ELECTRONIC **WORKSHOPS IN COMPUTING**

Series edited by Professor C.J. van Rijsbergen

Jessie Kennedy and Peter Barclay (Eds)

Interfaces to Databases (IDS-3)

Proceedings of the 3rd International Workshop on Interfaces to Databases, Napier University, Edinburgh, 8-10 July 1996

Paper:

Visualising Large Data Sets in Molecular Biology

J. Boyle, H. Horch and M. Scharf

Published in collaboration with the
British Computer Society

Visualising Large Data Sets in Molecular Biology

J.Boyle,
Robert Gordon University,
Aberdeen, Scotland

H.Horch and M.Scharf
Biocomputing,EMBL,
Heidelberg,Germany.

Abstract

As the wealth of information in molecular biology continues to grow there is a pressing need for powerful graphical tools to aid in the comprehension of this mass of data. This paper discusses a number of such tools that have been developed at EMBL that aim to provide data visualisation and graphical interface environments so that the molecular biologist can perform analysis of biological information more easily.

A number of scientific disciplines that were once distinct from computer science are now inextricably interwoven with it - such as molecular modeling, computational chemistry, astrophysics, and fluid dynamics. These are disciplines for which visualization is not simply a tool, but an enabling technology. [Computer Graphics World, July, 1989]

1 Introduction

Until a few years ago there was a limited number of graphical visualisation tools that were being developed for molecular biology. There has been a rise in interest in developing graphical tools for bioinformatics, including specialised workshops [6]. Typically the graphical tools that have been developed are for well defined tasks (e.g. reference browsers (Entrez), multi-sequence alignment tools (GDE), protein displayers (Rasmol), genomic data viewers (Chromoscope)[10]). With the steady growth of molecular biology data general and powerful tools are starting to be developed. These tools take their ideas from the fields of data (and information) visualisation and graphical interface design. These fields are still developing techniques to browse and examine information [7], by using a number of different techniques.

The tools that are discussed in this paper show a number of different visualisation techniques. All the systems discussed have been implemented.

- Scatter (Scharf)[8] and Plot (Boyle). These systems provide different methods to visualise n-dimensional data. These systems were developed to provide an overview of high-dimensional data, so that patterns and irregularities in the data could be seen easily. These are generic data visualisation tools.

- Aliview (Horch), ProtQuiz (Scharf) and CheckSeq (Boyle)[4].

  Aliview is a multiple sequence alignment tool which allows for customisation of portions of the visualisation to show characteristics of the data (e.g. secondary structure, sequence identity, hydrophobicity). ProtQuiz is a protein displayer which uses a linked view between the 3D mode and the sequence mode. CheckSeq is a genome viewer, which provides a general overview facility coupled with a standard sequence display. These systems illustrate the power of simple ideas such as brushing, panning, and multimodal displays. These provide specialised visualisation tools for specific tasks.

2 Bioinformatics

Bioinformatics is a field which has emerged to deal with the vast quantities of data which are becoming available for analysis. Bioinformatics studies ways to access, interact and analyse this data.

