Eur. J. Biochem. *254,* 135−143 (1998)
© FEBS 1998

Different derivations of knowledge-based potentials and analysis of their robustness and context-dependent predictive power

Marianne ROOMAN and Dimitri GILIS

Ingénierie Biomoléculaire, Service de Chimie Organique, Université Libre de Bruxelles, Brussels, Belgium

(Received 13 January 1998) − EJB 98 0045/3

The possibility of defining effective potentials from known protein structures, which are sufficiently accurate to be used for protein-structure-prediction purposes, is investigated. Three types of distance potentials and three types of backbone torsion potentials are defined, based on propensities of amino acid pairs to be separated by a given spatial distance or to be associated to a backbone torsion angle domain. Their differences reside in the way the physical correlations between the amino acids and the conformational states are extracted from the bulk interactions due to the presence of many residues in a protein. For the distance potentials, a physical meaning can be associated to the different definitions, given that some of the potentials favor hydrophobic interactions and others favor interactions between oppositely charged residues. The performance of the different torsion and distance potentials in structure prediction procedures, in particular native-fold recognition and evaluation of protein stability changes upon point mutations, is analyzed. It appears to differ according to the specific proteins and protein environments. In particular, one of the distance potentials performs better than the others for membrane proteins and in protein regions involving charged residues, but less well in other protein regions. Furthermore, the dependence of the potentials on the characteristics of the proteins from which they are derived is analyzed. It is shown that the dependence of the potentials on the length, amino acid composition and secondary-structure content of the proteins from the dataset is either very limited or rather strong, according to the type of potential. The results obtained suggest that the main problem limiting the performance of database-derived potentials is their lack of universality: each potential describes with satisfactory accuracy only the interactions present in certain protein environments.

Keywords: folding free energy; fold recognition; stability changes upon mutation; protein length.

The success of protein-structure prediction from the amino acid sequence is limited by deficiencies in the conformational search procedures aiming at finding the global free energy minimum and in the effective potentials used to evaluate the free energies of the conformations. Recently, a number of solutions to the former problem have been proposed which suggests the possibility of getting a satisfactory solution in the near future. The most promising procedure is constraint-based exhaustive search [1, 2] or branch-and-bound algorithm [3]. It consists of an 'intelligent' search, generating first the structures that are most likely to be the lowest energy ones. It has the non-negligible advantage of yielding the global minimum with respect to a given energy function, usually in a reasonable time.

The second problem, the design of a satisfactory effective potential, seems less obvious to solve. Most effective potentials developed for structure prediction purposes, in particular fold recognition [4, 5], *ab initio* structure prediction [6, 7] and evaluation of stability changes upon mutation [8−11], are derived from a dataset of known protein structures. They are obtained either by optimization of a pre-determined potential function to known protein sequence and structure data [12−14], or are derived from joint frequencies of sequence elements and structural states in the dataset, e.g. from frequencies of amino acids in contact [6, 15, 16], separated by a certain distance [17, 18], solvent accessible [18, 19] or adopting certain values of backbone dihedral angles [20, 21]. These potentials are particularly well suited for structure prediction, where certain degrees of freedom may be neglected or must be neglected to keep computer time within reasonable limits.

Database-derived potentials are mean force potentials; they thus focus on certain interactions and average out the others. Their derivation is moreover subject to several approximations whose justification is not always obvious [22]. In spite of this, the results that these potentials yield for structure prediction are surprisingly good at first sight. In native fold recognition procedures in particular, they perform quite well. This test has however been shown to be relatively easy. It constitutes only a zero-test, which many potentials are able to fulfill [18]. Database-derived potentials perform less well in homologous fold recognition or *ab initio* prediction, for example. The shortcomings of the potentials have now been established unambiguously, owing to the development of search procedures that are able to reach with certainty the global energy minimum. Using these procedures, it has been shown that the conformations corresponding to the global minima of the database-derived potentials are often quite different from the native structures [3].

Recently, several analyses have been devoted to evaluate the quality of database-derived potentials. In particular, Godzik et

Correspondence to M. Rooman, Ingénierie Biomoléculaire, Service de Chimie Organique, Université Libre de Bruxelles, CP165, av. F. Roosevelt 50, B-1050 Brussels, Belgium

Fax: +32 2 650 36 06.

E-mail: mrooman@ulb.ac.be

Abbreviation. DSSP, dictionary of secondary structures in proteins.

al. [23] analyzed several contact potentials published in the literature and showed that some of them differ drastically, as monitored by correlation coefficients of less than 0.5; they also found the surprising and unexplained result that crystallographic and NMR structures yield very different energy parameters. Continuing this analysis, Skolnick et al. [24] studied the importance of neglecting chain connectivity in deriving effective potentials. Using a model of short chains composed of two types of monomers on a square lattice, Thomas and Dill [25] investigated the dependence of contact potentials on the set of proteins from which they are derived, and came to the conclusion that the dependence is very strong. This result was contradicted by Bahar and Jernigan [26], who observe only a weak dependence.

We attempt here to further clarify the meaning and quality of effective potentials. Following the formalism described earlier by one of us [22], we propose different definitions of distance-dependent residue-residue interaction potentials and of backbone torsion potentials, which correspond to different ways of extracting the relevant sequence-structure correlations from the bulk interactions. These approximations lead to potentials where the correlations between residues and/or structural states are differently taken into account. The predictive power of the different potentials is compared, using native fold recognition procedures and prediction algorithms of stability changes upon point mutations. Furthermore, the robustness of database-derived potentials against modifications of the dataset from which they are derived is examined by varying the length, secondary-structure content and amino acid composition of the dataset proteins. Finally, the implications of these results for the possibility of defining sufficiently accurate database-derived potentials are discussed.

RESULTS

Deriving potentials from known protein structures. *Formalism.* To derive effective potentials from known protein structures, it is necessary to make several basic hypotheses. First, it must be assumed that protein sequences S can be divided into sequence elements s and that conformations C can be described in terms of conformational states c. Typically, sequence elements s are single residues or residue pairs, and conformational states c are ranges of backbone dihedral angle values, of spatial distances between residues or of solvent accessibilities. Furthermore, it must be assumed that the relative frequency of these sequence elements and conformational states in the ensemble of native protein structures is equal to their relative frequency in the equilibrium conformations of a single protein. Such potentials may be considered as mean force potentials, because some of the degrees of freedom are averaged out. For example, when the conformational states are domains of backbone dihedral angles, all the side chain degrees of freedom and the backbone ones that correspond to the same domain are averaged out. When they are ranges of spatial distances between, say, C^β atoms, all the main and side chain degrees of freedom that are consistent with the distance constraints are averaged out.

