

CrossLink: visualization and exploration of sequence relationships between (micro) RNAs

Tobias Dezulian^{1,*}, Martin Schaefer¹, Roland Wiese², Detlef Weigel³ and Daniel H. Huson¹

¹Department of Algorithms in Bioinformatics, Center for Bioinformatics Tübingen, Tübingen University, D-72076 Tübingen, Germany, ²yWorks GmbH, Research and Development, D-72070 Tübingen, Germany and ³Department of Molecular Biology, Max-Planck-Institute for Developmental Biology, D-72076 Tübingen, Germany

Received February 11, 2006; Revised March 25, 2006; Accepted March 27, 2006

ABSTRACT

CrossLink is a versatile tool for the exploration of relationships between RNA sequences. After a parametrization phase, CrossLink delegates the determination of sequence relationships to established tools (BLAST, Vmatch and RNAhybrid) and then constructs a network. Each node in this network represents a sequence and each link represents a match or a set of matches. Match attributes are reflected by graphical attributes of the links and corresponding alignments are displayed on a mouse-click. The distributions of match attributes such as *E*-value, match length and proportion of identical nucleotides are displayed as histograms. Sequence sets can be highlighted and visibility of designated matches can be suppressed by real-time adjustable thresholds for attribute combinations. Powerful network layout operations (such as spring-embedding algorithms) and navigation capabilities complete the exploration features of this tool. CrossLink can be especially useful in a microRNA context since Vmatch and RNAhybrid are suitable tools for determining the antisense and hybridization relationships, which are decisive for the interaction between microRNAs and their targets. CrossLink is available both online and as a standalone version at <http://www-ab.informatik.uni-tuebingen.de/software>.

INTRODUCTION

Explicitly visualizing sequences and their relationships as a network provides concise and intuitive exploration possibilities. In this respect, CrossLink nicely complements the software CLANS (1) which uses a network to visualize sequence similarity between amino acid sequences. CrossLink delegates the determination of sequence relationships to the established

tools BLAST (2), Vmatch (3) and RNAhybrid (4). Users versed with these tools will appreciate that (almost) all tool-specific parameters may be set from within CrossLink. Furthermore, CrossLink allows relationships determined by distinct tools to be visualized within the same network. Both BLAST and Vmatch can detect local sequence similarity in both sense and antisense directions and are suitable for a wide range of scenarios. BLAST is a standard tool using a fast seed-and-extend strategy. Vmatch employs a suffix array-based approach that permits constraints on the match length and on the number of mismatched bases within a match. RNAhybrid is a specialized tool that can predict potential binding sites of microRNAs in large target RNAs using an extension of the classical RNA secondary structure prediction algorithm (5). In general, RNAhybrid finds the energetically most favorable hybridization sites of a small RNA in a large RNA.

Although CrossLink can be put to use in many scenarios amenable to the above tools, it can be especially useful in a microRNA context: microRNAs interact with target transcripts by complementary base-pairing and can be classified into families on the basis of sequence similarity—relationships that can be detected by using Vmatch/RNAhybrid and BLAST, respectively (cf. examples below).

DESIGN

Balancing flexibility and complexity, CrossLink allows the user to independently specify three different kinds of relationship searches, each with its own strategy (BLAST, Vmatch and RNAhybrid) and a set of parameters. To this end, CrossLink's input consists of two sets of RNA, A and B, each provided in the FASTA format. The first kind of similarity search, S_{AA} , is performed between all sequences of set A, yielding the set of matches M_{AA} . Likewise, similarity searches S_{AB} and S_{BB} are performed to yield the set of matches M_{AB} (between all sequences of set A and all sequences of set B) and the set of matches M_{BB} (between all sequences of set B),

*To whom correspondence should be addressed. Tel: +49 7071 2970454; Fax: +49 7071 295148; Email: dezulian@informatik.uni-tuebingen.de

respectively. For clarity, a color scheme is associated with each kind of match: reddish colors frame the parameter input controls for S_{AA} as well as the match representations of M_{AA} in the network, corresponding alignment windows and histograms. Similarly, S_{AB} and M_{AB} are associated with greenish colors and S_{BB} and M_{BB} are associated with bluish colors (Figure 1). Within each color scheme, shades indicate the orientation of each match: a dark shade is associated with matches in sense orientation and a light shade is associated with matches in antisense orientation.

