Identification and characterization of regulatory evolution in yeast

Matt Edwards

Abstract

Phenotypic change arises from differences in protein-coding genes and the regulatory connections determining their patterns of expression. Computational biology has focused heavily on changes in protein sequences and on discovering methods and patterns of transcriptional regulation, but comparatively less effort has been applied to determining instances and modes of regulatory evolution. This study identifies, at a large scale, specific genes and gene families in yeast that have undergone particular patterns of regulatory evolution. Specifically, this project (1) predicts transcriptional regulatory connections (motif instances) in four yeast species, (2) identifies and characterizes genes and functional families experiencing regulatory divergence or conservation, and (3) compares the discovered regulatory divergence to expression divergence (between four *sensu stricto* yeasts) as estimated in previous work. We can use this to predict genes that have switched regulation while maintaining similar expression and genes that have switched expression while maintaining similar regulation.

Background

Previous work has focused on the evolution of expression, sometimes focusing on a single pathway ([18]) or within one species ([27]). Regulatory studies have examined single gene families or a few motif instances ([15], [3]). This project examines the evolution of regulation at a larger scale than these previous studies, albeit at a greatly reduced resolution. It is this tradeoff that differentiates this project from previous work. Many previously published works began with experimental evidence and linked it to motif changes, but this work finds motif instances and uses them as a starting point for further biological investigation. In other words, the proposed changes will not be experimentally validated in this work - only insofar as they agree with existing expression datasets. This project compares expression and regulatory changes genome-wide, rather than finding regulatory explanations for specific instances of regulatory change ([13]). The genome-wide binding analysis using multiple existing motif datasets, combined with a new regulation divergence metric, enables novel analyses of gene categories and correlations. However, the low signal-to-noise ratio of regulatory motifs, even when starting with a known set of consensus sequences, will present challenges.

Methods and Results

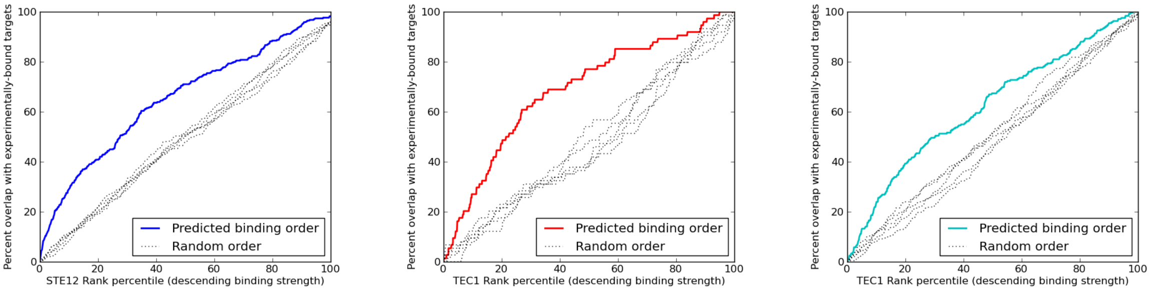
Motif Instance Identification

A compendium of computationally- or experimentally-validated motifs (position weight matrices) was assembled (from [1], [14], [5], [28]). A total of 193 motifs were used, out of the approximately 200 known transcription factors in *S. cerevisiae* ([1]). If the datasets overlapped for a particular factor, the curated or filtered set ([5] then [14]) was arbitrarily picked as authoritative. The binding motifs for these factors were all determined in *S. cerevisiae*, so the analysis continues based on the assumption that binding motifs will not dramatically change in the other species we consider. This is generally true, but novel (not known in *S. cerevisiae*) motifs are detectable ([13] and [3]); our method will miss these factors. Probabilities of a transcription factor binding to promoter regions (defined as 600 bases upstream of the translation start site) were determined using the GOMER (Generalizable occupancy model of expression regulation) tool ([12]). GOMER uses a thermodynamic model designed to model cooperative effects between multiple weak binding sites instead of strict binary or single-site matching. When comparing binding probabilities across genes, GOMER will be more reliable for factors that bind (or at least have sites) in multiple, possibly weak, locations in the promoter, rather than at a single location. While this biases against some factors, it decreases the method's susceptibility to noise. Based on this observation, it is important to note that this procedure in some sense assesses binding potential, rather than actual binding (as discussed in [14]). Another important point (and limitation) is that this step operates at the sequence level for a single factor at a time: competitive effects, nucleosome positioning, and conservation are not taken into account. This is for two main reasons: simplicity and the ability to detect changing binding events across species. The authors in [9] report high rates of binding site turnover and species-specific loss, and this single-species approach, while imprecise, hopes to capture this effect.

