BATMAS30: Amino Acid Substitution Matrix for Alignment of Bacterial Transporters

Roman A. Sutormin,[1*] Aleksandra B. Rakhmaninova,[2] and Mikhail S. Gelfand[1,2]
[1]*State Scientific Center GosNIIGenetica, Moscow, Russia*
[2]*Integrated Genomics, Moscow, Russia*

ABSTRACT **Aligned amino acid sequences of three functionally independent samples of transmembrane (TM) transport proteins have been analyzed. The concept of TM-kernel is proposed as the most probable transmembrane region of a sequence. The average amino acid composition of TM-kernels differs from the published amino acid composition of transmembrane segments. TM-kernels contain more alanines, glycines, and less polar, charged, and aromatic residues in contrast to non-TM-proteins. There are also differences between TM-kernels of bacterial and eukaryotic proteins. We have constructed amino acid substitution matrices for bacterial TM-kernels, named the BATMAS (BActerial Transmembrane MAtrix of Substitutions) series. In TM-kernels, polar and charged residues, as well as proline and tyrosine, are highly conserved, whereas there are more substitutions within the group of hydrophobic residues, in contrast to non-TM-proteins that have fewer, relatively more conserved, hydrophobic residues. These results demonstrate that alignment of transmembrane proteins should be based on at least two amino acid substitution matrices, one for loops (e.g., the BLOSUM series) and one for TM-segments (the BATMAS series), and the choice of the TM-matrix should be different for eukaryotic and bacterial proteins. Proteins 2003; 51:85–95.** **© 2003 Wiley-Liss, Inc.**

Key words: comparative analysis; transport proteins; amino acid substitution matrix; evolution; transmembrane segments

INTRODUCTION

The growth of databases describing various characteristics of proteins, such as amino acid sequence, spatial structure, function, functional domains, etc., allows one to describe new proteins, at least at the first approximation, comparing the sequences under analysis to already known ones. Most comparative techniques involve alignment of amino acid sequences that, in turn, depends on amino acid substitution matrices. Thus, it is crucial to develop adequate substitution matrices for different functional regions of proteins.

The best known and the most commonly used substitution matrices are the BLOSUM and PAM series, obtained by statistical analysis of large samples of amino acid sequences.[1,2] It becomes increasingly clear that in order to align proteins with non-standard physical and chemical characteristics and amino acid composition, specific matrices are required. Among such proteins is the group of transmembrane (TM) hydrophobic proteins. The idea that transmembrane proteins should be aligned using two different matrices at the same time, one for hydrophobic membrane segments and the other for hydrophilic loops, was repeatedly discussed. TM-specific scoring matrices derived using PHDhtm, an algorithm predicting TM-segments in multiple alignment by neural networks, were published in Ng et al.[3] and Muller et al.[4] A substitution matrix for highly homologous TM-proteins based on SwissProt annotations was constructed,[5] and then the Dayhoff mutation model was applied to derive matrices for comparison of more distant proteins. In all these studies, bacterial and eukaryotic proteins were combined into a single sample. As it will be shown below, statistical properties of bacterial and eukaryotic TM-segments differ and thus the transmembrane proteins of eubacterial and eukariotic origin should be considered separately.

The main problem arising during construction of substitution or score matrices for transmembrane proteins is the fact that in most cases it is not known what part of a protein actually resides within the membrane. The reason is that transmembrane proteins crystallize poorly, and thus only a few such proteins have known spatial structures determined by the X-ray analysis.[6,7] Different methods for prediction of transmembrane segments yield contradictory results when applied to the same sequence; for a typical example see Figure 1.

At the same time, a large number of known transmembrane proteins allows one to apply the comparative analysis for verification of predicted TM-segments using various criteria of consistency. A somewhat similar approach was used to predict TM-segments by consensus methods.[8,9] We use two criteria: agreement between five different TM-

Grant sponsor: Howard Hughes Medical Institute; Grant number: 55000309; Grant sponsor: INTAS; Grant number: 99-1476; Grant sponsor: Ludwig Institute for Cancer Research; Grant number: CRDF RB0-1268; Grant sponsor: Russian Fund of Basic Research; Grant number: 00-15-99362.

*Correspondence to: Roman A. Sutormin, State Scientific Center GosNIIGenetica, Moscow, 113545, Russia. E-mail: sutor_ra@mail.ru

Received 15 March 2002; Accepted 12 September 2002

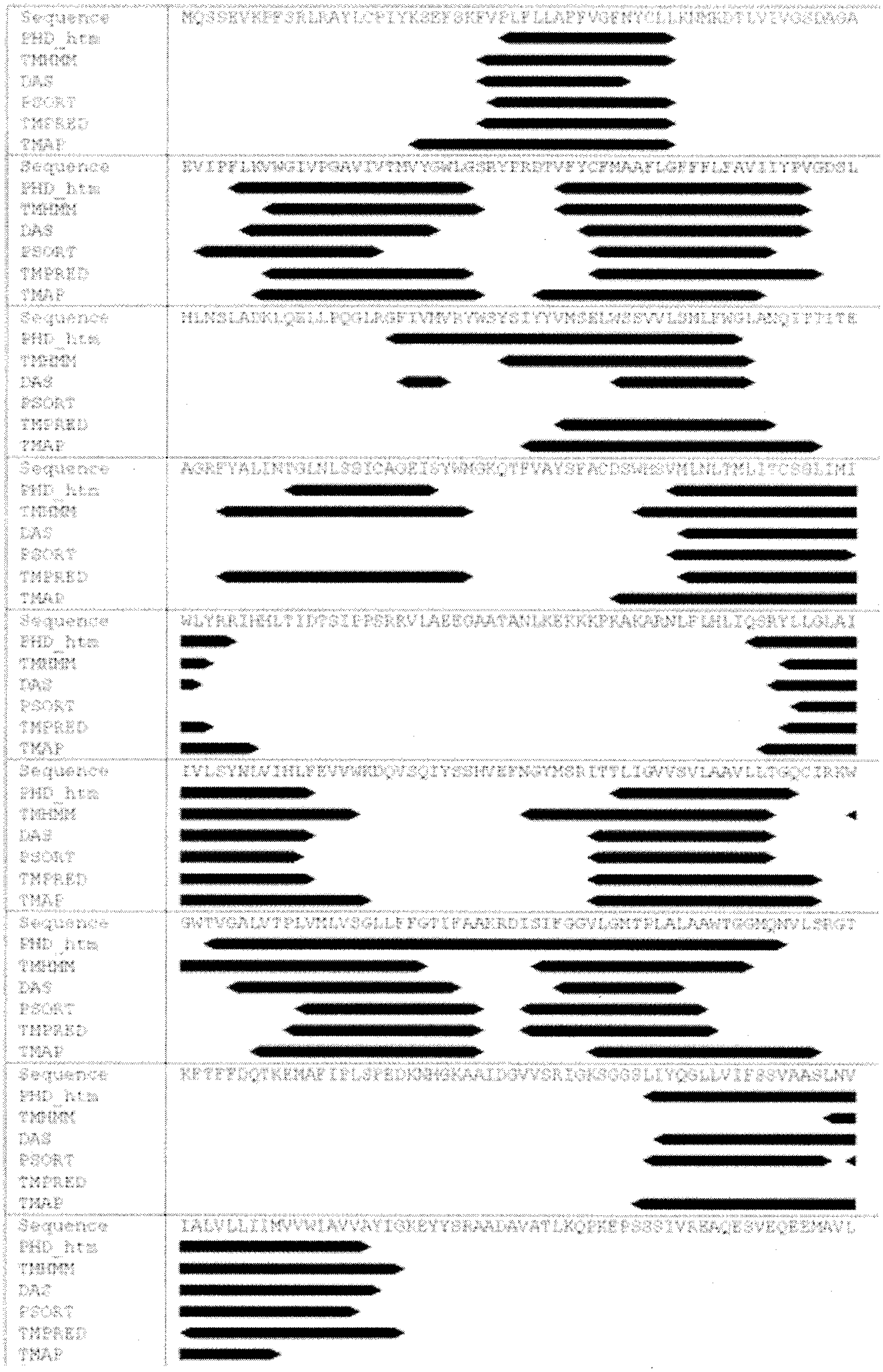

Fig. 1. Prediction of TM-segments for protein Q9Z7U0 by six methods.

segment prediction algorithms, and then consistency of predictions for homologous proteins. The filtered aligned TM-segments are used to construct a specific amino acid substitution matrix. Matrices were constructed for three independent representative sets: eubacterial secondary transporters (class TC.2A by the Saier-Paulsen classification[10–12]), eubacterial ABC-transporters (TC.3A.1), and eukaryotic secondary transporters (TC.2A).

