Biochemical and Biophysical Research Communications 386 (2009) 559–562

Contents lists available at ScienceDirect

Biochemical and Biophysical Research Communications

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

A graph model based study on regulatory impacts of transcription factors of *Drosophila melanogaster* and comparison across species

Feng Tian [a,b], Jia Chen [b], Suying Bao [c], Lin Shi [a], Xiangjun Liu [a,b,*], Robert Grossman [b,*]

[a] School of Medicine, Tsinghua University, Beijing 100084, PR China
[b] Dept. of Mathematics, Statistics, & Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA
[c] College of Bioinformatics Science and Technology, Harbin Medical University, Harbin 150081, PR China

ARTICLE INFO

Article history:
Received 22 May 2009
Available online 16 June 2009

Keywords:
ChIP chip
Co-regulation
Regulatory distance

ABSTRACT

Transcription factor binding sites and the *cis*-regulatory modules they compose are central determinants of gene regulation. The gene regulations in some model species have been well addressed. However, not as much is known about the fly due to the lack of experimental data. To study the transcription regulation of *Drosophila melanogaster* genes, we analyzed the regulation data from ChIP chip experiments as well as the regulatory database. A graph-based approach is applied to study the impacts of each transcription factor to the regulatory network. The model is also applied to *Saccharomyces cerevisiae* and *Homo sapiens* to study the behaviors of transcription factors in different species. Gene ontology annotations were used for further studies of the biological significance of studied transcription factors.

Transcriptional regulation is a key component of gene regulation, which plays a major role in all forms of cellular differentiation and function. To understand the mechanisms that regulate gene expression, it is important to identify and define the network of *cis*-acting DNA regulatory elements, which can be viewed as the regulatory code wired within the genome. One important method for *cis*-regulatory element detection is based on the concept of the *cis*-regulatory module (CRM) [1,2]. Regulated spatial and temporal control of gene transcription is a fundamental process for all metazoans. Critical to this process is the interaction of transcription factors (TFs) with specific *cis*-regulatory DNA sequences. These regulatory sequences—for instance, enhancers and promoters—are organized in a modular fashion, with each module containing one or more binding sites for a specific combination of TFs [1]. Each module regulates a particular temporal-spatial pattern of gene expression that is a subpart of the entire expression pattern of its associated gene; at the molecular level, each contains a series of binding sites, called transcription factor binding sites (TFBSs) for a specific complement of TFs. Most of the extensively studied CRMs, in particular the enhancers of the *Drosophila* early patterning genes, consist of a dense cluster of TFBSs containing multiple occurrences of binding sites for a small number of transcription factors [2–4].

Despite the importance of *cis*-regulatory elements for many areas of biology, the majority of CRMs are still not known and, of those that are, relatively few have been characterized in detail [5]. For *Drosophila melanogaster*, CRMs are associated with fewer than 2% of the total genes, and fewer than 1% of *Drosophila* genes currently have annotated TFBS data [6].

Chromatin Immunoprecipitation (ChIP) followed by microarray analysis (ChIP-on-chip) is a very powerful global method to identify the genomic regions bound by a transcription factor. However, it is hard to decide the regulated genes even when the TF binding sites are determined, because the regulated genes could be at any location of the binding site. Some approaches have been developed to identify the relationships among genes or proteins using gene ontology terms based on graph theory [7,8]. By combining the ChIP chip data and gene ontology information, we identified high confidence putative target genes of TFs and the regulatory network.

Materials and methods

Data overview. Affymetrx *Drosophila* tiling arrays were obtained from the supplementary materials associated with eight publications, which provided ChIP chip experiments associated with 27 transcription factors. The regulatory data from the Redfly database, which is a collection of known *Drosophila* transcriptional CRMs and TFBSs, is also included for study. *Saccharomyces cerevisiae* transcription regulation data were retrieved from the YEASTRACT database and human regulatory information was obtained from the TRANSFAC public data. Details about the datasets used in this study are listed in Table 1.

