BIOINFORMATICS

Vol. 14 no. 7 1998
Pages 617–623

JOY: protein sequence-structure representation and analysis

Kenji Mizuguchi[1], Charlotte M. Deane[1], Tom L. Blundell[1], Mark S. Johnson[2] and John P. Overington[3]

[1]Department of Biochemistry, University of Cambridge, Tennis Court Road, Cambridge CB2 1GA, UK, [2]Department of Biochemistry, University of Turku, Turku, Finland and [3]Computational Chemistry and Protein Crystallography, Pfizer Central Research, Sandwich, Kent CT13 9NJ, UK

Received on March 18, 1998; accepted on May 15, 1998

Abstract

Motivation: *JOY is a program to annotate protein sequence alignments with three-dimensional (3D) structural features. It was developed to display 3D structural information in a sequence alignment and to help understand the conservation of amino acids in their specific local environments.*
Results: *The JOY representation now constitutes an essential part of the two databases of protein structure alignments: HOMSTRAD (http://www-cryst.bioc.cam.ac.uk/~homstrad) and CAMPASS (http://www-cryst.bioc.cam.ac.uk/~campass). It has also been successfully used for identifying distant evolutionary relationships.*
Availability: *The program can be obtained via anonymous ftp from torsa.bioc.cam.ac.uk from the directory /pub/joy/. The address for the JOY server is http://www-cryst.bioc.cam.ac.uk/cgi-bin/joy.cgi.*
Contact: *kenji@cryst.bioc.cam.ac.uk*

Introduction

The comparative study of molecular sequences and structures has provided many insights into protein folding, stability and evolution. The display of sequence alignments (e.g. Parry-Smith and Attwood, 1992; Barton, 1993) can be very useful for encapsulating the essential features of a particular fold or family, and highlighting features of particular biological interest (e.g. active site residues, glycosylation sites). However, when the three-dimensional (3D) structure of a particular sequence is known, much important information is lost either in the representation as a single sequence, or as an alignment. There is, therefore, a need to marry the sequence and structural information within a common representation.

As part of our research into protein evolution, we have developed a program, JOY, to annotate automatically sequence alignments with particular structural features. The basic idea of the program was to display, in a highly compact notation, the local environment of a residue in terms of a number of structural descriptors that are considered to be among the most important to the fold. A typesetting code (Figure 1) was developed that can unambiguously represent the structural environment, with emphasis given to residues that have many interactions with other residues within the protein (and are thus likely to be important to the fold). JOY was initially developed (Overington *et al*., 1990, 1992) as an aid in the interpretation of COMPARER alignments (Sali and Blundell, 1990; Zhu *et al*., 1992) where many of the structural features displayed by JOY are used in the structural alignment procedure. The program has since been used in a number of other research projects (e.g. Hoffren *et al*., 1991; Sali and Blundell, 1993; Emsley *et al*., 1994; Srinivasan *et al*., 1996) and the representation is particularly appropriate for discussions of sequence-structure features within a single family.

solvent inaccessible	UPPER CASE	X
solvent accesible	lower case	x
positive ϕ	*italic*	*x*
cis–peptide	breve	x̆
hydrogen bond to other sidechain	tilde	x̃
hydrogen bond to mainchain amide	**bold**	**x**
hydrogen bond to mainchain carbonyl	underline	x̲
disulphide bond	cedilla	ç
α–helix	red	x
β–strand	blue	x
3_{10}–helix	maroon	x

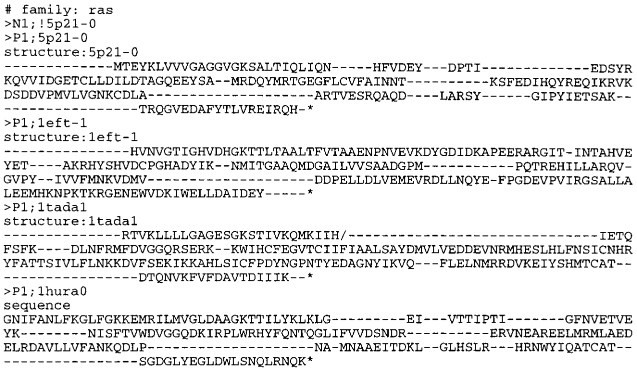
Fig. 1. Key to formatted alignments. Tilde and breve are not displayed in an HTML output. Colours are only for colour PostScript and HTML outputs.

