BIOINFORMATICS

Vol. 15 no. 4 1999
Pages 305–308

ESPript: analysis of multiple sequence alignments in PostScript

Patrice Gouet[1,*], Emmanuel Courcelle[2], David I. Stuart[1,3] and Frédéric Métoz[4]

[1]Laboratory of Molecular Biophysics, The Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK, [2]Groupe de Cristallographie Biologique, Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse Cedex, France, [3]Oxford Centre for Molecular Sciences, New Chemistry Building, South Parks Road, Oxford OX1 3QT, UK and [4]Institut de Biologie Structurale, 41 Avenue des Martyrs, 38027 Grenoble Cedex 1, France

Received on October 22, 1998; revised on December 15, 1998; accepted on December 16, 1998

Abstract

Motivation: *The program ESPript (Easy Sequencing in PostScript) allows the rapid visualization, via PostScript output, of sequences aligned with popular programs such as CLUSTAL-W or GCG PILEUP. It can read secondary structure files (such as that created by the program DSSP) to produce a synthesis of both sequence and structural information.*

Results: *ESPript can be run via a command file or a friendly html-based user interface. The program calculates an homology score by columns of residues and can sort this calculation by groups of sequences. It offers a palette of markers to highlight important regions in the alignment. ESPript can also paste information on residue conservation into coordinate files, for subsequent visualization with a graphics program.*

Availability: *ESPript can be accessed on its Web site at http://www.ipbs.fr/ESPript. Sources and helpfiles can be downloaded via anonymous ftp from ftp.ipbs.fr. A tar file is held in the directory pub/ESPript.*

Contact*: gouet@ipbs.fr*

Introduction

The Internet allows biologists to browse a variety of ever-growing databases on-line. This enables them to search, compare and retrieve protein sequences [e.g. from the SWISSPROT bank (Bairoch and Apweiler, 1997)] and three-dimensional (3D) structures [from the Protein Data Bank (PDB) (Bernstein *et al*., 1977)]. The number of entries deposited with these databases is increasing rapidly (1500 additional entries in the PDB for 1997–98), increasing the probability that any sequence will be homologous to one whose 3D structure is known. In the absence of a determined tertiary structure, secondary structures can be predicted with reasonable reliability from amino acid sequences by software such as PHD (Rost, 1996).

ESPript, Easy Sequencing in PostScript, is a program in the tradition of ALSCRIPT (Barton, 1993), which renders sequence similarities and secondary structure information for analysis and publication purposes. Most of the assignments are made by default in ESPript and a user familiar with the program can obtain Figure 1 in a few minutes. ESPript is not a sequence editor like CINEMA (Attwood *et al*., 1997) or CLUSTAL X (Thompson *et al*., 1997), but it can help to optimize an alignment, by displaying on the same figure the secondary structure information (observed or predicted) of each aligned sequence.

A first version of the program was produced in 1993, at the Institut de Biologie Structurale, Grenoble. Since then, ESPript has been rewritten in the Laboratory of Molecular Biophysics, Oxford, and is now developed in the Groupe de Cristallographie Biologique, Toulouse. ESPript's input consists of pre-aligned sequences and files defining secondary structures. Its output is a colourful PostScript file. The distributed package consists of the program source, a cgi script using a library from Lincoln D.Stein, Cold Spring Harbour Laboratory, for use with a WWW server, a manual written in hypertext, and examples related to a study made on orbiviruses in Oxford (Grimes *et al*., 1998).

General description of the program

The program can read up to 98 sequences aligned on 2000 columns, retaining the alignment of the input file. Sequences can be displayed on up to 10 pages of PostScript. Parameters

*Present address: Institut de Pharmacologie et de Biologie Structurale, 205 route de Narbonne, 31077 Toulouse Cedex, France