For a single genome and transcription factor, GOMER produces binding probabilities for each promoter. In the next section we will discuss how to normalize across factors and genomes, but here we check the predictions against experimental data. We compare the binding predictions against two ChIP-chip datasets: a comparison of two factors across three yeast genomes ([3]) and a comparison of many factors in one yeast genome ([14]). Figure 1 shows examples of similarity between the computational predictions and previous experimental data.

Measuring Regulatory Evolution

Given motif-promoter binding probabilities for a set of motifs and genes across multiple species, a regulatory divergence metric can be applied between any pair of genes. In this analysis, we will use two pairwise divergence metrics: a discrete and a continuous measurement. We will call them RD_D and RD_C, respectively. The discrete measurement RD_D is obtained by rounding each of



(a) (b) (c)

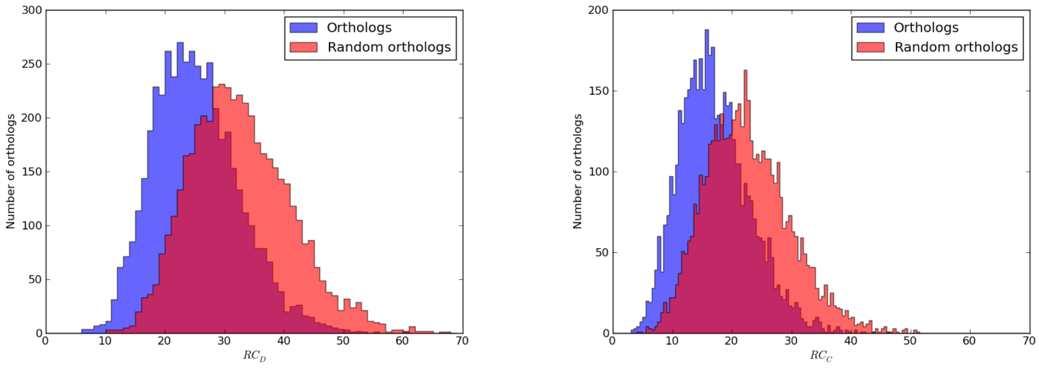
Figure 1: Rank order analysis (as in [3]) showing correlation between GOMER binding probabilities with experimental observations of (a) STE12 in *S. cerevisiae* ([14]) (b) TEC1 in *S. bayanus* ([3]), and (c) TEC1 in *S. mikatae* ([3])

the binding probabilities to "bound" and "unbound," given a particular significance cutoff. We obtain the cutoff for a given factor in a genome by assuming that an empirical null model obtained from shuffled motifs follows a log-normal distribution. We compute the cutoffs independently for each genome and factor, in order to avoid large-scale compositional differences or small changes in binding specificity. In order to include weak binding events, we use the cutoff $p = 0.1$. RD_D is then the number of bound factors that have changed regulation between a pair of genes (and can be calculated across or within species). This is a simplistic measurement, and among other things, assumes that each factor binds independently. However, it provides us a starting point and has been similarly applied in the literature ([27], though only 50 motifs were used in that study, each requiring perfect matches). The continuous measurement RD_C is the sum of log-fold binding probability changes for each factor, similar to some microarray analyses (as in [19]). There is a large range of improbable binding events, so to eliminate the uninformative effect of going from improbable to very improbable we round all improbable binding probabilities up to the median binding probability.

