ProbCons: Probabilistic consistency-based multiple sequence alignment

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

January 15, 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

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*436863544454451335634333542003335440000922

beta globin       ENFRLLGNVLVCVLAHHF.GKEFTPPVQAAYQKVVAGVANALAHKYH.....
myoglobin         KYLEFISECIIQVLQSKH.PGDFGADAQGAMNKALELFRKDMASNYKELGFQG
neuroglobin       SSFSTVGESLLYMLEKCL.GPAFTPATRAAWSQLYGAVVQAMSRGWD..GE..
soybean           PQFVVVKEALLKTIKAAV.GDKWSDELSRAWEVAYDELAAAIKKA.....
rice              AHFEVVKFALLDTIKEEVPADMWSPAMKSAWSEAYDHLVAAIKQEMKPAE...
Consistency       437448444982585423053655445 4*5546542644675432200 1000
```

(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
                    :  :   ::  :        :      * .     .   :
```

ProbCons: Probabilistic consistency-based multiple sequence alignment

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

January 15, 2020