A Maximum-Likelihood Formulation and EM Algorithm for the Protein Multiple Alignment Problem

Valentina Sulimova[1], Nikolay Razin[2], Vadim Mottl[3],
Ilya Muchnik[4], and Casimir Kulikowski[5]

[1] Tula State University, Lenine Ave. 92, 300600, Russia, Tula
[2] MIPT, Kerchenskaya St.1A, 117303, Russia, Moscow
[3] Computing Center of the RAS, Vavilov St.40, 119333, Russia, Moscow
[4] Rutgers University, DIMACS, New Brunswick, NJ 08901
[5] Rutgers University, Department of Computer Science, New Brunswick, NJ 08901
vsulimova@yandex.ru, nrmanutd@gmail.com, vmottl@yandex.ru,
muchnikilya@yahoo.com, kulikows@cs.rutgers.edu

Abstract. A given group of protein sequences of different lengths is considered as resulting from random transformations of independent random ancestor sequences of the same preset smaller length, each produced in accordance with an unknown common probabilistic profile. We describe the process of transformation by a Hidden Markov Model (HMM) which is a direct generalization of the PAM model for amino acids. We formulate the problem of finding the maximum likelihood probabilistic ancestor profile and demonstrate its practicality. The proposed method of solving this problem allows for obtaining simultaneously the ancestor profile and the posterior distribution of its HMM, which permits efficient determination of the most probable multiple alignment of all the sequences. Results obtained on the BAliBASE 3.0 protein alignment benchmark indicate that the proposed method is generally more accurate than popular methods of multiple alignment such as CLUSTALW, DIALIGN and ProbAlign.

Keywords: Multiple alignment problem, protein sequences analysis, EM-algorithm, HMM, common ancestor.

1 Introduction

The problem of multiple alignment of protein sequences is a fundamental problem for modern bioinformatics. It arises from applications such as secondary and tertiary structure prediction [1], reconstructing complex evolutionary histories [2, 3], locating conserved motifs and domains [4], and constructing phylogenetic trees [5].

The bioinformatics literature is replete with diverse alignment methods and tools. However, only few of them, such as multidimensional dynamic programming [6], have a mathematically strict problem formulation followed by a sound

T.M.H. Dijkstra et al. (Eds.): PRIB 2010, LNBI 6282, pp. 171–182, 2010.

optimization procedure. Those with mathematical formulations which try to take into account information about protein evolution [7] are NP-hard and cannot be applied for aligning more than a few sequences [8]. Approximations which are not based on evolutionary trees and stars [9] and other fast heuristics, such as approaches like those which include a large family of progressive alignments [10, 11], are less biologically relevant.

Profile-based algorithms with iterative updating [12] and HMM-based approaches [13–16] have an essential common disadvantage: their results strongly depend on the initial approximation. An additional problem which is typical for HMM-based multiple alignments is that of deciding on how to select model parameters.

In this paper, we consider a new approach to the problem of multiple alignment on the basis of the simplest probabilistic model of protein evolution built as a relatively straightforward generalization of Margaret Dayhoff's PAM model (Point Accepted Mutation) developed for the alphabet of single amino acids $A = (\alpha^1 \dots \alpha^{20})$ [17]. It is assumed that the amino acid sequences $\boldsymbol{\omega_j} = (\omega_{jt} \in A,\, t = 1 \dots N_j)$ forming the set to be processed jointly $\Omega^* = \{\boldsymbol{\omega}_j, j = 1 \dots M\}$ are results of independent random Markov chains of insertions/substitutions applied to some unknown n-length ancestor sequences $\boldsymbol{\vartheta_j} = (\vartheta_{ji}, i = 1 \dots n), j = 1 \dots M$, specific for each $\boldsymbol{\omega}_j$ of greater length, $n \leq \min\{N_j, j = 1 \dots M\}$. The elements of the hidden sequences ϑ_{ji} are a priori assumed to be randomly and independently chosen by nature according to a sequence of n unknown probability distributions over the set of 20 amino acids $\vartheta_i \in A$.

The goal of the analysis is to estimate these probability distributions as the sought-for n-length profile playing the role of a model of the given protein set.

Such a result is not in itself a multiple alignment, but any instance of the j-th insertion/substitution transformation cuts out a n-length subsequence from the corresponding amino acid sequence $\boldsymbol{\omega}_j = (\dots \tilde{\omega}_{jt_1} \dots \tilde{\omega}_{jt_i} \dots \tilde{\omega}_{jt_n} \dots)$, which is associated with the successive elements of the supposed ancestor $(1 \dots n)$. This process will generate a vast diversity of versions of how these positions could be assembled into n relatively conserved columns.

