Inferring Ideal Amino Acid Interaction Forms From Statistical Protein Contact Potentials

Piotr Pokarowski,[1]* Andrzej Kloczkowski,[2] Robert L. Jernigan,[2] Neha S. Kothari,[2] Maria Pokarowska,[3] and Andrzej Kolinski[4]

[1]*Institute of Applied Mathematics and Mechanics, Warsaw University, Warsaw, Poland*
[2]*Laurence H. Baker Center for Bioinformatics and Biological Statistics, Iowa State University, Ames, Iowa*
[3]*Faculty of Geodesy and Cartography, Warsaw University of Technology, Warsaw, Poland*
[4]*Laboratory 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 $e_{ij} = h_i + h_j$, $1 \le i, j \le 20$, 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.**

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

INTRODUCTION

Statistical pairwise contact potentials (CPs) of protein residues, have been derived either by using the quasi-chemical 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 repre-

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

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

Received 1 July 2004; Accepted 3 October 2004

Published online 1 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20380

sents 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 $e_{ij} = c_0 - h_ih_j + q_iq_j$. 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 $c_0 - h_ih_j$ (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

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

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 $\|x\| = \sqrt{<x,x>}$. 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 σ_x is its standard deviation, respectively. The correlation between vectors x and y is defined as

$$cor(x,y) = \frac{< x - \bar{x},\ y - \bar{y} >}{n\sigma_x\sigma_y} = \frac{< x - \bar{x},\ y - \bar{y} >}{\|x - \bar{x}\|\ \|y - \bar{y}\|}. \quad (2)$$

We define also the distance between normalized vectors $dist(x,y) = \|y^N - x^N\|/\sqrt{n}$, where n is the dimensionality of vectors x and y (210 in our case). Obviously $dist(x, y)$ 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 $err(x, y) = \|y - x\| / \|y\|$.

It is worth noting that all of the above defined measures of the quality of approximation (*cor*, *dist, err*) are invariant to multiplication by a scalar and are optimized by the solution of the least squares problem (Eq. 1).[50] Additionally it is easily seen that $dist^2(x, y) = 2 - 2cor(x, y)$.

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

$$\tilde{e}(h)_{i,j} = h_i + h_j \text{ (Hp)} \quad (3a)$$

$$\tilde{e}(h)_{i,j} = c_0 + c_1h_ih_j \text{ (Hp.Dx)} \quad (3b)$$

$$\tilde{e}(h,q)_{i,j} = h_i + h_j + c_0q_iq_j \text{ (Hp.pH)} \quad (3c)$$

$$\tilde{e}(h,q)_{i,j} = c_0 + c_1h_ih_j + c_2q_iq_j \text{ (Hp.Dx.pH)}. \quad (3d)$$

The above approximations are related to each other as follows:

$$\text{(Hp)} \rightarrow \text{(Hp.Dx)} \rightarrow \text{(Hp.pH)} \rightarrow \text{(Hp.Dx.pH)} \quad (4)$$

The relation (a) → (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 $n = 20$, $s_i = \sum_j e_{i,j} + e_{ii}$ and $c_0 = (\sum_{ij:i \leq j} e_{i,j})/(n+1)$. 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 isoelectric points designated here by pH. To check the dependence of solutions of Eq. (3c) on the starting points, we interchanged vectors h and q. This optimization is denoted as (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 (Hp.pH) and (pH.Hp) are essential.

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

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

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 (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 (Hp.pH) used for approximation (3c). The last approximation, (3d) (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.TS = (N_{ij}) and lN.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 Shakhnovich[25] 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

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.

