Inferring Ideal Amino Acid Interaction Forms From Statistical Protein Contact Potentials

Piotr Pokarowski  
*Warsaw University*

Andrzej Kloczkowski  
*Iowa State University*

Robert L. Jernigan  
*Iowa State University, jernigan@iastate.edu*

Neha S. Kothari  
*Iowa State University*

Maria Pokarowska  
*Warsaw University of Technology*

Follow this and additional works at:  [http://lib.dr.iastate.edu/bbmb_ag_pubs](http://lib.dr.iastate.edu/bbmb_ag_pubs)

Part of the [Biochemistry Commons](http://lib.dr.iastate.edu/bbmb_ag_pubs), [Biophysics Commons](http://lib.dr.iastate.edu/bbmb_ag_pubs), [Molecular Biology Commons](http://lib.dr.iastate.edu/bbmb_ag_pubs), and the [Structural Biology Commons](http://lib.dr.iastate.edu/bbmb_ag_pubs)

The complete bibliographic information for this item can be found at [http://lib.dr.iastate.edu/bbmb_ag_pubs/168](http://lib.dr.iastate.edu/bbmb_ag_pubs/168). For information on how to cite this item, please visit [http://lib.dr.iastate.edu/howtocite.html](http://lib.dr.iastate.edu/howtocite.html).
Inferring Ideal Amino Acid Interaction Forms From Statistical Protein Contact Potentials

Abstract
We have analyzed 29 different published matrices of protein pairwise contact potentials (CPs) between amino acids derived from different sets of proteins, either crystallographic structures taken from the Protein Data Bank (PDB) or computer-generated decoys. Each of the CPs is similar to 1 of the 2 matrices derived in the work of Miyazawa and Jernigan (Proteins 1999;34:49–68). The CP matrices of the first class can be approximated with a correlation of order 0.9 by the formula $e_{ij} = h_i + h_j$, $1 \leq i, j \leq 20$, where the residue-type dependent factor $h$ is highly correlated with the frequency of occurrence of a given amino acid type inside proteins. Electrostatic interactions for the potentials of this class are almost negligible. In the potentials belonging to this class, the major contribution to the potentials is the one-body transfer energy of the amino acid from water to the protein environment. Potentials belonging to the second class can be approximated with a correlation of 0.9 by the formula $e_{ij} = c_0 - h_i h_j + q_i q_j$, where $c_0$ is a constant, $h$ is highly correlated with the Kyte–Doolittle hydrophobicity scale, and a new, less dominant, residue-type dependent factor $q$ is correlated (~0.9) with amino acid isoelectric points pI. Including electrostatic interactions significantly improves the approximation for this class of potentials. While, the high correlation between potentials of the first class and the hydrophobic transfer energies is well known, the fact that this approximation can work well also for the second class of potentials is a new finding. We interpret potentials of this class as representing energies of contact of amino acid pairs within an average protein environment.

Keywords
protein folding, protein structure prediction, threading, residue-based contact potentials, statistical potentials, knowledge-based potentials

Disciplines
Biochemistry | Biophysics | Molecular Biology | Structural Biology

Comments
This is the peer reviewed version of the following article: Pokarowski, Piotr, Andrzej Kloczkowski, Robert L. Jernigan, Neha S. Kothari, Maria Pokarowska, and Andrzej Kolinski. “Inferring ideal amino acid interaction forms from statistical protein contact potentials.” PROTEINS: Structure, Function, and Bioinformatics 59, no. 1 (2005): 49-57, which has been published in final form at doi: 10.1002/prot.20380. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Self-Archiving.

Authors
Piotr Pokarowski, Andrzej Kloczkowski, Robert L. Jernigan, Neha S. Kothari, Maria Pokarowska, and Andrzej Kolinski

This article is available at Iowa State University Digital Repository: http://lib.dr.iastate.edu/bbmb_ag_pubs/168
Inferring Ideal Amino Acid Interaction Forms From Statistical Protein Contact Potentials

Piotr Pokarowski1,*, Andrzej Kloczkowski2, Robert L. Jernigan2, Neha S. Kothari2, Maria Pokarowska3, and Andrzej Kolinski4

1Institute of Applied Mathematics and Mechanics, Warsaw University, Warsaw, Poland 2Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa 3Faculty of Geodesy and Cartography, Warsaw University of Technology, Warsaw, Poland 4Laboratory of Theory of Biopolymers, Faculty of Chemistry, Warsaw University, Warsaw, Poland

Abstract

We have analyzed 29 different published matrices of protein pairwise contact potentials (CPs) between amino acids derived from different sets of proteins, either crystallographic structures taken from the Protein Data Bank (PDB) or computer-generated decoys. Each of the CPs is similar to 1 of the 2 matrices derived in the work of Miyazawa and Jernigan (Proteins 1999;34:49–68). The CP matrices of the first class can be approximated with a correlation of order 0.9 by the formula $e_{ij} = h_i + h_j$, 1 ≤ $i$, $j$ ≤ 20, where the residue-type dependent factor $h$ is highly correlated with the frequency of occurrence of a given amino acid type inside proteins. Electrostatic interactions for the potentials of this class are almost negligible. In the potentials belonging to this class, the major contribution to the potentials is the one-body transfer energy of the amino acid from water to the protein environment. Potentials belonging to the second class can be approximated with a correlation of 0.9 by the formula $e_{ij} = c_0 - h_i h_j + q_i q_j$, where $c_0$ is a constant, $h$ is highly correlated with the Kyte–Doolittle hydrophobicity scale, and a new, less dominant, residue-type dependent factor $q$ is correlated (~0.9) with amino acid isoelectric points pI. Including electrostatic interactions significantly improves the approximation for this class of potentials. While, the high correlation between potentials of the first class and the hydrophobic transfer energies is well known, the fact that this approximation can work well also for the second class of potentials is a new finding. We interpret potentials of this class as representing energies of contact of amino acid pairs within an average protein environment. Proteins 2005;59:49–57.

