

# 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

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References

# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References

- Multiple sequence alignment (MSA) → 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 → 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 → 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) → 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

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References

# Consistency-based methods

- Based on : “prevention is the best medicine”
- Combines iterative and progressive approaches with probabilistic models:
  - ① Uses **Hidden Markov Models** to calculate matrices for matching residues in pairwise alignments.
  - ② Uses information about multiple sequence alignment as it is being generated to guide the pairwise alignments
  - ③ Multiple alignment via tree-based **progressive alignment**
  - ④ Errors at early stages in the alignment are alleviated by **post-processing steps** such as iterative refinement

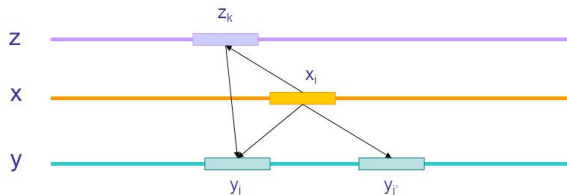
# Consistency-based methods

## Imaging this biological scenario

Sequence  $x \rightarrow x_i$

Sequence  $y \rightarrow y_i$

Sequence  $z \rightarrow z_k$



- If  $x_i$  aligns with  $z_k$  and  $z_k$  aligns with  $y_i$ , then  $x_i$  should align with  $y_i$
- Consistency-based techniques **score pairwise alignments** in the context of **information about multiple sequences**



# Table of contents

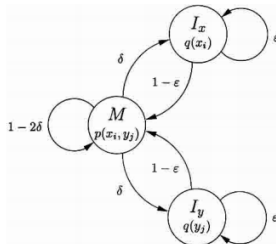
- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm**
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References

- **ProbCons[?]** is a pair-hidden Markov model-based progressive alignment algorithm that differs from most typical approaches in its use of **maximum expected accuracy** rather than Viterbi alignment, and of the probabilistic consistency transformation to incorporate multiple sequence conservation information during pairwise alignment.
- **Hidden Markov Models (HMMs)** in sequence analysis are based on a strong probabilistic model that **includes a representation of INDELs** (insertions and deletions, i.e. gaps).
- The HMM describing families of related sequences are called **profile HMMs**

# Algorithm overview

- In profile HMMs the residues in each position of the alignment can be in one of three possible states:
  - ① **Match:** represent conserved position
  - ② **Insert:** represent small stretches of nonspecific sequence
  - ③ **Delete:** correspond to gaps and represent the absence of a conserved residue
- Each state has associated:
  - ① **Emission probability:** correspond to the probability of observing each amino acid at that particular position of the alignment
  - ② **Transition probability:** describes the frequency of observing a match, insertion or deletion in column  $i+1$  given the state column  $i$ .

# Algorithm overview



- Emission probabilities, which correspond to traditional substitution scores, are based on the BLOSUM62 matrix.
- Transition probabilities, which correspond to gap penalties, are trained with unsupervised Expectation-Maximization (EM)
  - $\pi_{insert}$ : initial insertion probability parameter
  - $\delta$ : insertion start probability parameter
  - $\epsilon$ : insertion extension probability parameter
- The resulting parameters ( $\delta = 0.019931$ ,  $\epsilon = 0.79433$ ,  $\pi_{insert} = 0.19598$ ) are used as the default for the program.

## ProbCons [?]

- Given  $m$  sequences  $\rightarrow S = \{s^{(1)}, \dots, s^{(m)}\}$ .
- 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

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

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 in an alignment  $a^*$ .
  - Computing posterior probabilities in pair-HMMs [?].
- 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

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

where  $a$  is the alignment that maximizes the expected accuracy by dynamic programming.

- Set

$$E(x, y) = \mathbf{E}_{a^*}(\text{acc}(a, a^*)|x, y) . \quad (1)$$

## Step 3: Probabilistic consistency transformation

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

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

where  $F(x_i, y_j, z_k) = \mathbf{P}(x_i \sim z_k \in a^* | x, z) \times \mathbf{P}(z_k \sim y_j \in a^* | z, y)$ .

- In matrix form:

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

- **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 hierarchically.
  - Sum-of-pairs.
  - Gap penalties  $\rightarrow 0$ .
- Progressive alignment.
  - Randomly partition alignment into two groups of sequences.
  - Realign.
  - This step can be iterated.