Matrix	Qa	Qm	Qp	HLPL	SKOb	SKOa	SJKG	MJPL	MJ3h	MJ2h	TS	BT	BFKV	TD	TEI	TEs	RO	MS	MJ1	MJ3	GKS	B2	B1	B3	B5	VD	BL	B4	MSBM
Qa	-	38	68	71	71	73	66	74	61	72	70	84	84	82	92	98	100	127	127	126	142	137	125	140	122	116	127	142	150
Qm	93	-	44	59	66	66	67	81	69	87	83	77	80	89	94	98	108	113	113	111	125	121	116	123	112	109	121	134	150
Qp	77	90	-	53	62	57	67	78	77	94	92	74	73	92	97	101	115	105	104	99	113	110	108	110	104	104	115	129	150
HLPL	75	82	86	-	48	33	42	61	61	80	73	61	68	92	92	98	114	101	101	95	114	112	100	114	100	99	111	127	150
SKOb	75	78	81	88	-	40	53	61	65	78	70	70	66	88	92	93	113	118	118	113	127	122	119	122	106	115	121	129	150
SKOa	74	78	84	95	92	-	28	48	61	68	70	63	59	85	91	99	109	113	113	108	124	122	112	123	107	109	113	134	150
SJKG	78	78	78	91	86	96	-	47	52	60	64	61	65	83	89	98	105	116	115	110	129	125	115	126	111	108	114	134	150
MJPL	73	68	69	81	81	88	89	-	42	59	68	62	64	76	84	98	103	115	114	110	135	125	113	130	104	108	114	137	150
MJ3h	81	76	70	81	79	81	86	91	-	47	64	53	65	72	80	90	97	111	112	105	132	122	114	126	102	107	113	133	150
MJ2h	74	62	56	68	70	77	82	82	89	-	49	82	77	81	93	104	109	137	138	135	157	148	138	148	127	120	128	146	150
TS	76	66	58	73	75	76	80	77	80	88	-	91	83	88	98	105	115	139	139	140	156	151	141	148	130	123	132	145	150
BT	65	70	73	81	75	80	82	81	86	66	59	-	63	86	89	93	105	87	86	74	105	95	89	102	86	96	98	122	150
BFKV	65	68	73	77	78	82	79	80	79	70	65	80	-	76	92	99	108	109	108	104	121	114	107	117	87	107	111	131	140
TD	66	60	57	58	61	64	65	71	74	67	61	63	71	-	98	109	100	121	120	120	139	132	118	136	114	116	114	146	140
TEI	58	56	53	58	57	59	60	64	68	57	52	61	57	52	-	92	112	119	118	114	129	122	119	125	113	118	120	135	140
TEs	52	52	49	52	56	51	52	52	60	46	45	57	51	41	57	-	118	118	116	113	128	121	119	120	108	115	126	127	140
RO	50	42	34	36	36	41	45	47	53	41	34	44	42	49	37	31	-	120	121	118	131	128	114	133	103	124	125	137	150
MS	19	36	45	49	30	36	33	34	38	7	4	63	41	27	30	31	28	-	24	56	66	68	71	80	82	82	98	113	150
MJ1	19	36	46	49	31	36	34	35	38	5	3	63	42	28	30	32	27	97	-	53	65	67	69	78	81	88	95	112	150
MJ3	21	39	51	55	36	42	40	40	45	9	2	73	46	29	36	36	31	85	86	-	50	46	55	66	72	93	91	108	150
GKS	-1	22	36	34	20	23	17	9	13	-23	-22	44	27	4	17	18	14	78	79	87	-	49	71	66	86	105	104	107	140
B2	6	27	39	37	25	26	22	22	26	-10	-14	55	35	13	26	27	17	77	77	89	88	-	78	57	80	109	106	102	140
B1	21	33	42	50	29	37	34	36	35	5	0	60	43	31	29	30	34	75	76	85	75	70	-	91	75	98	86	119	140
B3	3	25	40	35	26	24	20	16	21	-9	-10	48	32	7	22	28	12	68	70	78	78	83	59	-	77	115	114	91	150
B5	26	37	46	50	43	43	38	46	48	20	16	63	62	35	36	41	47	66	67	74	63	68	72	70	-	109	111	102	140
VD	33	41	46	51	34	41	41	42	43	28	24	54	43	33	31	34	24	66	61	56	45	41	52	34	41	-	110	131	160
BL	19	26	33	38	27	36	35	35	36	18	12	52	38	36	28	21	22	52	55	58	45	44	63	35	39	39	-	137	150
B4	-1	10	17	20	17	11	10	6	12	-7	-5	25	14	-7	9	20	6	36	37	41	43	48	29	58	48	15	6	-	150
MSBM	-12	-12	-10	-18	-18	-15	-16	-16	-14	-11	-12	-17	-1	2	-4	-4	-5	-15	-15	-9	-2	0	-2	-5	-3	-24	-14	-6	-