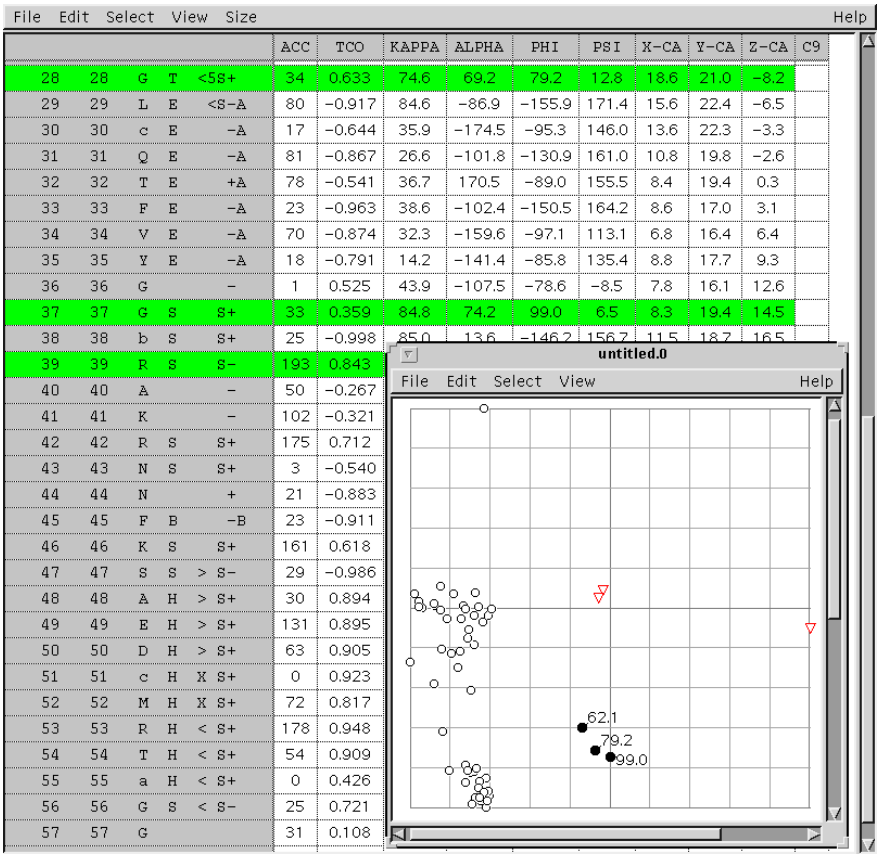

Figure 1: Showing the linked scatter viewer and table viewer. A selection in one window is automatically brushed into the other. The scatter view uses shape, colour, x-position and y-position to visualise high-dimensional data

The data comes from different laboratories and represents all facets of biological information. The systems discussed in this paper have been designed to visualise a wide range of data types:

Experimental data. This is calculated numeric information, usually about macro-molecules. It is typically highly dimensional numeric data.

Sequence data. There are two types of sequence data

Nuceotide Sequence data. This is the single largest data set in the biological sciences. It represents the actual sequence of nucleotides of the DNA in a number of different organisms (humans, mouse, hemophilius influenzea). The length of the sequences can easily be in excess of 250,000. Parts of the sequence can code for macromolecules (proteins, tRNA or rRNA), while other parts will contain control information (promoters,enhancers). There are four basic types of nucleotide.

Amino Acid sequence data. This contains information about the sequence of the amino acids in a protein. There are twenty basic amino acids which can be used to build a protein. These amino acids can be grouped into different types to show different characteristics (charge, hydrophobicity, secondary structure).

Structure data. This is information about the spatial position, type and charge of the atoms in a molecule. Typically the structure is shown as a three dimensional representation.

3 The Systems

3.1 Experimental Data

A lot of information in the biological sciences is n-dimensional, and this data can be visualised in a number of ways. A typical metaphor is a table based system (spreadsheet), which offers a familiar method for viewing the data. However it is difficult to get an overview of the data, or to see any patterns by using such an approach (and scale soon becomes a problem when there a couple of thousand items, each with n-dimensions to be visualised). There are a number of ways to visualise n-dimensional data, and as with most systems it is the task that dictates which method should be used. Most systems use a method of encoding data, so that the high-dimensional data can be mapped down to 2 or 3 dimensions.

If, for example, we wished to visualise some values for each amino acid in a protein (phi/psibond angles, different dssp scores, accessibility) then just using a table would not offer the best solution. However, if the table is coupled with a standard scatter plot viewer then this offers two different *modes* of display. Each *mode* can be used for different tasks: the table viewer can be used to examine the data item values individually and in depth, whilist the scatter viewer can be used to give an overall view of the data so that patterns and irregularities can be easily observed (see **Figure 1**).

We designed and implemented two different systems to visualise n-dimensional table data:

Scatter . This system has been coupled with the Genequiz viewer [8], so that when data items are selected in one viewer this selection is echoed in the other viewer modes (this is often referred to as *brushing*) (see **Figure 1**). Scatter visualises n-dimensional data by using a *bendiktine* approach [2]. Data is mapped on to intrinsic (colour, shading, shape) and extrinsic (x and y) positions. This allows for a visualisation of up to any 5 dimensions at one time - it is possible to change which dimensions are mapped on to which property at run time so other dimensions can be seen. For high-dimensional data this is a problem for the system.

Plot . To visualise n-dimensional data an alternative approach was needed. A simpler method of visualising high-dimension data is to use a triangular plot [1]. All dimensions of data are plotted against all other dimensions, so it is possible to see not only all the data but also how it is interrelated. When a point is selected in one graph, then this selection is echoed (brushed) in all the other graphs (see **Figure 2**).

These two systems are both designed to aid in the generic visualisation of high-dimensional data.

