BIOINFORMATICS

Vol. 17 no. 8 2001
Pages 686–699

Amino acid similarity matrices based on force fields

Zsuzsanna Dosztányi and Andrew E. Torda*

Research School of Chemistry, Australian National University, Canberra, ACT 0200, Australia

Received on December 18, 2000; revised on April 9, 2001; accepted on April 19, 2001

ABSTRACT

Motivation: We propose a general method for deriving amino acid substitution matrices from low resolution force fields. Unlike current popular methods, the approach does not rely on evolutionary arguments or alignment of sequences or structures. Instead, residues are computationally mutated and their contribution to the total energy/score is collected. The average of these values over each position within a set of proteins results in a substitution matrix.

Results: Example substitution matrices have been calculated from force fields based on different philosophies and their performance compared with conventional substitution matrices. Although this can produce useful substitution matrices, the methodology highlights the virtues, deficiencies and biases of the source force fields. It also allows a rather direct comparison of sequence alignment methods with the score functions underlying protein sequence to structure threading.

Availability: Example substitution matrices are available from http://www.rsc.anu.edu.au/~zsuzsa/suppl/matrices.html.

Supplementary information: The list of proteins used for data collection and the optimized parameters for the alignment are given as supplementary material at http://www.rsc.anu.edu.au/~zsuzsa/suppl/matrices.html.

Contact: zsuzsa@rsc.anu.edu.au

INTRODUCTION

Protein sequence comparisons lie at the heart of biology from the macroscopic to the molecular. At the macroscopic level, they may be used to define the phylogeny of species. At the molecular level, these comparisons are the first tool used to predict a protein's structure or function. Similarity or substitution matrices provide a quantitative basis for sequence comparisons. In this work, we show how these substitution matrices can be obtained from force fields acting on three-dimensional structures.

Similarity matrices may be obtained via different routes, but simply, they provide a measure for how similar different types of residues are to each other. The simplest example, an identity matrix, would say that a Glu is the same as a Glu, but different to a Trp. A more sophisticated matrix might be a 20×20 table which says that a Glu is similar to an Asp (both acidic residues), but very different to a large hydrophobic residue such as a Trp. The success or failure of sequence comparison depends entirely on the underlying comparison matrix.

The most popular matrices rely on evolutionary arguments. Similar proteins can be aligned and mutations counted. Looking at a series of alignments, it would be clear by inspection that Asp and Glu residues often occupy corresponding sites, but only rarely would a Trp residue be found there. Historically, initial alignments were done by hand on very similar proteins (Dayhoff *et al.*, 1978) and extrapolated to less related proteins. Modern matrices are constructed by automated alignments from large sequence databases (Gonnet *et al.*, 1992; Jones *et al.*, 1992b) and attempts have been made to remove the dependence on extrapolations by counting mutations in conserved blocks in less related proteins (Henikoff and Henikoff, 1992).

One can completely avoid reliance on similar sequences or conserved blocks by using structural superpositions (Johnson and Overington, 1993; Risler *et al.*, 1988). Structurally similar proteins can be aligned, even when sequence identity is insignificant (Russell *et al.*, 1997; Prlic *et al.*, 2000; Blake and Cohen, 2001). This approach has the property that it will use information from convergent evolution, rather than only divergent sequences. It has the disadvantages that structural alignments and thresholds for aligned sites are less well characterized and the database of known structures is a fraction of the size of sequence databases and the number of alignable structure pairs even smaller.

If one wishes to avoid reliance on evolution, it should be possible to use pure chemical information to estimate the similarity of amino acids. Physico-chemical properties such as hydrophobicity, volume, composition and secondary structure preferences have served as the

*To whom correspondence should be addressed.

basis for similarity measures (Grantham, 1974; Miyata *et al.*, 1979; George *et al.*, 1990; Mohana Rao, 1987). Intuitively, this should be reasonable since it has been argued that hydrophobicity and volume are the dominant factors underlying substitutions in evolution (Grantham, 1974; Tomii and Kanehisa, 1996). Continuing in this vein, there are also matrices summarizing amino acid similarity based on characteristics such as sequential neighbourhood (Tüdős *et al.*, 1990), conformational preferences (Niefind and Schomburg, 1991) or contact frequency (Miyazawa and Jernigan, 1993).

One disadvantage of these methods is that they require setting thresholds and making decisions as to the relative importance of properties. For example, one could use text book measures of hydrophobicity and molar volume to compare amino acids, but there may be no natural way to scale the individual terms. There are, however, alternative approaches which could solve this problem. Firstly, Kann *et al.* (2000) treat the issue of substitution matrix construction as one of numerical optimization. This can be done with minimal reliance on evolutionary assumptions. Alternatively, one can use the information in molecular mechanics force fields which contain a natural weighting of different physical contributions. A force field's parameters exactly define a weighting of everything from Lennard–Jones parameters to bond angle constants. Here, we show how this literature knowledge can be used in building a substitution matrix quite automatically.

The problem could be approached in terms of computational mutations and corresponding energy changes (Wang *et al.*, 1996). This would provide a simple measure of dissimilarity between amino acids, but it would not provide and obvious recipe for diagonal elements in a substitution matrix. The approach used here comes from analogy with sequence to structure alignment methods. A site in a protein is labelled by the type of its native residue. A score (energy) function provides a direct measure of compatibility of every type of amino acid at that position (Berezovsky *et al.*, 2000). With some set of calibration proteins, one can average over all corresponding energy values to obtain an element for a substitution matrix. For example, the average score for putting a Trp into an Ala site comes from the Trp energy averaged at every site in the protein library which has a native Ala. The approach naturally leads to a measure for self-similarity as native residue scores are calculated. This is essential for diagonal elements of the final matrix.

The methodology is independent of any particular force field, but most of the example calculations are based on one low-resolution, 'knowledge-based' score function built for protein fold recognition. To demonstrate the force field independence, an amino acid substitution matrix was also built from a Boltzmann-based potential of mean force, taken from the literature.

The most interesting application of the similarity matrix construction procedure may not be the final matrix. Instead, it may be the ability to compare force fields with force fields and with similarity matrices. Fold recognition force fields are often used for sequence to structure alignment, but there is no simple way to compare them with the matrices used for sequence to sequence alignment. The matrix derived from a force field, however, is easy to compare with a conventional substitution matrix. Using principal components analysis, one can look for dominant terms in the force fields as well as substitution matrices. Furthermore, the method can be used to compare force fields with each other, even when they have radically different formulations. The results give an example of this calculation.

METHODS

Energy/score calculations and matrix construction

To calculate a substitution matrix, one needs to be able to calculate a score or energy due to a single residue within the context of some protein. Since we are using strictly additive force fields, the total score can be decomposed into contributions from individual residues. Firstly, we say that in the native protein, a refers to a site which has amino acid type A as the native residue and the energy contribution from this site and residue is $E_{a,A}$. If one were to change the type of the residue at position a, to B, the energy would be $E_{a,B}$.

The substitution matrix, $\mathbf{M}$ contains elements $m_{A,B}$ which represents the compatibility of a residue of type B in a site originally occupied by a residue of type A. Considering the entire library of proteins, there are K_a sites with a native residue A, so we can write

$$m_{AB} = 1/K_a \sum_{a=1}^{K_a} E_{a,B}. \qquad (1)$$

Already it is clear that the final matrix will not be symmetric. $m_{A,B}$ shows the compatibility of residues of type B with sites where native structures have an A type residue, whereas $m_{B,A}$ represents $E_{b,A}$ energies.