Another key design consideration for the original program was that it should not require the use of colours (for page charges can sometimes be prohibitive) and that it should use a ‘standard’ text processing language so that integration with manuscripts is easy. Initially, LaTeX (Lamport, 1994) was chosen as the output device, but the development of a wide variety of text and graphics processing tools has required a more flexible system. It also became clear that colours could make a qualitative difference in the analysis by adding valuable information. Here we describe a new version of JOY, with significant enhancements, including PostScript and HTML output and a WWW server. The program now plays an essential role in the construction of the two databases of protein structure alignments: HOMSTRAD (HOMologous STRucture Alignment Database; Mizuguchi *et al.*,1998b; http://www-cryst.bioc.cam.ac.uk/~homstrad) and CAMPASS (CAMbridge database of Protein Alignments organised as Structural Superfamilies; Sowdhamini *et al.*, submitted; http://www-cryst.bioc.cam.ac.uk/~campass). The new representation of alignments has also proved to be useful for analysing distantly related proteins.

Description of JOY

JOY is not an alignment program. It acts as a post-processor to a structural (or optionally sequence) alignment program, taking an alignment file and producing annotated alignments. JOY takes an input alignment (or a single sequence) in a format similar to that of the NBRF/PIR format (Figure 2). Several extensions have been introduced to allow easy labelling of the alignment and mixing sequence and structural information. A character sequence can be mixed with amino acid sequences by using the delimiter line '>T1;text'. The line that follows '>P1;' or '>T1;text' must be one of 'sequence', 'structure' or 'text'. The 'sequence' means a sequence with no structural information. The 'structure' means that there is structural information for the sequence and thus the output alignment will include annotations. The 'text' must follow the line '>T1;text'. In the usual NBRF/PIR format, gaps are indicated by hyphens ('-'), but JOY also accepts slashes ('/'). The latter is translated to a blank in the output alignment (Figure 3) and is normally used to indicate a chain break (e.g. missing electron density, multi-chain alignments). The same alignment format is used in the modelling program MODELLER (Sali and Blundell, 1993).

JOY produces a number of output files. The 'featured' alignment can be output in four different formats. The black and white PostScript file is designed for inexpensive print out, while the colour PostScript file (Figure 3) contains more information and is easier to capture various features (see Figure 1). The LaTeX (Lamport, 1994) file is mainly for further text processing and integration with manuscripts. The HTML file is easily visualized with a WWW browser and is used for the HOMSTRAD and CAMPASS databases. In addition to the featured alignment, JOY produces a 'template' file containing data for a profile-based sequence search program PSLAVE (Johnson *et al.*, 1993). JOY also outputs a list of ungapped blocks derived from the input alignment, which can be used as initial equivalences for the structure comparison program MNYFIT (Sutcliffe *et al.*, 1987).



