*Protein Science* (1995), *4*:2107–2117. Cambridge University Press. Printed in the USA.


# Are proteins ideal mixtures of amino acids? Analysis of energy parameter sets

ADAM GODZIK,[1] ANDRZEJ KOLIŃSKI,[1,2] AND JEFFREY SKOLNICK[1]

[1] Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037
[2] Department of Chemistry, University of Warsaw, Pasteura 1, 02-049 Warsaw, Poland

(RECEIVED April 17, 1995; ACCEPTED July 27, 1995)

**Abstract**

Various existing derivations of the effective potentials of mean force for the two-body interactions between amino acid side chains in proteins are reviewed and compared to each other. The differences between different parameter sets can be traced to the reference state used to define the zero of energy. Depending on the reference state, the transfer free energy or other pseudo-one-body contributions can be present to various extents in two-body parameter sets. It is, however, possible to compare various derivations directly by concentrating on the "excess" energy—a term that describes the difference between a real protein and an ideal solution of amino acids. Furthermore, the number of protein structures available for analysis allows one to check the consistency of the derivation and the errors by comparing parameters derived from various subsets of the whole database. It is shown that pair interaction preferences are very consistent throughout the database. Independently derived parameter sets have correlation coefficients on the order of 0.8, with the mean difference between equivalent entries of $0.1kT$. Also, the low-quality (low resolution, little or no refinement) structures show similar regularities. There are, however, large differences between interaction parameters derived on the basis of crystallographic structures and structures obtained by the NMR refinement. The origin of the latter difference is not yet understood.

**Keywords:** empirical parameter sets; protein structure database; simplified energy calculations

Every known protein, under the appropriate environmental conditions, folds to its native structure, which, according to the "thermodynamic hypothesis," is at the global minimum of its free energy surface (Anfinsen, 1973). In principle, it should be possible to build a model of a protein, develop a formula for its total energy, and search for a global free energy minimum using the computational tools of statistical mechanics. At present, however, this approach is not able to solve the protein folding problem in general, i.e., to predict a previously unknown structure of a protein with a known sequence. There are a number of reasons for this, the most important being the inadequacy of the energy function (Novotny et al., 1984). This last problem can become even more severe, when the protein model is simplified to reduce the computational time needed for the calculations (Skolnick & Kolinski, 1989), and when it is technically feasible to study the whole folding pathway for a medium-sized protein. In such a simplified description, the protein model is built from units equivalent to various collections of heavy atoms (such as side chains or functional groups), and the interaction energy between these units should be treated as a potential of mean force obtained by averaging over all omitted degrees of freedom, rather than as a potential energy (Hill, 1956). It is only for simple motifs (Kolinski & Skolnick, 1994a) or proteins with exaggerated local patterns (Kolinski et al., 1995) that the simplified models can successfully predict a protein structure.

Known protein structures (Bernstein et al., 1977; PDB, 1994) contain a wealth of information about the interaction preferences of amino acids. It has long been recognized that some amino acids have the tendency to be buried in the protein interior (Kendrew et al., 1958). This fact was used in the derivation of many empirical hydrophobicity scales (Cornette et al., 1987). There are also pairs of residues that are often found interacting with each other, ion pairs being the best known example (Barlow & Thornton, 1982; Bryant & Lawrence, 1991). Many such preferences were noticed throughout the years and recently exhaustive classifications were published (Sali & Blundell, 1990; Singh & Thornton, 1990). Efforts to analyze and understand these preferences led to the derivation of the side-chain–side-chain effective potentials (Levitt, 1976; Tanaka & Scheraga, 1976; Warme & Morgan, 1978; Narayama & Argos, 1984; Miyazawa & Jernigan, 1985; Wilson & Doniach, 1989; Hendlich

Reprint requests to: Adam Godzik, Department of Molecular Biology, The Scripps Research Institute, 10666 N. Torrey Pines Road, La Jolla, California 92037; e-mail: adam@scripps.edu.

et al., 1990; Godzik et al., 1992; Jones et al., 1992; Bryant & Lawrence, 1993; Bauer & Beyer, 1994; Koliński & Skolnick, 1994b; Wallqvist & Ullner, 1994). The very fact that there are so many independent derivations, resulting in dramatically different parameter sets, suggests that this problem is far from being understood. Various derivations were never systematically compared to each other, neither on the level of parameter sets, nor on the level of derivation protocols. We intend to fill this gap with the present publication.

If there were no specific amino acid interactions in proteins, then the distribution of amino acids between the interior and exterior of the protein and the distribution of interacting pairs, triplets, etc., would depend only upon the system's geometry and on the relative concentrations of residues of a given type. This statement, in fact, requires clarification, as will be discussed below, because there are several different systems fitting this description. The existence of several different "random" systems is the origin of a great deal of misunderstanding and considerable confusion, as far as derivations of parameter sets are concerned.

It is usually assumed that a nonrandom distribution results from the existence of an energetical term that favors a particular side-chain arrangement over others. In a simplified description, where some degrees of freedom have been averaged out, such terms can be conveniently described as a potential of mean force (Hill, 1956). In principle, it is possible to calculate such potentials by performing long simulations at the atomic level and then averaging over the fast degrees of freedom we are not interested in (Clementi, 1980). In reality, this is not practical because of the number of computations involved and also because our understanding of protein behavior on the atomic level is insufficient. However, if we make the crucial assumption that residues in an ensemble of proteins follow a Boltzmann distribution describing their location, mutual interaction, etc., then we can estimate the potential of mean force by analyzing the distribution of their occurrence. For instance, it has been shown that the distribution of ion pairs is quantitatively related to Coulombs law, albeit the apparent temperature is too high (Bryant & Lawrence, 1991).

It must be noted, however, that the existence of a Boltzmann-like distribution of residue–residue interactions in proteins is far from being obvious. The original derivation of the Boltzmann distribution is done for a system in thermodynamic equilibrium (Hill, 1956). On the other hand, a database of protein structures is a collection of different systems, each in its own respective global free energy minimum (one has a biological ensemble rather than a statistical-mechanical ensemble). It is not at all clear what type of distribution would be followed by such an ensemble. It is only for a random energy model of proteins (Derrida, 1981) and under several other strong assumptions that it is possible to prove that indeed the distribution of residue–residue interactions in proteins is Boltzmann-like with respect to the energy of that interaction (Gutin et al., 1992).

In this contribution, we describe in detail various derivations of interaction energy parameter sets and compare them to each other, both on the level of derivation details, as well as on the level of final parameter sets. In addition, a derivation of a particular parameter set, used in the topology fingerprint-based inverse folding program (Godzik et al., 1992, 1993) is described in detail in the Appendix. To make the comparison possible, all parameter sets are decomposed into "ideal" and "excess" parts. Finally, the derivation consistency is checked by comparing sets derived from various subsets of the structural database.

## Results

### *Folding stages*

For the purpose of the following analysis, the process of protein folding, which starts from the completely unfolded chain ($U$) and ends with the final, native structure ($N$), is divided into conceptual steps. These may or may not have anything to do with the actual protein folding process.

In the first step, a protein changes from a completely unfolded chain to a compact globule, roughly the size of the final protein. We can view this structure as resulting from the existence of a generic "compacting" potential. The entire protein is uniformly packed and can be well described as a randomly packed droplet. We shall call this state $U_{compact}$.

In the next step, interactions between amino acid side chains and water are switched on. The protein separates into a hydrophobic core and a hydrophilic surface layer, each with a composition different from the protein as a whole. As yet, there are no interactions between side chains. Therefore, the distribution of side chain contacts both inside the protein and in the surface layer can be described by a random mixing approximation. We shall call this state $U_{phil\text{-}phob}$.

In the third step, interactions between side chains are "switched on." But only interactions between pairs of two identical side chains assume their correct value. Interactions between two different side chains are approximated by the arithmetic mean of the pair interaction between identical side chains, according to the formula

$$E_{ij}^{ideal} = \frac{E_{ii} + E_{jj}}{2}.$$

The distribution of side chain contact is now similar to that in an ideal liquid. This state will be called $U_{ideal}$.

In the final step, the correct distribution of pairs is formed by "switching on" an excess energy of interacting pairs, i.e.,

$$E_{ij}^{excess} = E_{ij} - \frac{E_{ii} + E_{jj}}{2}.$$

This state is the native state $N$.

