FUNCTIONAL TRENDS IN STRUCTURAL CLASSES OF THE DNA BINDING DOMAINS OF REGULATORY TRANSCRIPTION FACTORS

RACHEL PATTON MCCORD[1,4] AND MARTHA L. BULYK[*,1,2,3,4]

[1]Division of Genetics, Department of Medicine, [2]Department of Pathology, Brigham & Women's Hospital and Harvard Medical School, Boston, MA 02115
[3]Harvard/MIT Division of Health Sciences & Technology (HST), Harvard Medical School, Boston, MA 02115
[4]Harvard University Graduate Biophysics Program, Cambridge, MA 02138
Email: rpmccord@fas.harvard.edu, mlbulyk@receptor.med.harvard.edu

The DNA-binding domain (DBD) structure of a regulatory transcription factor (TF) is important in determining its DNA sequence specificity, but it is unclear whether a relationship exists between DBD structure and general TF biological function or regulatory mechanism. We observed moderate enrichment of functional annotation terms among TFs of the same structural class in *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster,* or *Mus musculus*, suggesting some preference for TFs of similar structures in the regulation of similar processes. In yeast, we also found trends among TF structural classes in phenomena including gene expression coherence, DNA binding site motif similarity, the general or specific nature of TFs' regulatory roles, and the position of a TF in a gene regulatory network. These results suggest that the biophysical constraints of different TF structural classes play a role in their gene regulatory mechanisms.

1. Introduction

The concepts that structure leads to function and that form follows function are common principles throughout biology[1]. In the study of gene regulation, TFs can be classified based on the structures of their DBDs, domains that mediate their interaction with specific DNA sequences[2,3]. These structural class designations have been used to infer the sequence specificity of a TF, predict binding sites and potential target genes, and infer biological function based on these target genes[4-7]. Since TF sequence specificities have been used to infer TF functional properties, it follows that members of a given TF structural class might have similar biological roles, and that the structure of a DBD could be used directly to predict the functions of uncharacterized TFs. Indeed, previous studies have identified instances of enrichment of a particular TF structural class in the regulation of a certain biological process. For example, homeodomains are enriched within genes involved in *C. elegans* neuronal function[8]. However, a

---

*[*] Corresponding author

large-scale analysis to determine the extent of functional enrichment within different TF structural classes has not been described previously.

TFs of the same class might also share other gene regulatory properties, such as their position in gene regulatory networks, the similarity or divergence and information content of their DNA binding site motifs, or co-expression across diverse conditions. Analysis of such regulatory features will elucidate ways in which the biophysical properties of a DBD structure might inform its modes of regulation.

Here, we investigate enrichment for common biological function among members of different TF structural classes in *E. coli, S. cerevisiae, D. melanogaster,* and *M. musculus*. We find several examples of modest functional enrichment among TFs of the same structural class in bacteria, yeast, fly, or mouse. Target genes of yeast TFs within some structural classes are also observed to share similar functions. In a few cases, the biological functions enriched for a particular structural class appear to be conserved across species. Using numerous genome- and proteome-wide datasets available in *S. cerevisiae*, we relate this observed functional enrichment to other regulatory mechanisms. Our results suggest that different modes of gene regulation are used by different TF structural classes. The functional relationships found here identify cases in which DBD structure could be used to predict TF biological function, suggest different ways in which structural classes partition functional roles, and inform future studies of the link between TF structure and function and the evolution of TF regulatory roles.

2. Methods

2.1. Data Sets Used in This Study

TFs and DBD Structural Classes

The TFs and structural class assignments for *E. coli* were obtained from GenProtEC[9], last updated on Dec 7, 2004. The structural classes of 421 known and predicted *S. cerevisiae* TFs[10] were assigned based on annotation in Pfam[11] and DBD[12] databases. For subsequent analyses, we considered only the subset of TFs from this initial list that belonged to known DBD structural classes with 4 or more members. *D. melanogaster* TFs and structural classifications were downloaded from FlyBase on July 11, 2006[13]. Mouse TF information and DBD

assignments were derived from a set of known TFs listed in Gray *et al.*[14]. All TFs and structural class assignments are listed in Supplementary Table 1[†].

