BioSystems 99 (2010) 79–81

Contents lists available at ScienceDirect

BioSystems

journal homepage: www.elsevier.com/locate/biosystems

ModuleMaster: A new tool to decipher transcriptional regulatory networks

Clemens Wrzodek[a,*], Adrian Schröder[a], Andreas Dräger[a], Dierk Wanke[b], Kenneth W. Berendzen[b], Marcel Kronfeld[a], Klaus Harter[b], Andreas Zell[a]

[a] *Center for Bioinformatics Tübingen (ZBIT), University of Tübingen, Sand 1, 72076 Tübingen, Germany*
[b] *Center for Plant Physiology Tübingen (ZMBP), University of Tübingen, Auf der Morgenstelle 1, 72076 Tübingen, Germany*

ARTICLE INFO

Article history:
Received 8 May 2009
Received in revised form 19 August 2009
Accepted 30 September 2009

Keywords:
Gene regulation
cis-Regulatory modules
Regulatory sequence analysis
Matrix scan
Transcription factors

ABSTRACT

In this article we present ModuleMaster, a novel application for finding *cis*-regulatory modules (CRMs) in sets of co-expressed genes. The application comes with a newly developed method which not only considers transcription factor binding information but also multivariate functional relationships between regulators and target genes to improve the detection of CRMs. Given only the results of a microarray and a subsequent clustering experiment, the program includes all necessary data and algorithms to perform every step to find CRMs. This workbench possesses an easy-to-use graphical user interface, together with job-processing and command-line options, making ModuleMaster a sophisticated program for large-scale batch processing. The detected CRMs can be visualized and evaluated in various ways, i.e., generating GraphML- and R-based whole regulatory network visualizations or generating SBML files for subsequent analytical processing and dynamic modeling. *Availability*: ModuleMaster is freely available to academics as a webstart application and for download at http://www.ra.cs.uni-tuebingen.de/software/ModuleMaster/, including comprehensive documentation.

1. Introduction

Deciphering gene regulatory relationships from microarray experiments is one of the key disciplines in bioinformatics. It is generally accepted that genes, coding for proteins which are involved in the same step of a metabolic pathway, are usually co-regulated (Segal et al., 2003). These genes mostly share common regulatory elements in their promoter sequences—so-called *cis*-regulatory modules (CRMs). CRMs are functional combinations of transcription factor binding sites that act in close proximity in regulatory sequences and regulate the expression of multiple genes (Segal et al., 2003). Many methods for clustering genes and conditions based on microarray data are available (Supper et al., 2007; Reiss and Baliga, 2006). These methods overcome the disadvantages of earlier clustering approaches by performing a biclustering, which yields overlapping clusters of genes that are co-expressed under given subsets of conditions. Here, a condition can be a different microarray experiment, another timepoint or any other expression value.

Based on any clustering results, we developed a novel workbench, which includes all algorithms and organism specific binding site information that is required for a successful detection of CRMs in sets of co-expressed genes. Searching for modules can be done by using the ModuleSearcher algorithm (Aerts et al., 2000) that only takes into account transcription factor binding sites (TFBSs), or the newly developed ModuleMaster algorithm (Schröder et al., submitted for publication). This new algorithm considers not only TFBSs but also linear relationships between genes and transcription factors based on mRNA expression values using a multi-objective genetic algorithm. The ModuleMaster algorithm has been validated on various datasets and improves the quality of the resulting CRMs. For example, a transcription factor that does not correlate well with the genes in a cluster is unlikely to be a transcription factor for this gene cluster under the given conditions. Whereas a transcription factor whose transcription level shows a high correlation to all genes in a cluster is likely to be an activating regulator for this cluster.

Currently, there are only few workbenches for *cis*-regulatory analysis freely available. TOUCAN 2 is one of the most popular open source approaches (Aerts et al., 2005). TOUCAN 2 uses the ModuleSearcher algorithm to scan for modules and does not consider experimental data. Furthermore, it performs all jobs on remote web services and therefore, no automated large-scale analysis of clustering data is possible with TOUCAN 2. Another example is

* Corresponding author.
E-mail addresses: clemens.wrzodek@uni-tuebingen.de (C. Wrzodek), adrian.schroeder@uni-tuebingen.de (A. Schröder), andreas.draeger@uni-tuebingen.de (A. Dräger), dierk.wanke@zmbp.uni-tuebingen.de (D. Wanke), kenneth.berendzen@zmbp.uni-tuebingen.de (K.W. Berendzen), marcel.kronfeld@uni-tuebingen.de (M. Kronfeld), klaus.harter@zmbp.uni-tuebingen.de (K. Harter), andreas.zell@uni-tuebingen.de (A. Zell).

