BIOINFORMATICS

Vol. 16 no. 9 2000
Pages 760–766

PHAT: a transmembrane-specific substitution matrix

Pauline C. Ng[1], Jorja G. Henikoff[2] and Steven Henikoff[2,*]

[1]Department of Bioengineering, University of Washington, Seattle, WA 98195, USA and [2]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N, Seattle, WA 98109-1024, USA

Received on February 23, 2000; revised on April 14, 2000; accepted on April 18, 2000

Abstract

Motivation: *Database searching algorithms for proteins use scoring matrices based on average protein properties, and thus are dominated by globular proteins. However, since transmembrane regions of a protein are in a distinctly different environment than globular proteins, one would expect generalized substitution matrices to be inappropriate for transmembrane regions.*

Results: *We present the PHAT (predicted hydrophobic and transmembrane) matrix, which significantly outperforms generalized matrices and a previously published transmembrane matrix in searches with transmembrane queries. We conclude that a better matrix can be constructed by using background frequencies characteristic of the twilight zone, where low-scoring true positives have scores indistinguishable from high-scoring false positives, rather than the amino acid frequencies of the database. The PHAT matrix may help improve the accuracy of sequence alignments and evolutionary trees of membrane proteins.*

Availability: *http://www.blocks.fhcrc.org/∼pauline*

Contact: *steveh@muller.fhcrc.org*

Introduction

Given a protein sequence, database searching for homologues can be used to infer the protein's function and structure. Database searching and other alignment algorithms for proteins use an amino acid substitution matrix to score protein alignments. The matrix contains log-likelihood scores that reflect how likely one amino acid is substituted for another; a positive score indicates that the substitution is favored over a chance event, a negative score indicates the substitution is less likely to occur than predicted by chance alone. Typically, amino acids with similar physiochemical properties have positive scores while amino acids that are unlike each other have negative scores. A substitution score for amino acid i to j can be calculated from alignment data by: $s_{ij} = \lambda^{-1} \ln(q_{ij}/(p_i p_j))$ where λ is a scaling factor, q_{ij}'s are target or observed frequencies of amino acid pairs taken from alignments and p_i's are the background frequencies (Altschul, 1991). The widespread use of database searching and other protein alignment tools in modern biology underscores the importance of using substitution matrices that most accurately resemble biological reality.

The point accepted mutation (PAM) and blocks substitution matrices (BLOSUM) are the two most popular matrix series (Dayhoff, 1978; Henikoff and Henikoff, 1992). The PAM matrix is computed by counting mutations between closely related sequences and an inferred common ancestral sequence to obtain PAM 1 target frequencies. The PAM 1 scores are extrapolated by matrix multiplication to get a matrix series corresponding to evolutionary distance. The PAM 1 matrix estimates scores for accepting one mutation per 100 positions; the PAM 170 is constructed by multiplying PAM 1 by itself 170 times and estimates 170 accepted point mutations per 100 positions. In 1992, Jones and colleagues modified the collection of PAM 1 counts (Jones *et al.*, 1992). They automated the process by counting substitutions between pairs of highly similar sequences from a protein sequence database to obtain target frequencies (q_{ij}'s). Background frequencies (p_i's) were taken from the database and the JTT PAM series was computed.

Whereas the PAM model is based on closely related sequences, the BLOSUM model is based on blocks, which are multiple alignments of distantly related but conserved regions (Henikoff and Henikoff, 1991). The target frequencies (q_{ij}'s) are calculated by counting the substitutions observed in blocks. Instead of using background frequencies from a protein database, the p_i's are calculated as the marginal frequencies: $p_i = \Sigma_j q_{ij}$. Closely related sequences are downweighted by clustering based on the percentage of identical residues. The BLOSUM series is constructed by varying this cluster percentage. For example, BLOSUM 62 is derived from counts between clusters of sequence segments that are less than 62% identical.

---

[*] To whom correspondence should be addressed.

Matrices based on different models that have similar relative entropies can be compared to each other (Altschul, 1991). The relative entropy (H) of a matrix is defined as the average information per aligned residue pair and is calculated by: $H = \Sigma_i \Sigma_j q_{ij} s_{ij}$. By comparing BLOSUM matrices to other matrices with similar relative entropy, it was shown that the BLOSUM series performed as well as or better than other published matrices (Pearson, 1995; Henikoff and Henikoff, 1993).

For transmembrane proteins, the hydrophobic environment for amino acids located in the lipid bilayer is very different from the aqueous cytosolic and extracellular compartments. Thus a matrix specialized for transmembrane regions should work better than matrices which have been generalized for all proteins.

