A New Generation of Statistical Potentials for Proteins

Y. Dehouck, D. Gilis, and M. Rooman
Unité de Bioinformatique génomique et structurale, Université Libre de Bruxelles, 1050 Brussels, Belgium

ABSTRACT We propose a novel and flexible derivation scheme of statistical, database-derived, potentials, which allows one to take simultaneously into account specific correlations between several sequence and structure descriptors. This scheme leads to the decomposition of the total folding free energy of a protein into a sum of lower order terms, thereby giving the possibility to analyze independently each contribution and clarify its significance and importance, to avoid overcounting certain contributions, and to deal more efficiently with the limited size of the database. In addition, this derivation scheme appears as quite general, for many previously developed potentials can be expressed as particular cases of our formalism. We use this formalism as a framework to generate different residue-based energy functions, whose performances are assessed on the basis of their ability to discriminate genuine proteins from decoy models. The optimal potential is generated as a combination of several coupling terms, measuring correlations between residue types, backbone torsion angles, solvent accessibilities, relative positions along the sequence, and interresidue distances. This potential outperforms all tested residue-based potentials, and even several atom-based potentials. Its incorporation in algorithms aiming at predicting protein structure and stability should therefore substantially improve their performances.

INTRODUCTION

Somewhere between the time-consuming semiempirical force fields (1–3) and the oversimplified Gō-like potentials (4–7), statistical energy functions, extracted from databases of known protein structures, are prime tools for the in silico study of proteins (8–12). They present the advantage of being easily adaptable to any level of simplification of protein representation, and have been successfully used in many applications, ranging from structure prediction to sequence design. Though there has been a considerable increase in the number of resolved protein structures since the first approaches of this type were described, no major improvement in predictive power could be drawn from the larger size of the databases (13–15). Indeed, increasing the database size beyond a few hundred proteins appears to yield no significant advantage in the case of the simple potentials that are still very commonly used nowadays, which are based on a limited number of sequence and structure descriptors.

In the last few years, a number of more complex potentials have been designed with the aim of exploiting more efficiently the large amount of available structural data and dealing with couplings between different structural features. Among those, let us cite distance or contact potentials that depend on the solvent accessibility of the residues (16,17), on the conformation of their main chain (18), or on the relative orientation of their side chains (19–21). On the other hand, potentials describing the propensities of the different amino acid types to adopt certain backbone conformations, which simultaneously take into account the nature and/or conformation of several neighboring residues, have also been developed (16,22,23). A major difficulty that frequently arises in such studies is related to the fact that the number of proteins in the database becomes rapidly too small when increasing the complexity of a potential. One faces a delicate choice: the use of a more complex potential can be quite advantageous for common values of the sequence and structure descriptors (e.g., Ala-Ala pair associated with α-helical conformations), and pretty disastrous in other cases (e.g., Trp-Trp pair associated with some rare turn conformations). The usual answer to this dilemma consists in drastic limitations of the description of the conformational space, for example by restricting the backbone to three possible conformations, the solvent accessibility to two different bins, or by deriving contact potentials rather than distance-dependent ones.

We present here a general derivation scheme that allows one to bypass this issue, and to build statistical energy functions based simultaneously on several sequence and structure descriptors without altering the efficiency of the elementary contributions when the values taken by these descriptors are not frequent enough in the database of known protein structures. We apply our procedure to generate statistical potentials based on the correlations among amino acid types, backbone conformations, and solvent accessibilities of residues close to each other in the sequence and/or in space. The resulting energy function displays a strongly improved ability to discriminate genuine proteins from decoy models. All potentials presented in this article are freely available at http://babylone.ulb.ac.be/StatPots.

METHODS

Sequence and structure descriptors

The backbone conformation of the residue at position i, t_i, is defined by the values of the torsion angles (ϕ,φ,ω). These values are grouped in seven

Submitted December 9, 2005, and accepted for publication February 28, 2006.

Address reprint requests to Yves Dehouck, E-mail: ydehouck@ulb.ac.be.

0006-3495/06/06/4010/08 $2.00

doi: 10.1529/biophysj.105.079434

domains corresponding to distinct regions on the Ramachandran map (22,24). The solvent accessibility of the residue at position i, a_i, is defined as the ratio of its solvent-accessible surface in the considered structure (as computed by DSSP (25)) and in an extended tripeptide Gly-X-Gly (26). These values are grouped in five discrete domains: $a_i \leq 5\%$, $5\% < a_i \leq 15\%$, $15\% < a_i \leq 30\%$, $30\% < a_i \leq 50\%$, and $50\% < a_i$. The interresidue distance d_{ij} is computed between the average side-chain centroids, noted C^{μ}, of the residues at positions i and j. The C^{μ} corresponds to the geometric center of heavy side-chain atoms of a given amino acid type, averaged over all side-chain conformations in a data set of known structures (16). The distances d_{ij} between 3 Å and 8 Å are grouped into 25 bins of 0.2 Å width; two additional bins describe distances smaller than 3 Å and larger than 8 Å, respectively. Finally, the sequence descriptor s_i corresponds to the nature (1 of 20 amino acids) of the residue at position i.

Protein structure data set

An initial set of 1522 high-resolution (≤2 Å) x-ray structures of protein chains with <20% pairwise sequence identity was extracted in October 2003 from the website ''Culling the PDB by Resolution and Sequence Identity'' (27) (http://dunbrack.fccc.edu/Guoli/pisces_download.php). All structures containing more than 5% heteroatoms or nonnatural residues were excluded. This led to a final set of 1403 protein chains. Furthermore, to ensure that the data set used to derive the potentials includes the proper, active, quaternary conformations of the selected proteins, the coordinates were taken from the ''Protein Quaternary Structure'' server (28) (http://pqs.ebi.ac.uk).

Correction for sparse data

All database-derived potentials and coupling terms presented here can be generically written as $\Delta W = -kT \ln (n^{obs}/n^{exp})$, where n^{obs} is the number of observations of a given association of sequence and structure descriptors in the data set of known protein structures, and n^{exp} is the corresponding number expected in a reference state. To deal with the limited size of the data set, a correction for sparse data (29) is applied: $(n^{obs}/n^{exp}) \rightarrow ((\sigma + n^{obs})/(\sigma + n^{exp}))$, where σ is an adjustable parameter, taken equal to 20 for local potentials, and 10 for distance potentials (see Results for the definition of local and distance potentials). This correction ensures that the potentials tend to 0 when the number of observations in the data set is too small.