0303-2647/$ – see front matter © 2009 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.biosystems.2009.09.005

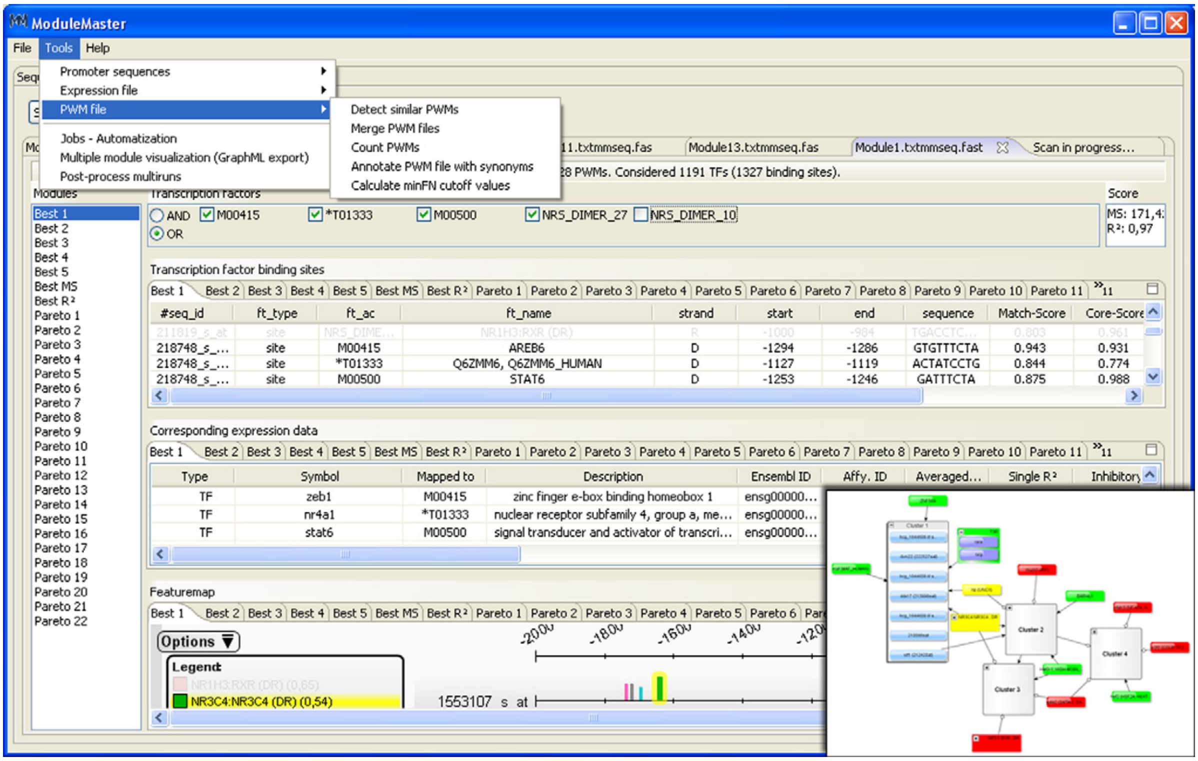
ModuleMiner, which is only available as a web interface (Loo et al., 2008). ModuleMiner enhances the ModuleSearcher by removing all configuration options and automatically estimating an optimal number of transcription factors for a module in a given set of sequences. Just like ModuleSearcher this approach does only consider TFBSs and no experimental data. A few approaches have been proposed earlier that try to include experimental data to search for motifs (Bussemaker et al., 2001; Roven and Bussemaker, 2003; Das et al., 2004). Most of these approaches do not specifically search for TFBSs but for motifs of a fixed length and none of these approaches correlates the expression of putative regulatory transcription factors to the expression of their target genes. In addition, these methods only search for motifs and not for modules. For these reasons, we developed ModuleMaster to provide a workbench for automated and effective large-scale analysis and included a new algorithm that considers not only TFBS, but also experimental data for an improved detection of CRMs.