Figure 2 shows RD_D and RD_C between *S. cerevisiae* and *S. bayanus* orthologs. As expected, this metric shows on average a lower regulatory divergence between orthologous pairs than between arbitrary gene pairs. There is a large overlap, but this is similar to expression divergence measurements ([13]) and can also be attributed to the imprecise binding predictions. For subsequent analyses we use the RD_C measurement.

Characterizing Regulatory Evolution

The RD metrics can be used to find originally misannotated orthologs. By comparing the original ortholog annotations to a customized orthology database ([24]), we find the highest 10% of orthologs by RD_C contain a third (7) of the miscalled orthologs (enriched with $p < 0.05$). As a



(a)

(b)

Figure 2: (a) RD_D and (b) RD_C between 4690 *S. cerevisiae* - *S. bayanus* true and shuffled ortholog pairs

verification of the orthology database ([24]), we note that it contains the correct versions of the misannotated orthologs reported in [13].

The lowest 5% orthologs by RD_C are enriched in regulatory and metabolic processes ([4]), as seen in the following table. The set of orthologs showing high regulatory conservation is also enriched for essential genes ($p < 0.006$) ([10]). The highly conserved ribosomal genes were excluded from this analysis. This largely matches our expectations and similar work related to gene family size ([25]).

Gene ontology annotation	Corrected p-value
regulation of cellular process	<0.003
metabolic process	<0.003
regulation of biological process	<0.006

The set of orthologs showing high regulatory divergence (highest 15% by RD_C) is enriched for terms including response to chemical and abiotic stimulus and catabolic processes (see the following table). The divergent genes are also significantly depleted for essential genes (p <5e-4).

Gene ontology annotation	Corrected p-value
cellular process	<7e-5
response to stimulus	<5e-4
organic acid metabolic process	<0.003
catabolic process	<0.01

Extending Regulatory Evolution to Phylogenies

We can extend the regulatory divergence metric to phylogenies, here using a very simple (though incorrect) heuristic. We compute the regulatory divergence over a phylogeny as the pairwise RD_C scores between all extant species. This requires knowing orthology relationships between the species and can use only one-to-one orthology relationships. This heuristic counts certain branches of the tree multiple times and is not informed by the actual phylogeny (branch lengths), but is simply an approximation that can be decomposed into RD_C scores. We calculated this extended RD_C for four *sensu stricto* yeasts, *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, and *S. kudriavzevii.* Figure 3 shows these quantities plotted against the expression divergence calculated between these four species as measured in a variety of stress-related microarray experiments ([19]). There is no detectable correlation between regulatory and expression divergence in this dataset ($r = 0.006$), and sets of genes in various cases (high regulatory and expression divergence, low regulatory and expression divergence) did not have significant gene annotation enrichments. This could be due to the noisy process of motif identification and heuristic nature of the all-pairs RD_C, but it could also be due to inherent difficulties in predicting expression level from regulatory sequences. As comparison, the authors in [27] found only a very weak correlation between shared motifs and expression correlation when comparing paralogs in *S. cerevisiae.* To eliminate the dependence on the all-pairs RC_D, the same analysis could be applied to paralogs in a single species (as in [27]) or to other cross-species expression datasets ([15] or [7]). Alternately, the current dataset could be split into pairwise expression divergences and compared separately.

Discussion

This analysis has defined a new procedure for calculating regulatory divergence between gene pairs. It extends previous work by using a large library of known motifs and utilizing a tool capable of modeling weak and cooperative binding, rather than exact sequence similarity. Sets of genes were identified that have experienced both high and low regulatory evolution between *S. cerevisiae* and *S. bayanus*, along with functional annotations and other characteristics. The regulatory divergence was extended in an approximate fashion to multiple species and compared to an existing expression divergence dataset, without clear results. Further analysis and more comprehensive datasets are needed to explore this relationship.