Decoy sets

To assess the performances of the potentials, we evaluate their ability of singling out correct sequence-structure matches out of sets of decoy models. Three groups of decoys sets are considered. The first, noted D^1_{str}, includes 25 proteins (30,31), each associated with hundreds of alternative structures generated by different modeling methods (4state_reduced (32): 1ctf, 1r69, 1sn3, 2cro, 4pti and 4rxn ; fisa (33): 1fc2-c, 1hdd-c, 2cro ; fisa_casp3 (33): 1bg8-a, 1bl0, 1jwe ; lattice-ssfit (31): 1ctf, 1dkt-a, 1fca, 1nlk, 1pgb, 1trl-a ; lmds (34): 1ctf, 1dtk, 1fc2-c, 1igd, 1shf-a, 2cro, 2ovo). The second group, noted D^2_{str}, includes 25 proteins (35), each associated with ~2000 alternative structures generated by the Rosetta structure prediction method (1a32, 1ail, 1am3, 1cc5, 1cei, 1hyp, 1flb, 1mzm, 1r69, 1utg, 1ctf, 1dol, 1orc, 1pgx, 1ptq, 1tif, 1vcc, 2fxb, 5icb, 1bq9, 1csp, 1msi, 1tuc, 1vif, 5pti). The third group, noted D_{seq}, includes 50 proteins (1ptq, 1d0d, 2igd, 1g2b, 1orc, 1hz6, 1i27, 1hoe, 1luz, 1ugi, 1aba, 1cy5, 1lpl, 1mk0, 1h7m, 1bm8, 1l8r, 1lyq, 1o13, 1gmx, 1cew, 1hxi, 1nyc, 1by2, 1lsl, 1o7i, 1gnu, 1fc3, 1mai, 1dzo, 1lwb, 1huf, 1nwz, 3nul, 1cuo, 1jf8, 1p0z, 1mdc, 1vsr, 1gmi, 1eca, 1j9b, 1kmt, 1mzg, 1oz9, 1h6h, 1l2h, 1srv, 2hbg, 1amx), each associated with 1000 decoys obtained by maintaining the structure and randomizing the amino acid sequence with fixed amino acid composition. To render the test more challenging, only a fraction of the sequence was modified. This fraction was chosen randomly between 25% and 100%, independently for each decoy.

To avoid any bias toward the native structure or wild-type sequence that might result from the presence of similar proteins in the data set, an extended jackknife procedure is applied: we remove the target protein, as well as all proteins sharing more than 20% sequence identity with the target, from the database before deriving the potentials.

Performance measures

We use five different measures to evaluate the ability of the potentials to discriminate the native structure from the decoys:

1. The success rate S_1 is the percentage of proteins, in each group of decoys, for which the free energy of the correct sequence-structure association is smaller than the free energies computed for all decoys.
2. $\langle Z \rangle$ is the average Z-score, over all proteins in a group of decoys. The Z-score is defined as $Z = (\Delta W_c - \langle \Delta W \rangle)/\sigma_{\Delta W}$, where ΔW_c is the free energy of the correct sequence-structure association, $\langle \Delta W \rangle$ is the average free energy of all sequence-structure associations, and $\sigma_{\Delta W}$ is the associated standard deviation. Energy functions discriminating well the genuine protein from the decoys are characterized by a very negative Z-score.
3. S_{-1} is the percentage of proteins with a Z-score lower than −1 (19). This measure may be more useful than S_1 when the test is challenging, for instance when the decoys and the native structures or sequences are very similar.
4. $\langle Z^x \rangle$ evaluates the ability of the potentials to select the decoys that are closest from the native among the complete decoy set. Z^x is defined as $(\langle \Delta W \rangle_{5\%} - \langle \Delta W \rangle)/\sigma_{\Delta W}$, where $\langle \Delta W \rangle_{5\%}$ is the average free energy computed on a subset including 5% of the decoys (19). This subset contains the decoys with the lowest root mean-square deviation from the native structure, or the decoys with the largest sequence identity with the wild-type in the case of decoys generated by sequence randomization.
5. S^x_{-1} is equal to the percentage of proteins for which Z^x is lower than −1 (19).

RESULTS

General derivation scheme

A form commonly used for statistical potentials derived from a set of protein structures is

$$\Delta W(c_1, c_2) = -kT \log \frac{P(c_1, c_2)}{P(c_1)P(c_2)}, \tag{1}$$

where c_1 is an amino acid type and c_2 a structure descriptor (e.g., a torsion angle or solvent accessibility domain) of the same or a neighboring residue, and P are their relative frequencies of occurrence in the structure data set. Similarly, considering two sequence descriptors c_1 and c_2 and one structure descriptor c_3, we have

$$\Delta W(c_1, c_2, c_3) = -kT \log \frac{P(c_1, c_2, c_3)}{P(c_1)P(c_2)P(c_3)}, \tag{2}$$

where, for example, c_1 and c_2 are amino acid types at positions i and j along the sequence and c_3 is the spatial distance between them.

This form can easily be generalized. First, c_1, c_2, and c_3 can be any sequence or structure descriptor. For example, all three can correspond to torsion angle domains, or c_1 can correspond to an amino acid type, c_2 to a solvent accessibility domain, and c_3 to a torsion angle domain. A second way to generalize this form is to consider higher order potentials involving n sequence and structure descriptors. We then get

$$\Delta W(c_1, c_2, \ldots, c_n) = -kT \log\frac{P(c_1, c_2, \ldots, c_n)}{P(c_1)P(c_2)\ldots P(c_n)}. \qquad (3)$$

Increasing n reduces the number of observations of each combination of the c_i's in the data set and the statistical significance of the frequencies $P(c_1,c_2,\ldots,c_n)$. When the number of observations is too small, the correction for sparse data (see Methods) becomes important and the potential tends to zero, leading to a complete loss of information. A straightforward solution to this problem involves decomposing the potential into different coupling terms $\Delta\tilde{W}$, and applying the correction for sparse data to each of them separately. In particular, for $n = 3$:

$$\begin{aligned}\Delta W(c_1, c_2, c_3) = {} & \Delta\tilde{W}(c_1, c_2) + \Delta\tilde{W}(c_2, c_3) + \Delta\tilde{W}(c_3, c_1) \\ & + \Delta\tilde{W}(c_1, c_2, c_3),\end{aligned} \qquad (4)$$

where the $n = 2$ coupling terms coincide with the ordinary potentials $\Delta\tilde{W}(c_1,c_2) = \Delta W(c_1,c_2)$, and the $n = 3$ coupling term is defined as

$$\Delta\tilde{W}(c_1, c_2, c_3) = -kT \log\frac{P(c_1, c_2, c_3) \quad P(c_1)P(c_2)P(c_3)}{P(c_1, c_2)P(c_2, c_3)P(c_3, c_1)}. \qquad (5)$$

This $n = 3$ coupling term measures the correlation between the three sequence and structure descriptors c_1, c_2, and c_3, independently of the correlations between c_1 and c_2, c_2 and c_3, and c_3 and c_1. More generally, we can define n-potentials ΔW in terms of all $k \leq n$ coupling terms $\Delta\tilde{W}$:

$$\Delta W(c_1, c_2, \ldots, c_n) = \sum_{k=2}^{n} \sum_{\substack{i_1,\ldots i_k=1 \\ i_1<\ldots<i_k}}^{n} \Delta\tilde{W}(c_{i_1}, c_{i_2}, \ldots, c_{i_k}), \qquad (6)$$

where the n-coupling terms describing correlations between n descriptors are defined as

$$\Delta\tilde{W}(c_1, c_2, \ldots, c_n) = - kT \log \prod_{k=n,n-2,n-4,\ldots} \frac{\prod\limits_{\substack{i_1,\ldots,i_k=1 \\ i_1<\ldots<i_k}}^{n} P(c_{i_1}, c_{i_2}, \ldots, c_{i_k})}{\prod\limits_{\substack{i_1,\ldots,i_{k-1}=1 \\ i_1<\ldots<i_{k-1}}}^{n} P(c_{i_1}, c_{i_2}, \ldots, c_{i_{k-1}})}. \qquad (7)$$

To ensure that each contribution is counted only once, the total free energy of a protein of sequence S and structure C, $\Delta W(C,S)$, is defined as the sum of the total contributions of all coupling terms of order $k \leq n$:

$$\Delta W(C, S) = \sum_{k=2}^{n} \sum_{\substack{i_1,\ldots i_k=1 \\ i_1<\ldots<i_k}}^{n} \sum_{(c_{i_1}, c_{i_2}, \ldots, c_{i_k}) \subset (C,S)} \Delta\tilde{W}(c_{i_1}, c_{i_2}, \ldots, c_{i_k}), \qquad (8)$$

where the third sum goes over all combinations of the $(c_{i_1},c_{i_2},\ldots,c_{i_k})$ descriptors present in the protein. The value chosen for n depends on the structural descriptors and the level of detail that one wishes to take into account, and also on the limitations arising from the finite size of the database.