The expression above defines a mean of an energy distribution, but it does not give an idea of the reliability. Assuming a normal distribution, this is well quantified by the standard deviation:

$$\sigma_{A,B} = \sqrt{\frac{\sum_{i=1}^{K_a} (E_{a,A} - \langle E_{a,A} \rangle)^2}{K_a}} \qquad (2)$$

where the summation runs over all K_a sites of native type A and $\langle E_{a,A} \rangle$ is equal to m_{AB} by definition.

Initial calculations suggested that the force fields have some systematic preferences, reflecting the zero-level of certain interactions. For example, in some force fields,

a hydrophobic residue is always energetically preferred, almost independent of the exact environment. In order to study this effect, a special normalization was performed. For each matrix element, the average over its column and row were subtracted:

$$m_{A,B}^{\mathrm{Norm}} = m_{A,B} - 1/N_A \sum_{p=1}^{N_A} m_{A,P} - 1/N_A \sum_{q=1}^{N_A} m_{Q,B} \quad (3)$$

where A and B refer to the type of amino acid, P and Q are the types of amino acid indexed by p and q, respectively, and N_A is the number of amino acid types.

Comparison and analysis of matrices

Principal component analysis and hierarchical clustering were applied to analyze the relationship between different matrices and to represent the similarity of amino acids in a given matrix. The principal component analysis of the original matrices was performed using the SPlus package (Becker *et al.*, 1988) while an unrooted tree was constructed using the FITCH algorithm in the Phylip package (Felsenstein, 1993).

Comparing amino acid properties requires the definition of a distance between amino acids. This was defined by the Euclidean distance between pairs of rows within a matrix:

$$d_{A,B} = \sqrt{\sum_{P=1}^{N_A} (m_{A,P} - m_{B,P})^2} \quad (4)$$

where the type of residue P can assume each of the N_A types. The original data matrix is not symmetric, so different results would be obtained based on distances between rows or columns. In this work, rows were chosen such that $d_{A,B}$ is based on having A as the native residue.

As well as comparing amino acids, whole substitution matrices were compared using the principal component projection and the clustering procedures. This requires the definition of a distance between matrices and was based on the correlation coefficient of the two matrices ($\mathbf{X}$ and $\mathbf{Y}$):

$$d_{\mathbf{X},\mathbf{Y}} = 1 - \mathrm{corr}(\mathbf{X}, \mathbf{Y}) \quad (5)$$

where

$$\mathrm{corr}(\mathbf{X}, \mathbf{Y}) = \left(\sum_{p=1}^{N_A} \sum_{q=1}^{N_A} (X_{P,Q} - \langle \mathbf{X} \rangle)(Y_{P,Q} - \langle \mathbf{Y} \rangle) \right) \Bigg/ \left(\sqrt{\sum_{p=1}^{N_A} \sum_{q=1}^{N_A} (X_{P,Q} - \langle \mathbf{X} \rangle)^2 \sum_{p=1}^{N_A} \sum_{q=1}^{N_A} (Y_{P,Q} - \langle \mathbf{Y} \rangle)^2} \right)$$

in which $X_{P,Q}$ represents an element in the matrix, and N_A is the number of amino acids. The angle bracket denotes the arithmetic average of the matrix elements. The correlation coefficient was chosen as the distance measure because, unlike the Euclidean distance, it does not change upon a linear transformation of the matrix.

As has been pointed out (May, 1999; Johnson and Overington, 1993), these methods are sufficient to highlight the overall relationship between matrices but not fine details. The result depends on the details of the algorithm used, the distance measure and the number of elements. We also experienced the instability of the results, nevertheless the general conclusions remained similar and were supported by independent analyses and by previously published works (Johnson and Overington, 1993; Tomii and Kanehisa, 1996). The result can be distorted if there are members of the set which are very different from the rest. If outliers were found, the analyses were repeated with these members omitted.

Alignment accuracy of matrices

The constructed substitution matrices were compared with literature examples for their ability to perform sequence alignments. A test set consists of lists of pairs of proteins. For each pair, there is some reference alignment, assumed to be correct. For these comparisons, five test sets, compiled by three different groups, were taken from the literature and listed in Table 2. In each case, the reference alignment was constructed from structural alignment.

Test sets based on structural alignments have several consequences. Firstly, the level of sequence identity can be low. Secondly, a structure based alignment may not agree with any sequence based alignment. Lastly, within any sequence, only certain segments may have corresponding regions (equivalenced residues) in the partner sequence. The quoted results are based on the correctly aligned residues among equivalenced positions, when residue a_i in the first protein is aligned with residue b_j in both the reference and the predicted alignment. The alignment accuracy is the average number of these correctly aligned pairs as the percentage of the equivalenced positions. No attempt was made to work with easier test sets (greater sequence similarity) nor to measure the performance in database searches. However, earlier analysis showed that matrices performing well in pairwise alignment tests are also good candidates for database searching, although the optimal gap parameters usually differ (Vogt *et al.*, 1995).

The alignments were constructed using the Gotoh algorithm (Gotoh, 1982) with affine gap penalties ($u + k * v$, where u is the gap opening, v the gap widening penalty and k the length of the gap). The simplex algorithm was used (Press *et al.*, 1992) to adjust gap opening and widening penalties as well as the minimum matrix value so as to optimize the alignment accuracy for a given test set. Since there will be many local minima in parameter space, one certainly cannot guarantee that optimal values were found.

Fortunately, the algorithm is very robust if good starting points are chosen. The parameters presented for each test set reached the highest accuracy among 30 different trials.

Force fields and energy calculations

Most calculations were carried out using the force field implemented in the sausage program (Huber *et al.*, 1999) which was constructed by optimizing parameters for protein fold recognition (Huber and Torda, 1998). This was used to build the SM_SAUSAGE and SM_SAUS_NORM substitution matrices. A smaller set of calculations, for comparison, used the Boltzmann-based potential of mean force implemented in THREADER2.5 (Jones *et al.*, 1992a).

Both force fields could be summarized by saying that the total energy, E^{tot} is the sum of pairwise and solvation terms:

$$E^{\text{tot}} = E^{\text{pair}} + wE^{\text{solv}} \tag{6}$$

and w is some weighting term. The sausage force field is strictly additive, so energy contribution $E_{a,A}$ can be extracted from those terms involving site a. The THREADER potential of mean force is not a conventional conservative force field in that w has a dependence on the energy distribution across a library. The energies, $E_{a,A}$, were calculated from self-alignment and manipulation of library coordinate files so as to delete the relevant residue. This relies on an approximation that w does not vary between $E_{a,A}$ and $E_{a,B}$ for some A, B pair of residue types. THREADER is not solely based on force field since it incorporates an additional sequence similarity term built into its solvation potential.

For the main calculations with the sausage force field, the protein data set was taken from the December 1998 release of PDB_select representative non-homologous proteins (Hobohm and Sander, 1994) where no two proteins had more than 25% sequence identity to each other. After removing members with chain breaks or resolution worse than 2.5 Å, 703 proteins remained, with a total of 132 877 sites to be mutated. The comparison matrix built from the THREADER program was based on the common subset of the 703 proteins and the 1096 proteins provided as the default fold library of THREADER (tdb_mar99). This common subset contained 380 proteins with 81 227 positions. The protein lists are available at http://www.rsc.anu.edu.au/~zsuzsa/suppl/list_of_proteins.html.

Collection of matrices

Table 1 lists the matrices either calculated or taken from the literature and used for comparisons. IDENTITY refers to a naive identity matrix (diagonal elements set to 10, off-diagonal to 0). All literature matrices were downloaded from the AAIndex collection (http://www.genome.ad.jp/dbget/aaindex.html) (Tomii and Kanehisa, 1996; Kawashima and Kanehisa, 2000), except RUSSELL_RH which was downloaded from the authors' web site (http://www.bmm.icnet.uk/AH/). Matrices that were supplied as dissimilarity matrices were converted into similarity matrices simply by subtracting each element from the maximum value of the matrix. If the original matrix contained two types of cysteines, they were averaged; other non-typical amino acids were omitted from the comparison.