Under these assumptions, the conditional probability $\Pi(C|S)$ of finding a sequence S in a conformation C can be approximated as a product of conditional probabilities $\Pi(c|s)$ of the sequence elements s and conformational states c included in S and C. An estimation of these probabilities $\Pi(c|s)$ can be obtained in terms of probabilities P of s and c approximated by their relative frequencies observed in the dataset of known protein structures. It is tempting to suppose the conditional probabilities $\Pi(c|s)$ and $P(c|s)$ to be equal, where $P(c|s) \equiv P(c, s)/P(s)$ with $P(s)$ the probability of s and $P(c, s)$ the probability of joint occurrence of s and c. However, it has been shown that they are not equal [22], essentially because the probabilities P do not contain a correction for the many-body effect arising from the presence of other residues than those in s and screening out the correlations between s and c. This biasing effect is clearly understood when the structural states c are ranges of inter-residue distances. In that case, the most frequent states c are those in which residues are not in contact. Thus, if the probability of finding a sequence S in a conformation C would be expressed as a product of the $P(c|s)$ values, the most frequent conformations C would be fully extended. This is in obvious contradiction with observation, and $\Pi(c|s)$ is thus not equal to $P(c|s)$.

We dispose of two conditions to determine the form of the Π values: (a) the Π values correspond to probabilities defined from relative frequencies of s and c in the dataset and (b) they must satisfy the condition $\Pi(c) = \text{constant}$. The first condition simply amounts to requiring that the Π values are derived from the structure dataset. The second condition corresponds to the hypothesis that protein conformations are exclusively determined by their amino acid sequence, thus that all the information about the tertiary structure is encoded in the sequence. The form of $\Pi(c|s)$ proposed in [22], which satisfies these two conditions, is

$$\Pi(c|s) = \frac{g(c,s)}{\sum_c g(c,s)}, \text{ where } g(c,s) = \frac{P(c,s)}{P(c)P(s)} . \qquad (1)$$

The probability of finding a sequence S in a conformation C, approximated by the product of the Π values, is related to the free energy $G^S(C)$ and the partition function Z^S by Boltzmann's law:

$$G^S(C) \approx -kT \sum_{i,j} \log \Pi(c_i|s_j) - kT \log Z^S \qquad (2)$$

where k is Boltzmann's constant, T a conformational temperature [27] and the indices i and j indicate the positions of the structural states and sequence elements along the sequence. As this expression contains the partition function, it cannot be completely evaluated. The quantity that can be evaluated is the folding free energy $\Delta G^S(C)$, defined as the free energy difference between a conformation C and a denatured state. Following [22], we define the denatured state as a state in which sequence and structure are uncorrelated; when the conformational states c are inter-residue distances, this state corresponds to the ensemble of conformations with no residue-residue contacts. This yields

$$\Delta G^S(C) \approx -kT \sum_{i,j} \log g(c_i,s_j) = -kT \sum_{i,j} \log \frac{P(c_i,s_j)}{P(c_i)P(s_j)} . \qquad (3)$$

Other solutions. We now show that, when the sequence elements s are amino acid pairs (a_j,a_k), Eqn (1) is not the only solution for $\Pi(c|s)$ that satisfies the aforementioned conditions, in particular $\Pi(c) = \text{constant}$. Indeed, defining Π in terms of a correlation function g:

$$\Pi(c_i|a_j,a_k) = \frac{g(c_i,a_j,a_k)}{\sum_{c_i} g(c_i,a_j,a_k)} \qquad (4)$$

the following expressions represent different solutions of $\Pi(c_i|a_j,a_k)$:

$$\text{I.} \quad g(c_i,a_j,a_k) = \frac{P(c_i,a_j,a_k)}{P(c_i)P(a_j,a_k)} \qquad (5)$$

$$\text{Ia.} \quad g(c_i,a_j,a_k) = \frac{P(c_i,a_j,a_k)}{P(c_i)P(a_j)P(a_k)} \qquad (6)$$

$$\text{Ib.}\quad g(c_i,a_j,a_k) = \frac{P(c_i,a_j,a_k)}{[P(c_i)P(a_j,a_k) + P(a_j)P(c_i,a_k)+P(a_k)P(c_i,a_j)]/3} \tag{7}$$

$$\text{II.}\quad g(c_i,a_j,a_k) = \frac{P(c_i,a_j,a_k)}{[P(a_j)P(c_i,a_k) + P(a_k)P(c_i,a_j)]/2} \tag{8}$$

$$\text{III.}\quad g(c_i,a_j,a_k) = \frac{P(c_i,a_j,a_k)}{[P(a_j,a_k)P(c_i,a_j)P(c_i,a_k)]/[P(c_i)P(a_j)P(a_k)]}\,. \tag{9}$$

The numerators are identical in these five expressions and represent the probability of joint occurrences of two amino acids a_j and a_k and a conformational state c_i. The denominators represent different ways of approximating the correlations between a_j, a_k and c_i present in the numerator. If the events a_j, a_k and c_i were independent, these five expressions would all be equal to one.

Expression I is equivalent to Eqn (1) and corresponds to the solution proposed in [22]. In this case, the correlation function g measures the correlations between the residue pairs (a_j,a_k) and the conformational state c_i. Expression Ia is very similar to I, as the joint probability $P(a_i,a_j)$ of amino acid pairs is not very different from the product $P(a_i)P(a_j)$ of the probabilities of the individual amino acids; we checked that the correlation coefficient between $P(a_i,a_j)$ and $P(a_i)P(a_j)$ is equal to 0.98 and that the regression line has a slope of 0.98. Expressions Ib, II and III are obtained by approximating $P(c_i,a_j,a_k)$ in different ways in terms of $P(a_j,a_k)$, $P(c_i,a_k)$ and $P(c_i,a_j)$. In expression II, the measured correlation is between (c_i,a_j), i.e. a conformational state at position i and an amino acid at position j, and a_k, i.e. an amino acid at position k. Expression Ib can be considered as the average between I and II. Expression III is much more different: g measures the strength of the triplet correlations (c_i,a_j,a_k) relative to the pair correlations (a_j,a_k), (c_i,a_j) and (c_i,a_k).

Further variations can be constructed by replacing in II and III $P(a_j,a_k)$ by $P(a_j)P(a_k)$. Apart from these variations, Eqns (5−9) represent all the solutions that are symmetric in the amino acids a_j and a_k and that do not take solvent molecules explicitly into account. Inserting these expressions into Eqn (3), we obtain the folding free energy of types I, II and III:

I.

$$\Delta G^s(C) \approx -kT \sum_{i,j,k} \log \frac{P(c_i,a_j,a_k)}{P(c_i)P(a_j,a_k)} \tag{10}$$

II.

$$\Delta G^s(C) \approx -kT \sum_{i,j,k} \log \frac{P(c_i,a_j,a_k)}{[P(a_k)P(c_i,a_j) + P(a_j)P(c_i,a_k)]/2} \tag{11}$$

III.

$$\Delta G^s(C) \approx -kT \sum_{i,j,k} \log \frac{P(c_i,a_j,a_k)}{[P(a_j,a_k)P(c_i,a_j)P(c_i,a_k)]/[P(c_i)P(a_j)P(a_k)]}\,. \tag{12}$$

In what follows, only these three types of folding free energies will be considered. The energies obtained with expressions Ia and Ib (Eqns 6−7) are dropped, because they are not sufficiently different from those obtained with expressions I and II.

