Efficient estimation of emission probabilities in profile hidden Markov models

Virpi Ahola1,∗, Tero Aittokallio2, Esa Uusipaikka1 and Mauno Vihinen3,4

1Department of Statistics, FIN-20014 University of Turku, Finland, 2Turku Centre for Computer Science, FIN-20520 Turku, Finland, 3Institute of Medical Technology, FIN-33014 University of Tampere, Finland and 4Research Unit, Tampere University Hospital, FIN-33520 Tampere, Finland

Received on November 19, 2002; revised on March 17, 2003; accepted on June 13, 2003

ABSTRACT

Motivation: Profile hidden Markov models provide a sensitive method for performing sequence database search and aligning multiple sequences. One of the drawbacks of the hidden Markov model is that the conserved amino acids are not emphasized, but signal and noise are treated equally. For this reason, the number of estimated emission parameters is often enormous. Focusing the analysis on conserved residues only should increase the accuracy of sequence database search.

Results: We address this issue with a new method for efficient emission probability (EEP) estimation, in which amino acids are divided into effective and ineffective residues at each conserved alignment position. A practical study with 20 protein families demonstrated that the EEP method is capable of detecting family members from other proteins with sensitivity of 98% and specificity of 99% on the average, even if the number of free emission parameters was decreased to 15% of the original. In the database search for TIM barrel sequences, EEP recognizes the family members nearly as accurately as HMMER or Blast, but the number of false positive sequences was significantly less than that obtained with the other methods.

Availability: The algorithms written in C language are available on request from the authors.

Contact: virahola@utu.fi

1 INTRODUCTION

Profile hidden Markov model (HMM) has been proved to be a very powerful method in various tasks in biological sequence analysis. It has been especially successful in performing sequence database search, detecting remote homologies and building sequence alignments (Krogh et al., 1994; Eddy, 1998; Karplus et al., 1998; Park et al., 1998; Gough et al., 2001; Madera and Gough, 2002).

Profile HMMs originate from the profile analysis, which is based on the use of a profile usually generated from the sequence or structure based alignments of probe sequences and amino acid comparison table (Gribskov et al., 1987, 1990). The essence of the profile analysis is that the information concerning the conservation of the residues is incorporated into the profile, whereby the analysis is able to detect structural similarities and homologies to the sequence family. The underlying idea and the objective of the profile HMM is exactly the same as in the profile analysis, the difference being that the HMM is a well-formulated probability model. The conserved positions of the alignment are modelled by match states, and information about the character of the allowed side chain at that position is included in the emission distribution at each match state. The gap penalties for insertions and deletions, by which positions of the conserved regions are controlled, are provided by transition probabilities from/to insert and delete states.

In HMM models, emission probabilities of all 20 amino acids are estimated in all emitting states, and thus the number of estimated parameters can be enormous. For example, if the model includes 300 emitting states, the number of emission parameters is 5700. However, the majority of estimated emission parameters are actually ‘noise’, that is, probabilities of uninteresting or unconserved residues. The number of emission parameters is especially large in profile HMMs, in which each conserved position in the sequence is modelled by a different match state. The phenomenon is related to overfitting, which occurs when there is not enough data to obtain good estimates for the model parameters, and consequently the model will not generalize adequately to new data. The simplest way to overcome the overfitting is to add some artificial counts called pseudocounts to the count of each amino acid in the alignment (Durbin et al., 1998). In the Bayesian framework, this simple pseudocount method corresponds to the use of Dirichlet distribution as a prior. The pseudocount method can be generalized to the use of linear combination
of pseudocounts which corresponds to the use of Dirichlet mixture prior distribution (Sjölander, 1996).

In this paper, we present an alternative, likelihood-based approach to the problem of reducing the parameter space in HMMs. The advantage of the new method is that it explicitly takes into account conservation of the alignment, whereas the use of Dirichlet mixture distribution generally considers the different alignment environments. We consider the conserved emission parameters as effective and the uninteresting ‘noise’ as ineffective. We show how the residues can be divided into effective and ineffective ones, and introduce a method for efficient emission probability (EEP) estimation. Accuracy and variance of the EEP estimates as well as the number of misclassifications is studied using simulations. In the practical studies, the prediction power of the EEP method is demonstrated with sequences from 20 protein families. Finally, the method is used for sequence database search of TIM barrel sequences. The results were compared with the searches performed by HMMER (Eddy, 1998) and Blast (Altschul et al., 1990) software.

2 SYSTEMS AND METHODS

2.1 Profile hidden Markov model

The profile HMM architecture (Durbin et al., 1998) has three classes of states: the match state, the insert state and the delete state. The match and the insert states always emit a symbol, whereas the delete states are silent. The model always starts from the begin state and ends with the end state. The model length is determined by the number of positions, that is, the number of match–insert–delete state triplets between the begin and the end states.