TABLE I. Characteristics of Clusters Used for Construction of the BATMAS Series of Amino Acid Substitution Matrices

	Clustering threshold					
	30	40	50	60	70	80
Clusters	213	322	345	319	270	237
Proteins per cluster (average)	9.5	5.0	3.6	3.0	2.7	2.5
Protein pairs	6854	2356	909	429	176	496
Amino acid pairs in kernels	1,254,754	449,147	177,371	87,186	35,610	97,794

METHODS

The Main Sample: Bacterial Transporters of Class TC.2A

Initially, all bacterial members of class TC.2A from the literature[10–12] and web-sites (http://www-biology.ucsd.edu/~msaier/transport/, http://www.biology.ucsd.edu/~ipaulsen/transport/) were collected. The size of the main sample was 1,312 sequences from 101 families of bacterial proteins.

Four of these families are members of the MFS-superfamily (438 sequences), five families are members of the APC-superfamily (103 sequences), and four families belong to the RND-superfamily (114 sequences). The CPA3 and NFE families include complex multicomponent transport systems consisting of more than two polypeptide chains and were not considered.

Then, each sequence from the basic sample was used as a seed for BLAST[13] homology search in eubacterial genomes in the ERGO system[14] (http://ergo.integratedgenomics.com/ERGO/). Only genomes satisfying the criterion of sufficient completeness were considered: the genome had to encode more than 500 genes in at most ten contigs. Thirty-one such genome were selected: *Aquifex aeolicus, Brucella melitensis, Bacillus subtilis, Buchnera sp. APS, Campylobacter jejuni, Chlamydia muridarum, Chlamydia pneumoniae, Chlamydia trachomatis, Deinococcus radiodurans, Escherichia coli, Fusobacterium nucleatum, Haemophilus influenzae, Helicobacter pylori, Lactococcus lactis, Mycoplasma genitalium, Mycoplasma pneumoniae, Mycobacterium tuberculosis, Neisseria meningitidis, Pasteurella multocida, Pseudomonas aeruginosa, Rhodobacter capsulatus, Rickettsia prowazekii, Salmonella typhi, Salmonella typhimurium, Streptococcus pyogenes, Synechocystis sp., Thermotoga maritima, Treponema pallidum, Ureaplasma urealyticum, Vibrio cholerae, Xylella fastidiosa.* This resulted in an additional 860 homologs to the basic sequences (identity $\geq$ 30%, E-value $\leq 10^{-10}$). The final sample contained 2,172 proteins.

Clustering and Alignment

The sample was divided into clusters using the nearest neighbor procedure with the percent identity of the BLAST alignment serving as the measure of closeness. A series of clusters was constructed with the lower value of threshold (MIN_IDENT) set to 30, 40,..., 80%. When the size of a cluster exceeded 50 proteins, a cluster was divided into several clusters by raising the lower threshold of clustering. That happened four times for MIN_IDENT = 30%, once for MIN_IDENT = 40%, and never for MIN_IDENT $\geq$ 50%. Clusters with MIN_IDENT = 30% generally coincided with the families of bacterial transporters according to the Saier-Paulsen classification.[10–12] At that, very large families were broken into 2 or 3 clusters, and 235 sequences were not included in any cluster. Characteristics of the clusters are shown in Table I. Then each cluster was aligned using CLUSTALW.[15] Correctness of alignments with low identity percentage is discussed below in Results and Discussion.

Determining the Transmembrane Kernels

We define transmembrane kernels (TM-kernels) as parts of the sequences consistently predicted to be transmembrane segments. Two conditions were used: agreement of several prediction algorithms and consistency of predictions for homologous proteins (see Fig. 2).

- A position in an amino acid sequence is *tentatively transmembrane* (TM-residue) if this position is predicted to belong to a TM-segment by at least three servers out of five: TMHMM[16] (http://www.cbs.dtu.dk/services/TMHMM-2.0/), TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html), DAS[17] (http://www.sbc.su.se/~miklos/DAS/maindas.html), TMAP[18] (http://www.mbb.ki.se/tmap/), PSORT[19] (http://psort.nibb.ac.jp/form.html). Web variants of these programs with default settings were used. *TM-runs* of each sequence were defined as groups of adjacent TM-residues of the sequence. *TM-kernels* in a cluster were defined as groups of adjacent columns in the multiple alignment if each column contained at least 60% of TM-residues. TM-kernels in a protein were defined as groups of positions that belong to the TM-kernel of the cluster. Kernels with gaps were allowed in order to retain data in cases when a large part of some protein in alignment is missing. Gaps within kernels are rare. In particular, for clusters with MIN_IDENT = 30% the number of kernel positions corresponding to gaps does not exceed 1.4%. For clusters with MIN_IDENT $\geq$ 40%, positions with gaps constitute less than 0.7% of all kernel positions. Thus, the constructed alignments are consistent and likely phylogenetically adequate.

Construction of the BATMAS Matrices

In each cluster, all pairs of sequences with identity from MIN_IDENT through MIN_IDENT+10% were considered. The alignment of the pair induced by the cluster

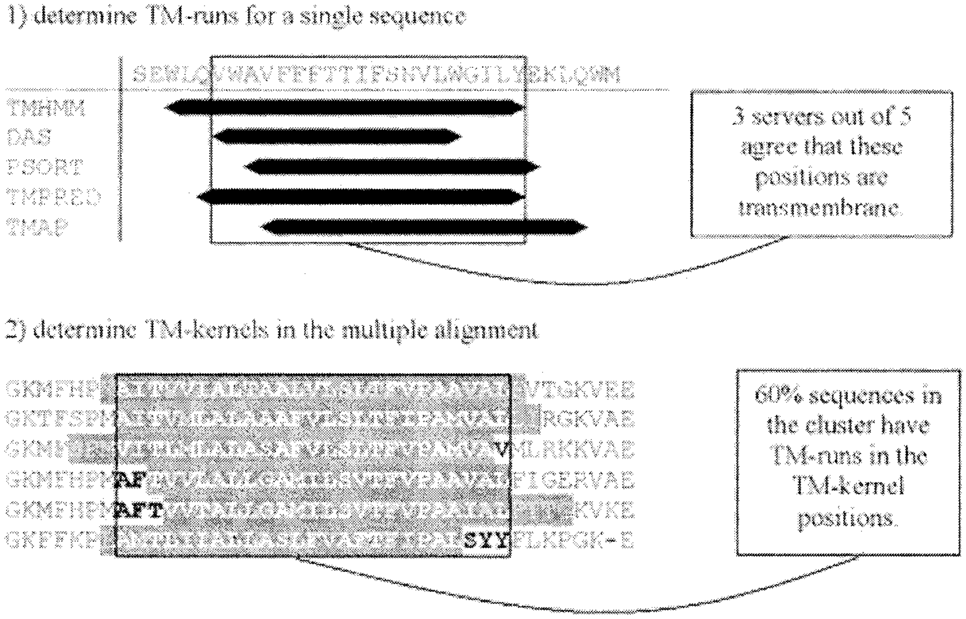
1) determine TM-runs for a single sequence
SEWLQVWAVFFFTTIFSNVLWGILYEKLQWM
TMHMM
DAS
PSORT
TMPRED
TMAP
3 servers out of 5 agree that these positions are transmembrane.
2) determine TM-kernels in the multiple alignment
GKMFHP
VTGKVEE
GKTFSPM
RGKVAE
GKM
V
MLRKKVAE
GKMFHPM
AF
FIGERVAE
GKMFHPM
AFT
KVKE
GKFFKP
SYY
FLKPGK-E
60% sequences in the cluster have TM-runs in the TM-kernel positions.