🟨 50 ▨ 60 🟥 70 ▨ 80 🟩 90 🟩 100

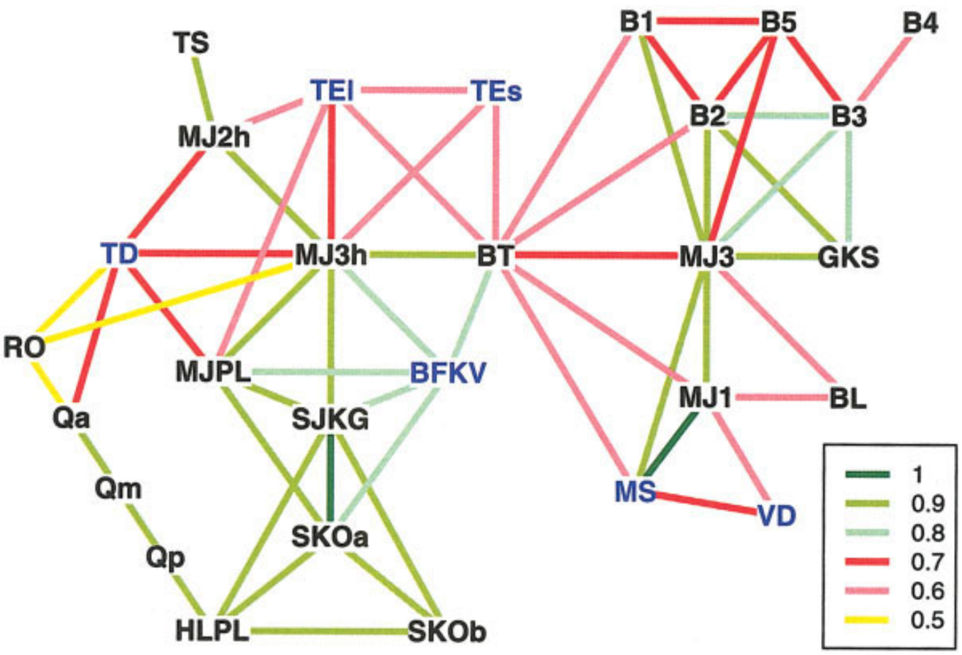

Fig. 1. Graphical illustration of correlations among different protein potentials. Coloring scheme is the same as in Table I.

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.[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.[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.[34,39]
- MSBM—optimization-derived potential developed by Micheletti et al.[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 (*cor*) between potentials, while the entries above the diagonal list the mean Euclidean distances between the normalized potentials (*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, lN.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 $cor(a,b) \geq 0.9$ and $cor(b,c) \geq 0.9$, then $cor(a,c) \geq 0.8$. 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 *cor*, *dist*, and *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 can be found at http://www.mimuw.edu.pl/~pokar. The last column in Table II displays correlations between vectors *q* and isoelectric points of amino acids pI (pH).[43–47] 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 (Hp) are mostly above 0.9, values only slightly smaller than the correlations with (Hp.Dx.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 (Hp) formula (correlation coefficient range is from 0.2 to 0.3, except 0.5 for GKS). Using the (Hp.Dx) formula, the correlation increases to about 0.8, and to about 0.9 for the (Hp.Dx.pH) formula. In the last formula the c_1 coefficients are negative, while the c_2 coefficients have positive values. Thus, the (Hp.Dx.pH) formula can be written in simplified form as

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

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,

TABLE II. One-Body Approximations to Protein Contact Potentials (CPs)

