Computational methods for transcriptional regulation

Eric D Siggia

How is the information from a thousand gene-expression arrays, the location of more than two hundred regulatory factors, and nine sequenced genomes to be integrated into a global view of the regulatory network in budding yeast? Computational methods that fit incomplete noisy data provide the outlines of regulatory pathways, but the errors are not quantified. In the fly, embryonic patterning has proved amenable to computational prediction, but only when the DNA-binding preferences of the relevant factors are taken into account. In both these model organisms, simply restricting attention to regulatory sequences that align with related species (i.e. 'conserved') discards much information regarding what is functional.

Addresses
Center for Studies in Physics and Biology, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA
Corresponding author: Siggia, Eric D (siggiae@rockefeller.edu).

Current Opinion in Genetics & Development 2005, **15**:XXX–XXX

This review comes from a themed issue on
Chromosomes and expression mechanisms
Edited by Barbara Meyer and Jonathan Widom

0959-437X/$ – see front matter

Introduction

This review focuses on recent advances in the computation of regulatory interactions in yeast and fly, from multiple data sources. The standards of success shift when the theoretical sciences encroach on biology, from merely getting something right to not getting anything wrong. Success should be measured not by whether the bias on a coin can be detected after enough tosses but, rather, by the accuracy in predicting each event. Thus, my focus is on methods — and where they fail — and on model systems in which constructing transgenics is easier and we are closer to knowing all inputs to a gene. Regulatory information ultimately has to be integrated into models of cellular response. So far, quantitative predictive network models (e.g. for the cell cycle in yeast [1] or pair-rule patterning in fly [2]) have emerged only from a close reading of the literature, from inspiration, or from a focused study on a particular phase of development [3]. The methods reviewed are somewhat sequence-centric, in response to the current emphasis on comparative genomics [4] and the technologies engendered by the human hap-mapping project, which make it possible to rapidly compare the genomes of similar organisms.

Computational strategies

The trend in recent years, most noticeably in yeast, has been the integration of multiple types of data; with no two studies using identical algorithms from start to finish. Genes were first clustered based on ad hoc measurement of similarity in their expression profiles over multiple experiments, and then the clusters were analyzed for common sequence motifs or gene functional classes. Then, better error models were used that parameterized specific steps and/or processes in gene expression analysis [5]; the error in an experiment–gene–probe triplet is a combination of three fitting parameters, each dependent on a single variable. The clustering step was superseded by direct correlation between gene ontologies or sequence and the expression data establishing regulatory interactions (reviewed in [6]).

The available data types can be schematized using coordinate axes that represent the regulatory sequence, RNA expression and genomic location of regulatory proteins. A common strategy for their integration is to use one data type at a time, refine the predictions using an orthogonal type of data, and then iterate. For instance, in their study [7], Bar-Joseph *et al.* first determine putative sets of co-expressed genes on the basis of stringent protein localization data. Subsequently, they reduce this set further on the basis of actual expression data and, finally, include genes that bind the same factors with less stringency — provided that they fit the common expression profile for the cluster. Algorithms using sequence and expression are explored elsewhere [8,9].

The alternative is a model encompassing several properties at once (e.g. sequence and expression [10•,11•,12]). Segal *et al.* [12] write the multivariate probability as the product of conditional probabilities that express (i) the probability that a regulatory protein binds to the sequence of interest, (ii) the probability that a gene belongs to a 'coexpression class' if specific regulatory proteins are bound, (iii) the probability distribution for RNA expression contingent on a 'coexpression class' and experiment. There are clearly multiple ways of formulating and parameterizing the conditional probabilities, and then some iterative improvement has to be done on all the parameters to maximize the total probability, which is only guaranteed to find a local optimum.

To illustrate some of the choices, consider just the subsidiary problem of identifying DNA sequence motifs. One can chose a rich description of the site (e.g. as a matrix of base frequencies at each site) and then optimize by a heuristic search, or one can use a simple pattern (e.g. all 7 base strings with a few degenerate symbols) but then do an exhaustive search that is guaranteed to find the optimum. The probability model can be constructive and intrinsic to the sequence being searched (e.g. by computing the probability that the motif occurs by chance when sampling random bases with a frequency computed from the sequence itself) or, alternatively, it can be a discriminatory model (e.g. the motif is present in most of the regulatory regions of genes involved in a given process and is uncommon elsewhere).