The algorithm yields the posterior distribution over the set of possible multiple alignments relevant to the given set of proteins, covering the large variety of versions of how these positions can lead to n relatively conserved columns. So we can easily find the most probable multiple alignment.

2 Dayhoff's PAM Model of Evolution within the Amino Acid Alphabet

The formulation of the multiple alignment problem considered in the present paper is based on the pioneering model of amino acid evolution Point Accepted Mutation (PAM) introduced by M. Dayhoff in 1978 [17]. The PAM model represents predispositions of amino acids towards mutual mutative transformations as a square matrix of conditional probabilities that amino acid α^i will be substituted at the next step of evolution by amino acid α^j :

$$\boldsymbol{\Psi} = \big(\psi(\alpha^j|\alpha^i), \alpha^i, \alpha^i \in A\big)(20 \times 20), \sum_{\alpha^j \in A} \psi(\alpha^j|\alpha^i) = 1. \tag{1}$$

The main probabilistic assumption underlying the PAM model is that the Markov chain defined by the transition matrix $\boldsymbol{\Psi}$ possesses the two classical properties:

- ergodicity, namely, existence of a final probability distribution over the set of amino acids $\xi(\alpha^j) = \sum_{\alpha^i \in A} \xi(\alpha^i)\psi(\alpha^j|\alpha^i)$,
- and reversibility $\xi(\alpha^i)\psi(\alpha^j|\alpha^i) = \xi(\alpha^j)\psi(\alpha^i|\alpha^j)$.

3 Model of the Common Origin of a Set of Proteins

Let Ω be the set of all finite amino acid sequences $\boldsymbol{\omega} = (\omega_t, t = 1, \ldots, N)$, $\omega_t \in A = \{\alpha^1, \ldots, \alpha^{20}\}$. We shall use also the notation $\Omega_n = \{\boldsymbol{\omega} = (\omega_t, t = 1, \ldots, N), \omega_t \in A, N = n\}$ for the set of all sequences having a fixed length n.

We proceed from the following probabilistic assumptions on the common origin of the proteins to be analyzed jointly $\Omega^* = \{\boldsymbol{\omega}_j, N_j \geq n, j = 1, \ldots, M\}$. These assumptions are essentially based on those taken in [18], aimed at an evolution-based pairwise comparison of proteins. On the one hand, we simplify them, because we use here only one particular class of described in [18] random transformations of sequences. But, on the other hand, we generalize this model because several amino acid sequences can be jointly processed here instead of just two.

Hypothesis 1. *Each of the amino acid sequences in the given set* $\Omega^* = \{\boldsymbol{\omega}_j = (\omega_{jt}, t = 1, \ldots, N_j), j = 1, \ldots, M\}$ *is considered as having evolved from its specific hidden ancestor* $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1, \ldots, n) \in \Omega_n$ *through independent known random transformations represented by the family of conditional probability distributions* $\varphi_{jn}(\boldsymbol{\omega}|\boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{jn}(\boldsymbol{\omega}|\boldsymbol{\vartheta}_j) = 1$.

Hypothesis 2. *Let the length* n *of the random ancestors* $\boldsymbol{\vartheta}_j \in \Omega_n$ *be fixed, and their elements* ϑ_{ji} *be drawn from the alphabet of amino acids in accordance with a common sequence of unknown independent probability distributions* $\big(\beta_i(\vartheta), \vartheta \in A\big)$, $\sum_{\vartheta \in A} \beta_i(\vartheta) = 1$.

Each of these distributions is completely represented by a 20-dimensional vector of probabilities $\boldsymbol{\beta}_i = (\beta_i^1, \ldots, \beta_i^{20}) \in \mathbb{R}^{20}$, $\sum_{k=1}^{20} \beta_i^k = 1$. It should be noticed that the sequence of distributions $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_i, i = 1, \ldots, n) \in \mathbb{R}^{20n}$ corresponds to the notion of the probabilistic profile, which is commonly adopted in bioinformatics.

