Published in final edited form as:

Methods Mol Biol. 2016 ; 1361: 375–389. doi:10.1007/978-1-4939-3079-1_21.

Reconstruction and analysis of the evolution of modular transcriptional regulatory programs using Arboretum

Sara A. Knaack¹, Dawn A. Thompson², and Sushmita Roy^{1,3}

¹Wisconsin Institute for Discovery, University of Wisconsin at Madison, Madison, WI, USA

²Broad Institute of MIT and Harvard, Cambridge, MA, USA

³Department of Biostatistics and Medical Informatics, University of Wisconsin at Madison, Madison, WI, USA

Summary

Comparative functional genomics aims to measure and compare genome-wide functional data such as transcriptomes, proteomes and epigenomes across multiple species to study the conservation and divergence patterns of such quantitative measurements. However, computational methods to systematically compare these quantitative genomic profiles across multiple species are in their infancy. We developed Arboretum, a novel algorithm to identify modules of co-expressed genes and trace their evolutionary history across multiple species from a complex phylogeny. To interpret the results from Arboretum we developed several measures to examine the extent of conservation and divergence in modules and their relationship to species life-style, *cis*-regulatory elements and gene duplication. We applied Arboretum to study the evolution of modular transcriptional regulatory programs controlling transcriptional response to different environmental stresses in the yeast Ascomycota phylogeny. We found that modules of similar patterns of expression captured the transcriptional responses to different stresses across species, however, the genes exhibiting these patterns were not the same. Divergence in module membership was associated with changes in lifestyle and specific clades and that gene duplication was a major factor contributing to the divergence of module membership.

Keywords

Arboretum; Gaussian mixture model (GMM); phylogeny; gene tree; multi-clustering; evolution; transcriptional modules; regulatory networks

1. Introduction

Comparative functional genomics aims to compare functional data such as transcript, protein, metabolite, transcription factor occupancy and histone modification levels across multiple species. With advances in genomics such studies are becoming routine in unicellular (1–5) and multicellular organisms (6–10). However, computational approaches to compare these data across species are not well developed. For example, an important

question in comparative studies of transcriptional regulation is to examine the extent of conservation and divergence of transcriptional modules, defined as sets of genes that are co-expressed across multiple conditions. Such modules represent genes that are often co-regulated and constitute entire or parts of a pathway. However, existing approaches to identify and compare modules across species have been limited to two (11) or three species (12, 13). The challenge with comparing multi-species data from a complex phylogeny with large number of species is that often there is no one-to-one mapping of genes between species, because of gene duplication and losses. Such relationships are most naturally represented by a “gene tree” with leaf nodes representing genes in extant species, and internal nodes representing ancestral species. To address that challenge we developed Arboretum, a novel computational algorithm to systematically identify modules of co-expressed genes across species and study their evolution.

Two unique aspects of Arboretum are (1) the ability to reconstruct the hidden ancestral module membership, and (2) explicitly model gene trees that are not necessarily identical to the species tree because of gene duplication and loss events. The ability to handle gene duplication events, a major mechanism by which networks can rewire (14), allows us to work with large phylogenies with complex many-to-many relationships. The inferred ancestral module assignments can be used to trace the evolution of module memberships. In addition to the algorithm itself, we developed several metrics to examine evolutionary conservation and divergence of sets of genes across species. We applied Arboretum to study patterns of conservation and divergence in yeast Ascomycota species under different stress conditions (15, 16). We found that although the general patterns of expression were conserved across the different species, the genes exhibiting these changes were the same. In particular, we found several cases of genes that were orthologous in sequence but exhibited change in expression level. We found that gene duplication played a major role in this divergence of module membership. In this chapter, we describe the steps of applying Arboretum and downstream analyses to study the evolution of modular regulatory programs to new phylogenies.

