BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 22 2010, pages 2900–2901
doi:10.1093/bioinformatics/btq545

Genome analysis

Advance Access publication September 24, 2010

SylArray: a web server for automated detection of miRNA effects from expression data

Nenad Bartonicek and Anton J. Enright*

EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Summary: A useful step for understanding the function of microRNAs (miRNA) or siRNAs is the detection of their effects on genome-wide expression profiles. Typically, approaches look for enrichment of words in the 3′UTR sequences of the most deregulated genes. A number of tools are available for this purpose, but they require either in-depth computational knowledge, filtered 3′UTR sequences for the genome of interest, or a set of genes acquired through an arbitrary expression cutoff. To this end, we have developed *SylArray*; a web-based resource designed for the analysis of large-scale expression datasets. It simply requires the user to submit a sorted list of genes from an expression experiment. SylArray utilizes curated sets of 3′UTRs to attach sequences to these genes and then applies the Sylamer algorithm for detection of miRNA or siRNA signatures in those sequences. An intuitive system for visualization and interpretation of the small RNA signatures is included.

Availability: SylArray is written in Perl-CGI, Perl and Java and also uses the R statistical package. The source-code, database and web resource are freely available under GNU Public License (GPL). The web server is freely accessible at http://www.ebi.ac.uk/enright/sylarray.

Contact: aje@ebi.ac.uk

Received on May 25, 2010; revised on August 26, 2010; accepted on September 21, 2010

1 INTRODUCTION

Non-coding RNAs can have a profound influence on global mRNA expression. Small RNAs such as small interfering RNAs and microRNAs bind to 3′UTRs of mRNAs as a part of the RNA-induced silencing complex (Tang, 2005; Yazgan and Krebs, 2007). This binding event stimulates cleavage or degradation of the target molecule (Lim *et al*., 2005). Efficiency of silencing is determined by various factors, the most important one being complementarity to the target mRNA 3′UTR in a conserved region of 6–8 nt, also called the 'seed' (Grimson *et al*., 2007). Analysis of the overrepresentation of particular seeds in 3′UTRs across an experimental gene set can be used to detect the influence of specific miRNAs and siRNAs on an expression profile (Farh *et al*., 2005).

Recently, the Sylamer algorithm was devised to assess significantly over or under-represented words in 3′UTR sequences according to a sorted gene list (van Dongen *et al*., 2008). The original algorithm required the user to provide both an ordered gene list of sufficient size and a matching set of 3′UTR sequences. In order to make this approach accessible to non-expert users we have developed SylArray.

SylArray combines a curated database of 3′UTRs with a user-friendly interface to visualize and detect the influence of miRNAs or siRNAs on gene expression profiles. Several other tools have been published with similar aims, such as DIANA-mirExTra (Alexiou *et al*., 2010) and GeneSet2miRNA (Antonov *et al*., 2009). SylArray allows for analysis of larger number of species, containing a premade database of both filtered 3′UTRs and coding sequences for seven model organisms. We will continue to update and maintain the SylArray database with new arrays and platforms as they become available.

2 IMPLEMENTATION

The starting point for analysis is a sorted gene list obtained by the user from a gene-expression experiment. The goal of SylArray is to take this list through the full protocol of miRNA and siRNA seed enrichment analysis using a simple and intuitive interface.

2.1 Algorithm

SylArray provides a database and a user interface to the Sylamer algorithm (van Dongen *et al*., 2008). It measures significance of word enrichment in 3′UTRs of ranked gene sets based on hypergeometric *P*-values.

2.2 Database

The SylArray MySQL database stores curated 3′UTR sequences and microarray platform information for a large variety of sources. This was created through the R software environment and the Bioconductor package *biomaRt*. Currently, 3′UTR sequences are available for: *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Xenopus tropicalis*, *Danio rerio*, *Drosophila melanogaster* and *Caenorhabditis elegans.* Sequences were obtained from ENSEMBL for all available BioMart gene sets of the following id types: *affymetrix*, *agilent*, *illumina*, *symbol, refseq*, *unigene*, *entrez gene* and *ensembl*. The full list of gene sets can be found as a supplementary material on the website. The longest UTR was selected for each of the Ensembl gene IDs and the sequences were cleaned of low complexity regions using the DUST algorithm with default parameters (Morgulis *et al*., 2006). The gene sets were then individually purged of repetitive and redundant sequences by using the RSAT purge-sequence interface to Vmatch (Thomas-Chollier *et al*., 2008). Purging has been previously shown to be essential for reliable statistical sequence analysis (Defrance *et al*., 2008).

*To whom correspondence should be addressed.