Fig. 2. Construction of TM-kernels.

alignment and the TM-kernels for this cluster were used to compute the number of matching amino acid pairs. Then each element of the substitution matrix was divided by the sum of all matrix elements. Thus the count matrix was converted to the frequency matrix normalized to 1.

Thus, we obtained a series of matrices for different values of MIN_IDENT, named BATMAS30, BATMAS40 …BATMAS80. As shown in Table I, BATMAS30 was constructed using more then 1,250,000 amino acid pairs. In order to construct the BLOSUM62 matrix, approximately the same number of amino acid pairs were used.[1] The number of amino acid pairs used in Jones et al.[5] and Ng et al.[3] to construct TM-matrices is not given in the original publications.

The Control Sets

In order to verify the robustness of the obtained results, we also considered two control sets. The first set was represented by eubacterial ABC-transporters. It was constructed using 116 eubacterial ABC transporters listed at the Saier web site expanded as described in Methods. The final sample of ABC transporters consisted of 760 proteins.

The second control set was represented by eukaryotic secondary transporters. Initially, this set consisted of eukaryotic proteins of class TC.2A listed on the same web-site, except for the ones annotated as the proteins of intracellular membranes (mitochondrial, lysosomal, chloroplast envelope, etc). Then eukaryotic proteins from SwissProt (release 40.0) were used to expand the initial set using the procedure described in Methods.

Construction of Dendrograms

In order to determine the functional role of amino acids, we constructed dendrograms reflecting behavior of amino acids in TM-kernels. The iterative procedure was as follows:

- for all pairs of amino acids *i,j* compute $l_{ij} = f_{ij}/(d_i \cdot d_j)$, where f_{ij} is the substitution frequency (taken from the BATMAS matrices), d_i is the amino acid probability,

$$d_i = \sum_{j=1}^{20} f_{ij} \qquad (1)$$

- merge amino acids *i,j* corresponding to the maximum value l_{ij} into a group and then treat this group as a degenerate amino acid;
- recompute the substitution frequencies and the amino acid probabilities.

RESULTS AND DISCUSSION

TM-Segments and TM-Kernels in the Main Sample

The frequency distribution of lengths of TM-segments predicted by different servers and of TM-kernels is shown in Figure 3. The average length of a TM-kernel is 18 amino acids. That is shorter than the average length of a TM-segment (20–21 amino acids).

Most TM-kernels are central parts of TM-segments. However, very short (1–4 amino acids) and very long TM-kernels exist. A short TM-kernel arises when a position is alternatively ascribed to two adjacent TM-segments by different servers in different proteins. A long kernel arises when different servers disagree about the location of a short loop between two long TM-segments in related proteins. The contribution of extra-short and extra-long TM-kernels to the matrices is negligibly small (in the case

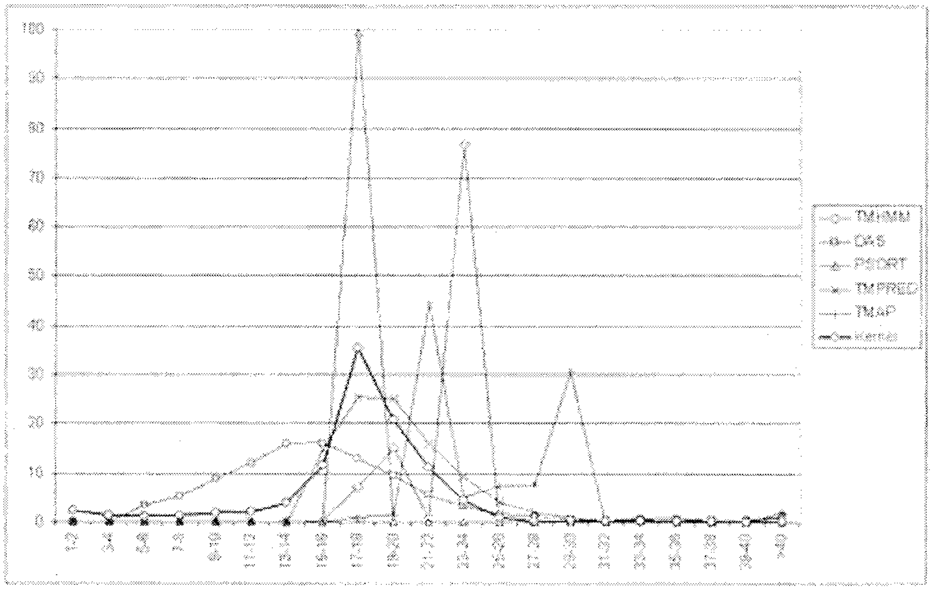

Fig. 3. Distribution of lengths of TM-segments, predicted by different servers, and of TM-kernels.

of BATMAS30 the TM-kernels of 1–4 amino acids account for 0.58% and TM-kernels longer than 35 amino acids account for 0.67% of all positions).

Amino Acid Composition of TM-Kernels From the Main Sample

The average amino acid composition of proteins, of the TM-segments according to Jones et al.[5] and Ng et al.,[3] and of the TM-kernels is given in Table II. As expected, the fraction of hydrophobic amino acids in both TM-kernels and TM-segments is markedly higher than in proteins in general.

However, the amino acid composition of the TM-kernels clearly differs from that of TM-segments. Indeed, the TM-kernels contain less polar and charged residues than the predicted TM-segments. Thus, the content of D, E, H, K, R, N, and Q totals 5% in the TM-kernels, but is 12–13% in the TM-segments. Interestingly, the content of negatively charged (D, E) and positively charged (R, K) residues almost coincides in the TM-kernels, whereas in the TM-