2. Materials

Arboretum is a multi-species clustering approach and is based on a generative probabilistic model that simultaneously infers the modules across species, while exploiting the relatedness of species encoded in the tree topology. Arboretum is implemented in C++ and is publicly available as open source code at (<http://pages.discovery.wisc.edu/~sroy/arboretum/>). To compile and run Arboretum, the three requirements are a C++ compiler, adequate processor memory (<1GB) and disk space (a standard desktop machine is sufficient), and one third-party library called the GNU scientific library to be installed prior to compiling. The following paragraphs describe the key steps in installing and applying the Arboretum approach to study evolutionary dynamics of transcriptional modules using an example dataset of five yeast species: *S. cerevisiae*, *C. glabrata*, *K. lactis*, *K. waltii*, and *C. albicans*. The expression data used here measure the transcriptional response of gene expression to a heat shock stress and were originally studied in the Arboretum publication (16).

3. Methods

3.1. Downloading and installation of Arboretum

The box below shows the set of commands that are needed to have a running version of Arboretum. The steps are described for a unix command line terminal. 1. Get the source code using the `wget` command. 2. Extract Arboretum code. 3. Change to the directory using `cd` and compile the code using `make`.

```
1. wget
```

```
http://pages.discovery.wisc.edu/~sroy/arboretum/gmm\_crossspeciescluster.tgz
```

```
2. tar -xvfz gmm_crossspeciescluster.tgz
```

```
3. cd gmm_crossspeciescluster
```

```
4. make
```

```
5. cd ../
```

3.2. Applying Arboretum to cross-species data

The inputs to running Arboretum requires us to specify: (a) expression data and input module assignments, (b) species and gene trees, (c) location of results, (d) base species, (e) initialization of the transition probabilities. Example inputs are available from the website which can be obtained from the Arboretum website using:

```
wget
```

```
http://pages.discovery.wisc.edu/~sroy/arboretum/example\_inputs.tgz
```

```
tar xvfz example_inputs
```

Once all inputs and configuration files have been organized, an example usage of Arboretum is as below:

```
../gmm_crossspeciescluster/incAncClust specorder_allclade.txt  
OGid_members.txt 5 species_prob_heat8spec.txt cluster_conf.txt rand  
result_dir learn Scer uniform 0.8
```

We now describe each of the arguments to the Arboretum program in detail below. The first three arguments, `specorder_allclade.txt`, `OGid_members.txt` and `species_prob_heat8spec.txt` constitute the phylogenetic information, including the species and gene tree relationships, specified to Arboretum. Here the `specorder_allclade.txt` and `OGid_members.txt` files present the gene orthology information. The `OGid_members.txt` file specifies the sets of gene orthology relationships

for the genes of each species. An example of the OGID file is available in `example_inputs`. Each OG is denoted by an ID, notated in the form `OG#_#`, which represents a gene tree with the first number in the ID, and the duplication level of the gene tree as the second number in the ID. In `OGid_members.txt` each line in the file, each line is associated with an orthogroup of genes, and begins with an OGID string, and is followed with the string of gene names from the respective species, each separated by a comma. The order of the genes from the species is in the order of the species as listed in `specorder_allclade.txt`. Where a specific species does not have a gene represented in a particular orthogroup “NONE” will appear instead of a gene name. Example lines from these files is as follows:

`specorder_allclade.txt:`

```
Scer
Cgla
Kwal
Klac
Calb
```

`OGid_members.txt:`

```
OG16_1  YLL043W,CAGL0C03267g,Kwal55.20572,KLLA0E0055Og,NONE
OG16_2  YLL043W,CAGL0C03267g,Kwal55.20572,KLLA0E0055Og,NONE
OG16_3  YLL043W,CAGL0C03267g,Kwal33.15269,NONE,NONE
OG16_4  YLL043W,CAGL0C03267g,Kwal33.15269,NONE,NONE
OG16_5  YFL054C,CAGL0E03894g,Kwal55.20572,KLLA0E0055Og,NONE
```

The species tree is used to define the phylogenetic relationships of the species and is specified by `species_prob_heat8.txt` file, which has the following format:

`species_prob_heat8spec.txt:`
#Child LeftorRight Parent

```
Seer    left    Anc4
Cgla    right   Anc4
Anc4     left    Anc9
Anc8     right   Anc9
Kwal     left    Anc8
Klac     right   Anc8
Anc9     left    And 1
Calb     right   And 1
```

The gene trees are expected to be directory named `data/TREES` relative to the working directory of the issued Arboretum command. If a species and gene trees are not known, we recommend using resources where such data are available, e.g., in Paranoid (17) or Ensemble Compara (<http://useast.ensembl.org/info/genome/compara/index.html>) (see **Note 1**).