In addition to orientation, each match has the following attributes: E -value, length and the proportion of identical nucleotides within the alignment when the match was determined by using BLAST or Vmatch; minimal free energy (MFE), length and the proportion of paired nucleotides within the alignment when the match was determined by using RNAhybrid. For each match set, a visualization option panel (Figure 2) is provided that uses a histogram for each match attribute to display the corresponding value distribution. Sense matches and antisense matches are tallied separately in each histogram. Note that the E -value and MFE attribute histograms run on a logarithmic scale and the length and identity/paired proportion histograms run on a linear scale. Serving a 2-fold purpose, the visualization option panel also allows manipulation of the network: a threshold may be set for each attribute and a specified combination of thresholds then determines which matches will be considered for analysis and represented as links in the network and which will be suppressed. This feature allows the user to rapidly focus on matches with interesting characteristics. A threshold is set by adjusting a slider for each attribute and selecting a combination mode. Two

combination modes are available: in conjunction mode (logical 'AND') only matches that pass all thresholds will be displayed. In disjunction mode (logical 'OR') only matches that pass at least one of the thresholds will be displayed. Whether the threshold acts as a cutoff for smaller or higher values of an attribute is specified by a radio button setting located on the left and right of each attribute histogram. In addition, all sense and/or antisense matches may be suppressed for a given match set. Exploration can further be focused on an arbitrary selection of sequences by removing all remaining sequences (along with their relationships) from the exploration session using the menu bar (►View►Remove all unselected nodes). All histograms are accordingly recalculated on the basis of the remaining relationships.

An exploration session involves three phases that occur in order: first, during a parametrization phase, the two input files are chosen and for each of the three relationship searches a strategy (BLAST, Vmatch or RNAhybrid) is selected and corresponding parameters are specified. Next, in the search phase, CrossLink uploads all necessary information to the server and the search is performed remotely. Upon completion the results are passed back. During the final exploration phase the resulting network is visualized and relationships can be explored. A reset button permits the user to jump back to the parametrization phase with the current parameters.

Any two sequences can give rise to several distinct local sequence similarities. Representing each match by its own link may clutter up the network visualization when many sequence pairs each yield a multitude of local matches. Therefore, each match set can independently be displayed in either 'single