TABLE II. Amino Acid Composition of Different Matrices (in %)†

		A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
I.	SWISS-PROT (Ng et al.)[3]	7.6	1.7	5.3	6.4	4.1	6.8	2.2	5.8	5.9	9.4	2.4	4.5	4.9	4.0	5.1	7.1	5.7	6.6	1.2	3.2
	BLOSUM62	7.4	2.5	5.4	5.4	4.7	7.4	2.6	6.8	5.8	9.9	2.8	4.5	3.9	3.4	5.2	5.7	5.1	7.3	1.3	3.2
II.	Jones et al.[5]	10.51	2.19	0.89	0.97	7.77	7.58	1.68	11.88	1.12	16.35	3.33	1.85	2.60	1.41	1.57	5.68	5.23	11.95	2.23	3.24
	PHDhtm (Ng et al.)[3]	8.8	2.6	1.4	1.0	9.3	5.7	1.1	11.0	0.9	16.0	4.1	2.2	3.2	1.2	2.1	6.5	5.3	11.0	1.9	4.7
III.	TC.2A eubacterial	12.76	1.30	0.40	0.42	8.37	9.77	0.32	11.93	0.45	18.37	4.31	1.13	2.63	0.79	0.59	5.21	5.41	11.54	1.70	2.62
	TC.3A.1	13.07	0.97	0.36	0.46	6.96	8.61	0.20	12.29	0.46	19.44	3.67	1.09	2.94	1.08	0.76	5.01	5.45	13.04	1.50	2.64
	TC.2A eukaryotic	9.21	2.79	0.49	0.65	10.10	8.97	0.37	12.97	0.52	15.17	3.42	2.16	2.27	1.26	0.56	6.58	5.22	11.42	2.09	3.79
IV.	Ptm-s (Jones et al.)[5]	1.42	0.89	0.17	0.18	1.64	1.02	0.64	1.75	0.19	1.65	1.34	0.41	0.67	0.41	0.30	0.99	1.03	1.64	1.69	1.01
	Ptm-s (PHDhtm)	1.19	1.06	0.26	0.18	1.96	0.77	0.42	1.62	0.16	1.62	1.65	0.49	0.82	0.35	0.41	1.13	1.04	1.51	1.44	1.46
	Ptm-k (TC.2A eubacterial)	1.72	0.53	0.07	0.08	1.77	1.32	0.12	1.75	0.08	1.86	1.73	0.25	0.68	0.23	0.11	0.91	1.06	1.58	1.29	0.82
	Ptm-k (TC.3A.1)	1.76	0.40	0.07	0.09	1.47	1.16	0.08	1.81	0.08	1.97	1.48	0.24	0.76	0.32	0.15	0.87	1.07	1.79	1.13	0.82
	Ptm-k (TC.2A eukaryotic)	1.24	1.13	0.09	0.12	2.13	1.21	0.14	1.91	0.09	1.53	1.37	0.48	0.59	0.37	0.11	1.15	1.03	1.56	1.58	1.18

†I: Average protein; II: TM-segments; III: TM-kernels; IV: propensity to the membrane; $P_{tm\text{-}s} = f_{tm\text{-}s} / f_{mean}$, $P_{tm\text{-}k} = f_{tm\text{-}k} / f_{mean}$, where $f_{tm\text{-}k}$, $f_{tm\text{-}s}$, f_{mean} are the frequencies of amino acid residues in TM-kernels, TM-segments, and average proteins (BLOSUM62), respectively.

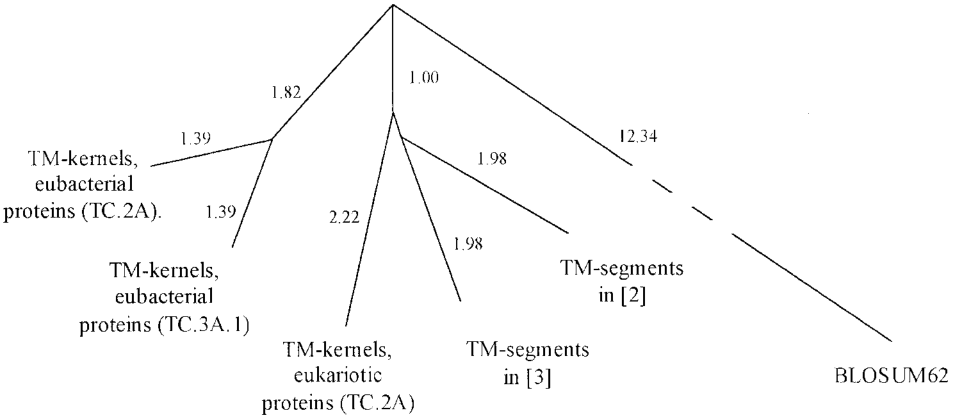
1.82
1.00
12.34
1.39
1.39
1.98
2.22
1.98
TM-kernels, eubacterial proteins (TC.2A).
TM-kernels, eubacterial proteins (TC.3A.1)
TM-kernels, eukariotic proteins (TC.2A)
TM-segments in [3]
TM-segments in [2]
BLOSUM62

Fig. 4. Dendrogram of the amino acid composition of all proteins, TM-segments, and TM-kernels. The tree reflects the matrix of pairwise distances between the vectors of the amino acid frequencies (in %) in the Euclid metric. The UPGMA method was used as implemented in the PHYLIP package.[22]

segments, the fraction of R and K equals twice the fraction of D and E.[5]

The relative content of hydrophobic residues in the TM-kernels and the TM-segments also differs markedly. The TM-kernels contain less W. This agrees with the observation that W is usually located at the ends of transmembrane α-helices.[20] The same, but to a lesser degree, holds for Y. The most important feature of the TM-kernels is the high content of A and G.

The differences between the amino acid content of the TM-kernels and the TM-segments can be illustrated as follows:

$$L > F > I > M > A > V > G > W \gg Y > P \text{ (TM-kernels)}$$

$$I > W > L > F = V > A > M \gg G > Y \gg P \text{ (TM-segments [5])}$$

$$F > M > L = I > V > Y > W \gg A \gg P > G \text{ (TM-segments [3])} \quad (2)$$

where amino acids are ordered by propensity to the TM-kernels ($P_{tm\text{-}k}$) or the TM-segments ($P_{tm\text{-}s}$) (see Table II).

The same observations hold also for the control set of eubacterial ABC-transporters. The only considerable difference is the decreased frequency of F, M, and the increased frequency of V.

On the other hand, the amino acid composition of the TM-kernels of eukaryotic transporters differs from the amino acid composition of the TM-kernels of bacterial transporters from both samples. As shown in Table II, the eukaryotic TM-kernels contain considerably more C, F, N, W, and Y, and less A and L, than the bacterial ones. Interestingly, the amino acid composition of the transmembrane proteins given in Jones et al.[5] and Ng et al.[3] is close to the amino acid composition of the eukaryotic transporters (see Fig. 4). Thus, it looks reasonable to use different TM-matrices for alignment of bacterial and eukaryotic proteins.

Consistency of the BATMAS Series

The evolution of amino acid substitution probabilities in time can be described by Evolutionary Markov Processes (EMP) [21]. Such processes satisfy a number of realistic assumptions.

Let $X(t)$ be an amino acid observed at time t, and let the probability of observing amino acid (i) be π_i not depending on time t. Then the probability of transition of the amino acid (i) at the time T into the amino acid (j) at the time $T+t$ is $p_{ij}(t)$ depending only on the time difference t. Thus, the transition matrix $p_{ij}(t_1)$ corresponding to the time t_1 has to be taken to the power t_2/t_1 in order to obtain the transition matrix $p_{ij}(t_2)$ corresponding to the time t_2. The total of elements in every row of any $p_{ij}(t)$ equals 1. The substitution matrix $m_{ij}(t)$ can be derived by multiplying every row of $p_{ij}(t)$ by π_i. Vice versa, $p_{ij}(t)$ can be derived by normalizing every row of $m_{ij}(t)$ to 1. This allows one to change the evolutionary time of EMP represented by some given substitution matrix. Here the PAM (Point Accepted Mutations) value was considered as a measure of evolutionary time, so that 1 PAM corresponds to the transition time $t = 1$.

In this evolutionary model, the substitution matrix has the following property. If the evolutionary time tends to infinity, then the matrix tends to the substitution matrix in which the probability of substitution of amino acid (i) by amino acid (j) equals $\pi_i \cdot \pi_j$. Thus a simple way to check the consistency of a substitution matrix is to tend the evolutionary time to infinity and to compare the matrix with the random substitution matrix generated by the observed amino acid frequencies. For every matrix of the BATMAS and BLOSUM series, the difference between the random matrix and the infinity time matrix does not exceed 10^{-8} in each cell, the only exclusion being BATMAS80.