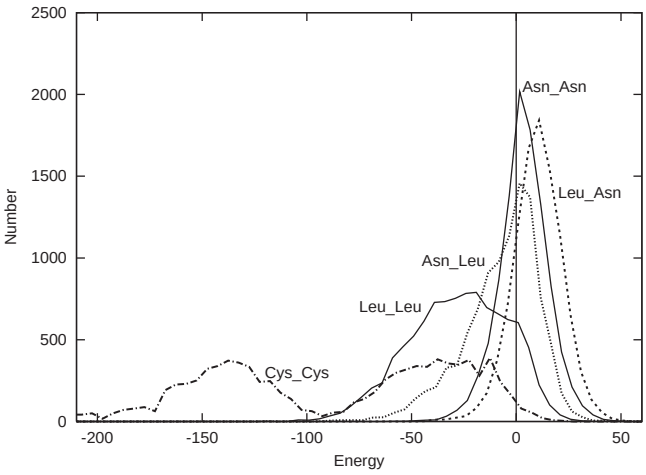

Fig. 1. Example for the energy distributions $E_{a,A}$ using the sausage force field. The examples are given for three native energy distributions (Cys_Cys, Leu_Leu and Asn_Asn) and for the replacement of Leu with Asn (Leu_Asn) and Asn with Leu (Asn_Leu). The more negative energy represents more favourable interactions.

RESULTS

Distribution of mutation energies

Before considering substitution matrices, the data collection process itself provided some information about the sausage force field. There were $20 \times 20 = 400$ types of substitution considered, but for illustration, Figure 1 shows the data from example hydrophobic and polar amino acids, Leu and Asn. The curves Leu_Leu and Asn_Asn are for the native residues and their averages lead to diagonal elements in the similarity matrix. The Asn_Leu curve shows the distribution of Leu energies when placed in native Asn sites.

This selection of curves shows the properties seen in the other 395 curves. They are generally well approximated by a Gaussian distribution, although the Asn_Leu curve does show some skew. The glaring exception to the style of distribution is the Cys data. The Cys_Cys distribution is clearly bimodal as can be seen in Figure 1. This comes out

Table 1. Amino acid substitution matrices compared in this study

Matrix name	Short name	Reference	Basis
SM_SAUSAGE	SA	Present study	Force field based (sausage)
SM_SAUS_NORM	SN	Present study	Force field based and normalized (sausage)
SM_THREADER	TH	Present study	Force field based (THREADER)
SM_THREAD_NORM	TN	Present study	Force field based and normalized (THREADER)
DAYHOFF	DA	Dayhoff *et al.* (1978)	Sequence comparison (PAM250)
FENG	FE	Feng *et al.* (1985)	Genetic code and sequence similarity
FITCH	FI	Fitch (1966)	Genetic code
GONNET	GO	Gonnet *et al.* (1992)	Sequence comparison (PAM250)
GRANTHAM	GR	Grantham (1974)	Physical property indeces
BLOSUM	BL	Henikoff and Henikoff (1992)	Sequence comparison (BLOSUM62)
IDENTITY	ID	–	(10, 0)
JOHNSON	JO	Johnson and Overington (1993)	Structure based sequence comparison
JONES	JD	Jones *et al.* (1992b)	Sequence comparison (PET91)
LEVIN	LE	Levin *et al.* (1986)	Sequence comparison by secondary structure
MCLACHLAN	MC	McLachlan (1971)	Sequence comparison
MIYATA	MI	Miyata *et al.* (1979)	Physical property indices
MIYAZAWA	MJ	Miyazawa and Jernigan (1993)	Contact potential
RAO	RA	Mohana Rao (1987)	Structural and physical property indices
RISLER	RI	Risler *et al.* (1988)	Structure based sequence comparison
RUSSELL_RH	RU	Russell *et al.* (1997)	Structure based sequence comparison (remote hom.)
TUDOS	TU	Tüdős *et al.* (1990)	Sequentional neighbourhood

of the data quite automatically, but can be attributed to the two populations of cysteines (disulfide and non-disulfide bonded). Cys residues have been treated as two separate type of residues by other workers (Overington *et al.*, 1992; Johnson and Overington, 1993), but in this process, it is quite automatic. The energy data can be simply divided into two subsets based on whether the native cysteine was involved in a disulfide bond or not. This leads to three entries in the tables below. One has distinct residue types for half-cystine (O) and free cysteine (J). For comparison with literature matrices and analysis, there is a conventional Cys residue which comes from treating Cys data as a single residue type.

The next feature of the distributions may be unique to this methodology. Not only do the distributions of Figure 1 have a mean, they have very distinct widths. Even by eye, the Leu_Leu distribution is much broader than the corresponding Asn_Asn curve. These differences, quantified according to equation (2) are collected in Table 3 for the SM_SAUSAGE matrix (corresponding values for the SM_THREADER matrix are at http://rsc.anu.edu.au/~zsuzsa/suppl/matrices.html). The table shows that Cys has the highest deviation. This is a very real effect. The energy distribution is very wide since it really includes two types of residue (free and disulfide bonded). At the level of sequence comparison, it is generally unknown which Cys one is dealing with.

Looking at the rest of Table 3, there is a strong correlation between the absolute size of an energy and the standard deviation. Even after accounting for this, the distribution of energies tends to be larger for hydrophobic residues, especially aliphatic examples. Without quantifying this properly, the trend suggested by Figure 1 is not unusual. The distributions involving Leu are always broader than those involving Asn. One could impose a physical interpretation. The aliphatic hydrophobic residues may have the most neighbours and the environment with the greatest variability. Whatever the reason, the energy distributions are wider and statistically, the entries in the final matrices less reliable.

Similarity matrix features and comparisons

Aside from the distributions of scores, the force field-derived similarity matrices have some clear properties. The sausage-derived matrix, SM_SAUSAGE, is given in Table 4. It would appear that replacing almost anything by a hydrophobic residue is favourable. This can be seen

Table 2. List of alignment test sets with their reference

Test set	Structural alignment program	Sequence identity (I)	Number of sequence pairs	Percentage of equivalenced residues (%)
MEDIUM (Russell *et al.*, 1997)	STAMP[a]	$25 < \%I < 50$	89	81
REMOTE (Russell *et al.*, 1997)	STAMP[a]	$\%I \leqslant 25$	94	54
HELLM (Domingues *et al.*, 2000)	ProSup[b]	$\%I \leqslant 30$	127	62
BASIC/family (Rychlewski *et al.*, 2000)	DALI[c]	$\%I \leqslant 30$	327	73
BASIC/superfamily (Rychlewski *et al.*, 2000)	DALI[c]	$\%I \leqslant 30$	195	56

The structural alignments were made by: [a] Russell and Barton (1992), [b] Feng and Sippl (1996), [c] Holm and Sander (1998). The MEDIUM, REMOTE and the two BASIC sets contain only pairs which are classified as homologues according to SCOP (Murzin *et al.*, 1995). The sequence identity range gives the cutoff values used to collect the pairs. The percentage of equivalenced residues give the ratio of the number of structurally aligned positions and the length of the first sequence.

