PROGRESS IN BIOLOGICAL SCIENCES

Comparing the Bidirectional Baum-Welch Algorithm and the Baum-Welch Algorithm on Regular Lattice

Vahid Rezaei[1, 4],Sima Naghizadeh[2], Hamid Pezeshk[3, 4, *], Mehdi Sadeghi[5] and Changiz Eslahchi[6]

[1]*Department of Mathematics, K.N.Toosi University of Technology. Tehran, Iran.*
[2]*The National Organization for Educational Testing (NOET), Ministry of Science, Research and Technology.*
[3]*School of Mathematics, Statistics and Computer Science, College of Science, University of Tehran, Iran.*
[4]*Bioinformatics Research Group, School of Computer Science, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran.*
[5]*Department of Biophysics, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran.*
[6]*Faculty of Mathematical Science, Shahid-Beheshti University, G. C., Tehran, Iran.*

Received: June, 14, 2012; Accepted: August, 28, 2012

Abstract

A profile hidden Markov model (PHMM) is widely used in assigning protein sequences to protein families. In this model, the hidden states only depend on the previous hidden state and observations are independent given hidden states. In other words, in the PHMM, only the information of the left side of a hidden state is considered. However, it makes sense that considering the information of the both left and right sides of a hidden state can improve the assignment task. For this purpose, bidirectional profile hidden Markov model (BPHMM) can be used. Also, because of the evolutionary relationship between sequences in a protein family, the information of the corresponding amino acid in the preceding sequence of residues in the PHMM can be considered. For this purpose the hidden Markov random field on regular lattice (HMRFRL) is introduced. In a PHMM, the parameters are defined by the transition and emission probability matrices. The parameters are usually estimated using an EM (Expectation-Maximization) algorithm known as Baum-Welch algorithm. In this paper, the bidirectional Baum-Welch algorithm and theBaum-Welch algorithm on regular lattice are defined for estimating the parameters of the BPHMM and the HMRFRL respectively. We also compare the performance of common Baum-Welch algorithm, bidirectional Baum-Welch algorithm and the Baum-Welch algorithm on regular lattice by applying them to the real top ten protein families from Pfam database. Results show that using the lattice model for sequence assignment increases the number of correctly assigned protein sequences to profiles compared to BPHMM .

Keywords: Profile Hidden Markov Model, EM Algorithm, Bidirectional Baum-Welch Algorithm, Regular Lattice.

Introduction

A central problem in genomics is to determine functions of newly discovered proteins using the information contained in their amino acid sequences (Gribskov *et al.*, 1987). The fundamental assumptions are that homologous protein sequences have similar functions and similar structures (Churchil1, 1989). Based on this assumption homologous sequences have been grouped into known protein families. There are many systematic methods that have been developed to assign sequences to protein families (Pearson and Sierk, 2005). One of the most important methods to recognize homologous sequences from a protein family is profile hidden Markov Models (PHMM) (Krogh *et al.*,1994). The PHMM uses hidden Markov model (HMM) to provide a better method for dealing with gaps

* Corresponding author: pezeshk@khayam.ut.ac.ir
Tel.: +98-21-61112917; Fax: +98-21-66412178

found in protein sequences (Rabiner, 1989). The PHMM is a representation of multiple sequence alignment of protein families in profiles. The five components of HMM are defined as follows (Eddy , 1998):

States: a set of states:$S=\{S_1, S_2, S_3,..., S_N\}$

Emissions: a set of symbols that may be observed $O= \{O_1,O_2,...,O_M\}$

Transition probabilities: a matrix A which its entries represent the probability of transition from hidden state s_i to hidden state s_j:

$$a_i(j) = P(S_{t+1} = s_j | S_t = s_i),\ 1\leq i, j \leq N$$

Emission probabilities: a matrix B which its entries denote the probabilities of amino acid residue o_k being emitted by state s_j:

$$b_j(k) = P(O_t=o_k \,| S_t = s_j), 1 \leq j \leq N, 1\leq k \leq M$$

Initial state distribution: it is the probability that s_i is a start state:

$$\pi_i = P(S_1=s_i), 1\leq i \leq N$$

In the PHMM it is assumed that the amino acid sequences are emitted from the hidden states. Note that the PHMM is specified as a triple λ=(A, B, π) where A is the transition probability matrix, B is the emission probability matrix and π is the vector of initial values. The Baum-Welch algorithm which is a special type of EM, algorithm is used for estimating these parameters (Bilmes, 1998). In the Baum-Welch algorithm it is assumed that parameters are estimated based on the information of the left side of a hidden state. But it makes sense to assume that the performance of protein recognition and assignment to PHMM's can be improved by considering information of the both sides of a hidden state. We call this the bidirectional profile hidden Markov model (BPHMM). Following Aghdam (Aghdam *et al.*, 2010), we use the Baum-Welch algorithm twice (from left to right and right to left) or the bidirectional Baum-Welch algorithm for estimating the parameters of BPHMM.

Also, one of the major limitations in PHMM is the assumption that successive amino acid residues are independent (Rabiner, 1989). So, the probability of a protein sequence, is usually written as the product of probabilities of amino acid residues $\{O_1, ..., O_M\}$, i.e.,

$$P(O_1, O_2,..., O_M)= \prod_i P(O_i)$$

But in protein families it is assumed that protein sequences are descended from a common ancestor (Wang and Jiang, 1994). A Multiple sequence alignment (MSA) can be used to assess the shared evolutionary origins (Just, 2001). From the resulting MSA, the phylogenetic analysis can be conducted. Many progressive alignment programs use a guild tree, which is similar to the phylogenetic tree (see Felenstein , 2004; Letunic and Bork, 2006). Using the MSA, the sequences in a protein family are arranged due to their order of evolution (Ortet and Bastein, 2010). Based on this assumption the protein sequences in the final MSA are determined by the guild tree. So, the information of the corresponding amino acid in the preceding sequence in the PHMM can be considered. We call this model the hidden Markov random field on regular lattice (HMRFRL). So, Baum-Welch algorithm on regular lattice for estimating the parameters of HMRFRL is defined. It should be noted in the HMRFRL, not only the information of the left side of a hidden state, but also the information of corresponding amino acid located above the residue in a sequence of residues should be considered.

In this paper, first both the PHMM and BPHMM are reviewed. Then the bidirectional Baum-Welch algorithm and the Baum-Welch algorithm on regular lattice for estimating the parameters of the BPHMM and HMRFRL are introduced. The necessary preliminaries including notations and extensions are presented. We show how we may modify the existing Baum-Welch algorithm to be able to use a lattice framework. Finally, we compare the results of applying both algorithms on some real data from the Pfam database (Finn *et al.*, 2010).

Material And Method

The PHMM And BPHMM

The profile hidden Markov model (PHMM) is a useful method to determine distantly related proteins by sequence comparison (Gribskov *et* *al.*, 1987). The PHMM is a linear structure of three states Match (M), Delete (D), and Insert (I). The number of the states must be determined by using a MSA on a protein family. Here we assume that K is the number of match states in the PHMM. So the total number of states is 3K+3. A commonly used rule is to set K equal to the number of columns including more than half of the amino acid characters (Eddy, 1998). 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 *et al.* (2002), using the plan 7 construction (Figure1), we estimate the transition probabilities, A, and the emission probabilities, B. This construction has no transitions from D to I, or from I to D. Figure 1 illustrates a typical PHMM. In order to improve the prediction accuracy of assigning a sequence to a PHMM, following Aghdam (Aghdam *et al.*, 2010), we consider the information of the both sides of hidden states in the PHMM called bidirectional profile hidden Markov Model (BPHMM).As mentioned in section 1, in a PHMM, just the information of the left side of a hidden state e.g.$a_i(j) = P(S_{t+1} = s_j | S_t = s_i)$ is considered. So, for considering the information of the right side of a hidden state, the PHMM is conversed.These two PHMM can be integrated using a mixture density.Therefore in the BPHMM we have two PHMM and each of these PHMM has its own A and B matrices. The first PHMM is a Left to Right model (L-R-M) that considers the left to right transition between states in PHMM. The other PHMM is called (R-L-M) since the right to left transitions among states is considered. As a result, for each sequence two probabilities under L-R-M and R-L-M models are obtained. Using a mixture density, the probability of a sequence can be calculated by:

$$P(O|\lambda)=q_1P(O|\lambda_{L\text{-}R})+q_2P(O|\lambda_{R\text{-}L}),$$

in which O denotes the sequence. q_1 and q_2 are nonnegative parameters which must be estimated by an iterative process. In this paper, the both q_1 and q_2 are assumed to be equal to ½. However, other non-negative values for these parameters (a weighted average) could be also admissible.$\lambda_{L\text{-}R}$ and $\lambda_{R\text{-}L}$ indicate the parameters of L-R-M and R-L-M models, respectively. So, the Baum-Welch algorithm is used twice for parameter estimation.

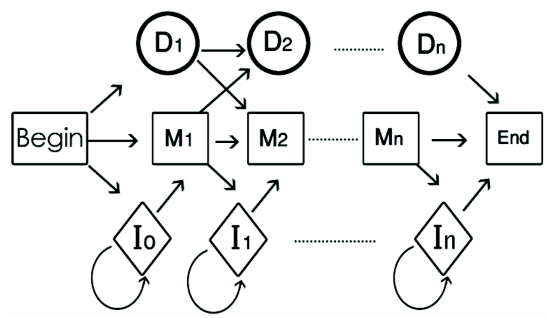

Figure 1. The Plan 7 Construction

The Bidirectional Baum-Welch Algorithm

An important question in the PHMM construction is how to estimate values of its emission probabilities ($B_{(3K+3)\times 20}$) and transition probabilities ($A_{(3K+3)\times(3K+3)}$). The Baum-Welch algorithm is usually used for estimating parameters of PHMM. The Baum-Welch algorithm defines an iterative procedure for estimation parameters that computes maximum likelihood estimators for the unknown parameters given observation (Blims , 1998). The algorithm finds $\lambda^*=\text{argmax}_\lambda\ L(\lambda;O)$ where $\lambda=(A,B,\pi)$denotes the parameters and O indicates the amino acid sequences in PHMM. The steps of Baum-Welch algorithm are as follows;

(a) Define variable $\gamma_t(i)$ as the probability of being in state s_i at time t, given the observation sequence $O_1 O_2 ... O_t$.

$$\gamma_t(i)=P(S_t=s_i|O_1O_2...O_t)$$

(b) Calculate

$$\xi_t(i,j)=P(S_t=s_i,S_{t+1}=s_j|O_1O_2...O_t)$$

(c) Calculate the expected number of transitions from state s_i to state s_j by $\sum_t \xi_t(i,j)$

(d) Calculate the expected number of transitions out of state s_i by $\sum_t \gamma_t(i)$

(e) Calculate the expected number of times that an observation o_k occurs in state s_i by $\sum_{t.o_t=o_k}\gamma_t(i)$

(f) Calculate the expected frequency in state s_i at time t=1 by $\gamma_1(i)$

Then the estimation of parameters can be found by:

$$\hat{a}_i(j) = \frac{\text{Expected number of transitions from state } s_i \text{ to state } s_j}{\text{Expected number of transitions out of state } s_i} = \frac{\sum_t \xi t(i,j)}{\sum_t \gamma t(i)}$$

$$\hat{b}_j(k) = \frac{\text{Expected number of times observation } o_k \text{ occurs in state } s_j}{\text{Expected number of transitions out of state } s_j} = \frac{\sum_{t.o_t=o_k} \gamma\, t(j)}{\sum_t \gamma t(j)}$$

$\hat{\pi}_i$ = Expected frequency of state s_i attime t=1)=γ_1(i).

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 a way that the transition probabilities into Match states are larger than transition probabilities into other states.In BPHMM, the Baum-Welch algorithm is used twice for estimating the parameters. So, using a mixture density, the Baum-Welch algorithm can be defined as a Bidirectional Baum-Welch algorithm.

