

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

1 Introduction

2 Examples

Table of contents

1 Introduction

2 Examples

- 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 Examples

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).

(a) Praline multiple sequence alignment

```

beta globin      .....MVLHTPEEKSAVTALNKGKV..NVDVEYDGLGLRLVL.PWTQRFPEF.PG
myoglobin       .....MGLSDGQELVGLKQGVKEADI.PHQHGVKILRLKPGHPSTLEKFDK..PK
neuroglobin     .....MERPEKFDLIRIQSNRAVSRSLEHGTVLARLFLAEPDLPLDQVNC..FL
soybean         .....MVAFPTQDALVSSSFESAFKANI.QPVYSSVFTYSLELAKAAKDLF..PCL
rice            0ALNVEDNNNAVAVSPSEQELALVLSKVALLKDSANALRFFPLFPEVPAASQOMF..PL
Consistency     0000000000014265438257934571363343624453686433*3534*50063

```

beta globin DLSTPDAVMGNPKVFKV^ΔGLGAFSDGLAHLNDLKGFTATSL^Δ.MCDKLH...VDP
 myoglobin HLKSEDEKAS^ΔEDLKKHGAVLTALGGI^ΔKKKGHHIEIKPLAQ^Δ.NATIKHK...IPV
 neuroglobin GFSSPEDCLSS^ΔPEFLDRIKVM^ΔVIDAA^ΔTNVEDLSLEYLSA^ΔLSGRN^ΔAVG...VKL
 soybean A.NGVDP^Δ.TNPKLTGHARKLFA^ΔLVDRKQGL^ΔKASGT^ΔTVDA^Δ...LSGVAKQAVT^Δ
 rice R.NSDVPLEKGLTKAMSV^ΔPTVNCIA^ΔEAKL^ΔRKAGVT^ΔIVDR^ΔTKLRLKGL^ΔKKYGVGD^Δ
 Consistency 3166354224776635436863542445444513365343335420333544000922

```

beta globin      ENFRLIGNVLVCYLAHHP.GKEFTPPVQAAQKVVAGVYANALAHKYH....
myoglobin       KYLEFISCEIQLVLYSLK.PGDFGADAGMAMNKLELFKIDMASNYKELGFGQ
neuroglobin     SSFSTVGESLIYLMYLEKCL.GPAFTPATRAAGSKVLGVAGVAMSRGWD..GE
soybean         PQFVVVKEALLKTIKAAAY.GDKVSELSRWAVFYDELAAAIKKA.....
rice            AHFEVVKVALLDITKEEVPADMSSPAMKSAWSEAYDELHVAALIKEMKPAE..
Consistency     437448444442585423053365545454545454642644654320100

```

(c) PROBCONS

```

beta globin  M-----VHLTPEEKSAVTALMGKVND--EVGGEALGRLLVVYPWTQFFES-FG
myoglobin    M-----GLSDGEWLQVLMGVKEVADIPHQGVQLVILFKPGHKPTLEKDFK-PK
neuroglobin  M-----ERPELPIRQSMVAVRSPLELHGTVFARLFALEPDLPLLQFNCR
soybean      M-----VAFTEKQDALVSSSEAFKANI PQYSVVFVTTILEKAPAAKDLPSYF-LA
rice         MVLVEDNNVAVVSFEQEQEALVLKSMAILKDSANIALRFLKIFEPVAPASQMFSS-LR

```

beta globin DLSTPDAVMGNPKYAKGVGAFSDGLAHLD---NLK---GTFATLSLSEGLDKLHVDP
 myoglobin HLKSEDEMKASREDLKHGHTVITAIAGGI---LKKKGHHB---AEIKPLAQSGLKRRHKIPV
 neuroglobin QFSDSE---TSSPEFLDRLIKRMVLVIDAATVNDLSLSE---EYIASLQGRKHRAV---GK/L
 soybean GNVDP---CLNPKLTGHAELKFLALVRDSSQQLKASGTVTV---ADAALGSYVHQAQ---AT/L
 rice NSDVP---LEKNPKLKTAMSVFVMTCEAAQLRQKAGVTVREDTTLKRLGATLKY---G/GD

beta globin	ENFRLLGVLVLCVLAHFF-GKEFTPPVQAAYQKVVAGVANALAHK-----YH
myoglobin	KYLEFISCEIIQVLQSKH-PGDFGADAQGAMNKALELFRKIDMASNYKELGPFQG
neuroglobin	SSPSTVGESLLYMLEKCL-GPAFTPATRAAWSQLYGVAVQAMSRG---W-DGE
soybean	PQFVFWKEALLKTIKAAY-GDKMSDELSEWSEVAYDELAIAAI-----KA
rice	AHFVVKVALLDTIKEEVPSMSPAMKSAWEAYYDLHVAIAIKOE---MKPAP