Table 3. Standard deviation of SM_SAUSAGE matrix elements

	C	V	I	L	F	M	Y	A	W	H	T	R	P	Q	S	G	N	E	K	D
C	94.8	24.6	26.9	23.8	25.8	20.5	21.8	16.9	21.0	17.4	13.5	11.7	19.2	11.8	14.5	16.2	12.9	11.9	11.7	16.2
O	94.6	23.1	25.4	22.4	26.1	20.1	22.2	16.7	21.4	16.3	13.2	11.6	17.5	11.1	14.2	15.2	12.9	12.3	11.2	15.1
J	39.3	26.5	28.9	25.6	21.8	20.8	20.4	15.7	19.8	11.3	11.0	10.8	15.0	10.2	12.0	13.7	11.7	10.6	10.3	14.6
V	36.2	27.9	29.7	26.0	22.4	21.3	20.4	15.4	20.3	10.8	10.1	10.4	13.3	9.7	10.5	12.5	11.8	10.6	10.2	14.2
I	34.5	25.9	27.7	24.3	20.9	20.1	19.3	14.6	19.0	10.6	10.1	10.5	13.3	9.6	10.1	12.0	11.5	10.5	10.2	13.6
L	35.1	24.2	26.2	23.2	20.2	19.3	18.6	14.3	18.2	10.7	10.6	10.6	13.9	10.0	11.1	12.3	11.6	10.7	10.3	13.7
F	36.0	23.7	25.5	22.5	19.6	19.2	18.4	14.4	18.3	10.9	10.7	10.7	13.8	9.8	11.4	12.9	11.8	10.6	10.0	14.1
M	34.7	23.5	25.7	23.4	19.7	19.7	18.4	14.8	18.4	10.2	10.3	10.5	13.8	9.8	10.7	12.5	11.3	10.4	10.0	13.8
Y	36.1	22.7	24.6	21.8	19.0	18.6	17.8	14.1	17.7	10.8	10.5	10.7	13.8	9.9	11.4	13.2	11.7	10.6	10.1	14.0
A	37.5	26.4	29.0	25.9	22.3	21.5	20.5	16.9	20.7	10.9	11.1	11.1	14.8	10.6	12.1	14.0	12.7	11.5	11.0	15.4
W	38.0	23.0	25.0	22.2	19.4	18.6	18.4	14.2	17.8	11.0	11.0	11.0	14.3	10.0	11.6	13.0	11.4	10.7	10.1	13.8
H	35.2	19.5	21.4	19.5	17.5	16.9	16.5	13.4	16.5	10.7	10.6	10.4	13.9	9.8	12.5	12.9	11.6	10.8	10.1	13.8
T	32.7	22.0	23.9	21.1	18.5	18.1	17.2	14.1	17.8	10.5	11.7	10.4	13.2	9.7	15.0	12.6	11.8	10.6	10.1	14.6
R	28.1	18.1	19.8	17.8	15.7	15.6	15.5	12.5	15.6	10.1	10.4	10.5	13.7	9.8	11.7	12.5	11.5	10.6	9.9	14.0
P	32.9	20.4	22.1	19.3	17.4	16.0	16.2	12.3	16.3	9.6	9.3	9.0	11.7	8.5	11.0	11.6	10.4	9.5	9.1	12.7
Q	28.6	18.0	19.5	17.8	15.6	15.6	15.4	12.6	15.2	9.8	10.3	10.3	13.7	9.8	10.9	12.0	11.0	10.5	9.9	13.8
S	33.9	19.9	21.6	19.2	17.4	16.9	16.7	14.2	17.2	10.7	12.3	10.5	14.3	10.1	16.1	13.1	12.0	11.0	10.2	15.1
G	38.3	19.4	21.2	19.1	17.5	16.6	16.8	15.6	17.3	10.7	10.6	10.5	13.7	10.1	12.4	14.8	11.6	10.8	10.4	13.9
N	29.1	18.2	19.7	17.8	15.8	15.8	15.7	12.6	16.2	10.4	10.2	10.4	13.8	9.7	12.1	13.1	11.6	10.5	10.1	14.5
E	24.9	16.4	17.8	16.4	14.4	14.6	14.5	12.4	14.5	9.7	10.0	10.4	13.8	9.6	10.9	11.7	11.2	10.7	9.9	14.1
K	25.6	16.8	18.3	16.6	14.6	14.7	14.2	11.7	14.6	9.6	10.0	10.0	13.5	9.3	11.0	12.0	11.0	9.9	9.4	13.0
D	26.8	16.0	17.3	15.9	14.2	14.5	14.5	12.6	14.6	10.1	10.3	10.4	14.2	9.6	12.5	12.6	11.3	10.4	9.9	14.7

The native cysteine residues were divided into two subsets depending on their covalent state. The matrix elements were also calculated on these subsets, resulting in a separate row for disulfide bonded (O) and free cysteines (J). The amino acids are ordered according to the first principal component of the substitution matrix.

by the hydrophobic columns (Val, Ile, Leu, ...) consisting of entirely favourable scores. The trend is present to a lesser extent in the SM_THREADER matrix. This reflects the goal of the original force field. In protein sequence to structure alignments, it is usually desirable to find the placement of residues which maximizes hydrophobic contacts and forms a hydrophobic core. This may not be appropriate for sequence comparisons, so the trend has been removed by the normalization given in equation (3). For the sausage force field, this results in SM_SAUS_NORM, given in supplementary material (http://www.rsc.anu.edu.au/suppl/matrices.html).