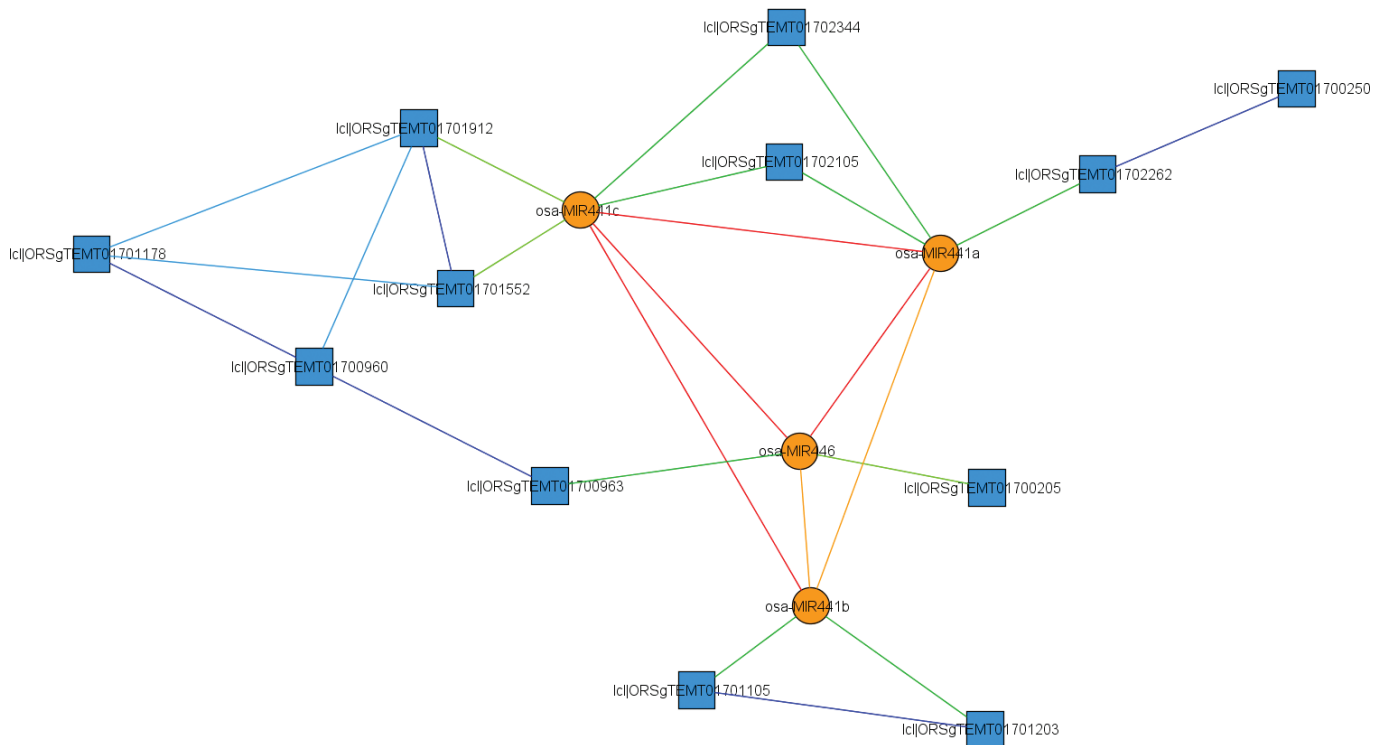


Figure 1. Network of sequence sets A (red nodes) and B (blue nodes) with corresponding matches of set M_{AA} , M_{AB} and M_{BB} represented by links in reddish, greenish and bluish colors, respectively.

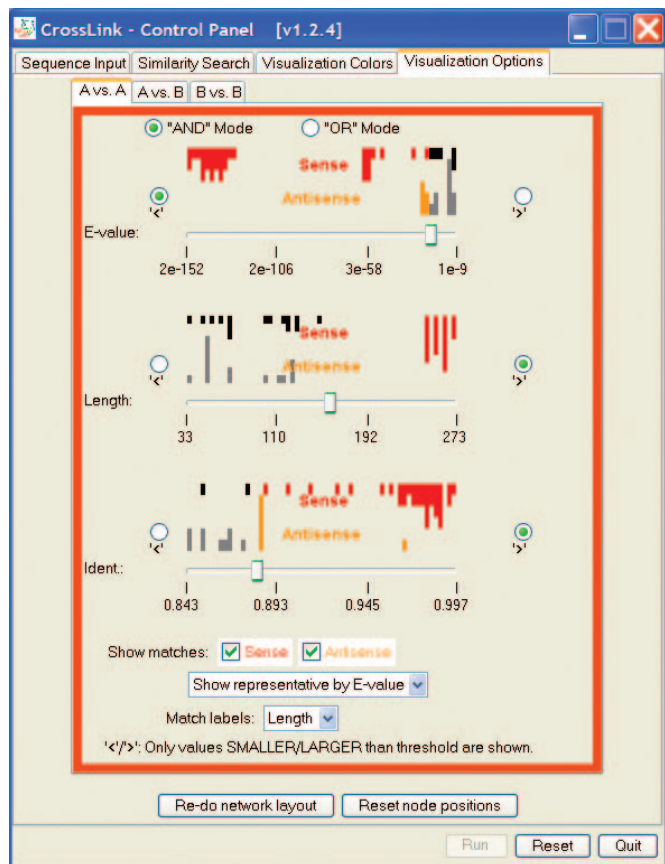


Figure 2. The visualization options panel for matches in set M_{AA} displaying the histograms associated with each match attribute.

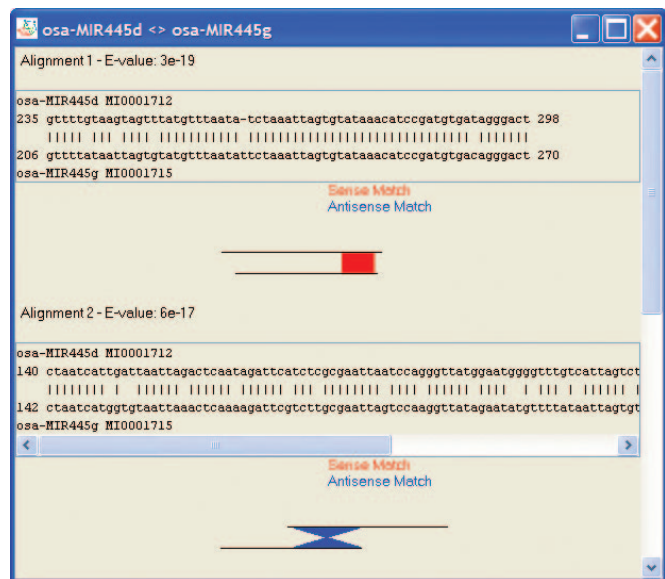


Figure 3. Alignment window showing two separate matches between one pair of sequences.

