Generalized Baum-Welch Algorithm Based on the Similarity between Sequences

Vahid Rezaei[1,2]*, Hamid Pezeshk[2,3], Horacio Pérez-Sa'nchez[4]

[1] Department of Mathematics and Statistics, Faculty of Financial Science, University of Economic Sciences, Tehran, Iran, [2] School of Computer Science, Institute for Research in Fundamental Science (IPM), Tehran, Iran, [3] School of Mathematics, Statistics and Computer Science, University of Tehran, Iran, [4] Computer Science Department, Catholic University of Murcia (UCAM), Murcia, Spain

Abstract

The profile hidden Markov model (PHMM) is widely used to assign the protein sequences to their respective families. A major limitation of a PHMM is the assumption that given states the observations (amino acids) are independent. To overcome this limitation, the dependency between amino acids in a multiple sequence alignment (MSA) which is the representative of a PHMM can be appended to the PHMM. Due to the fact that with a MSA, the sequences of amino acids are biologically related, the one-by-one dependency between two amino acids can be considered. In other words, based on the MSA, the dependency between an amino acid and its corresponding amino acid located above can be combined with the PHMM. For this purpose, the new emission probability matrix which considers the one-by-one dependencies between amino acids is constructed. The parameters of a PHMM are of two types; transition and emission probabilities which are usually estimated using an EM algorithm called the Baum-Welch algorithm. We have generalized the Baum-Welch algorithm using similarity emission matrix constructed by integrating the new emission probability matrix with the common emission probability matrix. Then, the performance of similarity emission is discussed by applying it to the top twenty protein families in the Pfam database. We show that using the similarity emission in the Baum-Welch algorithm significantly outperforms the common Baum-Welch algorithm in the task of assigning protein sequences to protein families.

Citation: Rezaei V, Pezeshk H, Pérez-Sa'nchez H (2013) Generalized Baum-Welch Algorithm Based on the Similarity between Sequences. PLoS ONE 8(12): e80565. doi:10.1371/journal.pone.0080565

Editor: Jérémie Bourdon, Université de Nantes, France

Received June 17, 2013; **Accepted** October 4, 2013; **Published** December 20, 2013

Copyright: © 2013 Rezaei et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: vhrezaei@ipm.ir

Introduction

Structure and function determination of newly discovered proteins, using the information contained in their amino acid sequences, is one of the most important problems in genomics [1]. Often, but certainly not always, as the homologous proteins have similar sequences and structures, they have similar functions [2]. The profile hidden Markov model (PHMM) can be applied to determine the related proteins by sequence comparison [3]. The parameters of a PHMM are of two types; transition and emission probabilities. Under a PHMM, there are two assumptions made for transition and emission probabilities as follows:

1. The t^{th} hidden state, given the $(t-1)^{th}$ hidden state, is independent of previous states.
2. The t^{th} observation depends only on the t^{th} state.

The PHMM is specified as a triplet $\lambda = (A,B,\Pi)$ where A is the transition probability matrix, B is the emission probability matrix and Π is the vector of initial probabilities. An important task in assigning a new protein sequence to a protein family is to estimate the parameters of the PHMM. The Parameters of a PHMM (transition probability matrix and emission probability matrix) can be estimated in two ways: they can be estimated either from the aligned sequences or unaligned sequences using the Baum-Welch algorithm [4].

The Baum-Welch algorithm works by guessing initial parameter values, then estimating the likelihood of the observation under the current parameters. This likelihood then will be used to re-estimate the parameters iteratively until a local maximum is reached. The Baum-Welch algorithm finds $Max_{\lambda}\ P(\text{observation} \mid \lambda)$ by considering only the information on the previous state of a hidden state. In other words, in the process of the Baum-Welch algorithm, it is assumed that given states the observations are independent and only the dependency between hidden states is considered. So, the dependency between observations can be combined with the PHMM. For this purpose, the multiple sequence alignment (MSA) which is a representative of a PHMM can be considered. In this paper the ClustalW program which is the current implementation of MSA is used for consideration of the dependency between observations.