Note that it is not always necessary or advantageous to fully decompose the potential functions like in Eqs. 4 and 6. In particular, the coupling terms of the type $\Delta\tilde{W}(s_1,s_2)$, with s_1 and s_2 being single residues, may reasonably be overlooked. For example, a relevant and commonly used distance potential $\Delta W'$ (s_1,s_2,d_{12}) may be defined as

$$\begin{aligned}\Delta W'(s_1, s_2, d_{12}) &= \Delta W(s_1, s_2, d_{12}) - \Delta\tilde{W}(s_1, s_2) \\ &= -kT \log\frac{P(s_1, s_2, d_{12})}{P(s_1, s_2)\,P(d_{12})}.\end{aligned} \qquad (9)$$

More generally, we denote by $\Delta W'$ potentials comprising only some of the couplings included in ΔW.

Local potentials and couplings

A first application of our general derivation scheme consists in defining local potentials reflecting the correlations among characteristics of residues that are close to each other along the sequence. We focus here on three different residue characteristics: its type s, its backbone conformation t, and its solvent accessibility a (see Methods).

Among the local $n = 2$ coupling terms of the type $\Delta\tilde{W}(c_1,c_2)$ defined in Eqs. 1 and 7, let us consider first $\Delta\tilde{W}_{ts}(t_i,s_j)$, where c_1 is taken to be the backbone conformation of the residue at position i (t_i) and c_2 the type of the residue at position j (s_j). We assume that this effective energy depends only on the relative positions of the residues along the sequence (i–j), and not on the precise positions i and j. The total free energy of a given sequence S in a structure C, according to this potential, is computed by summing $\Delta\tilde{W}_{ts}(t_i,s_j)$ over all pairs of positions i and j in S that satisfy the condition $|i–j| \leq F_{LOC}$, where F_{LOC} is an adjustable parameter taken here equal to 2. This energy function is similar to previously described backbone torsion potentials (16,22,23,36). We also compute all other $n = 2$ coupling terms (except $\Delta\tilde{W}_{ss}(s_i,s_j)$, which depends only on the sequence), i.e., $\Delta\tilde{W}_{as}(a_i,s_j)$, $\Delta\tilde{W}_{at}(a_i,t_j)$, $\Delta\tilde{W}_{aa}(a_i,a_j)$ and $\Delta\tilde{W}_{tt}(t_i,t_j)$. Note that when c_1 and c_2 correspond to the same structure or sequence descriptor, the condition $|i–j| \leq F_{LOC}$ becomes $1 \leq i–j \leq F_{LOC}$.

We would like to stress that summing the energy contributions of all pairs (c_1,c_2) yields only an approximation of the total free energy of a protein. Indeed, the contributions $\Delta\tilde{W}_{ts}(t_i,s_j)$ and $\Delta\tilde{W}_{ts}(t_i,s_k)$ are in general not independent. Moreover, using simultaneously $\Delta\tilde{W}_{ts}(t_i,s_j)$ and $\Delta\tilde{W}_{as}(a_i,s_j)$ can be advantageous but introduces some redundancy since the solvent accessibility of a residue is related to its backbone conformation. To overcome these dependencies, we must add the $n = 3$ coupling terms $\Delta\tilde{W}_{tts}(t_i,t_j,s_k)$, $\Delta\tilde{W}_{tss}(t_i,s_j,s_k)$, $\Delta\tilde{W}_{ttt}(t_i,t_j,t_k)$, $\Delta\tilde{W}_{aas}(a_i,a_j,s_k)$, $\Delta\tilde{W}_{ass}(a_i,s_j,s_k)$, $\Delta\tilde{W}_{aaa}(a_i,a_j,a_k)$, $\Delta\tilde{W}_{aat}(a_i,a_j,t_k)$, $\Delta\tilde{W}_{att}(a_i,t_j,t_k)$ and $\Delta\tilde{W}_{ats}(a_i,t_j,s_k)$. They are defined on the basis of Eq. 5 so as to be additive to, and exclusive of, the lower order coupling terms (Eq. 4). The interdependence of the different $n = 3$ coupling terms can, in turn, be corrected by the use of $n = 4$ coupling terms.

We assessed the predictive power of the different $n = (2,3,4)$ coupling terms, independently and in combination, on the three groups of decoy sets described in Methods. The performance measures obtained are given in Table 1 for the basic potentials $\Delta\tilde{W}_{\rm ts}$ and $\Delta\tilde{W}_{\rm as}$ and for the most efficient linear combination of the local coupling terms, named $\Delta W'_{\rm LOC}$:

$$\begin{aligned}\Delta W'_{\rm LOC} &= \Delta\tilde{W}_{\rm ts} + \Delta\tilde{W}_{\rm tts} + \Delta\tilde{W}_{\rm tss} + \Delta\tilde{W}_{\rm ttts} + \Delta\tilde{W}_{\rm as} + \Delta\tilde{W}_{\rm aas}\\ &\quad + \Delta\tilde{W}_{\rm ass} + \Delta\tilde{W}_{\rm aaas} + \Delta\tilde{W}_{\rm ats} + \frac{1}{2}(\Delta\tilde{W}_{\rm at} + \Delta\tilde{W}_{\rm aat} + \Delta\tilde{W}_{\rm att} + \Delta\tilde{W}_{\rm aaat}\\ &\quad + \Delta\tilde{W}_{\rm aatt} + \Delta\tilde{W}_{\rm attt} + \Delta\tilde{W}_{\rm tt} + \Delta\tilde{W}_{\rm ttt} + \Delta\tilde{W}_{\rm aaa}).\end{aligned} \tag{10}$$

Overall, the predictive power of $\Delta W'_{\rm LOC}$ is quite impressive: each performance measure indicates a markedly better discrimination of the correct sequence-structure association than with the basic potentials. The only exception is $S^{\rm x}_{-1}$, which slightly decreases in the $D_{\rm seq}$ set.