Matrix	Hp			Hp.Dx			Hp.pH			pH.Hp			Hp.Dx.pH			Hp.Dx.pH*			cor.Hp	nb.Hp	cor.pH
	cor	dist	err	cor	dist	err	cor	dist	err	cor	dist	err	cor	dist	err	cor	dist	err			
Qa	91	42	41	91	42	41	95	32	31	95	32	31	95	31	30	95	31	30	76	27	62
Qm	86	53	51	86	53	51	91	43	42	91	43	42	93	38	37	91	43	42	66	26	60
Qp	80	64	49	80	63	48	83	58	45	92	41	42	92	41	42	85	55	42	71	26	61
HLPL	82	59	42	83	58	41	85	54	38	93	38	28	93	38	28	86	53	38	74	26	91
SKOb	91	43	40	91	42	39	96	41	38	96	28	27	97	24	13	92	41	37	84	26	89
SKOa	86	53	51	87	51	49	89	48	46	94	33	33	95	32	33	90	45	44	85	26	86
SJKG	88	50	48	89	47	45	90	46	44	95	32	31	95	32	31	91	41	40	92	36	88
MJPL	88	50	11	90	44	10	90	46	10	95	32	7	95	32	7	93	38	8	93	12	84
MJ3h	86	53	49	93	39	37	88	49	46	96	30	28	96	30	28	95	31	30	97	19	92
MJ2h	98	19	8	99	14	6	99	14	6	99	11	5	99	11	5	99	11	5	97	25	92
TS	98	18	4	98	17	4	99	16	4	99	17	4	99	16	4	99	16	4	87	25	91
BT	62	87	77	82	59	56	69	78	71	88	48	46	90	46	44	90	46	44	95	25	93
BFKV	71	76	69	80	63	58	74	72	66	88	49	46	88	49	46	85	55	52	92	12	79
TD	72	75	66	75	71	63	77	68	60	79	66	59	82	60	54	80	63	57	93	34	82
TEI	59	91	80	61	89	79	66	82	75	71	76	70	71	76	70	69	79	72	93	19	76
TEs	56	94	82	64	84	76	61	89	79	71	76	70	71	76	70	69	79	71	83	26	67
RO	73	73	66	75	70	64	93	39	37	93	39	37	96	29	28	96	29	28	74	34	10
MS	22	125	97	68	80	73	42	108	91	71	76	70	79	65	61	79	65	62	95	1	94
MJ1	22	125	13	70	78	9	41	109	12	72	74	9	80	63	8	80	63	8	95	1	94
MJ3	21	126	98	80	63	60	41	109	91	82	60	57	90	45	44	90	45	44	94	1	95
GKS	50	100	86	85	56	53	56	93	82	87	50	48	90	44	43	90	44	43	96	1	95
B2	34	115	94	86	52	51	46	104	89	88	48	47	94	35	35	94	35	35	96	1	95
B1	21	126	98	68	80	73	50	78	72	83	58	55	83	58	55	83	58	55	93	34	91
B3	31	117	95	86	52	50	35	114	93	88	50	48	89	46	45	89	46	45	97	1	93
B5	26	122	96	87	51	50	26	122	96	88	50	48	98	17	4	87	51	49	98	5	98
VD	29	119	96	47	103	88	45	105	90	55	94	83	59	90	80	59	90	80	77	36	86
BL	10	134	100	52	97	85	51	99	86	53	97	85	75	71	66	75	71	66	93	36	71
B4	28	120	96	64	85	76	68	80	73	64	85	76	86	53	38	64	85	76	85	30	100
MSBM	33	116	94	66	82	75	66	82	75	66	82	75	100	8	2	66	82	75	51	9	100
N.TS	84	57	36	91	43	28	84	57	36	91	44	28	91	43	27	91	43	27	66	1	100
IN.TS	89	47	10	89	47	10	89	47	10	88	49	10	90	46	10	82	59	12	67	1	100
N.MJ2	82	60	37	91	43	27	82	60	37	90	44	28	92	41	26	92	41	26	66	1	100
IN.MJ2	91	43	4	91	42	4	91	42	4	91	43	4	93	34	3	86	52	4	66	1	100

🟨 50 🟪 60 🟥 70 ▨ 80 🟩 90 🟩 100

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.