Based on the MSA, one-by-one dependencies between corresponding amino acids of two current sequences that model the similarity between them can be appended to the PHMM. This approach in spirit is similar to the works proposed by Holmes [5], Qian and Goldstein [6] and Siepel [7] where a PHMM is augmented with phylogenetic trees. In their approach, the evolutionary information is appended to the PHMM. They considered the dependency between sequences based on the fact that all the current sequences (external nodes in the guide tree or phylogenetic tree) are dependent upon their ancestral sequences.

Based on their idea, there is no dependency between two current sequences.

But in our approach, the dependency between two current sequences based on the similarity between them can be appended to the PHMM. Based on the fact that with a MSA, the sequences are biologically related, we can use the MSA to find the areas of similarity between two current sequences. So, the MSA is used for consideration of the one-by-one dependency between observations. In other words, the dependency between corresponding amino acid located above the residue and the residue can be combined with the PHMM. Therefore the new parameters of PHMM called similarity emission (SE) probabilities are created and should be estimated.

It should be noted that the similarity emission probabilities are estimated from the MSA and then combined with the common emission probabilities estimated from Baum-Welch algorithm to generalize the Baum-Welch algorithm. In other words, both aligned and unaligned sequences are used to generalize the Baum-Welch algorithm: aligned sequences for estimation of the similarity emission probabilities and unaligned sequences for estimation of the common emission and transition probabilities.

In this paper, we first construct a PHMM. Then using a MSA, we model the similarity emission (SE) matrix for consideration of the similarity information and generalize the Baum-Welch algorithm. We finally compare the results of applying the similarity emission to the Baum-Welch algorithm with the results of the commonly used emission for sequence alignment. For this purpose we use real data from the top twenty protein families in the Pfam database [8].

Materials and Methods

2.1 The PHMM

The profile hidden Markov model (PHMM) is a useful method to determine distantly related proteins by sequence comparison [3]. The PHMM is a linear structure of three states named; Match (M), Delete (D), and Insert (I). Therefore we need to decide how many states exist in a PHMM. In other words, how many match states do we have in a family? Here we assume that K is the number of match states in the PHMM. A commonly used rule is to set K equal to the number of columns of the MSA including more than half of the amino acid characters. Note that the number of match states is related to the length of the MSA [9]. So, the total number of M, D and I states is 3K. Begin and End states which emit no output symbols are introduced as dummy states [9]. Since there is an Insert state for each transition, there should be a transition from Begin called I_0. Therefore the total number of states is 3K+3. Twenty amino acids are observed from Match and Insert states. Delete, Begin and End states are silent states because they do not emit any symbols.

Following Durbin [10], we estimate the transition probabilities, A, and the emission probabilities, B, using the plan7 construction (Figure 1). Unlike the original Krogh/Haussler and HMMER model architecture, Plan 7 has no D→I or I→D transitions. This reduction from 9 to 7 transitions per node in the main model is the origin of the codename Plan 7. Note that the transition probability a_{ij} is the probability of moving from state i to state j i.e.

$$a_{ij}=P(entering\ state\ q_j\ at\ time\ t+1|the\ process\ is\ in\ state\ q_i\ at\ time\ t)$$

and emission probability $b_j(k)$ is the probability of observation o_k being emitted from state s_j i.e.

$$b_j(k)=P(producing\ o_k\ at\ time\ t|the\ process\ is\ in\ state\ q_j\ at\ time\ t).$$

Parameters of a PHMM (transition probability matrix $A_{(3K+3)\times(3K+3)}$, and emission probability matrix $B_{(3K+3)\times 20}$) can be estimated using the Baum-Welch algorithm.