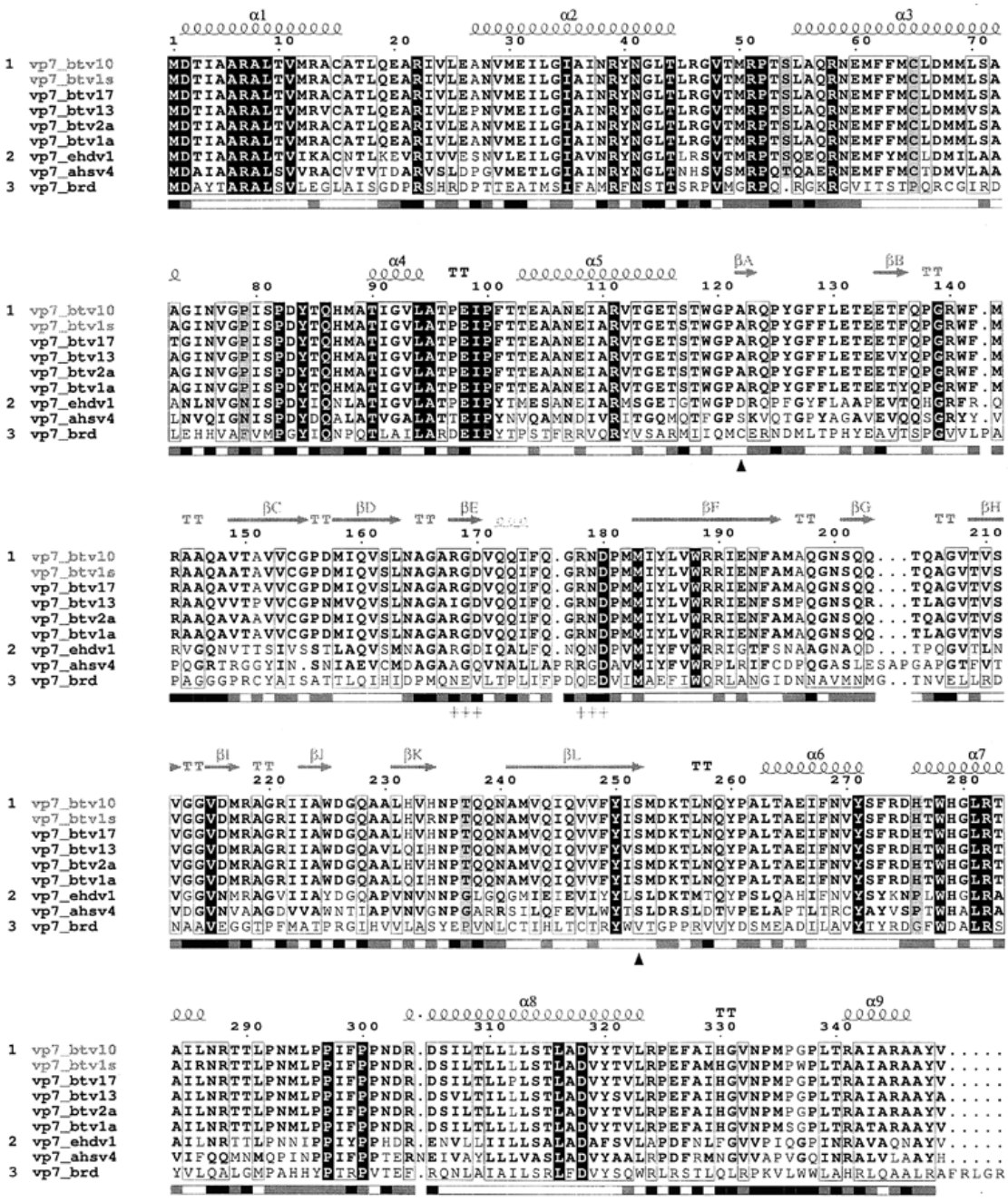

Fig. 1. An ESPript output, obtained from orbivirus sequences extracted from the SWISSPROT data bank (Bairoch and Apweiler, 1997) and aligned with CLUSTAL-W (Thompson *et al*., 1994). Sequences are divided into three groups according to similarity. Residues strictly conserved have a black background, residues well conserved within a group according to a Risler matrix (Risler *et al*., 1988) are indicated by black bold letters and the remainder are in regular black (an inner group score cannot be calculated for group 3 which is made of a single sequence; no residue is written with black bold letters in this group); residues conserved between groups are boxed and residues conserved within a group, but showing significant differences between groups, are on a grey light background. Symbols above blocks of sequences correspond to the secondary structure of protein VP7 of bluetongue virus serotype 10 (Grimes *et al*., 1995). This protein consists of a helical domain and a beta domain, coloured in black and grey, respectively. VP7 of bluetongue virus serotype 1 from South Africa shares the same secondary structure (Grimes *et al*., 1998) and the names of the two sequences are in red. Symbols below blocks of sequences show (i) the limits of the two domains as triangles, (ii) an RGD tripeptide which may be important in cell entry as stars and (iii) the relative accessibility of BTV-10 VP7 as rectangles (accessible residues are in black, intermediate in grey and buried in white).

are fed in through the standard input, which is divided into seven steps.