An observation sequence \( Y \) is considered to be a stochastic process with a finite set of symbols \( O = \{o_1, o_2, \ldots, o_z\} \). The state sequence, the path that goes through the model, is a finite-state Markov chain \( X \). The emitted symbols are assumed to be conditionally independent given the states. A more detailed description of the HMM can be found from Koski (2001).

When the estimation is based on the sequence alignment, the columns of the multiple alignment are assigned as match or insert states before the estimation, and thus, the path that generates the sequence is known. Generally, columns representing conserved positions are chosen as match states and the rest of the states as insert states. Here, we used a heuristic rule: columns containing more than 50% of gaps were modelled by insert states and the others by match states. The sequences including gaps in the match columns are modelled as visits in the delete states.

The profile HMM has two sets of parameters: transition probabilities \( a_{j,i} = P(x_j|x_{i-1}) \), and emission probabilities \( b_j(o_l) = P(y_l = o_l|x_j) \). In the protein family examples, the emission probabilities were estimated by using both the ML and EEP method. The maximum likelihood estimates are the relative frequencies of the residues in the match columns of the multiple alignment. The emission probabilities of insert states were assigned by the background distribution. The transition probabilities were attained by summing up the number of times each transition was used.

The probability of an observation sequence is the sum over all possible paths that go through the model and generate the given sequence. The likelihood of the model is thus

\[
L(\lambda; y) = P(y_0, \ldots, y_L) = \sum_{j_0, \ldots, j_L} L(j_0, \ldots, j_L),
\]

where \( \lambda \) is a vector of all parameters. Notice that the likelihood is dependent on the sequence length \( L \) and the number of states in the path \( J \).

The log-odds score

\[
l(\lambda; \phi; y) = \log \frac{L(\lambda; y)}{L(\phi; y)}.
\]

is a log-likelihood ratio between the estimated and the random model \( \phi \) (Durbin et al., 1998). The random (null) model (Barrett, 1997) has the same model structure, model length and transition probabilities as the estimated model, only the emission probabilities are replaced by the background distribution (see Table 1). The background distribution was formed by the amino acid composition in the SWISS-PROT database (Boeckmann et al., 2003).

2.2 Classification algorithm

The EEP method is based on the fact that in match states the emission probability distributions are conserved on some residues, whereas the other residues occur relatively seldom. In practice, however, the determination of conserved residues is variable. Therefore, we present an algorithm by which the residues are divided into conserved and unconserved, or effective and ineffective, as we call them.

The residues were assigned as effective or ineffective iteratively. At each iteration step, the residue with the largest relative frequency \( b^0_j \) of respect to its background probability \( b_j^0 \) was defined as effective or ineffective depending on a fixed threshold value, and the remaining probabilities were updated so that they again summed to one. The iteration was continued until the largest ratio fell below the threshold value. The classification algorithm, where \( J_1 \) denotes a set of effective, \( J_2 \) a set of ineffective residues, and \( J = J_1 \cup J_2 \) is elaborated below.

ALGORITHM 1.

1. Set \( J_1 = \emptyset \), \( J_2 = \{1, 2, \ldots, 20\} \), and choose threshold \( a \).
2. Calculate \( r_j = b_j^0 \) for all \( j \in J_2 \).
   If \( j^* = \arg\max_{j \in J_2} r_j \) and \( r_{j^*} \geq a \)
   then \( J_1 \leftarrow J_1 \cup \{j^*\} \) and \( J_2 \leftarrow J_2 \setminus \{j^*\} \).
3. Renormalize $b_j$ and $b_j^*$ for all $j \in J_2$.
4. Return to stage 2 and repeat steps 2 and 3 until $r_{j,e} < a$.

The renormalizing step is necessary because otherwise those residues with low background probability tend to be chosen as effective more often than those with high background probability.

2.3 EEP estimation method

The EEP method is constructed by using the log-likelihood function of the multinomial distribution function

$$ l = \sum_{j \in I} n_j \log b_j, $$

where $n_j$ is a frequency of an amino acid $j$. The constraints of the log-likelihood function are determined as

$$ \frac{b_{ij}}{b_j} = \frac{b_{ie}}{b_e}, $$

$$ \frac{\sum_{j \in I_e} b_j}{\sum_{j \in I} b_j} \leq c \frac{\sum_{j \in I_e} b_j}{\sum_{j \in I} b_j}, $$

$$ \sum_{j \in I} b_j + \sum_{j \in I} b_j = 1, $$

where $i, e \in J_2$, and $c$ is a constant. The first constraint ensures that the mutual ratios of the ineffective residues remain the same as in the background distribution. The second condition is only needed to make sure that the total proportion of the ineffective residues compared to the proportion of the ineffective ones does not increase too much when compared to the proportions in the background distribution. The optimization was performed with the Lagrange multipliers method (Luenberger, 1984); see the Appendix for details.

