BUTTERWORTH
HEINEMANN

VISTAS: A package for VIsualizing STructures And Sequences of proteins

D.N. Perkins* and T.K. Attwood†

**Departments of Biochemistry and Molecular Biology, The University of Leeds, Leeds, U.K.*
†University College London, London, U.K.

VISTAS is a suite of programs for protein sequence and structure analysis. The system allows the simultaneous display, in separate windows, of multiple sequence alignments, of known or model 3D structures, and of 2D graphic representations of sequence and/or alignment properties. The displays are fully integrated, and therefore manipulations in one window can be reflected in each of the others. Beyond its display facilities, VISTAS brings together a number of existing tools under a single, user-friendly umbrella: these include a fully functional interactive color alignment procedure, conserved motif selection, a range of database-scanning routines, and interactive access to the OWL composite sequence database and to the PRINTS protein fingerprint database. Exploration of the sequence database is thus straightforward, and predefined structural motifs from the fingerprint database may be readily visualized. Of particular note is the ability to calculate conservation criteria from sequence alignments and to display the information in a 3D context: this renders VISTAS a powerful tool for aiding mutagenesis studies and for facilitating refinement of molecular models.

Keywords: sequence alignment, database search, pattern recognition, structural motif, 3D-display, G protein-coupled receptors

Color plates for this article are on p. 62.

Address reprint requests to Dr. Perkins, Department of Biochemistry and Molecular Biology, The University of Leeds, Leeds LS2 9JT, U.K.
Received 11 August 1994; revised 12 October 1994; accepted 19 October 1994.

INTRODUCTION

Software packages for protein sequence analysis are now numerous (e.g., GCG,[1] STADEN,[2] ADSP,[3] and GDE)[4] Uniting a variety of analytical tools within a single system has several advantages, not least of which is that file formats are self-consistent, and a range of facilities can be explored, requiring smaller learning curves especially for novice users.

Formerly, programs offering three-dimensional (3D) structure display facilities were confined to specialized, molecular graphics packages (e.g., O[5] and QUANTA[6]), but a number of packages (e.g., Insight II,[7] GDE, and Cameleon[8]) now marry sequence analysis tools with 3D displays. Although these packages combine sequence alignment facilities with graph displays and 3D graphics in related ways, they differ in terms of database access, each interfacing preferentially with different primary sources; for example, in GCG, analysis facilities are linked to GenBank,[9] EMBL,[10] PIR,[11] and SWISS-PROT.[12] GDE is set up to search PIR, GenPept, GenUpdate and GenBank; and ADSP is linked to OWL,[13] a nonredundant composite of SWISS-PROT, a GenBank translation,[14] PIR, and NRL-3D.[15]

Searching a composite database is efficient—search times are smaller and resulting hitlists contain less noise. We have thus incorporated OWL into a new package, VISTAS, for VIsualizing STructures And Sequences. VISTAS embodies the analytical facilities of ADSP and interfaces with both OWL and its companion resource, the PRINTS database of protein motif fingerprints.[16]

MATERIALS AND METHODS

VISTAS is written in a modular form in ANSI standard C. Modules are compiled separately and finally linked to produce the executable image; this allows greater functional expandability and speeds bug fixing. Graphics and input/output facilities are provided by calls to the GL graphics library, but versions are also being implemented utilizing PEX and OpenGL.

Access is provided to the OWL database, with its query language DELPHOS and search routine SWEEP, to the PRINTS database, with its query language SMITE and search routine FINGER (see Figure 1), and to FASTA.[17] Structure files are read in PDB format.

Profiles may be generated to depict a variety of parameters, including hydropathy,[18–20] flexibility,[21] solvent accessible surface area,[22] and secondary structure prediction.[23] The variability at each position within an alignment may also be calculated.[24]

VISTAS is now accessible via the DRAL SEQNET facility, where it will shortly be superseded by an X Windows implementation. For further details of availability, please contact the authors.

PROGRAM OPERATION

Display windows

VISTAS provides two initial windows, one for the visualized structure and another for an alignment. The latter provides a restricted view of the alignment (10 or 25 sequences), but does not limit the number or length of sequences displayed (this is limited only

