



Visual representation of database search results: the RHIMS Plot

David M.A. Martin^{1,*}, Pamela Hill², Geoffrey J. Barton¹ and Andrew J. Flavell²

¹Post-Genomics and Molecular Interactions Centre, Wellcome Trust Biocentre, University of Dundee, Dundee DD1 5EH, UK and ²Plant Research Unit, University of Dundee at SCRI, Invergowrie, Dundee, DD2 5DA, Scotland, UK

Received on September 13, 2002; revised on December 21, 2002; accepted on January 14, 2003

ABSTRACT

Summary: An algorithm and software are described that provide a fast method to produce a novel, function-oriented visualization of the results of a sequence database search. Text mining of sequence annotations allows position specific plots of potential functional similarity to be compared in a simple compact representation.

Availability: The application can be accessed via a web server at <http://www.compbio.dundee.ac.uk>. The RHIMS software may be obtained by request to the authors.

Contact: d.m.a.martin@dundee.ac.uk

Sequence database searching is an integral part of modern molecular biology. A typical search with a single query sequence, may result in hits to many hundreds of sequences in the database, each with one or more high scoring pairwise alignments (HSPs). The interpretation of this voluminous output has been simplified by representing the output of BLAST (Altschul *et al.*, 1997) both as a multiple alignment of 'stacked bars' by the application Blixem (Sonhammer and Durbin, 1994) and the NCBI BLAST server (<http://www.ncbi.nih.gov/BLAST>) which extends this idea by adding an indication of the score for the represented HSP, and as a multiple sequence alignment in Blixem and Mview (Brown *et al.*, 1998). AV (Chi *et al.*, 1995) extends the stacked bars representation in a third dimension according to the score of the HSP.

This improved representation of output may still be voluminous and difficult to interpret. Accordingly, this application note describes a technique for generating a function-oriented view of a database search in the form of a composite histogram view of all HSPs. Each HSP is weighted according to the strength of the match against the database. The display can be refined by selecting HSP subsets from the search results to obtain a more informative visualization. In the example presented here the new technique has been used to identify and rapidly

distinguish between potential *Ty1-copia*, *Ty3-gypsy* and LINE group retrotransposons in a random set of sequences isolated from three *Vicia* bean species.

The database search is parsed into a list of high scoring pairwise matches. A similarity index (R) is calculated for each HSP by rescoring the probability (either p or E value) value as $R = \max\{-\log P, 0\}$. This effectively ignores alignments with probabilities greater than 1.

The total similarity index (I) at a given position (x) in the query sequence is calculated by summing the scores R for all pairwise alignments that overlap position x . I_x is then plotted versus x to represent the degree of similarity of position x to the database as a whole. This basic plot describes the strength of the overall match to the database along the length of the sequence. We call this measure the Relative Homology Index using Mathematical Summation (RHIMS).

The query sequence may contain multiple elements, each of which is separately similar to a number of database sequences from different sequence families. Selecting subsets of the database hits by, for example, keyword matching to the description field enables each alignment subset to be individually plotted and a direct visual comparison of the relative strength of similarity to be made.

In order to illustrate the technique, we have applied it to TBLASTX (Altschul *et al.*, 1997) searches made with 689 random plant sequences obtained from three *Vicia* bean species against the plant division of the EMBL nucleotide database (Stoesser *et al.*, 2002). Sequences with potentially interesting search results were rapidly identified by visual inspection of RHIMS plots corresponding to subsets of HSPs matching relevant keywords. Further examination at the sequence level was necessary to confirm a potential positive result from this first pass visual screening process.

Figure 1 illustrates an RHIMS plot of a TBLASTX search where the query sequence is a random *Vicia sativa* clone, mined by keyword. This analysis has allowed the sequence to be identified provisionally as a retrotranspo-

*To whom correspondence should be addressed.