Correction for sparse data. Due to the limited dataset, the statistics are not always reliable, especially for rare amino acid pairs and conformational states. To compensate for this, we use a correction which is a generalization of the correction originally introduced in [17]. It amounts to replace the correlation function g given in Eqns (5−9) by the following expression:

$$g(c_i,a_j,a_k) \rightarrow \frac{\sigma + n(a_j,a_k)g(c_i,a_j,a_k)}{\sigma + n(a_j,a_k)} \tag{13}$$

where $n(a_j,a_k)$ denotes the number of occurrences of the amino acid pair (a_j,a_k) in the dataset and σ is a parameter that we choose equal to 50, based on earlier tests [18].

Protein structure data. The dataset used to derive the mean force potentials contains 381 protein chains from the Brookhaven databank [28]. It corresponds to the set obtained by the procedure pdb—select [29, 30] on the databank version of 25 May 1996, with sequence identity lower than 25% and with NMR structures and structures with resolutions larger than 2.5 Å dropped. The set is sometimes restricted to the 217 proteins constituted of a single chain, for reasons explained below.

Distance potentials. Three types of distance potentials are obtained by considering in Eqns (10−12) the conformational states c to be inter-residue distances d_{ij}, measured between the two amino acids a_i and a_j:

I.

$$\Delta G^s(C) \approx -kT \sum_{i<j} \log \frac{P(d_{ij},a_i,a_j)}{P(d_{ij})P(a_i,a_j)} \tag{14}$$

II.

$$\Delta G^s(C) \approx -kT \sum_{i<j} \log \frac{P(d_{ij},a_i,a_j)}{[P(a_i)P(d_{ij},a_j) + P(a_j)P(d_{ij},a_i)]/2} \tag{15}$$

III.

$$\Delta G^s(C) \approx -kT \sum_{i<j} \log \frac{P(d_{ij},a_i,a_j)}{[P(a_i,a_j)P(d_{ij},a_i)P(d_{ij},a_j)]/[P(d_{ij})P(a_i)P(a_j)]}\,. \tag{16}$$

The inter-residue distances d_{ij} can be computed between C^α atoms, C^β atoms, side-chain centroids or any other atoms or pseudo-atoms, yielding somewhat different potentials [18]. Here we choose to compute the distances d_{ij} between average side-chain centroids, noted C^μ, defined as the average coordinate centers of all side-chain conformations of a given amino acid type observed in the protein dataset [18]; for Gly residues, the C^μ and C^α positions coincide. The inter-C^μ distances between 3 Å and 8 Å are divided into 25 bins of 0.2 Å width; all distances of more than 8 Å are merged into a single bin, and so are all distances of less than 3 Å. In deriving the potentials, pairs of consecutive residues ($j = i+1$) are not considered. For pairs separated by 1−6 sequence positions, probabilities are computed separately, yielding six distinct potentials describing local interactions along the chain. Pairs separated by more than seven positions along the sequence are all merged, leading to a non-local interaction potential. The so-defined distance potentials are referred to as C^μ-C^μ potentials.

Type I potential corresponds to the most widely used distance potential [17, 18], type III is similar to the residue-mediated effective contact energies of Bahar and Jernigan [26] and type II has to our knowledge never been considered before. The different behavior of these potentials is exemplified in Fig. 1, for selected residue pairs. It appears that potentials I and III differ most and that potential II is in some way the average between the two others.

Potentials I and III favor different kinds of interactions. Potential I favors hydrophobic interactions, as clearly seen in Fig. 1 for the Asp-Arg and Ile-Val pairs. In contrast, potential III favors salt bridge interactions relative to hydrophobic interactions. It nearly vanishes for all hydrophobic pairs (e.g. Ile-Val and Phe-Tyr) and has a pronounced minimum for oppositely charged residues (e.g. Asp-Arg). It has, moreover, a less pronounced minimum than potential I for disulfide bridges (Cys-Cys) and a less pronounced maximum for equally charged residues (e.g. Asp-Glu). For charged-polar interactions (e.g. Asp-Ser) potential III is favorable whereas I is unfavorable.

It is difficult to determine which of these mean force potentials is closest to the true potential, as we do not know exactly what the true potential is. For instance, the importance of electrostatic versus hydrophobic interactions is not fixed throughout

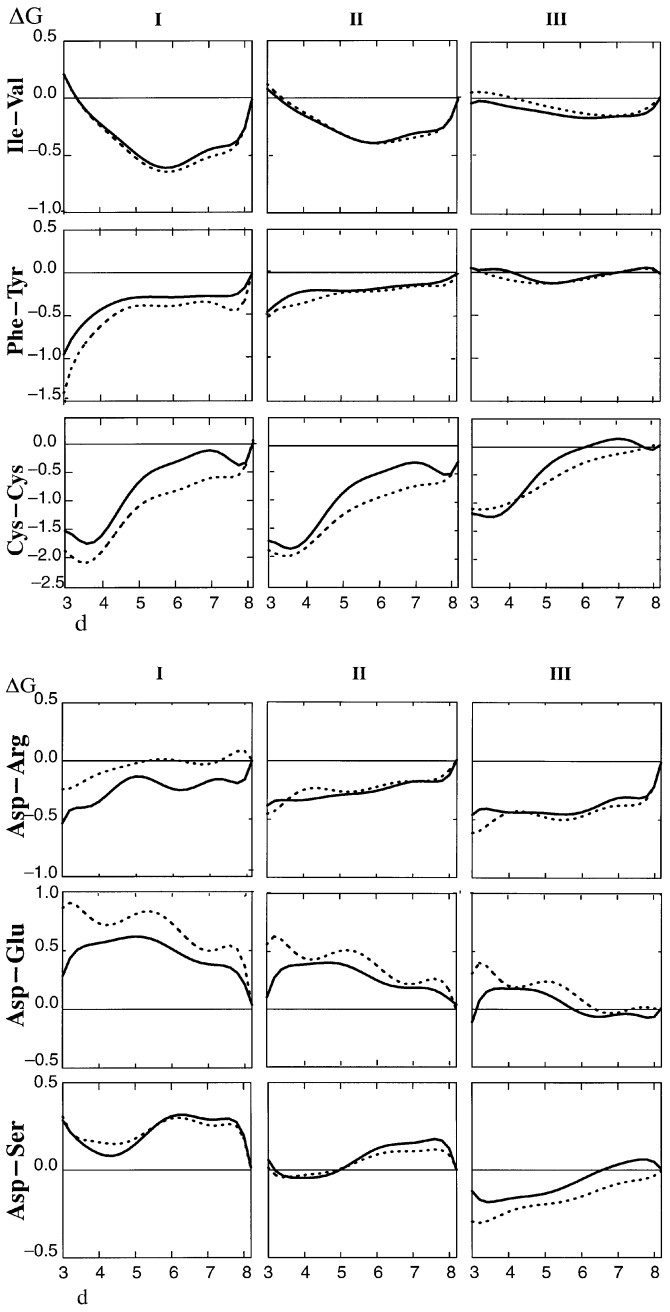

Fig. 1. C$^{\alpha}$-C$^{\alpha}$ distance potentials derived from the protein subsets containing small proteins (dashed line) and large proteins (solid line). The folding free energies ΔG are given as a function of the inter-C$^{\alpha}$ distance d (in Å), for six different amino acid pairs (6 rows) and for potentials I to III (3 columns) defined by Eqns (14−16). The inter-C$^{\alpha}$ distances are divided into bins of 0.2-Å width. The curves are slightly smoothed for aesthetic reasons. The potentials shown correspond to those describing non-local interactions along the chain (see text).

