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

Nikolay Razin
PhD Student, Moscow Institute of Physics and Technology, Russia

Valentina Sulimova
Assistant Professor, Tula State University, Russia

Vadim Mottl
Professor, Moscow Institute of Physics and Technology, Russia
Computing Center of the Russian Academy of Sciences

Ilya Muchnik, Casimir Kulikowski
Professors, Rutgers University, New Jersey, USA

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

Nikolay Razin
PhD Student, Moscow Institute of Physics and Technology, Russia

Valentina Sulimova
Assistant Professor, Tula State University, Russia

Vadim Mottl
Professor, Moscow Institute of Physics and Technology, Russia
Computing Center of the Russian Academy of Sciences

Ilya Muchnik, Casimir Kulikowski
Professors, Rutgers University, New Jersey, USA

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

Nikolay Razin
PhD Student, Moscow Institute of Physics and Technology, Russia

Valentina Sulimova
Assistant Professor, Tula State University, Russia

Vadim Mottl
Professor, Moscow Institute of Physics and Technology, Russia
Computing Center of the Russian Academy of Sciences

Ilya Muchnik, Casimir Kulikowski
Professors, Rutgers University, New Jersey, USA

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

Nikolay Razin
PhD Student, Moscow Institute of Physics and Technology, Russia

Valentina Sulimova
Assistant Professor, Tula State University, Russia

Vadim Mottl
Professor, Moscow Institute of Physics and Technology, Russia
Computing Center of the Russian Academy of Sciences

Ilya Muchnik, Casimir Kulikowski
Professors, Rutgers University, New Jersey, USA

Background

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

The main idea of our approach

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

The main idea of our approach

1. A simplest probabilistic model of protein evolution:
   relatively straightforward generalization the PAM model (developed by M. Dayhoff for the alphabet of single amino acids) onto amino acid sequences.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

The main idea of our approach

1. A simplest probabilistic model of protein evolution:
   relatively straightforward generalization the PAM model (developed by M. Dayhoff for the alphabet of single amino acids) onto amino acid sequences.
2. The amino acid sequences to be aligned are treated as results of independent random insertions/substitutions applied to random hidden ancestors of the same preset smaller length.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

The main idea of our approach

1. A simplest probabilistic model of protein evolution:
   relatively straightforward generalization the PAM model (developed by M. Dayhoff for the alphabet of single amino acids) onto amino acid sequences.
2. The amino acid sequences to be aligned are treated as results of independent random insertions/substitutions applied to random hidden ancestors of the same preset smaller length.
3. The immediate goal of the analysis is estimating the common probabilistic profile of the hidden ancestors as a sequence of independent probability distributions over the alphabet of amino acids.

Background

1. Only few of existing multiple alignment methods, such as multidimensional dynamic programming, are underlaid by a mathematically strict problem formulation.
2. However, mathematically sound methods are computationally too hard.
3. Fast heuristic algorithms are essentially less relevant from the biological point of view.

The main idea of our approach

1. A simplest probabilistic model of protein evolution:
   relatively straightforward generalization the PAM model (developed by M. Dayhoff for the alphabet of single amino acids) onto amino acid sequences.
2. The amino acid sequences to be aligned are treated as results of independent random insertions/substitutions applied to random hidden ancestors of the same preset smaller length.
3. The immediate goal of the analysis is estimating the common probabilistic profile of the hidden ancestors as a sequence of independent probability distributions over the alphabet of amino acids.
4. The algorithm yields the posterior distribution over the set of all multiple alignments. The most probable one of them is considered as the final result.

Margaret Dayhoff's PAM model of evolution within the amino acid alphabet

Margaret Dayhoff's PAM model of evolution within the amino acid alphabet

The set (alphabet) of amino acids: $A = \left\{ \alpha^1, ..., \alpha^{20} \right\}$

Margaret Dayhoff's PAM model of evolution within the amino acid alphabet

The set (alphabet) of amino acids: $A = \left\{ \alpha^1, ..., \alpha^{20} \right\}$

The PAM (Point Accepted Mutation) model of amino acid comparison represents predispositions of amino acids towards mutative transformations.

Margaret Dayhoff's PAM model of evolution within the amino acid alphabet

The set (alphabet) of amino acids: $A = \left\{ \alpha^1, ..., \alpha^{20} \right\}$

The PAM (Point Accepted Mutation) model of amino acid comparison represents predispositions of amino acids towards mutative transformations.

Markov chain of amino acid evolution represented by transition probabilities matrix for the accepted evolutionary step

$$\mathbf{\Psi} = \left(\psi(\alpha^j | \alpha^i), \alpha^i, \alpha^j \in A \right) (20 \times 20), \ \sum_{\alpha^j \in A} \psi(\alpha^j | \alpha^i) = 1 \text{ for all } \alpha^i \in A$$

Margaret Dayhoff's PAM model of evolution within the amino acid alphabet

The set (alphabet) of amino acids: $A = \left\{ \alpha^1, ..., \alpha^{20} \right\}$

The PAM (Point Accepted Mutation) model of amino acid comparison represents predispositions of amino acids towards mutative transformations.