Table 4. SM_SAUSAGE matrix

	C	V	I	L	F	M	Y	A	W	H	T	R	P	Q	S	G	N	E	K	D
C	111.4	31.5	31.0	27.4	35.3	26.1	30.3	19.2	25.9	11.4	1.5	−6.6	0.2	−6.3	2.6	−3.0	−7.0	−12.2	−12.3	−7.4
O	166.7	32.3	31.6	28.3	44.3	29.4	36.2	24.8	30.9	20.6	7.5	−4.6	10.4	−1.7	8.8	4.6	−3.2	−10.1	−9.3	−1.4
J	48.4	31.8	30.8	27.0	25.1	23.1	24.9	14.8	21.6	1.0	−4.6	−7.1	−10.1	−9.3	−3.1	−8.9	−10.4	−13.4	−13.5	−13.9
V	41.3	36.7	35.9	29.2	26.0	23.0	25.1	13.4	22.1	1.0	−3.9	−5.4	−8.4	−8.6	−5.5	−11.4	−10.7	−13.4	−12.6	−14.8
I	39.5	35.5	35.7	29.5	25.3	22.8	24.1	12.7	21.3	0.6	−4.7	−5.6	−8.9	−9.0	−6.2	−12.7	−11.3	−13.5	−12.6	−15.0
L	39.5	31.1	31.9	27.8	23.6	21.8	22.3	13.0	20.1	0.9	−5.3	−5.7	−9.2	−8.2	−5.1	−11.5	−10.1	−12.4	−12.3	−13.3
F	39.2	27.5	27.2	23.2	21.5	18.9	21.0	11.4	18.4	0.4	−5.3	−6.5	−8.0	−8.7	−4.1	−9.4	−9.3	−12.5	−12.2	−11.9
M	36.3	25.7	25.5	22.4	19.4	18.8	19.0	11.2	17.3	0.5	−5.7	−5.6	−8.7	−7.7	−4.2	−10.3	−8.8	−11.4	−11.2	−11.6
Y	37.4	24.6	23.7	19.9	19.1	16.6	19.8	10.2	16.9	−0.0	−5.3	−6.5	−7.7	−8.4	−3.9	−8.8	−8.6	−11.6	−11.7	−10.7
A	38.6	23.7	23.0	20.9	18.9	17.8	18.9	15.4	17.6	0.2	−5.2	−4.9	−9.3	−7.4	−1.6	−6.0	−8.1	−10.7	−10.9	−10.7
W	37.4	23.8	23.2	19.7	18.4	16.2	19.1	9.9	17.0	−0.3	−5.5	−6.9	−7.4	−8.6	−3.8	−9.0	−8.8	−11.7	−12.1	−10.7
H	30.5	16.1	14.1	11.8	12.3	10.5	13.2	6.8	9.8	−0.7	−5.7	−6.0	−8.2	−7.8	−2.4	−7.5	−6.1	−10.7	−10.5	−8.0
T	27.1	16.2	13.6	10.9	11.0	9.3	12.3	6.7	9.0	−1.5	−2.6	−5.4	−7.3	−7.3	0.8	−8.0	−5.2	−10.1	−10.1	−7.0
R	23.8	14.6	12.3	10.4	9.8	9.6	10.8	6.4	8.6	−1.1	−5.5	−3.1	−8.8	−6.5	−2.1	−8.3	−5.3	−10.0	−7.9	−9.0
P	26.7	14.3	12.5	9.1	10.1	8.2	10.0	4.9	8.9	−1.2	−5.3	−7.8	2.0	−8.5	−2.1	−6.2	−7.3	−10.0	−10.4	−7.5
Q	24.0	13.1	11.2	10.0	9.2	8.9	10.4	6.3	8.3	−1.4	−6.1	−4.8	−9.0	−6.0	−2.8	−8.4	−5.3	−8.3	−8.9	−6.8
S	28.2	11.7	8.8	7.9	8.9	7.8	10.5	7.3	7.6	−1.6	−3.0	−5.9	−7.1	−7.1	3.0	−5.3	−4.3	−9.3	−10.0	−4.9
G	30.7	11.1	7.8	6.0	8.8	7.1	9.8	7.3	8.4	−2.9	−7.3	−8.2	−7.1	−9.0	−1.5	1.2	−6.9	−10.8	−11.6	−8.4
N	23.5	11.5	8.6	7.1	7.8	6.2	9.5	4.0	6.4	−1.9	−5.2	−6.1	−8.7	−7.2	−1.3	−7.6	−3.4	−9.2	−10.1	−4.8
E	21.3	10.8	8.7	7.4	7.4	6.7	8.7	5.3	6.8	−1.8	−5.8	−5.3	−7.9	−5.8	−2.3	−7.7	−4.5	−6.0	−9.1	−3.7
K	20.3	10.9	8.5	7.0	7.0	7.2	8.1	5.3	5.8	−1.5	−5.6	−3.0	−8.2	−5.7	−1.7	−7.3	−4.4	−8.7	−6.7	−7.3
D	22.2	8.1	5.3	4.8	5.7	4.4	8.0	4.1	4.6	−2.7	−4.9	−7.2	−8.2	−6.5	−0.1	−7.1	−2.7	−6.0	−10.6	0.2

The native cysteine residues were divided into two subsets depending on their covalent state. The matrix elements were also calculated on these subsets, resulting in a separate row for disulfide bonded (O) and free cysteines (J). The amino acids are ordered according to the first principal component of the substitution matrix.

Given this collection of substitution matrices, they can now be compared with classical tables from the literature. Both the principal component analysis and the tree based on the hierarchical clustering show that the matrix based on the sausage force field is very different from the other matrices (Figure 2). This difference is much smaller after normalization using equation (3), and the resulting substitution matrix (SM_SAUS_NORM) is much more similar to others, including those based on evolutionary arguments. To form a better picture of the relationship between the matrices, the analysis was repeated omitting the four matrices (SM_SAUSAGE, RISLER, FITCH, IDENTITY) lying farthest away from the rest of the matrix set (see Figure 2).

When assessing the magnitude of distances between the matrices, one should note that matrices like DAYHOFF, JONES and GONNET (Dayhoff *et al.*, 1978; Gonnet *et al.*, 1992; Jones *et al.*, 1992b) or JOHNSON and RUSSELL_RH (Johnson and Overington, 1993; Russell *et al.*, 1997) are based generally on the same procedure applied on different data sets. All reflect counting interchanges of residues in a set of sequentially or structurally aligned pairs. In this light, it is surprising just how similar the force field based matrices are to literature examples, despite no alignments whatsoever being used in the construction process. SM_THREADER, in particular, results in a matrix which is well within the family including BLOSUM and GONNET matrices. The normalization process moves matrices from both force fields closer to conventional examples, but SM_SAUS_NORM does remain quite different. If anything, it is closest to the RUSSELL_RH matrix based on structural alignments of remote homologues (Russell *et al.*, 1997) and the MIYAZAWA matrix based on contact frequencies (Miyazawa and Jernigan, 1993). This may not be surprising. The two force fields are based on different construction philosophies which emphasize different aspects of sequence to structure alignment and fold recognition.

Decomposing into physico-chemical parameters

The analysis of matrices at the amino acid level gives some explanation for the observed similarities and differences between matrices of completely different origins. The substitution matrices were represented as a projection onto the first two principal components (Figure 3). Additionally, amino acids have been clustered and formed into trees in Figure 4. The first panel (a) of both figures simply show that in the sausage force field, Cys has the most unusual properties and is quite unique. To improve the resolution amongst the other amino acids, SM_SAUSAGE was reanalyzed with Cys omitted (panel (b) of Figures 3 and 4). The plot for the GONNET matrix is also shown for comparison.