* Corresponding authors. Fax: +86 10 62792995 (X. Liu), +1 312 355 0373 (R. Grossman).
E-mail addresses: frankliu@tsinghua.edu.cn (X. Liu), grossman@uic.edu (R. Grossman).

0006-291X/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.bbrc.2009.06.055

Table 1
Datasets used in this study.

Species	Data type	Number of TFs	Data source
S. cerevisiae	Based on experimental evidence	174	YEASTRACT [9]
D. melanogaster	Affymetrx Drosophila tiling arrays	96	Schwartz, Kahn et al. [10], Georlette, Ahn et al. [11], Isogai, Takada et al. [12], Matsumoto, Ukai-Tadenuma et al. [13], Kwong, Adryan et al. [14], Lee, Li et al. [15], Li, MacArthur et al. [16], Misulovin, Schwartz et al. [17], and Redfly [5]
H. sapiens	Based on experimental evidence and literature	300	TRANSFAC [18]

Affymetrix tiling array analysis. Affymetrix tiling arrays were analyzed with MAT [19]. The mock (non-specific antibody or no antibody) arrays were used as controls in the experiments. For MAT analysis, the parameters: BandWidth = 200, MaxGap = 100, MinProbe = 10, Pvalue = 1e-05 were used in the analyses. When there were more than two replicates for each experiment, Var = 1 was set.

Affymetrix BPMAP files were re-mapped to the most up-to-date full genome (including repeats) using xMAN [20]. These newly generated BPMAP files further stored the copy number of each 25-mer and removed probe redundancy to ensure that the same 25-mer map appeared no more than once within any 1 kb window along the genome. The UCSC (http://genome.ucsc.edu/) dm3 RepeatMasker and simple repeats files were downloaded and used to create a Repeat Library file for use with MAT [19]. The outputs of MAT were further filtered using FDR values for each binding. Probe FDR values of 0.1% were used.

Identification or retrieval of putative target genes (PTG). The *Drosophila* gene annotations were extracted from Flybase (release 5.8) [21,22], which contains the location information of 15,145 genes. The identification of putative target genes was based on the distance between the center of the binding site and the transcription start site (TSS) of each gene, and the closest gene was assigned as the putative target gene.

Graph theoretical regulatory distance study. A graph theory based approach was applied to study the relationship between each TF and the whole network. We define the *graph regulatory distance* as the distance between a transcription factor and a gene as follows. First, connect all TFs and their putative target genes to produce a graph. Define the distance between a TF and a gene as the smallest number of edges from the TF to the gene on the graph. Fig. 1 shows an example of the regulatory distance between TF 1 (blue node) and other TFs or genes. Sometime this distance is referred to informally as the Bacon distance.

To compute the graph theoretic regulatory distance, first, connect all the TFs and their putative target genes using the ChIP chip data. To find the regulatory distance between a TF and a gene or another TF, all possible paths between this TF and the gene or the other TF are computed. The shortest path required for the TF to reach the gene is defined as the regulatory distance for this TF and the target gene or TF.

We use $d(TF, g)$ to denote the regulatory distance between a transcription factor TF and one of its target genes g. Suppose there are a total of I transcription factors in a study, let TF_i denote the ith TF, $i = 1, 2, \ldots, I$. For each TF_i, there are a number of target genes this TF_i connects to, and let this number be denoted as J_i. Let $g_{i,j}$ de-