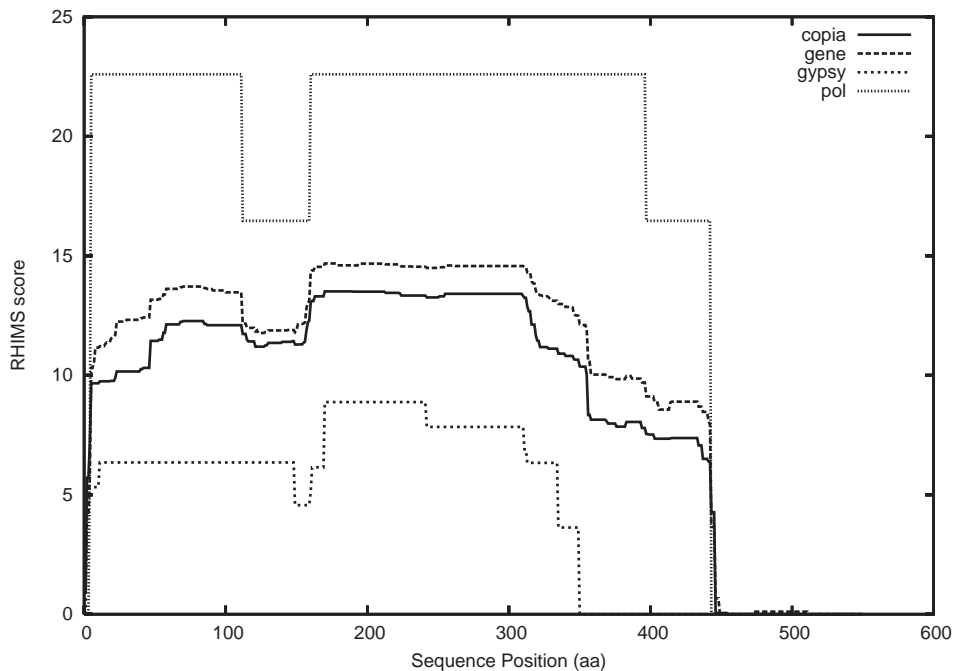


Fig. 1. The figure shows an RHIMS plot of a TBLASTX search against the plant division of the EMBL database. The query sequence is a random genomic clone from *Vicia sativa* and RHIMS scores for subsets of hits whose descriptions match certain keywords are shown. The example sequence shows considerable similarity to the *pol* region of retrotransposons and subsequent examination revealed that it contains well conserved motifs between base 175 and 425 (data not shown).

son with strong similarity to the *Ty1-copia* group but no significant similarity to *Ty3-gypsy* or *LINE* type groups of retrotransposable elements. (For a review of plant retrotransposable elements see Kumar and Bennetzen, 1999).

A Java applet for plotting and querying the data has been developed and incorporated into a web-based service. A BLAST database search, either user supplied or via an integrated search of the non-redundant protein database (nr) at NCBI, is parsed as input. Subsets of pairwise alignments can be selected by a combination of sequence position and text matching to the description field for the subject sequence.

The general visualization method presented here allows the rapid 'first pass' qualitative screening of database search reports which contain sequence alignment information and an appropriate scoring metric, so that the results may be prioritized for more detailed examination. Although powerful, the method is dependent on the quality and extent of the functional annotation in the sequence database and susceptible to potential overrepresentation of specific protein families. Subsequent visual inspection of the HSP lists obtained allows the keyword selection to be optimized and over-representation to be assessed manually.

ACKNOWLEDGEMENTS

We thank Professor Mike Ferguson for his encouragement and support. DMAM is supported by a JIF grant from The Wellcome Trust (grant no. 060269).

REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Brown,N.P., Leroy,C. and Sander,C. (1998) MView: a Web compatible database search or multiple alignment viewer. *Bioinformatics*, **14**, 380–381.
- Chi,E.H.-H., Barry,P., Shoop,E., Carlis,J.V., Retzel,E. and Riedl,J. (1995) Visualization of biological sequence similarity search results. In *Proceedings of the IEEE Conference on Visualization*. pp. 44–51.
- Kumar,A. and Bennetzen,J.L. (1999) Plant retrotransposons. *Annu. Rev. Genet.*, **33**, 479–532.
- Sonhammer,E.L. and Durbin,R. (1994) A workbench for large scale sequence homology analysis. *Comput. Appl. Biosci.*, **10**, 301–307.
- Stoesser,G., Baker,W., van den Broek,A., Camon,E., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V. et al. (2002) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **30**, 21–26.