 © The Author 2010. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016

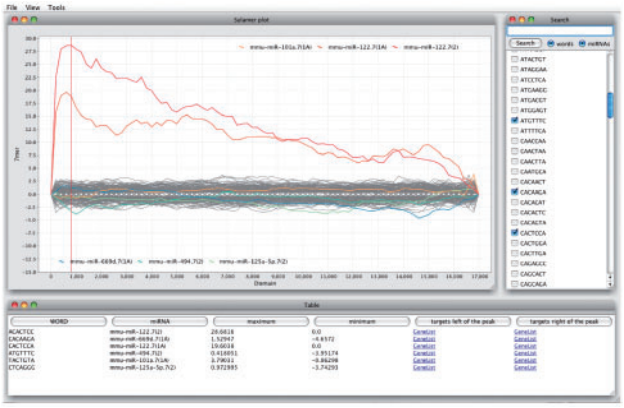

Fig. 1. Screen shot of SylArray with a sample analysis. Chart panel (top left): landscape plot for a mmu-miR-122 antagonism experiment (Elmen *et al*., 2008). Colored lines represent log of enrichment/depletion *P*-value for miRNA seeds in 3′UTRs, calculated in incremental parts of the submitted ranked gene list. Genes are sorted by change in expression, from most to least. Statistically significant depletion is observable in the first ∼800 upregulated genes (red vertical line from the peak selection tool). Search panel (top right): allows queries for individual words or known miRNA seeds. Table panel (bottom): provides information for selected words or miRNAs and allows access to the lists of genes on either side of the selected peak.

2.3 Web server

The web interface allows the user to upload an ordered gene list obtained from a microarray or other genome-wide experiment. Ideally, this list is sorted according to a measure of differential expression (e.g. log fold-change or *t*-statistic) and also contains a sufficient number of genes for statistical power. For each gene provided, curated 3′UTR sequences are retrieved from the SylArray database and submitted with the ranked list to Sylamer for analysis. User can also submit his own premade set of curated FASTA sequences to the web server for the further analysis. Once analysis is complete the results are passed to the Java-based Analysis Toolkit.

2.4 Analysis toolkit

The Analysis Toolkit is written in Java (1.5+ compatible) using the Eclipse framework via Web Start. It runs on all Java-supporting operating systems (Mac OS X, MS Windows, UNIX and Linux). The toolkit enables inspection and export of plots for all standard seed lengths (6mers, 7mers and 8mers) simultaneously (Fig. 1). Individual miRNAs can be highlighted and their behavior cross-referenced across different seed lengths. Furthermore, 'peak selection' tool enables access to individual genes that have contributed to the Sylamer miRNA signal. All the biological entities are linked to the corresponding databases and easily exported.

3 CONCLUSION

We have created a web server for the analysis, visualization and interpretation of miRNA and siRNA influence on gene expression profiles. This tool is both freely available and simple to use. This resource should allow researchers from a broad area of expertise to perform fast and accurate detection of small RNA signatures in their gene-expression datasets.

ACKNOWLEDGEMENTS

The authors would like to thank Cei Abreu-Goodger, Stijn van Dongen, Harpreet Saini, Sergei Manakov, Jose Assuncao, Matias Piipari, Leopold Parts and Matthew P. Davis for help and support during the web server development and manuscript preparation.

Conflict of Interest: none declared.

REFERENCES

Alexiou,P. *et al*. (2010) The DIANA-mirExTra web server: from gene expression data to microRNA function. *PLoS One*, **5**, e9171.

Antonov,A.V. *et al*. (2009) GeneSet2miRNA: finding the signature of cooperative miRNA activities in the gene lists. *Nucleic Acids Res.*, **37**, W323–W328.

Defrance,M. *et al*. (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat. Protocols*, **3**, 1589–1603.

Elmen,J. *et al*. (2008) LNA-mediated microRNA silencing in non-human primates. *Nature*, **452**, 896–899.

Farh,K.K. *et al*. (2005) The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science*, **310**, 1817–1821.

Grimson,A. *et al*. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.

Lim,L.P. *et al*. (2005) Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, **433**, 769–773.

Morgulis,A. *et al*. (2006) WindowMasker: window-based masker for sequenced genomes. *Bioinformatics*, **22**, 134–141.

Tang,G. (2005) siRNA and miRNA: an insight into RISCs. *Trends Biochem. Sci.,* **30,** 106–114.

Thomas-Chollier,M. *et al*. (2008) RSAT: regulatory sequence analysis tools. *Nucleic Acids Res.*, **36**, W119–W127.

van Dongen,S. *et al*. (2008) Detecting microRNA binding and siRNA off-target effects from expression data. *Nat. Methods*, **5**, 1023–1025.

Yazgan,O. and Krebs,J.E. (2007) Noncoding but nonexpendable: transcriptional regulation by large noncoding RNA in eukaryotes. *Biochem. Cell Biol.*, **85**, 484–489.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 31, 2016