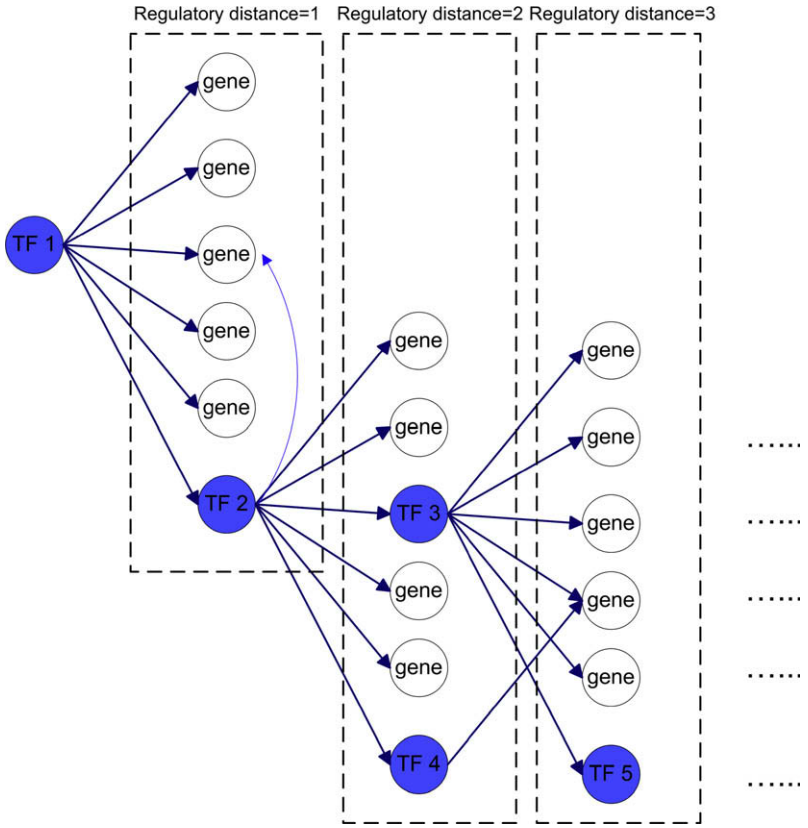

Fig. 1. Regulatory distance between TF and genes.

note the jth target gene of the ith TF, where i is from 1 to I and j is from 1 to J_i for each i.

Define:

$$m_{i,j} = d(TF_i, g_{i,j}).$$

Then for a given TF_i, we calculated the following regulatory distance values:

Average TF-specific regulatory distance = $Avg(m_{i,j})$, where i is fixed and j is from 1 to J_i.

Call this ARD_{TF_i}.

Maximum TF-specific regulatory distance = $Max(m_{i,j})$, where i is fixed and j is from 1 to J_i.

Call this MRD_{TF_i}.

For the entire collection of all TFs, we calculated the following values:

Maximum overall regulatory distance = $Max(MRD_{TF_i})$, where i is from 1 to I.

Call this MRD_{ALL}.

Clearly, MRD_{ALL} is also equal to $Max(m_{i,j})$ where i is from 1 to I and j is from 1 to J_i for each i.

Maximum TF-specific average regulatory distance = $Max(ARD_{TF_i})$, where i is from 1 to I.

Call this MRD_{TFs}.

Average overall regulatory distance = $Avg(m_{i,j})$, where i is from 1 to I and j is from 1 to J_i for each i.

Call this ARD_{ALL}.

GO relative specificity similarities assigning. The relations between gene ontology terms were downloaded from http://geneontology.org/. The gene ontology (GO) annotations of *Drosophila* genes were retrieved from Flybase [21,22]. Wu et al. [8] introduced an approach to reveal protein–protein interaction network based on the GO terms and graph theory. The approach identified 78% of the known yeast protein–protein interactions with the relative specificity similarity (RSS) of the TF-gene pair satisfying the criteria $RSS_{biological\_process} > 0.8$ and $RSS_{cellular\_component} > 0.8$. Here we used a similar approach to identify the high confidence putative target genes (HCPTGs) among the target genes generated based on the location of TSS.

The RSS score between two terms are calculated using the following formula:

$$RSS(\text{term}_i, \text{term}_j) = \frac{\max Depth^{GO}}{\max Depth^{GO} + \gamma} \times \frac{\alpha}{\alpha + \beta}$$

To calculate the RSS between two terms (let's say term_i and term_j), we first define the most recent common ancestor (MRCA) of term_i and term_j, which represents the most specific of all common ancestors of the term pair. Then α is the value of measuring how specific the MRCA of the two terms is according to the structure of the GO; β measures how relatively general term_i and term_j are in the GO; and γ measures the local distance between two terms relative to the MRCA. Finally $\max Depth^{GO}$ is the maximum distance from the root term of the GO to the leaf terms.