Strikingly, $\Delta W'_{\rm LOC}$ includes almost all $n = 2$ and $n = 3$ coupling terms. The only exception is $\Delta\tilde{W}_{\rm aa}$, which systematically drags down the predictive power when included in a combination of coupling terms. This follows from the fact that $\Delta\tilde{W}_{\rm aa}$ strongly favors situations in which residues close to each other in the sequence have similar solvent accessibilities, and therefore awards very negative energies to (partially) unfolded proteins. The best combination incorporates also several $n = 4$ coupling terms: $\Delta\tilde{W}_{\rm ttts}$, $\Delta\tilde{W}_{\rm aaas}$, $\Delta\tilde{W}_{\rm attt}$, $\Delta\tilde{W}_{\rm aatt}$, and $\Delta\tilde{W}_{\rm aaat}$. The other $n = 4$ coupling terms have a negative impact on the predictive power. This is most probably due to the limited size of the data set, which does not allow one to compute precisely enough the probabilities of observing simultaneously four sequence and/or structure descriptors. Also note that there are 20 types of sequence elements (s), whereas only 7 torsion (t) and 5 accessibility (a) domains. Coupling terms involving several sequence elements, such as $\Delta\tilde{W}_{\rm tsss}$ or $\Delta\tilde{W}_{\rm asss}$, do not appear in $\Delta W'_{\rm LOC}$ as they require larger data sets to extract reliable statistics.

TABLE 1 Performances of local and distance potentials and couplings

	Potential	$\langle Z\rangle$	S_1	S_{-1}	$\langle Z^{\rm x}\rangle$	$S^{\rm x}_{-1}$
$D^1_{\rm str}$	$\Delta\tilde{W}_{\rm ts}$	−2.69	40%	80%	−0.34	4%
	$\Delta\tilde{W}_{\rm as}$	−2.40	44%	80%	−0.45	16%
	$\Delta\tilde{W}_{\rm ts} + \Delta\tilde{W}_{\rm as}$	−3.44	64%	88%	−0.53	24%
	$\Delta W'_{\rm LOC}$	−4.16	76%	92%	−0.57	28%
	$\Delta\tilde{W}_{\rm sd} + \Delta\tilde{W}_{\rm sds}$	−3.27	72%	84%	−0.66	28%
	$\Delta W'_{\rm DIST}$	−4.65	80%	88%	−0.73	28%
	$\Delta W'_{\rm LOC} + \Delta W'_{\rm DIST}$	−5.25	84%	88%	−0.79	36%
$D^2_{\rm str}$	$\Delta\tilde{W}_{\rm ts}$	−1.45	8%	68%	−0.27	0%
	$\Delta\tilde{W}_{\rm as}$	−0.60	0%	44%	−0.26	0%
	$\Delta\tilde{W}_{\rm ts} + \Delta\tilde{W}_{\rm as}$	−1.84	20%	72%	−0.41	0%
	$\Delta W'_{\rm LOC}$	−2.06	20%	88%	−0.49	12%
	$\Delta\tilde{W}_{\rm sd} + \Delta\tilde{W}_{\rm sds}$	−1.80	16%	76%	−0.33	0%
	$\Delta W'_{\rm DIST}$	−2.32	28%	88%	−0.50	12%
	$\Delta W'_{\rm LOC} + \Delta W'_{\rm DIST}$	−2.65	36%	92%	−0.59	24%
$D_{\rm seq}$	$\Delta\tilde{W}_{\rm ts}$	−2.21	22%	100%	−1.54	100%
	$\Delta\tilde{W}_{\rm as}$	−2.29	50%	100%	−1.58	96%
	$\Delta\tilde{W}_{\rm ts} + \Delta\tilde{W}_{as}$	−2.22	26%	100%	−1.54	100%
	$\Delta W'_{\rm LOC}$	−2.57	80%	100%	−1.71	98%
	$\Delta\tilde{W}_{\rm sd} + \Delta\tilde{W}_{\rm sds}$	−2.75	64%	100%	−1.90	100%
	$\Delta W'_{\rm DIST}$	−2.64	48%	100%	−1.81	100%
	$\Delta W'_{\rm LOC} + \Delta W'_{\rm DIST}$	−2.74	84%	100%	−1.87	100%

The predictive power of the basic potentials and of the different combinations of coupling terms is evaluated on three groups of decoy sets, with five different measures (see Methods). The sequence-independent terms are not taken into account when $D_{\rm seq}$ is considered.

In principle, our derivation scheme does not give any reason to under- or overweight some coupling terms with respect to others. However, some contributions may be less/not relevant and should therefore not be included, for example because of the limited size of the data set (e.g., $\Delta\tilde{W}_{\rm tsss}$, $\Delta\tilde{W}_{\rm asss}$,. . .), the overstabilization of the unfolded state (e.g., $\Delta\tilde{W}_{\rm aa}$), or the uselessness of purely sequence terms (e.g., $\Delta\tilde{W}_{\rm ss}$). Furthermore, sequence-independent terms can be expected to yield interesting results when discriminating among nonprotein-like structures, and to be quite useless in applications such as threading experiments. Testing the potentials on decoy sets can reasonably well be considered as an intermediate case, which probably explains why we observed that underweighting these contributions by a ½ factor, in Eq. 10, is advantageous in terms of predictive power.

Distance potentials and couplings

A very popular category of statistical potentials is derived from the spatial distance distribution between residue types (e.g., 16,17,29,37). They are complementary to the local potentials presented above. It has been previously noted that such potentials do not represent the ''true'' energy of interaction between two residues (or two atoms) as if they where in a vacuum, but rather an effective energy including the influence of a mean protein and solvent environment (38,39). As a consequence, these potentials may depend on some characteristics of the proteins from which they are derived, such as their size (40–42) or their content in secondary structures (14,42–44). The idea of being more precise on the definition of the environment that is actually ''felt'' by the two interacting residues is not new (16–18), and can have a positive impact on the performances of the potentials. We show that the formalism presented in this article can be applied to define residue pair distance potentials that take appropriately into account the influence of the specific environment in which the two residues are located. This environment is here represented by backbone conformations and solvent accessibilities.

The $n = 2$ coupling term $\Delta\tilde{W}_{\rm sd}(s_{\rm i},d_{\rm ij})$ is a ''one-body'' distance potential that reflects the preferences of each type of residue to be located more or less close to other residues, whatever their type, and is therefore dominated by the hydrophobic effect. For residues close to each other along the sequence, i.e., $|i–j| \leq F_{\mathrm{DIS}}$ (taken here equal to 8), the frequencies and potentials are computed separately, whereas they are merged in a single class when $|i–j| > F_{\mathrm{DIS}}$. The total contribution to the free energy of a given sequence S in a structure C is computed by summing $\Delta\tilde{W}_{\mathrm{sd}}(s_{\mathrm{i}},d_{\mathrm{ij}})$ over all pairs of positions i and j in S that satisfy the condition $|i–j| > 1$.

On its own, $\Delta\tilde{W}_{\mathrm{sds}}(s_{\mathrm{i}},d_{\mathrm{ij}},s_{\mathrm{j}})$ is a two-body distance potential that excludes the one-body contributions reflecting the individual preferences of the two amino acids s_{i} and s_{j}. Such a potential has been presented previously and shown to describe more accurately the electrostatic interactions (42). In this case, by reason of symmetry, the condition $|i–j| > 1$ becomes $i–j > 1$ when computing the total free energy of a protein. Coupling $\Delta\tilde{W}_{\mathrm{sd}}(s_{\mathrm{i}},d_{\mathrm{ij}})$ with $\Delta\tilde{W}_{\mathrm{sds}}(s_{\mathrm{i}},d_{\mathrm{ij}},s_{\mathrm{j}})$ yields the common distance potential given in Eq. 9.