Markov chain of amino acid evolution represented by transition probabilities matrix for the accepted evolutionary step

$$\mathbf{\Psi} = \left(\psi(\alpha^j | \alpha^i), \alpha^i, \alpha^j \in A \right) (20 \times 20), \ \sum_{\alpha^j \in A} \psi(\alpha^j | \alpha^i) = 1 \text{ for all } \alpha^i \in A$$

Dayhoff's main assumptions on the Markov chain:

– ergodicity, namely, existence of a final probability distribution over A

$$\xi(\alpha^j) = \sum_{\alpha^i \in A} \xi(\alpha^i) \psi(\alpha^j | \alpha^i) \text{ for all } \alpha^j \in A$$

– reversibility, namely, invariance to time inversion

$$\xi(\alpha^i) \psi(\alpha^j | \alpha^i) = \xi(\alpha^j) \psi(\alpha^i | \alpha^j) \text{ for all } \alpha^i, \alpha^j \in A$$

Notations

$A = \left\{ \alpha^1, ..., \alpha^{20} \right\}$ – the set (alphabet) of amino acids

$\boldsymbol{\omega} = (\omega_t \in A, t = 1, ..., N_{\boldsymbol{\omega}})$ – amino acid sequence of length $N_{\boldsymbol{\omega}}$

n – an integer called the order of the multiple alignment, namely, the assumed number of common columns .

$\Omega_{\geq n}$ – the set of all amino acid sequences of length $N_{\boldsymbol{\omega}} \geq n$

Ω_n – the set of all amino acid sequences of fixed length $N_{\boldsymbol{\omega}} = n$

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j, N_j \geq n, \ j = 1, ..., M \right\}$ – the given finite set of amino acid sequences to be aligned

Hypothesis 1

Each sequence in $\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1, \ldots, N_j \geq n) \in \Omega^*_{\geq n}$ has evolved from a specific ancestor $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1, \ldots, n) \in \Omega_n$ through independent known random transformation $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j) = 1$.

Hypothesis 1

Each sequence in $\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1,..., N_j \geq n) \in \Omega^*_{\geq n}$ has evolved from a specific ancestor $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1,..., n) \in \Omega_n$ through independent known random transformation $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j) = 1$.

Hypothesis 2

The length n of the random ancestors $\boldsymbol{\vartheta}_j \in \Omega_n$ is fixed, and their elements $(\vartheta_{j1},..., \vartheta_{jn})$ are drawn from the amino acid alphabet A in accordance with a common sequence of independent probability distributions

$$\left(\beta_i(\vartheta), i = 1,..., n\right),\ \vartheta \in A,\ \sum_{\vartheta \in A} \beta(\vartheta) = 1.$$

The sequence of these distributions forms a probabilistic profile of the “fuzzy” common ancestor

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,..., n\right).$$

Hypothesis 1

Each sequence in $\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1,..., N_j \geq n) \in \Omega^*_{\geq n}$ has evolved from a specific ancestor $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1,..., n) \in \Omega_n$ through independent known random transformation $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j) = 1$.

Hypothesis 2

The length n of the random ancestors $\boldsymbol{\vartheta}_j \in \Omega_n$ is fixed, and their elements $(\vartheta_{j1},..., \vartheta_{jn})$ are drawn from the amino acid alphabet A in accordance with a common sequence of independent probability distributions

$$\left(\beta_i(\vartheta), i = 1,..., n\right),\ \vartheta \in A,\ \sum_{\vartheta \in A} \beta(\vartheta) = 1.$$

The sequence of these distributions forms a probabilistic profile of the “fuzzy” common ancestor

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,..., n\right).$$

Let $\left(p_n(\boldsymbol{\vartheta} \mid \overline{\boldsymbol{\beta}}), \boldsymbol{\vartheta} \in \Omega_n \right)$ be the respective parametric distribution family.

Hypothesis 1

Each sequence in $\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1,..., N_j \geq n) \in \Omega^*_{\geq n}$ has evolved from a specific ancestor $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1,..., n) \in \Omega_n$ through independent known random transformation $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j) = 1$.

Hypothesis 2

The length n of the random ancestors $\boldsymbol{\vartheta}_j \in \Omega_n$ is fixed, and their elements $(\vartheta_{j1},..., \vartheta_{jn})$ are drawn from the amino acid alphabet A in accordance with a common sequence of independent probability distributions

$$\left(\beta_i(\vartheta), i = 1,..., n\right),\ \vartheta \in A,\ \sum_{\vartheta \in A} \beta(\vartheta) = 1.$$

The sequence of these distributions forms a probabilistic profile of the “fuzzy” common ancestor

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,..., n\right).$$

Let $\left(p_n(\boldsymbol{\vartheta} \mid \overline{\boldsymbol{\beta}}), \boldsymbol{\vartheta} \in \Omega_n \right)$ be the respective parametric distribution family.

The first intermediate goal of the analysis

For the accepted family of transformation distributions $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, it is required to estimate the common probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,..., n\right)$ of the preset length n.