In this paper, we describe a matrix built from **p**redicted **h**ydrophobic **a**nd **t**ransmembrane (PHAT) regions of the Blocks database (Henikoff *et al.*, 1999). We demonstrate that in searches on queries consisting of putative transmembrane regions, the PHAT matrix significantly outperforms both the generalized matrices, BLOSUM and JTT PAM, as well as the JTT transmembrane matrix (Jones *et al.*, 1994). When nontransmembrane regions were included in the query, the PHAT matrix performed better than most of the generalized matrices. We attribute the success of the PHAT matrix to using background frequencies characteristic of the twilight zone, rather than background frequencies of the entire database. The PHAT matrix may be useful for structural alignments and phylogenetic trees of membrane proteins.

Methods

Matrix construction

In order to obtain target (q_{ij}) and background frequencies (p_i) for a transmembrane matrix series, membrane prediction methods were applied to the Blocks + database (Henikoff *et al.*, 1999). Block families were submitted to PHDhtm, a prediction program that achieves 86% transmembrane prediction accuracy per residue (Rost *et al.*, 1996). PHDhtm can accept multiple alignments but needs the context of neighboring residues to predict transmembrane topology. Therefore blocks for a given family were linked with the sequence of the intervening segments from a representative sequence of that family (Henikoff and Henikoff, 1997) to reconstruct a multiple alignment that could be submitted to PHDhtm. Out of 2935 families, 598 were predicted to contain transmembrane segments (20.3%); out of 8909 blocks, 844 had transmembrane segments (9%). Blocks lacking transmembrane regions were discarded, and nontransmembrane regions were removed from blocks containing putative transmembrane regions. Blocks containing more than one predicted transmembrane region were split. The resulting blocks containing only predicted transmembrane regions were clustered, and matrices were constructed by the BLOSUM method (Henikoff and Henikoff, 1992). The resulting matrix series was termed PHDhtm.

A second matrix series was built from hydrophobic blocks. Persson and Argos obtained propensity values of amino acids for transmembrane regions (Persson and Argos, 1994). When eight or more consecutive amino acids have an average transmembrane propensity value exceeding 1.23, the sequence is predicted to be the core of a transmembrane segment. These values were used to predict blocks that were entirely hydrophobic. The average transmembrane propensity value of an entire block was calculated. If the block's value exceeded 1.23, the entire block was used to construct a hydrophobic matrix. 514 of the 8909 blocks passed this criterion, 383 of which were also identified as transmembrane by PHDhtm. These hydrophobic blocks were clustered by percentage identity to construct the Persson–Argos matrix series.

Using the target frequencies from the PHDhtm matrix and background frequencies from the Persson–Argos matrix with corresponding relative entropy, scores for a third transmembrane matrix were calculated by $s_{ij}^{\mathrm{PHAT}} = \lambda^{-1} \ln(q_{ij}^{\mathrm{PHDhtm}}/(p_i^{P-A} p_j^{P-A}))$. We termed this matrix series PHAT because it was built from **p**redicted **h**ydrophobic **a**nd **t**ransmembrane regions of blocks.

Tested matrices

Matrices of similar relative entropies were compared, we compared BLOSUM 55 ($H = 0.5637$), JTT PAM 170 (Jones *et al.*, 1992) ($H = 0.5655$), and the JTT transmembrane matrix 170 ($H = 0.5655$) with our own matrices PHDhtm 80 ($H = 0.5550$) and Persson–Argos 80 ($H = 0.5725$). The PHAT 75/73 matrix was constructed from PHDhtm 75 ($H = 0.5007$) target values and Persson–Argos 73 ($H = 0.5038$) background frequencies. PHAT 75/73 has $H = 0.5605$, a relative entropy similar to the other matrices, so can be used for comparison. Although BLOSUM 62 has a higher relative entropy ($H = 0.6979$), we also used it as a test matrix since it is the default matrix in BLASTP searches.

Test set

To test the matrices, 100 sequences from 74 different Prosite (release 14.0) (Hofmann *et al.*, 1999) protein families documented as transmembrane were used as queries. Sequences closest to the consensus sequence for the blocks (Henikoff and Henikoff, 1997) corresponding to these Prosite families resulted in 74 of the 100 queries. Twenty six of the 74 Prosite families listed false negative sequences. From each of these, one false negative sequence was randomly chosen to be used as a query to give a total of 100 queries. The percentage of transmembrane residues in a query protein ranged from 2 to 82%; the number of predicted segments from 1 to 14.