There are two possible sets of solutions depending on the inequality

$$ \frac{n_j}{\sum_{j \in I} n_j^e} \leq \frac{\sum_{j \in I} b_j^e}{\sum_{j \in I} b_j^*}. $$

(3)

When the maximum likelihood estimates satisfy the constraint, that is, inequality (3) is true, the rescaled optimal probabilities $b_j^*$ are

$$ b_j^* = \frac{n_j}{\sum_{j \in I} n_j^e}, $$

for $j \in J_1$,

$$ b_j^* = \frac{b_j^e \sum_{j \in I} n_j^e}{\sum_{j \in I} b_j^e \sum_{j \in I} n_j^e}, $$

for $j \in J_2$.

Here, the estimates of the effective residues are the normal maximum likelihood estimates. The probabilities of the ineffective residues are estimated by dividing the sum of the remaining probability in proportion to the background probability. Hence, only one degree of freedom is needed to estimate all ineffective residues.

If the inequality (3) is not fulfilled, the probabilities are given by

$$ b_j^* = \frac{c \sum_{j \in I} b_j^e}{c \sum_{j \in I} b_j^e + \sum_{j \in I} b_j^* \sum_{j \in I} n_j^e}, $$

for $j \in J_1$,

$$ b_j^* = \frac{b_j^e \sum_{j \in I} n_j^e}{c \sum_{j \in I} b_j^e + \sum_{j \in I} b_j^*}, $$

for $j \in J_2$.

Notice that in this case the optimal solution is on the boundary of the parameter space. When the constant $c$ is large enough, the second solution will only be used when the sum of the effective probabilities approaches one. In practice, this solution is useful, because in the absence of data, the estimates produced by the ML method are simply zero, whereas the estimates of the EEP method are proportional to their background distribution. In this situation, the results of the EEP method bear similarities to the Bayesian approach, in which the prior distribution is a 1-component Dirichlet mixture distribution (Sjölander et al., 1996).

3 SIMULATION RESULTS

This chapter introduces the characteristics of the EEP estimates based on simulations. In order to study how successfully the EEP method classifies the residues as effective or ineffective, the percentages of misclassified residues were calculated.
The accuracy and variance of the EEP estimates were compared to the ML estimates. Finally, the robustness of the EEP method for the choice of the threshold value was examined.

Because of the conditional independence assumption of the HMM, protein sequences were simulated only in one match state. The theoretical simulation set was composed of three effective residues: alanine (35%), glycine (50%), and methionine (10%). The other residues were ineffective and were assigned by sharing the remaining probability in the same proportion as their background probabilities (see Table 1). One thousand simulations were run with various numbers of residues. Unless otherwise stated, the results presented here were obtained using sequences of 50 residues. The emission probabilities were estimated for the simulated residues with the ML and the EEP methods using three different threshold values: 1, 1.5 and 2.

### 3.1 False effective and ineffective residues

Table 1 shows the percentages of false classified residues. Among the effective residues, alanine and glycine were correctly classified through all simulations. The number of misclassified methionine residues increased from 0.8 to 2.9% as the threshold value was increased from 1 to 2. Hence, as the threshold value increases, the classification of effective residues whose probabilities are relatively low might fail. This problem, however, disappears as the number of estimated sequences increases.

When the threshold value was set to 1, cysteine, histidine, and tryptophan were misclassified in 5.3, 2.6 and 4.1% of the simulations, respectively, whereas the misclassification rate of the other ineffective residues was under 1%. The variations seem to be closely related to the background distribution such that residues with low background probabilities tend to be more often misclassified than the others.

### 3.2 Accuracy and variance of the estimates

The ML and the EEP estimates were compared by calculating the differences between the theoretical and the estimated probabilities. The kernel density estimates were used instead of histograms to illustrate the distribution of the differences (Fig. 1). We refer to Silverman (1986) for further details in kernel density estimation.

The estimates of the effective residues were rather accurate (Fig. 1, lower). Also, there were no great differences between the ML and the EEP estimates. Considering the ineffective residues, an evident difference between the ML and the EEP methods is the variance of the ineffective residue estimates (Fig. 1, upper). Due to the one degree of freedom that was needed to estimate all ineffective residues in the EEP method, the variances of the EEP estimates were clearly less than the variances of the ML estimates, whereas the variances of the ineffective residues remained unchanged. Additionally, when examining the ML method, the distribution of the kernel density estimates seemed to be positively skewed, which indicates that the ML method tends to overestimate the probabilities which are relatively small; the bias is, however, negligible.

### 3.3 Choosing the threshold value

In order to examine how the choice of the threshold value affects the estimates, the simulation data was estimated by using threshold values that differed from the true constant. When the threshold value was less than the true threshold, the sensitivity seemed to improve and the specificity to worsen. The opposite occurred when the threshold value was greater than the true threshold.

