BIOINFORMATICS APPLICATIONS NOTE
Vol. 20 no. 3 2004, pages 433–435
DOI: 10.1093/bioinformatics/btg444

VisCoSe: visualization and comparison of consensus sequences

Michael Spitzer[1,], Georg Fuellen[1], Paul Cullen[2,3] and Stefan Lorkowski[2,4]*

[1]Integrated Functional Genomics, IZKF, University Hospital Münster, Von-Esmarch-Str. 56, 48149 Münster, Germany, [2]Institute of Arteriosclerosis Research, Domagkstr. 3, 48149 Münster, Germany, [3]Ogham GmbH, Mendelstr. 11, 48149 Münster, Germany and [4]Institute of Biochemistry, Wilhelm-Klemm-Straße 2, 48149 Münster, Germany

Received on 23 July 2003; accepted on 10 September 2003
Advance Access Publication January 22, 2004

ABSTRACT

Summary: We introduce visualization and comparison of consensus sequences (VisCoSe) as a WWW service and a stand-alone command line Perl script for visualizing and comparing consensus sequences of protein and nucleotide sequences. VisCoSe is the only interface available that simultaneously calculates consensus sequences of multiple data sets and automatically compares these consensus sequences. Furthermore, VisCoSe allows visualization of chemical properties of amino acids.

Availability: http://viscose.ifg.uni-muenster.de/

Contact: michael.spitzer@uni-muenster.de

The identification of consensus sequences is an important step towards the identification of conserved motifs that may be characteristic parts of protein domains. The identification of well-defined protein motifs or domains allows the classification of proteins into families. Such classifications can be used to assign putative physiological roles to proteins (Schug *et al*., 2002). Well-known examples of such motifs are the Walker motifs of the ATP-binding cassette (ABC) domain, which is a characteristic of the abundant ABC protein family (Walker *et al*., 1982). The analysis of consensus sequences may also help to identify motifs that are a characteristic of a subfamily within a larger protein family. Systematic analyses of consensus sequences of protein subfamilies or groups of homologous proteins may reveal motifs that play an essential role in the physiological function of the groups investigated.

Much effort has gone into the identification of protein domains and the classification of proteins into families, but only a few tools such as Sequence Logos (Schneider and Stephens, 1990) or BOXSHADE (Hofmann and Baron; http://www.ch.embnet.org/software/BOX_form.html) have been published that allow calculation and visualization of consensus sequences. Using these tools, visualization of sequence alignments and extraction of information on conserved residues is often impossible. Although these tools produce high-quality postscript output, the visualization of a consensus sequence in a single window together with the exact conservation rates for each single column of the alignment is difficult, particularly for long sequences. Furthermore, none of these tools allows the comparison of multiple consensus sequences.

Visualization and comparison of consensus sequences (VisCoSe) provides an easy way for calculating and visualizing consensus sequences of nucleotide and protein sequences. It calculates and visualizes the comparison of multiple consensus sequences. VisCoSe allows both the visualization of consensus sequences and conservation rates for single as well as for multiple alignments/data sets. Unaligned data sets are aligned by VisCoSe using the multiple alignment program *fftnsi* from the MAFFT library (Katoh *et al*., 2002). VisCoSe is available as a stand-alone command line Perl script and as an HTML-based web application. The resulting graphical presentation of the consensus sequences and the conservation rates of each amino acid is clear and straightforward (Fig. 1). In addition, VisCoSe allows one to calculate and visualize the chemical properties of the amino acids in the consensus sequences by use of simplified amino acid alphabets. Another advantage is the presentation of the results by HTML. Since both the sequences and the alignments are shown continuously in full length, it is easy to identify highly conserved regions and to assign them to the sequence. For printing or publication purposes, it is possible to define line breaks in the output alignment.

Use of VisCoSe is simple and the following input options are available:

(1) A single unaligned data set of protein or nucleic acid sequences in FASTA format can be used. In this case, VisCoSe uses MAFFT to align the data set. Multiple data sets can be 'simulated' by using a single data set

*To whom correspondence should be addressed.