This profile is the common parameter of identical independent probability distributions of the hidden ancestors $\boldsymbol{\vartheta}_j$ for each of the observed amino acid sequences :

$$p_n(\boldsymbol{\vartheta}_j|\bar{\boldsymbol{\beta}}) = p_n(\vartheta_{j1},\dots,\vartheta_{jn}|\boldsymbol{\beta}_1,\dots,\boldsymbol{\beta}_n) = \prod_{i=1}^{n}\beta_i(\vartheta_{ji}). \tag{2}$$

So, it is assumed here that the entire given set of amino acid sequences $\Omega^* = \{\boldsymbol{\omega}_j, N_j \geq n, j = 1,\dots,M\}$ has evolved from the same hidden profile $\bar{\boldsymbol{\beta}}$.

Hypothesis 3. *The transformation $\varphi_{Nn}(\boldsymbol{\omega}|\boldsymbol{\vartheta})$ of the n-length ancestor $\boldsymbol{\vartheta}_j \in \Omega_n$ into some random protein $\boldsymbol{\omega}_j$ of random length $N_j \geq n$ is a concatenation of the two following random mechanisms.*

The first step of the transformation is a random choice of the structures $\boldsymbol{v} = (1 \leq v_1 \leq \cdots \leq v_n)$ of transformations independently for each of the resulting sequences $\boldsymbol{\vartheta} \to \boldsymbol{\omega}$, $v_n \leq N$, namely, assigning the positions $\boldsymbol{\omega} = (\dots\bar{\omega}_{v_1}\dots\bar{\omega}_{v_i}\dots\bar{\omega}_{v_n}\dots)$ into which the elements of the ancestor $\boldsymbol{\vartheta} = (\vartheta_1,\dots,\vartheta_n)$ will be mapped. These positions are called in [18] *key positions.* The apriori distributions of the respective key-position vectors $q_{Nn}(\boldsymbol{v}) = q_{Nn}(v_1,\dots,v_n)$ are assumed to take into account only the gaps between the key positions $v_i - v_{i-1}$ and be indifferent to the lengths of both tails v_1 and $N - v_n$. Distributions $q_{Nn}(\boldsymbol{v})$ are necessarily specific for each of the lengths $N_j, j = 1,\dots,M$, because of the constraints $v_n \leq N_j$:

$$\begin{aligned} &q_{N_jn}(\boldsymbol{v}|a,b) = \begin{cases} \propto \prod_{i=2}^{n} g(v_i - v_{i-1}|a,b), v_n \leq N_j, \\ = 0, v_n > N_j, \end{cases} \\ &g(v_i - v_{i-1}|a,b) \propto \begin{cases} 1, d_i = v_i - v_{i-1} = 1, \\ \exp\left[-c(a + b(v_i - v_{i-1}))\right], d_i > 1, \end{cases} \\ &a > 0, b > 0, c > 0. \end{aligned} \tag{3}$$

Such a distribution ranks one long gap as more preferable than several short ones adding up to the same length.

The second step is filling the key positions in the resulting sequences with random amino acids in accordance with Dayhoff's conditional mutation probabilities $\psi(\omega_{v_i}|\vartheta_i)$ (1). The structure-dependent conditional transformation distributions are assumed to be completely uniform relative to amino acids in other positions:

$$\eta_n(\boldsymbol{\omega}|\boldsymbol{\vartheta},\boldsymbol{v}) \propto \prod_{i=1}^{n}\psi(\omega_{v_i}|\vartheta_i), \tag{4}$$

where $v \in \mathbb{V}_{Nn}$ for each specific $N = N_j$, and $\mathbb{V}_{Nn}$ is the set of all n-length transformation structures with respect to the length of the sequence $1 \leq v_1 < \cdots < v_n \leq N$.

It follows from Hypotheses 3 that each transformation $\boldsymbol{\vartheta} \to \boldsymbol{\omega} = \boldsymbol{\omega}_j, N = N_j$, is defined as the mixture

$$\varphi_{Nn}(\boldsymbol{\omega}|\boldsymbol{\vartheta}) = \sum_{v \in \mathbb{V}_{Nn}} q_{Nn}(\boldsymbol{v})\eta_n(\boldsymbol{\omega}|\boldsymbol{\vartheta}, \boldsymbol{v}), \boldsymbol{\omega} \in \Omega_N, \tag{5}$$

and, in accordance with Hypotheses 2, the marginal conditional distribution of the sequence of length N is expressed as