We do not know much about an unfolded state $U$. Therefore, most derivations have used one of the states $U_{compact}/U_{phil\text{-}phob}/U_{ideal}$ as their points of origin. It is important to note that both in going from the state $U_{compact}$ to $U_{phil\text{-}phob}$ and from $U_{phil\text{-}phob}$ to $U_{ideal}$, the new interactions that are being introduced are effectively one-body interactions, i.e., there are only 20 parameters. For each amino acid, $k$, there is the energy of transfer from water to a mean protein environment $E_k$ ($U_{compact} \rightarrow U_{phil\text{-}phob}$) and a pair interaction energy between two identical residues $E_{ii}$ ($U_{phil\text{-}phob} \rightarrow U_{ideal}$). On the other hand, on going from the state $U_{ideal}$ to the native state, the interactions are two body and there are 190 parameters needed to describe them.

In the discussion above, secondary structure was not treated separately; instead, it was assumed that it will form in the native state. Some time ago it was suggested that compactness itself in-

**Table 1.** *Parameter sets analyzed in detail, together with the description of specific interaction definitions and other derivations details*[a]

| | Database size | Interaction center | Interaction definition | Threshold (Å) | *r* Dependence |
|---|---|---|---|---|---|
| Warme and Morgan (1978) | 21 | Heavy atoms | Atom-atom closer than threshold, residue-residue recalculated | Sum of VdW + 1.0 | No |
| Narayama and Argos (1984) | 44 | Heavy atoms | Atom-atom closer than threshold, residue-residue recalculated | 6 | No |
| Tanaka and Scheraga (1976) | 25 | Heavy atoms | At least one atom-atom closer than threshold | 6 | No |
| Miyazawa and Jernigan (1985) | 42 | Center of mass | Closer than threshold | 6.5 | No |
| Maiorov and Crippen (1992) | 109 | Cβ | Closer than threshold | 9 | No |
| Bryant and Lawrence (1991) | 141 | "Interaction center" | 1 Å intervals | 5 | Yes |
| Godzik et al. (1992) | 56 | Heavy atoms | At least one atom-atom closer than threshold | 4.5 | No |
| Hinds and Levitt (1992) | 56 | Lattice vertex | At least one atom-atom closer than threshold | 4.5 | No |
| Koliński and Skolnick (1992) | 56 | Center of mass | Atom-atom closer than threshold, residue-residue recalculated | 4.5 | No |

[a] Note that in a number of cases, the interaction center used in the energy calculations and the interaction definition used in the parameter derivation do not coincide. When the strength of an interaction is the same despite the number of atoms actually interacting, the interaction definition is denoted as "at least one atom-atom."

duces secondary structure (Chan & Dill, 1990). If so, secondary structure could spontaneously appear in the state $U_{compact}$. This suggestion was later proved incorrect (Hunt et al., 1994), which is corroborated by results from our laboratory (Koliński & Skolnick, 1992).

### *Comparison of parameter sets*

The single most important difference between various energy parameter sets, such as, for instance, sets reviewed in Table 1, is the difference in the calculation of $N_{expected}$ (see the Materials and methods for explanation of various abbreviations), and more specifically, the reference state defining the zero of the energy function. This statement is corroborated by the data in Table 2, which lists results of pairwise comparisons between various available parameter sets. It is clear that parameter sets can be divided into two groups, with a correlation coefficient of more than 0.5 within each group and almost no correlation between sets from different groups.

It is interesting to note that, on occasion, parameter sets derived by using apparently very different approaches can, in fact, be very similar, sometimes despite the authors' intentions. For instance, the parameters by Warme (Warme & Morgan, 1978) with a reference state $U_{compact}$ have a correlation of 0.74 with the set derived by Godzik et al. (1992) with the reference state $U_{phil\text{-}phob}$. Even more spectacular is the similarity between potentials derived from statistical analysis of protein structures from the first group (see Table 2) and a set derived by Maiorov and Crippen (1992). As discussed earlier, this set was not derived from a statistical analysis of the protein database but instead was

**Table 2.** *Results of pairwise comparisons between several publicly available parameter sets*

| | MJ_I | TS | HL | MC | BL | KS | GC | MJ_II | GKS | I/E[a] | Ideal[b] | Excess[c] | Burial[d] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Miyazawa and Jernigan I (1985) | – | | | | | | | | | 8.2 | 0.96 | 0.16 | 0.96 |
| Tanaka and Scheraga (1976) | 0.89 | – | | | | | | | | 13.2 | 0.97 | 0.06 | 0.88 |
| Hinds and Levitt (1992) | 0.82 | 0.85 | – | | | | | | | 1.6 | 0.81 | 0.13 | 0.89 |
| Maiorov and Crippen (1992) | 0.75 | 0.60 | 0.63 | – | | | | | | 1.4 | 0.70 | 0.19 | 0.86 |
| Bryant and Lawrence (1991) | 0.71 | 0.60 | 0.66 | 0.66 | – | | | | | 1.3 | 0.66 | 0.30 | 0.79 |
| Koliński and Skolnick (1992) | 0.57 | 0.48 | 0.73 | 0.54 | 0.69 | – | | | | 1.0 | 0.42 | 0.28 | 0.74 |
| Gregoret and Cohen (1990) | 0.19 | 0.10 | 0.29 | 0.36 | 0.40 | 0.56 | – | | | 0.7 | 0.02 | 0.78 | 0.66 |
| Miyazawa and Jernigan II (1985) | 0.29 | 0.02 | 0.31 | 0.25 | 0.29 | 0.57 | 0.66 | – | | 0.6 | 0.00 | 0.82 | 0.61 |
| Godzik et al. (1992) | 0.05 | 0.08 | 0.23 | 0.07 | 0.15 | 0.53 | 0.50 | 0.77 | – | 0.7 | 0.09 | 0.79 | 0.06 |
| Warme and Morgan (1978) | 0.09 | 0.06 | 0.28 | 0.22 | 0.31 | 0.58 | 0.55 | 0.74 | 0.63 | 0.6 | 0.04 | 0.79 | 0.46 |

[a] I/E, ratio between the ideal and excess part of the parameter set.
[b] Ideal, correlation between $E_{ij}^{ideal}$ and $E_{ij}$.
[c] Excess, correlation between $E_{ij}^{excess}$ and $E_{ij}$.
[d] Burial, correlation between the "ideal" part of the pair interaction parameter set and the hydrophobic energy.

optimized for recognition of native structures from the group of misfolded structures.

The difference between two groups, as identified in Table 2, can be traced to a single important decision regarding the reference state. If the state $U_{compact}$ is explicitly or implicitly used, then the interaction energy between buried residues includes the transfer energy from the surface of the protein to the protein interior. In other words, the apparent attraction between two residues may result from the fact that they are pushed together into the protein interior. On the other hand, pairs that are often exposed to solvent and thus are underrepresented in the core, end up being neutral or weakly interactive, even if in the protein interior they attract or repel each other. Accordingly, in parameter sets from the first group, the most attractive interactions are typically interactions between two hydrophobic residues. For instance, Phe-Phe interactions at −6.85 are the strongest in the MJ_I set. In the same set, the Glu-Glu interaction energy (−1.18) is almost the same as Lys-Glu (−1.60) or Arg-Arg (−1.39). On the other hand, in parameter sets from the second group, both trends are reversed. Hydrophobic residues in the core are often neutral to each other, and the strongest attraction is typically between oppositely charged groups, whereas groups with the same charge repel each other. For instance, in the MJ_II potential set, Lys-Glu is the strongest interaction at −0.96. It could never be mistaken for the Asp-Asp interaction (+0.04). In another set from this group, GKS (Godzik et al., 1992), Asp-Arg is the strongest attraction at −1.0, with Phe-Phe being mildly attractive at −0.3 and the strongest repulsion occurring between charged and hydrophobic residues.

Another difference between the two groups is seen when parameters are split into the "ideal" and the "excess" part according to Equations 5A and 5B. After such a decomposition is made, it is possible to ask the question, what is the correlation coefficient of the original parameter set to each of its parts? These coefficients are presented as the last two columns in Table 2. The difference between the two groups of parameter sets is again clearly visible, as in the first group the "ideal" part constitutes the dominant part of the total interaction energy. In contrast, the parameters from the second group are almost entirely composed from the "excess" term.