The next argument specifies the expression data and initial set of cluster assignments in the configuration file `cluster_conf.txt`, which is exemplified below:

```
cat cluster_conf.txt:
Seer      heat8/Scer_clusterassign.txt    heat8/Scer_expr.geneexp
Calb      heat8/Calb_clusterassign.txt     heat8/Calb_expr.geneexp
Cgla      heat8/Cgla_clusterassign.txt     heat8/Cgla_expr.geneexp
Kwal      heat8/Kwal_clusterassign.txt     heat8/Kwal_expr.geneexp
Klac      heat8/Klac_clusterassign.txt     heat8/Klac_expr.geneexp
```

The first column has the name of the species, the second the name of the input cluster assignment file for the species genes, and lastly the name of the expression data file for the expression measurements used for that species. The expression data themselves may need to be pre-processed as described in **Note 2**.

The input cluster assignments of individual species are needed to seed the Arboretum algorithm. These cluster assignments themselves are typically obtained by merging all the data across species into a single file and applying a k -means or a Gaussian mixture model algorithm (18) to cluster the merged data. In this case each row of the merged data represents the data values for those genes in a single ortholog set of genes across the represented species. The input cluster assignments for the ortholog set are used to assign initial cluster assignments to each of ortholog genes in each species. We emphasize that this is only an initial assignment for seeding Arboretum, and we do not advise interpreting such results as being more significant than that. This is because merging the data across species also requires us to deal with many-to-many mappings and in a flat merged clustering approach such complex mappings result in missing values (gene loss) or duplicated data (gene duplication) in the data rows.

The fifth argument is k , the number of modules. This is specified as user input. In **Note 3** we discuss possible ways to determine this argument. The sixth argument specifies how to initialize the initial set of module parameters and can be either “`rand`” or “`none`”. This option will tell the program to either use the input clustering assignments of genes only to define cluster sizes (`rand`), or to use the input cluster assignments verbatim for initialization (`none`). For “`rand`” Arboretum randomly partitions the data using these size distributions. In **Note 4**, we discuss inspecting results for different random initializations of the Arboretum algorithm using this argument.

The seventh argument is the `result_dir`, a name for the output directory for the results.

The eighth argument defines the mode of the algorithm. This option will generally be set to “`learn`”, which will run the clustering algorithm on the input expression data, or “`generate`” to invoke the generative model to produce data for simulation studies. The ninth argument is the name of a well-annotated base species (`Scer` in the example usage). This species is chosen to organize the output of Arboretum, especially the module assignments for the ancestral nodes.

The remaining two arguments relate more specifically to the functioning of the Arboretum multi-clustering algorithm. The 10th argument determines the mode of initializing the transition matrices for module membership of genes across species. These matrices will be of the size $k \times k$, where k is the number of modules specified in the analysis. There is one transition matrix for each non-root node in the tree and specifies the probability of a gene to maintain (diagonal entry) or switch its module membership (off-diagonal). There are two initialization methods for these matrices. If `uniform` is selected, the eleventh argument is a real-valued number and used to initialize the diagonal elements of the transition matrices for all species. If “`branch length`” is used, the value of the eleventh is interpreted as a file name with branch lengths and is used to initialize the transition matrices for each individual node. Typical settings for these two arguments are “`uniform 0.8`”.

3.3. Arboretum outputs and their interpretation

Once Arboretum has finished, outputs will be presented in the `result_dir` directory. These results include (1) inferred transition matrices, (2) inferred module assignments, for both ancestral and extant species.

