ProbCons: Probabilistic consistency-based multiple sequence alignment

Álvaro Huertas García
Diego Mañanes Cayero
Alejandro Martín Muñoz
Sara Dorado Alfaro

January 16, 2020

Table of contents

Table of contents

Introduction

- Multiple sequence alignment (MSA) $\rightarrow$ way of identifying and visualizing patterns of sequence conservation. It facilitates evolutionary and phylogenetic studies. There are many approaches to multiple sequence alignment:
  1. Exact methods.
  2. Progressive alignment (e.g., ClustalW).
  3. Iterative approaches (e.g., PRALINE, IterAlign, MUSCLE).
  4. Consistency-based methods (e.g., MAFFT, ProbCons).
  5. Structure-based methods: include information about one or more known 3D protein structures.

Introduction: method's approaches

- Dynamic programming $\rightarrow$ too inefficient for more than a few sequences. Instead, heuristic strategies: tree-based progressive alignment, sequences are assembled via several pairwise alignment steps. Errors at early stages propagate and may increase the likelihood of misalignment (alleviated by post-processing steps).
- Consistency-based techniques $\rightarrow$ use evidence from intermediate sequences to guide the pairwise alignment (adjusting the score for a residue pairing according to support from the position of a third sequence that aligns to the others). That is, multiple sequence information is used, as it is being generated.
- COFFEE (another consistency-based) $\rightarrow$ a library is computed by merging consistent CLUSTALW global and LALIGN local pairwise alignments to form three-way alignments, which are assigned weights. The score for the pairwise alignment is the sum of the weights of all alignments in the library containing that aligned residue pair.

Table of contents

Algorithm overview

ProbCons [DMBB05]

- Given m sequences $\rightarrow S = \{s^{(1)}, \ldots, s^{(m)}\}$.
- Maximum expected accuracy.
- Probabilistic consistency $\rightarrow$ MSA conservation information in the pairwise alignment.

1. Step 1: Computation of posterior probability matrices.
2. Step 2: Computation of expected accuracies.
3. Step 3: Probabilistic consistency transformation.
4. Step 4: Computation of the guide tree.
5. Step 5: Progressive alignment.
6. Step 6: Iterative refinement (post-processing OPTIONAL step).

Step 1: Computation of posterior probability matrices

- For $x, y \in S$, compute the matrix

$$P_{xy}(i,j) = \boldsymbol{P}(x_i \sim y_j \in a^* | x, y) \ ,$$

  where $1 \leq i \leq |x|$ and $1 \leq j \leq |y|$.
- Each position $P_{xy}(i,j)$ is the **posterior** probability that letters x_i and y_j are paired i an alignment a^*.
  - Computing posterior probabilities in pair-HMMs [DEKM98].
- Time complexity $O(m^2 L^2)$.
  - m is the number of sequences.
  - L is the length of each sequence.

Step 2: Computation of expected accuracies

- The expected accuracy is defined as

$$\boldsymbol{E}_{a^*}(acc(a, a^*)|x, y) = \frac{1}{\min\{|x|, |y|\}} \sum_{x_i \sim y_j \in a} P_{xy}(i, j) \ ,$$

  where a is the align*ment that maximizes the expected accuracy by dynamic programming.

- Set

$$E(x, y) = \boldsymbol{E}_{a^*}(acc(a, a^*)|x, y) \ . \tag{1}$$

Step 3: Probabilistic consistency transformation

- Reestimate quality scores $\boldsymbol{P}_{xy} \rightarrow$ probabilistic consistency transformation.
- Incorporate similarity of x and y to other sequences in S:

$$\boldsymbol{P}'(x_i \sim y_j \in a^* | x, y) = \frac{1}{|S|} \sum_{z \in S} \sum_{z_k \in z} F(x_i, y_j, z_k) \ ,$$

  where $F(x_i, y_j, z_k) = \boldsymbol{P}(x_i \sim z_k \in a^* | x, z) \times \boldsymbol{P}(z_k \sim y_j \in a^* | z, y)$.
- In matrix form:

$$P'_{xy} = \frac{1}{|S|} \sum_{z \in S} P_{xz} P_{zy} \ .$$

- **Optimization:** use sparse matrices ignoring entries $\leq \omega$ (threshold).
- This step can be iterated until convergence.