$$f_N(\boldsymbol{\omega}|\bar{\boldsymbol{\beta}}) = \sum_{v \in \mathbb{V}_{Nn}} q_{Nn}(\boldsymbol{v})\zeta_n(\boldsymbol{\omega}|\bar{\boldsymbol{\beta}}, \boldsymbol{v}), \boldsymbol{\omega} \in \Omega_N, \tag{6}$$

where

$$\zeta_n(\boldsymbol{\omega}|\bar{\boldsymbol{\beta}}, \boldsymbol{v}) = \sum_{\boldsymbol{\vartheta} \in \Omega_n} \eta_n(\boldsymbol{\omega}|\boldsymbol{\vartheta}, \boldsymbol{v})p_n(\boldsymbol{\vartheta}|\bar{\boldsymbol{\beta}}) \tag{7}$$

is the conditional distribution of a single random sequence with respect to the assumed structure $\boldsymbol{v} \in \mathbb{V}_{Nn}$ of its evolving from the unknown random ancestor of length n.

4 Maximum-Likelihood Estimation of the Common Profile

It follows from Hypothesis 1 that the joint distribution of independent sequences making the given set $\Omega^* = \{\boldsymbol{\omega}_j, j = 1 \dots M\}$ is the product of individual distributions (6)

$$F(\Omega^*|\bar{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j}(\boldsymbol{\omega}_j|\bar{\boldsymbol{\beta}}). \tag{8}$$

This is, in effect, a likelihood function with respect to the sought-for profile whose maximum-likelihood estimate will be given by the maximum point of this function:

$$\hat{\bar{\boldsymbol{\beta}}} = \arg\max_{\bar{\boldsymbol{\beta}}} \ln F(\Omega^*|\bar{\boldsymbol{\beta}}) = \arg\max_{\bar{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{v \in \mathbb{V}_{N_j n}} q_{N_j n}(\boldsymbol{v})\zeta_n(\boldsymbol{\omega}_j|\bar{\boldsymbol{\beta}}, \boldsymbol{v}). \tag{9}$$

The presence of a sum within the logarithm seems to hinder the maximization. But on the other hand, the set of sequences $\Omega^* = \{\boldsymbol{\omega}_j, j = 1 \dots M\}$ is the observable part of the two-component random object (Ω^*, Υ_n) whose hidden part $\Upsilon_n = (\boldsymbol{v}_j \in \mathbb{V}_{N_j n}, j = 1 \dots M)$ is the collection of the sequence-specific transformation structures.

This fact suggests the application of the Expectation-Maximization (EM) principle, which results, in this case, in the following iterative procedure $s = 1, 2, 3, \dots$, starting with an initial approximation $\bar{\boldsymbol{\beta}}_0 = (\boldsymbol{\beta}_{1,0}, \dots, \boldsymbol{\beta}_{n,0}) \subseteq \mathbb{R}^{20n}$.

Let $\bar{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s}, \dots, \boldsymbol{\beta}_{n,s})$ be approximation at step s, and

$$p_{it}(\bar{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) = P(v_{ij} = t|\bar{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) \tag{10}$$

be the a posteriori probability of the event $v_{ij} = t$ in the transformation structure $\boldsymbol{v}_j = (1 \le v_{j1} < \dots < v_{jn})$, which means that the i-th element $\boldsymbol{\beta}_{i,s}$ of the profile $\bar{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s}, \dots, \boldsymbol{\beta}_{n,s})$ is associated with the t-th element ω_{jt} of the j-th sequence $\boldsymbol{\omega}_j = (\omega_{j1}, \dots, \omega_{jN_j})$. The next value of the i-th element of the profile $\boldsymbol{\beta}_{i,s+1} = (\beta^1_{i,s+1}, \dots, \beta^{20}_{i,s+1}) \in \mathbb{R}^{20}$ is defined as

$$\begin{cases} (\beta^1_{i,s+1}, \dots, \beta^{20}_{i,s+1}) = \underset{(\beta^1_i, \dots, \beta^{20}_i) \in \mathbb{R}^{20}}{\arg\max} \sum\limits_{l=1}^{20} h^l_i \ln \sum\limits_{k=1}^{20} \psi(\alpha^l | \alpha^k) \beta^k_i, \\ \sum\limits_{k=1}^{n} \beta^k_i = 1, \, \beta^k_i \ge 0, \, k = 1, \dots, 20, \end{cases} \tag{11}$$

where $h^l_i = \sum\limits_{j=1}^{M} \sum\limits_{t=1}^{N_j} I[\omega_{jt} = \alpha^l] p_{it}(\bar{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j)$, and indicator function $I[\omega_{jt} = \alpha^l] = 1$ if the condition $\omega_{jt} = \alpha^l$ is met, or 0 if not. Solving this problem is provided by the well-known gradient projection algorithm.