As far as the accuracy of the estimates was concerned, sensitivity seemed to become more important than specificity: when the effective residues were misclassified, the estimates were less accurate than in the case when the classification of the ineffective residues failed. This is because the proportions of ineffective residues are very low, and therefore also the number of false effective residues stays low.
Table 2. List of 20 protein families with their PDB ID code, average sequence length (99% confidence interval), average % identity, the number of sequences in the training and restricted/whole positive test sets, the number of parameters and the percentage of false negative and positive sequences in the ML and the EEP estimated models

<table>
<thead>
<tr>
<th>Description</th>
<th>Structure</th>
<th>Average length 99% CI</th>
<th>Average % id</th>
<th>No. seq training restricted/whole test set</th>
<th>No. of parameters in ML/EEP</th>
<th>% false neg in ML/EEP restricted set</th>
<th>% false neg in ML/EEP whole set</th>
<th>% false pos in ML/EEP whole set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amino acid kinase family</td>
<td>1e19</td>
<td>244 (168, 321)</td>
<td>20</td>
<td>109/190/209</td>
<td>4464/748</td>
<td>0.5/0</td>
<td>2.4/1.9</td>
<td>0/0</td>
</tr>
<tr>
<td>α-Amylase, catalytic domain</td>
<td>1ppi</td>
<td>380 (299, 460)</td>
<td>20</td>
<td>54/580/688</td>
<td>6839/1168</td>
<td>9.7/3.1</td>
<td>12.1/6.1</td>
<td>0/0</td>
</tr>
<tr>
<td>Multicopper oxidase</td>
<td>1aoz</td>
<td>148 (111, 184)</td>
<td>16</td>
<td>127/328/378</td>
<td>2773/448</td>
<td>14.5/6.7</td>
<td>22.0/13.8</td>
<td>0/0</td>
</tr>
<tr>
<td>Globin</td>
<td>1iba</td>
<td>143 (130, 156)</td>
<td>37</td>
<td>79/968/1122</td>
<td>2849/473</td>
<td>0.2/0.2</td>
<td>3.7/3.6</td>
<td>0/0</td>
</tr>
<tr>
<td>Glutamine</td>
<td>1gpm</td>
<td>194 (145, 243)</td>
<td>23</td>
<td>140/281/299</td>
<td>3609/585</td>
<td>4.3/1.8</td>
<td>6.0/3.7</td>
<td>0/0</td>
</tr>
<tr>
<td>Aminotransferase classqq I</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyceraldehyde 3-phosphate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dehydrogenase, NAD binding domain</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Haloacid dehalogenase-like hydrolase</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trisphosphatase isomerase</td>
<td>1ph</td>
<td>251 (235, 267)</td>
<td>43</td>
<td>20/86/155</td>
<td>4768/709</td>
<td>0/0</td>
<td>9.0/8.4</td>
<td>0/0</td>
</tr>
<tr>
<td>Immunoglobulin domain</td>
<td>6fb</td>
<td>68 (43, 94)</td>
<td>18</td>
<td>113/807/836</td>
<td>1310/173</td>
<td>2.3/1.6</td>
<td>3.9/3.3</td>
<td>0/0</td>
</tr>
<tr>
<td>Protein kinase domain</td>
<td>1apm</td>
<td>271 (230, 313)</td>
<td>24</td>
<td>67/441/6266</td>
<td>4863/792</td>
<td>0/0</td>
<td>9.2/8.9</td>
<td>0/0</td>
</tr>
<tr>
<td>Reverse transcriptase</td>
<td>1hmv</td>
<td>220 (110, 330)</td>
<td>46</td>
<td>178/320/1129</td>
<td>4008/664</td>
<td>1.9/1.9</td>
<td>9.5/9.8</td>
<td>0/0</td>
</tr>
<tr>
<td>Retroviral asparyl protease</td>
<td>1da</td>
<td>102 (86, 117)</td>
<td>78</td>
<td>53/102/171</td>
<td>1918/300</td>
<td>1.8/1.7</td>
<td>2.5/2.3</td>
<td>0/0</td>
</tr>
<tr>
<td>Fibronecictype III domain</td>
<td>1tf</td>
<td>85 (71, 99)</td>
<td>19</td>
<td>108/287/3073</td>
<td>1671/244</td>
<td>0.7/0.17</td>
<td>2.1/1.4</td>
<td>0/0</td>
</tr>
<tr>
<td>7 transmembrane receptor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(rhodopsin family)</td>
<td>1f88</td>
<td>268 (194, 342)</td>
<td>19</td>
<td>64/257/3249</td>
<td>4920/845</td>
<td>0.3/0.4</td>
<td>13.9/14.2</td>
<td>0/0</td>
</tr>
<tr>
<td>ABC transporter</td>
<td>1b0u</td>
<td>189 (129, 249)</td>
<td>26</td>
<td>63/3967/4263</td>
<td>3533/548</td>
<td>0.05/0.05</td>
<td>2.4/2.6</td>
<td>0/0</td>
</tr>
<tr>
<td>EGF-like domain</td>
<td>1apo</td>
<td>35 (28, 42)</td>
<td>34</td>
<td>87/4480/4839</td>
<td>683/78</td>
<td>0/0</td>
<td>0.08/0.08/20.2</td>
<td>0/0</td>
</tr>
<tr>
<td>Elongation factor Tu GTP binding domain</td>
<td>1etu</td>
<td>226 (117, 335)</td>
<td>45</td>
<td>224/1788/2066</td>
<td>3913/586</td>
<td>4.0/5.9</td>
<td>5.5/7.1</td>
<td>0/0</td>
</tr>
<tr>
<td>MarR family</td>
<td>1jgs</td>
<td>105 (92, 119)</td>
<td>22</td>
<td>44/204/214</td>
<td>1975/356</td>
<td>0/0</td>
<td>0/0</td>
<td>0/0</td>
</tr>
<tr>
<td>PAS domain</td>
<td>2phy</td>
<td>66 (64, 88)</td>
<td>18</td>
<td>68/463/874</td>
<td>1253/204</td>
<td>7.8/4.5</td>
<td>24.7/24.9</td>
<td>0/0</td>
</tr>
<tr>
<td>Ubiquitin family</td>
<td>1aar</td>
<td>75 (69, 81)</td>
<td>60</td>
<td>96/737/890</td>
<td>1424/218</td>
<td>0.1/0.1</td>
<td>3.6/3.4</td>
<td>0/0</td>
</tr>
</tbody>
</table>