Steps 4, 5 and 6

- Hierarchical clustering.
  - Similarity measure $E(x, y)$ as defined in Equation (1).
  - WPGMA method.
- Align sequence groups hierarquically.
  - Sum-of-pairs.
  - Gap penalties $\rightarrow 0$.
- Progressive alignment.
  - Randomly partition alignment into two groups of sequences.
  - Realign.
  - This step can be iterated.

Table of contents

Some experiments with BAliBASE dataset

- The BAliBASE dataset:
  - 141 reference protein alignments.
  - Hand constructed alignmets from the literature.
  - 5 subsets with alignments of different characteristics.
  - Test alignmets are scored respect **core blocks** $\rightarrow$ reliable alignmets.
- **No universally accepted accuracy measure for protein alignmets.**
  - Sum-of-pairs score (SP).
  - Column score (CS).

Column reliability for BAliBASE

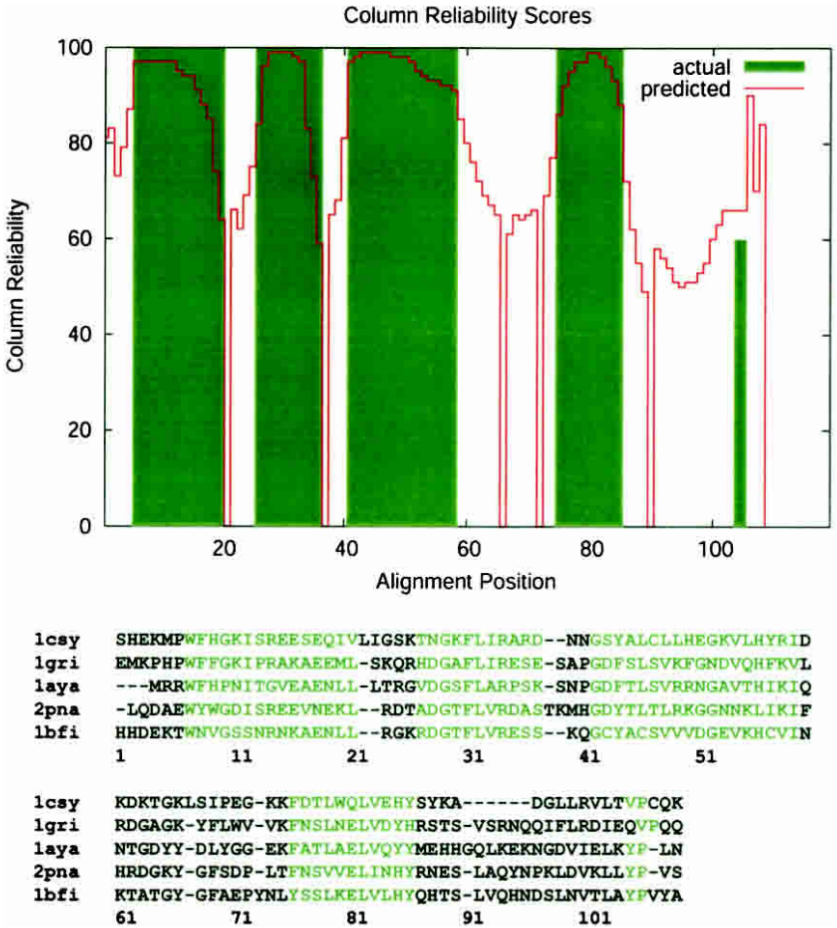

Image from [DMBB05].

At each position:

- Red line $\rightarrow$ predicted proportion of correct pairwise matches.
- Green Blocks $\rightarrow$ actual proportion of correct pairwise matches.