2.2 Considering The Similarity Between Sequences in the Baum-Welch Algorithm

The sequences appearing in the final multiple sequence alignment are written based on their similarity [10]. So, in a PHMM, the one-by-one dependency between corresponding amino acids of two current sequences can be considered. Therefore, we propose a model which considers the effect of the similarity information (the dependency between observation) as well as the effect of the hidden state on the previous state of an amino acid in a PHMM. For consideration of the similarity information, we introduce a similarity emission probability matrix based on the multiple sequence alignment. This matrix illustrates the similarity dependencies among the observations. Following the MSA, we assume that protein sequences consisting of 21 observations (20 amino acids and one gap) have been placed on a regular lattice. In other words, each observation is arranged as a site and a matrix with R rows and L columns (length of sequences) is obtained. This matrix is called the MSA matrix, in which the site position above the s = (r, c) is denoted by (r -1, c). Hence, we assume each site on the lattice has a dependency with the corresponding residue located at the above position. This scheme is a special case of the discrete state hidden Markov random field (HMRF) with 2-point cliques (Table 1). Note that the adjective 'hidden' refers to the states. The ingredients of this model are as follows:

1. S: a set of lattice points
2. s: a lattice point, $s \in S$, $s=(r,c)$, $1 \le r \le R$, $1 \le c \le L$
3. Emissions (O_s): an observation at point s
4. Hidden states (Q_s): the hidden state at point s
5. ∂s: the neighboring point of s (in this work, it is the above position of an amino acid)
6. Transition probabilities on the lattice: a matrix A with following entries:

$$a_s(i,j)=P(Q_s=i|Q_{s-1}=j),\ 1 \le i,j \le N, s \ge 2$$

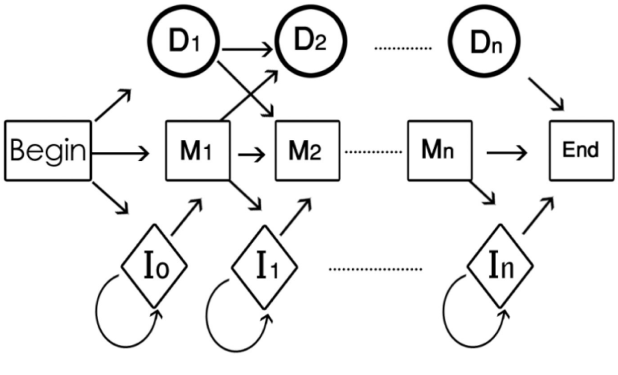

Figure 1. Plan 7 Construction.
doi:10.1371/journal.pone.0080565.g001

where, $s-1$ is a lattice point at previous state of s and N is the total number of hidden states.

7. Emission probabilities on the lattice: a matrix B with the following entries:

$$b_s(k,i)=P(O_s=k|Q_s=i),\ 1\leq i\leq N,\ 1\leq k\leq M,$$

where M is a set of symbols that may be observed.

8. Emission probabilities on the lattice based on the above position: a matrix E with the following entries representing the probabilities of the above position of an observation on the lattice:

$$E_s(i,k)=P(O_{\partial s}=i|O_s=k),$$

where $O_{\partial s}$ is an observation (amino acid or gap) at the above position.

9. Initial value: the probability of starting state at $s=(r,1)$, $\forall r\geq 1$:

$$\pi_s(j)=P(Q_s=j).$$

The likelihood of the parameters (λ) given the observations is:

$$\begin{aligned}
L(\lambda|O)=p(O|\lambda)&=\sum_q p(O|Q,\lambda)p(Q|\lambda)\\
&=\sum_q\prod_s p(O_s|Q_s,O_{\partial s})p(Q_s|Q_{s-1})\\
&=\sum_q\prod_s\frac{p(Q_s,O_{\partial s}|O_s)p(O_s)}{p(Q_s,O_{\partial s})}p(Q_s|Q_{s-1})\\
&=\sum_q\prod_s\frac{p(Q_s|O_s)p(O_{\partial s}|O_s)p(O_s)}{p(Q_s)p(O_{\partial s})}p(Q_s|Q_{s-1})\\
&=\sum_q\prod_s\frac{p(O_s|Q_s)p(O_{\partial s}|O_s)}{p(O_{\partial s})}p(Q_s|Q_{s-1})\\
&=\sum_{j,l}\prod_{o_s=i,o_{\partial s}=k}\frac{p(O_s=i|Q_s=j)p(O_{\partial s}=k|O_s=i)}{v}\\
&\times p(Q_s=j|Q_{s-1}=l)\\
&=\sum_{j,l}\prod_i\frac{b_s(i,j)E_s(k,i)}{v}a_s(j,l)\pi_s(j),
\end{aligned}\tag{1}$$

where v is a constant equal to $\frac{1}{21}=0.047$. It should be noted that in Equation (1), the Q_s and $O_{\partial s}$ are independent, because Q_s emits only O_s. Based on Equation (1), we wish to find $\lambda^*=(A,B,E,\Pi)$, where $\lambda^*=argmax_\lambda L(\lambda|O)$. The entries of matrices A, B, and the vector Π will be estimated through the following steps [11]:

Table 1. An example of the dependency between corresponding residues.

$X_{1,1}$	$X_{1,2}$	$X_{1,3}$
$X_{2,1}$	$X_{2,2}$	$X_{2,3}$
$X_{3,1}$	$X_{3,2}$ ⇑	$X_{3,3}$
$X_{4,1}$	$X_{4,2}$	$X_{4,3}$

doi:10.1371/journal.pone.0080565.t001

1. Define auxiliary forward variable $\alpha_s(i)$ which is the probability of the partial observation sequence $O_1,\cdots,O_s$ at lattice points $1,\cdots,s$ when it terminates at the state i:

$$\alpha_s(i)=P(O_1,\cdots,O_s|Q_s=i,\lambda)$$

2. Define backward variable $\beta_s(i)$ as the probability of the partial observation sequence $O_{s+1},\cdots,O_T$, given that the current state is i:

$$\beta_s(i)=P(O_{s+1},\cdots,O_T|Q_s=i,\lambda)$$

3. Calculate $\xi_s(i,j)$ as the probability of being in state i at lattice point s and in state j at lattice point $s+1$, given observations and model:

$$\xi_s(i,j)=P(Q_s=i,Q_{s+1}=j|O,\lambda)$$

This is the same as,

$$\xi_s(i,j)=\frac{P(Q_s=i,Q_{s+1}=j,O|\lambda)}{P(O|\lambda)}$$

Using forward and backward variables this can be expressed as,

$$\xi_s(i,j)=\frac{\alpha_s(i)a_s(i,j)\beta_{s+1}(j)b_s(o_{s+1},j)}{\sum_{i=1}^N\sum_{j=1}^N\alpha_s(i)a_s(i,j)\beta_{s+1}(j)b_s(o_{s+1},j)}$$

4. Define variable $\gamma_s(i)$ as the probability of being in state i at lattice point s, given the observations and model:

$$\gamma_s(i)=P(Q_s=i|O,\lambda)=\sum_j\xi_s(i,j)$$

In forward and backward variables this can be expressed by,

$$\gamma_s(i)=\frac{\alpha_s(i)\beta_s(i)}{\sum_{i=1}^N\alpha_s(i)\beta_s(i)}.$$

Now it is possible to use the Baum-Welch algorithm to maximize the quantity, $P(O|\lambda)$. The estimation of parameters based on iterative calculation can be obtained by the following expressions:

$$\hat{a}_s(i,j)=\frac{\sum_{s=1}^{T-1}\xi_s(i,j)}{\gamma_{s=1}^{T-1}}$$

$$\hat{b}_s(k,j)=\frac{\sum_{o_s=k}^T\gamma_s(j)}{\sum_{s=1}^T\gamma_s(j)}$$

$$\hat{\pi}_s(j)=\gamma_s(j), s=(r,1).$$

Hence, we have the matrices of parameter estimation $\hat{B}_{(3K+3)\times 20}$, $\hat{A}_{(3K+3)\times(3K+3)}$, and vector $\hat{\Pi}$.