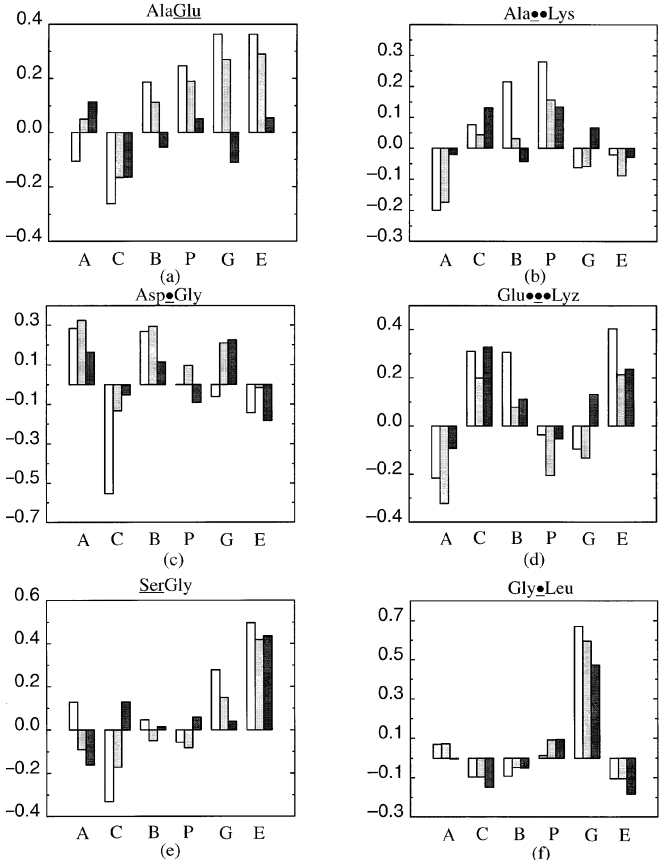

Fig. 2. Type I (white rectangles), II (light grey rectangles) and III (dark grey rectangles) backbone torsion potentials for selected amino acid pairs. The folding free energies ΔG are given for the six backbone torsion domains corresponding to the *trans* conformation, referred to as A and C for helical conformations with A being α-helical and C 3_{10}-helical, B and P for extended conformations with B corresponding more specifically to β-structures, and G and E for conformations with positive ϕ-angle values [20]. The amino acid pairs are represented by the name of the two amino acids, sometimes separated by '•' symbols to indicate unspecified amino acids along the sequence. The underlined residues indicate the positions of the backbone torsion domains whose energy values are computed.

the proteins, but depends on the environment: solvent-accessible salt bridges are not very favorable energetically, whereas fully buried ones are [31]. To give a more objective evaluation of the mean force potentials, we use them in structure prediction algorithms, as described in the next section.

Backbone torsion potentials. Backbone torsion potentials are obtained by considering the conformational states c to be domains of backbone torsion angles (φ, ψ, ω), noted t. As in Rooman et al. [20], we consider seven domains, six for the *trans* peptide bond conformation and one for the *cis* conformation.

Inserting these definitions into in Eqns (10−12), we obtain three types of backbone torsion potentials:

I.

$$\Delta G^{s}(C) \approx -kT \sum_{i,j,k} \frac{1}{\zeta_k} \log \frac{P(t_i,a_j,a_k)}{P(t_i)P(a_j,a_k)} \tag{17}$$

II.

$$\Delta G^{s}(C) \approx -kT \sum_{i,j,k} \frac{1}{\zeta_k} \log \frac{P(t_i,a_j,a_k)}{[P(a_j)P(t_i,a_k) + P(a_k)P(t_i,a_j)]/2} \tag{18}$$

III.

$$\Delta G^{s}(C) \approx -kT \sum_{i,j,k} \frac{1}{\zeta_k} \log \frac{P(t_i,a_j,a_k)}{[P(a_j,a_k)P(t_i,a_j)P(t_i,a_k)]/[P(t_i)P(a_j)P(a_k)]}. \tag{19}$$

The indices i, j, k satisfy $i-8 \leqslant j \leqslant k \leqslant i+8$ and cover a sequence window of 17 residues; ζ_k is a normalization factor ensuring that the contribution of each residue in the window is counted once and is equal to the window size except near chain ends. Only type I torsion potential (Eqn 17) has been described before [20, 22].

These three types of torsion potentials are depicted in Fig. 2 for selected residue patterns. It is noteworthy that torsion poten-

tials I and III differ most and that potential II is intermediate, like observed for C^{μ}-C^{μ} potentials. The similarity between torsion potentials I and II and between torsion potentials II and III can clearly be seen when correlating their energy values for all residue patterns and torsion angle domains. The correlation coefficients I−II, II−III and I−III are indeed equal to 0.95, 0.94 and 0.88, respectively. It can also be noted that, in all depicted examples, the torsion angle domain of minimum (or maximum) energy is the same for potentials I and II, but often differs for potential III.

That the similarity is highest between potentials I and II and between potentials II and III can also be deduced from their definitions (Eqns 17−19). Potential II reduces to potential I if one makes the approximations $P(t_i,a_j) \approx P(t_i)P(a_j)$ and $P(a_j,a_k) \approx P(a_j)P(a_k)$. The latter condition is nearly exact, as mentioned above, but the former is not and measures the correlation between a single residue and a torsion angle domain. The same two approximations allow to transform potential III into potential II. But, to transform potential III into I, one needs to make the approximation $P(t_i,a_j) \approx P(t_i)P(a_j)$ twice, thereby increasing the error and explaining the larger difference between potentials I and III than between I and II or II and III.

Contrary to the C^{μ}-C^{μ} distance potentials where types I to III could be related to the varying importance of hydrophobic and electrostatic interactions, there seems to be no physical interpretation for the different torsion potentials I to III. We can only give a statistical interpretation, which is that the measured correlations are between two residues (a_j,a_k) and a torsion angle domain (t_i) in torsion potential I, and between a residue (a_j) and a residue and a torsion angle domain (t_i,a_k) in potential II. In potential III, the correlation between a residue (a_j) and a residue and a torsion angle domain (t_i,a_k) is compared to the correlation between a residue (a_k) and a torsion angle domain (t_i). It seems at first sight that the definition of torsion potential I is the most meaningful; this will be confirmed in the subsequent sections.