The Baum-Welch Algorithm On Regular Lattice

In a PHMM, only the information on the previous state of a hidden state is considered. However, in a protein family, based on the construction of protein family, the input set of query sequences is assumed to have an evolutionary relationship, by which they are descended from a common ancestor. Therefore, we propose a model which considers the effect of the evolutionary information in a sequence of amino acids, as well as the effect of the hidden state on the previous state of an amino acid. This model is a special case of a discrete state hidden Markov random field, with one point neighborhood on the lattice. This describes a special case of first order neighborhood structure. So, we can extend this model for higher order neighborhood structure e.g. a six-regular lattice with considering the following equation introduced by Besag (1974) on a regular lattice:

$$p(z(s_{r,c}) = j \mid othersites) = \frac{\exp(u_j + \nu_j n_{jrc})}{1 + \sum_{i=1}^{s} \exp(u_i + \nu_i n_{irc})},\ j = 0,1,2,...,s,$$

where n_{jrc} is the number of sites with jth observations neighboring site (r,c),$\boldsymbol{u_i}$and $\mathbf{v_i}$ are parameters and $z(s_{r,c})$ isone of the m amino acids residue (multinomial data) in site (r,c). On a regular lattice the neighbors of the site with coordinates (r,c) can be denoted by {(r - 1, c), (r + 1, c),(r, c - 1),(r, c+1),(r+1,c+1), (r-1,c-1)}. This can be considered as six-regular lattice (see Rezaei *et al*, 2013).

In this section, for introducing the Baum-Welch algorithm on regular lattice, we need to define a new emission probability matrix by considering the information of the corresponding amino acid in the preceding sequence in the multiple sequence alignment (MSA). The MSA is a sequence alignment of three or more biological sequences such as protein, DNA, or RNA. Typically it is implied that the set of sequences share an evolutionary relationship, which means they are all descendants from a common ancestor. So, there is a relationship between phylogenies and the MSA. Phylogenetic is an area of research concerned with finding the genetic relationships between various organisms based on evolutionary relationship (Della Vedova, 2000). So, for considering the information of above amino acids, after performing an MSA on a protein family, we assume the protein sequences consisting of 21 observations (20 amino acids and one gap), are arranged in a regular lattice grid as shown by Figure 2.

So, the MSA matrix with R rows (the number of sequences) and C columns is obtained ($MSA_{R\times C}$). In other words, each amino acid is arranged as a site. This matrix is called the MSA matrix, in which the site above the (r, c) is denoted by (r – 1, c). Hence, we assume each site on the lattice has a dependency with the above residue. Using the MSA matrix, we want to build a new emission probability matrix in which the size of this matrix must be the same as the emission probability matrix with 3K+3 rows and 20 columns. So, the estimation method for the new emission probability matrix, shown by $\beta_{(3K+3)\times 20}$, is as follows:

The frequencies of ordered pairs of 20 amino acids and one gap, i.e., ($O_{r-1,c}$,$O_{r,c}$) in each column of MSA matrix are counted. In other words, in each column, for a given amino acid $O_{r,c}$ (20 amino acids), the position (r-1,c) can be filled with any of 20 types of amino acids or one gap (21 observations). So we have a 420 (20× 21) by C matrix. Dividing these frequencies by the sum of frequencies in each column the probabilities are estimated as follows:

$$\hat{\beta}_i(j)=\hat{P}(O_{r-1,c}=o_j|O_{r,c}=o_i) = \frac{\hat{P}(O_{r-1,c}=o_j,O_{r,c}=o_i)}{\sum_j \hat{P}(O_{r-1,c}=o_j,O_{r,c}=o_i)}$$

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

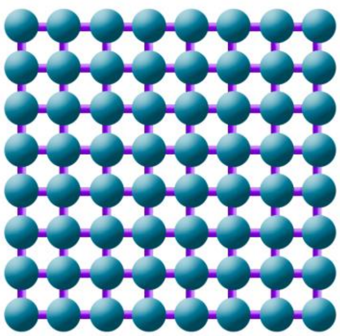

Figure 2. Regular Lattice

This step provides us with the matrix $\hat{\beta}_{420\times C}$. Maximum likelihood estimators are frequently used to estimate parameters of a distribution. That is to find the parameters which makes the probability (or the chance of) observing data maximum.In the first step of construction of matrix $\hat{\beta}$ the matrix $\hat{\beta}_{420\times C}$ is obtained. So, the highest probability for a given amino acid $O_{r,c}$, should be chosen so that the matrix $\hat{\beta}_{420\times C}$ is changed to matrix $\hat{\beta}_{20\times C}$.