Since the matrix E is a type of emission probability matrix, it should have the same size as the matrix $B_{(3K+3)\times 20}$. The estimation method for the matrix $E_{(3K+3)\times 20}$ is as follows:

In the MSA matrix, the frequencies of ordered pairs of 20 amino acids and the gap i.e. $(O_{\partial s}, O_s)$ in each column are determined. It should be noted that O_s represents the amino acids (20 types) at lattice point $s=(r,c)$ and $O_{\partial s}$ is the amino acids or one gap (21 types) located above O_s. In other words, for a given amino acid O_s, the position $(r-1,c)$ can be filled with any of 20 types of amino acids or the gap. hence, we can imagine of having a 420 (20×21) by L frequency matrix. After dividing these frequencies by the sum of frequencies in each column, the probabilities are estimated as follows:

$$\hat{E}_s(i,k)=\hat{P}(O_{\partial s}=i|O_s=k),$$

$$1\leq i\leq 21, \quad 1\leq j\leq 20$$

for which i and k are amino acids or the gap and $\hat{E}_s(i,k)$ is the conditional probability of i given k at lattice point s. This procedure produces the matrix $\hat{E}_{420\times L}$.

In each column of matrix $\hat{E}_{420\times L}$, for every set of 21 probabilities, the highest probability is chosen. In other words, the highest probability for a given amino acid O_s in the position (r,c) should be chosen. Then the matrix $\hat{E}_{420\times L}$ is reduced to a new matrix with 20 rows and L columns $(\hat{E}_{20\times L})$. After transposing the matrix $\hat{E}_{20\times L}$, the matrix $\hat{E}_{L\times 20}$ is obtained.

We assume that the L rows consist of Match and Insert states in which each Insert state can be repeated on its own several times. Using this assumption, we determine the Match states in $\hat{E}_{L\times 20}$ corresponding to Match states in $\hat{B}_{(3K+3)\times 20}$. Note that there are K Match states. In addition, the average values of the rows between each of two Match states in $\hat{E}_{L\times 20}$ are considered as Insert states. So, the matrix $\hat{E}_{L\times 20}$ is changed to the matrix $\hat{E}_{(2K+1)\times 20}$ with $2K+1$ Match and Insert states in a row. The Delete states are included by adding zeros to the rows of $\hat{E}$, so that $3K+1$ states are obtained.

Since the Begin and the End states are silent and do not emit any symbols, the two rows with zero number can be added at the beginning and the end of matrix $\hat{E}_{(3K+1)\times 20}$. Consequently the matrix $\hat{E}_{(3K+3)\times 20}$ is obtained. This matrix is the estimation of the matrix $\{P(O_{\partial s}|O_s)\}_{(3K+3)\times 20}$.

2.3 Similarity Emission Matrix

The Baum-Welch algorithm defines an iterative procedure for estimating the parameters. It computes maximum likelihood estimators for the unknown parameters given observation [11]. Since the Baum-Welch algorithm finds local optima, it is important to choose initial parameters carefully. In this paper we perform the algorithm with different initial parameters in such a way that the transition probabilities into Match states are larger than transition probabilities into other states. In order to improve the prediction accuracy of assigning sequences to protein families, we consider both emission probability matrices $\hat{E}_{(3K+3)\times 20}$ and $\hat{B}_{3K+3\times 20}$. We generalize the Baum-Welch algorithm by integrating the both emission probability matrices $\hat{E}_{(3K+3)\times 20}$ and $\hat{B}_{3K+3\times 20}$ called similarity emission matrix (SE). In what follows, we give the details:

1. Count the frequencies of ordered pairs of 20 amino acids and the gap, i.e., $(O_{\partial s}, O_s)$ in each column of the MSA matrix
2. Calculate the probability matrix $\hat{E}_{420\times L}$ of ordered pairs by dividing frequencies by the sum of frequencies in each column with elements:

$$\hat{E}_s(i,k)=\hat{P}(O_{\partial s}=i|O_s=k)$$

3. Choose the highest probability for each set of twenty one probabilities of each column of matrix $\hat{E}_{420\times L}$, to obtain the matrix $\hat{E}_{20\times L}$
4. Transpose the matrix $\hat{E}_{20\times L}$ to obtain the matrix $\hat{E}_{L\times 20}$
5. Write directly the values of Match states of L rows and the average values of Insert states between two Match states of the matrix $\hat{E}_{L\times 20}$ to obtain the matrix $\hat{E}_{(2K+1)\times 20}$. It should be noted the Match and Insert States will be obtained by using the multiple sequence alignment.
6. Add zero rows after each Match and Insert states to the $\hat{E}_{(2K+1)\times 20}$ and also two zero rows as Begin and End states to obtain the matrix $\hat{E}_{(3K+3)\times 20}$
7. Use Hadamard product that is the entry-wise product of $\hat{E}_{(3K+3)\times 20}$ and $\hat{B}_{(3K+3)\times 20}$ and then divide the entries by

Table 2. Top twenty protein families in pfam database.

profile	Number of sequence	
	Seed	Full
ABC tran	60	163029
RVT 1	155	126258
COX1	94	118265
GP120	24	105452
WD40	1842	101999
RVP	50	93675
zf-C2H2	195	88330
Response_reg	57	75322
Cytochorm B N	92	70463
HA TPase c	662	70410
BPD transp 1	81	70027
MFS_1	196	69503
Oxidored q1	33	60333
Pkinase	54	56691
Cytochrom_B_C	114	51006
RVT_thumb	41	50191
Adh short	230	50144
Acetyltransf 1	243	46279
Helicase_C	491	42435
HTH_1	1556	41545

doi:10.1371/journal.pone.0080565.t002

0.047 to get the estimated similarity emission $\hat{S}E_{(3K+3)\times 20}$ with the following entries:

$$\hat{P}(O_s|Q_s,O_{\partial s}) = \frac{\hat{p}(O_s|Q_s)\hat{P}(O_{\partial s}|O_s)}{\hat{P}(O_{\partial s})} \\ = \frac{\hat{b}_s(i,j)\hat{E}_s(k,i)}{0.047} \qquad (2)$$

2.4 Data Preparation

The Pfam is a well known database of protein families [8]. It is widely used to align new protein sequences to the known proteins of a given family. There are two components in Pfam: Pfam-A and Pfam-B. The entries of Pfam-A have high quality. As shown in Table 2, we use twenty families of Pfam-A for assigning the protein sequences to these families. In this paper due to computational challenges and round-off errors in estimating parameters, we selected just twenty protein families from Pfam database which called top twenty HMM.

Results and Discussion

To assess the performance of our method, ten sequences from each of the top twenty families are randomly removed. These ten removed sequences in each family are used as test sequences, while the other sequences form the training set. We repeat this procedure ten times. Since some of the protein families contain few proteins (likeGP120 and Oxidored q1), we choose just ten samples. Therefore, each time we have selected 200 sequences. In total 2000 sequences are randomly removed. Then we estimate the transition matrix $A_{(3K+3)\times(3K+3)}$, emission matrices $B_{(3K+3)\times 20}$ and $E_{(3K+3)\times 20}$ for each protein family. Given top twenty protein families, the score of each removed sequence belonging to each family are computed and compared. To score a sequence and assign it to one of the top twenty families, we use the logarithm of the probability score. It is defined by

$$log_2 \frac{prob}{null-prob} = log_2(prob) - log_2(null-prob) \qquad (3)$$

where prob is the probability of sequence based on parameter estimation and null-prob is equal to $(0.05)^T$ where T is the length of sequence. Since there are twenty amino acids, the probability of random occurrence of each of them is 0.05. Hence, for a sequence of L amino acids, the probability of random occurrence is $(0.05)^L$.