The pervasiveness of pairwise sequence comparisons might lead one to believe that this is a solved problem. But only for a certain class of scoring functions, those amenable to recursive evaluation, can the optimal solution be found; and its computed statistical significance ignores most biological knowledge. Even this restricted regime of confidence does not exist for multiple sequence alignment, which is a prerequisite for exploiting multiple genome sequences. Most comparative projects are organized around a reference species (e.g. human for primates or mammals) for which repeated sequencing and gap closure yields long continuous stretches of genome, whereas only shorter unordered continuous segments ('contigs') are available for the secondary genomes. If pairwise alignment between these contigs and the reference species suffices, then the problem is in principal solved by BLAST; but, in practice, BLAST will introduce spurious breaks into long stretches of similar sequence, along with the real ones caused by recombination and gene duplications.

Certain codes (see [13,14]) reassemble these fragmented alignments. They enumerate compact strong pairwise regions of similarity and then chain them together in an order- and strand-preserving manner in each of the species. This so called syntenous assembly is a very good filter for the correct placement of repetitive regions, which a program such as BLAST would align in many ways. The highest-scoring string of anchors is then selected, with perhaps some weight given to the quality of the global alignments preformed between corresponding anchors. In this way, a reasonable compromise is made between speed and sensitivity. Extensions of the basic method will handle duplications, inversions and translocations [15]. When multiple sequences are presented (the cystic fibrosis transmembrane conductance regulator (CFTR) region of 13 vertebrates has emerged as a test set [16]), the alignment is progressive, with the closest species aligned first. Various tradeoffs have to be made between how the phylogenetic relationships are modeled, how alignment scores between subtrees propagate to the next higher level and how gaps are handled [17]. Once the multiple alignment is performed, a more refined phylogeny score can be computed in a moving window for the aligned bases. In the case of vertebrates, this enabled selection of the ~5% of the sequence thought to be functional [16]. The cited methods are all well implemented, and the choice among them is a matter of personal familiarity and the nuances of the application.

When some alignment is possible between homologous regulatory regions, how is it best to extract protein binding sites? The simplest expedient of simply ignoring nonconserved sites will miss many functional sites (see below), and also will give undue weight to the sequences of evolutionarily close species. Using parsimony (which minimizes the number of changes) might seem reasonable, but it ignores phylogenetic branch length; sequence similarity between more distant organisms is stronger evidence for functionality than is similarity between close ones. A model for molecular evolution subject to the constraint that a fixed protein bind was formulated [18•] along the same lines as algorithms that measure the evolution of coding sequences constrained by the genetic code or codon biases etc. Standard search procedures were then generalized for this new similarity measurement [19]. Another evolutionary model for proteins [20], which explicitly fits the protein selection coefficient (fitness), was applied to DNA binding sites [21], but it then carries the implication that the expression of all genes regulated by the factor are subject to similar selection coefficients, which is not generally true.

Applications

Data Resources

DNA microarrays, which revolutionized genome biology, were quickly applied to genotyping [22], and to mapping the location of regulatory factors, (reviewed in [23]), histone acetylation patterns in yeast [24] and methylation patterns in fly [25]. The wealth of new applications has mitigated against repeating nominally the same experiment in two laboratories. In yeast experiments, in which the most extensive chromatin immunoprecipitation (ChIP) analysis is possible, the targets of the regulatory protein Gal4 compared well with those reported in the literature [reviewed in [23]), but comparison of two regulators of the G_1/S transition [26,27] yielded only a 20–30% overlap of the genes in each experiment judged significant.

An informative perspective on what constitutes meaningful variability in gene expression comes from experiments that compare sporulation between two different strains of budding yeast [28]. In each strain, expression of about 1600 genes varied during sporulation but only 900 were in common (and only 269 of these were detected using spotted arrays [29]). The sources of this variability were partially elucidated [30,31•] using arrays both to genotype and phenotype a wild and laboratory strain and to perform association studies by following all four haploid offspring produced from crossing these strains. In the first study [30], approximately 1500 genes were differentially expressed in vegetative growth, and the mutations responsible for these expression differences could be genetically mapped for 570 of these. Most of the responsible mutations mapped far from the gene whose expression changed, and a few of these were obviously linked to transcription factors.