(b) MUSCLE (3.6) multiple sequence alignment

```

beta globin -----MVHLTPEEKSAVTLWGKVNVD-----EVGGEALGRLLVVPWTRFFES-PG
myoglobin -----MGLSDGEWQLVLNVWGKVEADIPGHGQEVILRLFKGHPETLEKFDK-FK
neuroglobin -----MERPEELLRQSWAVRSPSFLHGTLVLFARLFALEPDLLPLFYQNCR
soybean -----MERPEKQALVSSFEAFKANIPQYSVVFITSILEKAPAKDLFSF-LA
rice MALVEDNNAVAVSFSEQEAQLKSLWALKDSDANIALRFLFKTIEVPASASQMSF-SL

```

beta globin DLSTPDVAMGNPKVAKGKGVGLG--SDGLAHLNLTNGKPTATLSLEA¹CKDLH--VDPE
 myoglobin HLKSEDEMKA²SELKKHGA³TVTLTAL--GGIKKKGHHEA⁴EKPLAQGNATKH⁵--IPVK
 neuroglobin GSGSPEDCI⁶SSPFLLE⁷HLKVMVLI⁸--DAATVLE⁹LSLSELY¹⁰LAGKHKRA¹¹GVKLT
 soybean NSGVPD--¹²TNPKLTGHAE¹³KLFA¹⁴LRD¹⁵SGAK¹⁶ASGTGGV¹⁷--AALGSGVH¹⁸AGKVL¹⁹
 rice NSDVP--LE²⁰KNPKLKT²¹RAMSV²²FMTECAA²³GL²⁴RGAKGVTV²⁵TR²⁶TLKLGAT²⁷HLAY²⁸GVGD²⁹

```

beta globin  NFRLGNVLVCVLAHHFGKE-FTPPVQAAYQKVVAGVANALAHKYH-----
myoglobin   YLEFISECIQIVLQSKHFGD-FGADAQAMNKALYLFGRKDMANSYKELGFGQ
neuroglobin SFSTVGESLLYMLEKCLGPA-FTPATRAANSQLYFGAVVQMSRGNWGE-----
soybean     QFVVVKELAKLTIKAAVQDK-WSDELRANFEVAYDELAAAIKKA-----
rice        HFVEVVKFALLDTIKEEVPADMNSPAMKSAHSEAYDHLVAAIKQEMKPAE--

```

(d) CLUSTAL FORMAT for T-COFFEE Version 5.13

```

beta globin -----MVHLTPEEKSAVTALMGKVNVD-----EVGGEALGRLLVVPWTQRFP-SG
myoglobin -----MGLSDGEGWLVLNVNGKVEADIPGHGQEVILRLPKFGPKETLEKDF-KF
neuroglobin -----MERPEKEDQLVSSSPRAVSNP1LEHVTFLARLFAEDPLLFLQYNCR
soybean -----MVAFTEPQALIRSSGPEAFKANIPQVSVYFTLILEKAPADLDS-FLA
rice -----MALVEDNNAVAFSPSEEGEALVLKSNALKDKD2SANIALRFPIKIFEPVAPASQMS-FLR

```

beta globin DLSTPDAVMGNPKYKAHGKKLVGASPDLAHLNDL---KGF---ATSLERDCKLHV
 myoglobin HLKSEDEMKAS**EDLKKHGATVLTAL---**GGILKKKGHEAE---**IKPLAQSHATKHKIP**
 neuroglobin QFSPSEDLSS**LDLHRLIKVMVLDATN**TVDEL---SS**LEYLSSAGRKH-R**AVGVG
 soybean NGVDP---**INPKLTGHAHLFALVDR**SSGQGLKASSTVAD---AALG**SVKQATV**E
 rice NSDVP---LE**INPKLKT**HAMSVFVMT**CEAAQLR**KAGVTVTRDT**TLKRLGATLKYGVG**

beta globin ENFRLLGNVLVCLAHHF-GKEFTPPVQAAQYQVVAVGVANALAHKYH----

myoglobin KYLEFISECIIQVLQSKH-PGDFGADAGQMNAKELFRKDMASNYKELGFQG

neuroglobin SSFSTVGESILYMLEKCL-GPAFTPATRAANSQLYGAVVQMSRGGW----E

soybean Q-FVVVKEALLTKIAAV-GDKMSSRANEVAYDIAAAIKK-----

rice H-PSVVKFALDITTKRPADMSDAMSSEKSAEYDTHLVAATKQ-----MKPAE

ProbCons: Probabilistic consistency-based multiple sequence alignment

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

January 15, 2020