Hypothesis 1

Each sequence in $\boldsymbol{\omega}_j = (\omega_{jt} \in A, t = 1,...,N_j \geq n) \in \Omega^*_{\geq n}$ has evolved from a specific ancestor $\boldsymbol{\vartheta}_j = (\vartheta_{ji} \in A, i = 1,...,n) \in \Omega_n$ through independent known random transformation $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, $\sum_{\boldsymbol{\omega} \in \Omega_{N_j}} \varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j) = 1$.

Hypothesis 2

The length n of the random ancestors $\boldsymbol{\vartheta}_j \in \Omega_n$ is fixed, and their elements $(\vartheta_{j1},...,\vartheta_{jn})$ are drawn from the amino acid alphabet A in accordance with a common sequence of independent probability distributions

$$\left(\beta_i(\vartheta), i = 1,...,n\right),\ \vartheta \in A,\ \sum_{\vartheta \in A} \beta(\vartheta) = 1.$$

The sequence of these distributions forms a probabilistic profile of the “fuzzy” common ancestor

$$\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,...,n\right).$$

Let $\left(p_n(\boldsymbol{\vartheta} \mid \overline{\boldsymbol{\beta}}), \boldsymbol{\vartheta} \in \Omega_n\right)$ be the respective parametric distribution family.

The first intermediate goal of the analysis

For the accepted family of transformation distributions $\varphi_{N_j}(\boldsymbol{\omega} \mid \boldsymbol{\vartheta}_j)$, it is required to estimate the common probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20}, i = 1,...,n\right)$ of the preset length n.

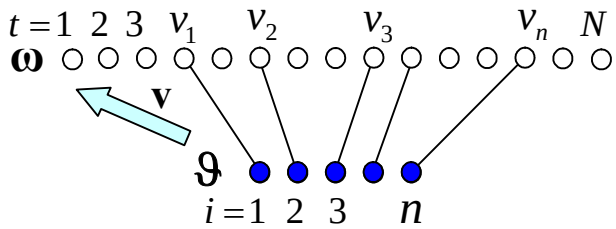
The final multiple alignment

Combination of individual pair-wise alignments of the given sequences with the found common profile.

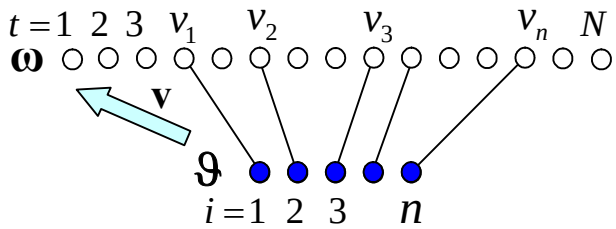
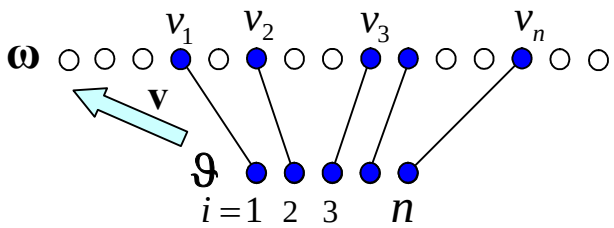
Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega|\vartheta_j)$:

Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega|\vartheta_j)$: Three constituents

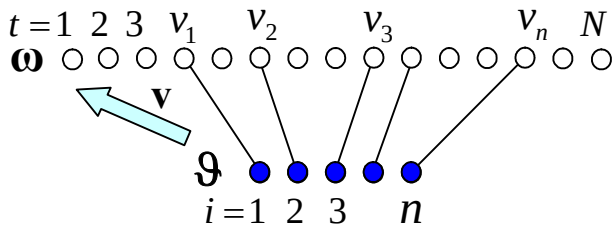
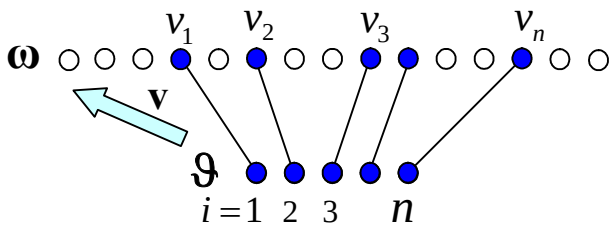
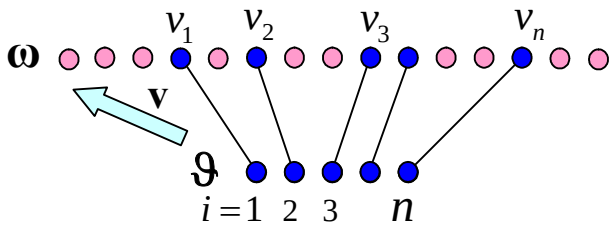
Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega|\vartheta_j)$: Three constituents

1. Random structure of the transformation: Unilateral alignment of the ancestor to the resulting sequence  	$\mathbf{v}=(v_1,\ldots,v_n)\in\mathbb{V}_{N\|n}$, $v_n\le N$ A preset probability distribution: $q_{N\|n}(\mathbf{v})=q_{N\|n}(v_1,\ldots,v_n)$