The earlier suggestion, that by using $U_{compact}$ as a reference state, the transfer energy between solvent and protein environment is "mixed in" to the pairwise interaction set is further corroborated by analyzing the 20-parameter "ideal" component of interaction energy. As seen by the correlation coefficient listed in the last column of Table 2, the "ideal" component is almost identical to a "hydrophobic" scale derived on the basis of analysis of composition change between a protein core and a protein surface (Godzik et al., 1992), which in turn is closely related to "transfer" hydrophobic scales (Cornette et al., 1987). This explains the strong similarity between various parameter sets. In particular, it helps to explain the results of the Crippen (Maiorov & Crippen, 1992) experiment in recognizing misfolded structures. It was proved repeatedly that the hydrophobic energy is very effective in recognizing correct (Godzik & Skolnick, 1992) or similar (Bowie et al., 1990) protein folds.

A very different picture is seen when various parameter sets are compared on the level of $E_{ij}^{excess}$, recalculated from various sets according to the Equation 5. Now almost all parameter sets are correlated with each other at the level of 50–70%, including parameters that previously belonged to different groups (see Table 3). This probably reflect differences in the interaction definitions and data sets.

### *Analysis of the consistency of the derivation*

The derivation of the empirical energy parameter set was carried out according to the procedure described in Appendix 1, using the database of high-quality crystallographic structures. The resulting parameter set is presented in Table 4. This set, together with all parameter sets discussed here, can also be downloaded via anonymous ftp from pub/adam directory at ftp.scripps.edu. Because the energy was obtained from the analysis of a statistical distribution according to Equation 1, the natural unit is $kT$. To test how much the results depend on the actual protein list, the current results were compared to the results of the same analysis performed on a preliminary list compiled for the PDB release 56, which contained 59 proteins (Godzik et al., 1992). The correlation between the two sets is shown in Figure 1, where each energy term is shown as a point $[x, y]$, where $x$ is the value derived from the small and $y$ from the large database. The correlation between the two sets is very good, with a correlation coefficient equal to 0.84 for the two-body terms and the mean difference between equivalent terms equal to $0.15kT$. However, the differences between individual contributions can be quite large, and, in some cases, they exceed $0.5kT$. Errors are largest for interactions between rare amino acids, where the number of cases in the small protein database was ap-

**Table 3.** *Results of pairwise comparisons between "excess" part of several publicly available parameter sets*

| | MJ_I | TS | HL | MC | BL | KS | GC | MJ_II | GKS |
|---|---|---|---|---|---|---|---|---|---|
| Miyazawa and Jernigan I (1985) | – | | | | | | | | |
| Tanaka and Scheraga (1976) | 0.67 | – | | | | | | | |
| Hinds and Levitt (1992) | 0.81 | 0.67 | – | | | | | | |
| Maiorov and Crippen (1992) | 0.60 | 0.33 | 0.51 | – | | | | | |
| Bryant and Lawrence (1991) | 0.52 | 0.47 | 0.62 | 0.32 | – | | | | |
| Koliński and Skolnick (1992) | 0.78 | 0.74 | 0.86 | 0.41 | 0.61 | – | | | |
| Gregoret and Cohen (1990) | 0.68 | 0.44 | 0.60 | 0.51 | 0.39 | 0.66 | – | | |
| Miyazawa and Jernigan II (1985) | 1.00 | 0.67 | 0.81 | 0.60 | 0.52 | 0.78 | 0.68 | – | |
| Godzik et al. (1992) | 0.77 | 0.54 | 0.75 | 0.42 | 0.46 | 0.61 | 0.61 | 0.77 | – |
| Warme and Morgan (1978) | 0.77 | 0.62 | 0.69 | 0.45 | 0.45 | 0.46 | 0.46 | 0.77 | 0.59 |

**Table 4.** *The parameter set derived in this paper*

| | Ala | Ser | Cys | Val | Thr | Ile | Pro | Met | Asp | Asn | Leu | Lys | Glu | Gln | Arg | His | Phe | Tyr | Trp |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Ala | 0.1 | 0.1 | 0.1 | −0.1 | 0.0 | −0.1 | 0.0 | 0.0 | 0.2 | 0.1 | −0.1 | 0.3 | 0.3 | 0.0 | 0.3 | 0.2 | −0.1 | −0.1 | 0.0 |
| Ser | 0.1 | −0.4 | 0.1 | 0.3 | −0.2 | 0.4 | −0.3 | 0.3 | −0.5 | −0.4 | 0.4 | −0.3 | −0.5 | −0.3 | −0.3 | −0.3 | 0.1 | −0.1 | 0.2 |
| Cys | 0.1 | 0.1 | −0.9 | 0.0 | 0.0 | 0.2 | −0.1 | 0.1 | 0.3 | −0.1 | 0.1 | 0.4 | 0.4 | 0.1 | 0.4 | 0.0 | 0.1 | 0.1 | 0.2 |
| Val | −0.1 | 0.3 | 0.0 | −0.2 | 0.2 | −0.1 | 0.1 | 0.1 | 0.6 | 0.3 | −0.1 | 0.4 | 0.4 | 0.2 | 0.4 | 0.5 | 0.0 | 0.1 | 0.1 |
| Thr | 0.0 | −0.2 | 0.0 | 0.2 | 0.0 | 0.1 | −0.2 | 0.1 | −0.3 | −0.3 | 0.2 | −0.1 | −0.2 | −0.2 | −0.1 | −0.1 | 0.1 | 0.0 | 0.2 |
| Ile | −0.1 | 0.4 | 0.2 | −0.1 | 0.1 | −0.1 | 0.2 | 0.0 | 0.5 | 0.4 | −0.1 | 0.2 | 0.5 | 0.2 | 0.3 | 0.4 | 0.0 | 0.1 | 0.1 |
| Pro | 0.0 | −0.3 | −0.1 | 0.1 | −0.2 | 0.2 | −0.3 | 0.0 | 0.0 | −0.3 | 0.1 | 0.1 | −0.1 | −0.4 | −0.3 | −0.1 | −0.1 | −0.4 | −0.4 |
| Met | 0.0 | 0.3 | 0.1 | 0.1 | 0.1 | 0.0 | 0.0 | −0.2 | 0.5 | 0.1 | 0.0 | 0.2 | 0.3 | 0.1 | 0.3 | 0.2 | −0.1 | −0.1 | −0.1 |
| Asp | 0.2 | −0.5 | 0.3 | 0.6 | −0.3 | 0.5 | 0.0 | 0.5 | −0.3 | −0.6 | 0.6 | −0.9 | −0.3 | −0.3 | −1.0 | −0.7 | 0.3 | −0.2 | 0.0 |
| Asn | 0.1 | −0.4 | −0.1 | 0.3 | −0.3 | 0.4 | −0.3 | 0.1 | −0.6 | −0.7 | 0.3 | −0.5 | −0.5 | −0.6 | −0.5 | −0.3 | 0.1 | −0.2 | 0.0 |
| Leu | −0.1 | 0.4 | 0.1 | −0.1 | 0.2 | −0.1 | 0.1 | 0.0 | 0.6 | 0.3 | −0.2 | 0.3 | 0.5 | 0.3 | 0.3 | 0.4 | −0.1 | 0.0 | 0.1 |
| Lys | 0.3 | −0.3 | 0.4 | 0.4 | −0.1 | 0.2 | 0.1 | 0.2 | −0.9 | −0.5 | 0.3 | 0.1 | −0.8 | −0.3 | 0.3 | 0.0 | 0.1 | −0.3 | −0.1 |
| Glu | 0.3 | −0.5 | 0.4 | 0.4 | −0.2 | 0.5 | −0.1 | 0.3 | −0.3 | −0.5 | 0.5 | −0.8 | −0.4 | −0.3 | −1.0 | −0.6 | 0.3 | −0.3 | 0.0 |
| Gln | 0.0 | −0.3 | 0.1 | 0.2 | −0.2 | 0.2 | −0.4 | 0.1 | −0.3 | −0.6 | 0.3 | −0.3 | −0.3 | −0.4 | −0.4 | −0.2 | 0.1 | −0.2 | −0.1 |
| Arg | 0.3 | −0.3 | 0.4 | 0.4 | −0.1 | 0.3 | −0.3 | 0.3 | −1.0 | −0.5 | 0.3 | 0.3 | −1.0 | −0.4 | −0.4 | −0.2 | 0.2 | −0.3 | −0.1 |
| His | 0.2 | −0.3 | 0.0 | 0.5 | −0.1 | 0.4 | −0.1 | 0.2 | −0.7 | −0.3 | 0.4 | 0.0 | −0.6 | −0.2 | −0.2 | −0.8 | 0.1 | −0.1 | 0.0 |
| Phe | −0.1 | 0.1 | 0.1 | 0.0 | 0.1 | 0.0 | −0.1 | −0.1 | 0.3 | 0.1 | −0.1 | 0.1 | 0.3 | 0.1 | 0.2 | 0.1 | −0.2 | 0.0 | −0.1 |
| Tyr | −0.1 | −0.1 | 0.1 | 0.1 | 0.0 | 0.1 | −0.4 | −0.1 | −0.2 | −0.2 | 0.0 | −0.3 | −0.3 | −0.2 | −0.3 | −0.1 | 0.0 | −0.1 | 0.1 |
| Trp | 0.0 | 0.2 | 0.2 | 0.1 | 0.2 | 0.1 | −0.4 | −0.1 | 0.0 | 0.0 | 0.1 | −0.1 | 0.0 | −0.1 | −0.1 | 0.0 | −0.1 | 0.1 | 0.0 |