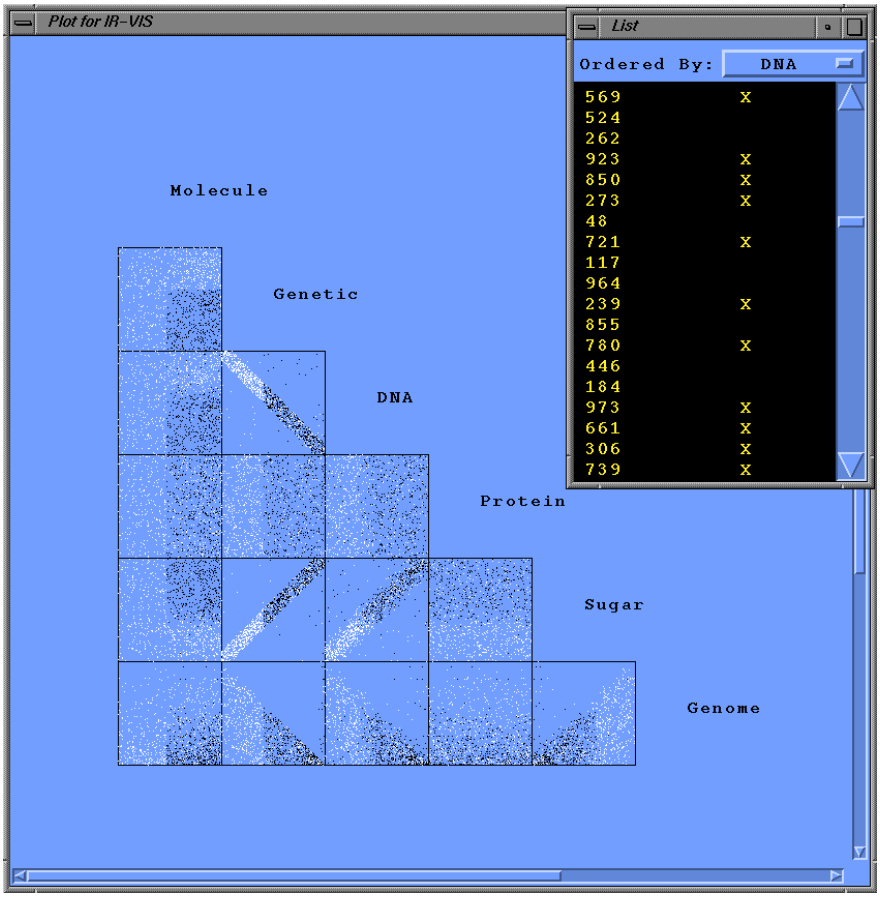

Figure 2: All N-dimensions are displayed in a series of graphs (all dimensions are plotted against all dimensions). When a selection is made this is echoed in all the other graphs. This implementation is based on a design suggested by the Imperial College Data Visualisation Group.

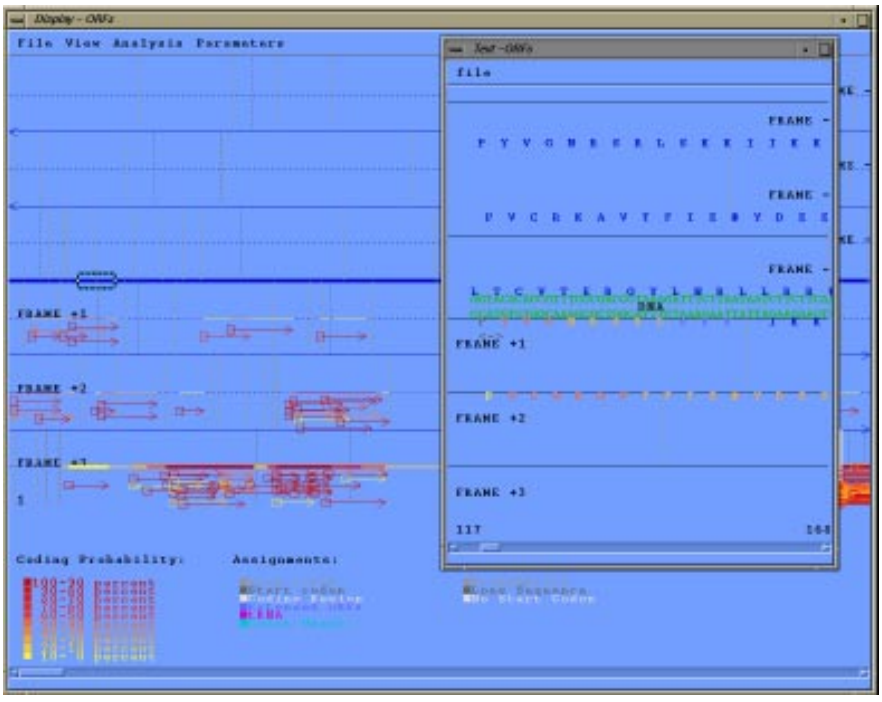

Figure 3: Checkseq is a genome viewer which shows the coding regions of all six frames of a DNA sequence (it can also be used to perform error checking within the sequence). One window provides a graphical overview of the genome, while the second window provides the actual nucleotide(and translated amino acid) sequence of the selected region.