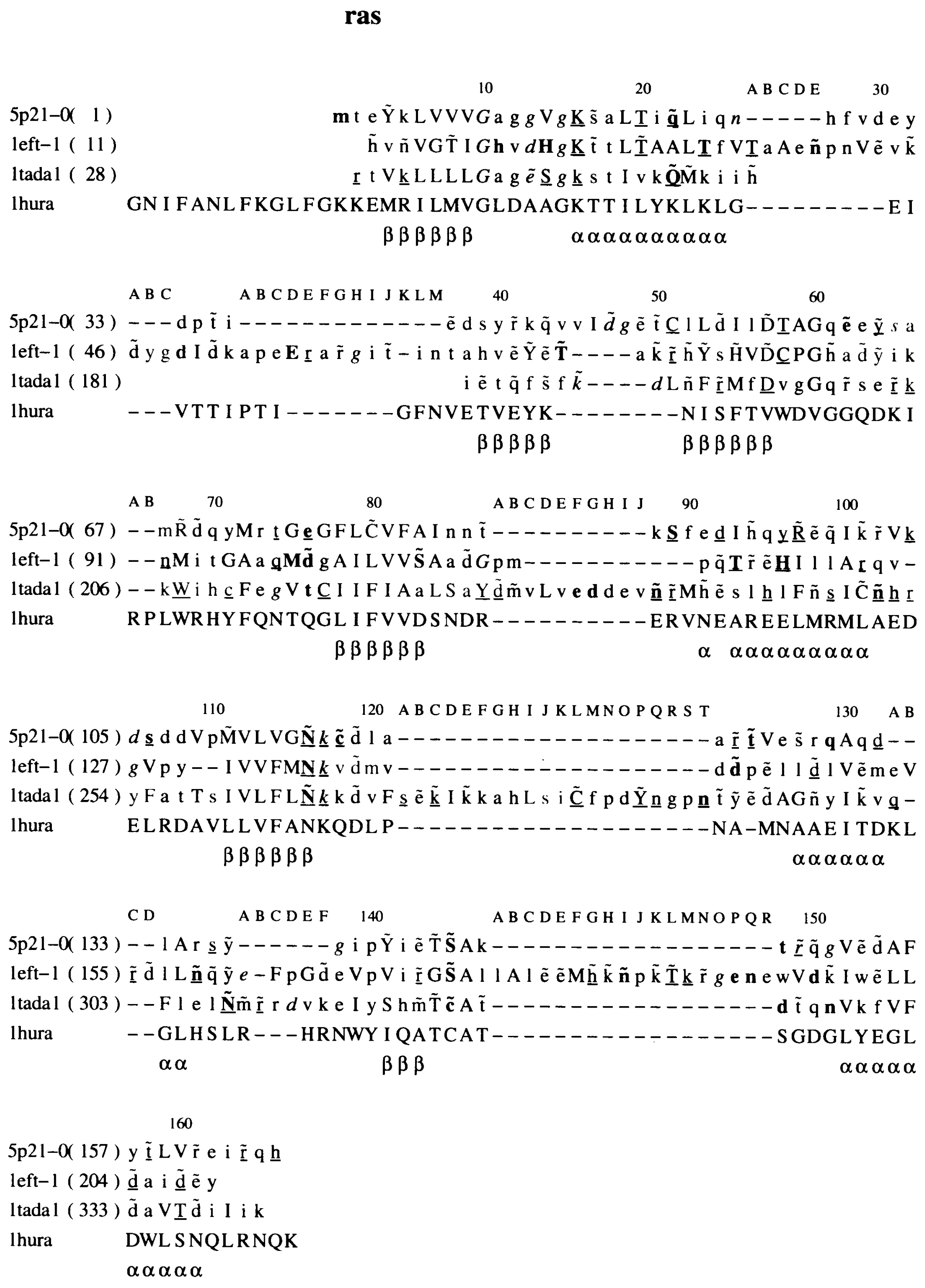
```
# family: ras
>N1;!5p21-0
>P1;5p21-0
structure:5p21-0
-------------MTEYKLVVVGAGGVGKSALTIQLIQN-----HFVDEY---DPTI-------------EDSYR
KQVVIDGETCLLDILDTAGQEEYSA--MRDQYMRTGEGFLCVFAINNT----------KSFEDIHQYREQIKRVK
DSDDVPMVLVGNKCDLA--------------------ARTVESRQAQD----LARSY------GIPYIETSAK--
----------------TRQGVEDAFYTLVREIRQH-*
>P1;1eft-1
structure:1eft-1
---------------HVNVGTIGHVDHGKTTLTAALTFVTAAENPNVEVKDYGDIDKAPEERARGIT-INTAHVE
YET----AKRHYSHVDCPGHADYIK--NMITGAAQMDGAILVVSAADGPM-----------PQTREHILLARQV-
GVPY--IVVFMNKVDMV--------------------DDPELLDLVEMEVRDLLNQYE-FPGDEVPVIRGSALLA
LEEMHKNPKTKRGENEWVDKIWELLDAIDEY-----*
>P1;1tada1
structure:1tada1
--------------RTVKLLLLGAGESGKSTIVKQMKIIH/------------------------------IETQ
FSFK----DLNFRMFDVGGQRSERK--KWIHCFEGVTCIIFIAALSAYDMVLVEDDEVNRMHESLHLFNSICNHR
YFATTSIVLFLNKKDVFSEKIKKAHLSICFPDYNGPNTYEDAGNYIKVQ---FLELNMRRDVKEIYSHMTCAT--
----------------DTQNVKFVFDAVTDIIIK--*
>P1;1hura0
sequence
GNIFANLFKGLFGKKEMRILMVGLDAAGKTTILYKLKLG---------EI---VTTIPTI-------GFNVETVE
YK--------NISFTVWDVGGQDKIRPLWRHYFQNTQGLIFVVDSNDR----------ERVNEAREELMRMLAED
ELRDAVLLVFANKQDLP--------------------NA-MNAAEITDKL--GLHSLR---HRNWYIQATCAT--
----------------SGDGLYEGLDWLSNQLRNQK*
```

Fig. 2. Sample input alignment for JOY. The alignment of the ras superfamily has been taken from the CAMPASS database and slightly modified for the purpose of illustration. The alignment is stored in a format similar to that of the NBRF/PIR format. A line beginning with the character 'C;' or '#' is a comment and ignored. In a special case where the word 'family:' follows a 'C;' or '#', the text after the 'family:' (in this example, ras) is used for the title of the output alignment. The line '>N1!5p21-0' indicates that the alignment will be labelled according to the PDB residue numbers of the structure corresponding to the code 5p21-0. The second line of the entry 1hura0 shows 'sequence', indicating that no structural information for this protein is used for annotations and that it will appear in the output alignment as a sequence only entry.