Now consider a series of substitution matrices $m_{ij}^{(k)}$. If these matrices are substitution matrices of a common EMP, each matrix $m_{ij}^{(k)}$ corresponds to some evolutionary time t_k, and $m_{ij}^{(k)} = m_{ij}(t_k)$. Thus, each matrix $m_{ij}^{(k)}$ can be

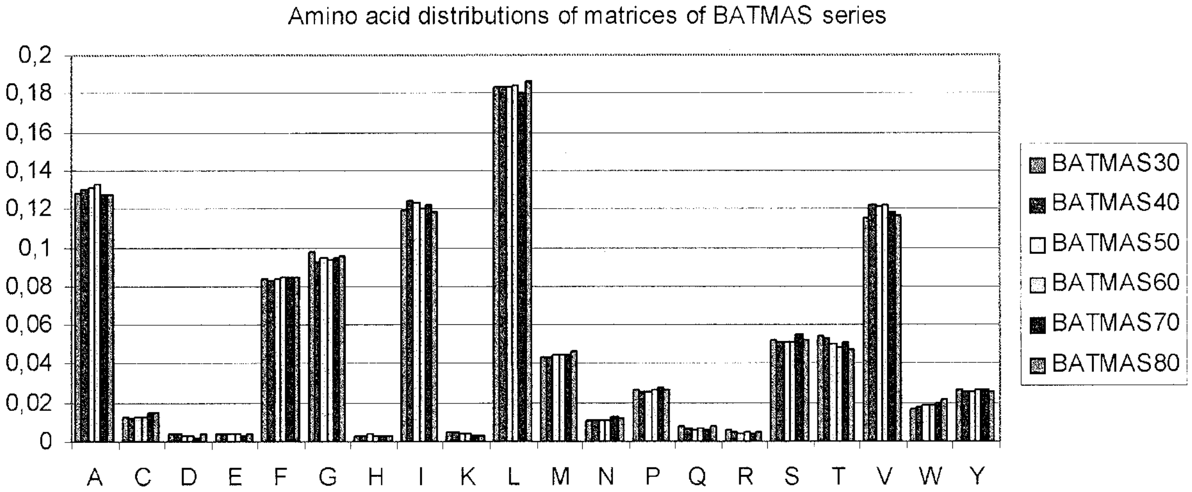

Fig. 5. Amino acid frequencies of matrices in the BATMAS series.

derived from any other matrix $m_{ij}^{(n)}$ in the series by changing the evolutionary time t_n to t_k.

A pair of elements occupying the same cell in two matrices can be described by a point in a plane. If the matrices are equal, these points will lie on the diagonal between points (0,0) and (1,1). We use the standard deviation of points from this line as a measure of matrix similarity. This is done separately for the diagonal and subdiagonal elements. Each pair of matrices was considered.

The described evolutionary model assumes conservation of the amino acid frequencies. As can be seen from Figure 5, the BATMAS matrices satisfy this criterion. Further, Table IIIa presents the results of matching of all pairs of matrices. The error for both the diagonal and subdiagonal elements does not exceed the respective values for the BLOSUM series at analogous transformations of evolutionary time (Table IIIb). Thus, the BATMAS series agrees well with the evolutionary model. The same holds for the matrices constructed for both control sets (data not shown). The relatively worse values observed for BATMAS70 and BATMAS80 can be explained by insufficient data and the lower strength of the comparative analysis at small evolutionary distances.

Comparison of Matrices

To compare different matrices, we calculated the correlation coefficient between subdiagonal elements of normalized matrices using the procedure of Tudos et al.[23] Elements of substitution matrices were normalized by substitution frequency expected from amino acid distribution (that is, the product of frequencies of two amino acids). Matrix RReM proposed in Tudos et al.[23] using a variety of physical and chemical characteristics of amino acid residues, and used in Cserzo et al.[24] for alignment of transmembrane proteins, is not normalized. The values of matrix PAM250 were converted to the antilogarithm. Correlation coefficients are shown in Table IV. Not surprisingly, the two maximum correlations coefficients are between two transmembrane matrices (BATMAS30 and PHDHTM95) and between three "general" evolutionary matrices (PAM250, BLOSUM62, Dayhoff), with matrix RReM being an outlier, but still closer to the general matrices.

Properties of BATMAS30

The frequencies of amino acid substitutions in the TM-kernels of proteins with 30–40% identity (matrix BATMAS30) are shown in Table V. This matrix markedly differs from both the standard BLOSUM62 matrix and the matrix for TM-segments described in Jones et al.[5]

The comparison between BATMAS30 and BLOSUM62 is meaningful, because the traces of these matrices are approximately equal: tr[BATMAS30] = 0.35, tr[BLOSUM62] = 0.33 (the trace is the sum of diagonal elements, tr[BATMAS30] = 0.35, tr[BLOSUM62] =0.33 (the trace is the sum of diagonal elements, $tr(a_{ij}) = \sum_{i=1}^{20} a_{ij}$ and equals the average identity of proteins used to construct the matrix). Table VI illustrates the differences between BATMAS30 and BLOSUM62. The matrix elements were normalized by the average amino acid distribution of the matrices (see Table VI footnote). The main differences of BATMAS30 from BLOSUM62 are an increase in conservation of the charged residues (D, E, K, R, H) and some polar residues (N, Q, P), and a decrease in conservation of hydrophobic residues, namely L, I, F, W, and V. In BATMAS30, W frequently matches to polar residues, namely R, K, and H. The same characteristics are typical for the TM-segments as a whole.[5] Unfortunately, the format of the data in Jones et al.[5] does not allow one to compare the matrices quantitatively. However, even qualitative comparison between our conclusions and the conclusions of Jones et al.[5] demonstrates considerable differences. In particular, according to Jones et al.,[5] L is the most conserved hydrophobic residue. However, the comparison between BATMAS30 and BLOSUM62 results in the opposite conclusion. If we order

TABLE III. Fitting Substitution Matrices to the EMP Model: (a) BATMAS series, (b) BLOSUM Series[†]