Testing the predictive power of the different potentials. *Structure prediction algorithms.* To compare the predictive power of the different types of distance and backbone torsion potentials and to analyze the effect of the modification of the balance between the dominating interactions, two prediction algorithms are used. The first, called metaFoRe [18], is a native fold recognition algorithm, which proceeds by threading sequences over all the structures from a dataset, without allowing insertions and deletions in the sequence, and identifies native sequence-structure matches on the basis of mean force potentials. To limit computer time, we use a smaller set than that used for deriving the potentials. It contains 141 protein chains from the Brookhaven databank [28], whose structure has been determined by X-ray crystallography to better than 2.5-Å resolution, and which exhibit less than 20% sequence identity (see [32] for a list).

The second prediction algorithm evaluates stability changes upon point mutations on the basis of database-derived potentials [10, 11]. The computed differences in folding free energies between mutant and wild-type structures are compared to experimentally measured values. Two sets of mutations are used: a set of 106 mutations of surface residues with solvent accessibility of at least 50%, whose folding free energy difference has been shown to be well predicted by backbone torsion potentials [10], and a set of 121 mutations of fully buried residues with solvent accessibility between 0 and 20%, whose folding free energy difference is well estimated by C^{μ}-C^{μ} distance potentials [11].

Predictive power of the different types of distance potentials. Using in turn the three C^{μ}-C^{μ} potentials I to III, given by Eqns (14−16), in the fold recognition algorithm metaFoRe, we find that potentials I, II and III identify the native sequence-structure match for 83%, 80% and 52% of the dataset proteins, respectively. Thus, on the average, potential I performs best and potential III worst, and potential II performs nearly as well as potential I. However, though the average performance of potential III is rather low, in some specific cases it performs better than the two other potentials. This is the case for chains L and M of the photosynthetic reaction center (1PRC). This protein is a membrane protein and its chains L and M, well recognized by potential III, are situated inside the membrane. The two other chains of the protein, C and H, situated at least in part outside the membrane, are, in contrast, better recognized by potential I.

These results are easily understood if one remembers that hydrophobic interactions have much less weight in potential III than in potential I and II. It seems thus that potential I and II are better suited for evaluating the folding free energy of non-membrane, globular proteins, with a hydrophobic core. However, potential III seems to yield a better folding free energy estimation for proteins in an apolar medium, such as membrane proteins.

Similar results are obtained with the algorithm predicting stability changes upon single-site mutations. The three C^{μ}-C^{μ} potentials I to III are used in turn to predict the folding free energy changes of 121 mutations of fully buried residues, and the computed values are correlated with the experimental ones. On the average, we find that potential I performs better on this set than potential II, which performs better than potential III. The correlation coefficient between measured and computed changes in folding free energies is indeed equal to 0.78, 0.74 and 0.67 for potential I, II and III, respectively.

Restricting the set of 121 mutations to the subset of 75 mutations where both the mutated and mutant amino acids are hydrophobic yet increases the difference in performance of the potentials: the correlation coefficient becomes equal to 0.63, 0.50 and 0.22 for potentials I, II and III. In contrast, on the 46 remaining mutations, which do not involve purely hydrophobic interactions, the three potentials behave roughly equally well, with correlation coefficients of 0.78, 0.75 and 0.79. On the subset of these 46 mutations where the mutant or mutated amino acids (or both) are charged, the correlation coefficient is equal to 0.82, 0.78 and 0.83. Thus, potential I is only superior for hydrophobic interactions; for non-hydrophobic interactions potentials I and III perform nearly equally well, with even a slightly better score for potential III.

Predictive power of the different types of backbone torsion potentials. To test the predictive power of the backbone torsion potentials I to III given by Eqns (17−19), we use the algorithm that predicts the stability changes of single-site mutations on a set of 106 mutations of solvent accessible residues. The results obtained with potentials I and II are almost similar: the correlation appears to be good except for 10 mutations − the same for potentials I and II − that are situated far from the regression line; as described in [10], these mutations seem to perturb the backbone conformation or to involve atypical interactions for surface residues. On the 96 remaining mutations, the correlation coefficients between computed and measured folding free energy changes are equal to 0.85 and 0.84 for type I and type II potentials, respectively. Potential II performs thus slightly less well than potential I, but remains predictive. The performance of potential III, in contrast, is not good at all. Its correlation coefficient is indeed equal to 0.45, thereby excluding this potential for prediction purposes. Because this potential does not seem to have a physical interpretation, we do not see on which subset of mutations it could perform better.

Using the native fold recognition algorithm metaFoRe, the same trend is observed. Type I torsion potential allows us to

recognize the native structure of 76% of the database proteins, which is a very high score if one remembers that this potential describes only local interactions along the chain, known to be unable to fold proteins. Type II potential recognizes 59% of the proteins and type III potential only 29%; they are thus significantly less well performing than type I potential.

Robustness of database-derived potentials. The dependence of knowledge-based potentials on the characteristics of the proteins from which they are derived is investigated. The considered characteristics are the length of the amino acid sequence, the secondary-structure content and the amino acid composition. For this analysis, the protein structures from the dataset are sorted as a function of one of these characteristics and are divided into three subsets, differing with respect to that characteristic. For example, for analyzing the dependence on the chain length, the dataset structures are sorted according to the length of the protein to which they belong and divided into three subsets containing small, medium-size and large proteins, respectively. The division into three subsets is performed in such a way that the number of residues in each subset is close to 1/3 of the total number of residues.

The effective potentials are derived separately on the three subsets. The difference between them is estimated by correlating the energy values of all sequence elements and structural states computed on one subset, with the equivalent values computed on another subset. The energy values computed from less than five observations are not taken into account in the correlation, to avoid non-physical sparse data effects. To render the dependence of the potentials as clear as possible, the correlations are performed on the potentials derived from the two subsets that differ most with respect to the considered characteristic. For the protein length, for example, it amounts to correlate the potentials derived from the subsets of smallest and largest proteins.

The linear regression lines are computed using the algorithm of least rectangles [33], which determines the coefficients a and b of the regression line $y = a + bx$ so as to minimize the sum of the surface areas of the rectangles:

$$\sum_{i=1}^{n} [y_i - a - bx_i] \left[x_i - \frac{y_i - a}{b} \right] \qquad (20)$$

where n is the number of points. This algorithm ensures that the optimal regression line is independent of the choice of the x and y variables, contrary to the usually employed algorithm of least squares. The coefficients a and b that minimize Eqn (20) are

$$b = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \langle y \rangle)^2}{\sum_{i=1}^{n} (x_i - \langle x \rangle)^2}} \qquad (21)$$

$$a = \langle y \rangle - b \langle x \rangle \qquad (22)$$

where $\langle x \rangle$ and $\langle y \rangle$ denote the mean of the x_i values and y_i values, respectively.

The considered mean force potentials are those described in the previous sections: the three types of C^μ-C^μ distance potentials given by Eqns (14−16) and the three types of backbone torsion potentials given by Eqns (17−19). The results are summarized in Table 1 and are described below.