parently too small. All values for more common amino acids, such as alanine or serine, differ by less than $0.1kT$.

As explained in the previous paragraph, in the current derivation, care was taken to separate the effects of the one-body hydrophobic interactions from the two-body interactions. This separation can be tested by looking at the dependence of the pair interaction energy on the hydrophobicity of both residues, using the statistical hydrophobicity scale (Godzik et al., 1992). As seen in Figure 2A, there is actually an anticorrelation (equal to −0.22). In contrast, as discussed earlier, these two effects were often not separated as illustrated in Figure 2b for the Miyazawa-Jernigan interaction energy parameter set (Miyazawa & Jernigan, 1985). In fact, this effect is so strong (correlation is equal to 0.91) that, in such cases, the pairwise interaction energy can actually be decomposed into a sum of one-body terms.

In the present derivation, the residue size was accounted for by introducing the "contact ratio." As before, it can be seen that in the current parameter set, there is no correlation ($r = -0.06$) between the value of the interaction parameter between pairs of residues and their sizes (Fig. 3A). This is not the case for parameter sets that lack this correction (Fig. 3B, $r = 0.89$).

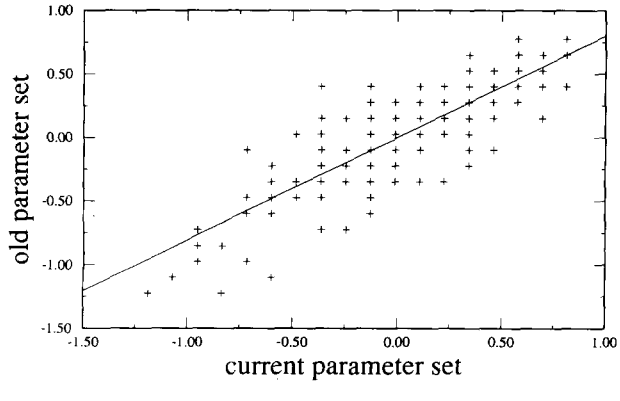


**Fig. 1.** Comparison between parameter sets derived on the basis of protein structure databases having 59 and 381 members.

The current database of structures is large enough that the consistency of the derivation can be checked by rederiving the parameter set for various subsets of the full database. The standard jackknife test is difficult to apply, because there are many related structures in the database. As discussed in the Materials and methods, a 50% sequence identity cutoff was used to build the current database, which left many homologous proteins. Also, there are many examples of significant structural similarity despite any sequence similarity. In the current database, there are more than 30 topologies with more than one example. Some of them, such as globins or TIMs, have more than 10 members. Therefore, two tests were performed that extended the idea of the jackknife test.

In the first test, the database was randomly divided into two subsets in such a way that all members of a given topological family are in one subset. The second test was to check how far the correction for the protein size really eliminates size effects. The whole database was divided into a set of large proteins with more than 210 residues and small proteins with less than 200 residues. The reason for this division was that no topological group was split into two subsets. In both cases, the agreement between the two independently derived parameter sets was very good, with a correlation coefficient better than 0.9 ($r = 0.91$ and $r = 0.92$, respectively).

Protein structures can also be divided according to the type of dominant secondary structure. In our database, we have 46 all-$\alpha$ proteins, where $\alpha$-helices constitute more than 40% of the total length of the proteins and the extended structure is not present. Similarly, there are 42 all-$\beta$ proteins. Comparing parameter sets derived from sets of all-$\alpha$ and all-$\beta$ proteins (see Fig. 4), we see that there is little correlation between the two-body interaction parameters in the two sets ($r = 0.34$). A closer

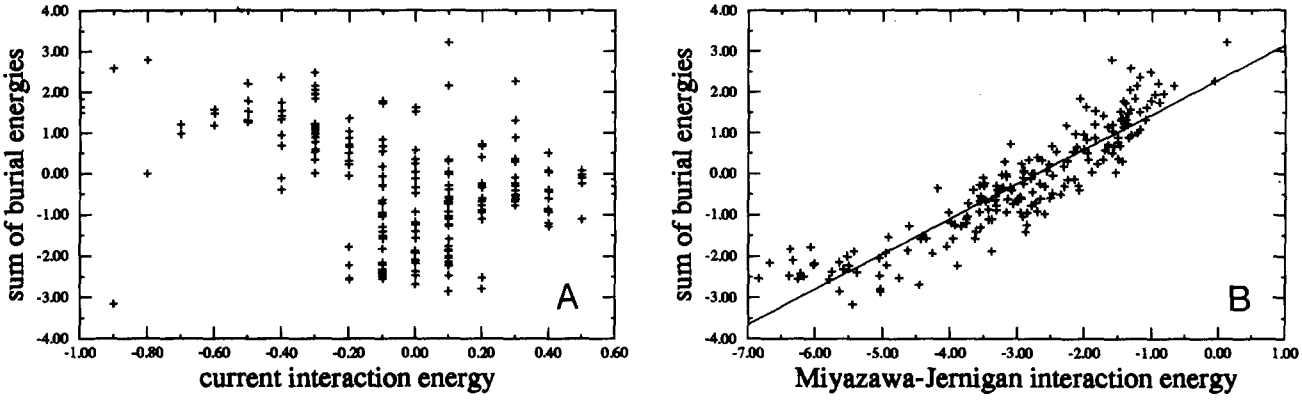


**Fig. 2.** **A:** Comparison between the two-body interaction parameter set developed here and the sum of one-body hydrophobicities of both interacting residues. **B:** The same plot for the Miyazawa–Jernigan parameters.

analysis of Figure 4 reveals that there are a few side chains that are responsible for most of the discrepancy. In general, hydrophobic residues behave in the same way in both types of proteins; for instance, Phe–Phe or Phe–Leu interaction having the same energy. On the other hand, polar residues in different secondary structure types for all practical purposes behave as two different side chains. In $\alpha$ proteins, the Glu–Glu, interaction is repulsive with an energy equal to $+0.6kT$, whereas in $\beta$ proteins it is attractive with the energy value equal to $-0.8kT$. Other examples of such dramatic changes involve pairs such as Arg–Cys ($-0.3$ versus $+1.0$), Lys–Cys (0.0 versus $+1.6$), or Ala–Arg ($+1.5$ versus $-0.2$). It is very tempting to rederive a parameter interaction set to account for this difference by introducing separate types of residues for different secondary structure types, such as Glu_in_the_$\alpha$_helix and Glu_in_the_$\beta$_strand.

As described in the Materials and methods, in addition to the database of low-resolution, highly refined protein structures (HIGH), two other databases were constructed: one containing structures obtained by refinement of NMR data (NMR) and the other has low-quality structures (LOW). It is interesting to test if there are any significant differences between these databases.

