MatrixPlot: visualizing sequence constraints

J. Gorodkin[1,2], H. H. Stærfeldt[1], O. Lund[1*], and S. Brunak[1]

[1]Center for Biological Sequence Analysis, Department of Biotechnology,
The Technical University of Denmark, Building 208, DK-2800 Lyngby, Denmark

[2]Department of Ecology and Genetics, The Institute of Biological Sciences,
University of Aarhus, Building 540, Ny Munkegade, DK-8000 Aarhus C, Denmark

May 19, 1999

Abstract

Summary: *MatrixPlot is a program for making high quality matrix plots, such as mutual information plots of sequence alignments and distance matrices of sequences with known three-dimensional coordinates. The user can add information about the sequences (e.g., a sequence logo profile) along the edges of the plot, as well as zoom in on any region in the plot.*
Availability: *MatrixPlot can be obtained on request, and can also be accessed online at*
http://www.cbs.dtu.dk/services/MatrixPlot/
Contact: *gorodkin@cbs.dtu.dk*

Introduction

The sequence logo (Schneider and Stephens, 1990) is a very useful tool for analyzing sequence alignments. However, the use of information content or relative entropy in conventional logos assumes that positions in the alignments are independent. Clearly, this is not the case for a structural alignment of RNA sequences, and therefore the sequence logo was extended to include mutual information between known basepair positions in the structural alignment (Gorodkin *et al.*, 1997). For the matrix plot program presented here, the user can combine the mutual information between all positions in the alignment, with a sequence information profile plotted along the edges of the matrix plot.

Matrix plots are also useful when analyzing single sequences. In proteins, for example, contact maps hold binary values indicating contact between the respective C^α atoms (using an arbitrary distance threshold). These plots are sometimes replaced by distance matrices, which indicate the physical distance between all C^α atoms in the protein. Such plots are used for RNA sequences, for instance when displaying the energy of a predicted RNA secondary structure. Along the edges of such a matrix plot, the user may include color encoded bars indicating various features, for example the protein secondary structural elements.

*Present address: Structural Bioinformatics Advanced Technologies A/S, Agern Alle 3, DK-2970 Hørsholm, Denmark

Implementation and web interface

MatrixPlot is written in `nawk` and runs on Unix platforms. The program accepts a matrix plot (`mp`) format, which contains a listing of values for any set of 2D coordinates. Bars and logo profiles are encoded by one number and a corresponding value. The values are represented by colors, except in the case of logo profiles for which they indicate the information content. The user can zoom in on a region in the plot. This is done by saving the relevant coordinates from the original matrix file, into a new matrix file, and use the matrix plot program again. The color scale in the matrix plot can be defined by the user.

MatrixPlot is accompanied by two other programs: i) *Inform* that computes the mutual information content of an alignment, and ii) *pdb2mp* which converts the Brookhaven Data Bank (PDB) (Bernstein *et al.*, 1977) format to the mp format used by the main program.

Inform can take the sequence alignment in fasta format, a simple "one sequence per line" (`align`) format, and the msf format. The sequence alphabet is not restricted to DNA, RNA or proteins, but can be user defined, for example, the protein secondary structure assignments. For RNA alignments, the mutual information can also be computed as in Gorodkin *et al.* (1997). That measure is a lower bound for the standard form of mutual information, but it is also more robust in the presence of noise caused by a small sample size (see the web page for a brief discussion).

Figure 1(a) shows an example of a mutual information plot of an RNA alignment extracted from the signal recognition database (Samuelsson and Zwieb, 1999). The alignment contains only 12 sequences, and still the principal covarying regions appear on the plot, in particular the main stem that has many basepairs between positions 75–150 and 250–325.

Distance matrices of available PDB entries can be made by the program pdb2mp. However, the user can also provide a sequence with 3D coordinates either in PDB format, or in a simple column format as described on the web page. For the online version, distance matrices of PDB entries, can be generated solely by specifying entry and chain. The user may include up to two bars on