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(94)00013-I

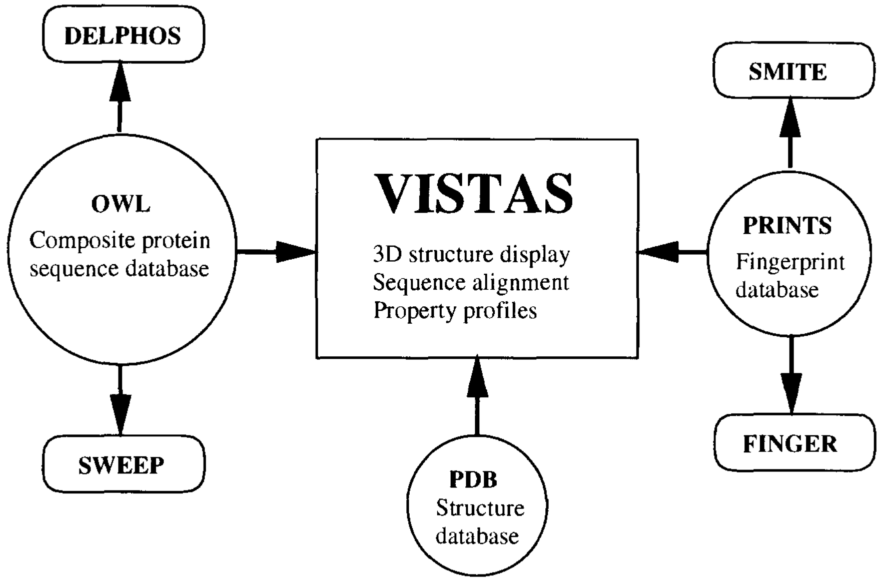

Figure 1. Relationship between the various components of VISTAS: databases are shown in circles, and search facilities in rounded boxes. In addition to interfaces to the OWL and PRINTS databases, VISTAS allows the display of 3D structures, sequence alignments, parameters calculated from the alignment (hydropathy, positional variability, etc.), and fingerprint profiles.

by system resources). Windows for depicting graphs of sequence and/or alignment properties and for plotting fingerprint profiles are optional. The displays are fully integrated, and therefore, the locations of specific conserved residues or motifs in an alignment, for example, can be visualized in a 3D context; elements of a structure can be located in an alignment; or characteristic features of, say, a hydropathy profile can be seen in both primary and tertiary environments. The program operates via two menus, which are accessed via the mouse buttons.

Graphics menu

This menu controls the display attributes, allowing the scale and orientation of the depicted structure to be changed, and offering a variety of display modes (e.g., C_α, main chain, space filling). Where appropriate, a ligand may also be depicted. The menu also allows control of the manner in which the various windows respond to each other—each may display or be colored according to separate properties, and any may be linked with any other to show the same properties.

Functions menu

The functions menu is concerned with manipulating the information depicted in the different windows, including (1) sequence manipulations, providing full facilities for interactive sequence alignment; (2) scanning procedures, providing access to both global and local similarity search methods (FASTA, SWEEP, ADSP, and FINGER); (3) direct interrogation of the OWL and PRINTS databases, providing interactive access via their query languages DELPHOS and SMITE; (4) motif selection, allowing highlighting of motifs chosen from any window; (5) motif excision, either for display and analysis, or for database searching; and (6) profile plotting, allowing selected motifs, which provide a "fingerprint" for the aligned protein family, to be scanned against any named sequence, giving an instant diagnosis of family membership.

Display colors

The default display is colored according to residue properties, using the following criteria: blue, polar positive; red, polar negative; green, polar neutral; gray, aliphatic hydrophobic; purple, aromatic hydrophobic; brown, Pro/Gly; and yellow, Cys. This scheme was chosen largely to be compatible with those used in standard physical modeling components. The displays can also be colored to depict hydropathy, positional variability calculated across the alignment, or secondary structure prediction.

The ability to distill information latent in large sequence alignments and to superpose it onto 3D models is illustrated in Color Plate 1, which depicts an analysis of G protein-coupled receptors (GPCRs). Here, a model is colored according to the positional variability calculated across an alignment of 60 sequences. With the exception of a well-conserved disulfide bridge at the extracellular end of the molecule, it is clear that the most conserved region of the structure occurs toward its cytoplasmic end,[25] and effectively outlines the region of ligand binding. Examined in more detail, the result pinpoints specific well-conserved positions in the structure and highlights the orientation of conserved helix faces. On this evidence, we can see deficiencies in the model (e.g., conserved helix faces pointing into the lipid environment) and can begin to make refinements accordingly.

DISCUSSION

Database access and the ability to construct sequence alignments are essential requirements of sequence analysts. Although structural data are still limited, where 3D structures are known it makes sense to integrate them with sequence information. Packages such as GDE, Cameleon, and VISTAS do just this, bringing together in self-consistent packages the essential equipment of the sequence analyst and providing the means for structural interpretation where such information is available.

VISTAS departs from these systems in two significant ways: first, it provides access both to a composite sequence database, making database searching efficient and convenient, and to a fingerprint database, which offers a unique perspective on pattern searching; second, it provides a framework within which users may display, analyze, and construct fingerprints for protein families in which they have a particular interest.

The ability to color alignment and structure displays according to different parameters calculated across an alignment lends VISTAS another important strength, allowing, for example, the three-dimensional locations of the most variable and conserved residues in an alignment to be visualized. This is exciting in a number of ways: it offers a rapid and informed means of selecting residues suitable for mutagenesis studies by revealing residues crucial to the structure and/or function of a protein; it provides a means of evaluating and refining existing molecular models (e.g., by pinpointing conserved faces of helices and suggesting more appropriate orientations); and it offers the ability, particularly for large superfamilies, to analyze conserved regions of individual subfamilies and to guide modeling studies accordingly. In this context, it has already provided a basis for evaluating the respective merits of GPCR models and has suggested useful refinements.