A survey of variation in gene expression in *Drosophila* [32•,33] found enhanced conservation in regulatory genes and enhanced variation for duplicated ones. More variability in gene expression was attributable to changes in regulatory sequences (*cis*) than to changes in factors (*trans*). A novel approach that directly compares expression from two genomes in a common host (gotten by crossing the species) with the parental expression also demonstrates a preponderance of *cis* over *trans* effects [34•].

Yeast

We now have the means to examine how access to the genomes of multiple species improves regulatory predictions, [35•,36–38]. To date, the analysis first imposes interspecies sequence conservation, then enumerates simple patterns and applies other filters [39]. One should not infer, however, that known protein binding sites are strongly biased towards blocks of sequence that align with other species; they are not (Table 1). Simply imposing strict sequence conservation ignores the fluidity of regulatory sequences [40]. A list of motifs obtained in this way is in fact no more complete than an earlier calculation [41] that used just a single genome but fit all upstream regions simultaneously (Table 2). It is not generally possible to multi-align the sensu-latu species with *Saccharomyces cerivisiae* [37], yet one would suppose that genes such as the cyclins are regulated in the same way. How to best exploit multi-species data and, thus, which species to sequence remain topics for research.

In their study [7], Bar-Joseph *et al.* used ChIP data to computationally extract core sets of factors and the genes they regulate. The gene list was then enlarged by adding those genes with similar expression profiles. Multiple cell growth conditions and transcription factor subcellular localization for a more extensive set of factors were subsequently combined with data on multiple species conservation [42•]. The functional significance of transcription factor localization is still unclear. For instance, the expression of four out of nine factors composing a transcriptional regulatory loop for the cell cycle inferred from factor binding [27] do not themselves vary during the cell cycle [43]. Regulatory networks were inferred from sequence and expression data [11•] and several predictions were verified experimentally. Sequence and expression data were integrated [10•] and, where possible, Boolean interactions were inferred and used to predict expression patterns of test sets of genes for validation. These methods led to the discovery of new regulatory interactions, but we still lack a sense of how the methods compare and what information they miss (but see [44]). From this point of view, studies that concentrate on a single pathway (e.g. the transition from respiration to fermentation [45] or MAPK signaling [46]) are perhaps more satisfying, because they provide a sense of closure and link transcription to a particular pathway.

Fly

The number of sequenced *Drosophila* species will soon exceed those in the *Saccharomyces* clade. (http://flybase.bio.indiana.edu/docs/news/announcements/drosboard/).

A sense of the genomic scale of the genus is provided by *Drosophila pseudoobscura*, which diverged from *Drosophila melanogaster* 30 million years ago, the genome of which is fully sequenced and annotated (http://www.hgsc.bcm.tmc.edu/projects/drosophila/). Even though evolution has almost completely randomized the synonymous codons, homologous genes are easy to locate. Approximately 40% of the nonrepetitive noncoding sequence can be aligned in a dense array of syntenous blocks. Half of the remaining ten *Drosophila* species being sequenced are less than 30 million years diverged from *D. melanogaster*, and were selected for their relevance to molecular evolution and ecology. [47] This dataset will provide a huge impetus for developing robust methods for comparative analysis.

The units of regulatory signal are modules between 100 bp and 1 kb. They can be 10 kb or more from the gene in either direction or be situated in introns and, thus, the first challenge is to locate these sequences. Given a multi-species alignment with homologous blocks every 100 bp or less in the noncoding regions, one can simply count the fraction of conserved bases in a sliding 0.5-1 kb window (perhaps scoring also the granularity [48]). Using as a test set either known intercellular signaling modules (E Emberly, unpublished) or blastoderm patterning modules from the literature [49] or from recent experiments [50•,51], simple sequence comparisons failed to distinguish most of these annotated modules from the surrounding sequence.

This accords with indirect evidence, from molecular evolution studies, that most of the noncoding euchromatic sequence in the fly genome is functional. The evidence is threefold: nonfunctional sequence (e.g. mobile elements and pseudogenes) is rapidly lost [52]; alignments between *D. melanogaster* and *D. pseudoobscura* reveal that the size of ~1 kb units of noncoding sequence is preserved; and three-way alignments including *Drosophila yakuba* demonstrate an excess of insertions over deletions [53]. In other model systems [53], simple sequence conservation is a useful screen for functional modules. In vertebrates, functional modules have been discovered this way [54], as they have in fly [51], but that does not imply that most modules can be found this way.