The principal component analysis shows a remark-

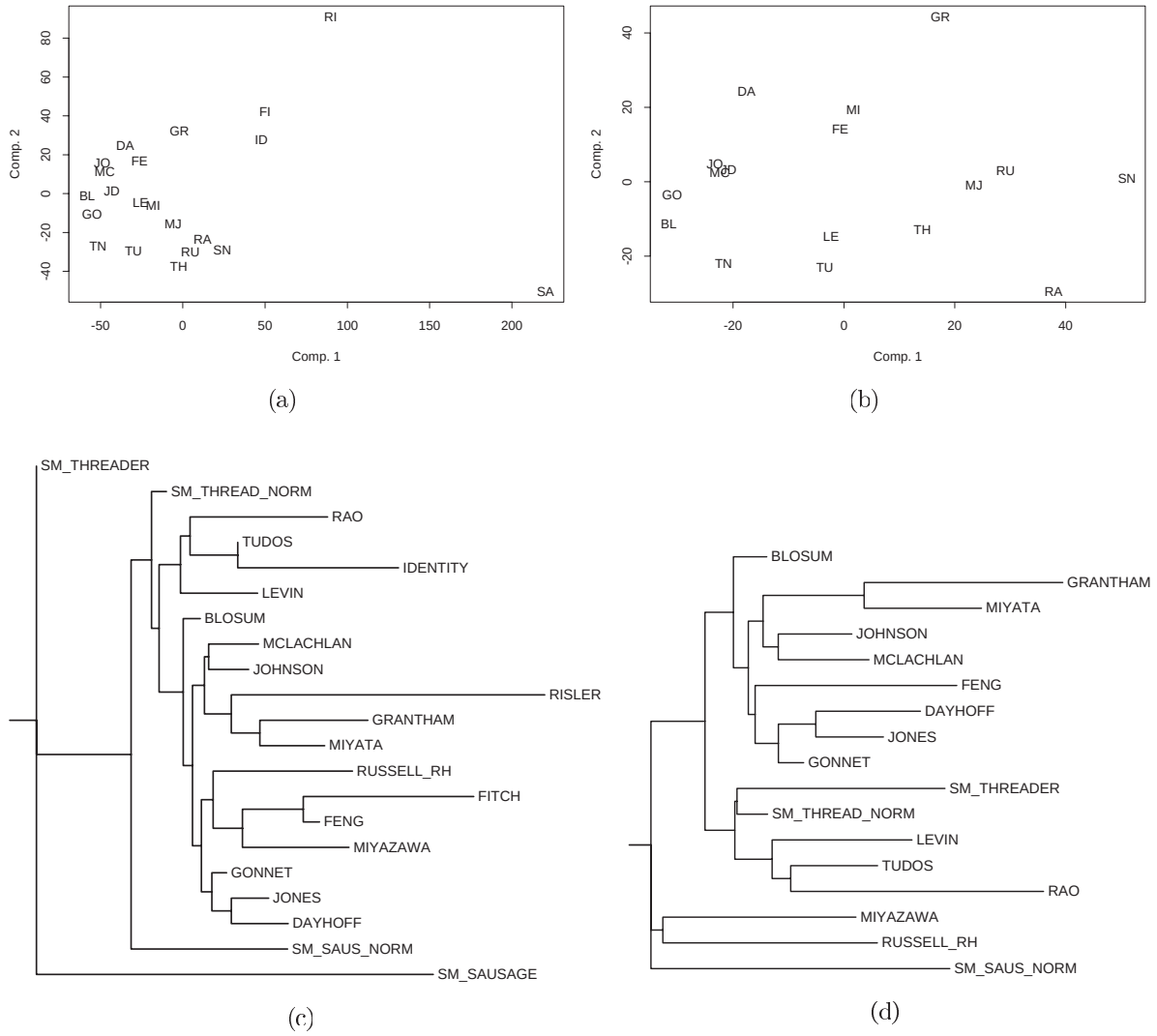

Fig. 2. Principal component analysis (a, b) and tree (c, d) representation of distances between matrices. On the left (a), (c) the distance was calculated for all matrices. Right panels, (b), (d) are the same data, but four outlying matrices (SM_SAUSAGE (SA), RISLER (RI), FITCH (FI), IDENTITY (ID)) have been omitted. The key for the other matrices: SM_SAUS_NORM (SN), SM_THREADER (TH), SM_THREAD_NORM (TN), DAYHOFF (DA), FENG (FE), GONNET (GO), GRANTHAM (GR), BLOSUM (BL), JOHNSON (JO), JONES (JD), LEVIN (LE), MCLACHLAN (MC), MIYATA (MI), MIYAZAWA (MJ), RAO (RA), RUSSELL_RH (RU) and TUDOS (TU). See Table 1 for the description of matrices. The tree on the left (c) is the best of 2587 computed, the sum of squares is 7.08 and the average percentage of standard deviation is 13.02. The corresponding values for panel (d) are 32 149, 4.14, 12.38.

able similarity between the different matrices, most pronounced along the first eigenvector (Figure 3). This essentially appears to be a hydrophobicity index with residues to the right being hydrophobic (Leu, Ile, Val, Trp, Phe, Tyr and Cys). The second group, to the left, is composed of charged and polar residues (Asp, Asn, Arg, Glu, Gln, Lys, His, Ser, Thr), as well as Gly and Pro. The correlation coefficient between the first eigenvector and the hydrophobicity scale of Nishikawa and Ooi (1986), which is based on the contact number of residues, is indeed above 0.9 for all three matrices. The two groups of amino acids are also separated in the hierarchical trees according to their hydrophobicity (Figure 4).

An interesting difference between evolutionary based and force field based matrices occurs in the treatment of aromatic amino acids. In the former, these are the most conserved amino acids (Johnson and Overington, 1993), and in the projection they form a separate group lying even

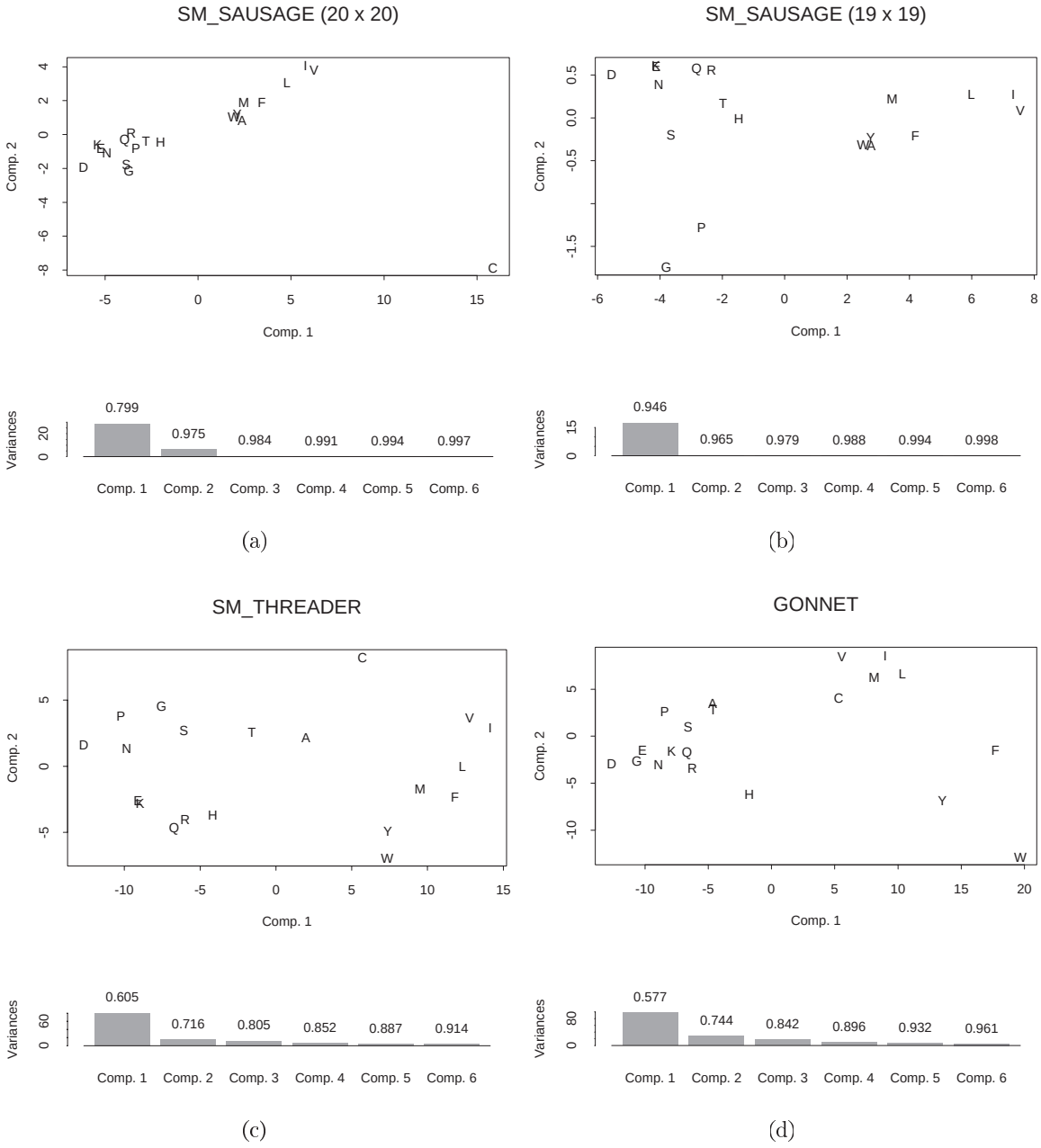

Fig. 3. The first two principal component representation of the SM_SAUSAGE substitution matrix for 20 × 20 elements (a), with 19 × 19 elements obtained by omitting cysteine (b), for the SM_THREADER (c) and GONNET (d) matrices. The barplot of the variances for the first six eigenvalues is also given for each matrix. The cumulative fraction of the variance is printed above each bar in the plot.

further away from the polar group than other hydrophobic amino acids (Figure 3).