To conclude, the overall bias was the smallest when the threshold was chosen to be one, even if the true threshold was greater than one. However, the greater the threshold, the smaller the number of free parameters. Therefore, the choice of the threshold should be done in accordance with the situation, as can be seen in the TIM barrel example below.

4 PRACTICAL RESULTS

This chapter demonstrates an example of the usage of the EEP method: the first example illustrates the ability of the ML and the EEP estimated models to separate the family members from the other sequences in 20 protein families; the second section studies the effect of the EEP threshold value on the log-odds scores in the TIM barrel family; the third example demonstrates the ability of the EEP method to recognize new TIM barrel sequences from the databases. The results of the database search are compared with ML, HMMER and BLAST searches from the same database. Additionally, a

4.1 Application to 20 protein families

The profile HMM explained in Section 2.1 was fitted to 20 protein families chosen from the Pfam database (Bateman et al., 2002). By means of positive and negative test sets the ability of the model to distinguish family members from the other sequences was studied. The percentage of false positive and negative sequences as well as the number of parameters in the models were compared between the ML and EEP estimated models.

The families chosen in the training sets represent various superfamilies in the SCOP database (Murzin et al., 1995). The average length of the training sequences varied between 35 and 380 amino acids, and the sequence identities were between 15 and 78% (see Table 2). Both the multiple alignments of the training sets and the unaligned sequences of the positive test sets were downloaded from the Pfam database. The sequences in the ‘seed’ and the ‘full’ data sets in the Pfam database were chosen as a training and a positive test sets, respectively. Negative test sets included 974 sequences of the Astral SCOP database representing each superfamily in the SCOP database (Chandonia et al., 2002). Additionally, a
subset of the positive test set was examined. This set consists of the sequences whose lengths were within the 99% confidence interval of the length of the training sequences. The confidence limits of sequence lengths and the number of sequences in the training and positive test sets are shown in Table 2.

The estimation of the 20 families was based on the multiple alignment. After deciding which of the columns were match and insert states, transition probabilities were calculated directly from the alignment. Emission probabilities were estimated by using ML and EEP estimation methods, where \( a = 2 \) was used as a threshold. After all model parameters were estimated, log-odds scores were calculated for the training sequences, the positive and the negative test sets. The percentages of false negative and false positive sequences when the log-odds score threshold value was chosen to be one are reported in Table 2.

In the length restricted test set, over 95% of the family members were detected in 17 families with the ML method and in 18 families with the EEP. In the whole test set, 95% of the sequences were found only in 9 of the ML estimated and 10 of the EEP estimated families. The 90% level was reached by 15 and 16 families of the ML and EEP estimated models, respectively. The error rates were almost identical in the ML and EEP methods; however, in the \( \alpha \)-amylase catalytic domain, multicopper oxidase, glutamine amidotransferase, and PAS domain the false negative error rates were considerably greater in the ML method than in the EEP method. Only in the elongation factor Tu GTP binding domain were the false negative error rates slightly greater in the EEP method than in the ML method. The reason why some of the sequences in the whole positive test set were not recognized was that because these sequences were much longer/shorter than the training sequences, they had several visits in the insert/delete states, which were not constantly used by the training sequences. Therefore, the log-odds scores were reduced under zero. The false positive error rates were zero in most example families.

Although the classification rates of the two methods were very similar, there was great difference in degrees of freedom (see Table 2); for example in the \( \alpha \)-amylase catalytic domain the number of free parameters was decreased from 6839 to 1168. The largest difference was shown in the EGF-like domain, in which the number of free parameters dropped 89% when the EEP method was used instead of the ML method. This indicates that the number of free parameters was significantly reduced when the EEP instead of the ML method was used, whereas the prediction power usually remained at the same level or was even improved.