There are several analyses that we perform to interpret the modules:

1. Heat maps to visualize patterns of expression in each module (Figure 1A); 2. Global module conservation between extant species (Figure 1B); 3. Patterns of divergence across a phylogeny; 4. Gene Ontology enrichment of modules; (5) Motifs of *cis*-regulatory elements, 6. Computation of various metrics to quantify the extent of conservation and divergence in modules. We will present several of the most significant examples here.

Figure 1A presents the five modules inferred using Arboretum for the five-species example data set presented here. From these plots we see the pattern of expression of each module is highly similar for all five species. An important aspect of interpretation of plots like this is to assess the similarity of gene expression within each module, across each species, for example, are genes in module 1 of *S. cerevisiae* similar to genes in module 1 of *K. waltii*, noting that in both species, the expression patterns of the modules are quite conserved.

This brings us to a key computational measure developed for measuring the similarity of gene content between modules across species. For each pair of species we estimate the significance of gene orthology relationships between each pair of modules within those two species (Figure 1B). These results are generated by using a Hyper-geometric test to assess the significance of overlap between all pairs of modules for these two species. For each pair of modules, one module from species A, and another from species B, we obtain a p-value for the significance of the similarity of the module gene groups. We use the negative log of that p-value as a score for the similarity of those modules. Here in Figure 1B, for each species pair among our species we have a 5 by 5 matrix of Hyper-geometric test p-value scores, represented here in heatmaps. The diagonal matrix elements, in red, illustrate that the set of modules are well conserved across species within each module. The off diagonal elements, in blue, show significant divergence relationships, where genes in one species have a given module membership, but their orthologs in another species are in a different module. Figure

2 shows one such example of a set of ortholog gene groups, spanning all five species that diverge in their expression and module membership. These plots demonstrate a coherent change in expression for these ortholog gene groups from an induced state (red) in *S. cerevisiae* and *C. glabrata*, to a repressed state (green) in the three remaining species. This observation represents a phylogenetically coherent change in expression, and as such these sets of orthologous genes represent an interesting example of changes in gene regulation in evolution. Possible explanations for such changes include change in transcription factor binding in the different species, and DNA sequence features such as *cis*-regulatory elements, and other genomic markers and can subsequently be examined to substantiate such observations to provide insight to the mechanisms behind such changes in gene expression.

To assess biological meaning of our modules, we perform enrichment analysis of the modules. We use Gene ontology (19) terms to assess the significance of overlap of genes in each module with known annotated categories. For the base species, we usually have the GO annotation available. For the non-base species including the ancestral species we use the ontology terms mapped to the orthologs of the base species. We use the Hyper-geometric test with the Benjamini Hotchberg procedure for multiple hypothesis correction. Once GO enrichments for all species have been completed, we examine the patterns of enrichment across the species per module as well as per term. Figure 3 shows the GO enrichments for module 1 and it can be seen that there is significant conservation of processes as well. This presentation highlights cases where specific biological processes are associated with a specific module across all species (e.g. ncRNA processing). Such patterns demonstrate evolutionary conservation of biological function for these modules, which Figure 1 shows to be highly conserved for co-expression and gene orthology. Figure 3 also shows examples of terms that are associated with a module in all species, but not necessarily the same module (e.g. chromatin organization), and this gives examples of processes that have likely diverged in their regulation between species. Similarly, we assess the modules for motif enrichments. For each species, we used a motif collection generated by the Cladeoscape algorithm (20). Once motifs have been identified across each module for each species, we visualize them in a similar way as the GO enrichment results.

In addition to assessing the orthology overlap significance, we have developed several measures to quantify extent of conservation and divergence between the modules. These include: **1.** Ancestral module conservation index (AMCI), **2.** Module expansion index (MEI), **3.** Module contraction index (MCI), **4.** Gene module divergence index (DI). All these indices are directly computed from the outputs of Arboretum. The AMCI metric uses the inferred transition matrix (one of the results of Arboretum), to assess the tendency of a species to maintain or switch its module assignment from its ancestor. This is a number between 0 and 1, and the closer it is to 1, the more likely is a species to maintain its ancestral module assignments, and the closer it is to 0 the more likely are modules going to diverge from its ancestral version.