To better characterize the force fields and substitution matrices, one can look at the variances for each eigenvector, describing the relative weights of the eigenvalues. Figure 3 shows that the main difference between the two force field based matrices is the dominance of hydrophobicity in the sausage force field. While the first eigenvalue of SM_SAUSAGE accounts for 90% of the variance, the corresponding values are 60 and 57% for the SM_THREADER and GONNET matrices, respectively. The same effect also appears in the tree representation (Figure 4), as the branch between the first and second group of amino acids is significantly longer for the

Fig. 4. The tree representation of the amino acids according to the SM_SAUSAGE matrix with 20 × 20 elements (a), with 19 × 19 elements omitting Cys (b), the SM_THREADER matrix (c) and the GONNET matrix (d). The number of trees computed, the sum of squares and the average percentage of standard deviation are the following: (a) SM_SAUSAGE (20 × 20) 21 339, 2.64, 8.35 (b) SM_SAUSAGE (19 × 19) 28 279, 2.04, 7.75 (c) SM_THREADER 32 561, 3.81, 10.04 (d) GONNET 20 817, 3.89, 10.14.

SM_SAUSAGE matrix compared to SM_THREADER or GONNET matrices. It has been suggested that other factors such as volume or composition play a significant role in substitution matrices (Grantham, 1974; Tomii and Kanehisa, 1996). In the sausage force field, these are obviously less important.

Alignment testing

Table 5 summarizes the ability of each substitution matrix to perform in sequence alignments. The test sets differed not only in the structural alignment method they used, but also in the similarity range they covered. The percentage of equivalenced residues ranged from 81% in the case of medium homologues of Russell *et al.* (1997) to 54% for the remote homologous test set made by the same authors. However, the most difficult test set was the BASIC superfamily set (Rychlewski *et al.*, 2000), despite the higher sequence similarity cut-off and higher percentage of equivalenced residues (Table 2). In agreement with literature tests (Vogt *et al.*, 1995; Johnson and Overington,

1993), the evolutionary based matrices performed very well, but generally the differences were quite small. Although it is apparent that the best matrix can be different for each test set, there is a high correlation of the ranks of matrices in the different tests. The optimized parameters are given in the supplementary material (http://www.rsc.anu.edu.au/∼zsuzsa/suppl/matrices.html).

The performance of substitution matrices based on the two force fields are quite different. The alignment accuracy using SM_THREADER is the best on two out of the five test sets, and the best overall, but SM_SAUSAGE does not perform so well. The normalization procedure (equation (3)) benefits SM_SAUSAGE, but if anything, is deleterious to SM_THREADER.

DISCUSSION

This work has presented example substitution matrices derived from force fields and prompts an obvious question. Do these matrices work as well as conventional ones? It would appear that the THREADER-based matrices are candidates for use with distantly related sequences. The true answer is that the best matrix is problem specific. Different substitution matrices perform differently on test sets with varying degrees of similarity. One could even pursue this and try different measures of performance such as remote homologue detection, avoidance of false positives and different measures of alignment quality. Certainly it seems safe to say that the new matrices are comparable to literature standards, despite their different construction method. Without debating details of test sets, it is also clear that the new matrices are of little use for obviously related sequences. They could only be useful for detection or alignment of remote homologues. Finally, when considering the substitution matrix performance, it would appear that the methodology may be worth applying to other force fields and more generally, to any kind of scoring methodology producing position specific substitution values. It could be that the THREADER-based matrices are the best possible, but it is more likely that there are even better matrices waiting to be built. This work simply presents the methodology for a new generation of substitution matrices which have less obvious connection with evolution.

Conventional substitution matrices are often characterized by the intended evolutionary distance. A PAM 30 matrix would be used for closely related proteins while a PAM 250 matrix would be appropriate for proteins which have greatly diverged (Altschul, 1993). In matrices based on structural alignments, the connection to evolution is less obvious, but there are thresholds in the selection of proteins for alignment and the sites within those proteins. In the force field based approach, there is nothing like evolutionary distance and none of the thresholds associated with structural alignments. In the case of the sausage based matrix there are no alignments used at any stage of the matrix construction. It is worth noting that the specific example of the THREADER score function does contain a sequence similarity term with implicit thresholds.

Unlike almost every other approach, the force field based methodology has the unusual property of producing asymmetric substitution matrices. This could be a feature or problem. In principle, this may be a problem when comparing two sequences of equal importance. In practice, it will be a problem using software which can only handle symmetric matrices. At the same time, there are applications where asymmetric matrices are entirely appropriate. If one has a parent sequence from which others may have diverged, then the asymmetry is quite natural. One may accept that different types of amino acid are conserved to different degrees. Statistically, residues of type A may be more conserved than residues of type B. It is entirely appropriate that changes of A to B are more heavily penalized than B to A. In principle, one should be able to answer the question of the merit of asymmetric substitution matrices by following Kann *et al.* (2000) and explicitly optimizing 400 parameters from an asymmetric matrix.

Aside from asymmetry, the methodology has the other interesting property of giving some estimate of what one might call reliability. The energy distributions are usually close to Gaussian curves with a characteristic width (quantified by the standard deviation). Ultimately, one should be able to use this information. Residue substitutions of A to B and C to D may both have similarity of 10 units, but the distribution may be much broader for one of them. In sequence alignment one would use the value of 10 in both cases, but in one case it may be far less significant. In the future, it would be of interest to see if there is any correlation between reliability of individual elements and the reliability of an alignment or sequence homologue detection.

With any knowledge-based method, there will always be the question as to whether the data contains any systematic biases or errors. Essentially, the method is automatic. The protein calibration list could be skewed in some manner, and it is certainly dominated by smaller water soluble proteins. The Cys_Cys data shows what happens when one has a distinct sub-population (one can see extra peaks in the energy distributions). The other energy distributions do not show extra peaks or even shoulders. There is however, a more serious concern. The force field-derived substitution matrices reflect all the errors, biases and statistical anomalies of the source force field. As an example, the SM_SAUSAGE matrix is more heavily dominated by hydrophobicity than other matrices and this reflects a bias in the source force field.

Rather than ask if the resulting substitution matrices are universally applicable tools for sequence analysis, one

Table 5. Alignment accuracy

Test sets Matrix names	MEDIUM	REMOTE	HELLM	BASIC family	BASIC superfamily
SM_THREADER	88.3	36.2	39.8	47.0	20.7
SM_THREAD_NORM	88.5	36.0	39.1	46.3	19.9
GONNET	88.7	34.4	38.6	47.3	21.4
BLOSUM	87.9	35.3	37.3	47.2	19.9
JOHNSON	88.3	35.0	37.6	46.2	18.5
MCLACHLAN	87.1	33.8	36.5	44.5	17.9
DAYHOFF	87.1	33.3	36.1	44.6	20.4
RAO	85.4	33.1	35.8	43.7	16.3
TUDOS	85.7	33.7	35.8	44.8	17.0
GRANTHAM	87.8	33.7	35.4	43.8	18.4
LEVIN	85.6	33.7	35.3	43.6	16.7
MIYATA	85.9	32.4	35.0	43.6	16.2
JONES	86.7	32.1	34.9	44.9	20.6
FENG	85.0	30.9	34.8	42.4	16.1
SM_SAUS_NORM	86.7	30.5	33.5	42.3	16.4
RISLER	84.8	30.1	32.9	42.5	16.8
MIYAZAWA	84.5	31.6	31.9	41.9	16.2
RUSSELL_RH	82.7	31.1	31.1	39.7	14.9
SM_SAUSAGE	84.1	29.4	30.4	36.7	14.4
IDENTITY	81.2	26.0	25.7	35.4	10.9
FITCH	80.2	27.4	25.4	32.8	9.2

The alignment accuracy was calculated as the average percentage of correctly aligned residues compared to structurally equivalent residues. Matrices are sorted according to their average performance on the five test sets. The force-field based matrices are printed in bold. The description of matrices is given in Tables 1 and 2 contains information on the different test sets.

can reverse the question. From analyzing the properties of a substitution matrix, what features of the force field stand out? By far the best example of this comes from the principal component analysis. The single most important term in the sausage force field simply reflects disulfide bonding. This could even be seen if one were to look at the raw force field parameters (data not shown). More interesting is the force field's dominance by hydrophobicity. Principal components do not come with physical labels, but the clustering in Figure 3 shows a grouping of amino acids by what appears to be very conventional hydrophobicity. This kind of analysis can now be used to compare force fields and substitution matrices. If one neglects cysteines, the first principal component (hydrophobicity) from the sausage-derived matrix accounts for 95% of the total variance of the matrix. The GONNET matrix is also dominated by hydrophobicity, but the relative weight of the first eigenvector is only 58%. From the point of view of force field construction, this is a convenient and useful data point. Given two force fields, one can say, for example, which features are emphasized in the better performing force field.