VISTAS is not a modeling package, but is a facility for assisting protein sequence analysis. It builds on an existing package,[26] providing direct database access and the ability to interpret the information latent in sequence alignments in a structural context.

ACKNOWLEDGMENTS

The work of D.N.P. has been partly supported by an MRC studentship. T.K.A. is a Royal Society University Research Fellow. This work benefits from the use of the DRAL SEQNET facility.

REFERENCES

1 Devereux, J., Haeberli, P., and Smithies, O. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 1984, **12**(1), 387–395

2 Staden, R. Methods to define and locate patterns of motifs in sequences. *CABIOS* 1988, **4,** 53–60

3 Parry-Smith, D.J., and Attwood, T.K. ADSP—a new package for computational sequence analysis. *CABIOS* 1992, **8**(5), 451–459

4 Smith, S.W., Overbeek, R., Woese, C.R., Gilbert, W., and Gillevet, P.M. The genetic data environment, an expandable GUI for multiple sequence analysis. *CABIOS* 1994 (in press)

5 Jones, T.A., Zou, J.Y., Cowan, S.W., and Kjeldgaard, M. Improved methods for building protein models in electron-density maps and the location of errors in these models. *Acta Cryst.* 1991, **47,** 110–119

6 Molecular Simulations, Inc., 200 Fifth Avenue, Waltham, MA 02154, U.S.A.

7 Biosym Technologies, Inc., 9685 Scranton Road, San Diego, CA 92121, U.S.A.

8 Oxford Molecular, Ltd., Magdalen Centre, Oxford Science Park, Oxford OX4 4GA, U.K.

9 Benson, D., Lipman, D.J., and Ostell, J. GenBank. *Nucleic Acids Res.* 1993, **21**(13), 2963–2965

10 Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J., and Cameron, G.N. The EMBL data library. *Nucleic Acids Res.* 1993, **21**(13), 2967–2971

11 Barker, W.C., George, D.G., Mewes, H.-W., Pfeiffer, F., and Tsugita, A. The PIR—international databases. *Nucleic Acids Res.* 1993, **21**(13), 3089–3092

12 Bairoch, A. and Boeckmann, B. The SWISS-PROT protein sequence data bank: Recent developments. *Nucleic Acids Res.* 1993, **21**(13), 3093–3096

13 Bleasby, A.J., Akrigg, D., and Attwood, T.K. OWL—a non-redundant composite protein sequence database. *Nucleic Acids Res.* 1994, **22**(17), 3574–3577

14 Fickett, J.W. Correct transmission of protein coding regions in GenBank. *Trends Biochem. Sci.* 1986, **11,** 190

15 Pattabiraman, N., Namboodiri, K., Lowrey, A., and Gaber, B.P. NRL-3D: A sequence-structure database. *Protein Seq. Data Anal.* 1990, **3,** 387–405

16 Attwood, T.K., Beck, M.E., Bleasby, A.J., and Parry-Smith, D.J. PRINTS—a database of protein motif fingerprints. *Nucleic Acids Res.* 1994, **22**(17), 3590–3596

17 Pearson, W.R. and Lipman, D.J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* 1988, **85,** 2444–2448

18 Kyte, J. and Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 1982, **157,** 105–132

19 Sweet, R.M. and Eisenberg, D. Correlation of sequence hydrophobicities measures similarity in 3-dimensional protein structure. *J. Mol. Biol.* 1983, **171,** 479–488

20 Engelman, D.M., Steitz, T.A., and Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 1986, **15,** 321–353

21 Ragone, R., Facchiano, F., Facchiano, A., Facchiano, A.M., and Colonna, G. Flexibility plot of proteins. *Protein Eng.* 1989, **2,** 497–504

22 Rose, G.D., Geselowitz, A.R., Lesser, G.J., Lee, R.H., and Zehfus, M.H. Hydrophobicity of amino acid residues in globular proteins. *Science* 1985, **229,** 834–838

23 Garnier, J., Osguthorpe, D.J., and Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 1978, **120,** 97–120

24 Risler, J.L., Delorme, M.O., Delacroix, H., and Henaut, A. Amino acid substitutions in structurally related proteins, a pattern recognition approach. *J. Mol. Biol.* 1988, **204,** 1019–1029

25 Attwood, T.K. and Findlay, J.B.C. Fingerprinting G-protein-coupled receptors. *Protein Eng.* 1994, **7**(2), 195–203

26 Akrigg, D., Attwood, T.K., Bleasby, A.J., Findlay, J.B.C., Maughan, N.A., North, A.C.T., Parry-Smith, D.J., Perkins, D.N., and Wootton, J.C. SERPENT—an information storage and analysis resource for protein sequences. *CABIOS* 1992, **8,** 295–296