This question is answered by Figure 5, where interaction energy parameter sets developed for various databases are compared. As illustrated in Figure 5, the set derived for the HIGH database is markedly different from the one derived for the NMR database with the correlation coefficient between both sets equal to 0.46. This finding could not be explained by differences in protein sizes (NMR-solved structures are usually smaller) nor by secondary structure type (there is a slight predominance of helical proteins in the NMR database). The correlation coefficient is even lower between NMR databases and a subset of HIGH, containing only small, helical proteins. In contrast to this result, the parameter sets derived from HIGH and LOW databases are surprisingly similar, with $r = 0.91$. Because both databases are independent, this is an additional confirmation of the robustness of our parameter set. Still, it is surprising that the increase in protein model accuracy in the HIGH database exerts such a small difference on the interaction parameter sets. This may be related to the on/off definition of interaction.

## Discussion

In this paper, we have compared various existing derivations of energy parameter sets used for energy calculations for simplified models of protein structures. We have shown that, depending on the state used as a reference point, existing sets can

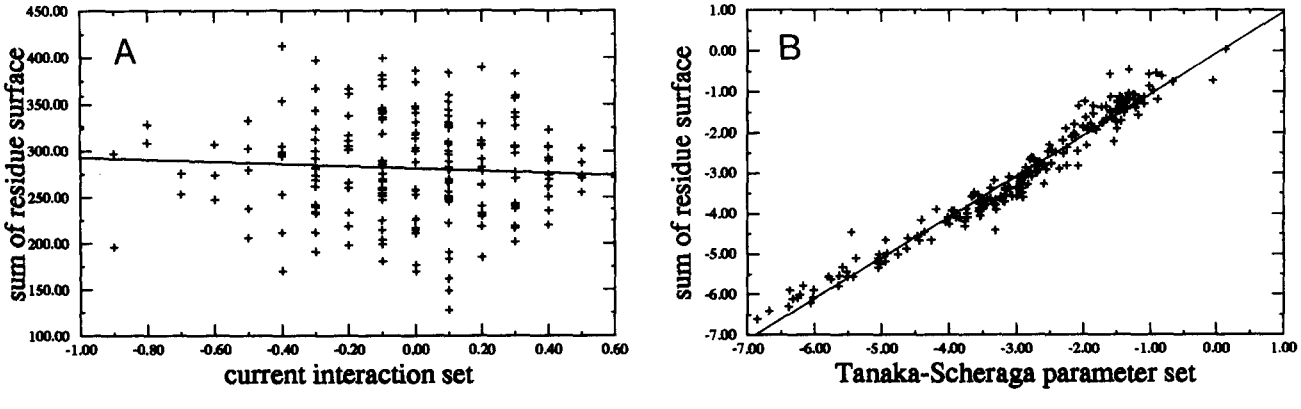


**Fig. 3.** **A:** Comparison between the two-body interaction parameter set developed here and the sum of side chain surface areas of both interacting residues. **B:** The same plot for the Tanaka–Scheraga parameters.

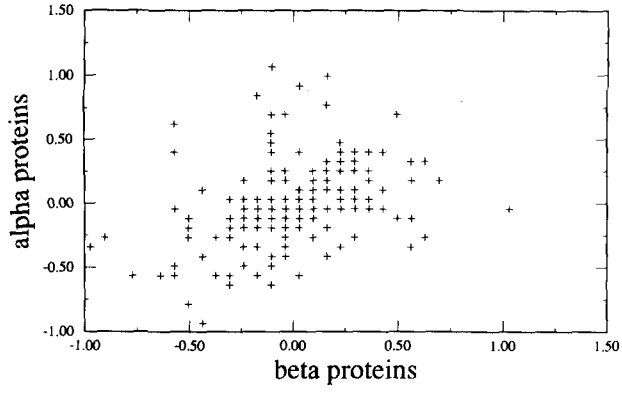


**Fig. 4.** Comparison between the two-body interaction parameter set derived for the subset of all-$\alpha$ and all-$\beta$ proteins.

emphasize different contributions to the total energy of the system. For this reason, it was virtually impossible to compare different derivations and study how other choices, such as interaction definition and database used for derivation influence the results. Introduction of $E_{ij}^{excess}$, which measures the difference between an actual protein and an "ideal" amino acid liquid and thus uses a well-defined reference point, makes it possible to compare at least one component of the different energy sets. At the same time, analysis of the remaining part of the interaction energy parameters is helpful in establishing what reference point was actually used in the derivation. Decomposition of the total two-body energy into the "ideal" and "excess" parts is important for understanding the derivation process and the physical meaning of parameters obtained in derivation but is not likely to influence how the parameters are used in actual calculations.

For all parameter sets, the "ideal" and "excess" part was calculated according to Equations 5A and 5B. It was shown that the "ideal" part in all cases is closely related to the amino acid transfer energy from water to the protein interior. By studying the "excess" part, it was shown that, indeed, apparently different parameter sets are clearly related and display very similar trends. It is possible to compare the relative strength of both

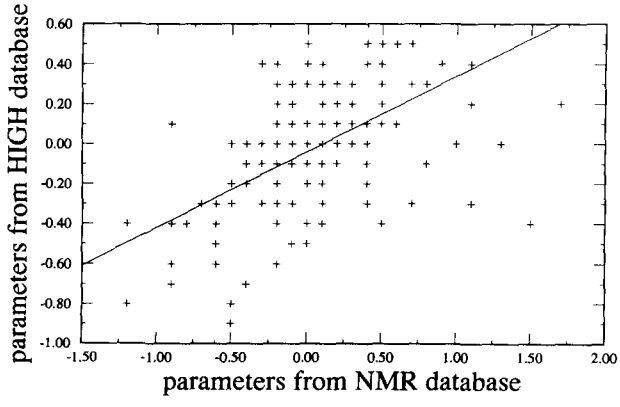


**Fig. 5.** Comparison between the two-body interaction parameter set derived for the subset of protein structures solved by x-ray crystallography and by NMR refinement.

contributions by calculating the mean value of $E_{ij}^{excess}$ and $E_{ij}^{ideal}$. Such comparison is presented in the I/E column of Table 2. With the exception of two derivations, which are almost entirely composed of transfer energy, these two contributions are of equal size. Thus, it is possible to answer the question posed in the title. In all derivations, both ideal and excess parts of the pair interaction are almost equal in strength, and therefore, as might be expected, proteins are definitely not ideal mixtures of amino acids.

At this point, it is difficult to assess which derivation or parameter set is better. Indeed, it is possible that some parameter sets are well suited for some applications and not for others. For instance, in folding simulations, parameters that were derived using a completely unfolded state ($U$) as a reference state should be used. But, if a generic compacting force is introduced (Kolinski & Skolnick, 1992), the parameters obtained with $U_{compact}$ as a reference state might be more appropriate. On the other hand, for threading calculations, parameters using $U_{phil\text{-}phob}$ might be the best.

In the second part of the paper, the internal consistency of the protocol used in the topology fingerprint inverse folding method was tested. It was shown that this parameter set achieves a good separation of one- and two-body terms and is properly corrected for the residue size and surface area. It was also shown that, in contrast to some estimates (Rooman & Wodak, 1988), the size of the database used to derive parameters does not significantly change the general trends but does change individual contributions.

It is not clear at this point what the reason is for the difference between parameters derived for NMR and crystal structures. Possible explanations include different protein environments (solution versus crystal) sampled by NMR or X-ray structures, as well as lack of clear quality assessment of NMR structures. This is an open question, requiring additional investigation.

## Materials and methods

### *Database preparation*

The latest edition of PDB includes more than 2,600 entries (PDB, 1994). There are, however, two serious problems that prohibit the direct use of PDB files for statistical analysis. First, there are many closely related or identical proteins in PDB, with, for instance more than 200 T4 lysozyme mutants and more than 100 closely related hemoglobin structures. The second problem is that the structure quality varies greatly among structural entries, from unrefined, initial models with serious errors in connectivity, packing, and even global topology, to high-quality final structures. The obvious solution is to create a PDB subset that would be as large as possible, but that would contain only unrelated protein structures of reasonable quality. Several such PDB subsets were built in the past (Hobohm et al., 1992; Hobohm & Sander, 1994), but they have usually neglected model quality ($R$ factor) (Hobohm et al., 1992). Furthermore, the process of elimination of similar proteins was always performed before other factors (resolution, the presence of prosthetic groups, technique used) were taken into account. The database prepared here contains 381 high-resolution (resolution better than 2.5 Å, residual factor better than 20%) proteins with a homology threshold of 50% (i.e., every two proteins from the database

have homology lower than 50% identical residues). In addition, two other databases were prepared; one contains only NMR-refined proteins and the other contains low-quality structures (high resolution, high $R$ factors). The NMR-derived structures were grouped separately because, at present, there are no established methods to assess their quality.