Comparison with other methods

Table 1. Performance of aligners on the BAliBASE benchmark alignments database

Aligner	Ref 1 (82)		Ref 2 (23)		Ref 3 (12)		Ref 4 (12)		Ref 5 (12)		Overall (141)		Time (mm:ss)
	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	SP	CS	
Align-m	76.6	n/a	88.4	n/a	68.4	n/a	91.1	n/a	91.7	n/a	80.4	n/a	19:25
DIALIGN	81.1	70.9	89.3	35.9	68.4	34.4	89.7	76.2	94.0	84.3	83.2	63.7	2:53
CLUSTALW	86.1	77.3	93.2	56.8	75.3	46.0	83.4	52.2	85.9	63.8	86.1	68.0	1:07
MAFFT	86.7	78.1	92.4	50.2	78.8	50.4	91.6	72.7	96.3	85.9	88.2	71.4	1:18
T-Coffee	86.6	77.4	93.4	56.1	78.5	48.7	91.8	73.0	95.8	90.3	88.3	72.2	21:31
MUSCLE	88.7	80.8	93.5	56.3	82.5	56.4	87.6	60.9	96.8	90.2	89.6	73.9	**1:05**
ProbCons	**90.1**	**82.6**	**94.4**	**61.3**	84.1	**61.3**	90.1	72.3	97.9	91.9	91.0	77.2	5:32
ProbCons-ext	90.0	82.5	94.2	59.1	**84.3**	61.1	**93.8**	**81.0**	**98.1**	**92.2**	**91.2**	**77.6**	8:02

Columns show the average sum-of-pairs (SP) and column scores (CS) achieved by each aligner for each of the five BAliBASE references. All scores have been multiplied by 100. The number of sequences in each reference is given in parentheses. Overall numbers for the entire database are reported in addition to the total running time of each aligner for all 141 alignments. The best results in each column are shown in bold.

Figure: Image from [DMBB05].

Table of contents

Examples: Comparison between methods

- MSA of distantly related globins (human beta globin, human myoglobin, human neuroglobin, soybean leghemoglobin, rice hemoglobin) using four different programs. Symbols: * complete conservation, : conservative substitutions, . less conservative substitutions. Programs differ in:
  - Align corresponding regions of alpha helical secondary structure (red lettering).
  - Align conserved histidines (open and black arrowhead). They are important in coordinating protein binding to the heme group $\rightarrow$ they should be aligned by all the programs. The open arrowhead histidine shows a complete conservation. The conservation of the black is only achieved by ProbCons and T-Coffee.
  - Create and place gaps (boxed regions).

(a) Praline multiple sequence alignment

```
beta globin      .........MVHLTPEEKSAVTALWGKV..NVDEVGGEALGRLLVVYPWTQRFFES.FG
myoglobin        ..........MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK.FK
neuroglobin      ............MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          .........MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS..FL
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS..FL
Consistency      0000000000142654382579345734633643436244536864 33*35344*50063

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSEL..HCDKLH....VDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTALGGILKKKGHHEAEIKPLAQS..HATKHK....IPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLEEYLASLGRKHRAVG....VKL
soybean          A.NGVDP..TNPKLTGHAEKLFALVRDSAGQL.KASGTVVADAA....LGSVHAQKAVTD
rice             R.NSDVPLEKNPKLKTHAMSVFVMTCEAAAQL.RKAGKVTVRDTTLKRLGATHLKYGVGD
Consistency      3166354224776653*43686354244544513356343335420033354400009 22

beta globin      ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH.....
myoglobin        KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean          PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA.......
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency      437448444982585423053365544 54*5546542644675432200100 0
```