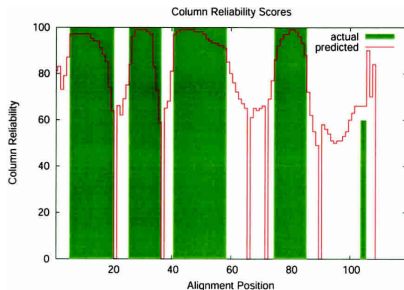
# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments**
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References

# 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** → reliable alignmets.
- **No universally accepted accuracy measure for protein alignmets.**
  - Sum-of-pairs score (SP).
  - Column score (CS).

# Column reliability for BALiBASE



```

1csy SHEKMPWFHGKISRREEQIVLGSKTNGKFLIRARD--NNGSYALCLLNEGKVLHYRID
1gri EMKPHPFPGKIPRAAEML-SKQRHGDGAFIRESE-SAPGDFLSVKFGNDVQHFKVL
1aya ---MRRWFHFNITGVAEHLI-LTRGVGGSFLARPSK-SNPGDFTLSVRENGAVTHIKIQ
2pna -LQDAEFTWGDISRREEVNEKI--RDTADGTFLVRDASTMHGDDYTLTLREGGNSKLIKIF
1bfi HHDEKTNVVGSSNRNKAENLI--RGKRGGTFLVRESS--KQGCYACSVVVDGEVHCVIN
      1      11      21      31      41      51

```

```

1csy KDKTKGLSIEEG-KKPTLMQLVEHYSYKA-----DGLLRVLTVFCQK
1gri RDGAGK-YFLAV-VKFNLSLNLVDYHRSTS-VSRHQQIFLRDIEQVTFQ
1aya NTGDIY-DLYGG-EKFATLAEVLQYIMEHHGQKKEKNGDVIELKYP-LN
2pna HRDGKY-GFSDP-LTFNSVVELINHYRNES-LAQYNPKLDVKLLYP-VS
1bfi KTATGY-GFAEPTNLYSSLKELVLYHQHTS-LVQHNDLSLNTLAYFVYA
      61      71      81      91     101

```

Image from [?].

At each position:

- Red line → predicted proportion of correct pairwise matches.
- Green Blocks → actual proportion of correct pairwise matches.

# Comparison with other methods

## ProbCons multiple alignment tool

**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	<b>1:05</b>
ProbCons	<b>90.1</b>	<b>82.6</b>	<b>94.4</b>	<b>61.3</b>	84.1	<b>61.3</b>	90.1	72.3	97.9	91.9	91.0	77.2	5:32
ProbCons-ext	90.0	82.5	94.2	59.1	<b>84.3</b>	61.1	<b>93.8</b>	<b>81.0</b>	<b>98.1</b>	<b>92.2</b>	<b>91.2</b>	<b>77.6</b>	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 [?].

# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples**
- 6 Drawbacks
- 7 Conclusions
- 8 References

# 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 → 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).

```

beta globin      .....MVHLTPEEKSAVTALGKGV.....NVDEGGEGALGRLNVVYPTWRQFFES.PG
myoglobin       .....MGLSDGQGLVNLVNVGKRVADVDPHGQGEVLRLPFGHPETLEKDFK.PK
neuroglobin     .....MERPELRLQSGRAVSRSPLEHGTVLRLFALEPAKDLPLFPQYNCR
soybean         .....MVAFPTQKDALVSSSFAFKANITQPVSVFVTSILEKAPAKDLFS..FL
rice            .....MLVEDYNNAIVASPSDEQALGKLKSNALKDSANLRFLPKLVPFVASPMQFS..FL
Consistency     000000000000142654382579345734633643436244536864333344*00063

```

beta globin DLSTPDAVMGNPKVKANGKVLGFAGSDGLAHLDNLKGTFTATSEL...NCDLHK...VDP  
myoglobin HLKSEDEMDKLEKDKKHGATVLTALGGLKKKGKHHGAEIPKLAQS...NATIKHK...IPV  
neuroglobin QFSPEFDCLSSPFLDRLRKVMIVDIAATNVHDELSLEYLSGLGRKHEAVG...VKL  
soybean A.NGVDP...TNPKLTGHAELKFLVRDSAGQLKASGTVVADAA...LGSVIAQKAVTD  
rice R.NSDVPELQKHLKTHAMSVFVNICAAAGQLRKAGKVTVRDITLKRGLKELKYGVGD  
Consistency 3166354224776653433686352445444513356343352400333544000922

```
beta globin      ENFRLGLGNVLVCYLAHHP.GKEFTPPVQAAQKVVAGVGNALAHKYH.....
myoglobin       KYLEFISECIQLVLSQKH.PGDFPADAQGMANKALELFRIDMASNIYKELGFQG
neuroglobin     SSFSTVGESLLMYMLEKCL.GPAFTPATRAASQLVXGAVQMSRWMD..GE...
soybean         PPFVTVVKSLALLKT IKAAY.GDKVKSLSRAWVEYDELAAI KKA.....
rice            AHPEVFKVPLDIT IEKVFPADMMSPAMSKASEAYDLHLVAALIKEMKPAE...
Consistency     4374844444225854230536355454545454646464444444444444
```