a	BATMAS30	BATMAS40	BATMAS50	BATMAS60	BATMAS70	BATMAS80
BATMAS30	—	t = −43 PAM	t = −72 PAM	t = −91 PAM	t = −106 PAM	t = −126 PAM
id: 35.1%	—	d = 0.000928	d = 0.000979	d = 0.001459	d = 0.001376	d = 0.001688
t: 131 PAM	—	s = 0.000165	s = 0.000235	s = 0.000251	s = 0.000265	s = 0.000144
BATMAS40	t = 43 PAM	—	t = −29 PAM	t = −48 PAM	t = −62 PAM	t = −83 PAM
id: 46.9%	d = 0.000789	—	d = 0.000994	d = 0.001517	d = 0.001915	d = 0.002469
t: 88 PAM	s = 0.000174	—	s = 0.000140	s = 0.000159	s = 0.000204	s = 0.000130
BATMAS50	t = 72 PAM	t = 29 PAM	—	t = −19 PAM	t = −34 PAM	t = −54 PAM
id: 58.6%	d = 0.000790	d = 0.000895	—	d = 0.000689	d = 0.001377	d = 0.001960
t: 60 PAM	s = 0.000219	s = 0.000129	—	s = 0.000116	s = 0.000136	s = 0.000101
BATMAS60	t = 91 PAM	t = 48 PAM	t = 19 PAM	—	t = −15 PAM	t = −35 PAM
id: 68.8%	d = 0.001192	d = 0.001396	d = 0.000699	—	d = 0.001438	d = 0.001885
t: 40 PAM	s = 0.000287	s = 0.000171	s = 0.000137	—	s = 0.000134	s = 0.000101
BATMAS70	t = 106 PAM	t = 62 PAM	t = 34 PAM	t = 15 PAM	—	t = −20 PAM
id: 78.2%	d = 0.001308	d = 0.001857	d = 0.001423	d = 0.001408	—	d = 0.001754
t: 26 PAM	s = 0.000310	s = 0.000232	s = 0.000191	s = 0.000168	—	s = 0.000075
BATMAS80	t = 126 PAM	t = 83 PAM	t = 54 PAM	t = 35 PAM	t = 20 PAM	—
id: 94.6%	d = 0.002333	d = 0.003416	d = 0.003048	d = 0.002850	d = 0.002144	—
t: 6 PAM	s = 0.000420	s = 0.000375	s = 0.000347	s = 0.000342	s = 0.000230	—
b	BLOSUM30	BLOSUM40	BLOSUM50	BLOSUM62	BLOSUM80	BLOSUM100
BLOSUM30	—	t = −61 PAM	t = −97 PAM	t = −123 PAM	t = −149 PAM	t = −182 PAM
id: 15.2%	—	d = 0.001273	d = 0.002034	d = 0.002464	d = 0.002915	d = 0.003196
t: 251 PAM	—	s = 0.000334	s = 0.000503	s = 0.000609	s = 0.000650	s = 0.000606
BLOSUM40	t = 61 PAM	—	t = −36 PAM	t = −63 PAM	t = −89 PAM	t = −121 PAM
id: 21.0%	d = 0.000980	—	d = 0.001031	d = 0.001477	d = 0.001897	d = 0.002195
t: 190 PAM	s = 0.000266	—	s = 0.000221	s = 0.000315	s = 0.000343	s = 0.000293
BLOSUM50	t = 97 PAM	t = 36 PAM	—	t = −27 PAM	t = −53 PAM	t = −85 PAM
id: 27.1%	d = 0.001342	d = 0.000884	—	d = 0.000804	d = 0.001545	d = 0.002146
t: 154 PAM	s = 0.000370	s = 0.000211	—	s = 0.000133	s = 0.000167	s = 0.000154
BLOSUM62	t = 123 PAM	t = 63 PAM	t = 27 PAM	—	t = −26 PAM	t = −58 PAM
id: 33.2%	d = 0.001499	d = 0.001120	d = 0.000710	—	d = 0.000997	d = 0.001870
t: 128 PAM	s = 0.000435	s = 0.000299	s = 0.000130	—	s = 0.000076	s = 0.000112
BLOSUM80	t = 149 PAM	t = 89 PAM	t = 53 PAM	t = 26 PAM	—	t = −32 PAM
id: 40.7%	d = 0.001687	d = 0.001319	d = 0.001291	d = 0.000924	—	d = 0.001109
t: 101 PAM	s = 0.000451	s = 0.000328	s = 0.000167	s = 0.000076	—	s = 0.000085
BLOSUM100	t = 182 PAM	t = 121 PAM	t = 85 PAM	t = 58 PAM	t = 32 PAM	—
id: 52.7%	d = 0.001827	d = 0.001476	d = 0.001694	d = 0.001569	d = 0.000992	—
t: 69 PAM	s = 0.000436	s = 0.000302	s = 0.000173	s = 0.000133	s = 0.000095	—

[†] *id* is the identity, *t* is the evolutionary time (in PAM units), *d* and *s* are the standard deviation of the diagonal and off-diagonal matrix elements, respectively.

TABLE IV. Correlation Coefficients Calculated for Different Tables[†]

	Dayhoff	Pam250	RReM	Blosum	Batmas
Pam250	0.72				
RReM	0.47	0.59			
Blosum	0.69	0.80	0.70		
Batmas	0.41	0.50	0.42	0.60	
Phdhtm	0.35	0.47	0.40	0.56	0.85

[†] Antilogarithms were calculated for Pam250; matrices BLOSUM62, BATMAS30, and PHDHTM95 were used.

the amino acids by decrease of the comparative conservation (see footnote to Table VI), it becomes clear that P and Y are more conserved in TM-kernels than in average proteins, but L is less conserved (P>Y>>M>A>W≈V≥I≈L≈F).

Thus, the degree of conservation of L is comparable to that of hydrophobic residues I, V, and F. In addition, in Jones et al.,[5] R and K often mutate in TM-segments (as compared to average proteins), unlike the superconserved N. It is shown in Table VI that N is more conserved in TM-kernels compared to average proteins, but to a lesser degree compared to R and K.

Some characteristics of TM-kernels have not been observed for TM-segments in Jones et al.[5] For example, as compared to the proteins in general, represented by BLOSUM62, W in TM-kernels is often replaced by not only positively charged residues, but also by D, Q, and P. High relative conservation of Y in TM-kernels also was not described in Jones et al.[5] for TM-segments.

A similar matrix (for protein pairs with 30–40% identity) was constructed for the sample of eubacterial ABC-transporters. All the above observations hold for this

TABLE V. BATMAS30, Frequencies of the Amino Acid Substitutions

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	**0.044743**	0.002195	0.000210	0.000289	0.004922	0.014598	0.000185	0.008738	0.000275	0.013223	0.003376	0.000745	0.001828	0.000607	0.000230	0.010893	0.007149	0.011367	0.000899	0.001115
C	0.002195	**0.001411**	0.000015	0.000011	0.000907	0.001007	0.000014	0.001198	0.000012	0.001925	0.000434	0.000147	0.000112	0.000045	0.000029	0.001063	0.000841	0.001328	0.000088	0.000192
D	0.000210	0.000015	**0.002089**	0.000210	0.000057	0.000211	0.000039	0.000081	0.000020	0.000098	0.000029	0.000270	0.000049	0.000047	0.000016	0.000184	0.000147	0.000108	0.000030	0.000047
E	0.000289	0.000011	0.000210	**0.001798**	0.000068	0.000216	0.000050	0.000152	0.000032	0.000255	0.000067	0.000071	0.000119	0.000211	0.000030	0.000164	0.000172	0.000159	0.000039	0.000044
F	0.004922	0.000907	0.000057	0.000068	**0.030824**	0.002367	0.000216	0.007472	0.000128	0.015687	0.003296	0.000336	0.000545	0.000334	0.000187	0.001629	0.002069	0.006648	0.001527	0.004528
G	0.014598	0.001007	0.000211	0.000216	0.002367	**0.053438**	0.000117	0.003239	0.000269	0.005232	0.001236	0.000622	0.000899	0.000269	0.000207	0.005994	0.002685	0.004046	0.000483	0.000543
H	0.000185	0.000014	0.000039	0.000050	0.000216	0.000117	**0.000764**	0.000109	0.000056	0.000273	0.000114	0.000159	0.000048	0.000136	0.000085	0.000164	0.000134	0.000174	0.000100	0.000238
I	0.008738	0.001198	0.000081	0.000152	0.007472	0.003239	0.000109	**0.033132**	0.000166	0.026516	0.005511	0.000393	0.001067	0.000389	0.000172	0.002298	0.004454	0.021668	0.001088	0.001429
K	0.000275	0.000012	0.000020	0.000032	0.000128	0.000269	0.000056	0.000166	**0.001549**	0.000340	0.000137	0.000143	0.000121	0.000176	0.000352	0.000246	0.000139	0.000201	0.000096	0.000087
L	0.013223	0.001925	0.000098	0.000255	0.015687	0.005232	0.000273	0.026516	0.000340	**0.071392**	0.009900	0.000689	0.001368	0.000929	0.000402	0.003503	0.005591	0.021754	0.002012	0.002575
M	0.003376	0.000434	0.000029	0.000067	0.003296	0.001236	0.000114	0.005511	0.000137	0.009900	**0.009407**	0.000373	0.000318	0.000324	0.000099	0.001235	0.001965	0.004226	0.000385	0.000662
N	0.000745	0.000147	0.000270	0.000071	0.000336	0.000622	0.000159	0.000393	0.000143	0.000689	0.000373	**0.004191**	0.000130	0.000254	0.000076	0.001100	0.000782	0.000458	0.000099	0.000248
P	0.001828	0.000112	0.000049	0.000119	0.000545	0.000899	0.000048	0.001067	0.000121	0.001368	0.000318	0.000130	**0.016227**	0.000147	0.000075	0.000830	0.000773	0.001286	0.000161	0.000197
Q	0.000607	0.000045	0.000047	0.000211	0.000334	0.000269	0.000136	0.000389	0.000176	0.000929	0.000324	0.000254	0.000147	**0.002424**	0.000134	0.000410	0.000404	0.000394	0.000115	0.000177
R	0.000230	0.000029	0.000016	0.000030	0.000187	0.000207	0.000085	0.000172	0.000352	0.000402	0.000099	0.000076	0.000075	0.000134	**0.002965**	0.000187	0.000153	0.000223	0.000115	0.000148
S	0.010893	0.001063	0.000184	0.000164	0.001629	0.005994	0.000164	0.002298	0.000246	0.003503	0.001235	0.001100	0.000830	0.000410	0.00018	**0.012604**	0.005302	0.003367	0.000360	0.000520
T	0.007149	0.000841	0.000147	0.000172	0.002069	0.002685	0.000134	0.004454	0.000139	0.005591	0.001965	0.000782	0.000773	0.000404	0.000153	0.005302	**0.014593**	0.005770	0.000346	0.000624
V	0.011367	0.001328	0.000108	0.000159	0.006648	0.004046	0.000174	0.021668	0.000201	0.021754	0.004226	0.000456	0.001286	0.000394	0.000223	0.003367	0.005770	**0.029987**	0.000954	0.001280
W	0.000899	0.000088	0.000030	0.000039	0.001527	0.000483	0.000100	0.001088	0.000096	0.002012	0.000385	0.000099	0.000161	0.000115	0.000115	0.000360	0.000346	0.000954	**0.007007**	0.001135
Y	0.001115	0.000192	0.000047	0.000044	0.004528	0.000543	0.000238	0.001429	0.000087	0.002575	0.000662	0.000248	0.000197	0.000177	0.000148	0.000520	0.000624	0.001280	0.001135	**0.010414**