In gene ontology, terms of the annotation of genes are divided into three categories, biological process, cellular component and molecular function. A biological process describes a series of events accomplished by one or more ordered assemblies of molecular functions; a cellular component defines a component of a cell, but with the requirement that it is part of some larger object; and, a molecular function describes activities, such as catalytic or binding activities that occur at the molecular level [23]. In this study, cellular component and molecular function terms are used for target gene filtering, and the algorithm was implemented using Perl scripts.

Table 2
Comparison of regulatory distances among species.

	S. cerevisiae	*D. melanogaster*	*H. sapiens*
Number of TFs in the study	174	96	300
Annotated genes used in the study	6270	15,145	40,608
Maximum overall regulatory distance (MRD_{ALL})	9	6	6
Maximum average regulatory distance (MRD_{TFs})	5.99	4.35	3.70
Average overall regulatory distance (ARD_{ALL})	3.36	2.57	2.16

Results and discussion

Graph theoretical regulatory distance reveals impact of each TF on regulatory network

The number of genes that a TF can reach is used to measure the coverage of regulated genes for this TF and its impact on the whole network. To quantify the importance of each TF (as determined by the graph regulatory distance), the number of genes that each TF can reach with different regulatory distances was calculated.

The *average regulatory distance* is defined as the average distance for a TF to reach other nodes on the graph. The *coverage* is defined as the number of genes or TFs reached by a TF. The coverage measures how many genes are regulated either directly or indirectly by a specified TF, while the average regulatory distance measures the average path required to reach other genes.

The TF tramtrack (Flybase ID: FBgn0003870), which is a *D. melanogaster* transcription factor associated with RNA polymerase II, has an average regulatory distance of 4.35. This means that tramtrack connects to other genes relatively indirectly. The details of this analysis for each species can be found in supplementary materials.

Table 2 shows the comparison of maximum regulatory distances for TFs among *S. cerevisiae*, *D. melanogaster* and *H. sapiens*. For *D. melanogaster*, transcription factors have larger coverage, and smaller average regulatory distance to reach their putative target genes, while the TFs of *S. cerevisiae* require longer paths to get to their target genes, which suggests that TFs in *S. cerevisiae* tend to regulate fewer genes directly. In *H. sapiens*, the average regulatory distance of TFs are smaller, which suggests that in higher organisms the functionalities of TFs trend to be more specific and direct.

Gene ontology validation of TFs co-relation

We applied the RSS score to the known 233 TFBS from Redfly, and 163 of them have been annotated with at least one GO category of 'molecular_function' or 'cellular_component'. A high RSS score between two genes indicates that they are highly co-related. The TFBS data from Redfly were taken as the true positives. According to the GO RSS, the TF-gene pairs are divided into three areas, namely high confidence areas as marked in red in Table 3, medium confidence areas marked in blue, and low confidence areas marked in grey. Results show that, 95 (58.3%) out of 163 TF-gene pairs in Redfly fall into the high confidence area, 56 (34.4%) fall into the medium confidence area, while only 12 (7.4%) fall into the low confidence area.

Table 3
The GO RSS scores for TF-gene pairs in Redfly.

		Cellular component		
		(0, 0.3]	(0.3, 0.8]	(0.8, 1]
Molecular function	(0, 0.3]	0	3	0
	(0.3, 0.8]	9	23	30
	(0.8, 1]	0	3	95

We calculated the putative target genes of each TF (introduced in the method section), and used GO RSS score filtering to find high confidence results. We applied this strategy to identify putative target genes of each TF and filtered the extracted target genes by the GO RSS score. Five hundred fifty six high confidence putative target genes and 1696 high confidence TF-putative target gene pairs were identified. We found that 69.1% (384 out of 556) of the high confidence putative target genes are regulated by two or more TFs. Note that using this method of filtering the putative target genes by GO RSS scores might cause the loss of true positives due to the lack of GO annotation or the criteria being too strong. However, we still found that more than half of the entire target genes are co-regulated. This ratio indicates that co-regulation in *D. melanogaster* is very common. The full list of the co-regulated genes can be found in our supplementary materials.