match representative mode’ or ‘multiple match representative mode’. In ‘single match representative mode’ each link between two network nodes represents a single match between the corresponding sequences. In the case of several matches

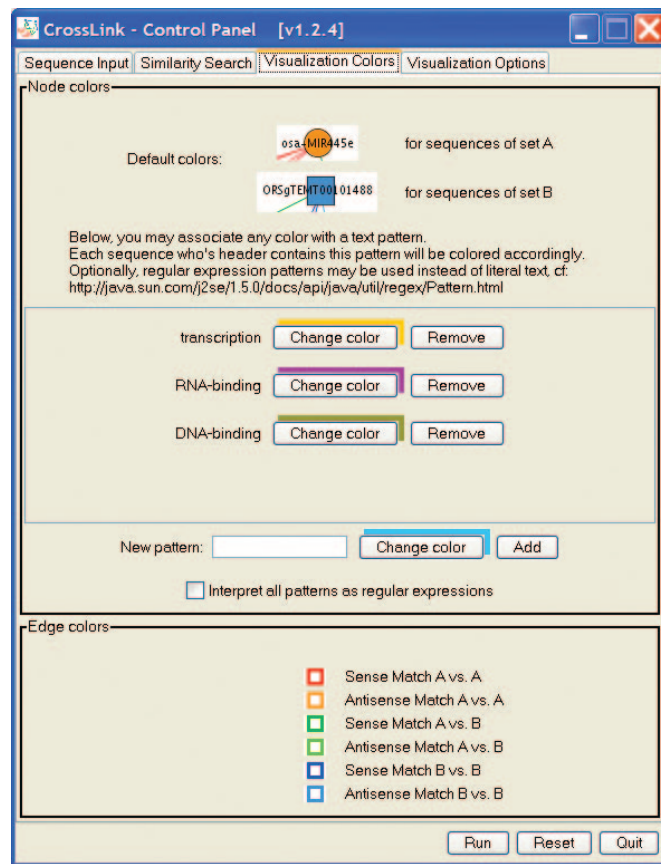


Figure 4. The visualization color panel, showing custom pattern–color associations in the center.

between this pair of sequences each is represented by its own link running side by side between the two nodes. In ‘multiple match representative mode’ a link between a pair of sequences represents all corresponding matches. One can select whether the representative of this match set should be the one with the smallest *E*-value/MFE, greatest length or highest identity/paired proportion—as this may be relevant for the mentioned attribute histograms.

Clicking on a node or link of the network spawns a separate window displaying detail information about the corresponding sequence or match(es) (Figure 3), respectively. Note that the alignment is displayed in text form exactly as output by the originating tool. Clicking on a subset of selected nodes spawns a separate window displaying the corresponding sequences in the FASTA format. This enables export of sequence subsets for further scrutiny using other tools.

By default, sequences of set A and set B are displayed as red and blue nodes, respectively, in the network. Arbitrary colors may be assigned to subsets of sequences using the following strategy: a color can be associated with a text pattern. Each sequence, which contains the text pattern literally as a substring in its FASTA header, will be colored accordingly. Optionally, the pattern may contain a regular expression that is matched accordingly. Any number of such pattern–color associations may be specified (Figure 4). A sequence thus associated with several colors will appear multicolored.