Boldface: diagonal elements.

matrix as well (data not shown). However, this matrix is less diagonal than the main one (the trace equals 0.29 compared to 0.35 for the main BATMAS30 matrix). This might be caused by variations in the evolutionary rate in TM-segments and loops in these two groups of proteins.

Functional Similarity of Amino Acids in TM-Kernels

Construction of dendrograms as described above is a convenient way to analyze similarity between amino acids from the evolutionary point of view. Two amino acids (*i*) and (*j*) are considered to be similar if the frequency of the substitution (*i*–*j*) exceeds the frequency of random matching of this pair expected given the amino acid frequencies.

The dendrograms were constructed for BATMAS30 (TM-kernels) and BLOSUM62 (all proteins) matrices (Fig. 6). One can see that the topologies of these dendrograms are different. It can be readily seen that in all proteins H clusters with aromatic amino acids, whereas in TM-kernels, it is closer to the group of positively charged or polar amino acids. Negatively charged D and E form one group in the TM-kernel dendrogram in contrast to the BLOSUM62 matrix where D is in the "aspartic" group with N and E is in the "glutamic" group with Q. C clusters with hydrophobic residues in BLOSUM62, but it is closer to small residues in BATMAS30. The dendrogram of BATMAS30 differs also from the dendrogram constructed in Jones et al.[5] where, in particular, C is closer to F, and F does not cluster with two other aromatic residues, Y and W.

CONCLUSIONS

The main difficulty in generating amino acid substitutions for transmembrane segments is the scarcity of experimental data. We have overcome it by application of comparative genomic analysis and several consistency checks. The derived BATMAS series of matrices, constructed by analysis of bacterial secondary transporters, is evolutionarily consistent and thus correspond to sequential snapshots of the evolutionary process.

The BATMAS matrices are similar to matrices constructed for transmembrane components of bacterial ABC transporters, but differ from eukaryotic transporter matrices. This could be explained by the fact that the fine structure of the outer membrane of eukaryotes and prokaryotes is different. In particular, the bacterial membranes never contain polyunsaturated fatty acids and do not normally contain sterols.

Even more drastic are the differences between the transmembrane matrices, including the published ones, and the general evolutionary series, BLOSUM and PAM. This is even less surprising, as the cytozolic environment of most proteins and the lipid membrane environment of transporters are very different. However, the RReM matrix, constructed based on physical and chemical properties, and used for alignment of transmembrane proteins, differs from both general and transmembrane matrices.

The BATMAS series of matrices have been constructed using a sample of bacterial transport proteins, Although

TABLE VI. Comparison Between the Normalized Matrices BATMAS30 and BLOSUM62[†]

	A	C	D	E	F	G	H	I	K	L	M	M	P	Q	R	S	T	V	W	Y
A	$\boxed{\mathbf{0.703}}$	1.513	0.751	0.733	1.010	1.110	0.805	0.904	0.619	0.940	0.873	0.899	0.711	0.797	0.510	1.107	1.053	0.819	1.013	0.611
C	1.513	$\boxed{\mathbf{0.427}}$	0.945	0.695	1.945	1.813	1.058	1.177	0.581	1.229	1.189	2.752	0.780	1.222	1.212	2.222	1.663	1.138	1.296	1.489
D	0.751	0.945	$\boxed{\mathbf{18.026}}$	7.614	0.550	0.868	4.380	0.523	1.405	0.475	0.450	3.908	0.815	1.709	1.153	0.982	0.986	0.712	1.550	1.313
E	0.733	0.695	7.614	$\boxed{\mathbf{19.078}}$	0.559	1.127	3.844	0.942	1.315	0.896	0.718	1.664	1.638	3.388	1.271	0.787	1.056	0.771	1.332	0.787
F	1.010	1.945	0.550	0.559	$\boxed{\mathbf{0.538}}$	0.846	1.261	0.802	1.021	0.884	0.897	0.936	0.907	1.620	1.028	0.845	0.914	0.914	0.837	0.747
G	1.110	1.813	0.868	1.127	0.846	$\boxed{\mathbf{0.815}}$	0.733	1.000	1.041	1.018	0.775	0.643	0.718	0.627	0.811	1.319	0.869	1.079	0.711	0.632
H	0.805	1.058	4.380	3.844	1.261	0.733	$\boxed{\mathbf{5.605}}$	0.857	4.902	1.216	1.362	3.695	1.173	4.806	5.155	1.355	1.487	1.516	3.202	1.604
I	0.904	1.177	0.523	0.942	0.802	1.000	0.857	$\boxed{\mathbf{0.584}}$	0.753	0.713	0.726	0.882	0.895	1.058	0.717	0.848	0.881	0.650	1.201	0.713
K	0.619	0.581	1.405	1.315	1.021	1.041	4.902	0.753	$\boxed{\mathbf{15.655}}$	0.934	1.120	2.994	1.422	3.104	6.348	1.117	0.724	0.652	3.152	1.353
L	0.940	1.229	0.475	0.896	0.884	1.018	1.216	0.713	0.934	$\boxed{\mathbf{0.558}}$	0.629	1.046	0.775	1.342	0.792	0.865	0.855	0.779	1.199	0.772
M	0.873	1.189	0.450	0.718	0.897	0.775	1.362	0.726	1.120	0.629	$\boxed{\mathbf{0.786}}$	1.701	0.676	1.150	0.627	0.874	1.066	0.672	0.862	0.782
N	0.899	2.752	3.908	1.664	0.936	0.643	3.695	0.882	2.994	1.046	1.701	$\boxed{\mathbf{4.630}}$	0.839	2.871	1.311	1.543	1.316	0.947	1.517	1.716
P	0.711	0.780	0.815	1.638	0.907	0.718	1.173	0.895	1.422	0.775	0.676	0.839	$\boxed{\mathbf{1.842}}$	1.165	0.968	0.792	0.753	0.997	1.834	0.711
Q	0.797	1.222	1.709	3.388	1.620	0.627	4.806	1.058	3.104	1.342	1.150	2.871	1.165	$\boxed{\mathbf{6.117}}$	2.018	1.020	1.162	0.890	1.910	1.330
R	0.510	1.212	1.153	1.271	1.028	0.811	5.155	0.717	6.348	0.792	0.627	1.311	0.968	2.018	$\boxed{\mathbf{12.831}}$	0.787	0.698	0.774	2.604	1.772
S	1.107	2.222	0.982	0.787	0.845	1.319	1.355	0.846	1.117	0.865	0.874	1.543	0.792	1.020	0.787	$\boxed{\mathbf{1.214}}$	1.165	0.877	1.025	0.703
T	1.053	1.663	0.986	1.056	0.914	0.869	1.487	0.881	0.724	0.855	1.066	1.316	0.763	1.162	0.698	1.165	$\boxed{\mathbf{1.027}}$	0.950	0.839	0.797
V	0.819	1.138	0.712	0.771	0.914	1.079	1.516	0.650	0.852	0.779	0.672	0.947	0.997	0.890	0.774	0.977	0.950	$\boxed{\mathbf{0.611}}$	1.168	0.661
W	1.013	1.296	1.550	1.332	0.837	0.711	3.202	1.201	3.152	1.199	0.862	1.517	1.834	1.910	2.604	1.025	0.839	1.168	$\boxed{\mathbf{0.648}}$	1.198
Y	0.611	1.489	1.313	0.787	0.747	0.632	1.604	0.713	1.353	0.772	0.782	1.716	0.711	1.330	1.772	0.703	0.797	0.661	1.198	$\boxed{\mathbf{1.534}}$