Keywords

protein folding; protein structure prediction; threading; residue-based contact potentials; statistical potentials; knowledge-based potentials

© 2005 Wiley-Liss, Inc.

*Correspondence to: Piotr Pokarowski, Institute of Applied Mathematics and Mechanics, Warsaw University, Banacha 2, 02-097 Warsaw, Poland. pokar@mimuw.edu.pl.
INTRODUCTION

Statistical pairwise contact potentials (CPs) of protein residues, have been derived either by using the quasichemical approximation from databases of proteins having known structures,\(^1\)–\(^{23}\) or by fitting their values to optimize the selection of the correct structures as the lowest energy forms in comparisons against sets of misfolded structures (decoys).\(^{24}\)–\(^{30}\) CPs have been increasingly heavily used over the last 20 years for ligand docking, fold recognition, and protein structure prediction from amino acid sequence (see review papers\(^{31}\)–\(^{33}\)). Analysis of results of the Critical Assessment of Techniques for Protein Structure Prediction (CASP) experiment shows that most of the successful groups use statistical CPs in their force fields for threading or ab initio protein structure prediction.\(^{34}\)–\(^{41}\)

Introducing the same level of coarse graining over structures, as is considered in protein sequences, has a major advantage for relating sequences to structures. It is well known that coarse graining of structures removes some of the specificity. For example, when the level of structural representation is 1 point per amino acid, then some of the details of backbone conformation are lost, but most of the information regarding side conformations is thrown away. Successful use of CPs coarse grained at this level relies upon the underlying assumption that the terms coming from the atomic details will be less important than the placement of the residues within the structure overall. While there has been no rigorous proof of this, there is now a large body of evidence in support of this view coming from the widespread use of the CPs.

In the present work, we first compare 29 different CPs currently used in computational biology. Each of these potentials is similar to one of the 2 matrices defined by Miyazawa and Jernigan.\(^{22}\) We then show that the actual contribution of specific two-body interactions to CPs is quite insignificant. The issue regarding higher body terms, of course, remains open.\(^{42}\) Nonetheless, all the known pairwise matrices of CPs can be surprisingly well approximated by simple functions of individual residue properties, such as hydrophobicity and electrostatic properties (in pH units given as isoelectric points pI),\(^{43}\)–\(^{47}\) for each pair of amino acids. We term such an approximation of the CP matrices a one-body approximation. Hydrophobicity represents the dominant factor in protein potentials, with other, less important factors being the energy of demixing of amino acids in a protein environment and electrostatic interactions. As we will see, the accuracy of the one-body approximation works significantly better for potentials derived from the quasi-chemical principle than for the potentials obtained from the optimization of the prediction of the native structures among decoys. This calls into question the quality of the decoys in general compared to the known structures. It is quite interesting that the frequencies of contacts between different amino acids can also be successfully approximated with the present method. Thus, hydrophobicity, demixing, and electrostatics are identifiable as fundamental properties defining potentials from the simple statistics of inter-residue pair contacts for proteins in the Protein Data Bank (PDB). Furthermore, an appropriate function form for combining these terms is obtained in the present work.

The one-body approximation helps us to comprehend the separation of the CPs into 2 classes. Potentials belonging to the first class are dominated by the one-body energies of
transfer of amino acids from water to a protein environment. The matrices \((e_{ij})\), \(1 \leq i, j \leq 20\) representing this class of CPs can be approximated with the formula \(e_{ij} = h_i + h_j\), where the residue-type dependent coefficient \(h\) strongly correlates with the frequency of occurrence of a given amino acid type inside proteins. Potentials belonging to the second class represent mostly energies of contacts of amino acids in a protein environment. The second class of potentials can be approximated with a correlation of order 0.9 by the formula \(e_{ij} = c_0 - h_i h_j + q_i q_j\). Here residue-type dependent factors \(h\) and \(q\) are highly correlated (both with a correlation of order 0.9) with the Kyte–Doolittle hydrophobicity scale and electrostatic property \(pI\), respectively, and \(c_0\) is a constant. The electrostatic properties of an amino acid are represented by its isoelectric point and measured in pH units. The electrostatic interactions are quite important for the second class of potentials but are completely negligible for the first class of potentials.

The high correlation between potentials of the first class and transfer energies is well known. It seems somewhat surprising that the one-body approximation works well also for the second class of potentials, because these potentials have frequently been derived by excluding hydrophobic interactions. It can be shown that the term \(c_0 - h_i h_j\) (including hydrophobicity and energy of demixing) describes the dominant property of amino acid interactions in the protein environment leading to attraction between hydrophobic/polar like-type residues and repulsion between unlike-type residues that gives the spatial segregation between a protein’s hydrophobic interior and polar surface. It is interesting that similar long-range interactions come from our minimal model of protein folding.\(^{48,49}\)

The explicit inclusion of hydrophobic and electrostatic interactions allows us to analyze and compare various statistical potentials for proteins. On the other hand, a statistical analysis of the frequencies of pairwise residue contacts in protein structures leads to the derivation of explicit forms, which can be correlated and compared against various experimental scales of hydrophobicity and \(pI\). Validating the potential form in this way permits us to comprehend these complex interactions.

**METHODS**

Because of the symmetry, we identify the matrix \(E = (e_{ij})\) of contact potentials with its upper diagonal part \((e_{ij})_{i \leq j}\). Our aim is to find a simple function \(\tilde{E}(h,q) = [\tilde{e}(h,q)]_{ij}\) of two 20-dimensional vectors \(h\) and \(q\) (properties of the 20 amino acids) that minimizes the sum of squares

\[
\sum_{i,j:i \leq j} \left[ e_{i,j} - \tilde{e}(h,q)_{i,j} \right]^2 \rightarrow \min_{h,q} \quad (1)
\]

This defines the well known least squares problem.