To restrict the database search to transmembrane segments, queries were filtered by HMMTOP, a transmembrane prediction method based on a hidden Markov model (Tusnady and Simon, 1998). Nontransmembrane segments were replaced by the character ‘X’. X-ed out sequences were subjected to the tests described below.

Database searching

Ungapped BLASTP (v. 1.4.7) searches (Altschul *et al.*, 1990) were executed on the X-ed out sequences containing only putative transmembrane regions to search against the SWISS-PROT database (release number 36). Reported sequences ($E < 1$) were compared with the sequences listed in the same Prosite family.

To carry out searches on the entire sequence of transmembrane protein, we employed a bipartite scheme, as introduced by Jones *et al.* (1994). In their bipartite scheme, a generalized matrix is used on the nontransmembrane regions and a test matrix (either transmembrane or generalized) on the transmembrane regions. SWAT (http://bozeman.mbt.washington.edu), a Smith–Waterman alignment tool that can accept profiles as input, was used for testing the entire sequence. Nontransmembrane regions, as predicted by HMMTOP, were given BLOSUM 62 scores and putative transmembrane regions were given values from the test matrix. The resulting profile was searched against the SWISS-PROT database (release number 36).

Matrix evaluation

In order to assess the performance of a matrix, the equivalence number was calculated for each search (Pearson, 1995). The equivalence number is the rank at which the number of false positives equals false negatives. A lower equivalence number indicates better performance. To test whether two matrices perform differently we used the sign rank test as described by Pearson (1995). Z-scores and corresponding p-values for this test are reported.

Results

Matrices based solely on transmembrane regions

Jones *et al.* (1994) built a matrix for transmembrane proteins using the PAM model in order to investigate the evolutionary constraints imposed by the lipid environment. Counts were taken from 3155 pairwise alignments of documented transmembrane segments. The matrix showed hydrophobic residues to be variable and polar amino acids to be highly conserved. Jones *et al.* proposed using a bipartite scheme for database searching. Using a general matrix for nontransmembrane regions and their transmembrane matrix for the transmembrane regions, the transmembrane protein bacteriorhodopsin was searched against membrane proteins extracted from the SWISS-PROT database (Bairoch and Apweiler, 1999). Higher z-scores were observed for two rhodopsin sequences with the bipartite scheme using the JTT transmembrane matrix compared to using only the generalized matrix.

```
    C   S   T   P   A   G   N   D   E   Q   H   R   K   M   I   L   V   F   Y   W
    0   0   0  -2   2  -1   0   0   0   2  -2  -2  -1   2  -1   1   0  -1  -4  -3 C
        1   0  -2   0   0  -1   0   0   2   2  -1   3   1  -1   0   0   0  -2   1 S
C   9       0  -2  -1   0  -1  -2  -1   2   1  -1   3   0  -1   1   0   2   1   0 T
S   2   6       0  -3   0   0  -1   2  -2   1  -1   7  -1  -1  -4   0   0   1   4 P
T   0   2   5       0   0   0  -3  -2   1   2  -3   0   1   0   2   0   2   1   2 A
P  -8  -3  -3  13       0   3  -3  -5   2   2  -3   3   3   1   4  -1   4   3  -1 G
A   1   2   1  -3   4      -2  -4   1   0   2  -2  -3   0   0   2   1   4   4   1 N
G  -2   1  -1  -3   1   8       0   0   2  -1  -3  -1  -1   0   3   0   4   0   2 D
N  -2   1   0  -3  -2  -1  11       0  -4   1  -4   2   1   1   3   0   5   5   1 E
D  -5  -2  -3  -4  -4  -1   3  15      -2  -4  -4  -1   2   2   1   3   4   2   5 Q
E  -5  -1  -3  -3  -3  -2   1   9  15       0  -6  -2   1   2   3   2   4  -2   2 H
Q  -3   0  -1  -2  -2  -1   3   3   4  11       4  -4  -4  -1   0  -1   1  -2  -9 R
H  -4  -1  -2  -5  -2  -3   5   2   2   4  13       0  -1   3   3   3   5  -1  -4 K
R  -4  -3  -3  -5  -4  -3  -1  -3  -2   2   0  14       0  -1   0  -1   0   0   0 M
K  -5   0   0   1  -2   0   2   1   2   4   1   6  14       0   1   0   0   1   0 I
M  -1  -2   0  -5  -1  -2  -3  -5  -4   0  -3  -4  -2   5      -1   0  -1   1   0 L
I  -3  -3  -1  -5  -1  -3  -4  -5  -4  -3  -4  -5  -3   1   3       0   1   1   0 V
L  -1  -3  -1  -5  -1  -3  -3  -4  -4  -2  -3  -4  -3   1   1   3      -2   0   6 F
V  -1  -2   0  -4   0  -3  -4  -4  -3  -2  -4  -5  -3   0   2   0   3      -2   5 Y
F   0  -2  -1  -6  -2  -3  -2  -4  -4  -2  -1  -5  -3  -1  -1   0  -1   5      -1 W
Y  -1  -3  -3  -6  -4  -4   1  -3  -2   0   4  -4  -1  -4  -4  -3  -4   2   9
W  -2  -4  -5  -4  -3  -4  -3  -5  -4   3   0  -3  -2  -3  -4  -2  -3   1   1  13
    C   S   T   P   A   G   N   D   E   Q   H   R   K   M   I   L   V   F   Y   W
```