Theorem 1. *The choice of* $\bar{\boldsymbol{\beta}}_{s+1} = (\boldsymbol{\beta}_{1,s+1} \cdots \boldsymbol{\beta}_{n,s+1})$ *in accordance with* (11) *provides that the inequality* $F(\Omega^* | \bar{\boldsymbol{\beta}}_{s+1}) > F(\Omega^* | \bar{\boldsymbol{\beta}}_s)$ *holds true at each step* s *while* $\nabla_{\bar{\boldsymbol{\beta}}} F(\Omega^* | \bar{\boldsymbol{\beta}}_s) \ne \mathbf{0}$ *; if* $\nabla_{\bar{\boldsymbol{\beta}}} F(\Omega^* | \bar{\boldsymbol{\beta}}_s) = \mathbf{0}$ *then* $F(\Omega^* | \bar{\boldsymbol{\beta}}_{s+1}) = F(\Omega^* | \bar{\boldsymbol{\beta}}_s)$.

Proof. The proof directly follows from the standard derivation and reasoning for EM algorithms.

Computation of posterior probabilities (10) is also a standard problem, in this case, in the theory of hidden Markov models, because the random transformation structure $\boldsymbol{v} = (1 \le v_1 < \dots < v_n)$ with independent gaps defined by (3) is a Markov process for each amino acid sequence in the data set under analysis $\Omega^* = \{\boldsymbol{\omega}_j, j = 1, \dots, M\}$.

5 Choosing Main Parameters of the Algorithm

The main parameters of the proposed algorithm are the length n of the common profile $\bar{\boldsymbol{\beta}} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_n)$ and the initial approximation for the profile $\bar{\boldsymbol{\beta}}_0 = (\boldsymbol{\beta}_{0,1}, \dots, \boldsymbol{\beta}_{0,n})$.

These parameters can be chosen by a number of different ways. For example it appears reasonable to take the value n which provides the minimum average entropy of the profile columns:

$$\hat{n} = \arg\min_n \Big(-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{n} \beta^k_i \ln \beta^k_i \Big). \tag{12}$$

This criterion satisfies the requirement of the final goal of the analysis, which is understood as finding the most conserved columns of amino acids in the given set of proteins.

When the likelihood function (8) has only one maximum, i.e., the set of its stationary points $\{\bar{\boldsymbol{\beta}} : \nabla_{\bar{\boldsymbol{\beta}}} F(\Omega^* | \bar{\boldsymbol{\beta}}) = \mathbf{0}\} \subseteq \mathbb{R}^{20n}$ is convex, the choice of the initial approximation $\bar{\boldsymbol{\beta}}_0 = (\boldsymbol{\beta}_{0,1}, \ldots, \boldsymbol{\beta}_{0,n})$ is not too significant. For instance, it is enough to take the sequence of uniform distributions over the set of amino acids $\boldsymbol{\beta}_{0,i} = (1/20, ..., 1/20) \in \mathbb{R}^{20}$, $i = 1, \ldots, n$.

However, when the sequences under analysis $\Omega^* = \{\boldsymbol{\omega}_j, j = 1 \ldots M\}$ have low identity, the likelihood function has a tendency to be not unimodal. In this paper, we choose both parameters n and $\bar{\boldsymbol{\beta}}_0 \in \mathbb{R}^{20n}$ at once by computing them using the multiple alignment obtained by some different method, for example ProbAlign. The number of columns without gaps in this alignment defines the length of the common profile n, and the distributions of amino acids in these columns are taken as the initial distributions $\boldsymbol{\beta}_{0,1}, \ldots, \boldsymbol{\beta}_{0,n}$. The efficiency of such approach is confirmed by results of experiments.

6 The Most Probable Multiple Alignment

The n-column profile $\hat{\bar{\boldsymbol{\beta}}}$ found as the maximum-likelihood estimate (9) of the fuzzy common subsequence of the assumed preset length n in the given set of proteins may be considered as the goal of their joint analysis. But the final a posteriori probabilities $p_{it}(\hat{\bar{\boldsymbol{\beta}}}, \boldsymbol{\omega}_j) = P(v_{ij} = t | \hat{\bar{\boldsymbol{\beta}}}, \boldsymbol{\omega}_j)$ (10) of the positions associated with each of the single amino acid sequences for successive elements of the supposed common ancestor $(1, \ldots, n)$ show a vast variety of versions of how these positions could be assembled into relatively conserved columns. This is the posterior distribution over the set of possible multiple alignments relevant to the given set of proteins.