Functional Annotations

Each *E. coli* protein was assigned MultiFun classifications according to the GenProtEC database, last updated on February 1, 2007[9]. Specific annotations were divided into corresponding broader categories (i.e., a protein annotated "1.3.5: Fermentation" would also be given the annotations "1: Metabolism" and "1.3: Energy metabolism (carbon)"). Multiple sources of gene annotations, including the Gene Ontology (GO)[15] and MIPS database[16], last updated in June 2005, were used to annotate yeast target genes. We used GO annotations for yeast, fly, and mouse TFs that were last updated on September 12, 2007[17]. To avoid circularity and annotation bias, we eliminated all GO annotations that were inferred from structure or from a non-traceable author statement (GO Evidence Codes ISS and NAS, respectively)[15].

Genome-wide Yeast Datasets

Yeast TF binding site motif sequences, target gene information, and motif information content values (IC; a measure of the specific vs. degenerate nature of the DNA sequences recognized by a TF) for 82 TFs were derived from a re-analysis[18] by MacIsaac *et al.* of the single most comprehensive set of yeast ChIP-chip data[19]. We considered TF binding sites identified at *p*<0.005 binding threshold in ChIP-chip that were also conserved in at least 2 other yeast species. We considered only those structural classes with at least 3 TFs with greater than 5 target genes in our target gene analyses. Yeast gene regulatory interaction data were derived from networks compiled by Yu *et al.*[20]. The 1,327 publicly available gene expression microarray datasets were compiled by McCord *et al.*[21]

2.2. Statistical Approaches

Functional Enrichment Evaluation

To evaluate functional enrichment among groups of TFs or their target genes in bacteria, yeast, fly and mouse, we calculated *p*-values using the hypergeometric distribution:

$$\text{Eqn. (1):} \quad P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}$$

[†] All supplementary files, figures, and scripts (implemented in Perl and Matlab) are available on our lab website at http://the_brain.bwh.harvard.edu/TFstr/

where G is the number of genes in the entire genome or in a defined background gene set, C is the number of genes in this background set with a particular functional attribute, and n is the size of the query set of TFs or target genes, of which k are known to possess the functional attribute.

We evaluated functional enrichment within DBD structural classes in mouse, fly, and yeast with respect to all TFs using the FuncAssociate algorithm[17], which estimates an adjusted p-value (p_{adj}) by comparing the enrichment in the query gene set to the frequency of this degree of enrichment among 1,000 randomly generated gene sets. We report results at $p_{adj}<0.05$. Our implementation of the hypergeometric distribution for *E. coli* allowed us to search for functional enrichment at many levels of the MultiFun[22] annotation hierarchy. Our threshold for functional enrichment in *E. coli* was an uncorrected $p<0.05$, but we also report p-values from a stringent Bonferroni correction.

To evaluate TF target gene functional enrichment in yeast, we employed the Funspec algorithm[23] to calculate p-values of target gene functional enrichment for each TF (p_{TF}) in a class, and then calculated the geometric mean of the p-values for each annotation term over all TFs in a structural class (p_{avg}). We controlled for a potential inflationary effect on this functional enrichment, resulting from the existence of paralogous TFs due to the ancient yeast genome duplication, by calculating filtered p_{avg} values that excluded paralogous gene pairs. Specifically, for classes containing paralogous pairs, we calculated all possible filtered p_{avg} values resulting from averaging p-values over all but one TF of a structural class by leaving out, one at a time, members of literature-defined paralogous TFs[24]. We report results for which the least significant of these filtered average p-values (max filtered p_{avg}) was less than 0.05.

Coherence Scores

Co-expression of a set of TFs or target genes in yeast was measured by *expression coherence (EC)*[25]. Briefly, we calculated the Pearson correlation coefficient between the expression profiles of every pair of yeast genes over 1,327 expression conditions[21]. Then, the EC was calculated as the fraction of correlation coefficients between foreground genes (TFs or target genes in a DBD class) that were in the top 5th percentile of correlations among background genes (all TFs or all genes). In the case of TF target genes, we considered only the pairwise correlations between targets of different TFs within the structural class to ensure that high expression coherence was not attributable solely to regulation of targets of a single TF. A p-value was estimated by calculating the EC scores of 10,000 randomly generated sets of genes identical in size to the

foreground set and then calculating the fraction of random sets with an EC greater than that of the foreground set of interest.