Fig. 1. The lower half of the matrix is the PHDhtm 80 matrix ($H = 0.5550$). The upper half of the matrix gives the difference between PHDhtm 80 and JTT transmembrane 170 ($H = 0.5599$).

Since the search by Jones *et al.* was limited to a transmembrane protein database rather than the entire database, we tested the JTT transmembrane matrix against the entire SWISS-PROT database, a more realistic situation. Using a large data set of transmembrane sequences, we found that the JTT transmembrane matrix performed poorly compared with the generalized matrices when searching against the entire SWISS-PROT database (Table 1).

The PHDhtm matrix, like the JTT transmembrane matrix, was built from transmembrane segments. Figure 1 shows the differences between the JTT transmembrane and PHDhtm matrices. The PHDhtm matrix performs similarly to the JTT transmembrane matrix for database searching. Both these matrices perform worse than the generalized matrices (Table 1).

Effect of background frequencies

Surprisingly, the Persson–Argos matrix based on hydrophobic blocks performed similarly to BLOSUM 55 and better than the PHDhtm matrix, which was built from predicted transmembrane regions (Table 1). A comparison of the Persson–Argos matrix with the PHDhtm matrix show that while many of the scores are similar, the largest differences are observed for the charged amino acids (K, R, H, D, E) (Figure 2). We noticed that the PHDhtm background frequencies for the charged amino acids were

Table 1. Performance results for ungapped BLASTP (v. 1.4.7) searches of predicted transmembrane regions of 100 protein sequences against the SWISS-PROT36 database. All matrices tested had similar relative entropies. Z-scores and p-values are calculated as described by Pearson (1995). See Methods for details

Test matrix	No. of queries BLOSUM 55 better	No. of queries test matrix better	No. of queries for which matrices performed the same	z-score	p-value
JTT Transmembrane 170	33	10	57	3.51	0.0002
PHDhtm 80	29	12	59	2.65	0.0040
Persson–Argos 80	16	23	61	−1.12	0.63

```
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
   2 -2 -1 -2 -1 -1 -1    0 -1  0 -1  0 -1 -1 -1  0 -2  0  1 C
      0 -1  1  0  0  1  1  1  0  0  2  0  0  0  0  0  0  0  0 S
C 11     1  0  0  0  0  0  1  0  0  0 -1  0  0  0  0  0  0  0 T
S  0  6    -1  3  2  1  2  1  1  2  2 -4  2  0  0  1  1  1  1 P
T -1  1  6     0  0  1  1  1  0  0  1  0  0  0  0  1  0  0  0 A
P-10 -2 -3 12     0  0  0  0  0  1  1 -3  0  1  0  1  1  1  0 G
A  0  2  1 -1  4    -2  0  1 -1  0  1 -1  0  0  1  1  1  0  1 N
G -3  1 -1 -2  1  8    -2 -1  0  1  4  1  0  0  0  0  0  0  2 D
N -2  2  0 -3 -1 -1  9    -3  0 -1  4  2  0  0  0  0  0 -1  1 E
D -3 -2 -3 -2 -4 -1  3 13    -2 -1  1  0 -1  0  0 -1  0 -1 -2 Q
E -5  0 -2 -2 -3 -2  2  8 13    -3  3  3  0  0  1  0 -1  0 -2 H
Q -4  0 -1 -1 -2 -1  2  3  3 10    -3  1 -1  0 -1  0 -1  1  0 R
H -5 -1 -2 -3 -2 -3  5  2  1  3 11    -5 -2 -2 -2  0 -2 -2  0 K
R -6 -2 -3 -3 -3 -3  0  0  2  3  2 11     1  1  1  0  1  1  0 M
K -5 -1 -2 -3 -3 -3  2  2  4  5  3  7 10     0  0  0  1  0  0 I
M -2 -2  0 -4 -1 -2 -3 -5 -4 -1 -4 -5 -4  6     0  1  0  0  0 L
I -4 -3 -1 -5 -1 -3 -4 -5 -5 -3 -5 -5 -6  2  4     0  0  0  0 V
L -2 -3 -1 -5 -1 -3 -2 -5 -4 -2 -3 -5 -5  2  1  3     1  0  0 F
V -1 -2  0 -4  1 -2 -3 -5 -4 -3 -4 -5 -4  1  2  1  3     1 -1 Y
F -2 -2 -2 -5 -2 -2 -1 -4 -4 -2 -1 -6 -5  0 -1  0 -2  6    -1 W
Y -2 -3 -3 -5 -4 -3  0 -3 -3 -1  3 -4 -3 -3 -4 -3 -4  3 10
W -2 -5 -6 -4 -4 -5 -3 -4 -3  1 -3 -4 -2 -3 -4 -3 -4  1  0 13
   C  S  T  P  A  G  N  D  E  Q  H  R  K  M  I  L  V  F  Y  W
```