In a similar way, it is possible to define sequence-independent distance potentials involving the backbone torsion angles, $\Delta\tilde{W}_{\mathrm{td}}$ and $\Delta\tilde{W}_{\mathrm{tdt}}$, or the solvent accessibilities, $\Delta\tilde{W}_{\mathrm{ad}}$ and $\Delta\tilde{W}_{\mathrm{ada}}$. The concomitant use of these three types of potentials is hazardous since the backbone conformation and solvent accessibility of a residue are clearly dependent on its amino acid type, and some contributions are therefore overcounted. To deal with this problem, we have to define higher order coupling terms. The highest order coupling term is in this case the $n = 7$ term $\Delta\tilde{W}_{\mathrm{atsdats}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$. Considering all the lower level coupling terms would lead to a very large number of energetic functions and hamper any intuitive understanding of their significance. Among these, we choose to disregard all distance-independent terms, as they are redundant with the local potentials defined in the previous section for $|i–j| \leq F_{\mathrm{LOC}}$, and the contributions for other i and j may reasonably be assumed to be negligible. Moreover, to avoid overloading the notations, two-body asymmetrical terms, such as $\Delta\tilde{W}_{\mathrm{ads}}(a_{\mathrm{i}},d_{\mathrm{ij}},s_{\mathrm{j}})$ or $\Delta\tilde{W}_{\mathrm{asds}}(a_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},s_{\mathrm{j}})$, are not considered independently but grouped with the closest symmetrical coupling term, here $\Delta\tilde{W}_{\mathrm{asdas}}(a_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},s_{\mathrm{j}})$. We thus define $\Delta\hat{W}_{\mathrm{asdas}}(a_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},s_{\mathrm{j}})$ as the sum of $\Delta\tilde{W}_{\mathrm{asdas}}(a_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},s_{\mathrm{j}})$ and all the lower order asymmetrical two-body terms. Note finally that, given the limited size of the database, $\Delta\tilde{W}_{\mathrm{atsd}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}})$ and $\Delta\hat{W}_{\mathrm{atsdats}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$ are computed as contact potentials, where d_{ij} takes only two possible values: lower or larger than 8 Å.

Overall, according to our performance test on the three groups of decoy sets, the best combination of distance potentials and coupling terms is $\Delta W'_{\mathrm{DIST}}$, defined as

$$\begin{aligned}\Delta W'_{\mathrm{DIST}} = {} & \Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{sds}} + \Delta\tilde{W}_{\mathrm{td}} + \Delta\tilde{W}_{\mathrm{tdt}} + \Delta\tilde{W}_{\mathrm{ad(SR)}} \\ & + \Delta\tilde{W}_{\mathrm{tsd}} + \Delta\hat{W}_{\mathrm{tsdts}} + \Delta\tilde{W}_{\mathrm{asd(SR)}} + \Delta\hat{W}_{\mathrm{asdas}} \\ & + \Delta\tilde{W}_{\mathrm{atd}} + \Delta\hat{W}_{\mathrm{atdat}} + \Delta\tilde{W}_{\mathrm{atsd}}, \end{aligned} \tag{11}$$

where the terms $\Delta\tilde{W}_{\mathrm{ad}}$ and $\Delta\tilde{W}_{\mathrm{asd}}$ are only included for short-range interactions (SR), that is, when the considered residues are separated by no more than F_{DIST} positions along the sequence. As shown in Table 1, the improvement of the predictive power with respect to the basic distance potential $\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{sds}}$ is substantial in the two decoy sets based on structural modifications (D^1_{str} and D^2_{str}). However, it appears that $\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{sds}}$ performs slightly better than $\Delta W'_{\mathrm{DIST}}$ in the third decoy set. Since these decoys are obtained by modifications of the sequence, the sequence-independent terms ($\Delta\tilde{W}_{\mathrm{td}}$, $\Delta\tilde{W}_{\mathrm{tdt}}$, $\Delta\tilde{W}_{\mathrm{ad}}$,...) are not taken into account in the evaluation of the energies, which may limit the necessity of using coupling terms such as $\Delta\tilde{W}_{\mathrm{tsd}}$, $\Delta\tilde{W}_{\mathrm{asd}}$, or $\Delta\tilde{W}_{\mathrm{tsdts}}$.

Interestingly, as with $\Delta W'_{\mathrm{LOC}}$, almost all coupling terms are included in the best performing combination, $\Delta W'_{\mathrm{DIST}}$. This provides a strong support to the legitimacy of our derivation procedure. The only exceptions are $\Delta\tilde{W}_{\mathrm{ada}}$ and $\Delta\hat{W}_{\mathrm{atsdats}}$. The former strongly favors situations where residues close in space have similar solvent accessibilities, which is a characteristic of both folded and unfolded states. The relevance of the latter is obviously compromised by the limited size of the data set. On the other hand, the terms $\Delta\tilde{W}_{\mathrm{ad}}$ and $\Delta\tilde{W}_{\mathrm{asd}}$ are only included for short-range interactions. Indeed, for long-range interactions, the separation in sequence is not explicitly taken into account, and $\Delta\tilde{W}_{\mathrm{ad}}$ merely reflects a trivial correlation: residues with a higher solvent accessibility have fewer contacts with other residues. For those residue pairs that do not benefit from the $\Delta\tilde{W}_{\mathrm{ad}}$ term, it also appears that $\Delta\tilde{W}_{\mathrm{asd}}$ is unnecessary, as its aim is to uncouple $\Delta\tilde{W}_{\mathrm{ad}}$ and $\Delta\tilde{W}_{\mathrm{sd}}$.

Combination of local and distance potentials

The combination of the best performing local and distance potentials, $\Delta W'_{\mathrm{LOC}}$ and $\Delta W'_{\mathrm{DIST}}$, improves their individual scores, as seen in Table 1. We did not address explicitly the issue of possible redundancies between these two types of potentials. However, in itself, the use of distance coupling terms significantly limits this problem. For example, a relatively strong correlation is observed between $\Delta\tilde{W}_{\mathrm{as}}$ and $\Delta\tilde{W}_{\mathrm{sd}}$, but $\Delta\tilde{W}_{\mathrm{as}}$ and ($\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{asd}}$) are only weakly correlated. Overall, the performances of the combination $\Delta W'_{\mathrm{LOC}} + \Delta W'_{\mathrm{DIST}}$ are very impressive, as exemplified by average Z-scores of -5.25, -2.65, and -2.74, on the three groups of decoy sets.

Comparison with other statistical potentials

A large number of knowledge-based potentials reflecting the preferences of the different amino acids (or of short stretches of amino acids) to adopt particular local conformations (16,22,23,36), to be more or less accessible to the solvent (16,17,45,46), or to be separated by a given spatial distance (16,17,29,30,37) have been described in the literature. However, to our knowledge, our approach is the first to integrate all these different types of contributions in a single energetic function while taking special care of their couplings. Moreover, on the local level, the nonadditivity of contributions related to pairs of residues, such as $\Delta\tilde{W}_{\mathrm{ts}}(t_{\mathrm{i}},s_{\mathrm{j}})$

and $\Delta\tilde{W}_{\mathrm{ts}}(t_{\mathrm{i}},s_{\mathrm{k}})$, is taken care of by the use of higher order coupling terms ($\Delta\tilde{W}_{\mathrm{tss}}(t_{\mathrm{i}},s_{\mathrm{j}},s_{\mathrm{k}})$, $\Delta\tilde{W}_{\mathrm{tts}}(t_{\mathrm{i}},t_{\mathrm{j}},s_{\mathrm{k}})$,. . .).