Accuracy of the approximation is measured by the correlation coefficient, the relative Euclidean distance and the mean Euclidean distance between normalized matrices \(E\) and \(\tilde{E}\). Specifically, let us denote the scalar product of vectors \(x\) and \(y\) as \(\langle x, y \rangle\) and the norm of \(x\) as \(\|x\| = \sqrt{\langle x, x \rangle}\). The normalization of \(x\) is given by the vector \(x^N = (x - \bar{x})/\sigma_x\), where \(\bar{x}\) is the
mean value of \( x \) and \( \sigma_x \) is its standard deviation, respectively. The correlation between vectors \( x \) and \( y \) is defined as

\[
\text{cor}(x, y) = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{n \sigma_x \sigma_y} = \frac{\langle x - \bar{x}, y - \bar{y} \rangle}{\|x - \bar{x}\| \|y - \bar{y}\|}. \tag{2}
\]

We define also the distance between normalized vectors \( \text{dist}(x, y) = \|y^N - x^N\| / \sqrt{n} \), where \( n \) is the dimensionality of vectors \( x \) and \( y \) (210 in our case). Obviously \( \text{dist}(x, y) \) is simply the root-mean-square difference between \( y^N \) and \( x^N \). In numerical analysis, it is popular to define the relative error of approximation of the vector \( y \) by \( x \) as \( \text{err}(x, y) = \|y - x\| / \|y\| \).

It is worth noting that all of the above defined measures of the quality of approximation (\( \text{cor}, \text{dist}, \text{err} \)) are invariant to multiplication by a scalar and are optimized by the solution of the least squares problem (Eq. 1). Additionally it is easily seen that \( \text{dist}^2(x, y) = 2 - 2 \text{cor}(x, y) \).

To approximate \( E \), we investigate the following 4 simple functions:

\[
\hat{e}(h)_{i,j} = h_i + h_j (Hp) \tag{3a}
\]

\[
\hat{e}(h)_{i,j} = c_0 + c_1 h_i h_j (Hp.Dx) \tag{3b}
\]

\[
\hat{e}(h, q)_{i,j} = h_i + h_j + c_0 q_i q_j (Hp.pH) \tag{3c}
\]

\[
\hat{e}(h, q)_{i,j} = c_0 + c_1 h_i h_j + c_2 q_i q_j (Hp.Dx.pH). \tag{3d}
\]

The above approximations are related to each other as follows:

\[(Hp) \rightarrow (Hp.Dx) \rightarrow (Hp.pH) \rightarrow (Hp.Dx.pH) \tag{4}\]

The relation (a) \( \rightarrow \) (b) means that formula (b) is more general than (a); the proof is given in the Appendix. The simplest additive approximation is given by the vector \( h \), which is often highly correlated with empirical hydrophobicities. Thus, we denote this approximation [Eq. 3(a)] by (Hp). The solution of the least squares problem for this case [Eq. 3(a)] leads to linear equations that can be solved analytically:

\[
h_i = (s_i - c_0) / (n+2) \tag{5}
\]

with \( n = 20, s_i = \sum e_{i,j} + e_{i,i} \) and \( c_0 = (\sum_{i,j} g e_{i,j}) / (n+1) \). All the other approximations given by Eqs. (3b)–(3d) lead to nonlinear least squares problems and require numerical solutions. We used the free software R (www.r-project.org) and Matlab with optimization toolbox (Mathworks, Inc; www.mathworks.com) in our computations. Four vectors were used as a starting solution for vector \( h \): the vector defined by Eq. (5), the diagonal (\( e_{ii} \)), the eigenvectors of \( E \), and the centered \( E \) (matrix obtained from \( E \) by subtracting the mean value) corresponding to the dominant eigenvalues. As the starting solution for \( q \) we used

Proteins. Author manuscript; available in PMC 2015 May 03.
To check the dependence of solutions of Eq. (3c) on the starting points, we interchanged vectors \( h \) and \( q \). This optimization is denoted as \((\text{pH.Hp})\). Generally, there is no significant dependence on starting vectors or the software used. The only exceptions were for MJ3h, TEl, B4, B5, and MSBM. On the other hand, for many CPs, the differences between solutions for \((\text{Hp.pH})\) and \((\text{pH.Hp})\) are essential.

Let us note that Eq. (3b) can be written in the following form:

\[
\tilde{e}(h, q)_{i,j} = h'_i + h'_j - c_2(h_i - h_j)^2/2, \quad (6)
\]

where \( h'_i = c_0/2 + c_1 h_i + (c_2/2)h_i^2 \), as shown by Li et al.\(^{51}\). Eq. (6) explains the physical nature of this approximation: The hydrophobic potential \( h' \) is supplemented by the energy of demixing, well known from the Hildebrand theory of solutions that favors structures with spatial segregation of amino acids. Thus, we use the notation \((\text{Hp.Dx})\) for approximation (3b). In the next approximation [Eq. (3c)], the hydrophobicity is supplemented by electrostatic interactions, where electrostatics are linearly related to the experimental isoelectric points \( \pi \) of amino acids,\(^{43-47}\) measured in pH units. This describes the abbreviation \((\text{Hp.pH})\) used for approximation (3c). The last approximation, (3d) \((\text{Hp.Dx.pH})\), contains all 3 elements—hydrophobicity, energy of demixing, and electrostatics—and therefore actually gives the best results.

The numerical experiments performed for a variety of known CPs have shown that more complex approximations are not necessary, and the inclusion of higher order terms in the Taylor expansion of the function \( e(h, q) \) did not in general lead to the significant increase of correlations with \( E \).