In each column of matrix $\hat{\beta}_{420\times C}$, the highest probability for a given amino acid $O_{r,c}$, should be chosen. In other words, in each column, the highest probability for each set of 21 amino acids or gap is chosen. Then the matrix $\hat{\beta}_{420\times C}$ is changed to matrix $\hat{\beta}_{20\times C}$.

Transposing the matrix $\hat{\beta}_{20\times C}$, the matrix $\hat{\beta}_{C\times 20}$is obtained. Because we want to build a new emission probability matrix, we have to change the matrix $\hat{\beta}_{C\times 20}$ to matrix $\hat{\beta}_{(3K+3)\times 20}$. For this purpose, we assume that the C rows consisting of Match and Insert states in which each Insert state can be repeated many times (Figure 1). In other words, the Match states in $\hat{\beta}_{C\times 20}$ are corresponding to Match states in $\hat{B}_{(3K+3)\times 20}$ and the rows are between two Match states in $\beta_{C\times 20}$ corresponding to Insert states in$\hat{B}_{(3K+3)\times 20}$.

In the matrix $\hat{\beta}_{C\times 20}$, each row that is a Match state should be chosen and the average values of Insert rows between two Match states are calculated. So, we have 2K Match and Insert states.

We add Delete state by using zero, so that 3K states are obtained. The average value of those rows which are before the M_1 is considered as I_0 state. Then the matrix $\hat{\beta}_{C\times 20}$ is changed to the matrix $\hat{\beta}_{(3K+1)\times 20}$ with 3K+1 Match, Insert and Delete states in rows. We add two zero rows to the $\hat{\beta}_{(3K+1)\times 20}$ as the Begin and the End states to obtain the matrix $\hat{\beta}_{(3K+3)\times 20}$.

Therefore, the likelihood of parameters on a regular lattice is defined by:

$$L(\lambda;O) = \sum_S P(O|S,\lambda)P(S|\lambda)= \sum_S \prod_{r,c} P(O_{r,c}|S_{r,c},O_{r-1,c})P(S_{r,c}|S_{r,c-1})$$

$$= \sum_S \prod_{r,c} \frac{P(S_{r,c}O_{r-1,c}|\ O_{r,c})\ P(O_{r,c})}{P(S_{r,c}O_{r-1,c})} \times P(S_{r,c}|S_{r,c-1})$$

$$= \sum_S \prod_{r,c} \frac{P(\ O_{r,c}\ |\ S_{r,c})\ P(O_{r-1,c}|O_{r,c})}{P(O_{r-1,c})} \times P(S_{r,c}|S_{r,c-1})$$

$$= \sum_S \prod_{o_i,o_j,s_i,s_j} \frac{P(O_{r,c}= o_j\ |\ S_{r,c}= s_i)\ P(O_{r-1,c}= o_i\ |O_{r,c}= o_j)}{0.048} \times P(S_{r,c}= s_i\ |S_{r,c-1}= s_j)$$

$$= \sum_S \prod_{i,j} \frac{b_i(j)\beta_j(i)}{0.048} \times a_j(i)\ \pi_i(1)$$

where $S_{r,c}$ and $O_{r-1,c}$ are independent, because $S_{r,c}$just emits $O_{r,c}$. It should be noted that, in Equation (1), $P(S_{r,c})$ and $P(O_{r,c})$ on rectangular lattice have the same meaning as $P(S_t)$$ and $P(O_t)$.Because there are 21 observations (20 amino acids and one gap), the $P(O_{r-1,c})$ is assumed to be equal to 1/21=0.048. Also the $\beta_j(i)$'s are the elements of matrix $\beta_{(3K+3)\times 20}$.