In this paper, due to computational challenges and round-off errors in estimating probabilities of $B_{(3K+3)\times 20}$ and $E_{(3K+3)\times 20}$, we have employed logarithm transformation instead of the direct multiplication of these probabilities:

$$log_2\hat{S}E = log_2\hat{b} + log_2\hat{E} - log_2 0.047.$$

The mean and standard error of the numbers of correctly assigned proteins to the top twenty protein families are shown in Table 3. Based on the results shown in Table 3, the assignment of sequences to the protein families using the $\hat{S}E_{(3K+3)\times 20}$ is considerably improved. For all protein families, more than half of the sequences are assigned correctly. In the task of assigning

Table 3. The mean and standard error of the numbers of correctly assigned sequences.

profile	Mean		Standard Error	
	Using $\hat{B}_{(3K+3)\times 20}$	Using $\hat{S}E_{(3K+3)\times 20}$	Using $\hat{B}_{(3K+3)\times 20}$	Using $\hat{S}E_{(3K+3)\times 20}$
ABC_tran	6.200	9.100	0.805	0.482
RVT 1	9.102	9.723	0.588	0.531
COX1	5.529	9.34	0.534	0.482
GP120	9.034	9.980	0.460	0.405
WD40	7.515	8.601	0.672	0.520
RVP	6.129	8.802	0.801	0.672
zf-C2H2	1.980	9.001	0.534	0.578
Response_reg	8.456	8.991	0.555	0.612
Cytochorm B N	7.800	8.901	0.850	0.601
HA TPase c	7.098	9.992	0.640	0.504
BPD transp 1	7.091	8.002	0.605	0.604
MFS_1	8.409	8.997	0.583	0.538
Oxidored q1	8.001	8.973	0.593	0.471
Pkinase	2.009	8.623	0.981	0.812
Cytochrom_B_C	8.032	9.010	0.524	0.503
RVT_thumb	6.839	8.902	0.835	0.561
Adh short	6.998	8.572	0.984	0.607
Acetyltransf 1	6.504	9.760	0.551	0.504
Helicase_C	7.228	8.423	0.682	0.634
HTH_1	1.734	7.991	0.609	0.684

doi:10.1371/journal.pone.0080565.t003

Table 4. The mean and standard error of the standard scores of assigning sequences to each protein family based on the emission matrix $\hat{B}_{(3K+3)\times 20}$ and similarity emission matrix $\hat{SE}_{(3K+3)\times 20}$.

profile	Mean		Standard Error	
	Using $\hat{B}_{(3K+3)\times 20}$	Using $\hat{SE}_{(3K+3)\times 20}$	Using $\hat{B}_{(3K+3)\times 20}$	Using $\hat{SE}_{(3K+3)\times 20}$
ABC_tran	−0.834	−0.503	0.054	0.043
RVT 1	−0.546	−0.504	0.213	0.113
COX1	0.789	0.881	0.085	0.054
GP120	0.115	0.234	0.085	0.079
WD40	0.356	0.487	0.076	0.065
RVP	0.244	0.307	0.082	0.058
zf-C2H2	−0.567	−0.523	0.048	0.043
Response_reg	−0.775	−0.709	0.061	0.062
Cytochorm B N	2.143	3.4452	0.233	0.231
HA TPase c	1.814	3.651	0.202	0.200
BPD transp 1	0.807	0.718	0.069	0.058
MFS_1	−0.213	−0.035	0.082	0.044
Oxidored q1	−0.403	−0.352	0.050	0.078
Pkinase	−0.046	0.567	0.070	0.065
Cytochrom_B_C	−0.749	−0.757	0.089	0.055
RVT_thumb	0.005	0.142	0.057	0.021
Adh short	−0.550	−0.523	0.079	0.078
Acetyltransf 1	0.453	0.501	0.053	0.059
Helicase_C	0.478	0.501	0.078	0.076
HTH_1	0.640	0.703	0.070	0.052