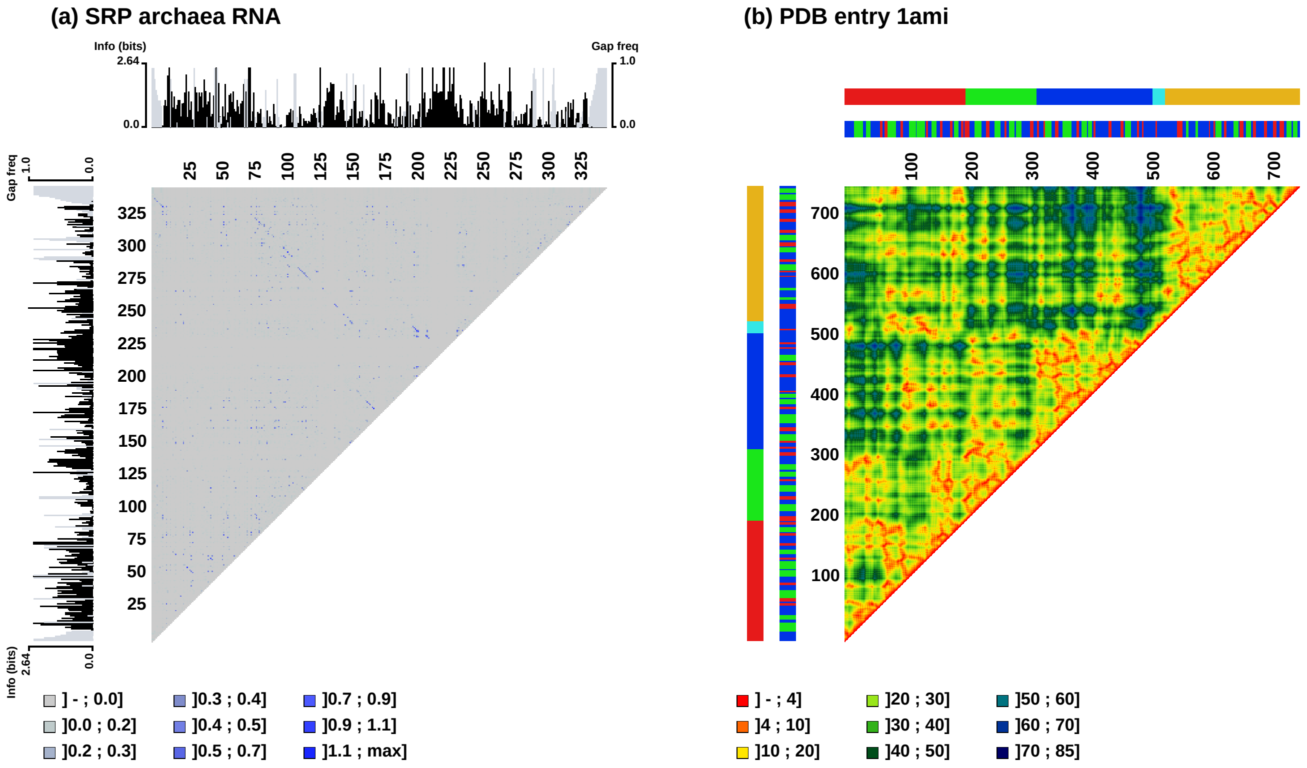

Fig. 1. Typical applications of matrix plots: (a) The mutual information content of an RNA alignment of SRP Archaea sequences. A combined plot of sequence information and gap frequencies is displayed along the edges. The mutual information is indicated by the color scale. The sequence information on the logo profiles is indicated in black, the gap frequencies in gray. (b) distance matrix of PDB entry 1ami. The outer bar indicates the four structural domains and a linker region. The inner bar indicates the secondary structure, with sheet in red, helix in green, and all other assignments in blue. The color scale of the matrix indicates the physical distances in Angstrom.

the edges of the matrix. The bars can also be replaced by text strings. Minimum distances for contact are indicated through the user defined color scale. Figure 1(b) displays the distance matrix of the four domain protein Aconitase (PDB entry `1ami`). The four domains clearly appear as compact regions along the diagonal, visible as a staircase motif on the plot.

MatrixPlot contains options to scale and move objects on the matrix plot. The program also has an option "colorgrain" for which the color plotted, is found by linear interpolation. The matrix plots can either be downloaded as .ps or .pdf files. The .mp file containing the values of the matrix plot can be downloaded as well.

A user's guide describes the features in greater detail; it will be updated as new features enters MatrixPlot. Future improvements include specification of the colors in the bars and real logos on the edges rather than profiles.

Acknowledgements

This work was supported by the Danish National Research Foundation.

References

Bernstein,F.C., Koetzle,T.G., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rogers,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.* **122**, 535–542. (http://www.pdb.bnl.gov/).

Gorodkin,J., Heyer,L.J., Brunak,S. and Stormo,G.D. (1997) Displaying the information contents of structural RNA alignments: the structure logos. *CABIOS,* **13**, 583–586. (http://www.cbs.dtu.dk/gorodkin/appl/slogo.html).

Samuelsson,T. and Zwieb,C. (1999) The signal recognition particle database (SRPDB). *Nucl. Acids Res.* **27**, 169–170. (http://psyche.uthct.edu/dbs/SRPDB/).

Schneider,T.D. and Stephens,R.M. (1990) Sequence logos: a new way to display concensus sequences. *Nucl. Acids Res.* **18**, 6097–6100.