(C) PROCONS

beta globin M-----VHLTPEEKSAVTALMGKVND--EVGGEALGRLLVVPWTQFFES-FG  
myoglobin M-----GLSDGEQWLVLNVGVKVEDIPGHGGVFLIRLFKGGPTLLEKDFPK  
neuroglobin M-----ERPEPEIRQGSWASRSPLSEGTFLVARFALEPDLFPLFYQNCR  
soybean M-----VAFTEKQDALVSSSFPAFKANI PQYSVVPYTSILEKAPAKDLPSF-**LA**  
rice MALVEDNNNAVAVSFSERQALVLKSKALKKDSANTALRFFLKIFEPVAPSASQSPSF-**LA**

beta globin DLSTPDVAMGNPKVKAHGKKVLGAFSDG LAHLD---NLK---GTFATLSLEKDKLVDP  
 myoglobin HKSEDEMKASDELKPKHGVTLTALGII---KKKGHHH---AEIKPLAQSHTKHKIPV  
 neuroglobin QLFSSPECDLSPKFLGTHIRKVLVMDIAVTTNDELSSLE---EYLALEGRHARVQ/GKL  
 soybean NGVDP---TNPKLTGHAELKFLVVRDSAGQLKASGTVV---ADAALGSVIAQK-A/TD  
 rice NSDVP---LEKNPKLKTAMSVFVMTCEAAQLRKAAGVTVRDITLLKRLGATHLKY-G/GD

```

beta globin  ENFRLNGNLVLCVLAHHF-GKEFTFPVQYAAQKVVGAVANALAHK-----YH
myoglobin   KYLEFISECIIVGLQSKH-PGDGGAQAGMINKALEFRDKMASNYKELGFGQ
neuroglobin SFSTFVTEGSLIYMLEKLI-GPAFTPATRAWSQLYGVAVQMSRQ-----W-DGE
soybean     PQFVVVKEALLKTIKAAY-GDKWSDLSRAWVEYADLAAAIK-----KA
rice        AHFEVVKFALLDTIKEEVPADMWSPAMKSAEYSDHVAALKQE-----MKPAE

```

```

beta globin -----MVHLTPEEKSAVTALGKGVND---EVGGEALGRLLVVPWTQRFFES-PG
myoglobin -----MGLSDGEQWLVLVGVKVEDIPHGGEVLRLPKFGKHPTLEKFFK-PK
neuroglobin -----MERPEEPLRLQSGRAVRSRSDPFLHGLFARLAPDPLLPLFYQNCR
soybean -----MVAFTEKQDALVSSSEAFKANKIPQSVVVFYTSILEKAPAKDLFSF-IA
rice      MALVDENNNAVVSFRRKRALVSKSLALKKSDANLRFPLTICVAPASCSMFSP-LR

```

beta globin DLSTPDVAMQNPVKVAKGKVLGAF--SDGLAHLNLEKGTFTSLSEGLKKDLK--VDPE  
 myoglobin HLKSEDEMKASDLKKKHGATVLTAL--GGILKKKGHHEAEIKPLQAGNATHK--IPVK  
 neuroglobin QFSDPEDLSSSPFLLKIRIKVMVLI--DAATNVNEDLSSLVYLAGSLGKHRAVQKVLDS  
 soybean GMSD----TFNPLKGHAELFALVRDASGKASGTSTV--AALGSVAGAGKATD  
 rice NSDWP--LEKNPKLKTAMSVFVMTCEAAAGLRLKAGKIVTRDITLKRIGATHLKGVDGA

```

beta globin  NP1RL2LN3VL4VC5LA6HH7F8G9KE10-F11TP12VQ13AA14Q15K16V17V18AG19VA20NA21L22HK23Y24-----
myoglobin    Y1LE2F3I4SE5C6I7Q8VL9SK10PG11D12-FG13AD14AG15MA16NK17LA18FL19R20KD21MA22SN23KE24L25GF26Q27G
neuroglobin  F1TP2TV3SE4LL5YL6IL7E8KL9GP10A11-F12PT13AK14RA15N16Q17LV18AG19VA20NA21SN22RG23ND24G25-----
soybean      Q1V2V3V4KE5AL6L7T8K9AA10V11G12-W13SE14LS15RA16N17EV18LL19VA20NA21L22AA23IK24KA25-----
rice         H1P2EV3V4KL5AL6LE7T8K9AA10V11G12-W13SE14LS15RA16N17EV18LL19VA20NA21L22AA23IK24KA25-----