**RESULTS AND DISCUSSION**

**Pairwise Contact Potentials Studied**

In this article, we have studied mostly new potentials developed since 1995, and used by groups that were the most successful in predicting protein structures from the amino acid sequence in recent CASP experiments. We have included also a few older, historically important potentials. The total number of potentials analyzed in this work is 29, listed and abbreviated as follows:

- **TS**—the oldest statistical potential derived by Tanaka and Scheraga.\(^{14}\) We also analyze the matrices \( N_{\text{TS}} = (N_{ij}) \) and \( \ln N_{\text{TS}} = [\log (N_{ij})] \), where \( N_{ij} \) is the number of contacts between amino acids \( i \) and \( j \).
- **RO**—the matrix developed by Robson and Osguthorpe.\(^ {15}\) This potential has been applied by Bates and coworkers\(^ {52}\) for threading and used quite successfully in CASP5.
- **BL**—distance-dependent statistical potential proposed by Bryant and Lawrence.\(^ {3}\) We have used the energies from the first bin only (contacts within 5 Å). Matrix BL has been used for threading,\(^ {37}\) and most recently by Fang and Shortle\(^ {35}\) in their ab initio method.
• TD—mixed quasi-optimization potential developed by Thomas and Dill.26

• MS—optimization-based potential derived by Mirny and Shakhnovich25 by the maximization of the harmonic mean of Z scores for decoys.

• VD—effective optimization-based potential constructed on the perceptron criterion proposed by Vendruscolo and Domany.28 The VD potential is based on the first optimization-derived potential of Maiorov and Crippen.24 The comparison of MS and VD is given in Vendruscolo et al.29

• BFKV—effective optimization-derived potential that is a modified version of VD.53

• MJ1, MJ1h, MJ2, MJ2h, MJ3, MJ3h—Miyazawa–Jernigan potentials published in 1985,18 1996,21 and 1999.22 Each Miyazawa–Jernigan article contains a derivation of 2 potentials: one including energy of transfer of amino acids from water to the protein environment (those are marked with the suffix “h”), and another for interactions in an average buried environment. Because MJ1h has a correlation of 0.97 with MJ2h, we have studied only MJ2h. A modified version of MJ1h potential has been used by Liwo and coworkers in the ab initio UNRES method.17,38 Because the last potential has a correlation of 0.97 with MJ1h, we have omitted it in the comparative analysis. Similarly, because potentials MJ2 and MJ3 are highly correlated (0.994), we have studied the newest potential, MJ3, only. It is worth mentioning that potentials MJ1h and MJ2h are the most frequently analyzed, modified, and used in protein structure predictions.5,12,16,19,20,51,54,55 Matrices N.MJ2 and lN.MJ2, with number of contacts and logarithms of the number of contacts, will also be investigated.

• BT—potential developed by Betancourt and Thirumalai,16 which is a modified version of MJ2h.

• TEl, TEs—effective optimization-derived potentials proposed by Tobi et al.27 based on the mixed perceptron–Z-score criterion. TEl and TEs are potentials obtained for large and small sets of decoys, respectively.

• MJPL, HLPL—potentials developed by Park and Levitt.5 MJPL is a modified version of MJ1h, while HLPL is an improvement of an earlier potential of Hinds and Levitt.2 These CPs are part of a hierarchical method of ab initio protein structure prediction.41

• GKS—quasi-chemical statistical potential of Godzik et al.4

• SJKG, SKOa, SKOb—quasi-chemical CPs of Skolnick et al.8,11

• Qa, Qm, Qp—new quasi-chemical potentials developed by Kolinski and coworkers,13 which depend on the relative orientation of side-chains of 2 contacting residues. Three different possible mutual orientations of the interacting side groups (a, antiparallel; m, intermediate; p, parallel) were considered.13 Qa, Qm, and Qp potentials were used in TOUCHSTONE, one of the most effective methods for structure prediction, as proven during the CASP5 experiment.40 Additionally, such environment-dependent potentials may include both the group
orientation and the two-state (compact/extended) main-chain conformation information.\textsuperscript{23} Since it leads to 40 × 40 matrices QaS, QmS, and QpS, these generalized potentials are not included here in the comparative analysis.

- **B1,…,B5**—the newest version of quasi-chemical potential developed in the research group of Baker. Earlier versions of this potential were discussed in Simons et al.\textsuperscript{9,10} The potential is distance dependent: Distance bins are denoted by increasing integer numbers. The potentials are a part of ROSETTA, currently the most successful protocol for ab initio prediction of protein structure from sequence.\textsuperscript{34,39}

- **MSBM**—optimization-derived potential developed by Micheletti et al.\textsuperscript{30}

### Comparative Analysis of One-Body Approximation for Pairwise Contact Potentials

Table I shows the results of calculations performed for all of the 29 potentials and matrices with numbers of contacts. The entries below the diagonal show correlation coefficients (\textit{cor}) between potentials, while the entries above the diagonal list the mean Euclidean distances between the normalized potentials (\textit{dist}). Potentials developed by the optimization methods are marked in blue. Figure 1 graphically illustrates the results from Table I. Each of the 28 potentials listed is represented by a node of the graph. (MSBM, N.TS, IN.TS, N.MJ2 and lN.MJ2 are not included because of their small correlations with other potentials.) All the strongest correlations of the order 0.9–1.0 are visualized as graph edges. Lower correlations have not been shown for reasons of clarity, since, often, if \textit{cor}(a, b) \geq 0.9 and \textit{cor}(b, c) \geq 0.9, then \textit{cor}(a, c) \geq 0.8. In the case where a given node (potential) is not correlated with other nodes by at least a value of 0.9, we show the first 2–4 edges connecting to nodes with the highest correlation (we use colors to indicate the different ranges of correlation). Table I and Figure 1 show that CPs can be clustered into 2 groups. The first cluster is centered on MJ3h and SJKG, and the second one around MJ3. Using a rule, that each potential in the group has to be correlated at the level of at least 0.9 with a neighbor, the following 2 sets clearly arise: \{TS, MJ2h, MJ3h, MJPL, SJKG, SKOa, SKOb, HLPL, Qp, Qa, Qm, BT\}, \{MJ3, MJ1, MS, GKS, B1, B2\}. Potentials in the second set (except MS) were designed to diminish the influence of hydrophobic interactions, by considering contacts for buried residues only, and by a proper definition of the reference state.