The term $\frac{\mathrm{bi(j)\beta j(i)}}{0.048}$ is considered as an element of emission probability matrix on regular lattice (ERL):

$$\mathrm{ERL(i,j)}= \frac{\mathrm{P}(O_{\mathrm{r,c}}=o_j|S_{\mathrm{r,c}}=s_i,O_{\mathrm{r-1,c}}=o_i)}{0.048}.$$

In other words we define the estimation of emission matrixon regular lattice (ERL) that is the entry-wise product (Hadamard product) of $\hat{\beta}_{(3K+3)\times 20}$ and$\hat{B}_{(3K+3)\times 20}$ and divide the entries by 0.048 as follows:

$$\widehat{ERL}_{(3K+3)\times 20} = \frac{\hat{\beta}_{(3K+3)\times 20} \times \hat{B}_{(3K+3)\times 20}}{0.048}$$

where the ".*" is the sign of Hadamard product. Using the ERL $_{(3K+3)\times 20}$ in common Baum-Welch algorithm instead of and $\mathrm{B}_{(3K+3)\times 20}$, the Baum-Welch algorithm on regular lattice is obtained.

Applying the Bidirectional Baum-Welch algorithm and Baum-Welch Algorithm on Regular Lattice to Real Data

In the last decades, systematic methods have been developed to assign sequence to protein families. Based on multiple sequence alignment (MSA) of protein family sequences, profile methods have been introduced to search databases for homologous sequences. The Pfam (Finn , 2010) is a high quality set of annotated multiple alignment and pre-built profile HMM (PHMM). It is widely used to align new protein sequences on the known proteins of a given family or to recognize new member of a protein family. For each protein domain family in Pfam, there is a seed alignment which is a manually verified multiple alignment of a representative set of sequences. The Pfam database contains 11912 families (Release 24.0, October 2009). There are two components in Pfam: Pfam-A and Pfam-B. The Pfam-A entries have high quality. Given a PHMM in Pfam and a protein sequence, one can compute the probability that this protein being generated by the PHMM and infer the family that the new protein belongs to. For this purpose, the parameters of PHMM using Baum-Welch algorithm should be estimated. In this paper, using both the bidirectional Baum-Welch and the Baum-Welch algorithm on regular lattice, parameters (emission and transition matrices) are estimated and protein sequences are assigned to protein families. As shown in Table 1, we select and use ten families from top twenty protein families of Pfam-A for assignment protein sequences to protein families.

Table 1. Top ten protein families from the Pfam database

ID	Accession	Number of Sequences	
		Seed	Full
RVT_1	PF00078	155	126258
WD40	PF00400	1842	101999
RVP	PF00077	50	93675
Cytochrom_B_N	PF00033	92	70463
HATPase_c	PF02518	662	70410
BPD_transp_1	PF00528	81	70027
Oxidored_q1	PF00361	33	60333
Pkinase	PF00069	54	56691
adh_short	PF00106	230	50144
Acetyltransf_1	PF00583	243	46279

Results And Discussion

To assess the performance of our method, ten sequences from each ten families are randomly removed.We repeat this procedure10 times.Therefore, each time we have selected 100 sequences. So, in total 1000 sequences are randomly removed. Then each time we estimate the transition Matrix ($A_{(3K+3)\ \times\ (3K+3)}$) and the emission matrix($B_{(3K+3)\times 20}$) and $\beta_{(3K+3)\times\ 20}$ for each family.Therefore the estimation of parameters of BPHMM are obtained. It should be noted, the results of applying bidirectional Baum-Welch algorithm (BBWA) are compared with the results of Baum-Welch algorithm on regular lattice (BWARL) and common Baum-Welch algorithm (CBWA). In this paper due to computational challenges and round-off errors in estimating parameters, we selected just ten sequences from each family and used the .Net Framework which is a software framework that runs primarily on Microsoft Windows. In this framework the source codes of Matlab software is combined with codes of .Net.