Among the local potentials based on backbone torsion angles that have been described earlier, let us cite the residue-to-torsion (22) and the torsion-to-residue (16) potentials, developed by one of us. As seen in Table 2, (a) and (b), both potentials can be expressed as simple combinations of the coupling terms $\Delta\tilde{W}_{\mathrm{ts}}$, $\Delta\tilde{W}_{\mathrm{tss}}$, and $\Delta\tilde{W}_{\mathrm{tts}}$. Miyazawa and Jernigan designed a more complex torsion potential (23), based on a reference state that is quite different from ours and on different values of the structural descriptors. A rigorous comparison of the two approaches is therefore difficult. However, a common feature is the expression of the energetic function as a sum of basic potentials and of higher order coupling terms defined so as to exclude the more basic contributions. In this sense, their potential can be compared to the combination of coupling terms $\Delta\tilde{W}$ given in Table 2 (c).

Most commonly used distance and contact potentials (16,29,37) can be written as a simple sum of $\Delta\tilde{W}$ coupling terms as described in Eq. 9, sometimes with a different reference state. In addition, more sophisticated distance potentials that take into account the solvent accessibilities or the conformations of the residues also appear as particular cases of our formalism. A first example is the ''C^{μ}-C^{μ} core/surface'' potential of Kocher et al. (16), which is derived separately for residue pairs that are buried or on the surface of the protein (Table 2 (d)). In the same line of thought, the energy function presented by Simons et al. (17) is composed of an environment term, comparable to $\Delta\tilde{W}_{\mathrm{as}}$ (with $F_{\mathrm{LOC}} = 0$), and a pair term based on the spatial distance separating two residues in specific environments and designed to avoid redundancy with the environment term. This energy function is equivalent to the combination given in Table 2 (e), where $\Delta\tilde{W}_{\mathrm{ass}}$ and $\Delta\tilde{W}_{\mathrm{asas}}$ are distance-independent contributions included in the distance potential, which do not correspond to local potentials since the sequence separation *i–j* is not taken into account. Furthermore, Zhang and Kim estimated contact energies between residue pairs, depending on the conformations of their main chain (ERCE: Environment-Independent Residue Contact Energies) (18). To do this, they combined the 20 amino acid types with 3 structural states (α-helix, β-sheet, and turn) to define an extended 60-residue alphabet. This approach can easily be translated into a combination of $\Delta\tilde{W}$ coupling terms, as described in Table 2 (f). Finally, several authors derived distance potentials from data sets containing only α- or only β-proteins (14,42–44). The basic potential defined in Eq. 9, when derived separately on a subset of the database (α- or β-proteins), becomes $-kT \ln(P(s_{\mathrm{i}},s_{\mathrm{j}},d_{\mathrm{ij}}|t_{\mathrm{i}},t_{\mathrm{j}})/P(s_{\mathrm{i}},s_{\mathrm{j}}|t_{\mathrm{i}},t_{\mathrm{j}})P(d_{\mathrm{ij}}|t_{\mathrm{i}},t_{\mathrm{j}}))$, where $(t_{\mathrm{i}},t_{\mathrm{j}})$ refers to the global secondary structure content of the protein. With such a definition, this distance potential is equivalent to the combination given in Table 2 (g).

TABLE 2 Correspondence with other statistical potentials

Potential	Corresponding combination of our coupling terms
(a) Residue-to-torsion (22)	$\Delta\tilde{W}_{\mathrm{ts}} + \Delta\tilde{W}_{\mathrm{tss}}/(2\ F_{\mathrm{LOC}} + 1)$
(b) Torsion-to-residue (16)	$\Delta\tilde{W}_{\mathrm{ts}} + \Delta\tilde{W}_{\mathrm{tts}}/(2\ F_{\mathrm{LOC}} + 1)$
(c) E^{sec} (23)	$\Delta\tilde{W}_{\mathrm{ts}} + \Delta\tilde{W}_{\mathrm{tt}} + \Delta\tilde{W}_{\mathrm{tt}s} + \Delta\tilde{W}_{\mathrm{ttt}} + \Delta\tilde{W}_{\mathrm{ttts}}$
(d) C^{μ}-C^{μ} core/surface (16)	$\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{asd}} + \Delta\tilde{W}_{\mathrm{sds}} + \Delta\hat{W}_{\mathrm{asdas}}$
(e) −log (*P*(sequence|structure)/*P*(sequence)) (17)	$\Delta\tilde{W}_{\mathrm{as}} + \Delta\tilde{W}_{\mathrm{as}s} + \Delta\tilde{W}_{\mathrm{asas}} + \Delta\tilde{W}_{\mathrm{sds}} + \Delta\tilde{W}_{\mathrm{asds}} + \Delta\tilde{W}_{\mathrm{asdas}}$
(f) ERCE (18)	$\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{td}} + \Delta\tilde{W}_{\mathrm{tsd}} + \Delta\tilde{W}_{\mathrm{sds}} + \Delta\tilde{W}_{\mathrm{tdt}} + \Delta\hat{W}_{\mathrm{tsdts}}$
(g) Distance potentials (only α- or only β-subsets) (14,42–44)	$\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{tsd}} + \Delta\tilde{W}_{\mathrm{sds}} + \Delta\hat{W}_{\mathrm{tsdts}}$

The generality of our approach is demonstrated by the fact that several previously described potentials can be expressed as a linear combination of some of our coupling terms. Note that, in some cases, the values taken by the structural descriptors and the formalism used to define the reference state are quite different from ours. As a consequence, these potentials are generally not identical to the corresponding combination of our coupling terms, but rather describe the same contributions in a slightly different way.

Regarding the increase in performances provided by our new derivation scheme, the results summarized in Table 1 are unambiguous: $\Delta W'_{\mathrm{LOC}}$, $\Delta W'_{\mathrm{DIST}}$, and especially $\Delta W'_{\mathrm{LOC}} + \Delta W'_{\mathrm{DIST}}$ are superior to common distance and local potentials such as $\Delta\tilde{W}_{\mathrm{sd}} + \Delta\tilde{W}_{\mathrm{sds}}$, $\Delta\tilde{W}_{\mathrm{a}s}$, and $\Delta\tilde{W}_{\mathrm{ts}}$. This comparison can be considered as fair, given that all these potentials are derived from the same data set, using the same type of reference state, structural descriptors, and adjustable parameters. Another way to assess the performances of the potentials is to look at previously published tests on the same groups of decoy sets. This comparison has nevertheless the drawback that the effects of derivation scheme, reference state, and other parameters are mixed.