4.2 Selecting EEP threshold

In order to select the appropriate threshold value for the EEP estimation of the TIM barrel (triosephosphate isomerase) family, the emission probabilities were estimated, in addition to
Table 3. TIM barrel sequences found from the SWISS-PROT database in ML, EEP, HMMER and Blast searches

<table>
<thead>
<tr>
<th>Entry name</th>
<th>Sequence length</th>
<th>ML log-odds</th>
<th>EEP log-odds</th>
<th>HMMER log-odds</th>
<th>HMMER E-value</th>
<th>BLAST E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TPIS ANASP</td>
<td>241</td>
<td>175.1</td>
<td>196.8</td>
<td>321.7</td>
<td>1.3e−103</td>
<td>2.5e−36</td>
</tr>
<tr>
<td>TPIS ARCFU</td>
<td>223</td>
<td>—</td>
<td>—</td>
<td>5.2</td>
<td>1.3</td>
<td>1.4 (0.003, 9.9)</td>
</tr>
<tr>
<td>TPIS BACHD</td>
<td>251</td>
<td>371.0</td>
<td>354.3</td>
<td>528.6</td>
<td>6.2e−171</td>
<td>1.85e−45</td>
</tr>
<tr>
<td>TPIS BRUME</td>
<td>254</td>
<td>201.7</td>
<td>193.6</td>
<td>347.2</td>
<td>6.7e−112</td>
<td>3.25e−36</td>
</tr>
<tr>
<td>TPIS CAUCR</td>
<td>253</td>
<td>161.9</td>
<td>171.9</td>
<td>284.1</td>
<td>2.2e−91</td>
<td>4.5e−30</td>
</tr>
<tr>
<td>TPIS CLOPA</td>
<td>19</td>
<td>8.6</td>
<td>7.7</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TPIS CLOPE</td>
<td>248</td>
<td>293.9</td>
<td>313.4</td>
<td>458.7</td>
<td>3.4e−148</td>
<td>—</td>
</tr>
<tr>
<td>TPIS ENTTHI</td>
<td>261</td>
<td>310.4</td>
<td>299.2</td>
<td>466.7</td>
<td>8.3e−151</td>
<td>4.5e−48</td>
</tr>
<tr>
<td>TPIS LACDL</td>
<td>252</td>
<td>269.3</td>
<td>281.4</td>
<td>451.6</td>
<td>8.6e−146</td>
<td>3.5e−45</td>
</tr>
<tr>
<td>TPIS LISIN</td>
<td>251</td>
<td>319.2</td>
<td>314.4</td>
<td>478.3</td>
<td>1.4e−154</td>
<td>2e−43</td>
</tr>
<tr>
<td>TPIS LISMO</td>
<td>251</td>
<td>322.7</td>
<td>312.7</td>
<td>479.1</td>
<td>8e−155</td>
<td>4e−43</td>
</tr>
<tr>
<td>TPIS METTH</td>
<td>229</td>
<td>—</td>
<td>—</td>
<td>5.5</td>
<td>0.97</td>
<td>1.25 (0.0009, 8.6)</td>
</tr>
<tr>
<td>TPIS MYCPU</td>
<td>248</td>
<td>231.3</td>
<td>222.2</td>
<td>359.9</td>
<td>4.9e−116</td>
<td>8e−40</td>
</tr>
<tr>
<td>TPIS NEIMA</td>
<td>251</td>
<td>155.7</td>
<td>181.5</td>
<td>283.5</td>
<td>3.4e−91</td>
<td>6.5e−37</td>
</tr>
<tr>
<td>TPIS NEIMB</td>
<td>251</td>
<td>153.8</td>
<td>178.9</td>
<td>288.3</td>
<td>9.4e−93</td>
<td>2.5e−36</td>
</tr>
<tr>
<td>TPIS NICPL</td>
<td>10</td>
<td>16.7</td>
<td>13.4</td>
<td>8.9</td>
<td>0.079</td>
<td>—</td>
</tr>
<tr>
<td>TPIS PIPNS</td>
<td>17</td>
<td>9.3</td>
<td>9.3</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>TPIS PYRFU</td>
<td>227</td>
<td>—</td>
<td>—</td>
<td>3.1</td>
<td>6.3</td>
<td>0.95 (0.001, 6.9)</td>
</tr>
<tr>
<td>TPIS RALSO</td>
<td>248</td>
<td>239.0</td>
<td>251.7</td>
<td>396.2</td>
<td>7.3e−128</td>
<td>2.5e−42</td>
</tr>
<tr>
<td>TPIS RHIL0</td>
<td>256</td>
<td>219.2</td>
<td>210.7</td>
<td>342.5</td>
<td>2.2e−110</td>
<td>6.5e−36</td>
</tr>
<tr>
<td>TPIS RHIME</td>
<td>256</td>
<td>211.3</td>
<td>197.9</td>
<td>348.2</td>
<td>3.1e−112</td>
<td>1.35e−38</td>
</tr>
<tr>
<td>TPIS RALTI</td>
<td>255</td>
<td>386.8</td>
<td>354.9</td>
<td>538.8</td>
<td>2.8e−174</td>
<td>2e−44</td>
</tr>
<tr>
<td>TPIS RALTY</td>
<td>255</td>
<td>393.0</td>
<td>358.7</td>
<td>543.1</td>
<td>1.1e−175</td>
<td>2e−44</td>
</tr>
<tr>
<td>TPIS STAAM</td>
<td>253</td>
<td>300.0</td>
<td>290.4</td>
<td>457.1</td>
<td>1.1e−147</td>
<td>8.5e−45</td>
</tr>
<tr>
<td>TPIS STRPN</td>
<td>252</td>
<td>274.9</td>
<td>281.8</td>
<td>443.6</td>
<td>2.7e−143</td>
<td>2.5e−41</td>
</tr>
<tr>
<td>TPIS SUSLO</td>
<td>227</td>
<td>—</td>
<td>19.4</td>
<td>—</td>
<td>3e−05</td>
<td>0.0235 (1e−6, 6.6)</td>
</tr>
<tr>
<td>TPIS SULTO</td>
<td>229</td>
<td>—</td>
<td>10.0</td>
<td>0.034</td>
<td>0.14 (0.00007, 3.1)</td>
<td></td>
</tr>
<tr>
<td>TPIS XYLFA</td>
<td>249</td>
<td>195.7</td>
<td>221.9</td>
<td>356.4</td>
<td>6.6e−115</td>
<td>9.5e−38</td>
</tr>
<tr>
<td>TPIS YERPE</td>
<td>255</td>
<td>355.3</td>
<td>326.2</td>
<td>516.4</td>
<td>5.7e−167</td>
<td>5.5e−44</td>
</tr>
</tbody>
</table>