Fig. 2. The lower half of the matrix is Persson–Argos 80 ($H = 0.5725$). The upper half of the matrix gives the differences between Persson–Argos 80 and PHDhtm 80 ($H = 0.5725$). Most of the difference values are 0; charged amino acids account for most of the large non-zero difference values (bold).

```
   A  R  N  D  C  Q  E  G  H  I  L  K  M  F  P  S  T  W  Y  V
A  5
R -6  9
N -2 -3 11
D -5 -7  2 12
C  1 -8 -2 -7  7
Q -3 -2  2  0 -5  9
E -5 -6  0  6 -7  1 12
G  1 -5 -1 -2 -2 -2 -3  9
H -3 -4  4 -1 -7  2 -1 -4 11
I  0 -6 -3 -5 -3 -3 -5 -2 -5  5
L -1 -6 -3 -5 -2 -3 -5 -2 -4  2  4
K -7 -1 -2 -5-10 -1 -4 -5 -5 -7 -7  5
M -1 -6 -2 -5 -2 -1 -5 -1 -4  3  2 -6  6
F -1 -7 -1 -5  0 -2 -5 -2 -2  0  1 -7  0  6
P -3 -7 -4 -5 -8 -3 -5 -3 -6 -4 -5 -4 -5 -5 13
S  2 -6  1 -4  1 -1 -3  1 -2 -2 -2 -5 -2 -2 -3  6
T  0 -6 -1 -5 -1 -3 -5 -1 -4 -1 -1 -6  0 -2 -4  1  3
W -4 -7 -5 -7 -4  1 -7 -5 -3 -4 -3 -8 -4  0 -6 -5 -7 11
Y -3 -6  2 -4 -1  0 -2 -3  3 -3 -2 -4 -2  4 -5 -2 -3  1 11
V  1 -7 -3 -5 -2 -3 -5 -2 -5  3  1 -8  1 -1 -4 -2  0 -4 -3  4
```

Fig. 3. The PHAT 75/73 matrix ($H = 0.5605$) constructed from PHDhtm 75 ($H = 0.5007$) target values and Persson–Argos 73 background frequencies ($H = 0.5038$). The PHAT 75/73 matrix was used for evaluating database searching performance.

lower than the Persson–Argos background frequencies, and both PHDhtm and Persson–Argos background frequencies for the charged amino acids were lower than the SWISS-PROT database (Table 2). This suggested to us that the differences in performance were due to differences in background frequencies.

Database searching requires not only identification of related sequences (sensitivity) but elimination of false positives (selectivity). The twilight zone is the region where high-scoring false positives overlap with related sequences. An improvement in database searching implies that there is better separation of false positives and related sequences in the twilight zone. We suspected the twilight zone of a search with a transmembrane query consisted of sequences with hydrophobic patches as well as transmembrane regions. We reasoned the Persson–Argos matrix outperformed the PHDhtm matrix (Table 1) because its background frequencies resembled the twilight zone, and hence, better separation could occur. Then to build a matrix for searching with a transmembrane region, we surmised that the scores should be calculated with target frequencies from transmembrane regions and background frequencies from hydrophobic regions. We built the PHAT matrix series using target frequencies from the PHDhtm matrices and background frequencies from the Persson–Argos matrix with corresponding relative entropy. The PHAT 75/73 matrix (Figure 3) with target frequencies from PHDhtm 75 and background frequencies from Persson–Argos 73 was subjected to tests described below.