Several potentials have been tested on the group of decoy sets D^{1}_{str} (30,47); the results are summarized in Table 3. According to this test, our distance potential $\Delta W'_{\mathrm{DIST}}$ is clearly superior to every other residue-based distance or contact potential given in Table 3, as indicated by all available measures except S_{-1} in the case of TE-13 and DFIRE-B. This difference is even more manifest when we consider the combination $\Delta W'_{\mathrm{LOC}} + \Delta W'_{\mathrm{DIST}}$. Table 3 also suggests that atom-based potentials perform on the average better than potentials considering only one interaction center per residue. Even so, the residue-based combination $\Delta W'_{\mathrm{DIST}}$ appears markedly more efficient than the RAPDF and KBP potentials. The good performances of the potentials DFIRE-A and DFIRE-B seem to result from the use of a particular reference state, defined in such a way that the effective energy associated to a pair of atoms (or residues) tends to zero when the distance separating them approaches 15 Å (47). Let us also note that another statistical potential, based on a detailed (atomic) representation of protein structures and designed to describe H-bonds as precisely as possible, has been recently tested on the D^{2}_{str} group of decoy sets (19). The results were slightly better than with our potentials ($\langle Z\rangle = -3.34$ and $S_{-1} = 92\%$, whereas $\langle Z\rangle = -2.65$ and

TABLE 3 Comparison with the performances of other statistical potentials

		$\langle Z \rangle$	S_1	S_{-1}
Our potentials (residue-based)	$\Delta W'_{\mathrm{LOC}}$	−4.16	76%	92%
	$\Delta W'_{\mathrm{DIST}}$	−4.65	80%	88%
	$\Delta W'_{\mathrm{LOC}} + \Delta W'_{\mathrm{DIST}}$	−5.25	84%	88%
Other distance or contact potentials (residue-based)	TE-13 (30)	−3.53	56%	100%
	MJ (13,30)	−2.82	44%	88%
	GKS (30,43)	−2.36	36%	80%
	BT (30,48)	−2.65	36%	84%
	HL (30,49)	−2.67	32%	88%
	BJ (30,37)	−2.75	60%	76%
	DFIRE-B (47)	−4.21	76%	96%
Other distance potentials (atom-based)	RAPDF (47,50)	−3.18	72%	84%
	KBP (47,51)	−2.91	60%	84%
	DFIRE-A (47)	−4.84	92%	92%

Results were obtained on the D^1_{str} group of decoy sets; data concerning the potentials derived by other groups were taken from the literature. TE-13, MJ, GKS, BT, HL, BJ (initials correspond to the authors' names), and DFIRE-B (distance-scaled, finite ideal-gas reference state) are contact or distance potentials between pairs of residues. RAPDF (residue-specific all-atom conditional probability discriminatory function), KBP (knowledge-based mean force interaction potential), and DFIRE-A are distance potentials between pairs of atoms (167 atom types are considered, according to the type of the residue to which the atom belongs).

$S_{-1} = 92\%$ are obtained with $\Delta W'_{\mathrm{LOC}} + \Delta W'_{\mathrm{DIST}}$). It is not surprising that better predictive capabilities can be obtained with potentials based on a more detailed structural representation, but it should be stressed that a higher level of detail inevitably induces drastic limitations of the application possibilities.

DISCUSSION

The most exciting result of this study is the definition of a general derivation scheme that allows one to define statistical potentials taking into account the interdependence of correlations among several different sequence or structure descriptors. To demonstrate its interest, we applied this formalism and generated combinations of local and distance potentials that perform strikingly well in discriminating genuine proteins from decoy models.

Our derivation scheme is mainly based on the decomposition of a complex potential into a sum of lower order terms, through the expression of products of probabilities. This decomposition gives the possibility to analyze independently each contribution and clarify its significance and importance. It also offers several valuable advantages in terms of predictive power. First of all, according to the choice of the sequence/structure descriptors, the decomposition may be absolutely necessary to avoid overcounting certain contributions. To clarify this point, let us focus on the correlations between one residue type, s, and two backbone conformations, t. The correct contribution to the total free energy of a protein is given by Eq. 8, in this particular case: $\Delta\tilde{W}_{\mathrm{tts}}(C,S) = \Sigma_{\mathrm{i,j}}\, \Delta\tilde{W}_{\mathrm{ts}}(t_{\mathrm{i}},s_{\mathrm{j}}) + \Sigma_{\mathrm{i,j}}\, \Delta\tilde{W}_{\mathrm{tt}}(t_{\mathrm{i}},t_{\mathrm{j}}) + \Sigma_{\mathrm{i,j,k}}\Delta\tilde{W}_{\mathrm{tts}}(t_{\mathrm{i}},t_{\mathrm{j}},s_{\mathrm{k}})$. In contrast, if the potential function $\Delta\tilde{W}_{\mathrm{tts}}(t_{\mathrm{i}},t_{\mathrm{j}},s_{\mathrm{k}})$ was not decomposed and was summed over all triplets of positions (i,j,k), each $\Delta\tilde{W}_{\mathrm{ts}}$ and $\Delta\tilde{W}_{\mathrm{tt}}$ contribution would be counted several times.

Secondly, the decomposition we propose allows one to deal much more efficiently with the limited size of the database since the correction for sparse data (see Methods) is applied to each coupling term rather than on the whole energy function. For example, the distance potential $\Delta W_{\mathrm{atsdats}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$ can be expressed as a sum of many n-coupling terms, ranging from $n = 2$ to $n = 7$, or computed directly from Eq. 3. If the database is large enough, these two possibilities are equivalent. But if the number of observations of a given combination of values of $(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$ is too small, the correction for sparse data will make $\Delta\tilde{W}_{\mathrm{atsdats}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$ tend to zero, but not $\Delta W_{\mathrm{atsdats}}(a_{\mathrm{i}},t_{\mathrm{i}},s_{\mathrm{i}},d_{\mathrm{ij}},a_{\mathrm{j}},t_{\mathrm{j}},s_{\mathrm{j}})$ unless it is computed directly through Eq. 3. In the latter case, the fact that the database is too small to reliably extract the higher order couplings actually leads to a consequent loss of valuable information about the lower order contributions. Finally, the decomposition makes it possible to modulate the reference state, by excluding some contributions (such as $\Delta\tilde{W}_{\mathrm{aa}}$, $\Delta\tilde{W}_{\mathrm{ada}}$,. . .) that do not appear to be relevant and decrease the overall predictive power.

The comparison with other potentials described in the literature underlines the generality of our approach, for previous potentials based on several sequence or structure descriptors can be expressed as particular cases of our formalism. This comparison also shows that we significantly raised the expectations regarding the predictive power of residue-based potentials. Indeed, our energetic functions even outperform some potentials that are based on a more detailed representation of protein structures at the atomic level.

Several improvements may still be envisaged. Indeed, our derivation scheme can easily be adapted to develop energy functions dealing with a more detailed representation of protein structures, or based on another, possibly more relevant, reference state. It is also straightforward to include additional structural descriptors, reflecting, for example, the relative orientations of interacting side chains or the relative positions of triplets of residues.

We thank J. M. Kwasigroch for his help with computers and web servers, and acknowledge support from the Communauté Française de Belgique through the Action de Recherche Concertée No. 02/07-289, and from the European Community through the Concerted Action Quality of Life 2001-3-8.4.

M.R. is research director at the Belgian National Fund for Scientific Research.