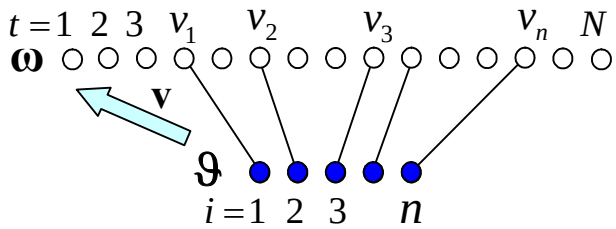
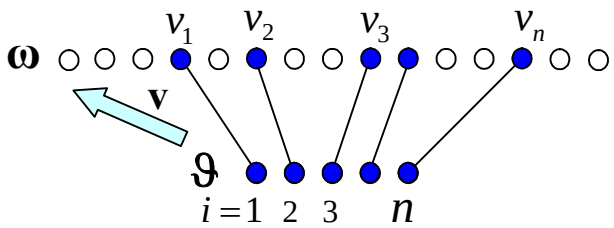
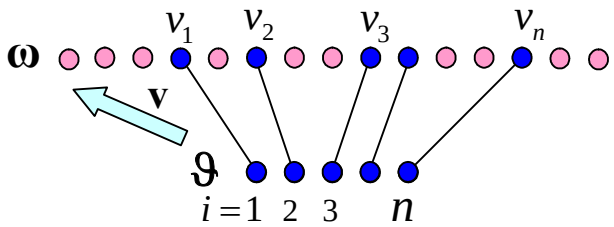
Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega \mid \vartheta_j)$: Three constituents

1. Random structure of the transformation: Unilateral alignment of the ancestor to the resulting sequence  	$\mathbf{v}=(v_1,\ldots,v_n)\in\mathbb{V}_{N\mid n},\quad v_n\le N$ A preset probability distribution: $q_{N\mid n}(\mathbf{v})=q_{N\mid n}(v_1,\ldots,v_n)$
2. Random key subsequence  	$\bar{\boldsymbol{\omega}}_{\mathbf{v}}=(\omega_{v_1},\ldots,\omega_{v_n}),$ $\eta_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}}\mid\vartheta,\mathbf{v})=\prod_{i=1}^{n}\underbrace{\psi(\omega_{v_i}\mid\vartheta_i)}_{\text{Dayhoff's conditional probabilities}}$

Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega \mid \vartheta_j)$: Three constituents

1. Random structure of the transformation: Unilateral alignment of the ancestor to the resulting sequence  	$\mathbf{v}=(v_1,\ldots,v_n)\in\mathbb{V}_{N\mid n},\quad v_n\le N$ A preset probability distribution: $q_{N\mid n}(\mathbf{v})=q_{N\mid n}(v_1,\ldots,v_n)$
2. Random key subsequence  	$\bar{\boldsymbol{\omega}}_{\mathbf{v}}=(\omega_{v_1},\ldots,\omega_{v_n}),$ $\eta_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}}\mid\vartheta,\mathbf{v})=\prod_{i=1}^{n}\underbrace{\psi(\omega_{v_i}\mid\vartheta_i)}_{\text{Dayhoff's conditional probabilities}}$
3. Random additional subsequence  	Absolutely randomly drawn amino acids

Random noncompressing transformation of the ancestor $\varphi_{N_j}(\omega \mid \vartheta_j)$: Three constituents

1. Random structure of the transformation: Unilateral alignment of the ancestor to the resulting sequence  	$\mathbf{v}=(v_1,\ldots,v_n)\in\mathbb{V}_{N\mid n},\quad v_n\le N$ A preset probability distribution: $q_{N\mid n}(\mathbf{v})=q_{N\mid n}(v_1,\ldots,v_n)$
2. Random key subsequence  	$\bar{\boldsymbol{\omega}}_{\mathbf{v}}=(\omega_{v_1},\ldots,\omega_{v_n}),$ $\eta_n(\bar{\boldsymbol{\omega}}_{\mathbf{v}}\mid\vartheta,\mathbf{v})=\prod_{i=1}^{n}\underbrace{\psi(\omega_{v_i}\mid\vartheta_i)}_{\text{Dayhoff's conditional probabilities}}$
3. Random additional subsequence  	Absolutely randomly drawn amino acids
All in all, we have the resulting parametric conditional distribution family of a single protein in terms of the unknown common probabilistic profile: $\zeta_{N\mid n}(\boldsymbol{\omega}\mid\bar{\boldsymbol{\beta}},\mathbf{v})$	

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \omega_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\beta} = \left(\beta_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\beta}) = \prod_{j=1}^{M} f_{N_j|n}(\omega_j \mid \overline{\beta})$

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Maximum-likelihood estimation of the common profile

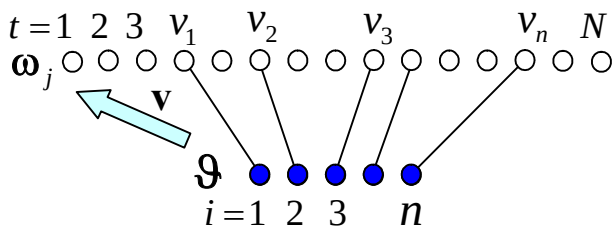
The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Random structure of the transformation: Unilateral alignment of the ancestor to the resulting sequence  	$\mathbf{v}_j = (v_{j,1},\ldots,v_{j,n}) \in \mathbb{V}_{N_j\|n},\ \ v_n \leq N_j$ A preset probability distribution: $q_{N\|n}(\mathbf{v}) = q_{N\|n}(v_1,\ldots,v_n)$

