FASMA: A Service to Format and Analyze Sequences in Multiple Alignments

Susan Costantini[1,2], Giovanni Colonna[2], and Angelo M. Facchiano[1,2]*

[1] Laboratory of Bioinformatics and Computational Biology, Institute of Food Science, CNR, via Roma 52 A/C, 83100 Avellino, Italy; [2] CRISCEB, Research Center of Computational and Biotechnological Sciences, Second University of Naples, via Costantinopoli 16, 80138 Naples, Italy.

Multiple sequence alignments are successfully applied in many studies for understanding the structural and functional relations among single nucleic acids and protein sequences as well as whole families. Because of the rapid growth of sequence databases, multiple sequence alignments can often be very large and difficult to visualize and analyze. We offer a new service aimed to visualize and analyze the multiple alignments obtained with different external algorithms, with new features useful for the comparison of the aligned sequences as well as for the creation of a final image of the alignment. The service is named FASMA and is available at http://bioinformatica.isa.cnr.it/FASMA/.

Key words: multiple alignment, sequence analysis, web tools

Introduction

Multiple alignments of nucleic acid and protein sequences are basic investigations widely used for a large number of studies, such as phylogeny, profile construction, structure prediction, and sequence/structure activity relationships. In the last years, many programs and services have been developed to align multiple sequences and to edit and analyze the obtained multiple alignments (*1*). The most commonly used methods for creating multiple sequence alignments are based on the progressive-alignment strategy, which first estimates a phylogenetic tree and then constructs an alignment of the sequences (named "profile") following the order in the tree. The best known of these methods is ClustalW (*2*). There are variants of this approach, such as T-Coffee, which builds a library of both local and global alignments of every pair of sequences and uses a library-based score for aligning two profiles (*3*), and MUSCLE, which applies horizontal refinement to the built alignment (*4*). Specific tools to help in the viewing and editing of the alignments have also been developed, including ALSCRIPT and CHROMA that visualize multiple sequence alignments (*5*, *6*), as well as SEAVIEW, Jalview, and CINEMA that allow editing of multiple sequence alignments (*7*–*9*).

In this study, we present a new web service devoted to the analysis and customized visualization of multiple sequence alignments, named FASMA, which offers new useful functions to the scientific community interested in sequence analysis and comparison.

Results and Discussion

FASMA can input multiple sequence alignments in a variety of common formats and offer many options able to format and visualize large multiple sequence alignments, evidencing specific residues or conserved patterns by default or user-provided coloring schemes. The user-friendly formatted alignment can be exported to create figures for manuscripts or presentations. Moreover, FASMA also enables users to compare and analyze the aligned sequences, not only for the global similarity but also in terms of conservation in specific positions and pairwise similarity.

Formatting options

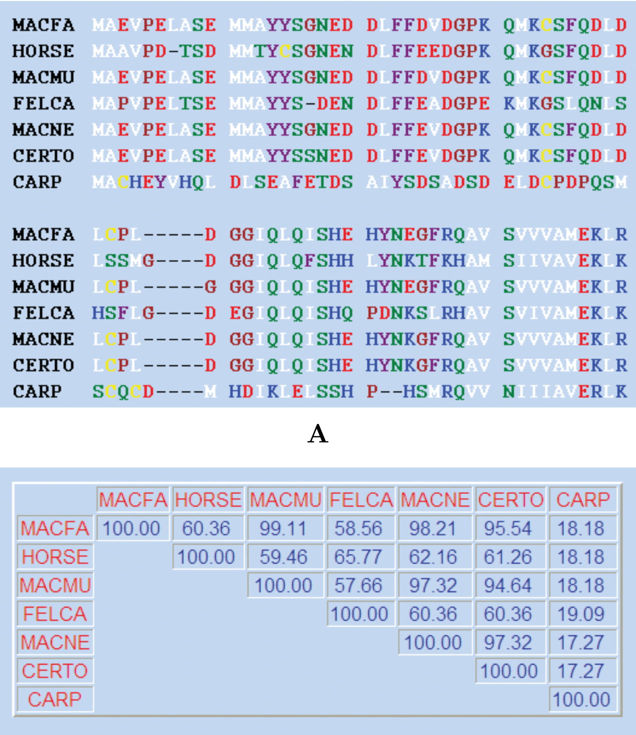
Users can choose one among three different styles to format the alignment. In the first format style, FASMA converts the alignment in GCG MSF format: it reports on each line the sequence names and 50 residues with an empty space between blocks of 10 nucleotides or amino acids. The gap regions are indicated with the period character ("."). The coloring

***Corresponding author.**
E-mail: angelo.facchiano@isa.cnr.it

scheme or a simple black and white option may be selected. The residues are colored according to the chemical species they represent. The default colors for the nucleotides are green for A, blue for C, orange for G, and red for T. The amino acids are reported using the following colors: polar positive (H, K, R) in blue, polar negative (D, E) in red, polar neutral (S, T, N, Q) in green, non-polar aliphatic (A, V, L, I, M) in white, non-polar aromatic (F, Y, W) in purple, C in yellow, and P and G in brown. In the second format style, users may customize the alignment format by choosing to report on each line the residues in blocks of tens and by indicating the number of amino acids or nucleotides for each line of the alignment, from 30 to 90 residues, or all residues of each sequence (Figure 1A). The color of residues can be chosen among five different options, including the simple black and white option, the coloring scheme present in the first format case, the hydrophobicity code according to Kyte and Doolitle (*10*), and two alternative color schemes in which users can define the color for every residue type or for specific residues present in functional positions (catalytic or active sites). Finally, users may also decide whether to indicate the gap regions with dash ("–") or period (".") characters, and whether to insert an empty space after blocks of 10 residues. In the third format style, users may choose the order for reporting the sequences in the multiple alignment and how to format it by choosing among the same options indicated in the second case. This option is very useful when the multiple alignment comprises many sequences and is difficult to compare sequences all together. In fact, it allows to collect the sequences in the basis of species (such as mammals, fish, and birds) or to easily compare some specific sequences that are interesting for the user. The formatted alignment obtained by any of the three styles can be saved and used by common editor and graphics software for creating figures for articles and presentations (see Figure 1A for example).