The similarity of DNA binding site motifs recognized by TFs in a structural class was measured by a metric we developed termed *"motif coherence"*, which we modeled after the expression coherence metric described above. The pairwise correlation coefficients between all motifs were calculated by the CompareACE algorithm[26], and then the motif coherence was calculated as the fraction of motif correlations within a structural class in the top 5th percentile of all motif correlations. A *p*-value for this coherence was estimated as for expression coherence, but here we considered 10 million random sets in order to allow estimation of *p*-values as low as 1.0×10^{-7} and thus to provide finer distinctions in the degree of motif coherence among structural classes with highly similar DNA binding domains.

Bottlenecks and Hubs

We classified yeast TFs as *"hubs"* if they were in the top 20% of the regulatory network degree distribution and as *"bottlenecks"* if they were in the top 20% of the betweenness distribution, as in Yu *et al.*[20]. The hypergeometric distribution (Eqn. 1) was used to assign a *p*-value to hub/bottleneck enrichment within a structural class by comparing the fraction of hubs/bottlenecks within a structural class to the fraction of hubs/bottlenecks over all TFs.

3. Results and Discussion

3.1 Functional Enrichment by TF Structural Class

We first searched for functional enrichment within a structural class by examining gene annotation terms assigned to the TFs themselves. Modest functional enrichment was seen for some structural classes in all 4 organisms, (see Table 1 for highlights of enriched annotations and Supplementary Table 2 for full results) though some classes in each organism showed enrichment for no biological functions, or only those common to most transcriptional regulatory proteins (*e.g.* "transcription, DNA dependent"). In *E. coli*, most classes showed some degree of functional enrichment; winged-helix TFs are enriched for roles in amino acid biosynthesis, while proteins with lambda repressor DBDs are enriched for carbohydrate metabolism functions. In fly, 40% of classes showed no specific enrichment, but classes like the HLH TFs and homeodomains are enriched for roles in the development of various systems. The minimal enrichment observed for 40% of mouse TF classes may be due to a lack of comprehensive GO annotation for most mammalian genes. However, as in fly,

Table 1. Highlighted examples of enriched functional annotation terms for DBD structural classes. k = number of TFs in structural class with the indicated annotation term; C = number of genes in background set (all TFs) with the indicated annotation term; p = p-value of functional annotation enrichment calculated using the hypergeometric distribution. p_{adj} = adjusted p-value calculated as described in Methods.

Annotation	k	C	p	p_{adj}
E. coli (N_{TFs} = 225)				
Winged Helix (FKH-like) (n=111):				
Repressor	69	101	2.2E-07	2.8E-05
Building block biosynthesis	10	12	0.015	1.9
Amino acid biosynthesis	8	9	0.016	2.0
Lambda repressor like (n=24):				
Carbohydrate catabolism	8	24	0.0012	0.16
Metabolism	15	79	0.0036	0.46
Yeast (N_{TFs} = 346)				
GATA (n=10):				
Regulation of nitrogen metabolism	4	4	3.6E-07	<0.001
Homeodomains (n=7):				
Mating-type specific transcriptional control	3	4	2.0E-05	0.003
G1/S-specific transcription in cell cycle	3	9	0.00041	0.05
Forkhead (n = 4)				
Positive regulation of progression through cell cycle	2	2	1.0E-04	0.018
G2/M-specific transcription in cell cycle	2	3	3.0E-04	0.031
Fly (N_{TFs} = 573)				
Homeodomains (n=102):				
System development	73	186	2.4E-16	<0.001
Pattern specification process	32	69	4.6E-08	<0.001
Cell fate specification	13	25	0.00022	0.022
Helix-Loop-Helix (HLH) (n=54):				
Sensory organ development	17	58	3.7E-07	0.004
Mouse (N_{TFs}=1,160)				
Homeodomains (n=221):				
Organ development	108	303	9.3E-16	<0.001
Central nervous system development	38	74	1.3E-10	<0.001
Endocrine system development	16	29	1.4E-05	0.004
Cell migration/motility	20	46	0.00011	0.018
Forkhead (FKH) (n=40):				
Organ development	25	303	1.2E-06	<0.001
Cell proliferation	11	72	1.5E-05	0.002
E2F (n=8):				
Regulation of progression through cell cycle	8	48	5.4E-12	<0.001
G1/S transition of mitotic cell cycle	3	4	9.0E-07	<0.001

some structural classes in mouse, such as homeodomains and forkhead TFs, are enriched for roles in organism development, and, as expected, the E2F TFs showed enrichment for roles in cell cycle control[27]. In *S. cerevisiae*, some structural classes (HLH, HSF, and others) showed no functional enrichment. Other classes are enriched for regulation of specific biological pathways, including GATA factors for regulation of nitrogen utilization, forkhead TFs in