This project made many simplifying assumptions, based on both dataset availability and modeling tractability. These assumptions, combined with the biological complexity of the problem, led to weak (noisy) measurements that only partially recovered functional information. Future directions include a focus on validation, both experimental (existing datasets as well as possible experiments) and computational. Working at the single-gene level with a large motif library limited

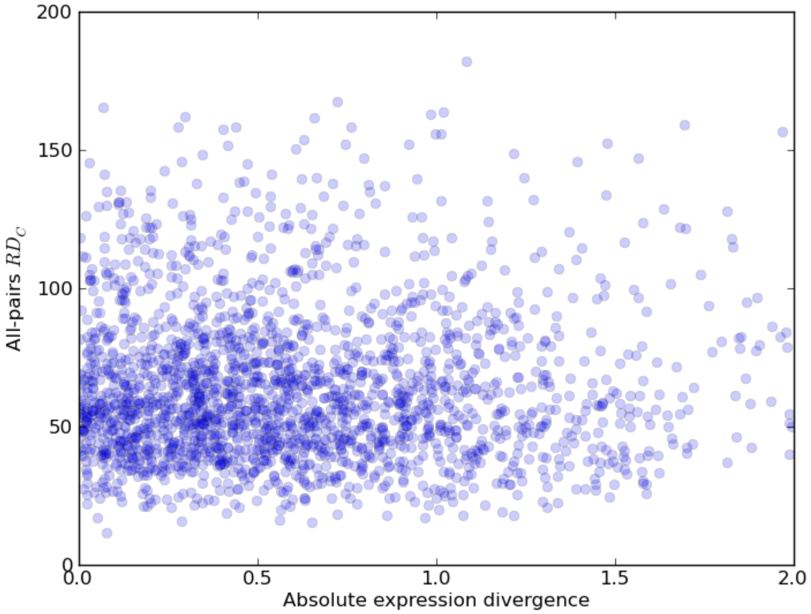


Figure 3: All-pairs RD_C and expression divergence ([19]) for 2141 orthogroups in four yeast species

the resolution of the analysis: starting with large expression datasets across species could allow *de novo* motif discovery, eliminating the bias towards *S. cerevisiae* transcription factors. Additionally, working with transcriptional modules would increase the signal of the informative features over the background noise. Within the existing framework, a more sophisticated binding probability estimator could improve results. A recent work has proposed a model that includes interactions between factors as well as nucleosomes ([26]). Nucleosome organization has been shown to have a strong effect of expression divergence ([21] and [11]). On the analysis side, comparison of the regulatory divergence between orthologs and paralogs could occur, as well as the comparison of regulatory divergence to functional similarity. This method could be extended to more fungi and perhaps used to identify regulatory differences separating pathogenic and nonpathogenic yeast (extending the analysis of [6]). Finally, examining the regulatory divergence of transcription factor genes could provide insight into the whole-network dynamics of regulatory evolution, rather than the view from a single gene. Certain factors may be more volatile, in terms of their ability to tolerate the addition and deletion of edges, than other factors.

The project largely follows the data generation and processing I laid out in my proposal, but some of the analyses were cut short both because of time and the quality of the actual predictions. For example, I did not observe a correlation between expression and regulatory divergence, or any enrichment in any interesting cases. This limited my ability to find illustrative examples or pursue informative cases further.

My reviewers were positive and had several insightful comments and suggestions. The reviews exceeded my expectations; they all seemed to have carefully read and thought about the proposed ideas. Two of them caught details I had intentionally left out because I was uncertain on the methods and urged me to clarify and expand my plans. I enjoyed reading other proposals and I felt like it added to the experience of the class. Also, it at least helped me in writing the initial report to know that various peers were going to read and comment on it. If anything, I thought the reviewers were (like me) too optimistic. They might be given more explicit license to be skeptical and push back on vague claims, making sure that each student has thought out precisely what they should be able to achieve. The most appropriate (at this point) feedback I remember is on the scope of the project, compared to the given timescale. In retrospect I underestimated how long the later parts of analysis, results compilation, and searching for compelling examples would take.