REFERENCES

1. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.
2. Halgren, T. A. 1995. Potential energy functions. *Curr. Opin. Struct. Biol.* 5:205–210.
3. Mackerell, A. D., Jr. 2004. Empirical force fields for biological macromolecules: overview and issues. *J. Comput. Chem.* 25:1584–1604.

4. Gō, N. 1983. Theoretical studies of protein folding. *Annu. Rev. Biophys. Bioeng.* 12:183–210.
5. Galzitskaya, O. V., and A. V. Finkelstein. 1999. A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA.* 96:11299–11304.
6. Alm, E., and D. Baker. 1999. Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures. *Proc. Natl. Acad. Sci. USA.* 96:11305–11310.
7. Munoz, V., and W. A. Eaton. 1999. A simple model for calculating the kinetics of protein folding from three-dimensional structures. *Proc. Natl. Acad. Sci. USA.* 96:11311–11316.
8. Wodak, S., and M. Rooman. 1993. Generating and testing protein folds. *Curr. Opin. Struct. Biol.* 3:249–259.
9. Sippl, M. J. 1995. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5:229–235.
10. Jernigan, R. L., and I. Bahar. 1996. Structure-derived potentials and protein simulations. *Curr. Opin. Struct. Biol.* 6:195–209.
11. Moult, J. 1997. Comparison of database potentials and molecular mechanics force fields. *Curr. Opin. Struct. Biol.* 7:194–199.
12. Russ, W. P., and R. Ranganathan. 2002. Knowledge-based potential functions in protein design. *Curr. Opin. Struct. Biol.* 12:447–452.
13. Miyazawa, S., and R. L. Jernigan. 1996. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.* 256:623–644.
14. Furuichi, E., and P. Koehl. 1998. Influence of protein structure databases on the predictive power of statistical pair potentials. *Proteins.* 31:139–149.
15. Melo, F., R. Sanchez, and D. Sali. 2002. Statistical potentials for fold assessment. *Protein Sci.* 11:430–448.
16. Kocher, J.-P., M. J. Rooman, and S. J. Wodak. 1994. Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches. *J. Mol. Biol.* 235:1598–1613.
17. Simons, K. T., C. Kooperberg, E. Huang, and D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.
18. Zhang, C., and S.-H. Kim. 2000. Environment-dependent residue contact energies for proteins. *Proc. Natl. Acad. Sci. USA.* 97:2550–2555.
19. Kortemme, T., A. V. Morozov, and D. Baker. 2003. An orientation-dependent hydrogen bonding potential improves prediction of specificity and structure for proteins and protein-protein complexes. *J. Mol. Biol.* 326:1239–1259.
20. Buchete, N. V., J. E. Straub, and D. Thirumalai. 2004. Orientational potentials extracted from protein structures improve native fold recognition. *Protein Sci.* 13:862–874.
21. Miyazawa, S., and R. L. Jernigan. 2005. How effective for fold recognition is a potential of mean force that includes relative orientations between contacting residues in proteins. *J. Chem. Phys.* 122:24901–24918.
22. Rooman, M. J., J.-P. A. Kocher, and S. J. Wodak. 1991. Prediction of backbone conformation based on seven structure assignments. Influence of local interactions. *J. Mol. Biol.* 221:961–979.
23. Miyazawa, S., and R. L. Jernigan. 1999. Evaluation of short-range interactions as secondary structure energies for protein fold and sequence recognition. *Proteins.* 36:347–356.
24. Ramachandran, G., and V. Sasilekharan. 1968. Conformation of peptides and proteins. *Adv. Protein Chem.* 23:283–438.
25. Kabsch, W., and C. Sander. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers.* 22:2577–2637.
26. Rose, G. D., A. R. Geselowitz, G. J. Lesser, R. H. Lee, and M. H. Zehfus. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science.* 229:834–838.
27. Wang, G., and R. Dunbrack. 2003. PISCES: a protein sequence culling server. *Bioinformatics.* 19:1589–1591.
28. Hendrick, K., and J. M. Thornton. 1998. PQS: a protein quaternary structure file server. *Trends Biochem. Sci.* 23:358–361.
29. Sippl, M. J. 1990. Calculation of conformational ensemble from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–883.
30. Tobi, D., and R. Elber. 2000. Distance-dependent, pair potential for protein folding: results from linear optimization. *Proteins.* 41:40–46.
31. Samudrala, R., and M. Levitt. 2000. Decoys‘R’Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.* 9:1399–1401.
32. Park, B., and M. Levitt. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392.
33. Simons, K. T., I. Ruczinski, C. Kooperberg, B. A. Fox, C. Bystroff, and D. Baker. 1999. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. *Proteins.* 34:82–95.
34. Keasar, C., and M. Levitt. 2003. A novel approach to decoy set generation: designing a physical energy function having local minima with native structure characteristics. *J. Mol. Biol.* 329:159–174.
35. Tsai, J., R. Bonneau, A. V. Morozov, B. Kuhlman, C. A. Rohl, and D. Baker. 2003. An improved protein decoy set for testing energy functions for protein structure prediction. *Proteins.* 53:76–87.
36. Kang, H. S., A. Kurochkina, and B. Lee. 1993. Estimation and use of protein backbone angle probabilities. *J. Mol. Biol.* 229:448–460.
37. Bahar, I., and R. L. Jernigan. 1997. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. *J. Mol. Biol.* 266:195–214.
38. Zhang, L., and J. Skolnick. 1996. How do potentials derived from structural databases relate to ‘‘true’’ potentials. *Protein Sci.* 7:1201–1207.
39. Shan, Y., and H.-X. Zhou. 2000. Correspondence of potentials of mean force in proteins and in liquids. *J. Chem. Phys.* 113:457–469.
40. Thomas, P. D., and K. A. Dill. 1996. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257:457–469.
41. Dehouck, Y., D. Gilis, and M. Rooman. 2004. Database-derived potentials dependent on protein size for *in silico* folding and design. *Biophys. J.* 87:171–181.
42. Rooman, M., and D. Gilis. 1998. Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power. *Eur. J. Biochem.* 254:135–143.
43. Godzik, A., A. Kolinski, and J. Skolnick. 1995. Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets. *Protein Sci.* 4:2107–2117.
44. Zhang, C., S. Liu, H. Zhou, and Y. Zhou. 2004. The dependence of all-atom statistical potentials on structural training database. *Biophys. J.* 86:3349–3358.
45. Bowie, J. U., R. Luthy, and D. Eisenberg. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science.* 253:164–170.
46. Summa, C. M., M. Levitt, and W. F. DeGrado. 2005. An atomic environment potential for use in protein structure prediction. *J. Mol. Biol.* 352:986–1001.
47. Zhou, H., and Y. Zhou. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 11:2714–2726.
48. Betancourt, M. R., and D. Thirumalai. 1999. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* 8:361–369.
49. Hinds, D. A., and M. Levitt. 1992. A lattice model for protein structure prediction at low resolution. *Proc. Natl. Acad. Sci. USA.* 89:2536–2540.
50. Samudrala, R., and J. Moult. 1998. An all-atom distance-dependent conditional discriminatory function for protein structure prediction. *J. Mol. Biol.* 275:895–916.
51. Lu, H., and J. Skolnick. 2001. A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins.* 44:223–232.