The a posteriori most probable one will be given by the solutions of separate optimization problems corresponding to single proteins $\boldsymbol{\omega}_j, j = 1 \ldots M$:

$$\begin{cases} \boldsymbol{v}_j = \underset{v_1, \ldots, v_n}{\arg\max} \prod_{i=1}^{n} p_{iv_i}(\hat{\bar{\boldsymbol{\beta}}}, \boldsymbol{\omega}_j), \\ v_{ji} \geq v_{j,i-1}, i = 2 \ldots n. \end{cases} \tag{13}$$

This is a standard dynamic programming problem.

7 Experimental Results and Discussion

7.1 Characteristic Features of the Proposed Alignment and Its Visual Representation

It should be noted, that the form of multiple alignment obtained in accordance with (13) is different from the most conventional form of multiple alignment. The proposed approach actually produces only n columns without gaps, each of which corresponds to the respective i-th $(i = 1 \ldots n)$ element of the alleged common ancestor of the sequences. Other amino acids are not aligned. An example of a visual representation of a multiple alignment produced in accordance with our approach is presented in Figure 1,b. In contrast, Figure 1,a shows the traditional form of the benchmark multiple alignment produced by biologists.

The main part of our alignment in Figure 1,b is separated from the rest at the left and at the right by three empty columns, each of which contains only gaps. Left fragments of the sequences, which precede the main part, are flushed right, whereas right fragments, following the main part, are flushed left. Amino acids located between the ungapped aligned columns are conventionally flushed to the centers of idle intervals.

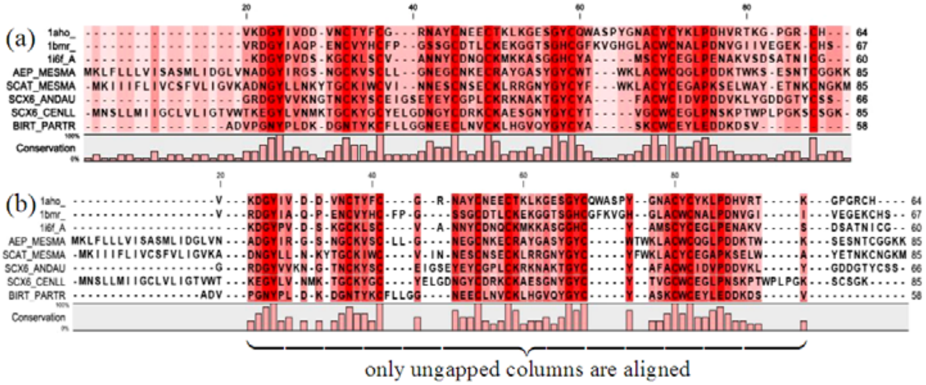

Fig. 1. Examples of multiple alignments: (a) manually-refined benchmark alignment and (b) alignment produced by the proposed approach

7.2 Alignment Benchmark

We tested our approach on a subset of BAliBASE 3.0 [20], which is the database of manually-refined multiple sequence alignments specifically designed for the evaluation and comparison of multiple sequence alignment programs.

For our tests we used families of short proteins from 3 different sets of BALiBase RV11, RV12 and RV20. The set RV11 contains equidistant families with sequence identity less than 20%, while RV12 contains equidistant families with sequence identity between 20% and 40%. Both of these sets lack sequences with large internal insertions (> 35 residues). The set RV20 contains families with > 40% similarity and an orphan sequence which shares less than 20% similarity with the rest of the family.

The main characteristics of the tested families are presented in Table 1.