Numbering of the alignment is possible in two different ways. First, a simple alignment-based numbering scheme is available. Second, a structure-based scheme labels the alignment according to the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Abola *et al.*, 1987) residue numbers of a designated structure. Insertion codes are automatically placed within the alignment as required. In this way, it is possible to label an alignment in a family-based way (as is done in Figure 3 for the ras superfamily).

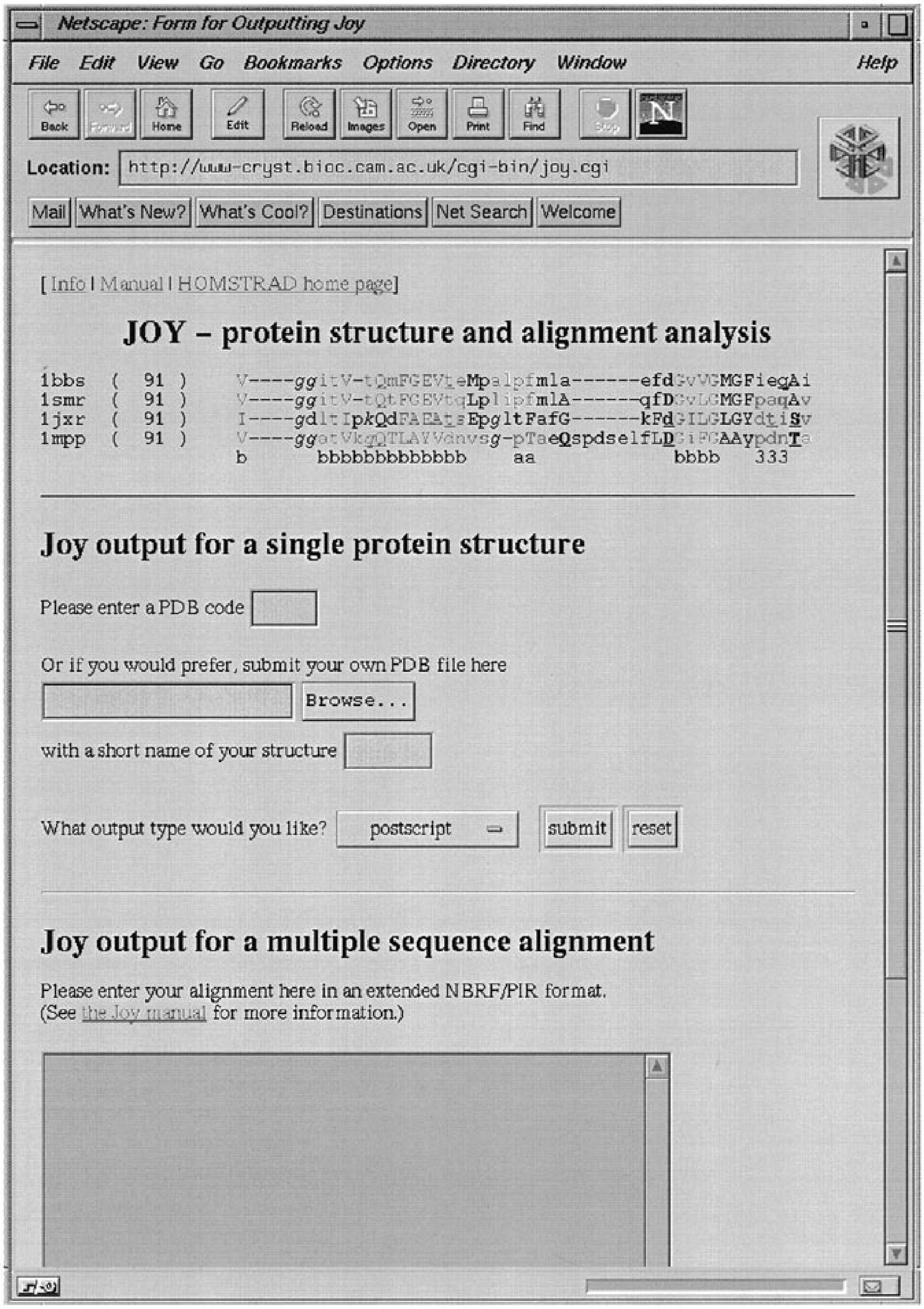
It is possible to display only a portion of the input alignment with the residues still in their proper structural environments. This option is useful when displaying aligned structural domains with each residue in its original environment within a whole protein. The segments to be displayed are read in from a separate input file.