Performance in searching

BLASTP database search results on transmembrane regions for the PHAT matrix were compared with the other matrices. The PHAT matrix performs significantly

Table 2. Amino acid composition $(p_i s)$ of the matrices. As expected, the percentage of the hydrophobic residues is higher in the PHDhtm and Persson–Argos matrices compared to BLOSUM and the SWISS-PROT database. Major differences in amino acid composition for the charged amino acids (K, R, H, D, E) are observed between Persson–Argos and PHDhtm

Matrix	Ala	**Arg**	Asn	**Asp**	Cys	Gln	**Glu**	Gly	**His**	lle
Persson–Argos	8.3	3.7	2.3	2.0	3.2	1.5	1.5	5.4	4.5	10
PHDhtm	8.8	2.1	2.2	1.4	2.6	1.2	1.0	5.7	1.1	11
SWISS-PROT	7.6	5.1	4.5	5.3	1.7	4.0	6.4	6.8	2.2	5.8
BLOSUM 62	7.4	5.2	4.5	5.4	2.5	3.4	5.4	7.4	2.6	6.8

Matrix	Leu	**Lys**	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Persson–Argos	14	2.6	3.6	8.1	3.3	6.5	6.3	2.5	3.8	10
PHDhtm	16	.9	4.1	9.3	3.2	6.5	5.3	1.9	4.7	11
SWISS-PROT	9.4	5.9	2.4	4.1	4.9	7.1	5.7	1.2	3.2	6.6
BLOSUM 62	9.9	5.8	2.8	4.7	3.9	5.7	5.1	1.3	3.2	7.3

better than the generalized matrices BLOSUM and JTT PAM, as well as the JTT transmembrane matrix for BLAST searches on queries consisting of transmembrane regions (Table 3). BLOSUM 55 was used in addition to BLOSUM 62 because it has a relative entropy similar to the transmembrane matrix. To show that the better performance of the PHAT matrix is independent of the search algorithm used, SWAT was used with the aforementioned matrices on queries consisting of only transmembrane regions. As expected, similar results were obtained (data not shown).

Blocks containing the test queries were represented in the data set used to build the PHAT matrix. To show that the success of the PHAT matrix was not due to overrepresentation of these blocks, blocks from families containing a test query were removed from the data set and the PHAT matrix reconstructed. Database searching using these reconstructed PHAT matrices gave similar results to the original PHAT matrix (data not shown). We also constructed a matrix from PHDhtm target values and SWISS-PROT database background frequencies. This matrix performed poorly in the BLAST searches (data not shown), supporting the notion that the twilight zone of the search consists of hydrophobic patches and transmembrane regions rather than a sample that is compositionally similar to the entire SWISS-PROT database.

Jones *et al.* (1994) introduced a bipartite scheme for transmembrane proteins by using generalized matrix values on nontransmembrane regions and a transmembrane matrix on the transmembrane. We applied this strategy using the SWAT program. When nontransmembrane regions were included in the search, performance was increased overall for all matrices, thereby decreasing the differences in performance between matrices (Table 4). The PHAT matrix still performed significantly better than BLOSUM 62 and the JTT transmembrane matrix, albeit with lower z-scores. Using the bipartite scheme, the PHAT matrix performed among the best of the matrices.

Protein sequence alignment

The Smith–Waterman alignments given by the generalized matrices were usually identical with that of the bipartite BLOSUM/PHAT matrix scheme. However, we noticed several examples in which the alignment resulting from the bipartite search using the PHAT matrix was more likely to be true than the alignment given by the generalized matrices. One example is in the heme copper oxidase family (PS00077), to which NORB_PSEAE and COX1_DIDMA both belong. It appears the alignment given by the BLOSUM matrices is incorrect since the Prosite patterns do not align (Figure 4). The alignment given by the bipartite scheme using the PHAT matrix has a lower Smith–Waterman z-value and a shorter alignment compared to the BLOSUM alignment. However, it is apparent that the ‘HH’ motif for COX1_DIDMA and NORB_PSEAE are aligned.