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.[48,49] 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 struc-

tures (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.[4] and by Li et al.[51] The present results are, however, stronger and more universal. The main point of the work of Godzik et al.[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 $e_{ij}^{excess} = e_{ij} - e_{ij}^{ideal}$, where $e_{i,j}^{ideal} = (e_{ii} + e_{jj})/2$ was not completed, as can be seen clearly from our present results. The authors found that the correlation of GKS with $e_{i,j}^{ideal}$ 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.[51] was the derivation of a better (in general) approximating vector than (e_{ii}) 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* $e_{ij} = -h_ih_j + 0.5\ q_iq_j$ correlates moderately well with members of the second group of CPs (*cor* = 0.66, 0.60, 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 is 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).

ACKNOWLEDGMENTS

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

REFERENCES

1. Sippl MJ. Calculation of conformational ensembles from potentials of mean force—an approach to the knowledge-based prediction of local structures in globular proteins. J Mol Biol 1990;213:859–883.
2. Hinds DA, Levitt M. Exploring conformational space with a simple lattice model for protein structure. J Mol Biol 1994;243:668–682.
3. Bryant SH, Lawrence CE. An empirical energy function for threading protein-sequence through the folding motif. Proteins 1993;16:92–112.
4. Godzik A, Kolinski A, Skolnick J. Are proteins ideal mixtures of amino acids?: analysis of energy parameter sets. Protein Sci 1995;4:2107–2117.
5. Park B, Levitt M. Energy functions that discriminate X-ray and near-native folds from well-constructed decoys. J Mol Biol 1996;258:367–392.
6. Bahar I, Jernigan RL. Inter-residue potentials in globular proteins and the dominance of highly specific hydrophilic interactions at close separation. J Mol Biol 1997;266:195–214.
7. Liwo A, Pincus MR, Wawak RJ, Rackovsky S, Oldziej S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations: 2. Parameterization of short-range interactions and determination of weights of energy terms by *Z*-score optimization. J Computational Chem 1997;18:874–887.
8. Skolnick J, Jaroszewski L, Kolinski A, Godzik A. Derivation and testing of pair potentials for protein folding: When is the quasichemical approximation correct? Protein Sci 1997;6:676–688.
9. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268:209–225.
10. Simons KT, Ruczinski I, Kooperberg C, Fox BA, Bystroff C, Baker D. Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins. Proteins 1999;34:82–95.
11. Skolnick J, Kolinski A, Ortiz A. Derivation of protein-specific pair

potentials based on weak sequence fragment similarity. Proteins 2000;38:3–16.