```

CLUSTAL FORMAT for T-COFFEE Version\_5.13

```
beta globin  -----MVHLTPEEKSAVTALMGKVIVND-----EVGGEALGRLLVVPWTQRFPE-SFG
myoglobin   -----MGLSDGEQWLVLNVKAVRSDLEIGHGVQLVRLFKPHGPTLETKEFD-KFK
neuroglobin -----MERPEPLIRQSRVAVRSFLEHGTVFLFARLFAEPDLLPLQFYNCR
soybean      -----MVAFTEKQDALVSSSPLEAKNIPOYSVVFVITILEKAPAAKDLFS-FLA
rice         MALVVDNNNAVASFSEQRQALVLKFSKANTATKPSVFTTFFETVADPSASOMES-FLA
```

```

beta globin DLSTPDVAMGNPKVKAHGKVLGAFSDGLAHLDNL---KGF--ATLSELHCDKHLVHP
myoglobin HLKSEDEMCASEDLLKKGATVLMV---GGILKKKGHEAE---TKPLAQSHATKHKIPV
neuroglobin QVSPEDCLSSDLPDTHIRKVTMLVIDAAITNVDEL---SSLERYLSALGRHK-RAGVGLK
soybean      NGDVP---TNPKLTQHAERKFLPVDIAAQKAGTIVVD---AALGYSAQKAVKATVP
rice         NSDVP---LEKNPKLKTAMSVFVMTCEAAQLRKLAKGVTVTRDTLLKRLGATLKYGVGIDA

```

beta globin ENFRLLGNLVLCVLAHHF-GKEFTTPVQAAYQKVVGAVANALAHKYH-----  
myoglobin KYLEFISCEIIQVLQSKH-PGDFGADAGAMNKALELPRKDMASHNYELGFGQ  
neuroglobin SSFSTVSGELLMYLEKLG-GPAFTPATRAANSQLYGAVVQASRWGDC-----E  
soybean CQ-FVVKKEALLTKAAV-GDKNSDELSRAMEVAYDELAATAIKK-----  
rice H-FEVVKFALLDTTKEEVPADMWSPAMKSAENSEYTHLVAAIKQE---MKPKE



# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks**
- 7 Conclusions
- 8 References

- **Computational weight:** The computation step of calculation of posterior probabilities takes time  $O(m^2L^2)$ , where  $m$  is the number of sequences and  $L$  is the length of each sequence.
- **M-Coffe (Meta-Coffe):** combines the output of 15 different sequence alignment methods(ProbCons included). M-Coffe employs a consistency-based approach to estimate a more accurate consensus alignment.
- **Structural methods:** adding structural information, even further accuracy is achieved.

# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions**
- 8 References

# Conclusions

The ProbCons algorithm uses an extremely simple model of sequence similarity (a three-state pair-HMM):

- 1 Makes no attempt to incorporate biological knowledge (i.e position specific gap scoring or rigorous evolutionary tree construction).
- 2 Use amino acid alphabet and BLOSUM emission probability matrices as protein-specific alignment information
- 3 Can be used to DNA alignment by changing the alphabet and the BLOSUM matrices with values for nucleotides.
- 4 The parameter used in the model are transparent ( $\pi_{insert}$ ,  $\delta$ ,  $\epsilon$ )
- 5 The training program as done automatically on unaligned sequences using Expectation-Maximization.
- 6 High accuracy: probabilistic consistency transformation and objective function.

# Table of contents

- 1 Introduction
- 2 Consistency-based methods
- 3 ProbCons: The algorithm
- 4 Experiments
- 5 Examples
- 6 Drawbacks
- 7 Conclusions
- 8 References**

# References I

# ProbCons: Probabilistic consistency-based multiple sequence alignment

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

January 16, 2020