A JOY server with a WWW interface has been developed (Figure 4). Users can submit either a PDB code or their own coordinate file. The output format can be specified by a menu and the formatted sequence can be retrieved in real time. Users can also submit a sequence alignment and retrieve the formatted alignment. In this case, only the atomic coordinates available from the PDB are used to format the user-supplied alignment.



```
                                      ras

                                      10        20     ABCDE  30
5p21-0(   1 )                mteỸkLVVVGaggVgKs̃aLTiq̃Liqn-----hfvdey
left-1 (  11 )                 h̃vñVGT̃IGhvdHgKt̃tLT̃AALTfVTaAeñpnVẽvk̃
1tada1 (  28 )                rtVkLLLLGagẽSgkstIvkQ̃M̃kiih̃
1hura           GNIFANLFKGLFGKKEMRILMVGLDAAGKTTILYKLKLG---------EI
                                ββββββ      αααααααααα

                ABC    ABCDEFGHIJKLM   40        50        60
5p21-0(  33 )   ---dpt̃i-------------ẽdsyr̃kq̃vvId̃gẽt̃ClLd̃IlD̃TAGqẽeỹsa
left-1 (  46 )  d̃ygdId̃kapeErar̃git̃-intahvẽỸẽT̃----ak̃r̃h̃ỸsH̃VD̃CPGh̃ad̃ỹik
1tada1 ( 181 )                       iẽtq̃fs̃fk̃----dLñFr̃MfDvgGqr̃ser̃k
1hura           ---VTTIPTI--------GFNVETVEYK-------NISFTVWDVGGQDKI
                                      βββββ        ββββββ

                AB    70        80     ABCDEFGHIJ  90        100
5p21-0(  67 )   --mR̃d̃qyMrtGeGFLC̃VFAInnt̃----------kSfedIh̃qyR̃ẽq̃Ik̃r̃Vk
left-1 (  91 )  --nMitGAaqMd̃gAILVVS̃Aad̃Gpm-----------pq̃Tr̃ẽHIllArqv-
1tada1 ( 206 )  --kWihcFegVtCIIFIAaLSaYd̃m̃vLveddevñr̃Mh̃ẽslhlFñsIC̃ñhr
1hura           RPLWRHYFQNTQGLIFVVDSNDR----------ERVNEAREELMRMLAED
                             ββββββ                α ααααααααα

                     110     120 ABCDEFGHIJKLMNOPQRST       130 AB
5p21-0( 105 )   dsddVpM̃VLVGÑkc̃d̃la--------------------ar̃t̃Ves̃rqAqd--
left-1 ( 127 )  gVpy--IVVFMNkvd̃mv--------------------dd̃pẽlld̃lVẽmeV
1tada1 ( 254 )  yFatTsIVLFLÑkkd̃vFsẽk̃Ik̃kahLsiC̃fpdỸngpnt̃ỹẽd̃AGñyIk̃vq-
1hura           ELRDAVLLVFANKQDLP--------------------NA-MNAAEITDKL
                      ββββββ                             αααααα

                CD     ABCDEF  140     ABCDEFGHIJKLMNOPQR 150
5p21-0( 133 )   --lArsỹ------gipỸiẽT̃S̃Ak------------------tr̃q̃gVẽd̃AF
left-1 ( 155 )  r̃d̃lLñq̃ỹe-FpGd̃eVpVir̃GS̃AllAlẽẽMh̃k̃ñpk̃T̃kr̃genewVdk̃IwẽLL
1tada1 ( 303 )  --FlelÑm̃r̃rdvkeIyShm̃T̃c̃At̃------------------dt̃qnVkfVF
1hura           --GLHSLR---HRNWYIQATCAT------------------SGDGLYEGL
                  αα            βββ                          ααααα

                  160
5p21-0( 157 )   yt̃LVr̃eir̃qh
left-1 ( 204 )  d̃aid̃ẽy
1tada1 ( 333 )  d̃aVTd̃iIik
1hura           DWLSNQLRNQK
                ααααα
```

Fig. 3. Sample output alignment of JOY. The colour PostScript output for the alignment shown in Figure 2 is displayed. The aligned proteins are: H-ras p21 (5p21-0), first domain of elongation factor Tu (1eft-1), first domain of transducin-α (1tada1) and ADP-ribosylation factor 1 (1hura). The number in parentheses after a protein code indicates the PDB residue number at the beginning of each alignment block. The top line shows alignment positions in 'ras' numbering.