Greater success in finding regulatory regions has been achieved by searching for clusters of binding sites for the proteins that together define a pathway (reviewed in [55]). Most studies use empirical thresholds for whatever degree of similarity that counts as a binding site match, and then count their number. Our work [56,18•] computes a binding free energy, so strong and weak sites contribute appropriately to the score with no additional parameters, and site overlaps are correctly handled.

Large-scale tests for some of these programs have recently become available. Anterior–posterior patterning in the embryo is a natural test because many of the relevant factors are known. The thousands of nuclei in the embryo provide a rich data source that enables phenomenological models for gene expression to be fit [57•]. Berman *et al.* [58•] tested 28 uncharacterized modules predicted from an earlier genome-wide scan and found that six at least partially recapitulated the expression of a neighboring gene when placed upstream of a reporter. This is more than expected by chance, and scoring for binding site synteny in *D. melongaster* versus that in *D. pseudoobscura*. In their study [50•], Schroeder *et al.* focus on the regulation of ~50 core genes in the segmentation hierarchy (themselves all encoding regulatory or signaling proteins). Thirteen out of sixteen predictions of new modules properly recapitulate the expression of a neighboring gene, and for many of the gap genes the entire expression pattern is accounted for by the enlarged set of modules.

Markstein *et al.* [59•] validated five out of fifteen genome-wide predictions of dorsal-regulated modules from an earlier study [60]. Then, three modules for dorsal-regulated genes exhibiting an expression pattern specific to the neuroectoderm were computationally analyzed for other binding motifs, which together with the dorsal-regulated genes were scanned across the genome [59•]. Two out of four newly identified modules were functional, the target genes being those known to express in the correct region.

All studies agree that a very good predictor for whether the genome-wide modules are functional is simply whether they are adjacent to a gene with the appropriate expression pattern. The incremental utility of interspecies comparisons depends on how the single species predictions are made (e.g. using our technique [18] and just *D. melanogaster* correctly classifies 75% of the 37 module predictions discussed in a later study [58•]). Presenting the second species to this code results in a ~30% improvement as judged against the fly *in situ* hybridization collection [61]. The moderate correlation between known binding sites and interspecies conserved sequence seen in Table 1 for yeast carries over to fly [49]. The number of genes necessary to pattern the blastoderm as defined by screens is much smaller than the number with a blastoderm pattern (~700 if one scales up the *in situ* collection to the entire genome) and it will be very interesting to see if these patterns are similar in *D. pseudoobscura* and are generated by a homologous module. Protein homology modeling suggests that the residues contacting DNA are the same in these two species for the segmentation gene transcription factors (A Morozov, unpublished).

It is rare to have a complete collection of transcription factor binding sites for a pathway, but Grad and coworkers [62•] have devised a strategy to discover coregulated modules within the regulatory sequence of coexpressed genes. Motif searches can also be usefully constrained if only the structural class of the regulatory protein is known [63].

Microarray data are available for both eye [64] and wing development [65] but, to date, have not been integrated with a computational analysis of the regulation. A survey of RNA expression during multiple stages in development has been performed [66•].

Conclusion

Bioinformatics, without much introspection, has transformed the question of biological function to one of probabilities with respect to some model of 'chance' (equated to the functionless). It is by no means obvious that evolution optimizes the information entropy of protein binding sites, as motif discovery algorithms assume. On the larger scale of a module or pathway, there might not be a static fitness function (the proverbial alpine 'landscape') on which evolution climbs forever upward. Rather, evolution might operate on kinetics, selecting networks that are quick to learn from example; the examples being generated by mutation and selection. There is in fact an extensive computer science literature on learning from examples, and this has revealed, not surprisingly, that it is easier to learn a series of conjunctions (A and (not B) and C) than a complicated collection of 'AND' and 'OR' relations. We see many conjunctions in biological circuits but, of course, these are also easier to learn experimentally.

Gene regulation operates when transcription factors 'recognize' their binding sites; our algorithms work by grouping similar sequences together. It is always possible to assign points in sequence-space to a predefined set of motifs by proximity (classification by protein recognition), yet there might not be a sufficient density of points to define the motifs (the task of clustering). The appeal of interspecies comparisons is to do what the organism cannot do; create a larger sample of points from the same distribution to make clustering based on the density of points feasible [67]. The relevant regulatory proteins need to have the same binding specificity for this approach to succeed.