1. The program asks first for the name of the multiple alignment file. Files generated by CLUSTAL-W (Thompson *et al*., 1994), GCG PILEUP (Wisconsin Package Version 9.0, GCG, Madison), MAXHOM (Sander and Schneider, 1991) and THREADER (Jones *et al*., 1992) are supported. The program offers the possibility of extracting a segment from the input sequences, and of choosing the number assigned to the first residue.
2. The names of one or two files containing secondary structure information can be specified. These files refer to the secondary structures of (i) the first sequence appearing on the PostScript output and (ii) one selected from the remaining sequences. Files generated by DSSP (Kabsch and Sander, 1983), STRIDE (Frishman and Argos, 1995) or PHD (Rost, 1996) are accepted. Helices are symbolized by squiggles, strands by arrows and turns by a T letter on the output. The program automatically numbers the secondary structural units. The relative accessibility can be indicated by symbols, if the secondary structures files were produced via DSSP or PHD.
3. The name of the PostScript output file is given. By default, the output name is that of the multiple sequence file with a '.ps' extension.
4. A scoring scheme for similarities is given. Fully conserved residues are shown on a red background. Similarity scores are calculated, by extracting all possible pairs of residues and by using a Risler (Risler *et al*., 1988), PAM250 (Dayhoff, 1978), BLOSUM62 (Henikoff and Henikoff, 1996) or identity scoring matrix. If $C(i,j)$ is the score for aligning a residue i with a residue j, a similarity score $Sc = S\ C(i,j)/S\ (i,j)$ is calculated for each column. Residues with a score above a user-defined threshold are written in red and boxed in blue, others are in black. Groups of sequences can be defined in step 7 below and additional scores calculated: an inner group score, ISc, equal to Sc within a group; a cross group score, XSc, for all possible pairs between residues of different groups; a total group score, TSc = ISc + XSc, and a difference group score, DSc = ISc – XSc. Residues conserved within a group appear in red (ISc above threshold), residues conserved between groups are boxed in blue (TSc above), residues conserved within a group, but significantly different from one group to the other, are written on yellow boxes (DSc above).
5. The plot layout is defined. The user can specify the size of the font, the number of residues per lines and the centring of the alignment on the paper. Sequence names are written in Times and one-letter code residues are written in Courier. Figure can be in colour or black and white. Portrait or landscape orientation in A4 or A3 formats are supported.
6. Symbols in different colours may be explicitly added at the bottom of the sequences blocks. Important residues, like the RGD segment in protein VP7 of bluetongue virus in Figure 1, can be highlighted (Grimes *et al*., 1995). It is also possible, at this stage, to change the default colours for sequence homology and secondary structure representation.
7. The last stage allows the user to define the displayed sequences and their order of appearance. Sequence groups can be selected to enhance striking similarities (Figure 1).

ESPript is easy to use. In the simplest case, the user merely runs the program on-line, specifies the name of the alignment file (part 1 above) and skips all other steps. This creates a PostScript file with information on sequence identities and similarities. For more complex cases, it is best to prepare a command file or use the html–user interface.

Implementation

The source

ESPript is written in FORTRAN77 and is developed on Silicon Graphics and DEC workstations at the IPBS, Toulouse. The present version, ESPript1.4, can be compiled using f77 or g77 and has been tested on most platforms (Unix, VMS, PC-Windows or Linux).

The html interface

A cgi script written in perl v 5.004 and relying upon CGI.pm v 2.39 or later is provided with the program. This script can be installed on a Web server, as has been done in Toulouse. The users can execute ESPript by filling the fields of an HTML form. Results are presented as hypertext links to PostScript files (or PDB files, if requested; see the paragraph below). These files can be viewed before retrieving, if the browser is properly configured. PostScript files can be converted into other graphics formats (jpeg, tiff, png), using a program such as GHOSTSCRIPT. It is also possible to declare printers available to local users, whilst their access is denied to remote users.