When the log-odds score threshold was set to 5, both the ML and EEP methods found 90 TIM barrel sequences. Of these sequences 24 (shown in Table 3) were not included in the data set obtained from the Pfam database. With the chosen log-odds score threshold, no false positive sequences were chosen, but 15 TIM barrels obtained a score that was under the threshold value. Figure 3 shows the log-odds scores of the EEP estimated model for all known TIM barrel sequences in the database.

The HMMER search with default options (E-value cutoff 10) suggested 110 new TIM barrel sequences. Of these sequences 15 were non-TIM barrels, and 27 were new TIM barrels, that is, not already included in the Pfam database (see Table 3). Furthermore, 10 known TIM barrel sequences were not detected. Blast search found all known TIM barrel sequences in the database, but was not capable of recognizing sequence fragments. In addition, 18 non-TIM barrels obtained by Blast an E-value by which they were classified as TIM barrels.

There were five sequences which were recognized as TIM barrels by the HMMER and the Blast but not by the ML and the EEP methods. The ML and the EEP found two sequences which were not detected by the other methods. These sequences were fragments whose lengths were only 17 and 19. Additionally, in the EEP method the log-odds score of TPIS_CANFA, whose length was only 5, was more than...
Fig. 4. The mean proportions of the ineffective amino acids calculated over all match states when $a = 2$ was used as a classification threshold. The upper line corresponds to the background distribution and the other lines to 20 protein families: α-amino acid kinase family, α-amylase, multicopper oxidase, globin, glutamine amidotransferase class II, glyceroldehyde 3-phosphate dehydrogenase, haloacid dehalogenase-like hydrolase, triosephosphate isomerase, immunoglobulin domain, protein kinase domain, reverse transcriptase, retroviral aspartyl protease, fibronectin type III domain, 7 transmembrane receptor, ABC transporter, EGF-like domain, elongation factor Tu GTP binding domain, MarR family, PAS domain and ubiquitin family.

5 DISCUSSION AND CONCLUSIONS

The profile HMMs are based on modelling the conserved positions of the multiple sequence alignment. In the conserved positions, however, the model does not separate the signal and the noise, but treats them in the same manner. The overfitting problem is typically avoided by replacing the observed amino acid distribution by its estimator, which is sometimes called regularizer (Karplus, 1995, ftp://ftp.cse.ucsc.edu/pub/tr/ucsc-crl-95-11.ps.Z). The regularizers are usually generated by two alternative approaches. Gribskov profile, average score method, and feature alphabets rely on substitution matrices or known structural environments (Tatusov et al., 1994; Gribskov et al., 1987), whereas simple pseudocounts and Dirichlet mixture distribution are generated by using statistical techniques for columns from a variety of multiple alignments (Sjölander, 1996). Both of these approaches incorporate some underlying characteristics of the alignment environments into the model.

In this paper, we introduce a new method to overcome the overfitting problem in HMMs. In our approach, amino acids are first divided into effective and ineffective and then estimated by assuming that the ineffective residues follow a background distribution. Thereby the EEP method explicitly models the conserved residues in the alignment instead of only considering the general characteristics of the amino acids. Moreover, it provides a flexible procedure for combining the methods scoring the residue conservation (Valdar, 2002) with the profile HMM.