The protein structure prediction community organizes regular blind prediction tests, and sequence databases are screened for proteins that are apt to yield new folds [68]. Although there are common microarray datasets, modeling studies do not provide a common metric of success. But in any fitting procedure, the variance of the residual error as a function of the number of parameters is a good start. There is also no formal discussion of the holes in genome-wide assays and how best to plug them with some complementary technology.

Currently, the prospect of finding a 'regulatory code' seems as remote as finding a folding code for proteins, and it is not yet clear whether the space of the possible is so large that the 'code' will just be a genome-wide list. The easiest codes to break are those with most redundancy which make least efficient use of the available 'bandwidth'. So while bacteria make do with a more limited regulatory 'vocabulary' than a vertebrate, genome bandwidth is at a premium and they are apt to use it more efficiently. Interspecies comparisons are an informative filter in vertebrates as to where functional sequence lies, but the diversity of single-celled life might provide more clues as to the architecture of regulatory sequence. Lacking any organizing principles, we can merely quote instances were single base changes in regulatory sequence cause human disease [69], and others where the regulatory sequence seems plastic [40].

Acknowledgement

Both Jon Widom and Mark Johnston made valuable comments on prior drafts of this review and my research has been supported by the NSF grant DMR-0129848, and the NIH GM066434.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Tyson JJ, Chen KC, Novak B: **Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell.** *Curr Opin Cell Biol* 2003, **15**:221-231.

2. Albert R, Othmer HG: **The topology of the regulatory interactions predicts the expression pattern of the segment polarity genes in *Drosophila melanogaster*.** *J Theor Biol* 2003, **223**:1-18.

3. Oliveri P, Davidson EH: **Gene regulatory network controlling embryonic specification in the sea urchin.** *Curr Opin Genet Dev* 2004, **14**:351-360.

4. Hardison RC: **Comparative genomics.** *PLoS Biol* 2003, DOI: 10.1371/journal.pbio.0000058.

5. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98**:31-36.

6. Li H, Wang W: **Dissecting the transcription networks of a cell using computational genomics.** *Curr Opin Genet Dev* 2003, **13**:611-616.

7. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA *et al.*:

Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 2003, **21**:1337-1342.

8. Wang W, Cherry JM, Botstein D, Li H: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci USA* 2002, **99**:16893-16898.

9. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network.** *Nat Genet* 2002, **31**:370-377.

•10. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **117**:185-198.

The authors fit a Bayesian model to a wide spectrum of expression profiles in yeast and highlight for a particularly favorable case of the ribosomal genes (many genes tightly regulated) where strict position and orientation effects can be learned for several cooperating proteins.

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

The authors describe a procedure to partition genes into disjoint coexpression sets and, for each of these, infer a tree of transcriptional regulation that best explains the sets. The potential regulators are the annotated transcription factors in the genome, and regulation is inferred only from co-variation (of either sign) between the expression of the regulator and its targets.

12. Segal E, Yelensky R, Koller D: **Genome-wide discovery of transcriptional modules from DNA sequence and gene expression.** *Bioinformatics* 2003, **19(Suppl 1)**:i273-i282.

13. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.** *Genome Res* 2003, **13**:721-731.

14. Bray N, Pachter L: **MAVID: constrained ancestral alignment of multiple sequences.** *Genome Res* 2004, **14**:693-699.

15. Brudno M, Malde S, Poliakov A, Do CB, Couronne O, Dubchak I, Batzoglou S: **Glocal alignment: finding rearrangements during alignment.** *Bioinformatics* 2003, **19(Suppl 1)**:i54-i62.

16. Margulies EH, Blanchette M, Haussler D, Green ED: **Identification and characterization of multi-species conserved sequences.** *Genome Res* 2003, **13**:2507-2518.

17. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED *et al.*: **Aligning multiple genomic sequences with the threaded blockset aligner.** *Genome Res* 2004, **14**:708-715.

•18. Sinha S, van Nimwegen E, Siggia ED: **A probabilistic method to detect regulatory modules.** *Bioinformatics* 2003, **19(Suppl 1)**:i292-i301.