Table II shows \textit{cor}, \textit{dist}, and \textit{err} between analyzed potentials and their one-body approximations. By comparing columns (Hp) with (Hp.Dx) and (Hp.pH) with (Hp.Dx.pH), we can estimate demixing energies, while the comparison of columns (Hp) with (Hp.pH) and (pH.Hp), and (Hp.Dx) with (Hp.Dx.pH) enables the evaluation of the strength of electrostatics in the protein potentials. Columns 17–19 contain errors of approximation by the formula (Hp.Dx.pH) for suboptimal solutions (\(h,q\)) that have significant correlations with hydrophobicity (Hp) and isoelectric points (pH). Column 20 of Table II shows correlations between approximating vectors \(h\) from the formula (Hp.Dx.pH) and the closest hydrophobicity scale (with negative sign). Forty hydrophobicity scales with correlations greater than 0.68 compared to the Kyte–Doolittle scale were selected from literature. Column 21 contains identifying numbers of the closest hydrophobicity scales. All numerical data and detailed references are available as Supplementary Materials to this article, which
The last column in Table II displays correlations between vectors $q$ and isoelectric points of amino acids pI (pH). The major conclusions from the analysis of Table II are as follows:

- All potentials (except TEl, TEs, VD, B4, and MSBM) can be quite well approximated by simple functions of one-body factors $h$ and $q$ that are highly correlated with hydrophobicities and isoelectric points of amino acids. Indeed, a correlation between approximating vectors $h$ and the closest hydrophobic scale is roughly 0.9, which is more than a mean correlation between 2 different hydrophobic scales. Because of this we may interpret vectors $h$ as statistical hydrophobicity scales. Hydrophobicity is the most dominant factor in protein potentials, much more important than electrostatics or demixing energy.

- By comparing rows in Table II, we may group together potentials having similar characteristics, similar to what was done earlier with Table I. The first group of potentials (rows Qa–BFKV) is dominated by the one-body transfer energy. Indeed, the correlation coefficients with the simplest approximation ($H_p$) are mostly above 0.9, values only slightly smaller than the correlations with ($H_p,D_x,pH$). The contributions from demixing and electrostatics are negligible. The nearest hydrophobicity scale to the vectors $h$ from this group is the Wertz–Scheraga frequency of occurrence for a given type residue inside proteins (scale no. 26). The second group of potentials (rows MS–B5) is poorly approximated by the ($H_p$) formula (correlation coefficient range is from 0.2 to 0.3, except 0.5 for GKS). Using the ($H_p,D_x$) formula, the correlation increases to about 0.8, and to about 0.9 for the ($H_p,D_x,pH$) formula. In the last formula the $c_1$ coefficients are negative, while the $c_2$ coefficients have positive values. Thus, the ($H_p,D_x,pH$) formula can be written in simplified form as

$$\tilde{e}(h,q)_{i,j} = c_0 - h_i h_j + q_i q_j.$$  \hspace{1cm} (7)

The large contributions of the demixing term means that the $c_0 - h_i h_j$ term describes the effects of interactions already present within the protein environment, where similar residues (hydrophobic or polar) are pairwise attractive, while the interactions between polar and hydrophobic residues are repulsive. Interestingly, similar interactions are necessary for proteinlike folding thermodynamics in our minimal model of proteins. The most correlated hydrophobicity scale with solutions for this group of potentials is the popular Kyte–Doolittle scale (scale no. 1).

- Potentials developed by optimization methods have significantly less hydrophobic character than do the quasichemical potentials, and additionally are less stabilizing. The only exceptions are MS (with correlation 0.97 with MJ1) and BFKV (with correlation of order 0.8 with MJ3h, MJPL and SJKG). Optimization-based approaches may additionally lead to many surprising counterintuitive results. For example, in the MSBM potential, the energy of the contact TRP-MET exceeds more than 20 times all other contact energies. That may be a reason that such
potentials are seldom used with much success for fold recognition, prediction of protein structure, or docking.

- Let us take a closer look at the potentials MJ3h, MJ3 and SJKG, which are centers of the groups. Potential MJ3h was derived by using the formula $e_{ij} = -\log(N_{ij}/C_{ij}) - h_i - h_j$, where $N_{ij}$ denotes the number of observed contacts and $C_{ij}$ the number of expected contacts between residues $i$ and $j$, while $h_i$ is a one-body potential highly correlated with hydrophobicity (“h” in the name of the potential refers to its hydrophobicity-driven nature). On the other hand, potential MJ3 has been derived from the formula $e_{ij} = -\log(N_{ij}/C_{ij})$, that could allow us to estimate the influence of $h$ terms on the one-body approximation. Interestingly, the potential MJ3 have a relatively negligible correlation with the Wertz–Scheraga scale, however, the results of our approximation show that hydrophobicity still remains, in a form highly correlated with the Kyte–Doolittle scale.