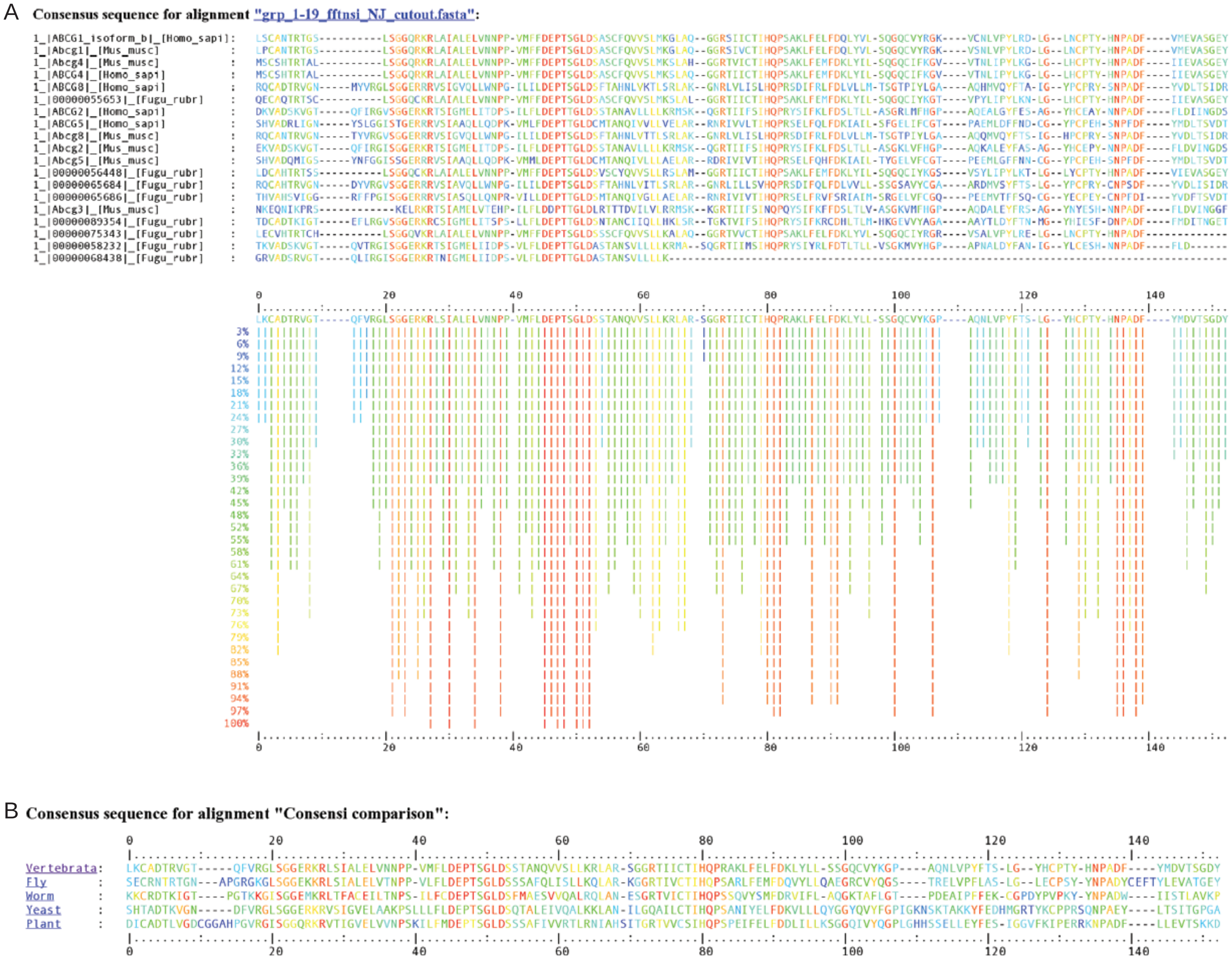
A
Consensus sequence for alignment "grp_1-19_fftnsi_NJ_cutout.fasta":
1_|ABCG1_isoform_b|_[Homo_sapi]
1_|Abcg1|_[Mus_musc]
1_|Abcg4|_[Mus_musc]
1_|ABCG4|_[Homo_sapi]
1_|ABCG8|_[Homo_sapi]
1_|00000055653|_[Fugu_rubr]
1_|ABCG2|_[Homo_sapi]
1_|ABCG5|_[Homo_sapi]
1_|Abcg8|_[Mus_musc]
1_|Abcg2|_[Mus_musc]
1_|Abcg5|_[Mus_musc]
1_|00000056448|_[Fugu_rubr]
1_|00000065684|_[Fugu_rubr]
1_|00000065686|_[Fugu_rubr]
1_|Abcg3|_[Mus_musc]
1_|00000089354|_[Fugu_rubr]
1_|00000075343|_[Fugu_rubr]
1_|00000058232|_[Fugu_rubr]
1_|00000068438|_[Fugu_rubr]
0
20
40
60
80
100
120
140
3%
6%
9%
12%
15%
18%
21%
24%
27%
30%
33%
36%
39%
42%
45%
48%
52%
55%
58%
61%
64%
67%
70%
73%
76%
79%
82%
85%
88%
91%
94%
97%
100%
B
Consensus sequence for alignment "Consensi comparison":
Vertebrata
Fly
Worm
Yeast
Plant