2. Key features

ModuleMaster is an advanced program for finding *cis*-regulatory modules in sets of co-expressed genes. It is capable of retrieving sequences, performing matrix scans on these sequences and finally, searching for *cis*-regulatory modules. It offers a huge list of options and features to customize it for various objectives. The whole process from clustering results to the final modules for all clusters can be fully automated. The program offers an easy-to-use graphical user interface, parallel job-processing and various command-line capabilities. ModuleMaster is able to analyze regulatory sequences of a large list of organisms, in contrast to previous approaches. Sequence retrieval can be done using the RSAT (Thomas-Chollier et al., 2008) or the EnsEMBL (Hubbard et al., 2009) database. For human sequences, a tissue filter based on the TiGER (Liu et al., 2008) database is implemented that automatically selects the correct gene if a gene name is ambiguous. Promoter sequences can be used to perform *cis*-regulatory analysis. The MATCH TM matrix scan algorithm (Kel et al., 2003) has been implemented and included into ModuleMaster. This algorithm searches for potential TFBSs in promoter sequences. All TFBSs, detected by the MATCH TM algorithm, will be reported along with various statistic measurements and visualizations. By including predicted binding sites according to Supper et al. (submitted for publication), next to experimentally validated binding sites from TRANSFAC® public (Matys et al., 2009) and PWMs generated from IUPAC sequences, e.g., from PLACE (Higo et al., 1999) or NUBIScan (Podvinec et al., 2002), ModuleMaster has a very comprehensive library of position weight matrices (PWMs). Hence, there is no need to search for unspecific motifs without designated targets in the promoter sequences of genes, which is still a disadvantage of other approaches. This is a key component that allows ModuleMaster to correlate the expression of transcription factors to the expression of other genes.

Individual cutoff scores to minimize the false negatives and organism dependent 4th order hidden Markov models are provided for all included PWMs and for PWMs from TRANSFAC® professional. If the user has access to a TRANSFAC® professional license, these PWMs can be integrated easily. For other PWMs, ModuleMaster provides automated methods to calculate cutoff scores and background models. In most cases, it is sufficient to select an organism and the collection of PWMs to be used. ModuleMaster automatically extracts and precalculates all necessary data and performs the matrix scan.

Based on binding sites and any microarray gene expression data, ModuleMaster can detect common *cis*-regulatory modules in sets of co-expressed genes. The program uses a multi-objective genetic algorithm from EvA2 (Kronfeld, 2008) to detect not only significant combinations of binding sites, but also plausible relationships



28 PWMs. Considered 1191 TFs (1327 binding sites).

○ AND ☑ M00415 ☑ *T01333 ☑ M00500 ☑ NRS_DIMER_27 □ NRS_DIMER_10
● OR

MS: 171,4:
R²: 0,97

#seq_id	ft_type	ft_ac	ft_name	strand	start	end	sequence	Match-Score	Core-Score
211819_s_at	site	NRS_DIME...	NR1H3:RXR (DR)	R	-1000	-984	TGACCTC...	0.803	0.961
218748_s_...	site	M00415	AREB6	D	-1294	-1286	GTGTTTCTA	0.943	0.931
218748_s_...	site	*T01333	Q6ZMM6, Q6ZMM6_HUMAN	D	-1127	-1119	ACTATCCTG	0.844	0.774
218748_s_...	site	M00500	STAT6	D	-1253	-1246	GATTTCTA	0.875	0.988

Type	Symbol	Mapped to	Description	Ensembl ID	Affy. ID	Averaged...	Single R²	Inhibitory
TF	zeb1	M00415	zinc finger e-box binding homeobox 1	ensg00000...				
TF	nr4a1	*T01333	nuclear receptor subfamily 4, group a, me...	ensg00000...				
TF	stat6	M00500	signal transducer and activator of transcri...	ensg00000...				