Dependence on protein length. Let us consider first the backbone torsion and C^μ-C^μ distance potentials of type I. These two potentials are found to exhibit somewhat different dependences on the length of the proteins from which they are derived: the torsion potential is almost totally independent of protein length whereas the C^μ-C^μ potential slightly depends on it. In the case of the torsion potential, the correlation coefficient between the energy values computed from the set of smallest and largest proteins is high (0.91) and the slope of the regression line is almost equal to 1 (Table 1 and Fig. 3a). In the case of the C^μ-C^μ potential, the correlation coefficient is also high (0.92), but the slope of the regression line is equal to 0.85 and thus significantly departs from 1 (Table 1 and Fig. 3b). This means that when the energy values are computed from large proteins, they are on the average smaller, by a factor of 0.85, than those computed from small proteins. Furthermore, as the correlation coefficient is high, the C^μ-C^μ potentials computed from the set of small and large proteins have similar shapes; the dependence on protein size seems thus to reduce to the multiplication by a global factor, independently of the particular amino acids and distance range.

However, a detailed analysis shows that the dependence on protein size of type I C^μ-C^μ potentials is not completely independent of the residue pairs (Fig. 1). For example, the Ile-Val potential, illustrating the potential of hydrophobic residue pairs, does not depend at all on protein length. The same is true for the Asp-Ser potential, with one charged and one hydrophobic residue. The potentials between two charged residues differ more significantly. In particular, when the charges are of opposite

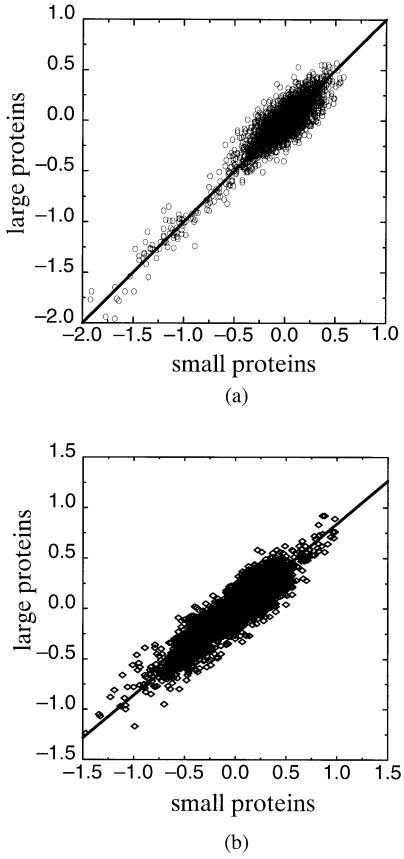

Fig. 3. Correlation between type I potentials computed from the subset of small proteins and the subset of large proteins. The regression lines are computed using the algorithm of least rectangles (Eqns 21−22). (a) Type I backbone torsion potential (Eqn 17). Each point represents the folding free energy value ΔG of an amino acid pair (a_j,a_k) and a torsion angle domain t_i. To avoid overloading the picture, only the energy values for the residue pairs (a_j,a_k) contained in a sequence window $[i-1,i+1]$ around residue i are considered. The equation of the regression line is $y = 1.01x$ and the correlation coefficient is equal to 0.91. (b) Type I C^μ-C^μ distance potential (Eqn 14). Each point represents the folding free energy value ΔG of an amino acid pair (a_j,a_k) and an inter-C^μ distance d_{ij}. The equation of the regression line is $y = 0.85x-0.01$ and the correlation coefficient is equal to 0.92.

Table 1. Dependence of mean force potentials on the characteristics of the proteins from which they are derived. For each potential, the correlation coefficient between the energy values computed from the two protein subsets that differ most with respect to the considered characteristic, is given, as well as the slope of the regression line (in parentheses). The considered characteristics are the length of the protein chain, the percentage of helices computed by DSSP [34], the percentage of β-structures computed by DSSP [34], the percentage of apolar residues (Ala, Cys, Ile, Leu, Met, Phe, Tyr, Trp, Val), and the percentage of charged residues.

Potential		Correlation coefficent for				
		length	percentage helix	percentage β-structure	percentage apolar	percentage charged
C^{α}-C^{α}	I	0.92 (0.85)	0.89 (1.06)	0.89 (0.95)	0.89 (1.02)	0.89 (1.10)
	II	0.86 (0.91)	0.83 (1.00)	0.83 (0.99)	0.83 (1.00)	0.83 (1.03)
	III	0.79 (0.93)	0.74 (0.96)	0.75 (1.05)	0.75 (0.98)	0.76 (0.95)
Torsion	I	0.91 (0.99)	0.83 (1.08)	0.84 (0.95)	0.88 (1.03)	0.89 (1.06)
	II	0.63 (1.01)	0.53 (1.03)	0.54 (0.99)	0.59 (1.02)	0.59 (1.04)
	III	0.35 (1.02)	0.26 (1.03)	0.27 (0.97)	0.32 (1.03)	0.32 (1.05)

sign, as in the Asp-Arg pair, the potential presents a deeper minimum in large than in small proteins. This can be explained by the fact that in small proteins charged residues are often located at the protein surface and that the formation of a salt bridge is much less stabilizing for solvated charged residues than for charged residues buried in the protein core [31]. The Cys-Cys potential presents a pronounced minimum for both small and large proteins, but the minimum is deeper for small proteins. This reflects the fact that small proteins are much more frequently stabilized by disulfide bridges.

Thus, according to the residue pair, type I C^{α}-C^{α} potential computed from large proteins is slightly larger, equal or smaller in absolute value than that derived from small proteins. On the average, it is somewhat smaller. This result can be interpreted as reflecting the fact that the stability of small proteins requires optimal residue-residue interactions, whereas large proteins can accommodate a larger number of interactions that are neither very favorable nor very unfavorable.

Type II and type III potentials are found to depend much more on protein length than type I potentials (Table 1). This is especially true for backbone torsion potentials. The correlation coefficients between energy values derived from large and small proteins are as low as 0.63 and 0.35 for type II and III torsion potentials respectively. One of the reasons of this strong dependence seems to be that the average of the absolute values of type III energies are lower than the corresponding type II values, which are themselves lower than the type I values: they are equal to 0.07, 0.10 and 0.16. Type III energy values seem close to the precision level of the potentials, so that the poor correlation can be attributed to noise effects. The lack of robustness of type II and particularly type III torsion potentials can be taken as an additional indication that these potentials have no physical significance.

In the case of the C^{α}-C^{α} potentials, the dependence of type II and type III potentials on protein size is measured by correlation coefficients of 0.86 and 0.79 (Table 1). The dependence is thus more limited, though larger than that of type I C^{α}-C^{α} potential. These results are consistent with the fact that the mean of the absolute values of type I, II and III energies are equal to 0.23, 0.36 and 0.12, respectively, and thus larger than the corresponding torsion energy values. Furthermore, the slopes of the regression line are larger for the type II and III potentials (0.91 and 0.93) than for the type I potential (0.85), thereby indicating that the dependence on protein length of type II and III C^{α}-C^{α} potentials does not reduce to the multiplication by a global factor, as it is the case for type I C^{α}-C^{α} potential.