12. Zhang C, Kim S-H. Environment-dependent residue contact energies for proteins. Proc Natl Acad Sci USA 2000;97:2550–2555.
13. Boniecki M, Rotkiewicz P, Skolnick J, Kolinski A. Protein fragment reconstruction using various modeling techniques. J Comput Aided Mol Des 2003;17:725–738.
14. Tanaka S, Scheraga HA. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. Macromolecules 1976;9:945–950.
15. Robson B, Osguthorpe DJ. Refined models for computer-simulation of protein folding—applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin-inhibitor. J Mol Biol 1979;132:19–51.
16. Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. Protein Sci 1999;8:361–369.
17. Liwo A, Oldziej S, Pincus MR, Wawak RJ, Rackovsky S, Scheraga HA. A united-residue force field for off-lattice protein-structure simulations: 1. Functional forms and parameters of long-range side-chain interaction potentials from protein crystal data. J Comput Chem 1997;18:849–873.
18. Miyazawa S, Jernigan RL. Estimation of effective interresidue contact energies from protein crystal structures—quasi-chemical approximation. Macromolecules 1985;18:534–552.
19. Park BH, Huang ES, Levitt M. Factors affecting the ability of energy functions to discriminate correct from incorrect folds. J Mol Biol 1997;266:831–846.
20. Keskin O, Bahar I, Badretdinov OB, Ptitsyn OB, Jernigan RL. Empirical solvent-mediated potentials hold for both intra-molecular and inter-molecular inter-residue interactions. Protein Sci 1998;7:2578–2586.
21. Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. J Mol Biol 1996;256:623–644.
22. Miyazawa S, Jernigan RL. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. Proteins 1999;34:49–68.
23. Kolinski A. Protein modeling and structure prediction with a reduced representation. Acta Biochim Pol 2004;51:349–371.
24. Maiorov VN, Crippen GM. contact potential that recognizes the correct folding of globular proteins. J Mol Biol 1992;227:876–888.
25. Mirny LA, Shakhnovich EI. How to derive a protein folding potential?: a new approach to an old problem. J Mol Biol 1996;264: 1164–1179.
26. Thomas PD, Dill KA. An iterative method for extracting energy-like quantities from protein structures. Proc Natl Acad Sci USA 1996;93:11628–11633.
27. Tobi D, Shafran G, Linial N, Elber R. On the design and analysis of protein folding potentials. Proteins 2000;40:71–85.
28. Vendruscolo M, Domany E. Pairwise contact potentials are unsuitable for protein folding. J Chem Phys 1998;109:11101–11108.
29. Vendruscolo M, Mirny LA, Shakhnovich EI, Domany E. Comparison of two optimization methods to derive energy parameters for protein folding: Perceptron and *Z* score. Proteins 2000;41:192–201.
30. Micheletti C, Seno F, Banavar JR, Maritan A. Learning effective amino acid interactions through iterative stochastic techniques. Proteins 2001;42:422–431.
31. Sippl MJ. Knowledge-based potentials for proteins. Curr Opin Struct Biol 1995;5:229–235.
32. Hao MH, Scheraga HA. Designing potential energy functions for protein folding. Curr Opin Struct Biol 1999;9:184–188.
33. Buchete NV, Straub JE, Thirumalai D. Orientational potentials extracted from protein structures improve native fold recognition. Protein Sci 2004;13:862–874.
34. Bradley P, Chivian D, Meiler J, Misura KM, Rohl C, Schief W, Wedemeyer WJ, Schueler-Furman O, Murphy P, Schonbrun J, Strauss C, Baker D. Rosetta predictions in CASP5: Successes, failures, and prospects for complete automation. Proteins 2003;53: 457–468.
35. Fang QJ, Shortle D. Prediction of protein structure by emphasizing local side-chain/backbone interactions in ensembles of turn fragments. Proteins 2003;53:486–490.
36. Jones DT, McGuffin LJ. Assembling novel protein folds from super-secondary structural fragments. Proteins 2003;53:480–485.
37. Panchenko AR, Marchler-Bauer A, Bryant SH. Combination of threading potentials and sequence profiles improves fold recognition. J Mol Biol 2000;296:1319–1331.
38. Pillardy J, Czaplewski C, Liwo A, Lee J, Ripoll DR, Kamierkiewicz R, Odziej S, Wedemeyer WJ, Gibson KD, Arnautova YA, Saunders J, Ye YJ, Scheraga HA. Recent improvements in prediction of protein structure by global optimization of a potential energy function. Proc Natl Acad Sci USA2001;98:2329–2333.
39. Shao Y, Bystroff C. Predicting interresidue contacts using templates and pathways. Proteins 2003;53:497–502.
40. Skolnick J, Zhang Y, Arakaki AK, Kolinski A, Boniecki A, Szilagyi A, Kihara D. TOUCHSTONE: a unified approach to protein structure prediction. Proteins 2003;53:469–479.
41. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. J Mol Biol 2000;300:171–185.
42. Munson PJ, Singh RK. Statistical significance of hierarchical multi-body potentials based on Delaunay tessellation and their application in sequence–structure alignment. Protein Sci 1997;6: 1467–1481.
43. Georgescu RE, Alexov EG, Gunner MR. Combining conformational flexibility and continuum electrostatics for calculating pK(a)s in proteins. Biophys J 2002;83:1731–1748.
44. Mehler EL, Fuxreiter M, Simon I, Garcia-Moreno EB. The role of hydrophobic microenvironments in modulating pKa shifts in proteins. Proteins 2002;48:283–292.
45. Sandberg L, Edholm O. A fast and simple method to calculate protonation states in proteins. Proteins 1999;36:474–483.
46. Tollinger M, Crowhurst KA, Kay LE, Forman-Kay JD. Site-specific contributions to the pH dependence of protein stability. Proc Natl Acad Sci USA 2003;100:4545–4550.
47. Laurents DV, Huyghes-Despointes BMP, Bruix M, Thurlkill RL, Schell D, Newsom S, Grimsley GR, Shaw KL, Trevi S, Rico M, Briggs JM, Antosiewicz JM, Scholtz JM, Pace CN. Charge–charge interactions are key determinants of the pK values of ionizable groups in ribonuclease Sa (pI = 3.5) and a basic variant (pI = 10.2). J Mol Biol 2003;325:1077–1092.
48. Pokarowski P, Kolinski A, Skolnick J. A minimal physically realistic protein-like lattice model: designing an energy landscape that ensures all-or-none folding to a unique native state. Biophys J 2003;84:1518–1526.
49. Pokarowski P, Droste K, Kolinski A. A minimal protein-like lattice model: an alpha-helix motif. 2004. Submitted for publication.
50. Rao CR. Linear statistical inference and its applications. New York: Wiley; 1973.
51. Li H, Tang C, Wingreen NS. Nature of driving force for protein folding: a result from analyzing the statistical potential. Phys Rev Lett 1997;79:765–768.
52. Contreras-Moreira B, Fitzjohn PW, Offman M, Smith GR, Bates PA. Novel use of a genetic algorithm for protein structure prediction: searching template and sequence alignment space. Proteins 2003;53:424–429.
53. Bastolla U, Farwer J, Knapp EW, Vendruscolo M. How to guarantee optimal stability for most representative structures in the Protein Data Bank. Proteins 2001;44:79–96.
54. Du R, Grosberg AY, Tanaka T. Models of protein interactions: how to choose one. Fold Des 1998;3:203–211.
55. Gan HH, Tropsha A, Schlick T. Lattice protein folding with two and four-body statistical potentials. Proteins 2001;43:161–174.