Legend:
NR1H3:RXR (DR) (0,65)
🟩 NR3C4:NR3C4 (DR) (0,54)

Fig. 1. Screenshot of ModuleMaster. The module-view is selected and statistics for the currently selected module are displayed. This includes, i.e., all TFBSs in the current module, all expression information and a visualization of the actual module (lower third of the picture). On the top, several tools that are available for processing of PWM files are shown. The lower right corner shows interactions between various clusters and transcription factors. This image section was automatically generated by ModuleMaster and has been visualized using yEd (yWorks, 2007).

between transcription factors and genes in the current cluster. The linear relationship is quantified with a multivariate linear regression. As a result, a list of significant modules is generated along with information regarding both score objectives and the actual module, and opportunities to further evaluate and validate each module. ModuleMaster reports all Pareto-optimal solutions, along with solutions that show significant scores in both objectives (significance of TFBSs and correlation of expression values). Besides these, modules that maximize only one objective will be reported, too. Every module is visualized (see Fig. 1 for an example), a list of all included TFBSs is given and all informations about the linear regression are reported. Based on this regression, ModuleMaster can successfully predict which transcription factor acts as a repressor and which acts as an activator (Schröder et al., submitted for publication). Besides this, ModuleMaster provides many tools to further process promoter sequences, PWM files and microarray expression files. For example: detecting similar PWMs, annotating expression probes with synonyms, retrieving promoters for all genes in an expression file, generate random datasets, retrieve sequences from a container file, and many more. ModuleMaster provides further support whenever additional information on a module is needed or a module should be validated. It provides tools which can draw relations between transcription factors and clusters, visualizes the expression correlation of genes and transcription factors, or visualizes the actual module. It is possible to create job files and let ModuleMaster automatically search for modules in a complete microarray dataset. To validate results, it is possible to perform multiruns and runs on random datasets. ModuleMaster compares all those runs automatically and generates comprehensive HTML reports that include all available information. An example can be found on the URL given in Section 4.

3. Performance

The implementation is fast and therefore allows the user to search for modules with more than 10 transcription factors in a large window (e.g., 500 bp) in reasonable time. To our knowledge, ModuleMaster is the first algorithm that is able to handle such a large search. For example, a simple module search (five transcription factors in a module) in a cluster of about 20 genes with about 1000 potential transcription factors can be done in less than 1 h on a 1.8 GHz dual core processor. The population size for this example was set to 100 and up to 10,000 generations were performed. ModuleMaster is able to take advantage of multicore and hyperthreading processors by automatically using every available actual and virtual processor.

4. Availability and requirements

ModuleMaster is entirely written in Java TM and runs on any operating system where a suitable Java Virtual Machine (JDK version 1.6 or newer) is installed. For the graphical user interface, the Standard Widget Toolkit (SWT) is required. A list of operating systems supported by SWT can be found at http://www.eclipse.org/swt/. We recommend launching the application by using the webstart link at http://www.ra.cs.uni-tuebingen.de/software/ModuleMaster/. The webstart version is always up to date and will automatically load all required libraries for any supported operating system.

Conflict of interest

None.

Acknowledgments

Funding: German Federal Ministry of Education and Research (BMBF) [National Genome Research Network (NGFNplus) under grant number 01GS08134; HepatoSys under grant number 0313080 L].

References

Aerts, S., Van Loo, P., Thijs, G., Moreau, Y., De Moor, B., 2000. Computational detection of cis-regulatory modules. Bioinformatics 19 (Suppl. 2), http://view.ncbi.nlm.nih.gov/pubmed/14534164.

Aerts, S., Van Loo, P., Thijs, G., Mayer, H., de Martin, R., Moreau, Y., De Moor, B., 2005. TOUCAN 2: the all-inclusive open source workbench for regulatory sequence analysis. Nucleic Acids Res. 33 (Web Server issue), W393–396, http://nar.oxfordjournals.org/cgi/content/abstract/33/suppl_2/W393.

Bussemaker, H.J., Li, H., Siggia, E.D., 2001. Regulatory element detection using correlation with expression, RECOMB '01. In: Proceedings of the Fifth Annual International Conference on Computational Biology, p. 86.

Das, D., Banerjee, N., Zhang, M.Q., 2004. Interacting models of cooperative gene regulation. Proc. Natl. Acad. Sci. 101, 16234–16239.