(b) MUSCLE (3.6) multiple sequence alignment

```
beta globin      ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin        -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin      ------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                           :   :   :   :  ..  .    .      ::   *     *.

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAF---SDGLAHLDNLKGTFATLSELHCDKLH--VDPE
myoglobin        HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAEIKPLAQSHATKHK--IPVK
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVI---DAAVTNVEDLSSLEEYLASLGRKHRAVGVKLS
soybean          NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice             NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                           . ..  *  .::        :                :    :

beta globin      NFRLLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH------
myoglobin        YLEFISECIIQVLQSKHPGD-FGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SFSTVGESLLYMLEKCLGPA-FTPATRAAWSQLYGAVVQAMSRGWDGE----
soybean          QFVVVKEALLKTIKAAVGDK-WSDELSRAWEVAYDELAAAIKKA--------
rice             HFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE---
                  :  :  ::  :        :      * .     .   :
```

(c) PROBCONS

```
beta globin      M----------VHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFES-FG
myoglobin        M----------GLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDK-FK
neuroglobin      M------------ERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          M----------VAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSF-LA
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFSF-LR
                 *         * :   :   :   :  ..  .    .      ::   *     *.

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLD---NLK---GTFATLSELHCDKLHVDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTALGGI---LKKKGHHE---AEIKPLAQSHATKHKIPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDLSSLE---EYLASLGRKHRAV-GVKL
soybean          NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVV----ADAALGSVHAQK-AVTD
rice             NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKY-GVGD
                 .    :    . ..  *  .::        ::   .          *.  *      :

beta globin      ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHK------YH
myoglobin        KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRG---W-DGE
soybean          PQFVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIK--------KA
rice             AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                   :  :   ::  :        :      * .     .   :
```

(d) CLUSTAL FORMAT for T-COFFEE Version_5.13

```
beta globin      ----------MVHLTPEEKSAVTALWGKVNVD--EVGGEALGRLLVVYPWTQRFFE-SFG
myoglobin        -----------MGLSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFD-KFK
neuroglobin      ------------MERPEPELIRQSWRAVSRSPLEHGTVLFARLFALEPDLLPLFQYNCR
soybean          ----------MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFS-FLA
rice             MALVEDNNAVAVSFSEEQEALVLKSWAILKKDSANIALRFFLKIFEVAPSASQMFS-FLR
                           :   :   :   :  ..  .    .      ::   *     *.

beta globin      DLSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL---KGTF---ATLSELHCDKLHVDP
myoglobin        HLKSEDEMKASEDLKKHGATVLTAL---GGILKKKGHHEAE---IKPLAQSHATKHKIPV
neuroglobin      QFSSPEDCLSSPEFLDHIRKVMLVIDAAVTNVEDL---SSLEEYLASLGRKH-RAVGVKL
soybean          NGVDP----TNPKLTGHAEKLFALVRDSAGQLKASGTVVAD----AALGSVHAQKAVTDP
rice             NSDVP--LEKNPKLKTHAMSVFVMTCEAAAQLRKAGKVTVRDTTLKRLGATHLKYGVGDA
                 .         . ..  *  .::        :                *.  *

beta globin      ENFRLLGNVLVCVLAHHF-GKEFTPPVQAAYQKVVAGVANALAHKYH------
myoglobin        KYLEFISECIIQVLQSKH-PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin      SSFSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGAVVQAMSRGWDG----E
soybean          Q-FVVVKEALLKTIKAAV-GDKWSDELSRAWEVAYDELAAAIKKA--------
rice             H-FEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQE---MKPAE
                   :  :   ::  :        :      * .     .   :
```

Table of contents

References I

Richard Durbin, Sean R Eddy, Anders Krogh, and Graeme Mitchison, *Biological sequence analysis: probabilistic models of proteins and nucleic acids*, Cambridge university press, 1998.

Chuong B Do, Mahathi SP Mahabhashyam, Michael Brudno, and Serafim Batzoglou, *Probcons: Probabilistic consistency-based multiple sequence alignment*, Genome research **15** (2005), no. 2, 330–340.

ProbCons: Probabilistic consistency-based multiple sequence alignment

Álvaro Huertas García
Diego Mañanes Cayero
Alejandro Martín Muñoz
Sara Dorado Alfaro

January 16, 2020