### *Parameter derivation*

As mentioned earlier, there have been many attempts to calculate empirical interaction parameters from a database of known protein structures (Levitt, 1976; Tanaka & Scheraga, 1976; Warme & Morgan, 1978; Narayama & Argos, 1984; Miyazawa & Jernigan, 1985; Wilson & Doniach, 1989; Hendlich et al., 1990; Godzik et al., 1992; Jones et al., 1992; Bryant & Lawrence, 1993; Bauer & Beyer, 1994; Koliński & Skolnick, 1994b; Wallqvist & Ullner, 1994). All basically followed the line of reasoning presented below.

We are interested in estimating the interaction energy $E$ between a side chain of type $i$ and a side chain of type $j$. In a real system, we can see $N^{ij}_{observed}$ such interactions. In a system where the actual interaction energy equals zero, this number would be equal to $N^{ij}_{expected}$. If we assume a Boltzmann-like distribution of interacting pairs, then the magnitude of this energy can be estimated from:

$$E = -kT \ln\left(\frac{N_{observed}}{N_{expected}}\right). \quad (1)$$

$N^{ij}_{observed}$ depends on the definition of the event. For instance, for a yes/no contact interaction definition, one simply counts the number of residue pairs that are closer than a threshold value. A much more difficult problem is how to obtain the value of $N^{ij}_{expected}$. We do not have any data about protein systems where interaction energies are equal to zero. Therefore, we have to estimate this number by creating a model of such a system and calculating the number of interacting pairs in such a system. Unfortunately, any of the states $U_{compact}/U_{phil\text{-}phob}/U_{ideal}$ fit this description. At this juncture, various derivations make different choices, which are mostly responsible for differences between parameter sets. The usual choice is to assume that in a "noninteracting" system, the number of interactions between $i$ and $j$ is proportional to a product of two variables (e.g., a mole fraction), one of which is a function of $i$, and one of $j$. Therefore, the expected number of AB interactions is equal to

$$N^{ij}_{expected} = qN_T x_i x_j \quad (2)$$

where $q$ is a residue coordination number and $N_T$ is a total number of residues; $x_i$ and $x_j$ could describe the mol fraction of a residue $i$ or $j$, respectively. Here, we assume that there are $N_T * x_j$ residues of the type $j$ and that each of them has $q$ neighbors. Therefore, there are $q * N_T * x_j$ residues (of any type) interacting with residues of the type $j$, and $x_i$ of which are of type $i$. This derivation closely follows the spirit of the Flory–Huggins mean field theory of polymer solutions (Flory, 1953), which analyzed interactions between a polymer solute and a solvent. In this classic derivation, several assumptions were made, a number of which do not hold for protein systems. For instance, the system was assumed to be infinite and no boundary effects were considered.

The necessity of averaging over a large number of small systems of different sizes complicates the intuitive derivation described above. For instance, boundary effects, which force certain residues away from the protein/water interface, introduce an effective attraction between such residues. But the magnitude of this effect changes with the system size, being strongest for very small systems. When averaging over systems of various sizes, such effects must be treated separately—otherwise, extraction of true pairwise interaction parameters would be impossible.

There are other considerations that potentially can make statistical analysis of proteins difficult. Proteins are not uniformly packed; there are sizable cavities inside them, and densely packed regions are intermingled with more sparsely packed areas. Each of the 20 amino acids has a different size and different connectivity (some are branched, some have rings, etc.). Some, but not all of these effects can be accounted for by introducing the "contact fraction" instead of the mole fraction used in Equation 2. This, in principle, should be calculated separately for every protein to account for the size and density differences between proteins. The contact fraction reduces to a mole fraction for residues that have the same coordination number, and it is a variable intermediate between a surface and a volume fraction, which were suggested in various extensions of FH theory (Ben-Naim & Mazo, 1993; Holtzer, 1994).

### *Parameter set derivations*

As mentioned earlier, there have been many independent derivations of residue–residue interaction parameter sets. Almost all utilize Equation 1, or a close equivalent, to estimate parameters from the analysis of the set of known protein structures. The only derivation that used a different approach was that of Crippen (Maiorov & Crippen, 1992). In his approach, the parameter set was optimized for a particular task: recognition of native structures from a group of misfolded structures. For a selected group of 37 proteins, more than 10,000 alternative conformations were created by taking the combination of a structure of one protein with the sequence of another. The requirement that all the native sequence/structure combinations are lowest in energy results in a set of inequalities that can be solved by iteration. The resulting parameter set will be compared below to those obtained by statistical analysis of interactions in proteins. All of the other derivations still differ in several respects.

#### *The set of protein structures used for parameter derivation*

Older derivations used smaller sets and usually did not consider structure quality. It is only recently that the number of structures available for analysis has increased to the point that the derivation can be repeated independently for various subsets of the whole database. Such subsets may include small, large, all-$\beta$, or all-$\alpha$ proteins.

#### *Definition of the interactions*

Possibilities include the following.

1. The use of various definitions of the interaction sites and various threshold distances for defining the interaction: C$\alpha$s

(Wilson & Doniach, 1989); C$\beta$s (Sippl & Weitckus, 1992); specially introduced "interaction centers" (Bryant & Lawrence, 1993); centers of mass of the side chains (Miyazawa & Jernigan, 1985); side chain heavy atoms (Tanaka & Scheraga, 1976; Warme & Morgan, 1978; Godzik et al., 1992; Hinds & Levitt, 1992).

2. The use of the distance-dependent potential of mean force, which is: a continuous function of $r$ (Wilson & Doniach, 1989; Sippl & Weitckus, 1992); defined in several discrete "bins" (Bryant & Lawrence, 1993); on/off information (Tanaka & Scheraga, 1976; Warme & Morgan, 1978; Miyazawa & Jernigan, 1985; Godzik et al., 1992; Hinds & Levitt, 1992; Maiorov & Crippen, 1992).

### *Level of interaction information*

Some methods used statistics for heavy-atom interactions to derive atom-atom interaction parameters. These were later recalculated to obtain the residue-residue interactions (Koliński & Skolnick, 1994b). Others calculate the residue-residue interaction parameters directly (Tanaka & Scheraga, 1976; Godzik et al., 1992).

### *Calculation of $N^{ij}_{expected}$*

As we will attempt to show below, the most important differences involve using different states ($U_{compact}$, $U_{phil\text{-}phob}$, $U_{ideal}$) as a reference point.

Detailed information about the particular choices made in different derivations are summarized in Table 1. All publicly available parameter sets were recovered from the literature and compared to each other. A compilation of the interaction parameter sets discussed in this paper is available via anonymous ftp (file adam.potentials on the pub/adam directory of ftp.scripps.edu) or can be obtained from the authors.

The first observation that can be made is that there are huge differences between various parameters sets (see Table 2 and the discussion in the next section). Therefore, various assumptions made during the derivation process are clearly very important.