Given ten protein families, the 1000 removed sequences are added to all families and the scores of the sequences belonging to each family based on values of estimations are computed and compared. To score a sequence and assign it to one of the top ten profiles, we use the log-odds score. It is defined by

$$log_2 \frac{\text{prob}}{\text{null} - \text{prob}}$$

whereprob is the probability of sequence based on parameter estimation and null-prob is equivalent $(0.05)^{(\text{lengofsequence})}$. Since there are 20 amino acids, then the probability of random occurrence of each of them is 0.05 and then for a sequence of L amino acids the probability of random occurrence is $(0.05)^{L}$. The mean of the numbers of correctly assigned proteins to the top ten protein families are shown in Table 2. Based on the results shown in Table 2, the mean of correct assignment of sequences to the protein families usingBaum-Welch algorithm on regular latticeare increased in all cases.Since the Baum-Welch algorithm finds local optima, it is important to choose initial parameters carefully.So, the number of assigning sequences to original profile can be increased by choosing proper values of the initial transition and emission probability matrices. In addition to the number of correctly assigned sequences, the mean of normalized scores for each ten removed sequences in most families (8 families), using Baum-Welch algorithm on regular lattice, are more than those obtained by bidirectional Baum-Welch algorithm (Figure 3). Note that after calculating each score based on each algorithm, the average value of scores corresponding to each family are calculated. Then they are normalized. Therefore based on each algorithm the scores are different together.

Table 2. The Mean of the Numbers of Protein Sequences Assigned Correctly.

ID	Using CBWA	Using BBWA	Using BWARL
RVT_1	9.010	9.125	9.463
WD40	7.557	7.889	8.610
RVP	5.614	6.015	7.808
Cytochrom_B_N	6.896	7.986	8.929
HATPase_c	7.523	7.125	8.981
BPD_transp_1	6.004	7.278	7.941
Oxidored_q1	8.052	8.781	8.878
Pkinase	1.034	2.107	6.334
adh_short	5.998	6.000	7.589
Acetyltransf_1	8.525	8.264	8.536

Not surprisingly, when we use more information due to evolutionary relationship between sequences, the results get better.