Discussion and conclusion

ESPript offers a few tricks not routinely available in other programs, which can be used to build up information on the output file. It is possible to obtain an output from different files of aligned sequences (Figure 2a) or to select sequences for similarity calculations, which are not displayed on the PostScript (Figure 2b). Information from two or more secondary structure files can be entered and related to a se-

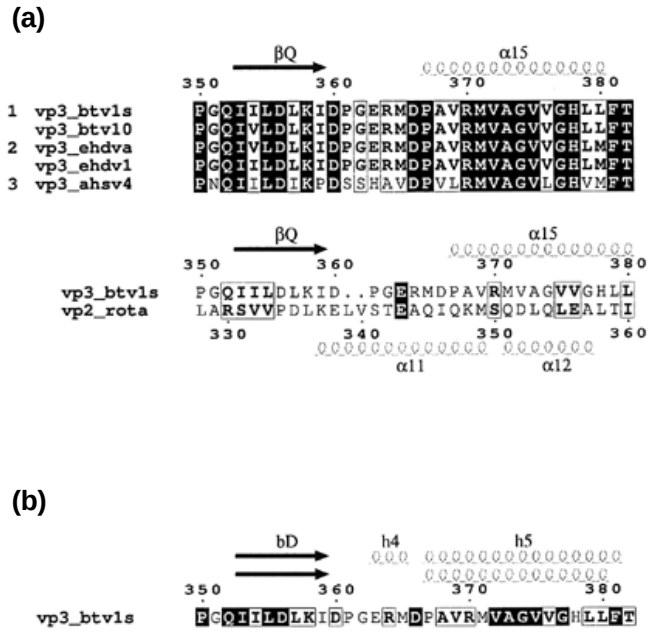

Fig. 2. ESPript outputs extracted from a study made on orbiviruses showing (**a**) the capability of the program to generate a PostScript file from several multiple aligned sequences files and (**b**) its capability to map similarity information on a single sequence for conciseness.

quence chosen by the user (Figure 2a and b). In addition, a PDB file can be produced with temperature factors replaced by similarity scores calculated with ESPript. This file can be passed to a graphics program to represent conserved areas by a colour code.

One can get even more from the program. The PostScript generated by ESPript starts with the definition of a new font, where letters correspond to drawing commands. The program relies on the manipulation of arrays of characters. ESPript generates an output made up of a succession of lines, containing commands to draw letters, digits or symbols in PostScript. It is quite easy to insert a subroutine in the program, to translate new information such as a list of contacts generated by X-PLOR (Brünger *et al*., 1987) into an array of characters. Such a modification was used to generate a figure published in Grimes *et al*. (1998), where residues involved in intermolecular contacts are pointed out by a letter code. This feature has been implemented in the latest version of the program, ESPript 1.4, released in September 1998.

Acknowledgements

ESPript is available thanks to Jean-Pierre Samama, in charge of the Groupe de Cristallographie at the IPBS, Toulouse. Catherine Mazza, EMBL Grenoble, and Jean-Denis Pedelacq, IPBS Toulouse, helped to test the program.

References

Attwood,T.K., Payne,A.W.R., Michie,A.D. and Parry-Smith,D.J. (1997) A Colour INteractive Editor for Multiple Alignments—CINEMA. *EMBnet.news*, **3**.

Bairoch,A. and Apweiler,R. (1997) The SWISS-PROT protein sequence data bank and its supplement TREMBL. *Nucleic Acids Res*., **25**, 31–36.

Barton,G.J. (1993) ALSCRIPT a tool to format multiple sequence alignments. *Protein Eng*., **6**, 37–40.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol*., **112**, 535–542.

Brünger,A.T., Kuriyan,J. and Karplus,M. (1987) Crystallographic R factor refinement by molecular dynamics. *Science*, **235**, 458–460.

Dayhoff,M. (1978) *Atlas of Protein Sequences and Structure*. National Biomedical Research Foundation, Washington, DC, pp. 345

Frishman,D. and Argos,P. (1995) Knowledge-based secondary structure assignment. *Proteins*, **23**, 566–579.

Grimes,J., Basak,A.K., Roy,P. and Stuart,D. (1995) The crystal structure of bluetongue virus VP7. *Nature*, **373**, 167–170.

Grimes,J., Burroughs,N., Gouet,P., Diprose,J.M., Malby,R., Zeintara,S., Mertens,P.P.C. and Stuart,D. (1998) The atomic structure of the bluetongue virus core. *Nature*, **395**, 470–478.

Henikoff,J.G. and Henikoff,S. (1996) Blocks database and applications. *Methods Enzymol*., **266**, 88–105.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) A new approach to protein fold recognition. *Nature*, **358**, 86–89.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J. Mol. Biol*., **204**, 1019–1029.

Rost,B. (1996) PHD: predicting one-dimensional protein structure by profile based neural networks. *Methods Enzymol*., **266**, 525–539.

Sander,C. and Schneider,R. (1991) Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*., **24**, 4876–4882.