The major advantage of the EEP method is the decrease in the dimension of the parameter space. In protein sequence alignments, the decrease is significant because in conserved positions only a few residues can be considered as effective. The study with 20 well-defined protein families indicates that the EEP method is able to detect sequences on average with 98% sensitivity and 99% specificity. The sensitivity proved to be better than in the ML method, even if the number of emission parameters was reduced to even 11% of the original.

The decrease in the number of parameters is of particular importance when using likelihood-based statistical methods. For example, in the long profile HMMs, Akaike information criteria cannot be adequately used for model selection because the likelihood of the model does not increase in the same proportion as the number of the parameters. In the Bayesian statistics, on the contrary, the large number of parameters is technically more easily controlled, and thus, it is typically not considered a problem. However, there is no reason why the EEP method could not be used in conjunction with the Bayesian approach by first choosing the effective residues, and thereafter defining the prior distribution only for the effective residues.

As a consequence of the reduction of the parameter space, the variance of the ineffective residues decreases without influencing variance of the effective residues. This improvement is significant when calculating confidence
intervals for emission probabilities. In practice, shortening the confidence intervals improves the prediction power of the model and thus results in more sensitive database search results. At this moment, the results in the small test sets suggest that the EEP method is a potent method for the database search, but further research is needed.

When using EEP method for the database search, one has to choose the threshold value for the estimation. The simulations with several threshold values showed that, because sensitivity is usually a more important characteristic than specificity, the estimates are accurate, even though the threshold value is less than the ‘true’ threshold. Furthermore, even if the estimates of the ineffective residues are inaccurate, it does not affect the model’s ability to detect the members of protein family.

In this paper, the alignment was supposed to be known, and thus, the method greatly reminds the ordinary profile analysis (Gribskov et al., 1987, 1990). Therefore, the time-consuming Baum–Welch iteration was not applied. In general, the EEP method can be generalized for situations where the appropriate alignment is not available, but it is found by iterating the HMM. In this case, the EEP method should additionally reduce the computational effort demanded by the Baum–Welch algorithm.

In this paper, we used background distribution as a criterion to divide residues into effective and ineffective ones. The major disadvantage of this approach is its inability to take into account the physical and chemical characteristics of the amino acids, and thus, it ignores the relationships among the amino acids. Therefore, future work in this area should be focused on developing methods that can reliably identify conserved residues in multiple sequence alignments.

ACKNOWLEDGEMENTS

The authors are grateful to referees for useful comments. This work was supported by a grant from graduate school in Computational Biology, Bioinformatics, and Biometry (ComBi). Financial support from the Academy of Finland, National Technology Agency of Finland, and The Medical Research Fund of Tampere University Hospital is gratefully acknowledged.

REFERENCES


APPENDIX

The Lagrange multipliers method was used to discover the estimation formulae for the emission probabilities. By using
the scaling equation $b_j = b_e b_j^0 / b_e^0$ for all $j \in J_2$ with a fixed $e \in J_2$, the two other constraints defined in Section 2.3 can be rewritten as

$$\sum_{j \in J_1} b_j \leq \frac{b_e}{b_e^0} \sum_{j \in J_1} b_j^0,$$

$$\sum_{j \in J_1} b_j + \frac{b_e}{b_e^0} \sum_{j \in J_2} b_j^0 = 1,$$

and the Lagrange function for the log-likelihood (2) subject to the above equality and inequality constraints take the form

$$Q = \sum_{j \in J_1} n_j \log b_j + \sum_{j \in J_2} n_j \log \frac{b_j^0}{b_e^0} + \sum_{j \in J_2} n_j \log b_e$$

$$+ \mu \left( \sum_{j \in J_1} b_j - c \frac{b_e}{b_e^0} \sum_{j \in J_1} b_j^0 \right)$$

$$+ \lambda \left( 1 - \sum_{j \in J_1} b_j - b_e \frac{b_e}{b_e^0} \sum_{j \in J_2} b_j^0 \right).$$

Here, $\mu \geq 0$ and $\lambda \in \mathbb{R}$ are the undetermined Lagrange multipliers. The restricted maximization of the likelihood with respect to $b_j$ can be done by solving the following set of equations, known as the Karush–Kuhn–Tucker optimality conditions:

$$\frac{\partial}{\partial b_j} Q = \frac{n_j}{b_j} - \lambda + \mu = 0, \quad \text{for } j \in J_1,$$

$$\frac{\partial}{\partial b_e} Q = \sum_{j \in J_2} \frac{n_j}{b_j} - \lambda \sum_{j \in J_1} \frac{b_j}{b_j^0} - \mu c \sum_{j \in J_1} \frac{b_j^0}{b_e^0} = 0,$$

$$\mu \left( \sum_{j \in J_1} b_j - c \frac{b_e}{b_e^0} \sum_{j \in J_1} b_j^0 \right) = 0,$$

$$1 - \sum_{j \in J_1} b_j - b_e \frac{b_e}{b_e^0} \sum_{j \in J_2} b_j^0 = 0.$$