Location: http://www-cryst.bioc.cam.ac.uk/cgi-bin/joy.cgi

[Info | Manual | HOMSTRAD home page]

JOY – protein structure and alignment analysis

```
1bbs  (  91 )    V----ggitV-tQmFGEVteMpalpfmla-------efdGvVGMGFiegAi
1smr  (  91 )    V----ggitV-tQtFGEVtqLplipfmlA-------qfDGvLGMGFpaqAv
1jxr  (  91 )    I----gdltIpkQdFAEAtsEpgltFafG-------kFdGILGLGYdtiSv
1mpp  (  91 )    V----ggatVkgQTLAYVdnvsg-pTaeQspdselfLDGiFGAAypdnTa
                 b       bbbbbbbbbbbbb      aa               bbbb   333
```

Joy output for a single protein structure

Please enter a PDB code

Or if you would prefer, submit your own PDB file here

with a short name of your structure

What output type would you like? postscript

Joy output for a multiple sequence alignment

Please enter your alignment here in an extended NBRF/PIR format.
(See the Joy manual for more information.)

Fig. 4. WWW interface to the JOY server. The top part shows textports and an option menu for the formatting of a single protein structu re. Below is a textport for the submission of a multiple alignment.

JOY requires a series of datafiles containing information about secondary structures, solvent accessibility and hydrogen bonding. These are produced automatically from a PDB file by supporting programs. These programs and the choice of structural features displayed are described below.

Structural features represented by JOY

Solvent accessibility

The partitioning of residues between a polar aqueous phase and a generally hydrophobic phase (the core of a globular protein) is established to be a major determinant in the process of protein folding. Residues in the solvent-inaccessible core of a protein are more conserved and are thus more useful for identifying distant evolutionary relationships. The program PSA is used to calculate the relative solvent-accessible surface area of all residues in a protein. The program uses an implementation of the algorithm of Lee and Richards (1971). Residues are defined as inaccessible by comparison to an extended conformation, and by default a 7% relative accessibility cut-off is applied (Hubbard and Blundell, 1987). The cut-off value can be set as a command line argument of JOY. Since solvent accessibility is known to be one of the most important attributes of a residue, a striking distinction is required in the typeset alignment. For this reason, we display buried residues in upper case and accessible residues in lower case.

Secondary structure and main chain conformation

Secondary structure assignments are calculated with the program SSTRUC [D.K.Smith, unpublished; now also part of the PROCHECK suite of programs (Laskowski *et al.*, 1993)]. This program calculates the secondary structural state according to the definition of Kabsch and Sander (1983). It also provides information about the main chain dihedral angles ϕ, ψ and ω. In a colour PostScript and an HTML output of JOY, repeating elements of secondary structure (α, 3_{10} and π helices, and β strands) are shown in different colours. In a black and white PostScript file, secondary structures can be displayed, by specifying a command line option, underneath each sequence. By default, the consensus secondary structure is shown underneath the output alignment. The definition of consensus is that a fraction of >0.7 is in a particular conformational state at a given position.

The ϕ and ψ angles are used to assign an ‘Efimov conformational assignment’ for a residue (Efimov, 1986). Efimov assignments can be displayed underneath each sequence. This annotation is useful for the display and identification of particular structural motifs such as β turns (Wilmot and Thornton, 1990).

Our studies into the structural constraints on residue substitution during divergent evolution identified another important feature related to main chain conformation (Overington *et al.*, 1990). The presence of a positive main chain ϕ angle places severe restriction on what residue types may be accommodated at a particular position. Residues with a positive ϕ angle are shown in italics in the formatted alignment (Figure 1). These positions (often occupied by glycine residues, the only residue that lacks a side chain and can therefore readily adopt conformations on either side of the ϕ angle) are often found as the C-terminal cap to an α helix (Richardson and Richardson, 1988), at key positions with β hairpins (Sibanda and Thornton, 1985) or within β bulges (Richardson, 1981). We also show, for structural interest, the positions of *cis* peptide bonds with a breve above the residue (not in an HTML output; see Figure 1).

