy of a protein with respect to a completely unfolded state. In this case the solvent-residue PMF is determinant in lowering the number of proteins with positive folding free energy. Gapless threading experiments have been performed for testing PMF. The 1671 structures of the database have been threaded with $\sim 3.18 \times 10^8$ decoys. The performance of the knowledge-based PMF with respect to threading has been evaluated using different definition of the arbitrary parameters of the model, such as contact radius and threshold exposure volume. Variation of these parameters within reasonable ranges does not alter significantly the results. The solvent-residue PMF gives a relevant contribution in stabilizing the native structure with respect to the decoys produced by gapless threading. In fact, the average amount of solvent induced stabilization of native structures with respect to decoys is $\sim 0.16\ RT$ per residue. Misclassified proteins (i.e., structures for which there exists at least one decoy with free energy below the native structure energy) are frequently found only when proteins have a small number of residues. For intermediate and large proteins, misclassification occurs very rarely, either because the threaded map is taken accidentally from a nearly perfect homologous of the native fold or because the native structure is characterized by a large share of secondary structure.

Our results show that the inclusion of the solvent-residue interaction does improve the overall performance of the PMF with respect to discriminating native structures. However, our results show also that knowledge-based pair PMFs with explicit inclusion of the solvent are not able alone to provide a completely reliable tool to discern native folds from decoys even when generated by simple threading. We therefore expect that the use of more efficient sampling methods for generating alternative folds, such as Monte Carlo or molecular dynamics simulations, would further reduce the reliability of the knowledge-based potentials. Other factors should be taken into account in building a potential function for protein folding. For example, from gapless threading experiments, we found that nearly all the medium or large misclassified proteins contain a large amount of secondary structure. This feature suggests that secondary structure formation is an important "structural adjustment process" of a predefined three-dimensional structure coded in the primary sequence. On the basis of the results discussed in the present work, we are studying on the possibility of building a minimal coarse-grained model made of three parts: i) knowledge-based two-body residue-residue interactions, ii) one body solvent-residue interactions, and iii) a *parametric* term driving secondary structure formation, such as the three particle term for optimum packing of thick chains proposed recently by Banavar et al.[29] or the two-body hydrogen bond term proposed by Irbach et al.[46] The parameters entering in this secondary structure term could be *trained* on a learning set of known protein folds so as to yield the native structure as a global minimum. At the present time, as a preliminary approach, we are attempting to include the secondary structure contribution as a *known term.* In particular, starting from an unfolded structure where secondary structure is formed (namely, imposed by geometric constraints on the basis of the known native structure), we would want to investigate on the possibility of obtaining the native structure by using only the residue-residue and residue-solvent contributions to the free energy. Such empirical coarse-grained model should be then tested and validated on the available experimental protein structures.

ACKNOWLEDGMENTS

We are grateful to Tina Trapane for her helpful suggestions in improving the article.

REFERENCES

1. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rogers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
2. Schmitt S, Kuhn D, Klebe G. A new method to detect related function among proteins independent of sequence and fold homology. J Mol Biol 2002;323:387–406.
3. Oldfield TJ. Data mining the Protein Data Bank: residue interactions. Proteins 2002;49:510–528.
4. Mitchell JBO, Smith J. D-amino acid residues in peptides and proteins. Proteins 2003;50:563–571.
5. Dudev T, Lin YL, Dudev M, Lim C. First-second shell interactions in metal binding sites in proteins: a PDB survey and DFT/CDM calculations. J Am Chem Soc 2003;125:3168–3180.
6. Goliaei B, Minuchehr Z. Exceptional pairs of amino acid neighbors in alpha-helices. FEBS Lett 2003;537:121–127.
7. Eyal E, Najmanovich R, Edelman M, Sobolev V. Protein side-chain rearrangement in regions of point mutations. Proteins 2003;50: 272–282.
8. Lu L, Lu H, Skolnick J. MULTIPROSPECTOR: an algorithm for the prediction of protein-protein interactions by multimeric threading. Proteins 2002;49:350–364.
9. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
10. Brocchieri L, Karlin S. How are close residues of protein structures distributed in primary sequence? P Natl Acad Sci USA 1995;92:12136–12140.
11. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79–96.
12. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
13. Vijayakumar M, Zhou HX. Prediction of residue-residue pair frequencies in proteins. J Phys Chem B 2000;104:9755–9764.
14. Burley SK, Petsko GA. Amino-aromatic interactions in proteins. FEBS Lett 1986;203:139–143.
15. Mitchell JBO, Nandi CL, McDonald IK, Thornton JM, Price SL. Amino/aromatic interactions in proteins—is the evidence stacked against hydrogen-bonding. J Mol Biol 1994;239:315–331.
16. Karlin S, Zuker M, Brocchieri L. Measuring residue associations in protein structures—possible implications for protein-folding. J Mol Biol 1994;239:227–248.
17. Brocchieri L, Karlin S. Geometry of interplanar residue contacts in protein structures. P Natl Acad Sci USA 1994;91:9297–9301.
18. McGaughey GB, Gagné M, Rappé AK. Pi-stacking interactions—alive and well in proteins. J Biol Chem 1998;273:15458–15463.
19. Chelli R, Gervasio FL, Procacci P, Schettino V. Stacking and t-shape competition in aromatic-aromatic amino acid interactions. J Am Chem Soc 2002;124:6133–6143.
20. Hao MH, Scheraga HA. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
21. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of