Since these data suggest that the PHAT matrix may perform better for aligning transmembrane proteins, we tested the ability of the PHAT matrix to align transmembrane proteins whose structures are known. We tested pairs of transmembrane proteins with known structures: the photosynthetic reaction center L and cytochrome c oxidase. Identical alignments were obtained from the bipartite scheme using BLOSUM 62 on the nontransmembrane regions and either BLOSUM 62, BLOSUM 55

Table 3. PHAT performance results for ungapped BLASTP (v. 1.4.7) searches of the transmembrane regions of 100 protein sequences against SWISS-PROT 36. All matrices tested had similar relative entropies ($H \approx 0.56$) except for BLOSUM 62 ($H = 0.70$)

Test matrix	No. of queries PHAT 75/73 better	No. of queries test matrix better	No. of queries for which matrices performed the same	*z*-score	*p*-value
BLOSUM 62	36	6	58	4.63	< 0.0001
BLOSUM 55	35	5	60	4.74	< 0.0001
JTT PAM 170	39	3	58	5.56	< 0.0001
JTT Transmembrane 170	44	0	56	6.63	< 0.0001

Table 4. SWAT results using the bipartite scheme. BLOSUM 62 values were used on nontransmembrane regions. Values from the test matrix were used for transmembrane regions predicted by HMMTOP. Comparison tests were done as described in Table 1

Test matrix	No. of queries PHAT 75/73 better	No. of queries test matrix better	No. of queries for which matrices performed the same	*z*-score	*p*-value
BLOSUM 62	15	4	81	2.52	0.0059
BLOSUM 55	8	11	81	-0.69	0.7451
JTT PAM 170	15	7	78	1.71	0.0436
JTT Transmembrane 170	19	4	77	3.13	0.0009

or the PHAT 75/73 matrix on the transmembrane regions. From this small data set, we are unable to determine whether using the bipartite scheme with the PHAT matrix gives better alignments overall.

Discussion

The PHAT matrix presented here performs significantly better in database searches than generalized matrices and the JTT transmembrane matrix on queries consisting of transmembrane regions. When nontransmembrane regions were included in the queries and a bipartite scheme employed, the disparity between searching performance using different matrices on transmembrane regions was reduced. An explanation for these results is that transmembrane regions may not contribute significantly to the score. This could be because transmembrane regions are not well conserved in comparison with nontransmembrane regions and/or they represent a small fraction of the region of alignment. Using the PHAT matrix specialized for transmembrane regions improves database searching, and the PHAT matrix may improve alignment of distantly related transmembrane proteins. The PHAT matrix may also be useful for pairwise and multiple alignment applications such as evolutionary trees. However, there is insufficient data to test this because too few homologous transmembrane protein structures are available.

A key step in constructing the PHAT matrix series was to use target frequencies of transmembrane regions and background frequencies of hydrophobic patches. As can be seen in Table 2, the background frequencies between Persson–Argos and PHDhtm are similar except for the charged amino acids. Without replacing the background frequencies, the substitution scores for conservation of charged amino acids are extremely high. This may increase the possibility of a spurious match with a hydrophobic patch that may have a composition similar to the transmembrane segment, and by chance have charged amino acids in the appropriate positions. When matching a transmembrane segment with all other sequences in the database, there is little chance that the sequence will match with sequences that have ‘normal’ amino acid frequencies of a generalized protein because the query itself has an abnormal amino acid content of more hydrophobic amino acids. The twilight zone for searching

```
(a)

NORB_PSEAE|Q59647 NITRIC-OXIDE REDUCTASE SUBUNIT B
Length: 466 Score: 120  z: 7.94  E: 1.54

Subject    91 PKLAWILFWVFAAAGV--LTILGYLLVPYAGLARLTGNELWPTMGREFLE 138
Query     228 PILYQHLF                                        -   276

Subject   139 QPTISKAGIVIVALGFLFNVGMTV-LRGRKTAISMVLMTGLIGLALLFLF 187
Query     277              -  MFTVGLDVDTRAYFTSATMIIAIP-TGVKVFSWL 324

Subject   188 SFYNPENLTRDKFYW                           ----     232
Query     325 ATLHGGNIK------WSPAMLWALGFIFLFTIGGLTGIVLANSSLDIVLH 368

Subject   233                            YFWIGVPGYWL---WLGSVFSAL 279
Query     369 DTYYVVAHFHYVLSMGAVFAIMGGFVHWFPL-FTGYMLNDMWAKIHFFIM 417

Subject   280 ---EPLPFFAMVLFAFNTINRRRRRDYPNRAVALWAMGTTVMAFL 321
Query     418 FVGVNLTFFPQHFLGLSGMPRRYS-DYPD-AYTMWNVVSSIGSFI 460

(b)

NORB_PSEAE|Q59647 NITRIC-OXIDE REDUCTASE SUBUNIT B
Length: 466 Score: 81  z: 5.49  E: 35.8

Subject   238               -     ---   YFWIGVPGYWLWLGSVFSALEPLP 283
Query     266                           MFTVGLDVDTRAYFTSATMIIAIP 315

Subject   284 FFAMVLFAFNTINRRRRRDYPNRAVALWAMGTTVMAFLGAGVWGFMHTLA 333
Query     316 TGVKVFSWLATLHGGNIKWSP---AMLWALGFIFLFTIG-GLTGIVLANS 361

Subject   334 PVNYYTHGTQLTAAHGHMAFYGAYAMIVMTIISYAMP 370
Query     362 SLDIVLHDTYYVVAHFHYVLSMGAVFAIMGGFVHWFP 398
```