- By comparing potentials SJKG, SKOa, and SKOb, one may notice that the composition corrected potential (SKOb) that was derived to increase its specificity is actually less specific than the simplest quasi-chemical potential obtained from the same set of protein structures (SJKG), though, of course, it could be more effective for the prediction of protein structure. We define here the specificity of CPs, through low correlations with their one-body approximations. We may notice that potentials SKOb and Qa are roughly equivalent, which means that an antiparallel orientation does not add to the specificity. The specificity of two-body interaction is, however, increased for parallel orientations (compare SKOb with Qp). Also, the inclusion of the backbone geometry characteristics increases specificity of the potential (data not shown).

The last four rows of Table II show the correlation between the number of contacts $N$ or as log $N$ and one-body approximations for potentials TS and MJ2. It is seen that the frequency of contacts of amino acids can be well approximated by hydrophobicity and electrostatic properties. Note that hydrophobicity and electrostatics in the CPs result not from sophisticated manipulations of the reference state (extensively studied in the past in the literature) but simply from the frequencies of contacts in protein structures. Interestingly, vectors $h$ approximating $N$ or log $N$ correlate less with hydrophobicity than vectors $h$ approximating CPs. To the contrary, vectors $q$ approximating $N$ or log $N$ correlate more strongly with pI (pH) than do the corresponding vectors approximating CPs. Comparison of columns (Hp.Dx) and (Hp.Dx.pH) shows that electrostatic interactions are almost negligible.

Approximation of the CPs by one-body amino acid functions was studied earlier by Godzik et al.\textsuperscript{4} and by Li et al.\textsuperscript{51} The present results are, however, stronger and more universal. The main point of the work of Godzik et al.\textsuperscript{4} was to compare known potentials and to discover their relationships with hydrophobicity (Table II). Their second aim, namely, the derivation of the excess part of the potential $e_{ij}^{\text{excess}} = e_{ij} - e_{ij}^{\text{ideal}}$, where $e_{ij}^{\text{ideal}} = (e_{ii} + e_{jj})/2$ was not completed, as can be seen clearly from our present results. The authors found that the correlation of GKS with $e_{ij}^{\text{ideal}}$ is only 0.21, and that led to their mistaken conclusion that their potential was more specific than, for example, the TS potential having a correlation.
0.98. The reason for this poor correlation lies in the formulation of the ideal values. Indeed, our Table II shows that GKS potential can be well approximated by one-body functions by including both hydrophobicity and electrostatics.

The major advantage of the work of Li et al.\textsuperscript{51} was the derivation of a better (in general) approximating vector than \((e_{ij})\) and a richer approximating formula \((Hp.Dx)\). The approximating vector they used was the eigenvector of the dominant eigenvalue of the matrix obtained from the matrix of the potential with the mean value subtracted. However, such an approximation can sometimes be significantly worse than the optimal one (e.g., for MSBM, we obtain in this way \(cor = 0.66\) instead of 0.997, which was found with the optimization formula \(Hp.Dx.pH\)).

**CONCLUSIONS**

It has been shown that all analyzed CPs can be divided into two groups, regardless of having completely different derivation origin. Most of these knowledge-based statistical potentials could be well approximated by appropriate combinations of one-body components. The one body approximation suggests the two following ideal amino acid interaction forms:

- Let \(h\) be a vector composed of the normalized Wertz–Scheraga interior frequency coefficients with negative signs. Then the formula \(e_{ij} = h_i + h_j\) gives a potential that belongs to the first group (e.g., correlation with MJ2h and TS are 0.90 and 0.88, respectively).

- Let \(h\) be a vector composed of the normalized Kyte–Doolittle coefficients with negative signs and \(q\) be the normalized isoelectric point (pH) vector. Then the potential \(e_{ij} = -h_i h_j + 0.5 q_i q_j\) correlates moderately well with members of the second group of CPs \((cor = 0.66, 0.60, \text{ and } 0.59\) for MJ3, MJ1, and B2, respectively).

From a practical point of view, the accurate one-body approximations of CPs provided in this work could be very useful in some applications, especially for 3-dimensional threading algorithms. On the other hand the lack of “excess” contributions to the pairwise potentials (that cannot be approximated by the one-body component) strongly suggests that an efficient structure-specific, knowledge-based pairwise potential is still to be designed. This means that there are opportunities to develop different further types of potentials (perhaps multibody).

**Supplementary Material**

Refer to Web version on PubMed Central for supplementary material.

**ACKNOWLEDGMENTS**

Our thanks to Sanzo Miyazawa for helpful discussions, and to the Reviewers for insightful comments.

Grant sponsor: Polish Research Council KBN; Grant number: PBZ-KBN-088/P04/2003 (to A. Kolinski).
REFERENCES


APPENDIX

1. In order to prove the relation \((H_p) \rightarrow (H_p.D_x)\), let us first notice that \((H_p.D_x)\) is equivalent to the following formula:

\[
a_0 + a_1(h_i' + h_j') + a_2 h_i' h_j'.(H_p.D_x.2) \quad (A1)
\]

Obviously \((H_p.D_x.2)\) can be transformed to \((H_p.D_x)\) with \(a_0 := c_0, a_1 := 0, a_2 := c_1, h' := h\). To obtain the inverse transformation \((H_p.D_x)\) to \((H_p.D_x.2)\) let us denote:

\[
c_0 := a_0 - a_2^2 / a_2, c_1 := a_2, c_2 := -a_1 / a_2, h_i' := h_i' - c_2. \quad (A2)
\]

Then:
For any given vector $h'$ and coefficients $a_0$, $a_1$, and $a_2$, expression $(Hp.Dx.2)$ can be written as $(Hp.Dx)$, with $h$, $c_0$, and $c_1$ given by A2. Now it is enough to show that $(Hp) \rightarrow (Hp.Dx.2)$. In expressions $(Hp.Dx)$ and $(Hp.Dx.2)$ coefficients $c_1$ and $a_2$ are nonzero, taking the limit $a_2 \rightarrow 0$ and substituting, we obtain from $(Hp.Dx.2)$ an expression that is infinitely close to $(Hp)$.