The algorithm presented here generalizes the Ahab code [58] to exploit block-aligned multi-species data **[Au Q1: Sorry! I can't just use a reference number as part of the text. Is it OK as I've edited this time?]**. Motifs in blocks are scored according to an evolution model for sequence constrained to bind a protein, whereas occurrences in unaligned regions are counted as independent events. Positional correlations between sites can also be learned and scored.

19. Sinha S, Blanchette M, Tompa M: **PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, DOI: 10.1186/1471-2105-5-170.

20. Halpern AL, Bruno WJ: **Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies.** *Mol Biol Evol* 1998, **15**:910-917.

21. Moses AM, Chiang DY, Kellis M, Lander ES, Eisen MB: **Position specific variation in the rate of evolution in transcription factor binding sites.** *BMC Evol Biol* 2003, **3**:19.

22. Winzeler EA, Richards DR, Conway AR, Goldstein AL, Kalman S, McCullough MJ, McCusker JH, Stevens DA, Wodicka L, Lockhart DJ *et al.*: **Direct allelic variation scanning of the yeast genome.** *Science* 1998, **281**:1194-1197.

23. Wyrick JJ, Young RA: **Deciphering gene expression regulatory networks.** *Curr Opin Genet Dev* 2002, **12**:130-136.

24. Kurdistani SK, Tavazoie S, Grunstein M: **Mapping global histone acetylation patterns to gene expression.** *Cell* 2004, **117**:721-733.

25. van Steensel B, Henikoff S: **Epigenomic profiling using microarrays.** *Biotechniques* 2003, **35**:346-50,352-4,356-7.

26. Iyer VR, Horak CE, Scafe CS, Botstein D, Snyder M, Brown PO: **Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF.** *Nature* 2001, **409**:533-538.

27. Simon I, Barnett J, Hannett N, Harbison CT, Rinaldi NJ, Volkert TL, Wyrick JJ, Zeitlinger J, Gifford DK, Jaakkola TS *et al.*: **Serial regulation of transcriptional regulators in the yeast cell cycle.** *Cell* 2001, **106**:697-708.

28. Primig M, Williams RM, Winzeler EA, Tevzadze GG, Conway AR, Hwang SY, Davis RW, Esposito RE: **The core meiotic transcriptome in budding yeasts.** *Nat Genet* 2000, **26**:415-423.

29. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282**:699-705.

30. Brem RB, Yvert G, Clinton R, Kruglyak L: **Genetic dissection of transcriptional regulation in budding yeast.** *Science* 2002, **296**:752-755.

•31. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, Mackelprang R, Kruglyak L: **Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors.** *Nat Genet* 2003, **35**:57-64.

The authors make a case for linkage analysis as an approach for discovering regulatory groupings. Of the 2300 genes showing linkage, only 25% linked to a marker within 10 kb of the genes and, hence, most variation is *trans*; but the *trans*-acting loci do not correlate with recognized transcription factors. Several co-varying gene clusters were traced to mutations that are particular to a laboratory strain.

•32. Rifkin SA, Kim J White KP: **Evolution of gene expression in the *Drosophila melanogaster* subgroup.** *Nat Genet* 2003, **33**:138-44.

The authors compare gene expression changes accompanying metamorphosis among four strains of *D. melanogaster, Drosophila simulans* and *Drosophila yakuba* – the latter two are diverged from *D. melanogaster* by 2.3 and 5 million years, respectively. Approximately half of the annotated genes varied during this transition and half of these showed some variation among the lineages. The data permit simple models of selection and drift to be fit, and the correlation between magnitude of change and evolutionary stability was also highlighted.

33. Gu Z, Rifkin SA, White KP, Li WH: **Duplicate genes increase gene expression diversity within and between species.** *Nat Genet* 2004, **36**:577-579.

•34. Wittkopp PJ, Haerum BK, Clark AG: **Evolutionary changes in *cis* and *trans* gene regulation.** *Nature* 2004, **430**:85-88.

The authors mate *D. simulans* and *D. melanogaster* and then examined transcripts for genes that are differentially expressed between the two species. In the hybrid, their sequencing technology permitted them to determine the level of each allele (*cis* differences), whereas comparison with the parent species defined *trans* changes.

•35. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES: **Sequencing and comparison of yeast species to identify genes and regulatory elements.** *Nature* 2003, **423**:241-254.

To screen for regulatory sequence, the authors ranked all six-base (with a gap) core motifs as to whether they were contained in four-way aligned blocks in the upstream regions of the species they sequenced. They subsequently filtered out motifs with significant occurrence in coding regions and 3′ of genes.