The methodology presented here raises many questions. It can be used to generate data for two kinds of cysteines which would not be possible with a conventional approach and less simple with a structure alignment approach. Can a sequence comparison method take advantage of this? In cases where one has a true reference sequence, can one take advantage of the natural asymmetry in the substitution matrices? Would other force fields produce even better substitution matrices? Given that sequence and structure based alignments do not produce identical results, is it possible that evolution and force field-based matrices are better for different problem areas? These issues suggest the direction of future work.

ACKNOWLEDGEMENTS

We are grateful to James B. Procter for the simplex code used in the optimization calculations and Dr Thomas Huber for providing the code used for calculating the sausage force field scores. Professor David T. Jones freely provided the THREADER executable.

REFERENCES

Altschul,S.F. (1993) A protein alignment scoring system sensitive at all evolutionary distances. *J. Mol. Evol.*, **36**, 290–300.

Becker,R.A., Chambers,J.M. and Wilks,A.R. (1988) *The New S Language: a Programming Environment for Data Analysis and Graphics*. Wadsworth.

Berezovsky,I.N., Esipova,N.G. and Tumanyan,V.G. (2000) Hierarchy of regions of amino acid sequence with respect to their role in the protein spatial structure. *J. Comput. Biol.*, **7**, 183–192.

Blake,J.D. and Cohen,F.E. (2001) Pairwise sequence alignment below the twilight zone. *J. Mol. Biol.*, **307**, 721–735.

Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. Matrices for detecting distant relationships. In Dayhoff,M.O. (ed.), *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3, National Biomedical Research Foundation, Washington DC, pp. 345–358.

Domingues,F.S., Lackner,P., Andreeva,A. and Sippl,M.J. (2000) Structure-based evaluation of sequence comparison and fold recognition alignment accuracy. *J. Mol. Biol.*, **297**, 1003–1013.

Felsenstein,J. (1993) *PHYLIP (Phylogeny Inference Package) Version 3.5c*, Distributed by the author, Department of Genetics, University of Washington, Seattle.

Feng,D.F., Johnson,M.S. and Doolittle,R.F. (1985) Aligning amino acid sequences: comparison of commonly used methods. *J. Mol. Evol.*, **21**, 112–125.

Feng,Z.K. and Sippl,M.J. (1996) Optimum superimposition of protein structures: ambiguities and implications. *Fold. Des.*, **1**, 123–132.

Fitch,W.M. (1966) An improved method of testing for evolutionary homology. *J. Mol. Biol.*, **16**, 9–16.

George,D.G., Barker,W.C. and Hunt,L.T. (1990) Mutation data matrix and its uses. *Meth. Enzymol.*, **183**, 333–351.

Gonnet,G.H., Cohen,M.A. and Benner,S.A. (1992) Exhaustive matching of the entire protein sequence database. *Science*, **256**, 1443–1445.

Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.

Grantham,R. (1974) Amino acid difference formula to help explain protein evolution. *Science*, **185**, 862–864.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10 915–10 919.

Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.

Holm,L. and Sander,C. (1998) Dictionary of recurrent domains in protein structures. *Proteins*, **33**, 88–96.

Huber,T., Russell,A.J., Ayers,D. and Torda,A.E. (1999) Sausage: protein threading with flexible force fields. *Bioinformatics*, **15**, 1064–1065.

Huber,T. and Torda,A.E. (1998) Protein fold recognition without Boltzmann statistics or explicit physical basis. *Protein Sci.*, **7**, 142–149.

Johnson,M.S. and Overington,J.P. (1993) A structural basis for sequence comparisons. An evaluation of scoring methodologies. *J. Mol. Biol.*, **233**, 716–738.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992a) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992b) The rapid generation of mutation data matrices from protein sequences. *Comput. Appl. Biosci.*, **8**, 275–282.

Kann,M., Qian,B. and Goldstein,R.A. (2000) Optimization of a new score function for the detection of remote homologs. *Proteins*, **41**, 498–503.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Levin,J.M., Robson,B. and Garnier,J. (1986) An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.*, **205**, 303–308.

May,A.C. (1999) Towards more meaningful hierarchical classification of amino acid scoring matrices. *Protein Eng.*, **12**, 707–712.

McLachlan,A.D. (1971) Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c551. *J. Mol. Biol.*, **61**, 409–424.

Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.

Miyazawa,S. and Jernigan,R.L. (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. *Protein Eng.*, **6**, 267–278.

Mohana Rao,J.K. (1987) New scoring matrix for amino acid residue exchanges based on residue characteristic physical parameters. *Int. J. Pept. Protein Res.*, **29**, 276–281.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Niefind,K. and Schomburg,D. (1991) Amino acid similarity coefficients for protein modeling and sequence alignment derived from main-chain folding angles. *J. Mol. Biol.*, **219**, 481–497.

Nishikawa,K. and Ooi,T. (1986) Radial locations of amino acid residues in a globular protein: correlation with the sequence. *J. Biochem. (Tokyo)*, **100**, 1043–1047.

Overington,J.P., Donnelly,D., Johnson,M.S., Sali,A. and Blundell,T.L. (1992) Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*, **1**, 216–226.

Press,W.H., Teukolsky,S.A., Vetterling,W.T. and Flannery,B.P. (1992) *Numerical Recipes in C*. Cambridge University Press.

Prlic,A., Domingues,F.S. and Sippl,M.J. (2000) Structure-derived substitution matrices for alignment of distantly related sequences. *Protein Eng.*, **13**, 545–550.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol.*, **204**, 1019–1029.

Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **14**, 309–323.

Russell,R.B., Saqi,M.A., Sayle,R.A., Bates,P.A. and Sternberg,M.J. (1997) Recognition of analogous and homologous protein folds: analysis of sequence and structure conservation. *J. Mol. Biol.*, **269**, 423–439.

Rychlewski,L., Jaroszewski,L., Li,W. and Godzik,A. (2000) Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.*, **9**, 232–241.

Tomii,K. and Kanehisa,M. (1996) Analysis of amino acid indices and mutation matrices for sequence comparison and structure

prediction of proteins. *Protein Eng.*, **9**, 27–36.
Tüdős,E., Cserző,M. and Simon,I. (1990) Predicting isomorphic residue replacements for protein design. *Int. J. Pept. Protein Res.*, **36**, 236–239.
Vogt,G., Etzold,T. and Argos,P. (1995) An assessment of amino acid exchange matrices in aligning protein sequences: the twilight zonerevisited. *J. Mol. Biol.*, **249**, 816–831.
Wang,Y., Lal,L., Li,S., Han,Y. and Tang,Y. (1996) Position-dependent protein mutant profile based on mean force field calculation. *Protein Eng.*, **9**, 479–484.