The MCI and MEI metrics were designed to assess how modules diverge along a phylogenetic tree. The MCI metric measures the module contraction tendency, which happens when a module in a species loses genes compared to its ancestral version. The MEI metric measures the tendency of a module to gain genes in a child species compared to this

ancestral version. Finally, the gene module divergence index is a measure to assess the extent to which a gene switches its module assignments, and is defined as the number of switches divided by the total number of species the gene is present. Table 1 presents the calculated AMCI scores, the MCI scores for the most repressed and most induced expression modules (1 and 5, respectively), and the MEI scores for the same modules for this example analysis. Such results are given for each of the five extant species, and the (internal) ancestral species nodes (excluding the last common ancestor).

Acknowledgments

SR is supported in part by a NSF ABI CAREER award (DBI-1350677). SK is supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM5T15LM007359). This work was also supported by NIH grant 2R01CA119176-01 and a SPARC grant from the Broad Institute.

4. Notes

1. Generating species and gene trees if not available

Often for a new phylogeny a species tree and gene trees may not be available. Such data can in general be obtained from public resources such as from ENSEMBL release 75 gene trees from ENSEMBL Compara resource. Our own gene trees were created using the SEMPHY (21) algorithm and the species tree were relearned for some subset of species by first doing a multiple sequence alignment of genes that are present in all species using MUSCLE (22) and learning trees using PAML (23).

2. Expression data pre-processing

The chief requirement for the input expression data is that the values can be modeled with Gaussian (normal) distributions. Typically the data used in Arboretum is gene expression data from a microarray or RNA-seq experiments. As microarray data is typically a comparative log ratio value of measurements from an experimental sample and a reference or background standard, such data is typically usable as is. In some cases, there might be missing values in the microarray. We recommend throwing away genes with >50% missing data. For the remaining genes with <50% missing, values can be interpolated using the mean calculated from the non-missing samples of that gene. For RNA-seq data, per-gene fragments per kilobase of transcript per million mapped reads (FPKM) expression values can be first Laplace-transformed, and then log-two-transformed. These expression data would finally then be row zero-meaned for Arboretum (rows corresponding to genes). To be explicit, the data for each species in the phylogeny being studied will have an independent input expression data file. When needed it is suggested to apply a row zero-mean transformation to data values across multiple measurements of a single species to facilitate the ability of the algorithm to model the variation in the data.

3. Determining the number of modules

An important question in applying Arboretum is to specify the number of modules, k . We do this in two ways: (a) use penalized maximum likelihood scores where the penalty is proportional to the number of parameters, which grows with the number of modules, (b) Use

enrichment of GO and motifs to see if the clusters represent meaningful partitions. Manual inspection of the patterns is also helpful to determining the number of modules. Figure 4 shows the average penalized log likelihood plot for a range of module number settings, k , and for multiple randomized input cluster assignment. An un-penalized and penalized log likelihood score is generated by each instance of an Arboretum run. The un-penalized log likelihood is log of the inferred probability of the algorithmic model to describe the final cluster assignments of a clustering run. The penalized log likelihood corrects for the complexity of the model for the number of genes and species, and modules involved. We see that the penalized log likelihood suggests that a $k > 12$ is not going to be useful to describe the data. The attentive reader may wonder why the results in Figure 1 and Figure 2 are not for $k=12$ but $k=5$. We use this as an upper bound and use a combination of enriched terms and manual inspection to determine the final number of modules. The final determination for the optimal number of modules is subject to manual inspection of the expression patterns in the inferred modules, the ortholog coherence of modules across species, and finally the significance (including number and uniqueness) of GO enrichment terms for the results generated with varying values of k , hence $k=5$ was chosen in our example.