APPENDIX

1. In order to prove the relation (Hp) → (Hp.Dx), let us first notice that (Hp.Dx) is equivalent to the following formula:

$$a_0 + a_1(h_i{}' + h_j{}') + a_2 h_i{}' h_j{}'. \text{ (Hp.Dx.2)} \qquad \text{(A1)}$$

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

$$c_0 := a_0 - a_1^2/a_2,\ c_1 := a_2,\ c_2 := -a_1/a_2,\ h_i' := h_i' - c_2. \tag{A2}$$

Then:

$$\begin{aligned} c_0 + c_1 h_i h_j &= c_0 + c_1(h_i' - c_2)(h_j' - c_2) \\ &= a_0 - a_1^2/a_2 + a_2(h_i' + a_1/a_2)(h_j' + a_1/a_2) \\ &= a_0 - a_1^2/a_2 + a_2 h_i' h_j' + a_1(h_i' + h_j') + a_1^2/a_2 \\ &= a_0 + a_1(h_i' + h_j') + a_2 h_i' h_j'. \end{aligned} \tag{A3}$$

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) → (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 $h_i : = a_0/2 + a_1 h_i'$, 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 $h_i' : = c_0/2, c_0' : = c_1,$ *and* $q_i' : = h_i$, we can transform Eq. (3b) to the (Hp.pH) form $h_i + h_j + c_0 q_i q_j$. This proves the relation (Hp.Dx) → (Hp.pH).
3. The relation (Hp.pH) → (Hp.Dx.pH) is derived similarly as (Hp) → (Hp.Dx) by proving first that (Hp.Dx.pH) is equivalent to the following formula:

$$a_0 + a_1(h_i' + h_j') + a_2 h_i' h_j' + a_3(q_i' + q_j') + a_4 q_i' q_j'. \tag{A4}$$

Assuming that $c_0 := a_0 - a_1^2/a_2 - a_3^2/a_4$, $c_1 := a_2$, $c_3 : = a_4$, $c_2 := -a_1/a_2$, $c_4 := -a_3/a_4$, $h_i := h_i' - c_2$, and $q_i := q_i' - c_4$, 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).