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Let $\overline{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s},\ldots,\boldsymbol{\beta}_{n,s})$ be approximation to the solution at step s. Then, the a posteriori probabilities of the events $\mathbf{v}_{j,i} = t$ are completely defined: $p_{it}(\overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) = P(v_{j,i} = t \mid \overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j)$

Maximum-likelihood estimation of the common profile

The general scenario once again:

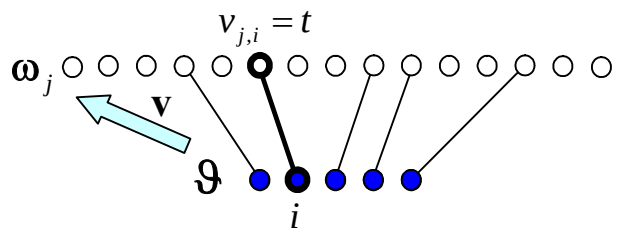
$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Let $\overline{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s},\ldots,\boldsymbol{\beta}_{n,s})$ be approximation to the solution at step s. Then, the a posteriori probabilities of the events $\mathbf{v}_{j,i} = t$ are completely defined: $p_{it}(\overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) = P(v_{j,i} = t \mid \overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j)$



$\mathbf{v}_j = (v_{j,1},\ldots,v_{j,n}) \in \mathbb{V}_{N_j|n},\quad v_{j,n} \leq N_j$

$\boxed{\mathbf{v}_{j,i} = t}$

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Let $\overline{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s},\ldots,\boldsymbol{\beta}_{n,s})$ be approximation to the solution at step s. Then, the a posteriori probabilities of the events $\mathbf{v}_{j,i} = t$ are completely defined: $p_{it}(\overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) = P(v_{j,i} = t \mid \overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j)$

The EM procedure boils down to independent computing each column $\left(\boldsymbol{\beta}_{i,s+1} = (\beta^1_{i,s+1},\ldots,\beta^{20}_{i,s+1}),\ 0 \leq \beta^k_{i,s+1} \leq 1 \right)$ of the best common profile $\overline{\boldsymbol{\beta}}_{s+1} = (\boldsymbol{\beta}_{1,s+1},\ldots,\boldsymbol{\beta}_{n,s+1})$ at the next step:

$$(\beta^1_{i,s+1},\ldots,\beta^{20}_{i,s+1}) = \underset{\substack{(\beta^1_i,\ldots,\beta^{20}_i) \in \mathbb{R}^{20} \\ \sum_{k=1}^{20} \beta^k_i = 1,\ \beta^k_i \geq 0}}{\arg\max} \sum_{l=1}^{20} \sum_{j=1}^{M} \sum_{t=1}^{N_j} I\left[\omega_{jt} = \alpha^l \right] p_{it}(\overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) \ln \sum_{k=1}^{20} \psi(\alpha^l \mid \alpha^k) \beta^k_i,$$

Maximum-likelihood estimation of the common profile

The general scenario once again:

$\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ N_j \geq n,\ j = 1,\ldots,M \right\}$ – the given finite set of amino acid sequences independently generated from the unknown probabilistic profile $\overline{\boldsymbol{\beta}} = \left(\boldsymbol{\beta}_i \in \mathbb{R}^{20},\ i = 1,\ldots,n \right)$ to be estimated.

Thus, the likelihood function is the product: $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \prod_{j=1}^{M} f_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}})$

The likelihood estimate: $\hat{\overline{\boldsymbol{\beta}}} = \arg\max_{\overline{\boldsymbol{\beta}}} \ln F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}) = \arg\max_{\overline{\boldsymbol{\beta}}} \sum_{j=1}^{M} \ln \sum_{\mathbf{v}_j \in \mathbb{V}_{N_j|n}} q_{N_j|n}(\mathbf{v}_j) \zeta_{N_j|n}(\boldsymbol{\omega}_j \mid \overline{\boldsymbol{\beta}}, \mathbf{v}_j)$

The essence of the iterative EM (Expectation-Maximization) procedure aimed at solving this optimization problem is based on the fact that the given set of proteins $\Omega^*_{\geq n} = \left\{ \boldsymbol{\omega}_j,\ j = 1,\ldots,M \right\}$ is considered as the observable part of the two-component random object $(\Omega^*_{\geq n}, \Upsilon_n)$ whose hidden part $\Upsilon_n = (\mathbf{v}_j \in \mathbb{V}_{N_j|n},\ j = 1,\ldots,M)$ is the collection of sequence-specific transformation structures.