Table 2: Highlighted examples of enriched functional annotation terms among target genes (N_{tar}) of yeast TFs. The values of p_{avg} and *max filtered p_{avg}* were calculated as defined in Methods. All genes in the *S. cerevisiae* genome (N_{bg}) were used as the background gene set in the *p*-value calculations.

Annotation	p_{avg}	*max filtered p_{avg}*
Yeast Target Genes (N_{bg} = 6,267)		
GATA (N_{TF} = 6; N_{tar} = 177):		
Nitrogen and sulfur metabolism	0.0004	5.1E-05
Myb (N_{TF} = 5; N_{tar} = 525):		
Cell growth and/or maintenance	0.0009	0.00090
Protein metabolism	0.0019	0.0019
Ribosome biogenesis	0.0037	0.0037
APSES (N_{TF} = 4; N_{tar} = 530):		
Cell wall	0.0003	0.0013
Mitotic cell cycle and cell cycle control	0.0043	0.0016
Cell fate	0.0003	0.00050
Cys₂His₂-type Zinc Finger (Zf-C2H2) (N_{TF} = 13; N_{tar} = 824):		
Stress response	0.0083	0.059

cell cycle progression, and homeodomain factors in mating type determination and the cell cycle.

The availability of ChIP-chip data for many yeast TFs allowed us to extend our analysis to the annotations of target genes of yeast TFs (see Table 2 for highlights and Supplementary Table 3 for full results). We observed that the GATA TFs and their target genes are both enriched for the same biological functions: nitrogen and sulfur metabolism. Consideration of target genes also provided additional functional information for several classes, including cell cycle and cell fate target gene enrichment for the APSES TFs, stress response for the C2H2 zinc finger (Zf-C2H2) TFs, and cell growth and protein biosynthesis for the Myb factors. We found that most of the enriched annotations were robust to paralog removal, so functional enrichment is not solely attributable to paralogous TFs resulting from the ancient yeast whole genome duplication[24].

We observed a few instances of functional enrichments that were consistent across organisms. In particular, homeodomain TFs in yeast are enriched for roles in the mating type determination, and the homeodomain TFs in fly and in mouse are enriched for roles in similar cell fate specification and development. Additionally, some basic transcription-related processes are shared across species: HMG factors are enriched for roles in chromatin architecture in both yeast and mouse. However, conservation of functional enrichment for members of a TF structural class is small, suggesting that, in most cases, functional specialization of structural classes arose according to different selective pressures in each of these organisms' evolutionary histories.

3.2 TF and Target Gene Expression Coherence (EC)

Observable functional enrichment within TF structural classes in several organisms suggests that other regulatory features of TFs might relate to this functional enrichment and vary across DBD structures. Since co-expression is often used to infer functional relationships between genes, we hypothesized that structural classes exhibiting functional annotation enrichment might also be co-expressed or exhibit co-expression of their target genes. Thus, we evaluated the EC of TFs or target genes within each structural class in yeast over 1,327 expression conditions (Figure 1). We found a range of EC across TF structural classes, suggesting further distinctions in the regulatory roles of different structural classes. As predicted, many classes with functional enrichment (Zf-C2H2, GATA, Myb, and Forkhead) do show strong EC, particularly among target genes. However, other TFs with enriched functional annotations (APSES, homeodomains) do not exhibit significant EC.
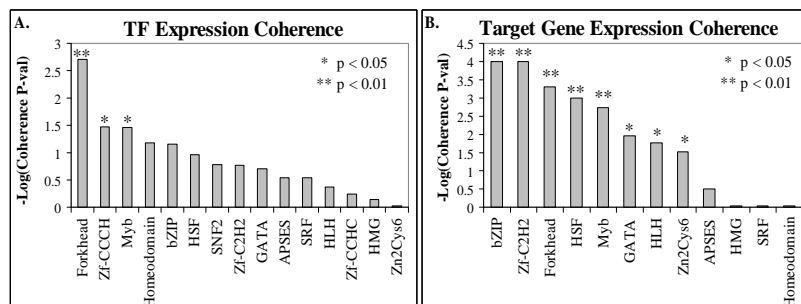

Figure 1. Significance of Expression Coherence scores for A) TFs, and B) TF target genes across structural classes in yeast.