Fig. 4. SWAT alignment of NORB_PSEAE (Subject) with COX1_DIDMA (Query). The PROSITE pattern for the heme copper oxidase family PS00077 is [YWG]-[LIVFYWTA](2)-[VGS]-H-[LNP]-X-V-x(44,47)-H-H (bold). (a) Shows the alignment given using BLOSUM 62 on the transmembrane regions in the bipartite scheme; (b) shows the alignment using the PHAT 75/73 matrix on the transmembrane regions. Note the 'HH' motif is aligned in (b) and not in (a). JTT PAM 170 and BLOSUM 55 gave the same alignment as (a). NORB_PSEAE was not found with the JTT transmembrane matrix.

with transmembrane segments is not a random sample of the whole database, but rather of hydrophobic patches and transmembrane regions. By changing the background frequencies to reflect this, the PHAT matrix outperformed generalized matrices and other transmembrane matrices for searching on transmembrane regions. In general, one should consider using the background frequencies characteristic of alignments found in the twilight zone rather than of those in the entire database when making a specialized substitution matrix.

Acknowledgements

We would like to thank Elizabeth Greene and Phil Green for thoughtful discussions and Chuck Staben for suggesting that we investigate this problem. Harmit Malik, Kavita Garg, and Max Robinson provided comments on the manuscript. P.C.Ng is supported by a fellowship from NSF. This work was supported by a grant from NIH.

References

Altschul,S., *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S. (1991) Amino acid substitution matrices from an information theoretic perspective. *J. Mol. Biol.*, **219**, 555–565.

Bairoch,A. and Apweiler,R. (1999) The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999. *Nucleic Acids Res.*, **27**, 49–54.

Dayhoff,M. (1978) *Atlas of Protein Sequence and Structure*. vol. 5, suppl.3 National Biomedical Research Foundation, Washington, D.C., pp. 345–358.

Henikoff,S. and Henikoff,J. (1991) Automated assembly of protein blocks for database searching. *Nucleic Acids Res.*, **19**, 6565–6572.

Henikoff,S. and Henikoff,J. (1992) Amino acid substitution matrices from protein blocks. *PNAS*, **89**, 10915–10919.

Henikoff,S. and Henikoff,J. (1993) Performance evaluation of amino acid substitution matrices. *Proteins: Structure, Function, and Genetics*, **17**, 49–61.

Henikoff,S. and Henikoff,J. (1997) Embedding strategies for effective use of information from multiple sequencealignments. *Protein Science*, **6**, 698–705.

Henikoff,S., Henikoff,J. and Pietrokovski,S. (1999) Blocks+: a non-redundant database of protein alignment blocks derived from multiplecompilations. *Bioinformatics*, **15**, 471–479.

Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.

Jones,D., Taylor,W. and Thornton,J. (1992) The rapid generatiion of mutation data matrices from protein sequences. *CABIOS*, **8**, 275–282.

Jones,D., Taylor,W. and Thornton,J. (1994) A mutation data matrix for transmembrane proteins. *FEBS Lett.*, **339**, 269–275.

Pearson,W. (1995) Comparison of methods for searching protein sequence databases. *Protein Science*, **4**, 1145–1160.

Persson,B. and Argos,P. (1994) Prediction of transmembrane segments in proteins utilising multiple sequencealignments. *J. Mol. Biol.*, **237**, 182–192.

Rost,B., Fariselli,P. and Casadio,R. (1996) Topology prediction for helical transmembrane proteins at 86% accuracy. *Protein Science*, **5**, 1704–1718.

Tusnady,G. and Simon,I. (1998) Principles governing amino acid composition to integral membrane proteins:applications to topology prediction. *J. Mol. Biol.*, **283**, 489–506.