The most regulated gene mirror (Flybase ID: FBgn0014343) with the highest number of co-regulators has been identified with the criteria of GO RSS score. This gene plays an important role in the *D. melanogaster* embryonic development via the syncytial blastoderm, and positively regulates transcription and peripheral nervous system development. This further supports the co-regulation and the importance of this gene between the regulators during the development of *D. melanogaster*.

Conclusion

In this study, we identified the TFBSs of *D. melanogaster* TFs as well as their co-regulated genes. Using the GO RSS, the regulation relationship has been further validated. We introduced the graph regulatory distance to study the impact of TFs in the regulatory network. We also compared the graph regulatory distance for different species, and noted that there is some evidence that higher species have shorter average graph regulatory distance.

Acknowledgments

This work was funded in part by the Chicago Biomedical Consortium with support from The Searle Funds at The Chicago Community Trust and the Chinese Student-Exchange Program.

References

[1] E.H. Davidson, The Regulatory Genome: Gene Regulatory Networks in Development and Evolution, Academic Press, Burlington, MA, 2006.
[2] M.I. Arnone, E.H. Davidson, The hardwiring of development: organization and function of genomic regulatory systems, Development 124 (1997) 1851–1864.
[3] E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C.T. Brown, C.B. Livi, P.Y. Lee, R. Revilla, A.G. Rust, Z. Pan, M.J. Schilstra, P.J. Clarke, M.I. Arnone, L. Rowen, R.A. Cameron, D.R. McClay, L. Hood, H. Bolouri, A genomic regulatory network for development, Science 295 (2002) 1669–1678.
[4] M.D. Schroeder, M. Pearce, J. Fak, H. Fan, U. Unnerstall, E. Emberly, N. Rajewsky, E.D. Siggia, U. Gaul, Transcriptional control in the segmentation gene network of *Drosophila*, PLoS Biol. 2 (2004) E271.
[5] M.S. Halfon, S.M. Gallo, C.M. Bergman, REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in *Drosophila*, Nucleic Acids Res. 36 (2008) D594–D598.
[6] C.M. Bergman, J.W. Carlson, S.E. Celniker, *Drosophila* DNase I footprint database: a systematic genome annotation of transcription factor binding sites in the fruitfly, *Drosophila melanogaster*, Bioinformatics 21 (2005) 1747–1749.
[7] H. Wu, Z. Su, F. Mao, V. Olman, Y. Xu, Prediction of functional modules based on comparative genome analysis and gene ontology application, Nucleic Acids Res. 33 (2005) 2822–2837.
[8] X. Wu, L. Zhu, J. Guo, D.Y. Zhang, K. Lin, Prediction of yeast protein–protein interaction network: insights from the gene ontology and annotations, Nucleic Acids Res. 34 (2006) 2137–2150.
[9] M.C. Teixeira, P. Monteiro, P. Jain, S. Tenreiro, A.R. Fernandes, N.P. Mira, M. Alenquer, A.T. Freitas, A.L. Oliveira, I. Sa-Correia, The YEASTRACT database: a tool for the analysis of transcription regulatory associations in *Saccharomyces cerevisiae*, Nucleic Acids Res. 34 (2006) D446–D451.