7.3 Determining Prediction Accuracy

Given a true and an estimated multiple sequence alignment, the accuracy of the estimated alignment is usually computed using two measures: the sum-of-pairs (SP) and the true column (TC) scores. The SP score is a measure of the number of correctly aligned residue pairs divided by the number of aligned residue pairs in the true alignment, and TC is the number of correctly aligned columns divided by the number of columns in the true alignment. Both of them are standard measures of computing alignment accuracy. The source code of a

Table 1. Characteristics of the considered families

Set	Family name	Description	Number of sequences	Lengths	Number of columns without gaps in benchmark
RV11	1aab	high mobility group protein	4	83 – 91	76
	1aboA	SH3	8	52 – 193	47
	1bbt3	foot-and-mouth disease virus	6	186 – 283	150
	1csy	SH2	4	104 – 540	91
	1dox	ferredoxin [2fe-2s]	4	97 – 337	78
RV12	1axo	toxin II	8	58 – 85	51
	1fj1A	homeodomain	9	49 – 254	49
	1hfh	factor h	4	118 – 129	115
	1hpi	high-potential iron-sulfur protein	6	71 – 85	65
	1krn	serine protease	5	79 – 475	78
RV20	1idy	myb dna-binding domain	38	54 – 256	45
	1pamA	cyclodextrin	16	247 – 527	215
	1pgtA	glutathione	31	202 – 244	175
	1tvxA	pertussis toxin	29	64 – 167	50
	1ubi	ubiquitin	47	76 – 155	67

program for computing these scores is available for download at the BALiBase site [20]. However, this program is not accurate enough, it has a tendency to overstate the TC and SP scores and, moreover, to give values greater then 1, which is impossible given the definition of these scores.

In this connection, we use our implementation of the procedure for computing Bali-scores. It should be noticed that our procedure, in contrast to the original one, takes into account only pairs of amino acids which belong to the columns without gaps. This approach is much more appropriate for the principle of multiple alignment proposed in this paper but, as a rule, yields smaller values of scores.

7.4 Experimental Setup and Results

For each family under consideration, four multiple alignments were computed. Three of them were produced by the popular multiple alignment tools CLUSTALW, DI-ALIGN and ProbAlign, which were run on their respective servers. The value of the constant for the ProbAlign algorithm, called "the thermodynamic temperature", was chosen to be 5 as the most reasonable value according to publications [14]. The remaining parameters of this and other algorithms were set at their default values.

Finally, the 4-th alignment was produced in accordance with the proposed approach, started from the resulting alignment of ProbAlign as initial approximation.

The four-way comparison of SP and TC scores is presented in Table 2. The best values of scores are highlighted in bold font.

Table 2. Results of comparing multiple alignment procedures. TC/SP scores of multiple alignments produced by different algorithms.

Set	Family	CLUSTALW	DIALIGN	ProbAlign	The proposed approach
RV11	1aab	0.92/0.96	0.91/0.93	0.83/0.87	**0.99/0.99**
	1aboA	0.00/0.38	0.00/0.00	0.00/**0.54**	0.00/0.45
	1bbt3	0.00/0.20	0.00/0.00	**0.29/0.42**	0.28/0.36
	1csy	0.37/0.42	0.31/0.37	0.46/0.56	**0.51/0.56**
	1dox	0.00/0.24	0.40/0.46	0.62/0.71	**0.64/0.75**
RV12	1axo	0.29/0.54	0.54/0.64	0.69/0.87	**0.87/0.93**
	1fj1A	**1.00/1.00**	0.69/0.76	0.79/0.84	**1.00/1.00**
	1hfh	0.68/0.78	0.39/0.53	**0.78/0.85**	0.75/**0.85**
	1hpi	0.59/0.72	0.37/0.57	0.40/0.55	**0.75/0.82**
	1krn	0.53/0.69	0.47/0.68	0.60/0.75	**0.79/0.88**
RV20	1idy	0.00/**0.62**	0.00/0.00	0.00/0.33	0.00/0.60
	1pamA	0.43/0.77	0.29/0.58	**0.74/0.84**	0.69/0.83
	1pgtA	**0.47**/0.49	0.14/0.52	0.26/**0.69**	0.27/0.68
	1tvxA	0.00/**0.64**	0.00/0.00	0.00/0.41	0.00/0.46
	1ubi	0.00/**0.68**	0.00/0.03	**0.09**/0.49	0.08/0.48
	mean	0.35/0.61	0.30/0.41	0.44/0.65	**0.51/0.71**

As can be seen, in more than half of all the above cases our proposed approach yields the best results. The greatest success is achieved for families of the set RV12. But also for other families, the TC and SP scores of our approach are larger, in many cases, than scores of the main competitor ProbAlign. As a result, the average scores for the proposed approach are the best.

In addition, some interesting statistics computed from Table 2 are presented in Table 3 for comparing the proposed approach with the ProbAlign.