3.3 Regulatory Bottlenecks

Functional enrichment without EC within a structural class may indicate that members of this structural class regulate different phases of the same biological process. Alternatively, lack of EC among targets of the same structural class may arise from regulatory network complexity. We searched for significant trends in network topology among members
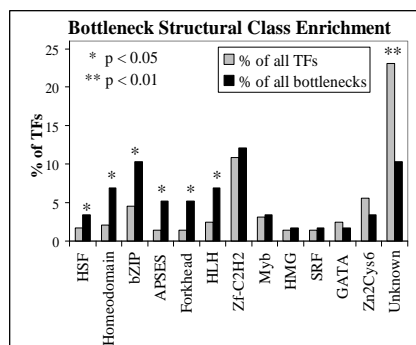

Figure 2. Bottleneck TFs within structural classes. Classes are ordered left to right from most enriched for bottlenecks to most depleted.

of a structural class within experimentally derived regulatory networks. Recent work has shown that "bottleneck" status (a measure of "betweenness", *i.e.*, how often regulatory pathways pass through a particular protein in a network graph) is a meaningful measure of the role of a TF in a regulatory network[20]. We found that certain TF structural classes are significantly enriched ($p < 0.05$) for bottlenecks (Figure 2). Interestingly, APSES and homeodomain TFs, two classes that showed functional enrichment but insignificant EC, are among those enriched for bottlenecks. Since bottleneck proteins often connect multiple biological modules[20], TFs in these classes may regulate genes within different specific pathways expressed at different times, but which all contribute to similar biological functions. Such a regulatory mode could explain the functional enrichment without significant EC observed for these TFs.

3.4 Motif Coherence (MC)

We hypothesized that TFs within structural classes that show functional enrichment should exhibit similarity in their DNA binding site motifs[28]. We observe variation in the degree of MC from one TF structural class to another. Structural classes with strong functional enrichment, even some that do not show significant EC, tend to have highly significant within-class MC (Figure 3). However, some classes with

Figure 3. Motif coherence by TF structural class.

functional enrichment (Myb, forkhead, homeodomain) do not have significant MC, suggesting that motif similarity is not the only factor contributing to similarity in function.

3.5 General vs. Specific Regulation

The binding mechanism of a particular DBD structure might be well-suited for a certain type of regulation, and thus, certain
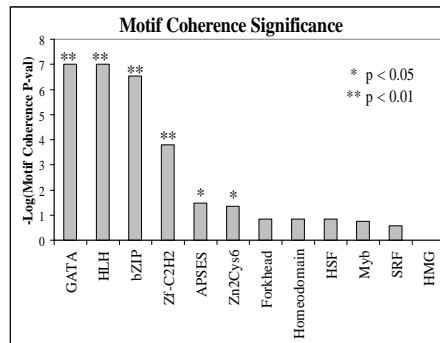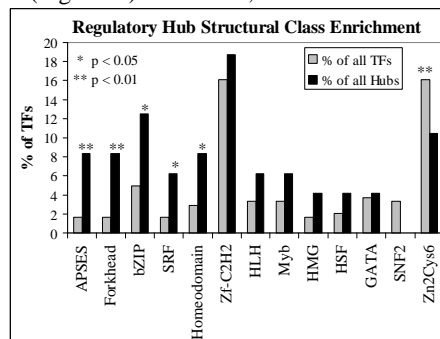

Figure 4. Regulatory hub enrichment within structural classes. Classes are ordered left to right from most enriched for hubs to most depleted.