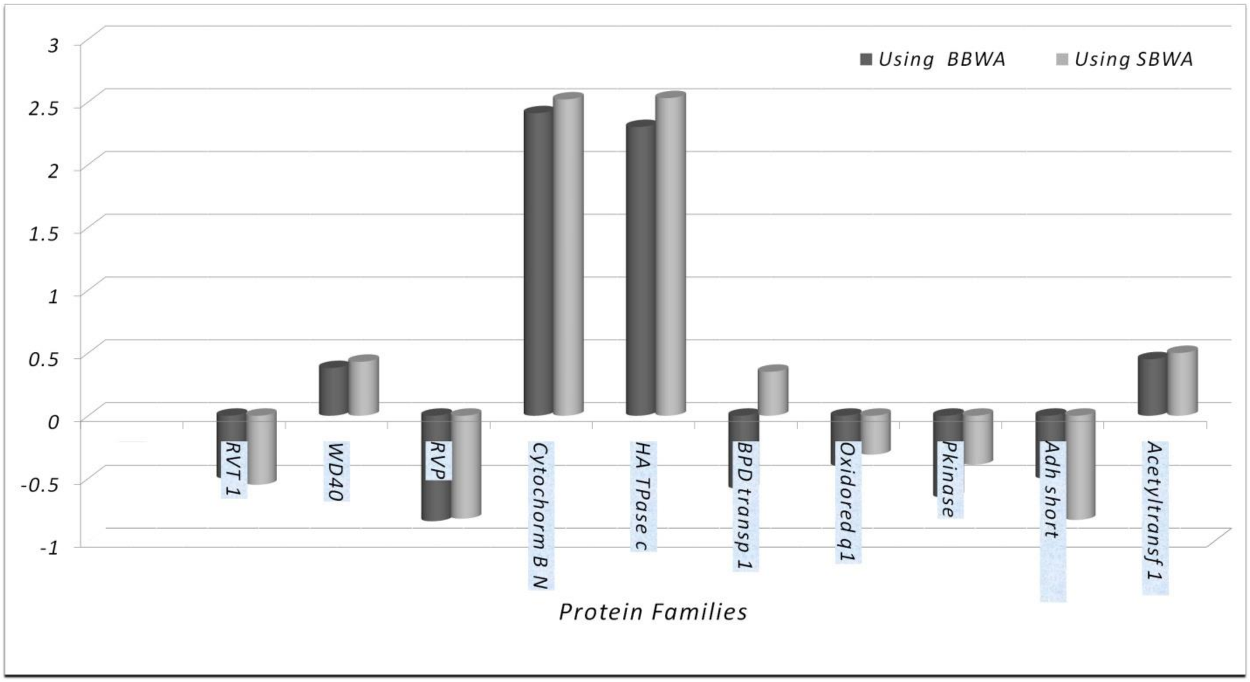

Figure 3. The Mean of Normalized Scores for each of the Ten Removed Sequences of Each Family Using BBWA and BWARL.

Acknowledgments

Vahid Rezaei is grateful to the Department of Mathematics atK.N.Toosi University of Technology. Hamid Pezeshk would like to thankthe Department of Research Affairs of University of Tehran forfinancial support (grant number: 6103016/01/07). He is also gratefulto the Biomath Research Group at the Institute for Research inFundamental Sciences. Changiz Eslahchi is supported in partby a grant from IPM (CS-1389-0-01).

References

1. Aghdam, R., Pezeshk, H., Malekpour, S., Shemehsavar, s., Sadeghi, M., and Eslahchi C, A. (2010). Bidirectional Bayesian Mont Carlo Approach for Estimating Parameters of a Profle Hidden Markov Model. *Applied Science Segment*, 1(2), APS/1531
2. Besag, J.(1974). Spatial Interaction and the Statistical Analysis of Lattice Systems, *Journal of the Royal Statistical Society*, Series B, 36(2), 192-236
3. 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
4. Churchill, G. (1989). Stochastic Models for Heterogeneous DNA Sequences. *Bulletin of Mathematical Biology*, 51(1), 79-94
5. Della Vedova, G. (2000). Multiple Sequence Alignment and Phylogenetic Reconstruction: Theory and Methods in Biological Data Analysis. PhD thesis, Citeseer
6. Durbin, R., Eddy, S., Krogh, A., and Mitchison, G. (2002). *Biological Sequence Analysis*. Cambridge university press Cambridge, UK:
7. Eddy, S. (1998). Profile Hidden Markov Models.

Bioinformatics, 14(9), 755-763

8. Felsenstein, J. (2004). Inferring Phytogenies. *Sunderland, Massachusetts*: Sinauer Associates
9. Finn, R., Mistry, J., Tate, J., Coggill, P., Heger, A., *et al.* (2010). The Pfam Protein Families Database. *Nucleic acids research*, 38(suppl 1), D211
10. Gribskov, M., McLachlan, A., and Eisenberg, D. (1987). Profile Analysis: Detection of Distantly Related Proteins. *Proceedings of the National Academy of Sciences*, 84(13), 4355-4358
11. Just, W. (2001). Computational Complexity of Multiple Sequence Alignment with Sp-score. *Journal of computational biology*, 8(6), 615-623
12. Krogh, A., Brown, M., Mian, I., Sjolander, K., and Haussler, D. (1994). Hidden Markov Models in Computational Biology: Applications to Protein Modeling. *Journal of molecular biology*, 235(5), 1501-1531
13. Letunic, I. and Bork, P. (2006). Interactive Tree of Life (itol): An Online Tool for Phylogenetic Tree Display and Annotation. *Bioinformatics*, 23(1), 127-128
14. Ortet, P. and Bastien, O. (2010). Where Does the Alignment Score Distribution Shape Come from? *Evolutionary Bioinformatics Online*, 6-159
15. Pearson, W. and Sierk, M. (2005). The Limits of Protein Sequence Comparison? *Current opinion in structural biology*, 15(3), 254-260
16. Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, 77(2), 257-286
17. Rezaeitabar, V., Pezeshk, H., Sadeghi, M., Eslahchi, Ch., (2013), “Assignment of Protein Sequence to Protein Profile Using Spatial Statistics”, *Communications in Mathematical and in computer chemistry*, 69(1), 7-24
18. Wang, L. and Jiang, T. (1994). On the Complexity of Multiple Sequence Alignment. *Journal of computational biology*, 1(4),337-348