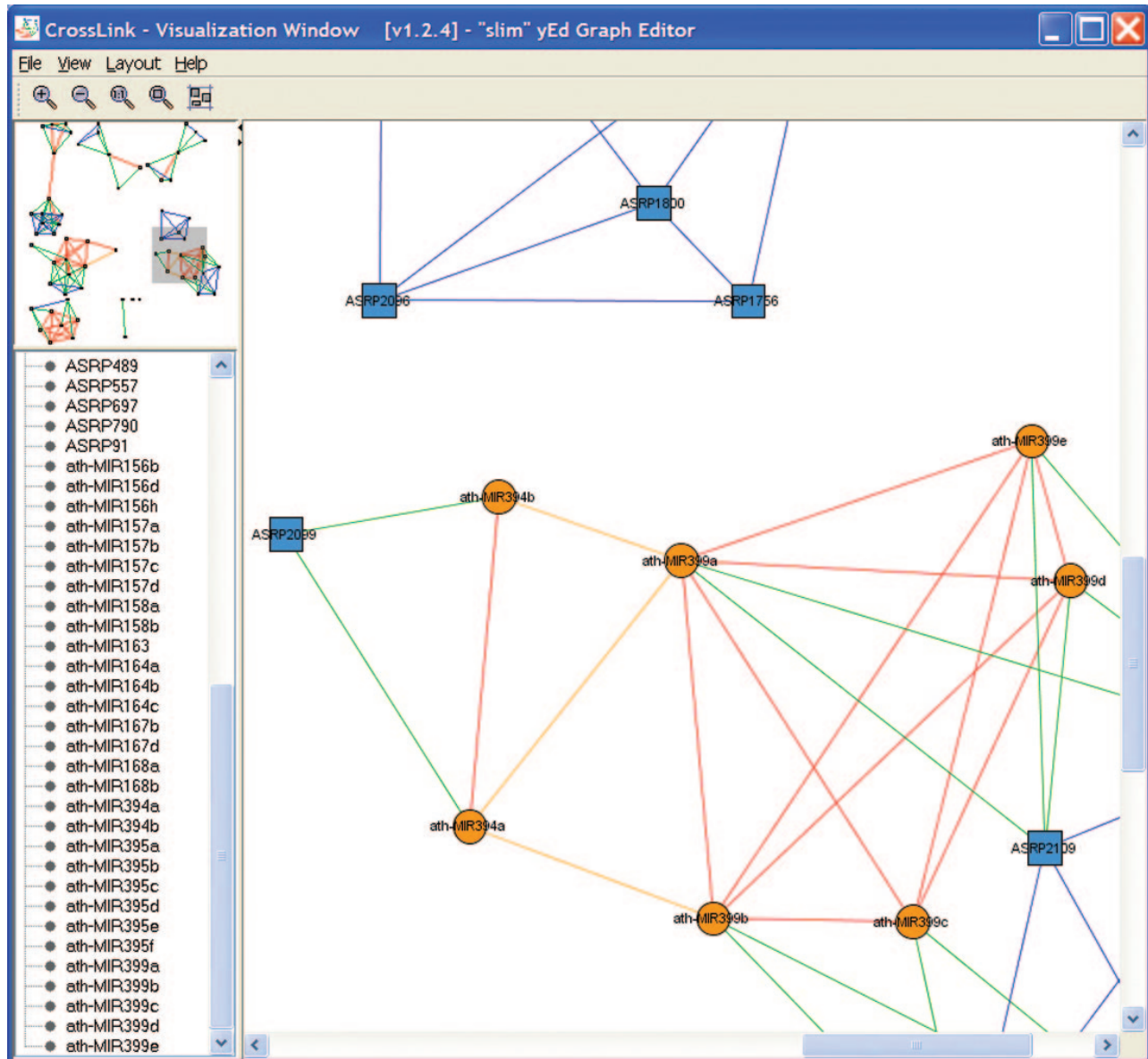


Figure 5. The visualization window, with an overview area on the top left and a sequence selection panel on the lower left.

To facilitate repeated exploration runs, the current parameter set can be named and saved as a configuration template. Any subsequent exploration task can be based on such a configuration template either ‘as is’ or after modification. Each configuration template contains the following parameters: each of the three sequence similarity search strategies including all parameters, the custom pattern–color associations and the two sequence input file names (as associated *default* file names). Note that, for consistency, selecting a different configuration template does not change the currently stated input file names. However, the default file names associated with the current template can be chosen explicitly.

A visualization window offers fast and powerful navigation of the network shown in the main view area (Figure 5): an overview area displays the currently visible clipping as a gray

rectangle, which can be dragged, focussing the main view area accordingly. The mouse wheel permits rapid zooming. Network nodes can be selected and moved. Double-clicking on a sequence in the sequence selection pane (Figure 5, lower left) centers the view onto this sequence. Dragging the mouse cursor over a sequence displays its FASTA header.

Several algorithms are available for network layout. The default layout algorithm is a Fruchterman–Reingold (6) spring-embedding, similar to the one used in the BioLayout (7) library, where each link acts as a spring pulling at the sequences it is attached to. A ‘Reset node positions’ Button undoes all node movements performed since the last application of a layout algorithm. CrossLink’s visualization component is based on the yFiles (8) graph library which provides the spring-embedding implementation.