biological processes. For example, structures that bind more degenerate sequences and/or have many potential binding sites in the genome might be utilized for general, housekeeping functions while structures that recognize highly specific binding sites might be used for processes requiring carefully restricted regulation. We examined trends in the information content (IC; a measure of motif specificity vs. degeneracy) and number of target genes recognized by TFs of each structural class[18]. We observed only modest variation in average motif IC between structural classes, but note that such variation tends to be anti-correlated with the average number of genes identified as bound in ChIP-chip experiments by TFs of the same class, as expected (Supplementary Figure 1). A clearer distinction between classes exists in the enrichment for regulatory hubs (proteins with the most connections in the regulatory network) within each structural class (Figure 4). Structural classes containing well-known "global" TFs (*i.e.*, those regulating many genes for broadly important functions) like the bZIP protein Gcn4 are significantly enriched for regulatory hubs, while those containing known "local" TFs (*i.e.,* those regulating a few genes for a specific function) like the Zn_2Cys_6 TF Gal4 are significantly depleted for such hubs. Thus, the global vs. local nature of these TFs appears to be a general feature of their structural class. Interestingly, structural classes with many regulatory hubs tend to be enriched for cell fate and cell cycle functions while those with fewer regulatory hubs tend to be involved in regulating the metabolism of specific nutrients such as nitrogen and carbohydrates.

4. Conclusions and Future Directions

We have found evidence for biological function enrichment among TFs in various structural classes in a wide range of organisms. We observed differences across structural classes in terms of regulatory features that may relate to this functional enrichment, including expression coherence, motif similarity, and regulatory network position. In addition to suggesting explanations for the observed functional enrichments, such regulatory feature differences indicate that different structural classes may have fundamentally different modes of gene regulation. Specifically, the data presented here suggest that different TF structural classes achieve regulatory specificity and avoid crosstalk in different ways. The combination of low motif coherence, low expression coherence, and lack of functional enrichment within some structural classes suggests that diversity in DNA recognition motifs allows different TFs of the same DBD class to participate in different biological functions and regulate distinct sets of target genes. In other structural classes, similar recognition motifs, high expression coherence, and functional enrichment suggest that harmful crosstalk is avoided

as TFs within a class act redundantly or supplementarily in the regulation of similar processes, as has been previously hypothesized in studies of the function of TFs with similar motifs[28]. Functional enrichment and high motif coherence paired with low expression coherence and an enrichment for regulatory bottlenecks suggests that, in yet other classes, TF function is partitioned into different modules so that all TFs in a class. Thus, though they bind similar motifs and participate in similar biological processes, they perform unique roles in the cell with precise functional specificity determined by their regulatory partners in the overall network.

These results offer a set of interesting correlations and potential distinctions in regulatory mechanism by structural class, but do not provide a mechanistic explanation for the existence of these correlations nor elucidate the causality or order of events that led to functional enrichment within certain TF structural classes. We can, however, note that certain structural classes, like the C2H2 zinc finger TFs, have retained their paralogs after yeast whole genome duplication at a much higher than average rate (Supplementary Figure 2). Interestingly, C2H2 zinc finger TFs have undergone expansion and neofunctionalization within diverse lineages[29,30]. Thus, we can hypothesize that the structural properties and corresponding regulatory mechanisms of certain structural classes made them more suited for neofunctionalization and expansion over evolutionary time.

The regulatory trends for different DBD structural classes could be used to improve gene function prediction. DBD structure is already used indirectly to predict TF function when biological roles are inferred from target genes that were in turn identified using binding sites predicted by structural homology[4,6]. The results presented here indicate that for certain TF structural classes, such as homeodomains in mouse, fly, and yeast, TF function prediction based on DBD structure is likely to be informative. For other TF classes, such as Myb domains in both fly and mouse, however, functional inferences from structure must be interpreted with caution. Likewise, our observed correlations of certain DBD structural classes with various regulatory properties suggest that such regulatory properties could also be included in predictions of TFs' regulatory roles. The resulting predictions of gene function could then be tested by directed experimentation. Beyond experimental testing to validate the predicted functions for novel or poorly characterized TFs, any TFs whose regulatory properties fall outside the general trends presented here could be investigated further to determine whether existing data and annotations have missed certain regulatory aspects of TF function that are expected for members of its structural class.