[10] Y.B. Schwartz, T.G. Kahn, D.A. Nix, X.Y. Li, R. Bourgon, M. Biggin, V. Pirrotta, Genome-wide analysis of polycomb targets in *Drosophila melanogaster*, Nat. Genet. 38 (2006) 700–705.
[11] D. Georlette, S. Ahn, D.M. MacAlpine, E. Cheung, P.W. Lewis, E.L. Beall, S.P. Bell, T. Speed, J.R. Manak, M.R. Botchan, Genomic profiling and expression studies reveal both positive and negative activities for the *Drosophila* Myb MuvB/dREAM complex in proliferating cells, Genes Dev. 21 (2007) 2880–2896.
[12] Y. Isogai, S. Takada, R. Tjian, S. Keles, Novel TRF1/BRF target genes revealed by genome-wide analysis of *Drosophila* Pol III transcription, EMBO J. 26 (2007) 79–89.
[13] A. Matsumoto, M. Ukai-Tadenuma, R.G. Yamada, J. Houl, K.D. Uno, T. Kasukawa, B. Dauwalder, T.Q. Itoh, K. Takahashi, R. Ueda, P.E. Hardin, T. Tanimura, H.R. Ueda, A functional genomics strategy reveals clockwork orange as a transcriptional regulator in the *Drosophila* circadian clock, Genes Dev. 21 (2007) 1687–1700.
[14] C. Kwong, B. Adryan, I. Bell, L. Meadows, S. Russell, J.R. Manak, R. White, Stability and dynamics of polycomb target sites in *Drosophila* development, PLoS Genet. 4 (2008) e1000178.
[15] C. Lee, X. Li, A. Hechmer, M. Eisen, M.D. Biggin, B.J. Venters, C. Jiang, J. Li, B.F. Pugh, D.S. Gilmour, NELF and GAGA factor are linked to promoter-proximal pausing at many genes in *Drosophila*, Mol. Cell. Biol. 28 (2008) 3290–3300.
[16] X.Y. Li, S. MacArthur, R. Bourgon, D. Nix, D.A. Pollard, V.N. Iyer, A. Hechmer, L. Simirenko, M. Stapleton, C.L. Luengo Hendriks, H.C. Chu, N. Ogawa, W. Inwood, V. Sementchenko, A. Beaton, R. Weiszmann, S.E. Celniker, D.W. Knowles, T. Gingeras, T.P. Speed, M.B. Eisen, M.D. Biggin, Transcription factors bind thousands of active and inactive regions in the *Drosophila* blastoderm, PLoS Biol. 6 (2008) e27.
[17] Z. Misulovin, Y.B. Schwartz, X.Y. Li, T.G. Kahn, M. Gause, S. MacArthur, J.C. Fay, M.B. Eisen, V. Pirrotta, M.D. Biggin, D. Dorsett, Association of cohesin and Nipped-B with transcriptionally active regions of the *Drosophila melanogaster* genome, Chromosoma 117 (2008) 89–102.
[18] V. Matys, O.V. Kel-Margoulis, E. Fricke, I. Liebich, S. Land, A. Barre-Dirrie, I. Reuter, D. Chekmenev, M. Krull, K. Hornischer, N. Voss, P. Stegmaier, B. Lewicki-Potapov, H. Saxel, A.E. Kel, E. Wingender, TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, Nucleic Acids Res. 34 (2006) D108–D110.
[19] W.E. Johnson, W. Li, C.A. Meyer, R. Gottardo, J.S. Carroll, M. Brown, X.S. Liu, Model-based analysis of tiling-arrays for ChIP-chip, Proc. Natl. Acad. Sci. USA 103 (2006) 12457–12462.
[20] W. Li, J.S. Carroll, M. Brown, S. Liu, xMAN: extreme MApping of OligoNucleotides, BMC Genomics 9 (Suppl. 1) (2008) S20.
[21] G. Grumbling, V. Strelets, FlyBase: anatomical data. Images and queries, Nucleic Acids Res. 34 (2006) D484–D488.
[22] M. Ashburner, R. Drysdale, FlyBase–the *Drosophila* genetic database, Development 120 (1994) 2077–2079.
[23] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. The gene ontology consortium, Nat. Genet. 25 (2000) 25–29.