Hydrogen bonds

Hydrogen bonds are calculated with the program HBOND (J.Overington, unpublished). HBOND identifies pairs of a heavy atom donor and acceptor, and calculates the associated energy and angles for each interaction. As we describe below, our main purpose for using the program is to identify side chain hydrogen bonds, but in a structure defined by X-ray crystallography side chain positions are often less accurate than the main chain positions, and it is often difficult to differentiate nitrogen and oxygen atoms in the side chain of asparagine and glutamine. Considering these ambiguities, we currently adopt a lenient definition of hydrogen bonds: a single distance cut-off of 3.5 Å. However, energy and angles printed out by HBOND can be used for refining the list of hydrogen bonds by further applying an energy-based or angular cut-off.

For the formatted alignment, the hydrogen bond data are split into four classes: (i) main chain to main chain (i.e. those responsible for secondary structure); (ii) side chain to main chain amide; (iii) side chain to main chain carbonyl; (iv) side chain to side chain/hetero atom group. Since the main chain-to-main chain hydrogen bond information is largely contained in the secondary structure representation, we do not explicitly display these data. It has been known that a buried side chain oxygen that is hydrogen bonded to a main chain amide proton plays a larger role in residue conservation than hydrogen bonds to a main chain carbonyl or to another side chain (Overington *et al.*, 1990). We therefore chose bold face to denote residues whose side chain is hydrogen bonded to a main chain amide. Residues whose side chain is hydrogen bonded to a main chain carbonyl, or to another side chain, are displayed as underlined and with a tilde, respectively (Figure 1). The deconvolution of these interactions can aid in the rapid identification of features such as helix N-terminal caps and potential salt bridges. Buried aspartic acid and threonine in the active site of aspartic proteinase exemplify the importance of hydrogen bonds to main chain groups (see the HOMSTRAD entry of aspartic proteinase). Alternatively, unusual features can be easily identified; for example, a buried but non-hydrogen bonded polar residue might suggest local errors within a model.

Half-cystine residues are considered as a separate class of side chain-to-side chain interaction. These are displayed with a cedilla (in an attempt to convey their attachment to another half-cystine residue). This information is also calculated by the HBOND program.

HBOND also calculates potential hydrogen bonds and covalent bonds between protein and ligand. This is useful for identifying ligand-binding sites and for recognizing different cofactors within a homologous family. This information, however, is not used for JOY formatting at the moment.

Implementation

JOY and the supporting programs are written in FORTRAN77 and ANSI C, and should compile and run on most UNIX platforms. Control of the program is through the use of UNIX command line arguments. These allow control of output formats, accessibility cut-off and other options. JOY assumes default filename extensions (e.g. .pdb for a PDB file, .sst for a secondary structure data file) to read in sequence and structural data. If JOY cannot find these files, it automatically runs appropriate supporting programs to generate them. The CGI program for the JOY server is written in PERL. The server runs on a Silicon Graphics Power Challenge with six 180 MHz R10000 processors.

Discussion and conclusion

Taken together, these modifications of the one-letter amino acid code allow rapid correlation of structural with sequence features and aid the visual identification of residues likely to be essential to the fold (Figures 2 and 3). Particularly visible in this representation are the buried polar residues known to be amongst the most conserved in evolution (Overington *et al*., 1990, 1992, 1993). One example is a buried glutamine conserved in a family of nerve growth factor (NGF) and related proteins. A JOY alignment of NGF, coagulogen and a *Drosophila* protein Spätzle enabled us to identify several key residues including this glutamine, which are conserved in these proteins, and this led to a hypothesis that Spätzle is also a member of this family (Mizuguchi *et al*., 1998a). This example shows that JOY is useful for analysing distantly related proteins, where the overall sequence identity is low but several key residues are conserved that play important roles in stabilizing the fold.

The JOY representation is now used in two databases of protein structure alignments. In HOMSTRAD, protein alignments are compiled at the level of homologous families, while CAMPASS presents alignments of more distantly related proteins. In both databases, JOY representation plays an essential role in characterizing each family or superfamily.

Acknowledgements

We thank our long-term collaborator Andrej Sali for many useful discussions. Pfizer Central Research, the SERC and the Imperial Cancer Research Fund are gratefully acknowledged for financial support. K.M. is a recipient of a Long Term Fellowship of the International Human Frontier Science Program Organization. T.L.B. is supported by a program grant from the Wellcome Trust.

References

Abola,E.E., Bernstein,F.C., Bryant,S.H., Koetzle,T.F. and Weng,J. (1987) Protein Data Bank. In *Crystallographic Databases—Information Content, Software Systems, Scientific Applications*, Data Commission of the International Union of Crystallography, Bonn/Cambridge/Chester.