It is particularly interesting to compare various procedures used to calculate $N^{ij}_{expected}$. In almost all cases, it was stated that the reference point is a protein where all specific interactions do not exist, but this could mean any of the systems, $U_{compact}$, $U_{phil\text{-}phob}$, $U_{ideal}$, or perhaps something entirely different. Various derivations used very different theoretical backgrounds and notations for their derivations; therefore, direct comparison is sometimes difficult. Below, we suggest a simple way that can be used to compare various energy sets. We examine the following ratio:

$$\frac{N^{ij}_{observed}}{\sqrt{N^{ii}_{observed}N^{jj}_{observed}}} \tag{3}$$

Using Equation 1, Equation 3 can be expressed as:

$$\frac{N^{ij}_{observed}}{\sqrt{N^{ii}_{observed}N^{jj}_{observed}}} = \frac{N^{ij}_{expected}}{\sqrt{N^{ii}_{expected}N^{jj}_{expected}}} \times \exp\left(-\frac{E_{ij} - \frac{E_{ii}+E_{jj}}{2}}{kT}\right). \tag{4}$$

As long as a functional form for the $N^{ij}_{expected}$ presented in Equation 2 holds, then the terms in the factorial of the exponential of Equation 4 cancel out exactly. Using the notation,

$$E^{excess}_{ij} = E_{ij} - \frac{E_{ii}+E_{jj}}{2} \tag{5A}$$

$$E^{ideal}_{ij} = \frac{E_{ii}+E_{jj}}{2} \tag{5B}$$

we arrive at the formula:

$$E^{excess}_{ij} = -kT\exp\left(\frac{N^{ij}_{observed}}{\sqrt{N^{ii}_{observed}N^{jj}_{observed}}}\right). \tag{6}$$

The variable $E^{excess}_{ij}$ has a number of interesting features. First, it can be derived directly from $N^{ij}_{observed}$ using Equation 6 without making any assumptions about how to calculate $N^{ij}_{expected}$. Next, it can be calculated from existing $E_{ij}$'s, thereby allowing us to compare various derivations, despite the fact that they all use different protocols to estimate $N^{ij}_{expected}$. Unfortunately, $E^{excess}_{ij}$ itself is not a very useful variable, because it describes the difference between states $U_{ideal}$ and $N$. One still needs to know transfer energies and the diagonal values, $E_{ii}$.

We summarize the protocols used to estimate $N^{ij}_{expected}$ for various derivations below.

The first derivation of a statistical parameter set using the database of then available proteins (25 proteins) was performed by Tanaka and Scheraga (1976), who used a relation:

$$N^{ij}_{expected} = N_{(n)i}N_{(n)j}N_T \tag{7}$$

$N_{(n)i}$ and $N_{(n)j}$ are global numbers of noninteracting residues $i$ and $j$, respectively, and, as before, $N_T$ is a total number of residues in the database. This derivation is based on the assumption that the pair interaction energy is related to an equilibrium constant between an $ij$ pair and separate, noninteracting residues $i$ and $j$. It is easy to realize that this way the state $U_{compact}$ is taken as a reference and the interactions contain a large transfer energy contribution.

The next derivation was due to Warme and Morgan (1978), who used a database of only 21 proteins. They employ a formula,

$$N^{ij}_{expected} = x_i x_j N_T \frac{\boldsymbol{q}_i \boldsymbol{q}_j}{\boldsymbol{q}} \tag{8}$$

where $\boldsymbol{q}_i$ is the mean number of interactions for a residue type $i$ (coordination number of residue type $i$), and $x_i$ is the mole fraction of such residues. As before, $N_T$ is the total number of residues. It is easy to see that Equation 8 is closely related to Equation 2, with the mean residue coordination number of $i$ and $j$ residues calculated as $\boldsymbol{q}_i\boldsymbol{q}_j/\boldsymbol{q}$. It is interesting to note that the coordination number for the residue type $i$ is calculated here as a mean number of atom-atom interactions summed over all heavy atoms for a given residue. In other words, if a residue $i$ has five atoms and a residue $j$ six, this interaction might count as 1 or 30, depending on their mutual position. Thus, $\boldsymbol{q}_i$ used here is different from the $q_i$, coordination number of the residue type $i$, used later in the Appendix. The interesting point of

this derivation is that, despite the fact the stated reference state is $U_{compact}$, as would be shown later, the actual reference state is $U_{phil\text{-}phob}$. The reason for this discrepancy is that the formula for $\boldsymbol{q}_i$ used in Equation 8 is strongly biased for buried residues that have more interactions.

Narayama and Argos (1984) "corrected" this formula to:

$$N^{ij}_{expected} = x_i x_j N_T N_i \frac{q_j}{q} \tag{9}$$

using the total number of residues $N_i$, instead of $q_i$. This way the interaction parameters are no longer symmetric, i.e., the interaction energy between a pair $[i,j]$ is not equal to the energy of the pair $[j,i]$. Because of this nonphysical asymmetry, this parameter set is not included in the subsequent analysis. At the same time, they have used $q_i$ instead of $\boldsymbol{q}_i$, thus switching back the reference state to $U_{compact}$.

The two approaches used in the earliest derivations were later repeated in many different variants by other groups. The most comprehensive derivation to date was done by Miyazawa and Jernigan (1985). They have, in fact, derived two energy parameter sets. The first, referred to as MJ_I in the tables below, describes the energy of creating a contact between residues of the types $i$ and $j$ by bringing them together starting from the unfolded state. Therefore, for this parameter set, state $U$ is used as the reference state. The second set gives the conditional energy of a formation of a contact between residues $i$ and $j$, given that both are already in a dense state, interacting with the "mean protein environment," which means that the state $U_{phil\text{-}phob}$ is used as a reference state. This set, denoted here as MJ_II, is very close in spirit to the derivation described in this paper.

Other derivations follow still different paths. Bryant and Lawrence calculate the expected number of contacts by permuting sequences in target structures. The permutation is done without paying attention to the burial/exposed status of the position, thus the state $U_{compact}$ is used as a reference. Koliński and Skolnick (1992) build their parameter set by calculating the interaction energy for atom-atom interactions and later rederive the residue-residue parameter set by averaging interaction energies over all residue pair geometries in the database. The assumption that all $[i,j]$ interactions in the database can be described by a single interaction energy, made indirectly in all other derivations, can be checked within this derivation by calculating histograms of energies of residue-residue interactions. In this derivation, the stated reference state is $U_{compact}$. However, the strongly interacting residues from the protein interior contribute more to the final parameter value, thus the actual reference state is to some extent moved in the direction of $U_{phil\text{-}phob}$. This is a similar but now much weaker effect, as was observed previously for the Warme derivation. Finally, Sippl (Sippl & Weitckus, 1992) and his followers (Jones et al., 1992) built a distance-dependent interaction function. In their derivation scheme, a parameter for a given distance for a given amino acid pair was calculated relative to parameters for other distances for the same pair. This way, short-distance interactions were calculated with the $U_{phil\text{-}phob}$ reference state, because interactions in the core dominate the statistics. On the other hand, long distance interactions were calculated using $U_{compact}$ as a reference state.

The parameter set used before in the topology fingerprint inverse folding algorithm (Godzik et al., 1992), and described in detail in the Appendix, used the state $U_{phil\text{-}phob}$ as its reference point. This is done by calculating $N^{ij}_{observed}$ only for buried residues. Also, care was taken to correct the derivation for the protein size and composition differences between proteins. The original derivation is repeated here for a larger set of proteins.

### *Correlation coefficients*

Throughout this paper, we repeatedly ask the question of how similar is one parameter set to another. To answer this question, we test the hypothesis that the two-parameter sets in question are related by a linear relation. Thus, parameters from the first sets are treated as $[x]$ values, parameters from the second set as matching $[y]$, and the linear regression analysis is performed to fit a line to a data set $[x,y]$. The correlation coefficient $r$ (Crow et al., 1960) is reported as a measure of similarity between the two sets with:

$$r = \frac{\langle xy\rangle - \langle x\rangle\langle y\rangle}{\sqrt{[\langle x^2\rangle - \langle x\rangle^2][\langle y^2\rangle - \langle y\rangle^2]}}. \tag{10}$$

## Acknowledgments

This research was supported in part by grant no. GM48835 of the Division of General Medical Sciences, the National Institutes of Health, and by University of Warsaw grant BST-502/34/95.

## References

Anfinsen CB. 1973. Principles that govern the folding of protein chains. *Science 181*:223-230.

Barlow DJ, Thornton JM. 1982. Ion-pairs in proteins. *J Mol Biol 168*: 867-885.

Bauer A, Beyer A. 1994. An improved pair potential to recognize native protein folds. *Proteins Struct Funct Genet 18*:254-261.

Ben-Naim A, Mazo RM. 1993. Size dependence of the solvation free energies of large solutes. *J Phys Chem 97*:10829-10834.

Bernstein FC, Koetzle TF, Williams GJB, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. 1977. The Protein Data Bank: A computer-based archival file for macromolecular structures. *J Mol Biol 112*:535-542.

Bowie JU, Clarke ND, Pabo CO, Sauer RT. 1990. Identification of protein folds: Matching hydrophobicity patterns of sequence sets with solvent accessibility patterns of known structures. *Proteins Struct Funct Genet 7*:257-264.