Table 3. Statistics computed from Table 2 for comparing the proposed approach with the ProbAlign

	TC / SP
The number of cases when our proposed approach is better or equal	11(73%) / 10(67%)
The mean increment of scores	0.112 / 0.127
The mean percentage increment of scores	23% / 21%
The mean decrement of scores	0.025 / 0.036
The mean percentage decrement of scores	6% / 7.1%

8 Conclusions

In this paper we have proposed and tested a new formulation of the multiple alignment problem. It is based on a deliberately simplified model of proteins evolution, which is a direct generalization of the PAM model for amino acids. For solving the respective optimization problem we have used an iterative procedure based on the EM-algorithm.

The first experiments show that the proposed approach outperforms other methods of multiple alignment by mean values of TC and SP scores. It does not yield the best scores for all considered cases, but it can be seen that, as a rule, our method shows small decreasing and large increasing of scores in contrast to other methods.

References

1. Rost, B., Sander, C., Schneider, R.P.: - an automatic server for protein secondary structure prediction. Computational Applications in Biosciences 10, 53–60 (1994)
2. Notredame, C.: Recent progresses in multiple sequence alignment: a survey. Pharmacogenomics 3(1), 131–144 (2002)
3. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, p. 356. Cambridge University Press, Cambridge (1998)
4. Attwood, T.K.: The PRINTS database: A resource for identification of protein families. Brief Bioinformatics 3, 252–263 (2002)
5. Saitou, N., Nei, M.: The neighbor-joining method: A new method for reconstructing phylo-genetic trees. Molecular Biology 212, 403–428 (1987)
6. Sankoff, D., Cedergren, R.J.: Simultaneous comparison of three or more sequences related by a tree. In: Sankoff, D., Kruskal, J.B. (eds.) Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison, pp. 253–263. Addison-Wesley, Reading (1989)
7. Altschul, S.F., Lipman, D.J.: Trees, stars, and multiple biological sequence alignment. SIAM J. Appl. Math. 49, 197–209 (1989)
8. Todd Wareham, H.: A simplified proof of the NP- and MAX SNP-hardness of multiple sequence tree alignments. J. Comput. Biol. 2(4), 509–514 (1995)
9. Carrillo, H., Lipman, D.: The multiple sequence alignment problem in biology. SIAM J. Appl. Math. 48, 1073–1082 (1988)
10. Notredame, C., Higgins, D.G., T-Coffee, H.J.: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205–217 (2000)
11. Subramanian, A.R., Kaufmann, M., Morgenstern, B.: DIALIGN-TX: Greedy and progres-sive approaches for segment-based multiple sequence alignment. Algorithms for Molecular Biology 3, 6 (2008)
12. Barton, G.J., Sternberg, M.J.E.: A strategy for the rapid multiple alignment of protein se-quences. J. Mol. Biol. 198, 327–337 (1987)
13. Durbin, R., Eddy, S., Krogh, A., Mitchison, G.: Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press, Cambridge (1998)
14. Roshan, U., Libesay, D.R.: Probalign: Multiple Sequence Alignment Using Partition Function Posterior Probabilities. Oxford University Press, Oxford (2005)
15. Do, C.B., Mahabhashyam, M.S., Brudno, M., Batzoglou, S.: ProbCons: Probabilistic Consis-tency-based Multiple Sequence Alignment. Genome Res. 15, 330–340 (2005)
16. Pei, J., Grishin, N.V.: PROMALS: Towards accurate multiple sequence alignments of dis-tantly related proteins. Bioinformatics 23, 802–808 (2007)

17. Dayhoff, M.O., Schwarts, R.M., Orcutt, B.C.: A model of evolutionary change in proteins. Atlas of Protein Sequences and Structures 5(suppl. 3), 345–352 (1978)
18. Sulimova, V., Mottl, V., Mirkin, B., Muchnik, I., Kulikowski, C.: A class of evolution-based kernels for protein homology analysis: A generalization of the PAM model. In: Proceedings of the 5th International Symposium on Bioinformatics Research and Applications, May 13-16, pp. 284–296. Nova Southeastern University, Ft. Lauderdale (2009)
19. Thompson, J.D., Koehl, P., Ripp, R., Poch, O.: BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins 61, 127–136 (2005)
20. BALiBASE3.0: A benchmark alignment database home page, http://www-bio3d-igbmc.u-strasbg.fr/~julie/balibase/index.html