In that vein, if I were to start this project over, I would not necessarily change the scope but would make sure it was more focused. In my proposal I implicitly hoped that decent methods would lead to obvious (and compelling) results, but in retrospect this is an unlikely scenario. I would have concentrated very closely on validation; planning analyses that can be checked against alternate data sources or other gold standards for verification. In this work I did not fully think about what the best case to plan towards was, and ended up with predictions that I had no completely compelling way to test. This was probably the most frustrating aspect for me as a computer scientist, that a perfect test dataset or validation oracle was not available (though in biology it rarely is) and that many of my predictions could only be compared against various random models. The most rewarding aspect was learning about a new problem, reading many papers in the area, and starting to recognize recurring names and labs. As far as time management, I would have focused on getting more of the writeup done earlier while development was occurring (even though this was the plan, it did not occur that way). Finally, in terms of aesthetics and personal preferences, I would have tried harder to come up with a more modeling or algorithmic project, rather than an almost purely analytical project.

Bibliography

[1] Badis et al. 2008. "A library of yeast transcription factor motifs reveals a widespread function for rsc3 in targeting nucleosome exclusion at promoters." Molecular Cell.

[2] Barrett et al. 2005. "NCBI GEO: Mining millions of expression profiles- database and tools." Nucleic Acids Research.

[3] Borneman et al. 2007. "Divergence of transcription factor binding sites across related yeast species." Science.

[4] Boyle et al. 2004. "GO::TermFinder- open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics.

[5] Bryne et al. 2008. "JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update." Nucleic Acids Research.

[6] Butler et al. 2009. "Evolution of pathogenicity and sexual reproduction in eight *Candida* genomes." Nature.

[7] Caudy et al. 2009. "Functional annotation of *S. bayanus* using a gene expression compendium." Submitted.

[8] Cliften et al. 2003. "Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting." Science.

[9] Doniger and Fay. 2007. "Frequent gain and loss of functional transcription factor binding sties." PLoS Computational Biology.

[10] Giaever et al. 2002. "Functional profiling of the *Saccharomyces cerevisiae* genome." Nature.

[11] Field et al. 2009. "Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization." Nature Genetics.

[12] Granek and Clarke. 2005. "Explicit equilibrium modeling of transcription-factor binding and gene regulation." Genome Biology.

[13] Guan et al. 2009. "Evolution of expression between two yeast species." Submitted.

[14] Harbison et al. 2004. "Transcriptional regulatory code of a eukaryotic genome." Nature.

[15] Ihmels et al. 2005. "Rewiring of the yeast transcriptional network through the evolution of motif usage." Science.

[16] Kellis et al. 2003. "Sequencing and comparison of yeast species to identify genes and regulatory elements." Nature.

[17] Rasmussen and Kellis. 2007. "Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes." Genome Research.

[18] Tanay, Regev, and Shamir. 2005. "Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast." Proceedings of the National Academy of Sciences.

[19] Tirosh et al. 2006. "A genetic signature of interspecies variations in gene expression." Nature Genetics.

[20] Tirosh et al. 2007. "Comparative analysis indicates regulatory neofunctionalization of yeast duplicates." Genome Biology.

[21] Tirosh et al. 2008. "On the relation between promoter divergence and gene expression evolution." Molecular Systems Biology.

[22] Tompa et al. 2005. "Assessing computational tools for the discovery of transcription factor binding sites." Nature Biotechnology.

[23] Tsong et al. 2006. "Evolution of alternative transcriptional circuits with identical logic." Nature.

[24] Wall, Fraser, and Hirsh. 2003. "Detecting putative orthologs." Bioinformatics.

[25] Wapinski, Pfeffer, Friedman, and Regev. 2007. "Natural history and evolutionary principles of gene duplication in fungi." Nature.

[26] Wasson and Hartemink. 2009. "An ensemble model of competitive multi-factor binding of the genome." Genome Research.

[27] Zhang et al. 2004. "How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution?" Trends in Genetics.

[28] Zhu et al. 2009. "High-resolution DNA binding specificity analysis of yeast transcription factors." Genome Research.