amino-acids—analysis of energy parameter sets. Protein Sci 1995; 4:2107–2117.

22. Thomas PD, Dill KA. Statistical potentials extracted from protein structures: how accurate are they? J Mol Biol 1996;257:457–469.
23. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
24. Zhang C. Extracting contact energies from protein structures: a study using a simplified model. Proteins 1998;31:299–308.
25. Vendruscolo M, Najmanovich R, Domany E. Protein folding in contact map space. Phys Rev Lett 1999;82:656–659.
26. Zhang C, Kim SH. Environment-dependent residue contact energies for proteins. P Natl Acad Sci USA 2000;97:2550–2555.
27. Kolinski A, Galazka W, Skolnick J. Monte Carlo studies of the thermodynamics and kinetics of reduced protein models: application to small helical, beta, and alpha/beta proteins. J Chem Phys 1998;108:2608–2617.
28. Maritan A, Micheletti C, Trovato A, Banavar JR. Optimal shapes of compact strings. Nature 2000;406:287–290.
29. Banavar JR, Maritan A, Micheletti C, Trovato A. Geometry and physics of proteins. Proteins 2002;47:315–322.
30. Holm L, Sander C. Mapping the protein universe. Science 1996;273: 595–602.
31. Fold classification based on Structure-Structure alignment of Proteins (FSSP). ftp://ftp.ebi.ac.uk/pub/databases/fssp. 2002.
32. Voronoi GM. Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire: Recherches sur les parallélloèdres primitifs. J Reine Angew Math 1908;134: 198–287.
33. Procacci P, Scateni R. A general algorithm for computing Voronoi volumes: application to the hydrated crystal myoglobin. Int J Quantum Chem 1992;42:1515–1528.
34. Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl MJ. Identification of native protein folds amongst a large number of incorrect models—the calculation of low-energy conformations from potentials of mean force. J Mol Biol 1990;216:167–180.
35. Kocher JPA, Rooman MJ, Wodak SJ. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. J Mol Biol 1994;235:1598–1613.
36. Vendruscolo M, Najmanovich R, Domany E. Can a pairwise contact potential stabilize native protein folds against decoys obtained by threading? Proteins 2000;38:134–148.
37. Miyazawa S, Jernigan RL. Estimation of effective inter-residue contact energies from protein crystal structures: quasi-chemical approximation. Macromolecules 1985;18:534–552.
38. Mirny L, Domany E. Protein fold recognition and dynamics in the space of contact maps. Proteins 1996;26:391–410.
39. Sippl MJ. Calculation of conformational ensembles from potentials of mean force—an approach to the knowledge-based prediction of local structures in globular-proteins. J Mol Biol 1990;213: 859–883.
40. Chandler D. Introduction to modern statistical mechanics. New York: Oxford University Press; 1987.
41. Park K, Vendruscolo M, Domany E. Towards an energy function for the contact map representation of proteins. Proteins 2000;40: 237–248.
42. Gervasio FL, Chelli R, Procacci P, Schettino V. The nature of intermolecular interactions between aromatic amino acid residues. Proteins 2002;48:117–125.
43. Stryer L. Biochemistry, 5th ed. New York: W.H. Freeman and Company; 1999.
44. Iverson TM, Arciero DM, Hsu BT, Logan MSP, Hooper AB, Rees DC. Heme packing motifs revealed by the crystal structure of the tetra-heme cytochrome c554 from Nitrosomonas europaea. Nat Struct Biol 1998;5:1005–1012.
45. Della Valle RG, Venuti E, Brillante A, Girlando A. Inherent structures of crystalline pentacene. J Chem Phys 2003;118:807–815.
46. Irback A, Samuelsson B, Sjunnesson F, Wallin S. Thermodynamics of alpha- and beta-structure formation in proteins. Biophys J 2003;85:1466–1473.