Let $\overline{\boldsymbol{\beta}}_s = (\boldsymbol{\beta}_{1,s},\ldots,\boldsymbol{\beta}_{n,s})$ be approximation to the solution at step s. Then, the a posteriori probabilities of the events $\mathbf{v}_{j,i} = t$ are completely defined: $p_{it}(\overline{\boldsymbol{\beta}}_{\cdot s}, \boldsymbol{\omega}_j) = P(v_{j,i} = t \mid \overline{\boldsymbol{\beta}}_{\cdot s}, \boldsymbol{\omega}_j)$

The EM procedure boils down to independent computing each column $\left(\boldsymbol{\beta}_{i,s+1} = (\beta^1_{i,s+1},\ldots,\beta^{20}_{i,s+1}),\ 0 \leq \beta^k_{i,s+1} \leq 1 \right)$ of the best common profile $\overline{\boldsymbol{\beta}}_{s+1} = (\boldsymbol{\beta}_{1,s+1},\ldots,\boldsymbol{\beta}_{n,s+1})$ at the next step:

$$(\beta^1_{i,s+1},\ldots,\beta^{20}_{i,s+1}) = \underset{\substack{(\beta^1_i,\ldots,\beta^{20}_i) \in \mathbb{R}^{20} \\ \sum_{k=1}^{20} \beta^k_i = 1,\ \beta^k_i \geq 0}}{\arg\max} \sum_{l=1}^{20} \sum_{j=1}^{M} \sum_{t=1}^{N_j} I\left[\omega_{jt} = \alpha^l \right] p_{it}(\overline{\boldsymbol{\beta}}_s, \boldsymbol{\omega}_j) \ln \sum_{k=1}^{20} \psi(\alpha^l \mid \alpha^k) \beta^k_i,$$

Theorem. This choice provides that the inequality $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_{s+1}) > F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_s)$ holds true at each step s while $\nabla_{\overline{\boldsymbol{\beta}}} F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_s) \neq \mathbf{0}$; if $\nabla_{\overline{\boldsymbol{\beta}}} F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_s) = \mathbf{0}$ then $F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_{s+1}) = F(\Omega^*_{\geq n} \mid \overline{\boldsymbol{\beta}}_s)$.

Choosing the length of the common profile

Each of n columns in the common profile is a probability distribution over the amino acid alphabet.
The idea: The most appropriate n must provide the minimum average entropy of these distribution:

$$\hat{n} = \arg\min_{n} \left(-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{20} \beta_i^k \ln \beta_i^k \right)$$

Choosing the length of the common profile

Each of n columns in the common profile is a probability distribution over the amino acid alphabet.
The idea: The most appropriate n must provide the minimum average entropy of these distribution:

$$\hat{n} = \arg\min_{n} \left(-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{20} \beta_i^k \ln \beta_i^k \right)$$

The most probable multiple alignment

The n-column profile $\hat{\bar{\beta}}$ found as the maximum-likelihood estimate:

$$\hat{\bar{\beta}} = \arg\max_{\bar{\beta}} \ln F(\Omega^*_{\geq n} \mid \bar{\beta})$$

The a posterior distribution over the set of possible multiple alignments relevant to the set of proteins:

$$p_{it}(\hat{\bar{\beta}}_{\cdot s}, \omega_j) = P(v_{j,i} = t \mid \hat{\bar{\beta}}_{\cdot s}, \omega_j).$$

Choosing the length of the common profile

Each of n columns in the common profile is a probability distribution over the amino acid alphabet.
The idea: The most appropriate n must provide the minimum average entropy of these distribution:

$$\hat{n} = \arg\min_{n} \left(-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{20} \beta_i^k \ln \beta_i^k \right)$$

The most probable multiple alignment

The n-column profile $\hat{\bar{\beta}}$ found as the maximum-likelihood estimate:

$$\hat{\bar{\beta}} = \arg\max_{\bar{\beta}} \ln F(\Omega^*_{\geq n} \mid \bar{\beta})$$

The a posterior distribution over the set of possible multiple alignments relevant to the set of proteins:

$$p_{it}(\hat{\bar{\beta}}_{\cdot s}, \omega_j) = P(v_{j,i} = t \mid \hat{\bar{\beta}}_{\cdot s}, \omega_j).$$

The idea: The a posteriori most probable alignment will be given by the solutions of separate optimization problems corresponding to single proteins ω_j, $j = 1, \ldots, M$:

$$\begin{cases} \mathbf{v}_j = (v_{j,1}, \ldots, v_{j,n}) = \arg\max_{v_1, \ldots, v_n} \prod_{i=1}^{n} p_{iv_{j,i}}(\hat{\bar{\beta}}, \omega_j), \\ v_{j,i} \geq v_{j,i-1}, i = 2, \ldots, n. \end{cases}$$

Choosing the length of the common profile

Each of n columns in the common profile is a probability distribution over the amino acid alphabet.
The idea: The most appropriate n must provide the minimum average entropy of these distribution:

$$\hat{n} = \arg\min_{n} \left(-\frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{20} \beta_i^k \ln \beta_i^k \right)$$

The most probable multiple alignment

The n-column profile $\hat{\bar{\beta}}$ found as the maximum-likelihood estimate:

$$\hat{\bar{\beta}} = \arg\max_{\bar{\beta}} \ln F(\Omega^*_{\geq n} \mid \bar{\beta})$$