doi:10.1371/journal.pone.0080565.t004

protein sequences, measuring the specificity is also important to prevent false positive prediction. Specificity is a statistical measure of the performance of a classification test, also known in statistics as classification function. Specificity measures the proportion of negatives which are correctly identified. This measure is closely related to the concepts of type II errors in testing a statistical hypothesis . Specificity relates to the ability of the test to identify negative results. This can also be written as:

$$Specificity = \frac{TN}{TN+FP},$$

where, TN is the number of True Negative and FP is the number of False Positive. In other words, specificity means how many of the true negatives are detected? Ideally, suitable method should have high specificity or a perfect predictor would be described as 100% specificity. The specificity on average is about 75% and 62% using similarity emission and common emission model respectively. In addition to the correct assignment, the mean of the standard assigning scores, based on the matrix $\hat{SE}_{3K+3\times 20}$ in most families are more than those obtained by the matrix $\hat{B}_{3K+3\times 20}$ (Table 4). The results presented in this paper show that considering a model which incorporates the similarity information of the corresponding amino acid located above a residue in a protein family will result in a notable improvement in assignment task. It should be noted that based on the MSA implemented by ClustalW, one-by-one dependencies between corresponding amino acids of two current sequences that model the similarity between them can be appended to the PHMM. In other words, we combine the similarity emission matrix obtained form the aligned sequences and common emission matrix obtained from the unaligned sequences to generalize the Baum-Welch algorithm.

Acknowledgments

Vahid Rezaei is grateful to the School of Computer Science at IPM (No. CS1391-4-21), and Department of Statistics and Actuarial Science, University of Waterloo. This work was completed when he was as a visiting Scholar at University of Waterloo. Hamid Pezeshk would like to thank the Department of Research Affairs of University of Tehran and Biomath research group of IPM.

Author Contributions

Conceived and designed the experiments: VR HP HPS. Performed the experiments: VR HP HPS. Analyzed the data: VR HP HPS. Contributed reagents/materials/analysis tools: VR HP HPS. Wrote the paper: VR HP HPS.

References

1. Karp R (2002) Mathematical challenges from genomics and molecular biology. Notices of the AMS 49: 544–553.
2. Sangar V, Blankenberg D, Altman N, Lesk A (2007) Quantitative sequence-function relationships in proteins based on gene ontology. BMC bioinformatics 8: 294.
3. Gribskov M, McLachlan A, Eisenberg D (1987) Profile analysis: detection of distantly related proteins. Proceedings of the National Academy of Sciences 84: 4355.
4. Baum L, Petrie T, Soules G, Weiss N (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. The annals of mathematical statistics 41: 164–171.

5. Holmes I, Bruno W (2001) Evolutionary hmms: a bayesian approach to multiple alignment. Bioin- formatics 17: 803–820.
6. Qian B, Goldstein R (2004) Performance of an iterated t-hmm for homology detection. Bioinfor- matics 20: 2175–2180.
7. Siepel A, Haussler D (2004) Combining phylogenetic and hidden markov models in biosequence analysis. Journal of Computational Biology 11: 413–428.
8. Finn R, Mistry J, Tate J, Coggill P, Heger A, et al. (2010) The pfam protein families database. Nucleic acids research 38: D211.
9. Eddy S (1998) Profile hidden markov models. Bioinformatics 14: 755.
10. Durbin R, Eddy S, Krogh A, Mitchison G (2002) Biological sequence analysis. Cambridge university press Cambridge, UK:.
11. Bilmes J (1998) A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. International Computer Science Institute 4: 126.