Analyzing options

FASMA offers three possible options to analyze multiple sequence alignments. In the first option, FASMA shows a table with the sequence names and the occurrence (expressed as number or frequency) of each nucleotide or amino acid in every sequence. In the second option, users can look for information about a specific position of the alignment. In the result page FASMA will report two tables. In the first table, for each sequence name, the nucleotide or amino acid type present in the required position is shown, using the coloring scheme mentioned above. In the second table, FASMA reports the number and the frequency in the selected position of each nucleotide or amino acid. This option allows users to know the sequence conservation for that position and verify in what groups (or species) the same residue is present. This analysis may be useful to highlight the conservation of a residue, which often indicates a structural and functional importance, not only for the whole alignment but also for sequence groups. This information can also be useful to select the order of the sequences in the alignment for the choice of the final format of visualization. In the third option, FASMA shows a table that reports for each sequence the identity percentage of the sequence between itself and the other (Figure 1B). With this option, users can quickly have a quantitative evaluation of the similarity of each sequence with respect to all the others present in the multiple alignment.



```
MACFA  MAEVPELASE MMAYYSGNED DLFFDVDGPK QMKCSFQDLD
HORSE  MAAVPD-TSD MMTYCSGNEN DLFFEEDGPK QMKGSFQDLD
MACMU  MAEVPELASE MMAYYSGNED DLFFDVDGPK QMKCSFQDLD
FELCA  MAPVPELTSE MMAYYS-DEN DLFFEADGPE KMKGSLQNLS
MACNE  MAEVPELASE MMAYYSGNED DLFFEVDGPK QMKCSFQDLD
CERTO  MAEVPELASE MMAYYSSNED DLFFEVDGPK QMKCSFQDLD
CARP   MACHEYVHQL DLSEAFETDS AIYSDSADSD ELDCPDPQSM

MACFA  LCPL-----D GGIQLQISHE HYNEGFRQAV SVVVAMEKLR
HORSE  LSSMG----D GGIQLQFSHH LYNKTFKHAM SIIVAVEKLK
MACMU  LCPL-----G GGIQLQISHE HYNEGFRQAV SVVVAMEKLR
FELCA  HSFLG----D EGIQLQISHQ PDNKSLRHAV SVIVAMEKLK
MACNE  LCPL-----D GGIQLQISHE HYNKGFRQAV SVVVAMEKLR
CERTO  LCPL-----D GGIQLQISHE HYNKGFRQAV SVVVAMEKLR
CARP   SCQCD----M HDIKLELSSH P--HSMRQVV NIIIAVERLK
```

	MACFA	HORSE	MACMU	FELCA	MACNE	CERTO	CARP
MACFA	100.00	60.36	99.11	58.56	98.21	95.54	18.18
HORSE		100.00	59.46	65.77	62.16	61.26	18.18
MACMU			100.00	57.66	97.32	94.64	18.18
FELCA				100.00	60.36	60.36	19.09
MACNE					100.00	97.32	17.27
CERTO						100.00	17.27
CARP							100.00

Fig. 1 Examples of output results. **A**. The formatted multiple alignments obtained with the second format style option. Color options allow to use predefined as well as user-defined colors. **B**. The table of amino acid identity percentage between all the aligned sequences.

Materials and Methods

The core of FASMA is a CGI script written in Perl language with the aim of producing an user-friendly web tool to format and analyze multiple sequence alignments provided in different formats. Users can select the sequence type indicating "DNA" or "Protein", choose the input format among those available types including ClustalW, GDE (*11*), FASTA (*12*), PIR, and GCG MSF, paste the alignment in the box, and choose the possible options for formatting or analyzing the sequences in the multiple alignment. The results appear immediately in an HTML page. The web pages report examples of input multiple alignments and output results. The service is available at http://bioinformatica.isa.cnr.it/FASMA/.

Acknowledgements

This work was partially supported by the CNR-Bioinformatics Project and the ISS-CNR "Oncoproteomica" Project (No. 527B/2A/10).

Authors' contributions

SC developed this service and prepared the manuscript. GC critically discussed and revised the project. AF conceived the idea of developing this service and assisted with manuscript preparation. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Notredame, C. 2002. Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics* 3: 131-144.
2. Thompson, J.D., *et al.* 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 22: 4673-4680.
3. Notredame, C., *et al.* 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* 302: 205-217.
4. Edgar, R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5: 113.
5. Barton, G.J. 1993. ALSCRIPT: a tool to format multiple sequence alignments. *Protein Eng.* 6: 37-40.
6. Goodstadt, L. and Pointing, C.P. 2001. CHROMA: consensus-based colouring of multiple alignments for publication. *Bioinformatics* 17: 845-846.
7. Galtier, N., *et al.* 1996. SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput. Appl. Biosci.* 12: 543-548.
8. Clamp, M., *et al.* 2004. The Jalview Java alignment editor. *Bioinformatics* 20: 426-427.
9. Parry-Smith, D.J., *et al.* 1998. CINEMA—a novel colour interactive editor for multiple alignments. *Gene* 221: GC57-63.
10. Kyte, J. and Doolitle, R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105-132.
11. Smith, S.W., *et al.* 1994. The genetic data environment: an expandable GUI for multiple sequence analysis. *Comput. Appl. Biosci.* 10: 671-675.
12. Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* 85: 2444-2448.