The a posterior distribution over the set of possible multiple alignments relevant to the set of proteins:

$$p_{it}(\hat{\bar{\beta}}_{\cdot s}, \omega_j) = P(v_{j,i} = t \mid \hat{\bar{\beta}}_{\cdot s}, \omega_j).$$

The idea: The a posteriori most probable alignment will be given by the solutions of separate optimization problems corresponding to single proteins ω_j, $j = 1,\ldots,M$:

$$\begin{cases} \mathbf{v}_j = (v_{j,1},\ldots,v_{j,n}) = \arg\max_{v_1,\ldots,v_n} \prod_{i=1}^{n} p_{iv_{j,i}}(\hat{\bar{\beta}}, \omega_j), \\ v_{j,i} \geq v_{j,i-1}, i = 2,\ldots,n. \end{cases}$$

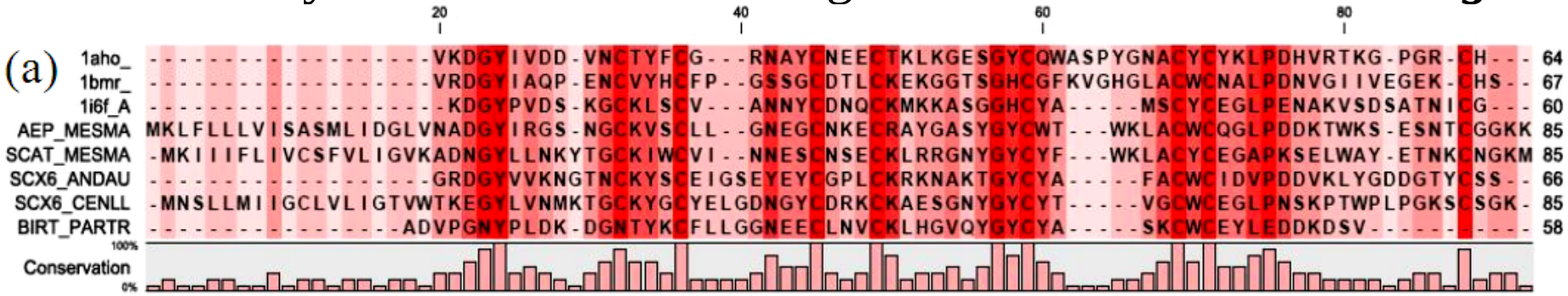
For each protein $j = 1,\ldots,M$, this is a standard dynamic programming problem.

Experimental setup

- Alignment benchmark: **BAliBase 3.0**

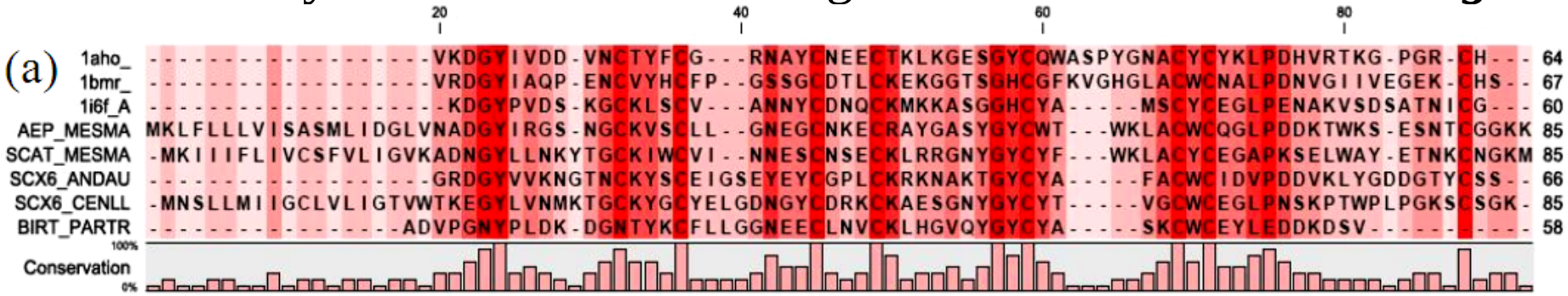
Experimental setup

- Alignment benchmark: **BAliBase 3.0**
- A manually-refined benchmark alignment – ***all columns are aligned***:

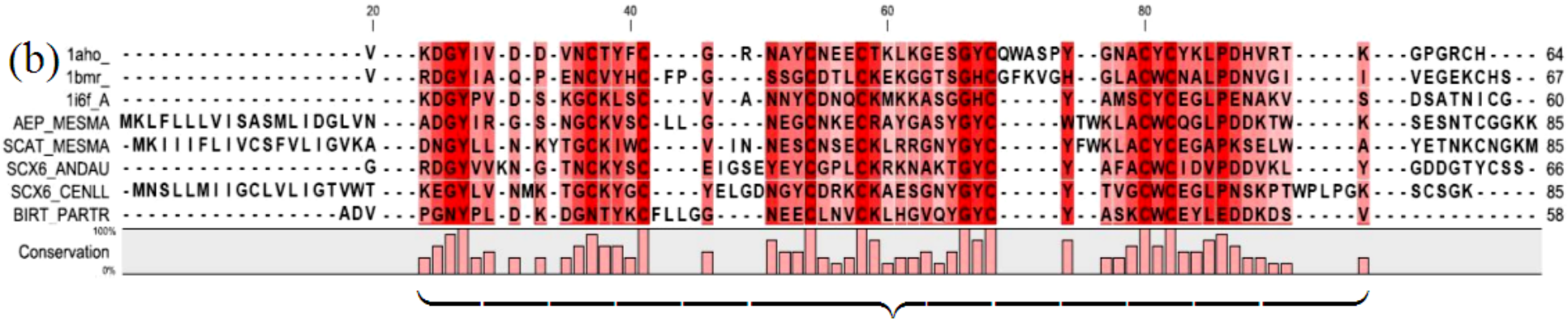


Experimental setup

- Alignment benchmark: **BAliBase 3.0**
- A manually-refined benchmark alignment – ***all columns are aligned***:

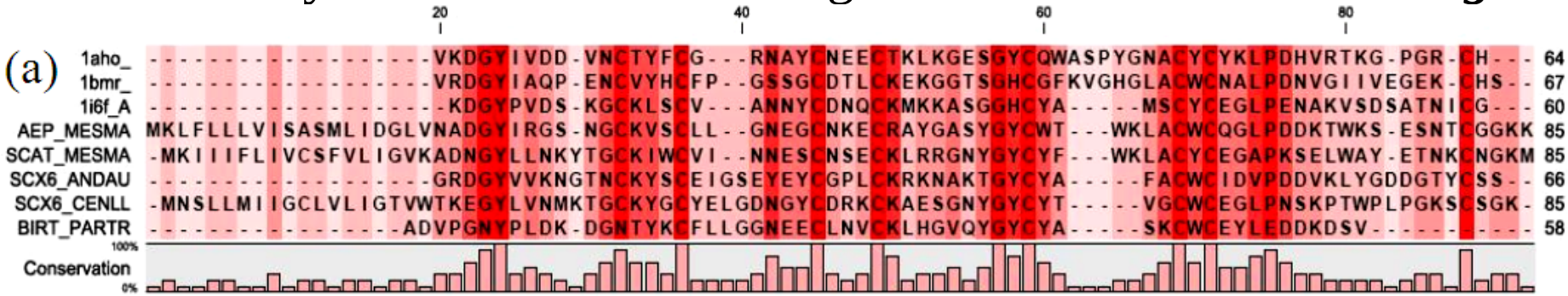


- Characteristic features of proposed alignment: ***only ungapped columns are aligned:***

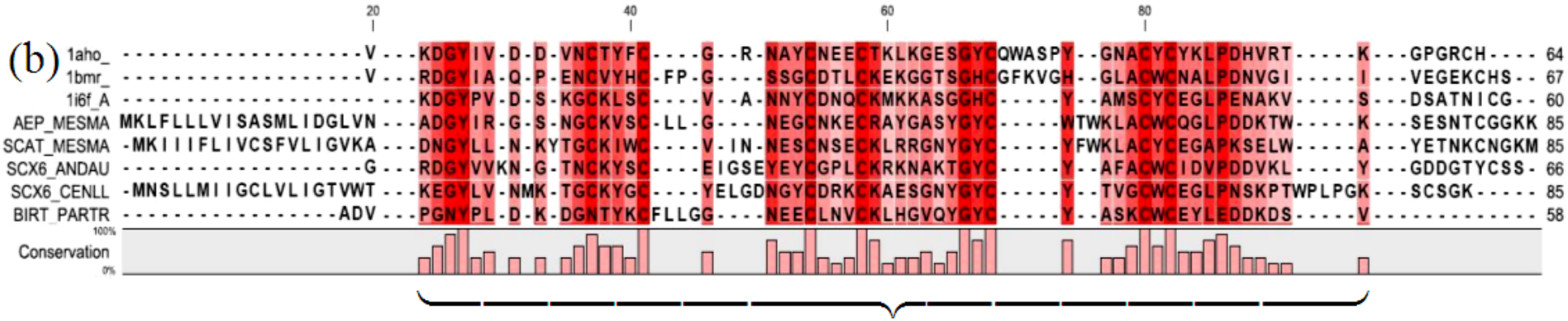


Experimental setup

- Alignment benchmark: **BAliBase 3.0**
- A manually-refined benchmark alignment – ***all columns are aligned***:



- Characteristic features of proposed alignment: ***only ungapped columns are aligned:***



- Prediction accuracy assessment:
  - **SP** – sum of pairs score,
  - **TC** – total column score.

Experimental comparison of multiple alignment procedures in BAliBase 3.0

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**

Experimental comparison of multiple alignment procedures in BAliBase 3.0

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**

Statistics of comparing the proposed approach with 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%

Conclusions

- The proposed formal approach to multiple alignment is based on a deliberately simplified model of protein evolution.
- The iterative procedure of solving the respective optimization problem is based on the well-known EM algorithm.
- The first experiments have shown that this approach outperforms, in average, 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 as a rule, our method shows small decreasing and large increasing of scores in comparison to other methods.