[†] Each element of the table is defined as $c_{ij} = a_{ij} / b_{ij}$ where a_{ij} and b_{ij} are the normalized elements of the substitution matrices BATMAS30 and BLOSUM62, respectively. The latter values are defined by $a_{ij} = f_{ij} / (p_i p_j)$, where f_{ij} is the frequency of matching pairs (i,j), p_i is the frequency of the amino acid (i). a_{ii} is the conservation of the amino acid (i) for the matrix (a_{ij}), which is the diagonal element. The values b_{ij} are defined similarly. c_{ij} measures the similarity of the substitution patterns of two matrices, BATMAS30 and BLOSUM62. Boxed: diagonal elements.

we have performed extensive testing of consistency of these series and compared them to other published matrices, we have not analyzed other classes of bacterial transmembrane proteins (outer transporters, respiration chain proteins, receptors, etc.). Although this has been caused mainly by technical problems (lack of well-curated data), we believe that the results of this study are not only interesting from the theoretical point of view, as they tell something new about evolution of transport systems, but also they can be used in practice, in particular, to align the transmembrane segments of bacterial transporters, to orient them rotationally relative to the membrane, and to study their specific properties, e.g., the functionality of residues involved in recognition of transported compounds.

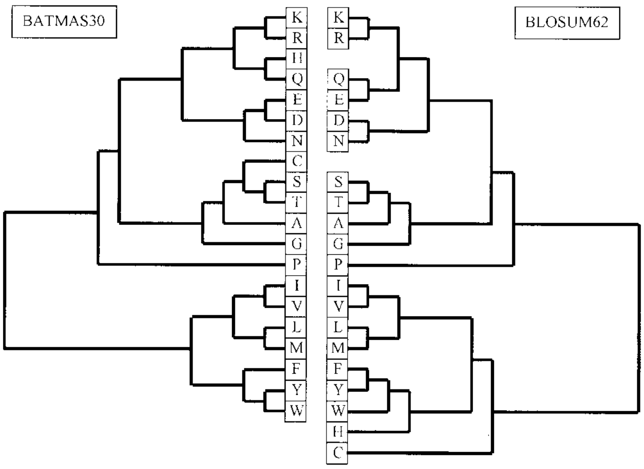

Fig. 6. Amino acid similarity dendrograms for matrices BATMAS30 and BLOSUM62.

ACKNOWLEDGMENTS

We thank Andrey A. Mironov, Vsevolod Yu. Makeev, Shamil Sunyaev, and Leonid Mirny for discussion and critique, Maria M. Bezruchenkova for assistance with the data set, and Maria Chkanikova for translation.

REFERENCES

1. Henikoff S, Henikoff J. Amino acid substitution matrices from protein blocks. PNAS 1992;89:10915–10919.
2. Dayhoff M, Schwartz R, Orcutt B. A model of evolutionary change in protein. Atlas of Protein Sequences and Structure 1978;5:345–352.
3. Ng PC, Henikoff JG, Henikoff S. PHAT: A transmembrane-specific substitution matrix. Bioinformatics 2000;16:760–766.
4. Muller T, Rahmann S, Rehmsmeier M. Non-symmetric score matrices and the detection of homologous transmembrane proteins. Bioinformatics. 2001;1:182–189
5. Jones DT, Taylor WR, Thornton JM. A mutation data matrix for transmembrane proteins. FEBS Letters 1994;339:269–275.
6. Okada T, Palczewski K. Crystal structure of rhodopsin: implications for vision and beyond. Curr Opin Struct Biol 2001;11:420–426.
7. Royant A, Nollert P, Edman K, Neutze R, Landau EM, Pebay-Peyroula E, Navarro J. X-ray structure of sensory rhodopsin II at 2.1-A resolution. Proc Natl Acad Sci USA 2001;98:10131–10136.
8. Parodi LA, Granatir CA, Maggiora GM. A consensus procedure for predicting the location of alpha-helical transmembrane segments in proteins. Comput Appl Biosci 1994;10:527–535.
9. Nilsson J, Persson B, von Heijne G. Consensus prediction of membrane protein topology. FEBS Lett 2000;486:267–269
10. Saier MH Jr. A functional-phylogenetic system for the classification of transport proteins. Cell Biochem 1999;Suppl 32–32, 84–94.
11. Saier MH Jr. A functional-phylogenetic classification system for transmembrane solute transporters. Microbiol Mol Biol Rev 2000; 64:354–411.
12. Paulsen IT, Sliwinski MK, Saier MH Jr. Microbial genome analyses: global comparisons of transport capabilities based on phylogenies, bioenergetics and substrate specificities. Mol Biol 1998;277:573–592.
13. Altschul S, Gish W, Miller W, Myers E, Lipman D. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
14. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr,

Kyrpides N, Fonstein M, Maltsev N, Selkov E. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 2000;28:123–125.

15. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.
16. Sonnhammer EL, von Heijne G, Krogh A. A hidden Markov model for predicting transmembrane helices in protein sequences. Proc Int Conf Intell Syst Mol Biol 1998;6:175–182.
17. Cserzo M, Wallin E, Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng 1997;10: 673–676.
18. Persson B, Argos P. Topology prediction of membrane proteins. Protein Sci 1996;5:363–371.
19. Klein P, Kanehisa M, DeLisi C. The detection and classification of membrane-spanning proteins. Biochim Biophys Acta 1985;815: 468–476.
20. Arkin IT, Brunger AT. Statistical analysis of predicted transmembrane alpha-helices. Biochim Biophys Acta 1998;1429:113–128
21. Muller T, Spang R, Vingron M. Estimating amino acid substitution models: a comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. Mol Biol Evol 2002;19:8–13.
22. Felsenstein J. Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods. Methods Enzymol 1996;266:418–427.
23. Tudos E, Cserzo M, Simon I. Predicting isomorphic residue replacements for protein design. Int J Pept Protein Res 1990;36: 236–239.
24. Cserzo M, Bernassau JM, Simon I, Maigret B. New alignment strategy for transmembrane proteins. J Mol Biol 1994;243:388–396.