Bryant SH, Lawrence CE. 1991. The frequency of ion-pair substructures in proteins is quantitatively related to electrostatic potential. A statistical model for nonbonded interactions. *Proteins Struct Funct Genet 9*: 108-119.

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

Chan HS, Dill KA. 1990. Origins of structure in globular proteins. *Proc Natl Acad Sci USA 87*:6388-6392.

Clementi E. 1980. Computational aspects for large molecular systems. Berlin/Heidelberg: Springer Verlag.

Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisi C. 1987. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J Mol Biol 195*:659-685.

Crow EL, Davis FA, Maxfield MW. 1960. *Statistic manual*. New York: Dover.

Derrida B. 1981. Random-energy model: An exactly solvable model of disordered system. *Phys Rev B 24*:2613-2626.

Flory PJ. 1953. *Principles of polymer chemistry*. Ithaca, New York: Cornell University Press.

Godzik A, Koliński A, Skolnick J. 1993. De novo and inverse folding predictions of protein structure and dynamics. *J Comput Aided Mol Des 7*:397-438.

Godzik A, Skolnick J. 1992. Sequence structure matching in globular pro-

teins: Application to supersecondary and tertiary structure prediction. *Proc Natl Acad Sci USA 89*:12098–12102.
Godzik A, Skolnick J, Koliński A. 1992. A topology fingerprint approach to the inverse folding problem. *J Mol Biol 227*:227–238.
Gregoret LM, Cohen FE. 1990. Novel method for rapid evaluation of packing in protein structures. *J Mol Biol 211*:959–974.
Gutin AM, Badretdinov AY, Finkelstein AV. 1992. Why is the statistics of protein structures Boltzman-like in Russian. *Mol Biol (Mosc) 26*:118–127.
Hendlich M, Lackner P, Weitckus S, Floeckner H, Froschauer R, Gottsbacher K, Casari G, Sippl M. 1990. Identification of native protein folds amongst a large number of incorrect folds. *J Mol Biol 216*:167–180.
Hill TL. 1956. *Statistical mechanics. Principles and selected applications.* New York: McGraw-Hill.
Hinds DA, Levitt M. 1992. A lattice model for protein structure prediction at low resolution. *Proc Natl Acad Sci USA 89*:2536–2540.
Hobohm U, Sander C. 1994. Enlarged representative of protein structures. *Protein Sci 3*:522–524.
Hobohm U, Scharf M, Schneider R, Sander C. 1992. Selection of representative protein data sets. *Protein Sci 1*:409–417.
Holtzer A. 1994. Does Flory-Huggins theory help in interpreting solute partitioning experiments? *Biopolymers 34*:315–320.
Hunt NG, Gregoret LM, Cohen FE. 1994. The origins of protein secondary structure. Effects of packing density and hydrogen bonding studied by a fast conformational search. *J Mol Biol 241*:214–215.
Jones DT, Taylor WR, Thornton JM. 1992. A new approach to protein fold recognition. *Nature 358*:86–89.
Kendrew JC, Bodo G, Dintiz HM, Parrish RG, Wyckoff H, Phillips DC. 1958. A three-dimensional model of the myoglobin molecule obtained by X-ray analysis. *Nature 181*:662–666.
Koliński A, Gałazka W, Skolnick J. 1995. Model of long range interaction in globular proteins. Computer design of idealized b-motifs. *J Chem Phys*. In press.
Koliński A, Skolnick J. 1992. Discretized model of proteins. I. Monte Carlo study of cooperativity in homopolypeptides. *J Phys Chem 97*:9412–9426.
Koliński A, Skolnick J. 1994a. Monte Carlo simulations of protein folding. II. Application to protein A, ROP, and crambin. *Proteins Struct Funct Genet 18*:353–366.
Koliński A, Skolnick J. 1994b. Monte Carlo simulations of protein folding. I. Lattice model and interaction scheme. *Proteins Struct Funct Genet 18*:338–352.
Levitt M. 1976. A simplified representation of protein conformations for rapid simulation of protein folding. *J Mol Biol 104*:59–107.
Maiorov VN, Crippen GM. 1992. Contact potential that recognizes the correct folding of globular proteins. *J Mol Biol 277*:876–888.
Miyazawa S, Jernigan RL. 1985. Estimation of effective interresidue contact energies from protein crystal structures: Quasi-chemical approximation. *Macromolecules 18*:534–552.
Narayama SV, Argos P. 1984. Residue contacts in protein structures and implication for protein folding. *Int J Pept Protein Res 24*:25–39.
Novotny J, Brucolleri R, Karplus M. 1984. An analysis of incorrectly folded protein models. Implications for structure prediction. *J Mol Biol 177*: 787–818.
PDB 1994. Quarterly Newsletter, No. 69, June 1994.
Rooman MJ, Wodak SJ. 1988. Identification of predictive sequence motifs limited by protein structure database size. *Nature 335*:45–49.
Sali A, Blundell TL. 1990. Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol 212*:403–428.
Singh J, Thornton JM. 1990. SIRIUS: An automated method for the analysis of the preferred packing arrangements between protein groups. *J Mol Biol 211*:595–615.
Sippl MJ, Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a database of known protein conformations. *Proteins Struct Funct Genet 13*:258–271.
Skolnick J, Koliński A. 1989. Computer simulations of globular protein folding and tertiary structure. *Annu Rev Phys Chem 40*:207–235.
Tanaka S, Scheraga HA. 1976. Medium and long range interaction parameters between amino acids for predicting three dimensional structures of proteins. *Macromolecules 9*:945–950.
Wallqvist A, Ullner M. 1994. A simplified amino acid potential for use in structure prediction of proteins. *Proteins Struct Funct Genet 18*:267–289.
Warme PK, Morgan RS. 1978. A survey of amino acid side-chain interactions in 21 proteins. *J Mol Biol 118*:289–304.
Wilson C, Doniach S. 1989. A computer model to dynamically simulate protein folding: Studies with crambin. *Proteins Struct Funct Genet 6*: 193–209.

## Appendix 1

We present the derivation of $N_{expected}$ for pair interactions between residues A and B in a database consisting of M proteins. The $k$th protein has a length $L(k)$, the number of residues of type A in this protein is equal to $number_A^k$, and the number of interactions in this protein is $N_{total}^k$. A residue at a position $n$ in a protein $k$ has $ncont_{total}^k$ interactions and the residue type is equal to $seq^k(n)$. In this derivation, only interactions between buried side chains are considered, to allow for the separation of one- and two-body effects (A. Godzik, in prep.). Capital letters $A$, $B$, and $C$ would denote a particular residue type, such as Gly, Ala, or Ser. The derivation proceeds as follows.

1. For every amino acid type $A$, the total number of interactions $S_A$ for this residue type in the whole database is calculated.

$$S_A = \sum_{k=1,M} \sum_{\substack{n=1,L(k) \\ seq^k(n)=A}} ncont_n^k. \tag{A1}$$

Here, the first summations run over all proteins in the database, the second over all positions in each protein, but only under the condition that a residue occupying this position is of the type $A$ and is buried. $S_A$, in turn, is used to calculate the mean number of interactions for every residue type.

$$q_A = \frac{S_A}{\sum_{k=1,M} number_A^k}. \tag{A2}$$

The assumption made here is that the mean number of contacts for every residue is constant throughout the database and this is the only value not calculated separately for every protein.

2. A "contact fraction" for every residue type $A$ is calculated for every protein $k$. Again, this ratio changes from protein to protein, due to variations in protein composition between proteins.

$$X_A^k = \frac{q_A number_A^k}{\sum_{B=1,20} q_B number_B^k}. \tag{A3}$$

3. The expected number of interactions between residues of type $A$ and $B$ is calculated as a product of contact fractions for residue types $A$ and $B$ in a given protein and the number of interactions in this protein.

$$n_{AB}^k = X_A^k X_B^k N_{total}^k. \tag{A4}$$

Steps 1–3 are repeated for every protein in the database.

4. $N_{expected}$ for the whole database is calculated by summing all $n_{expected}$ for individual proteins over all $M$ proteins. Values of $n_{AB}^k$ derived in steps 1–3 are used to obtain the final value of $N_{expected}^{AB}$,

$$N_{expected}^{AB} = \sum_{k=1,M} n_{AB}^k, \tag{A5}$$

which is used according to Equation 1 to yield the interaction parameter for a particular pair.