Higo, K., Ugawa, Y., Iwamoto, M., Korenaga, T., 1999. Plant cis-acting regulatory DNA elements (PLACE) database: 1999.n. Nucleic Acids Res. 27 (1), 297–300, http://view.ncbi.nlm.nih.gov/pubmed/9847208.

Hubbard, T.J.P., Aken, B.L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., Coates, G., Fairley, S., Fitzgerald, S., Fernandez-Banet, J., Gordon, L., Graf, S., Haider, S., Hammond, M., Holland, R.C.G., Howe, K., Jenkinson, A., Johnson, N., Kahari, A., Keefe, D., Keenan, S., Kinsella, R., Kokocinski, F., Kulesha, E., Lawson, D., Longden, I., Megy, K., Meidl, P., Overduin, B., Parker, A., Pritchard, B., Rios, D., Schuster, M., Slater, G., Smedley, D., Spooner, W., Spudich, G., Trevanion, S., Vilella, A., Vogel, J., White, S., Wilder, S., Zadissa, A., Birney, E., Cunningham, F., Curwen, V., Durbin, R., Fernandez-Suarez, X.M., Herrero, J., Kasprzyk, A., Proctor, G., Smith, J., Searle, S., Flicek, P., 2009. Ensembl 2009. Nucl. Acids Res. 37 (Suppl. 1), D690–697, doi:10.1093/nar/gkn828.

Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E., 2003. MATCH TM: a tool for searching transcription factor binding sites in dna sequences. Nucleic Acids Res. 31 (13), 3576–3579, doi:10.1093/nar/gkg585.

Kronfeld, M., 2008. EvA2 Short Documentation. Center for Bioinformatics Tübingen, University of Tübingen, http://www.ra.cs.uni-tuebingen.de/software/EvA2/.

Liu, X., Yu, X., Zack, D.J., Zhu, H., Qian, J., 2008. TiGER: a database for tissue-specific gene expression and regulation. BMC Bioinformatics 9, 271, doi:10.1186/1471-2105-9-271.

Loo, P.V., Aerts, S., B., Thienpont., Moor, B.D., Moreau, Y., Marynen, P., 2008. ModuleMiner—improved computational detection of cis-regulatory modules: are there different modes of gene regulation in embryonic development and adult tissues? Genome Biol. 9 (4), doi:10.1186/gb-2008-9-4-r66.

Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E., 2009. TRANSFAC® and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34 (Database issue), http://view.ncbi.nlm.nih.gov/pubmed/16381825.

Podvinec, M., Kaufmann, M.R., Handschin, C., Meyer, U.A., NUBIScan, 2002. An in silico approach for prediction of nuclear receptor response elements. Mol. Endocrinol. 16, 1269–1279.

Reiss, D.J., Baliga, N.S., Bonneau, R., 2006. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. BMC Bioinformatics 7, 280, doi:10.1186/1471-2105-7-280.

Roven, C., Bussemaker, H.J., 2003. REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. Nucleic Acids Res. 31, 3487–3490.

Schröder, A., Wrzodek, C., Dräger, A., Wanke, D., Berendzen, K.W., Wollnik, J., Harter, K., Zell, A., submitted for publication. ModuleMaster: a novel algorithm to detect functional combinations of regulators for sets of co-expressed genes. Biosystems.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Dana Botstein, D., Koller, N., Friedman, 2003. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. 34, 166–176.

Supper, J., Eichner, J., Wanke, D., Schröder, A., Henneges, C., Zell, A., submitted for publication. Predicting DNA-binding specificities of eukaryotic transcription factors—transferring functional data between proteins. Genome Biol.

Supper, J., Strauch, M., Wanke, D., Harter, K., Zell, A., 2007. EDISA: extracting biclusters from multiple time-series of gene expression profiles. BMC Bioinformatics 8, 334, doi:10.1186/1471-2105-8-334.

Thomas-Chollier, M., Sand, O., Turatsinze, J.-V., Janky, R., Defrance, M., Vervisch, E., Brohee, S., van Helden, J., 2008. RSAT: regulatory sequence analysis tools. Nucleic Acids Res. 304, doi:10.1093/nar/gkn304.

yWorks GmbH, 2003. yFiles Developer's Guide.