2. Now let us assume that we have the solution of $(Hp.Dx)$ for $c_0 + c_1 h_i h_j$. Making substitutions $h_i' := c_0/2, c_1' := c_1$, and $q_i' := h_i$, we can transform Eq. (3b) to the $(Hp.pH)$ form $h_i + h_j + c_{0q} q_i q_j$. This proves the relation $(Hp.Dx) \rightarrow (Hp.pH)$.

3. The relation $(Hp.pH) \rightarrow (Hp.Dx.pH)$ is derived similarly as $(Hp) \rightarrow (Hp.Dx)$ by proving first that $(Hp.Dx.pH)$ is equivalent to the following formula:

$$a_0 + a_1 (h_i' + h_j') + a_2 h_i h_j + a_3 (q_i' + q_j') + a_4 q_i q_j', \quad (A4)$$

Assuming that $c_0 := a_0 - a_2^2 / a_2 - a_3^2 / a_4, c_1 := a_2, c_2 := -a_1 / a_2, c_4 := -a_3 / a_4, h_i := h_i', c_2' := c_2$, and transforming $(Hp.Dx.pH)$ similarly, as in A1–A3, we obtain A4. In the limit $a_1 \rightarrow 0, a_3 \rightarrow 0$, we obtain from A4 an expression that is infinitely close to $(Hp.pH)$. 

Proteins. Author manuscript; available in PMC 2015 May 03.
Fig. 1.
Graphical illustration of correlations among different protein potentials. Coloring scheme is the same as in Table I.
TABLE I

Correlations between CPs for Lower Triangular Part and Distances Between Normalized Potentials for Upper Triangular Part, for a Convenience of Easy Comparison the Results are Multiplied by Factor 100 and the Coloring Scheme Explained Below the Table is Used. Potentials Derived by Optimization are Marked in Blue.
TABLE II