That types I, II and III C^{α}-C^{α} potentials exhibit different dependences on protein length is visible in Fig. 1. For the Asp-Arg pair, the difference is particularly marked: the dependence on protein length observed for derivation I disappears for derivations II and III. In contrast, for the Asp-Ser pair, potential III exhibits a dependence on protein length, whereas potentials I and II do not. The way potentials are normalized can thus affect their dependence on a given characteristic.

This detailed analysis shows that the dependence of potentials on protein size may be vanishing, very limited or rather large according to the type of potential and the normalization scheme. This conclusion explains the apparent disagreement between earlier studies, where distance potentials were found either to be independent on protein length [26], or to strongly depend on it [25]. We would like to add that a dependence on protein length can also appear for technical reasons, if one is not careful when deriving the potentials. In particular, it must be mentioned that the aforementioned results for the C^{α}-C^{α} distance potentials are not obtained from the complete dataset of 381 proteins but from the subset containing the 217 proteins composed of a single chain. When considering the full set mixing single-chain and multi-chain proteins, a significant dependence of type I C^{α}-C^{α} potential on protein size is found: the average of the energy differences computed from large and small proteins is equal to -0.32 (instead of 0.01 for single-chain proteins), and the average of the square of the energy differences is equal to 0.35 (instead of 0.11). The correlation coefficient is also slightly lower (0.88 instead of 0.92). The reason of the observed dependence on protein length is purely technical. When computing the C^{α}-C^{α} potential from multi-chain proteins, we take into account pairs of residues with one residue situated in one chain and the other in another chain. Since most of these residue pairs are not in contact, especially when the chains form different domains, the inclusion of these pairs amounts essentially to populating the non-contact bin, grouping the residues separated by more than 8 Å. This population is not counterbalanced by a population in the other bins because, when the sequences of the different chains are homologous, pairs of residues contained in the same chain are counted only once. As a result of the higher population in the non-contact bin, all the energy values are shifted by a positive number. It has to be stressed that for the torsion potential, the dependence on protein length is independent of whether single-chain or multi-chain proteins are used.

Dependence on secondary structure content. To analyze the dependence of the potentials on the secondary-structure content, the proteins from the dataset are sorted according to the proportion of their residues that are in helical conformation, using the definitions of the *Dictionary of secondary structure in proteins* (DSSP) [34]. Similar results are obtained when sorting the proteins according to the fraction of β-structure; the set containing the largest proportion of helices roughly coincides with the set containing the smallest proportion of β-structures.

Both backbone torsion and C^{α}-C^{α} potentials of type I present a small dependence on the secondary-structure content, which does not simply reduce to the multiplication by a global factor. The correlation coefficient between energy values derived from helical versus non-helical proteins is indeed equal to 0.83 and 0.89, for the torsion and C^{α}-C^{α} potentials respectively, and the slope of the regression line is 1.08 and 1.06 (Table 1). It is noteworthy that the C^{α}-C^{α} distance potential depends somewhat less on the secondary-structure content of the proteins from which they are derived than the backbone torsion potential; this can be related to the fact that the definition of torsion potentials involves domains of backbone torsion angles, which are directly related to secondary structures.

Type II and III potentials appear to depend much more on the secondary structure content than the type I potentials, as observed for protein length. Again, the dependence is much larger for the backbone torsion potentials, for the same reasons as those described above.

Dependence on amino acid composition. Similar results are obtained for the dependence of the potentials on the amino acid composition. The proteins from the dataset are sorted either according to the fraction of their residues that are charged, or according to the fraction of their residues that are hydrophobic. The dependence is non-zero but rather limited for type I C^{α}-C^{α} and backbone torsion potentials, as measured by correlation coefficients between 0.88 and 0.89 and slopes between 1.02 and 1.10 (Table 1). For type II and III C^{α}-C^{α} and backbone torsion potentials, on the contrary, the dependence is much more substantial.

DISCUSSION

Two main conclusions can be drawn from the above analysis. First, according to the chosen correction for the many-body effect responsible for the screening out of interactions between amino acids, the derived C^{α}-C^{α} distance potentials attach different weights to the different types of interactions, in particular to hydrophobic and electrostatic interactions. C^{α}-C^{α} potential I is dominated by the hydrophobic effect, while interactions between oppositely charged residues are predominant in potential III; potential II is the average between these two extremes. By analyzing existing contact potentials, Godzik et al. [23] already highlighted the existence of two groups of distance potentials, in which the most favorable interactions are either between hydrophobic amino acids or oppositely charged residues. C^{α}-C^{α} potential I seems thus to belong to Godzik's first group and potential III to the second. What we have shown here is that these different potentials result from different ways of correcting for the many-body effect. In potential I, the state with two given amino acids separated by a certain distance is compared with the state with any two amino acids separated by that distance. As charged residues are generally solvated and thus make few contacts, this normalization does not give much weight to electrostatic interactions. In potential III, the state with two given amino acids separated by a certain distance is compared with the state with each of the two amino acids separated by that distance from any other amino acid. Here, interactions between oppositely charged residues are very favorable, as they are compared with interactions of each of the charged residues with other residues.

The performances of the C^{α}-C^{α} distance potentials I, II and III in native fold recognition and prediction of stability changes upon mutation are found to differ, potential III having the lowest average score. However, a detailed analysis reveals that potential III performs better than potentials I and II in some specific cases, in particular for evaluating stability changes upon mutation of charged residues and for fold recognition of protein chains inside membranes where the hydrophobic effect is weakened. There are thus specific proteins, and specific protein regions, where C^{α}-C^{α} potential III performs better than the two others. The relative performance of the different distance potentials is thus context dependent.

For the backbone torsion potentials, the conclusions are somewhat different. Type I torsion potential is found to be very powerful, especially for predicting stability changes upon mutation of surface residues, whereas type III torsion potential, and to a lesser extent type II potential, have a much lower prediction score. This does not seem to differentiate torsion potentials from C^{α}-C^{α} potentials. What does differentiate them, however, is that we were unable to find subsets of residues, or particular protein environments, where torsion potential III performs better than torsion potential I. This leads to the tentative conclusion that only type I backbone torsion potentials is useful for prediction purposes.

The second main conclusion is that database-derived potentials depend either weakly or strongly on the characteristics of the proteins from which they are derived, according to the type of potential and normalization scheme. Backbone torsion potentials II and III show a strong but irrelevant dependence, as they have a weak predictive value and seem invalid for prediction. For the other potentials, i.e. type I backbone torsion and type I−III C^{α}-C^{α} potentials, the observed dependence is quite limited and seems insignificant compared with the imperfections due to the various approximations made when deriving the potentials, such as the assumption that all different interactions are independent. It is certainly not the dependence of these potentials on database size, secondary-structure content or amino acid composition that is responsible for their limited performance in structure prediction. To confirm this statement unambiguously, we used the fold-recognition procedure metaFoRe in conjunction with type I−III C^{α}-C^{α} potentials and type I torsion potential, where these potentials are derived either from the set of large proteins or from the set of small proteins. The results so obtained are almost undistinguishable from those shown before, where the potentials are computed on the full dataset. Thus, for all practical purposes, these potentials can be considered as independent of the characteristics of set of proteins from which they are derived, provided that the set contains sufficiently well resolved crystal structures with low or no sequence identity.