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.

Efimov,A.V. (1986) Standard conformations of a polypeptide chain in irregular protein regions. *Mol. Biol. (Mosk)*, **20**, 250–260.

Emsley,J., White,H.E., O'Hara,B.P., Oliva,G., Srinivasan,N., Tickle,I.J., Blundell,T.L., Pepys,M.B. and Wood,S.P. (1994) Structure of pentameric human serum amyloid P component. *Nature*, **367**, 338–345.

Hoffren,A.-M., Saloheimo,M., Thomas,P., Overington,J.P., Johnson,M.S. and Blundell,T.L. (1991) Modelling the lignin peroxidase LIII of *Phlebia radiata* using a knowledge-based approach. *J. Chim. Phys*., **88**, 2659–2662.

Hubbard,T.J. and Blundell,T.L. (1987) Comparison of solvent-inaccessible cores of homologous proteins: definitions useful for protein modelling. *Protein Eng*., **1**, 159–171.

Johnson,M.S., Overington,J.P. and Blundell,T.L. (1993) Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol*., **231**, 735–752.

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

Lamport,L. (1994) *LATEX: A Document Preparation System*, 2nd edn. Addison Wesley, Reading, Mass.

Laskowski,R.A., MacArthur,M.W., Moss,D.S. and Thornton,J.M. (1993) PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr*., **26**, 283–291.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*., **14**, 379–400.

Mizuguchi,K., Parker,J.S., Blundell,T.L. and Gay,N.J. (1998a) Getting knotted: a model for the structure and activation of Spätzle. *Trends Biochem. Sci*., **23**, 239–242.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998b) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, (in press).

Overington,J.P., Johnson,M.S., Sali,A. and Blundell,T.L. (1990) Tertiary structural constraints on protein evolutionary diversity: templates, key residues and structure prediction. *Proc. R. Soc. London Ser. B Biol. Sci*., **241**, 132–145.

Overington,J.P., Donnelly,D., Johnson,M.S., Sali,A. and Blundell,T.L. (1992) Environment specific amino acid substitution tables: Tertiary templates and prediction of protein folds. *Protein Sci*., **1**, 216–226.

Overington,J.P., Zhu,Z.-Y., Sali,A., Johnson,M.S., Sowdhamini,R., Louie,G.V. and Blundell,T.L. (1993) Molecular recognition in protein families: A database of aligned three-dimensional structures of related proteins. *Biochem. Soc. Trans*., **21**, 597–604.

Parry-Smith,D.J. and Attwood,T.K. (1992) ADSP—a new package for computational sequence analysis. *Comput. Applic. Biosci*., **8**, 451–459.

Richardson,J.S. (1981) The anatomy and taxonomy of protein structure. *Adv. Protein Chem*., **34**, 167–339.

Richardson,J.S. and Richardson,D.C. (1988) Amino acid preferences for specific locations at the ends of α-helices. *Science*, **240**, 1648–1652.

Sali,A. and Blundell,T.L. (1990) The definition of topological equivalence in homologous and analogous structures: A procedure involving comparison of local properties and structural relationships through dynamic programming and simulated annealing. *J. Mol. Biol*., **212**, 403–428.

Sali,A. and Blundell,T.L. (1993) Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol*., **234**, 779–815.

Sibanda,B.L. and Thornton,J.M. (1985) Beta-hairpin families in globular proteins. *Nature*, **316**, 170–174.

Sowdhamini,R., Burke,D.F., Huang,J.-F., Mizuguchi,K., Nagarajaram,H.A, Srinivasan,N., Steward,R.E. and Blundell,T.L. CAMPASS: A database of structurally aligned protein superfamilies. Submitted.

Srinivasan,N., Bax,B., Blundell,T.L. and Parker,P.J. (1996) Structural aspects of the functional modules in human protein kinase-C alpha deduced from comparative analyses. *Proteins*, **26**, 217–235.

Sutcliffe,M.J., Haneef,I., Carney,D. and Blundell,T.L. (1987) Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. *Protein Eng*., **1**, 377–384.

Wilmot,C.M. and Thornton,J.M. (1990) β-turns and their distortions: a proposed new nomenclature. *Protein Eng*., **3**, 479–493.

Zhu,Z.Y., Sali,A. and Blundell,T.L. (1992) A variable gap penalty function and feature weights for protein 3-D structure comparisons. *Protein Eng*., **5**, 43–51.