4. Stability of modules

Arboretum, like many clustering algorithms is prone to local minima, which might result in a sub-optimal solution. We recommend running Arboretum multiple times on the same data and inspecting the expression coherence of the modules to assess the reproducibility of the results and that they do not represent any aberrant patterns. This can be done by running Arboretum multiple times each time keeping the sixth argument set to “rand”, or setting it to “none” but using a different initial module assignment set each time. In practice we have seen that Arboretum produces very stable results.

References

1. Jensen LJ, Jensen TS, de Lichtenberg U, Brunak S, Bork P. Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*. 2006; 443:594–597. [PubMed: 17006448]
2. Gasch AP. Comparative genomics of the environmental stress response in ascomycete fungi. *Yeast* (Chichester, England). 2007; 24:961–976.
3. Wohlbach DJ, Thompson DAA, Gasch AP, Regev A. From elements to modules: regulatory evolution in Ascomycota fungi. *Current Opinion in Genetics & Development*. 2009; 19:571–578. [PubMed: 19879128]
4. Romero IG, Ruvinsky I, Gilad Y. Comparative studies of gene expression and the evolution of gene regulation. *Nat Rev Genet*. 2012; 13:505–516. [PubMed: 22705669]
5. Thompson DAA, Regev A. Fungal regulatory evolution: cis and trans in the balance. *FEBS letters*. 2009; 583:3959–3965. [PubMed: 19914250]
6. Brawand D, et al. The evolution of gene expression levels in mammalian organs. *Nature*. 2011; 478:343–348. [PubMed: 22012392]
7. Schmidt D, et al. Five-Vertebrate ChIP-seq Reveals the Evolutionary Dynamics of Transcription Factor Binding. *Science*. 2010; 328:1036–1040. [PubMed: 20378774]
8. Xiao S, et al. Comparative Epigenomic Annotation of Regulatory DNA. *Molecular Cell*. 2012; 149:1381–1392.
9. Barbosa-Morais NL, et al. The Evolutionary Landscape of Alternative Splicing in Vertebrate Species. *Science*. 2012; 338:1587–1593. [PubMed: 23258890]

10. Merkin J, Russell C, Chen P, Burge CB. Evolutionary Dynamics of Gene and Isoform Regulation in Mammalian Tissues. *Science*. 2012; 338:1593–1599. [PubMed: 23258891]
11. Tanay A, Regev A, Shamir R. Conservation and evolvability in regulatory networks: the evolution of ribosomal regulation in yeast. *Proc Natl Acad Sci U S A*. 2005; 102:7203–7208. [PubMed: 15883364]
12. Waltman P, et al. Multi-species integrative biclustering. *Genome Biol*. 2010; 11:R96+. [PubMed: 20920250]
13. Kuo D, et al. Evolutionary divergence in the fungal response to fluconazole revealed by soft clustering. *Genome Biol*. 2010; 11:R77. [PubMed: 20653936]
14. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic genetic switch. *Nature*. 2007; 449:677–681. [PubMed: 17928853]
15. Thompson DA, et al. Evolutionary principles of modular gene regulation in yeasts. *eLife*. 2013; 2doi: 10.7554/eLife.00603
16. Roy S, et al. Arboretum: Reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res*. 2013; doi: 10.1101/gr.146233.112
17. O'Brien KP, Remm M, Sonnhammer ELL. Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research*. 2005; 33:D476–80. [PubMed: 15608241]
18. Hastie, T., Tibshirani, R., Friedman, JH. *The Elements of Statistical Learning*. Springer; 2003. Corrected
19. Ashburner M, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000; 25:25–29. [PubMed: 10802651]
20. Habib N, Wapinski I, Margalit H, Regev A, Friedman N. A functional selection model explains evolutionary robustness despite plasticity in regulatory networks. *Mol Syst Biol*. 2012; 8
21. Wapinski I, Pfeffer A, Friedman N, Regev A. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics*. 2007; 23:i549–58. [PubMed: 17646342]
22. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*. 2004; 5:113. [PubMed: 15318951]
23. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007; 24:1586–1591. [PubMed: 17483113]