Fig. 1. (**A**) Parts of protein sequences of selected mammalian ABCG proteins were aligned and visualized using VisCoSe. The histogram below the consensus illustrates the grade of conservation. (**B**) Multiple alignment of consensus sequences based on ABCG protein data sets derived from vertebrate, fly, worm, yeast and plant proteomes. This figure can be viewed in colour as supplementary data at *Bioinformatics* online.

as input and specifying ‘groups’, which divide the data set into two or more single subsets. VisCoSe then produces sequence alignments for each subset. In this case, the resulting consensus sequences are subjected to multiple alignment.

(2) A pre-calculated alignment of protein or nucleotide sequences can be entered that can be produced from a method different than MAFFT, such as CLUSTAL W (Thompson *et al*., 1994) or DIALIGN (Morgenstern, 1994). Like before, groups can be defined to specify independent subsets.

(3) Using the stand-alone version, it is possible to enter two or more different data sets consisting of either aligned or unaligned sequences. In the latter case, the sequences are automatically aligned.

VisCoSe automatically calculates the consensus for each alignment following the relative majority rule and compiles the consensus into FASTA format. The grade of conservation at each position is compiled into a color scheme ranging from dark blue (indicating no conservation) to red (indicating high conservation) and a histogram indicating grades of conservation.

To access the results generated by VisCoSe, the following output is produced:

1. A zipped TAR archive containing all results and intermediate data such as aligned data subsets.
2. The sequence data set and user-defined groups of sequences as provided by the user.
3. The consensus of each user-defined group (or input data set) in FASTA format and in colorized HTML format.
4. The alignment of the consensus sequences colorized and illustrated according to their column-wise conservation rate.
5. The alignment and consensus sequences in a reduced code giving information on chemical properties of the amino acid residues if chosen.

The user controls the VisCoSe output using options mainly affecting realigning behavior of VisCoSe, the presentation of the alignment on which a consensus is based, the appearance of the conservation histogram and calculation of the consensus sequences in a reduced code illustrating chemical properties of the amino acid residues. This is described in more detail on the Web site.

Overall, VisCoSe is a helpful tool to identify conserved motifs from a set of sequences and to visualize them in a user-friendly format. To our knowledge, VisCoSe is the only interface available that simultaneously calculates consensus sequences of multiple data sets and automatically compares these consensus sequences. Thus, it is possible, e.g. to compare different subfamilies of large protein families or to compare members of protein families sampled from different groups of organisms.

ACKNOWLEDGEMENTS

P.C. and S.L. were supported by the European Union (grant no. QLG2-CT-1999-01007). We thank Reinhard Voss for advice and critical reading.

REFERENCES

Katoh,K., Misawa,K., Kuma,K. and Miyata,T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.*, **302**, 3059–3066.

Morgenstern,B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics* **15**, 211–218.

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.

Schug,J., Diskin,S., Mazzarelli,J., Brunk,B.P. and Stoeckert,C.J. (2002) Predicting gene ontology functions from ProDom and CDD protein domains. *Genome Res*., **12**, 648–655.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (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.

Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J*., **1**, 945–951.