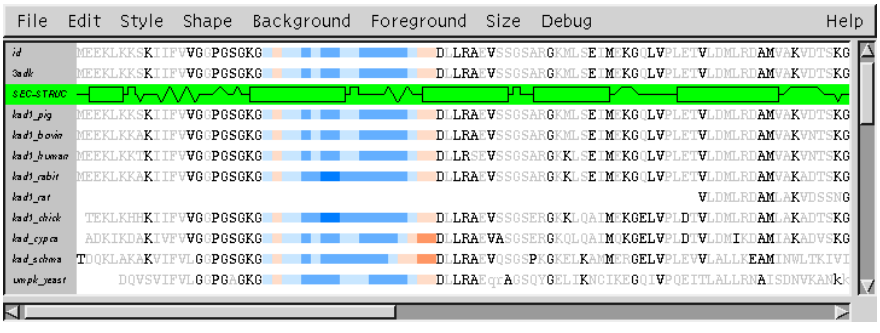

Figure 4: Aliview is a multiple sequence alignment tool. Any portion of the display can be visualised using a number of encoding techniques. In this example geometric shapes have been used to show secondary structure, colour has been used to show hydrophobicity, and intensity of the characters has been used to show sequence similiarity.

3.2 Sequence Data

Each type of sequence data (nucleotide and amino acid) can be stored as a long array of characters. However, to make use of this data, different types of visualisations have to be designed, as the useful information which can be extracted differs for each type.

3.2.1 Nucelotide Sequence Data

Nucleotide information contains information not only about the actual sequence of the genome, but also about what function parts of the sequence perform. For such data a mechanism of macro/micro views would enable one visualisation to provide an overview facility showing the high level functionality of the genome parts, while a second visualisation would show the actual sequence itself. A number of techniques have been suggested to extend and enhance the *zoom* functionality, e.g. fish eye view [5], magic lens [3]. At present *CheckSeq* uses two windows for a micro/macro view [9], so that it is possible to pan over the high level features to determine the actual sequence. A 'magic lens' implementation is being considered for the CheckSeq visualiser.

3.2.2 Amino Acid sequence

Amino acid sequences are typically compared against one another to determine similiarites between different proteins. Amino acids can be compared using a large number of criteria:

Hydrophobicity: the potential of the amino acid to attract water. Typically a protein has a large cluster of hydrophobic residues at its central core, and hydrophilic residues on its external interface (so it is soluble in water).

Secondary structure: a protein typically consists of large repeating structures (refered to as secondary structures). Proteins of similiar families have similar secondary structures.

Type: amino acids can be grouped into a number of different types (polar, non-polar, charges, acidic, basic). Amino acids of the same type have similar properties.

Mutation: the probabilities of an amino acid mutating to another amino acid have been calculated. This can be used to show how closely two proteins are related.

It is important to be able to see all these criteria, so patterns in protein family groups can be recognised. This requires a methodology of over-laying information onto the display. Multiple encoding of the data within the same visualisation allows the user to choose which visualisation best suits their needs. Aliview is a multiple-sequence alignment program which allows a user to choose how each item is to be displayed. Any portions of the display can be defined using any of the implemented techniques (see **Figure 4**).

3.3 ProtQuiz

The ProtQuiz (see **Figure 5**) graphical interface offers a protein structure viewer which uses a brushing technique so that when a user selects regions of a protein in the table view this is echoed in the 3D view (and visa versa). This allows for the integration of three dimensional graphics within a 2D framework.

4 Conclusion

A number of graphical interfaces have been introduced in this paper, each implemented for a specific task and using a variety of techniques.

Visualisation is becoming more important in bioinformatics as the amount of data which is becoming available is growing at a geometric rate. To be able to understand and comprehend this data biologists are looking to the visualisation community for ideas on how to portray this information. Techniques such as the use of multi-modal displays, macro/micro views, brushing and panning are frequently used, and are important features of an interface.