One-Body Approximations to Protein Contact Potentials (CPs)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Qa</td>
<td>91</td>
<td>42</td>
<td>41</td>
<td>96</td>
<td>32</td>
<td>31</td>
<td>99</td>
<td>31</td>
<td>30</td>
<td>76</td>
<td>27</td>
<td>62</td>
</tr>
<tr>
<td>Qm</td>
<td>86</td>
<td>53</td>
<td>51</td>
<td>91</td>
<td>43</td>
<td>42</td>
<td>93</td>
<td>38</td>
<td>37</td>
<td>91</td>
<td>43</td>
<td>42</td>
</tr>
<tr>
<td>Qp</td>
<td>80</td>
<td>64</td>
<td>49</td>
<td>83</td>
<td>58</td>
<td>41</td>
<td>85</td>
<td>54</td>
<td>38</td>
<td>93</td>
<td>38</td>
<td>28</td>
</tr>
<tr>
<td>HPL</td>
<td>62</td>
<td>50</td>
<td>42</td>
<td>65</td>
<td>54</td>
<td>38</td>
<td>93</td>
<td>38</td>
<td>28</td>
<td>96</td>
<td>53</td>
<td>38</td>
</tr>
<tr>
<td>SKQ</td>
<td>91</td>
<td>43</td>
<td>40</td>
<td>91</td>
<td>42</td>
<td>39</td>
<td>92</td>
<td>28</td>
<td>27</td>
<td>92</td>
<td>41</td>
<td>37</td>
</tr>
<tr>
<td>SKQa</td>
<td>86</td>
<td>53</td>
<td>51</td>
<td>87</td>
<td>57</td>
<td>49</td>
<td>89</td>
<td>48</td>
<td>38</td>
<td>94</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>SJKG</td>
<td>88</td>
<td>50</td>
<td>48</td>
<td>89</td>
<td>47</td>
<td>45</td>
<td>90</td>
<td>46</td>
<td>44</td>
<td>93</td>
<td>32</td>
<td>31</td>
</tr>
<tr>
<td>MJPL</td>
<td>88</td>
<td>50</td>
<td>11</td>
<td>90</td>
<td>44</td>
<td>10</td>
<td>95</td>
<td>46</td>
<td>10</td>
<td>93</td>
<td>38</td>
<td>37</td>
</tr>
<tr>
<td>MJ3n</td>
<td>86</td>
<td>53</td>
<td>49</td>
<td>93</td>
<td>39</td>
<td>37</td>
<td>88</td>
<td>49</td>
<td>48</td>
<td>90</td>
<td>30</td>
<td>28</td>
</tr>
<tr>
<td>MJ2h</td>
<td>89</td>
<td>19</td>
<td>8</td>
<td>93</td>
<td>14</td>
<td>6</td>
<td>99</td>
<td>14</td>
<td>6</td>
<td>93</td>
<td>11</td>
<td>5</td>
</tr>
<tr>
<td>DT</td>
<td>98</td>
<td>18</td>
<td>4</td>
<td>93</td>
<td>17</td>
<td>4</td>
<td>99</td>
<td>16</td>
<td>4</td>
<td>93</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>BFVK</td>
<td>76</td>
<td>76</td>
<td>69</td>
<td>80</td>
<td>63</td>
<td>58</td>
<td>74</td>
<td>72</td>
<td>68</td>
<td>84</td>
<td>49</td>
<td>46</td>
</tr>
<tr>
<td>TD</td>
<td>72</td>
<td>75</td>
<td>66</td>
<td>75</td>
<td>71</td>
<td>63</td>
<td>77</td>
<td>68</td>
<td>60</td>
<td>82</td>
<td>60</td>
<td>54</td>
</tr>
<tr>
<td>TEI</td>
<td>59</td>
<td>91</td>
<td>80</td>
<td>61</td>
<td>89</td>
<td>79</td>
<td>68</td>
<td>82</td>
<td>75</td>
<td>71</td>
<td>76</td>
<td>70</td>
</tr>
<tr>
<td>TEs</td>
<td>56</td>
<td>94</td>
<td>82</td>
<td>64</td>
<td>84</td>
<td>76</td>
<td>61</td>
<td>89</td>
<td>79</td>
<td>71</td>
<td>76</td>
<td>70</td>
</tr>
<tr>
<td>RO</td>
<td>73</td>
<td>73</td>
<td>66</td>
<td>75</td>
<td>70</td>
<td>64</td>
<td>93</td>
<td>39</td>
<td>37</td>
<td>93</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>MS</td>
<td>22</td>
<td>125</td>
<td>97</td>
<td>65</td>
<td>80</td>
<td>73</td>
<td>42</td>
<td>108</td>
<td>91</td>
<td>71</td>
<td>76</td>
<td>70</td>
</tr>
<tr>
<td>MJ</td>
<td>22</td>
<td>125</td>
<td>13</td>
<td>70</td>
<td>75</td>
<td>8</td>
<td>41</td>
<td>109</td>
<td>91</td>
<td>72</td>
<td>74</td>
<td>9</td>
</tr>
<tr>
<td>MJ3</td>
<td>21</td>
<td>126</td>
<td>98</td>
<td>80</td>
<td>63</td>
<td>60</td>
<td>41</td>
<td>109</td>
<td>91</td>
<td>82</td>
<td>60</td>
<td>57</td>
</tr>
<tr>
<td>GKS</td>
<td>50</td>
<td>100</td>
<td>86</td>
<td>85</td>
<td>56</td>
<td>53</td>
<td>58</td>
<td>93</td>
<td>82</td>
<td>67</td>
<td>50</td>
<td>48</td>
</tr>
<tr>
<td>B2</td>
<td>34</td>
<td>115</td>
<td>94</td>
<td>86</td>
<td>52</td>
<td>51</td>
<td>46</td>
<td>104</td>
<td>99</td>
<td>88</td>
<td>48</td>
<td>47</td>
</tr>
<tr>
<td>B1</td>
<td>21</td>
<td>126</td>
<td>98</td>
<td>68</td>
<td>80</td>
<td>73</td>
<td>50</td>
<td>78</td>
<td>72</td>
<td>83</td>
<td>58</td>
<td>55</td>
</tr>
<tr>
<td>B3</td>
<td>31</td>
<td>117</td>
<td>95</td>
<td>86</td>
<td>52</td>
<td>50</td>
<td>35</td>
<td>114</td>
<td>93</td>
<td>88</td>
<td>50</td>
<td>48</td>
</tr>
<tr>
<td>B5</td>
<td>26</td>
<td>122</td>
<td>96</td>
<td>87</td>
<td>51</td>
<td>50</td>
<td>26</td>
<td>122</td>
<td>96</td>
<td>87</td>
<td>51</td>
<td>49</td>
</tr>
<tr>
<td>VD</td>
<td>29</td>
<td>119</td>
<td>96</td>
<td>47</td>
<td>103</td>
<td>88</td>
<td>45</td>
<td>105</td>
<td>90</td>
<td>55</td>
<td>94</td>
<td>83</td>
</tr>
<tr>
<td>BL</td>
<td>10</td>
<td>134</td>
<td>100</td>
<td>52</td>
<td>97</td>
<td>85</td>
<td>51</td>
<td>99</td>
<td>86</td>
<td>75</td>
<td>71</td>
<td>66</td>
</tr>
<tr>
<td>BN</td>
<td>24</td>
<td>120</td>
<td>96</td>
<td>65</td>
<td>85</td>
<td>76</td>
<td>68</td>
<td>80</td>
<td>73</td>
<td>64</td>
<td>85</td>
<td>70</td>
</tr>
<tr>
<td>MBSM</td>
<td>33</td>
<td>116</td>
<td>94</td>
<td>66</td>
<td>82</td>
<td>75</td>
<td>65</td>
<td>82</td>
<td>75</td>
<td>68</td>
<td>65</td>
<td>75</td>
</tr>
<tr>
<td>N.TS</td>
<td>84</td>
<td>57</td>
<td>36</td>
<td>91</td>
<td>43</td>
<td>28</td>
<td>84</td>
<td>57</td>
<td>36</td>
<td>91</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>IN.TS</td>
<td>89</td>
<td>47</td>
<td>10</td>
<td>87</td>
<td>49</td>
<td>10</td>
<td>90</td>
<td>46</td>
<td>10</td>
<td>90</td>
<td>59</td>
<td>12</td>
</tr>
<tr>
<td>N.MJ2</td>
<td>85</td>
<td>57</td>
<td>36</td>
<td>91</td>
<td>43</td>
<td>28</td>
<td>91</td>
<td>43</td>
<td>27</td>
<td>91</td>
<td>43</td>
<td>27</td>
</tr>
<tr>
<td>IN.MJ2</td>
<td>89</td>
<td>47</td>
<td>10</td>
<td>90</td>
<td>44</td>
<td>10</td>
<td>91</td>
<td>43</td>
<td>27</td>
<td>91</td>
<td>43</td>
<td>27</td>
</tr>
</tbody>
</table>

Columns 2–16 contain values of optimal cor, dist, and err between particular CPs and their approximations. Columns 17–19 contain values of cor, dist, and err for suboptimal solutions (h,q), which have significant correlations with hydrophobicity (column 20) and isoelectric points pI (column 22). Column 21 contains numbers designating the closest hydrophobicity scales (details explained in the text). The scaling factor and the coloring scheme are the same as in Table I.