36. Kellis M, Birren BW, Lander ES: **Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*.** *Nature* 2004, **428**:617-624.

37. Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M: **Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting.** *Science* 2003, **301**:71-76.

38. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, Mohr C, Pohlmann R, Luedi P, Choi S *et al.*: **The *Ashbya gossypii***

genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* 2004, **304**:304-307.

39. Kellis M, Patterson N, Birren B, Berger B, Lander ES: **Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery.** *J Comput Biol* 2004, **11**:319-355.

40. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA: **The evolution of transcriptional regulation in eukaryotes.** *Mol Biol Evol* 2003, **20**:1377-1419.

41. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci USA* 2000, **97**:10096-10100.

•42. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J *et al.*: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**:99-104.

In this study, ChIP and interspecies conservation data are used to predict sequence motifs for over 200 yeast transcription factors. The motifs are then used to identify binding sites in promoters bound by the corresponding factor, and further filtered by presence in several other species. Regulatory inputs are found for 1300 genes.

43. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9**:3273-3297.

44. Horak CE, Luscombe NM, Qian J, Bertone P, Piccirrillo S, Gerstein M, Snyder M: **Complex transcriptional circuitry at the G_1/S transition in *Saccharomyces cerevisiae*.** *Genes Dev* 2002, **16**:3017-3033.

45. Wang Y, Pierce M, Schneper L, Guldal CG, Zhang X, Tavazoie S, Broach JR: **Ras and gpa2 mediate one branch of a redundant glucose signaling pathway in yeast.** *PLoS Biol* 2004, DOI: 10.1371/journal.pbio.0020128.

46. Zeitlinger J, Simon I, Harbison CT, Hannett NM, Volkert TL, Fink GR, Young RA: **Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signaling.** *Cell* 2003, **113**:395-404.

47. Powell JR: *Progress and Prospects in Evolutionary Biology.* Oxford University Press; 1997.

48. Elnitski L, Hardison RC, Li J, Yang S, Kolbe D, Eswara P, O'Connor MJ, Schwartz S, Miller W, Chiaromonte F: **Distinguishing regulatory DNA from neutral sites.** *Genome Res* 2003, **13**:64-72.

49. Emberly E, Rajewsky N, Siggia ED: **Conservation of regulatory elements between two species of *Drosophila*.** *BMC Bioinformatics* 2003, DOI: 10.1186/1471-2105-4-57.

•50. Schroeder MD, Pearce M, Fak J, Fan H, Unnerstall U, Emberly E, Rajewsky N, Siggia ED, Gaul U: **Transcriptional control in the segmentation gene network of *Drosophila*.** *PLoS Biol* 2004, DOI: 10.1371/journal.pbio.0020271.

This study enlarges the number of confirmed regulatory modules for the anterior–posterior patterning network by 40% by using a computational screen for binding site clusters to target experiments. The authors demonstrate a strong correlation between computed module composition and where it is expressed. In addition, the presence of both activators and repressors distinguishes functional modules from other regions with comparable scores.

51. Bergman CM, Pfeiffer BD, Rincon-Limas DE, Hoskins RA, Gnirke A, Mungall CJ, Wang AM, Kronmiller B, Pacleb J, Park S *et al.*: **Assessing the impact of comparative genomic sequence data on the functional annotation of the *Drosophila* genome.** *Genome Biol* 2002 DOI: 10.1186/gb-2002-3-12-research0086.

52. Petrov DA, Hartl DL: **High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups.** *Mol Biol Evol* 1998, **15**:293-302.

53. Sinha S, Siggia ED: **Sequence turnover and tandem repeats in *cis*-regulatory modules in *Drosophila*.** *Mol Biol Evol* 2004, DOI:10.1093/molbev/msi090

53. Yuh CH, Brown CT, Livi CB, Rowen L, Clarke PJ, Davidson EH: **Patchy interspecific sequence similarities efficiently identify positive *cis*-regulatory elements in the sea urchin.** *Dev Biol* 2002, **246**:148-161.

54. Cooper GM, Sidow A: **Genomic regulatory regions: insights from comparative sequence analysis.** *Curr Opin Genet Dev* 2003, **13**:604-610.