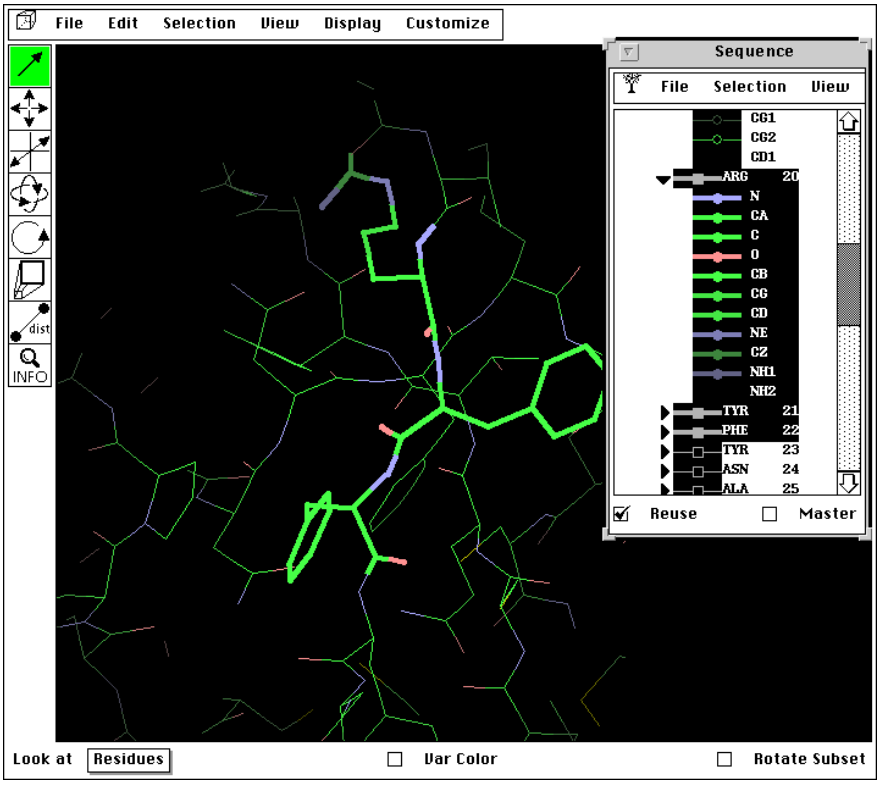

Figure 5: Protquiz is a graphical interface that demonstrates how multimodal displays can be linked, so that brushing is possible. When a user selects a region in the sequence, this selection is echoed in the protein display window.

These systems were not designed as experimental systems, but to be applications which can be used by molecular biology to aid in the analysis of biological information.

Designers of biological visualisation systems will not only have to use the most modern information visualisation techniques and paridigms, but will also have to lead research in this area of visualisation as they have an active and needful user community who require usable tools to be delivered. Conversly, as generic tools and techniques are being developed by biologists,these can be used in other fields apart from bioinformatics.

References

[1] R. Becker, P. Huber, W. Clevelane, and A. Wilks. Dynamic graphics for data analysis. *Stat Science*, 2, 1987.

[2] M. Benedikt. Cyberspace:some proposals. In M. Benedikt, editor, *Cyberspace: First Steps*, pages 273–302. MIT Press, 1993.

[3] E. Bier, M. Stone, K. Pier, W. Buxton, and T. DeRose. Toolglass and magic lenses: The see through interface. In *Proceddings of SIGGRAPH 93*, pages 73–80, 1993.

[4] J. Boyle, N. Brown, and P. Bork. Checkseq: Genome analysis, error checking and visualisation package. Submitted to CABIOS, 1996.

[5] K. Fairchild, L. Serra, N. Hern, L. Hai, and A. Leong. Dynamic fisheye information visualisations. In M. Gigante R. Earnshaw and H.Jones, editors, *Virtual Reality Systems*. London Academic Press, 1993.

[6] Graphical user interfaces in bioinformatics workshop. Available at http:// nimnet51.nimr.mrc.ac.uk/ mathbio/ t-flores/GUI-Bioinform/ meeting.html, 1994.

[7] Clayton H. Lewis. A research agenda for the nineties in human-computer interaction. *Human Computer Interaction (Hillsdale), Lawrence Erlbaum*, 5(2-3):125–143, 1990.

[8] Scharf M, Schneider R, Casari G, Bork P, Valencia A, Ouzounis C, and Sander C. Genequiz : a workbench for sequence analysis. in intelligent systems for molecular biology. In *ISMB94*, pages 348–353. Stanford CA - AAAI Press, 1994.

[9] E. Tufte. *Micro/Macro Readings*, chapter 2, pages 37–52. Graphics Press, 1990.

[10] J. Zhang, J. Ostell, and K. Rudd. Chromoscope: A graphic interactive browser for biological data expressed in the ncbi data model. In L. Hunter, editor, *27th Hawaii International Conference on System Sciences*, pages 58–67. IEEE Computer Society Press, 1994.