EXAMPLES

CrossLink provides three example configuration templates along with the corresponding sequence files. To try out CrossLink, one merely has to select one of the examples and press the 'Run' button. The following example scenarios are provided:

- *Example 1:* Sequence set A consists of all rice microRNAs of families 440–446 available from miRBase (9). Sequence set B contains a subset of repetitive rice sequences downloaded from the TIGR Rice Genome Annotation Database. It is immediately visible that, for example, the rice microRNA family 445 exhibits very close sequence similarity to a family of repetitive rice sequences. Initially displaying a multitude of links in a tangle, this example demonstrates the power of the interactive histograms to focus on relevant relationships.
- *Example 2:* Sequence set A consists of all *Arabidopsis* microRNA precursors available at miRBase. Sequence set B contains all (~2000) sequences contained in the *Arabidopsis* Small RNA Project Database (10) to date. Setting these two sets in relationship with each other allows one to assess which microRNA families have been sequenced by the ASRP project. This example also demonstrates CrossLink's ability to handle large sets of sequences and also shows the power of the spring-embedding algorithm in clustering microRNAs into families.
- *Example 3:* Sequence set A consists of the *Drosophila* microRNAs dme-miR-3, dme-miR-4 and dme-miR-5. Sequence set B contains all corresponding targets which have been predicted (with an *E*-value < 1) in a study by Rehmsmeier *et al.* (4), plus some randomly picked sequences from the same study that have not been predicted as potential targets of these microRNAs. This example demonstrates the use of RNAhybrid, for example, revealing that one sequence (accession no. CG15125) is simultaneously targeted by two different microRNAs. Furthermore, the capability of custom pattern–color associations is shown as each predicted target set of the Rehmsmeier *et al.* (4) study is associated with its own color (yellow, magenta and cyan for the targets of dme-miR-3, dme-miR-4 and dme-miR-5, respectively) and the non-targets are shown in blue.

AVAILABILITY

CrossLink is available both online and as a downloadable local version. Both versions require an installed Java Runtime Environment (JRE1.4.2 or later). To prevent overload of our server, the online version restricts the size of the two

input files to 1 MB. The local version requires locally installed NCBI BLAST, Vmatch and RNAhybrid tools and a TCSH command line. The CrossLink website at <http://www-ab.informatik.uni-tuebingen.de/software> provides a user manual including a quick start guide plus detailed descriptions of the example input data that CrossLink supplies.

ACKNOWLEDGEMENTS

We thank Norman Warthmann, Rebecca Schwab, Heike Wollmann and Matthias Zschunke for helpful suggestions. Furthermore, we are grateful to Stefan Kurtz for his Vmatch software (www.vmatch.de) and to Marc Rehmsmeier for supplying us with the *Drosophila* sequences. We especially appreciate that yWorks (www.yworks.com) provided us with yFiles library components. We would like to thank all users who helped to improve our software with their questions and feedback as well as two anonymous reviewers for their helpful comments. Funding to pay the Open Access publication charges for this article was provided by the Deutsche Forschungsgemeinschaft (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Frickey, T. and Lupas, A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
2. 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.
3. Kurtz, S., Choudhuri, J.V., Ohlebusch, E., Schlieiermacher, C., Stoye, J. and Giegerich, R. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. *Nucleic Acids Res.*, **29**, 4633–4642.
4. Rehmsmeier, M., Steffen, P., Hochmann, M. and Giegerich, R. (2004) Fast and effective prediction of microRNA/target duplexes. *RNA*, **10**, 1507–1517.
5. Zuker, M. and Stiegler, P. (1981) Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, **9**, 133–148.
6. Fruchterman, T.M. and Reingold, E.M. (1991) Force directed placement. *Softw. Pract. Exp.*, **21**, 1129–1164.
7. Enright, A.J. and Ouzounis, C.A. (2001) BioLayout—an automatic graph layout algorithm for similarity visualization. *Bioinformatics*, **17**, 853–854.
8. Wiese, R., Eiglsperger, M. and Kaufmann, M. (2001) yFiles: visualization and automatic layout of graphs. In Mutzel, P., Jünger, M. and Leipert, S. (eds), *Proceedings of the 9th International Symposium on Graph Drawing*. Springer, Berlin, pp. 453–454.
9. Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
10. Gustafson, A.M., Allen, E., Givan, S., Smith, D., Carrington, J.C. and Kasschau, K.D. (2005) ASRP: the *Arabidopsis* small RNA project database. *Nucleic Acids Res.*, **33**, D637–D640.