B. Shown for each species pair are the 5 by 5 matrices of negative logarithms of Hypergeometric test p-value results. The diagonal elements, shown on a 0–40 color scale, in each matrix represent the comparison of the gene sets in different species with the same expression pattern (e.g. Module 1 of *S. cerevisiae* with Module 1 of *C. glabrata*). When the modules are highly conserved for gene content (or coherent) these matrices are highly diagonal. The off-diagonal entries of the matrix show the significance of overlap of modules of different expression patterns. These are generally lower than the diagonal entries and are shown on a scale of 0–5.

Figure 2. Example of divergence of orthologous genes
Shown is a set of orthologous genes using *S. cerevisiae* systematic gene names that exhibit coordinated change in expression along the phylogeny. The left-most panel of shows the Arboretum-inferred module membership for each genes in the five extant species, and the inferred membership for the ancestral species (Anc4, Anc8 and Anc11). Note that these memberships are generally for module 4 for *S. cerevisiae* and *C. glabrata*, and module 2 for the related ortholog genes in *K. waltii*, *K. lactis* and *C. albicans*. The expression levels of these genes in each of the species are shown in the next set of five panels.

Figure 3. Gene Ontology (GO) process enrichment of Arboretum modules
Shown is a summary of GO enrichments results from the inferred Arboretum modules in our five species example. Specifically, each module in each species is tested for association with the curated set of GO enrichment terms. The enrichment results shown here have been selected using two criteria: **1.** a q-value threshold of <0.001 for each term and species-module association indicated, **2.** the number of genes associated with each specific term within each individual species module, be at least a 0.2 fraction of the number of genes in that module. The terms on the left represent examples of these high confidence term-module associations that are conserved in the same module (denoted by the color of the box) across all five species. The terms on the right are cases where a given biological process shows significant association with at-most one module in each species, but a different module in each species. These are cases where orthologous genes associated with a given process have diverged in their module membership.

Figure 4. Using penalized log likelihood scores to determine the upper bound number of modules in an Arboretum analysis

This figure shows the penalized likelihood scores determined from results for the same example data used here for $k=5$ to 15. Each data point in these plots represents the average scores for each setting of k , where the error bar represents the variation across 5 multiple random initializations of the input module assignments. The values for corrected log likelihood scores turn over at around $k=12$, which is indicated as the maximum number of modules to consider for this given input set of data.

Table 1
Calculated metrics for examining Arboretum module dynamics

This table highlights three measures: AMCI: the ancestral module conservation index, MCI: module conservation index, and MEI: module expression index. Calculated AMCI values of each non-root species node are reported in the right most column of this table. The variation of this measure across the species of this example analysis is from 0.63 for *C. glabrata* to 0.93 for Anc8. MCI is the propensity for the same ancestral module to have shrunk. This measure is calculated for each individual species module, with results provided in columns 2 and 3 for the most repressed (module 1) and induced (module 5) modules respectively. MEI for a module measures the degree to which genes join that module from a different module in the immediate ancestor. Here the trend for the left most two data columns show the MEI results for the example analysis for the same two modules (1 and 5), which show widely varying degrees of expansion across species.

Species	AMCI	MCI in Repressed Module	MCI in Induced Module	MEI in Repressed Module	MEI in Induced Module
<i>S. cerevisiae</i>	0.77	0.14	0.10	0.18	0.22
<i>C. glabrata</i>	0.63	0.44	0.36	0.52	0.48
<i>Anc4</i>	0.74	0.22	0.11	0.31	0.35
<i>K. waltii</i>	0.68	0.31	0.15	0.43	0.52
<i>K. lactis</i>	0.71	0.30	0.12	0.36	0.58
<i>Anc8</i>	0.97	0.00	0.01	0.15	0.00
<i>Anc9</i>	0.77	0.15	0.28	0.16	0.05
<i>C. albicans</i>	0.64	0.33	0.24	0.42	0.47