ENDscript: a workflow to display sequence and structure information

Patrice Gouet¹,* and Emmanuel Courcelle²

¹Laboratoire de Bio-Cristallographie, IBCP-CNRS UMR 5086, 7 passage du Vercors, 69367 Lyon Cedex 07, France and ²Laboratoire de Biologie Moléculaire et des Relations Plantes Microorganismes, BP 27 Chemin de Borde Rouge, 31326 Castanet Tolosan, France

Received on October 16, 2001; revised on November 4, 2001; accepted on December 17, 2001

ABSTRACT
Summary: ENDscript is a web server grouping popular programs such as BLAST, Multalin and DSSP. It uses as query the co-ordinates file of a protein in Protein Data Bank format and generates PostScript and png figures showing: residues conserved after a multiple alignment against homologous sequences, secondary structure elements, accessibility, hydropathy and intermolecular contacts. Thus, the user can relate quickly 1D, 2D and 3D information of a protein of known structure.
Availability: http://genopole.toulouse.inra.fr/ENDscript
Contact: p.gouet@ibcp.fr

BACKGROUND

ENDscript is a web tool in the tradition of PDBsum (Laskowski et al., 1997) a database of the known 3-dimensional structures of protein, and of the suite of programs developed in the group of Geoff Barton to perform sequence alignments (AMPS, Barton and Sternberg, 1987) and colouring (ALSCRIPT, Barton, 1993). ENDscript takes as input a protein co-ordinate file in Protein Data Bank format (Berman et al., 2000) and runs the following programs: DSSP (Kabsch and Sander, 1983) to extract secondary structure elements and to compute accessibility per residues, CNS (Brünger et al., 2000) to calculate distances for intermolecular contacts and protein–ligand interactions, BLAST (Altschul et al., 1997) to search databases such as the SWISS-PROT (Bairoch and Apweiler, 2000) or the PDBaa (sequences derived from the 3-dimensional Protein Data Bank) for homologous sequences, Multalin (Corpet, 1988) or ClustalW (Thompson et al., 1994) to perform multiple sequence alignments, ESPript (Gouet et al., 1999) BobScript (Esnouf, 1997) and MolScript (Kraulis, 1991) to render all this information in coloured PostScript files, ImageMagick (ImageMagick Studio LLC) to convert these files to images in png format. Such figures are sought routinely by scientists, who want to understand relations between the 3D structure of a protein and its sequence. ENDscript aims to help them to better deal with the ever growing flow of information due to the development of structural genomics. The ENDscript cgi interface is based on the one of ESPript which was completely rewritten. A few additional programs were written in Fortran and in Perl, to pipe information between the different stages of ENDscript. The server runs on Unix and Linux platforms.

RUNNING ENDSHEET

The user can upload the co-ordinates file of a protein from his workstation to the ENDscript server or enter a PDB identifier. X-ray structures are supported as well as NMR structures. A click on the RUN button of the interface leads to the first PostScript figure produced by ESPript. As an example, Figure 1(a) was obtained from the PDB file deposited under the code 1G94. It corresponds to a psychrophilic alpha-amylase solved with a bound hepta-saccharide and crystallized with one monomer in the asymmetric unit (Aghajari, 1998). Sequences of each protein chain contained in the PDB query (CHAIN A in our example) are checked by a home-made program named SPDB. Each monomer appears in the resulting figure in a one letter code sequence, displayed with additional information. Thus, secondary structure elements are shown above the sequence, along with grey stars marking residues modelled in alternate conformations in crystallographic structures. Relative accessibility is shown below the sequence by a blue-coloured bar. A second bar presents the figure of hydropathy, calculated from the sequence according to the algorithm of Kyte and Doolittle (1982). Intermolecular contacts are figured by characters on the bottom line. According to their colour and format, the user can deduce that a residue is involved in a short/long, crystallographic/non-crystallographic contact with a neighbouring molecule. A different code is used on the same line, to highlight residues in interaction...
with hetero-compounds as well as cysteins involved in disulphide bridges.

In a next step, the user can check a box labelled ‘enable the BLAST search’. After a click on the RUN button, a new ESPript figure is produced, displaying the result of a multiple alignment between the sequence of the first homomer of the query, against matching sequences detected by a BLAST search in either the SWISSPROT or the PDBaa databases. The user can show on the alignment, the secondary structure elements of all sequences are displayed, (c) a 3D representation of the full protein, coloured according to similarity from white (low) to black (high).

ACKNOWLEDGEMENT
We thank Dr Richard Haser and Dr Daniel Kahn for their support in the development of ENDscript.

REFERENCES

Fig. 1. Samples output of ENDscript obtained from the PDB file 1G94, showing (a) the partial sequence of 1G94 (448 residues in total) with (i) on the top secondary structures (squiggles are helices, arrows beta strands, TT letters beta turns) alternate residues (grey stars) (ii) on the bottom a bar of accessibility (black is accessible, grey intermediate, white buried), a bar of hydrophytis (black is hydrophobic, grey hydrophilic, white neutral) and a line of molecular contacts (letters are crystallographic contacts, quotes are protein-ligand interactions, digits are disulphide bridges), (b) an excerpt from a multiple alignment, obtained after a BLAST search against the Protein Data Bank sequence database (PDBaa); secondary structure elements of all sequences are displayed, (c) a 3D representation of the full protein, coloured according to sequence similarity from white (low) to black (high).