The trends we observed here may have been affected by incomplete or biased annotations. In the future, as more precise data on the DNA binding specificities of TFs from each structural class and the biological processes they

regulate become available[31], more concrete relationships between these features might be revealed. Analysis of other regulatory features, such as co-regulation within and between classes, other domains associated with a structural class, and the variability of TF and target gene expression could also further elucidate the role of DBD structure in TF function and regulatory mechanism.

5. Acknowledgments

References

1. Gaskell, W.H. *J Physiol* **7,** 1-80.9 (1886).
2. Narlikar, L. and Hartemink, A.J. *Bioinformatics* **22,** 157-63 (2006).
3. Luscombe, N.M., Austin, S.E., Berman, H.M., *et al. Genome Biol* **1,** REVIEWS001 (2000).
4. Tan, K., McCue, L.A., and Stormo, G.D. *Genome Res* **15,** 312-20 (2005).
5. Siggers, T.W. and Honig, B. *Nucleic Acids Res* **35,** 1085-97 (2007).
6. Kaplan, T., Friedman, N., and Margalit, H. *PLoS Comput Biol* **1,** e1 (2005).
7. Narlikar, L., Gordan, R., Ohler, U., *et al. Bioinformatics* **22,** e384-92 (2006).
8. Vermeirssen, V., Barrasa, M.I., Hidalgo, C.A., *et al. Genome Res* **17,** 1061-71 (2007).
9. Serres, M.H., Goswami, S., and Riley, M. *Nucleic Acids Res* **32,** D300-2 (2004).
10. Hu, Y., Rolfs, A., Bhullar, B., *et al. Genome Res* **17,** 536-43 (2007).
11. Bateman, A., Coin, L., Durbin, R., *et al. Nucleic Acids Res* **32,** D138-41 (2004).
12. Kummerfeld, S.K. and Teichmann, S.A. *Nucleic Acids Res* **34,** D74-81 (2006).
13. Grumbling, G. and Strelets, V. *Nucleic Acids Res* **34,** D484-8 (2006).
14. Gray, P.A., Fu, H., Luo, P., *et al. Science* **306,** 2255-7 (2004).
15. Harris, M.A., Clark, J., Ireland, A., *et al. Nucleic Acids Res* **32,** D258-61 (2004).
16. Mewes, H.W., Frishman, D., Guldener, U., *et al. Nucleic Acids Res* **30,** 31-4 (2002).
17. Berriz, G.F., King, O.D., Bryant, B., *et al. Bioinformatics* **19,** 2502-4 (2003).
18. MacIsaac, K.D., Wang, T., Gordon, D.B., *et al. BMC Bioinformatics* **7,** 113 (2006).
19. Harbison, C.T., Gordon, D.B., Lee, T.I., *et al. Nature* **431,** 99-104 (2004).
20. Yu, H., Kim, P.M., Sprecher, E., *et al. PLoS Comput Biol* **3,** e59 (2007).
21. McCord, R.P., Berger, M.F., Philippakis, A.A., *et al. Mol Syst Biol* **3,** 100 (2007).
22. Serres, M.H. and Riley, M. *Microb Comp Genomics* **5,** 205-22 (2000).
23. Robinson, M.D., Grigull, J., Mohammad, N., *et al. BMC Bioinformatics* **3,** 35 (2002).
24. Kellis, M., Birren, B.W., and Lander, E.S. *Nature* **428,** 617-24 (2004).
25. Pilpel, Y., Sudarsanam, P., and Church, G.M. *Nat Genet* **29,** 153-9 (2001).
26. Roth, F.P., Hughes, J.D., Estep, P.W., *et al. Nat Biotechnol* **16,** 939-45 (1998).
27. Kusek, J.C., Greene, R.M., Nugent, P., *et al. Int J Dev Biol* **44,** 267-77 (2000).
28. Itzkovitz, S., Tlusty, T., and Alon, U. *BMC Genomics* **7,** 239 (2006).
29. Huntley, S., Baggott, D.M., Hamilton, A.T., *et al. Genome Res* **16,** 669-77 (2006).
30. Chung, H.R., Lohr, U., and Jackle, H. *Mol Biol Evol* **24,** 1934-43 (2007).
31. Bulyk, M.L. *Curr Opin Biotechnol* **17,** 422-30 (2006).