It becomes thus increasingly clear that there does not exist a single database-derived potential of universal predictive value. The dominant interactions vary according to the position and environment in the parent protein, with the consequence that different definitions of distance potentials are better suited to different protein environments and that backbone torsion potentials perform better at the protein surface whereas distance potentials perform better in the core. Of course, several potentials do well in simple tests such as native fold recognition, but in more demanding tests, not any of the potentials does show a satisfactory performance. The main problem seems to be that the optimal potential, defined by the interactions that have to be considered explicitly and those that may be averaged over, and by the manner of extracting the relevant sequence-structure correlations from the bulk interactions, is highly context dependent: it is different at the surface and in the core, and it depends on the types of residues and secondary structures involved. The issue remains thus to design a sufficiently accurate potential function, or context-dependent combination of potential terms.

We thank Christian Lemer for discussions in the initial stages of this work. D. G. is a Research Assistant at the *Fonds pour la Formation à*

la Recherche dans l'Industrie et l'Agriculture (FRIA), and acknowledges partial support from the *Unité de Conformation des Macromolécules Biologiques*. M. R. is a Senior Research Associate at the Belgian National Fund for Scientific Research (FNRS).

REFERENCES

1. Yue, K. & Dill, K. A. (1995) Forces of tertiary structural organization of globular proteins, *Proc. Natl Acad. Sci. USA 92*, 146–150.
2. Yue, K. & Dill, K. A. (1996) Folding proteins with a simple energy function and extensive conformational searching, *Protein Sci. 5*, 254–261.
3. Lathrop, R. H. & Smith, T. F. (1996) Global optimum protein threading with gapped alignment and empirical pair score functions, *J. Mol. Biol. 255*, 641–655.
4. Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Forschauer, R., Gottsbacher, K., Casari, G. & Sippl, M. J. (1990) 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. 216*, 167–180.
5. Lemer, C. M. R., Rooman, M. J. & Wodak, S. J. (1995) Protein structure prediction by threading methods: evaluation of current techniques, *Proteins 23*, 337–355.
6. Tanaka, S. & Scheraga, H. A. (1975) Model of protein folding: inclusion of short-, medium-, and long-range interactions, *Proc. Natl Acad. Sci. USA 72*, 3802–3806.
7. Defay, T. & Cohen, F. E. (1995) Evaluation of current techniques for ab initio structure prediction, *Proteins 23*, 431–445.
8. Muñoz, V. & Serrano, L. (1994) Intrinsic secondary structure propensities of the amino acids, using statistical ϕ-ψ matrices: comparison with experimental data, *Proteins Struct. Funct. Genet. 20*, 301–311.
9. Miyazawa, S. & Jernigan, R. L. (1994) Protein stability for single substitution mutants and the extent of local compactness in the denatured state, *Protein Eng. 7*, 1209–1220.
10. Gilis, D. & Rooman, M. (1996) Stability changes upon mutation of solvent-accessible residues in proteins evaluated by database-derived potentials, *J. Mol. Biol. 257*, 1112–1126.
11. Gilis, D. & Rooman, M. (1997) Predicting protein stability changes upon mutation using database-derived potentials. Solvent accessibility determines the importance of local versus non-local interactions along the sequence, *J. Mol. Biol. 272*, 276–290.
12. Crippen, G. M. (1991) Prediction of protein folding from amino acid sequence over discrete conformations spaces, *Biochemistry 30*, 4232–4237.
13. Maiorov, V. N. & Crippen, G. M. (1992) Contact potential that recognizes the correct folding of globular proteins, *J. Mol. Biol. 227*, 876–888.
14. Mirny, L. A. & Shakhnovich, E. I. (1996) How to derive a protein folding potential? A new approach to an old problem, *J. Mol. Biol. 264*, 1164–1179.
15. Miyazawa, S. & Jernigan, R. L. (1985) Estimation of effective interresidue contact energies from protein crystal structures, *Macromolecules 18*, 534–552.
16. Godzik, A. & Skolnick, J. (1992) Sequence structure matching in globular proteins: application to supersecondary and tertiary structure prediction, *Proc. Natl Acad. Sci. USA 89*, 12098–12102.
17. Sippl, M. J. (1990) 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. 213*, 859–883.
18. Kocher, J.-P. A., Rooman, M. J. & Wodak, S. J. (1994) Factors influencing the ability of knowledge-based potentials to identify native sequence-structure matches, *J. Mol. Biol. 235*, 1598–1613.
19. Bowie, J. U., Lüthy, R. & Eisenberg, D. (1991) A method to identify protein sequences that fold into a known three-dimensional structure, *Science 253*, 164–170.
20. Rooman, M. J., Kocher, J.-P. A. & Wodak, S. J. (1991) Prediction of protein backbone conformation based on 7 structure assignments: influence of local interactions, *J. Mol. Biol. 221*, 961–979.
21. Kang, H. S., Kurochkina, N. A. & Lee, B. (1993) Estimation and use of protein backbone angle probabilities, *J. Mol. Biol. 229*, 448–460.
22. Rooman, M. J. & Wodak, S. J. (1995) Are database-derived potentials valid for scoring both forward and inverted protein folding, *Protein Eng. 8*, 849–858.
23. Godzik, A., Kolinski, A. & Skolnick, J. (1995) Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets, *Protein Sci. 4*, 2107–2117.
24. Skolnick, J., Jaroszewski, L., Kolinski, A. & Godzik, A. (1997) Derivation and testing of pair potentials for protein folding. When is the quasichemical approximation correct, *Protein Sci. 6*, 676–688.
25. Thomas, P. D. & Dill, K. A. (1996) Statistical potentials extracted from protein structures: how accurate are they, *J. Mol. Biol. 257*, 457–469.
26. Bahar, I. & Jernigan, R. L. (1997) Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation, *J. Mol. Biol. 266*, 195–214.
27. Pohl, F. M. (1971) Empirical protein energy maps, *Nature 237*, 277–279.
28. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meywe, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanoushi, T. & Tasumi, M. (1977) The protein data bank: a computer-based archival file for macromolecular structures, *J. Mol. Biol. 112*, 535–542.
29. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) Selection of a representative set of structures from the Brookhaven Protein Data Bank, *Protein Sci. 1*, 409–417.
30. Hobohm, U. & Sander, C. (1994) Enlarged representative set of protein structures, *Protein Sci. 3*, 522–524.
31. Matthews, B. W. (1991) Mutational analysis of protein stability, *Curr. Opin. Struct. Biol. 1*, 17–21.
32. Wintjens, R. T., Rooman, M. J. & Wodak, S. J. (1996) Automatic classification and analysis of $\alpha\alpha$-turn motifs in proteins, *J. Mol. Biol. 255*, 235–253.
33. Dagnelie, P. (1973) *Théorie et méthodes statistiques* vol. 1, 2nd edn, Presses agronomiques de Gembloux.
34. Kabsch, W. & Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, *Biopolymers 22*, 2577–2637.