55. Stathopoulos A, Van Drenth M, Erives A, Markstein M, Levine M: **Whole-genome analysis of dorsal-ventral patterning in the *Drosophila* embryo.** *Cell* 2002, **111**:687-701.

56. Rajewsky N, Vergassola M, Gaul U, Siggia ED: **Computational detection of genomic cis-regulatory modules applied to body patterning in the early *Drosophila* embryo.** *BMC Bioinformatics* 2002, DOI: 10.1186/1471-2105-3-30.

•57. Jaeger J, Blagov M, Kosman D, Kozlov KN, Manu, Myasnikova E, Surkova S, Vanario-Alonso CE, Samsonova M, Sharp DH *et al.*: **Dynamical analysis of regulatory interactions in the gap gene system of *Drosophila melanogaster*.** *Genetics* 2004, **167**:1721-1737.

The authors fit a model for the gap gene protein network using an empirical response function whose argument is the fitting matrix of protein interactions. They fit this to data at ~7 minute time intervals, and over most of the embryo (pole-specific factors are missing). They do not fit mutant data.

•58. Berman BP, Pfeiffer BD, Laverty TR, Salzberg SL, Rubin GM, Eisen MB, Celniker SE: **Computational identification of developmental enhancers: conservation and function of transcription factor binding-site clusters in *Drosophila melanogaster* and *Drosophila pseudoobscura*.** *Genome Biol* 2004, **5**:R61.

The authors examine the expression patterns of 28 uncharacterized modules found within their top 37 predictions for anterior–posterior patterned genes. Six of the 28 are active in the embryo, and mirror the expression of a neighboring gene, although two of these are neuroblast specific. They consider various criteria that would discriminate their active from inactive modules and find that binding site conservation in *D. pseudoobscura* works best.

•59. Markstein M, Zinzen R, Markstein P, Yee KP, Erives A, Stathopoulos A, Levine M: **A regulatory code for neurogenic gene expression in the *Drosophila* embryo.** *Development* 2004, **131**:2387-2394.

The authors test 15 genome-wide predictions of strong dorsal binding site clusters, generated by regular expression matches. This scan recovered one out of seven dorsal-regulated modules known at the time. Of the fifteen, five are functional and expressed at various levels in the nuclear dorsal gradient. The authors then predicted and tested other binding sites (two of which corresponded to Twist and Su(H)) common to three neural ectoderm genes, and scanned the genome with the entire set, recovering several functional modules that expressed in the expected region. From the data in this article, modules do not read their position in the dorsal gradient by the strength of the dorsal binding sites.

60. Markstein M, Markstein P, Markstein V, Levine MS: **Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the *Drosophila* embryo.** *Proc Natl Acad Sci USA* 2002, **99**:763-768.

61. Sinha S, Schroeder MD, Unnerstall U, Gaul U, Siggia ED: **Cross-species comparison significantly improves genome-wide prediction of cis-regulatory modules in Drosophila.** *BMC Bioinformatics* 2004, **5**:129.

•62. Grad YH, Roth FP, Halfon MS, Church GM: **Prediction of similarly-acting cis-regulatory modules by subsequence profiling and comparative genomics in *D. melanogaster* and *D. pseudoobscura*.** *Bioinformatics* 2004, **20**:2738-2750.

Without using prior knowledge about factor binding sites, the authors determine modules common to a set of co-expressed genes. They first screen rather liberally for regions of strongest interspecies conservation, and then use a Gibbs-like search that models the current set of modules as a fifth order Markov model, searches for similar modules within the filtered sequence, and then updates the Markov model. Validation was done against the fly *in situ* set.

63. Sandelin A, Wasserman WW: **Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics.** *J Mol Biol* 2004, **338**:207-215.

64. Michaut L, Flister S, Neeb M, White KP, Certa U, Gehring WJ: **Analysis of the eye developmental pathway in *Drosophila* using DNA microarrays.** *Proc Natl Acad Sci USA* 2003, **100**:4024-4029.

65. Butler MJ, Jacobsen TL, Cain DM, Jarman MG, Hubank M, Whittle JR, Phillips R, Simcox A: **Discovery of genes with highly restricted expression patterns in the *Drosophila* wing disc using DNA oligonucleotide microarrays.** *Development* 2003, **130**:659-670.

•66. Stolc V, Gauhar Z, Mason C, Halasz G, van Batenburg MF, Rifkin SA, Hua S, Herreman T, Tongprasit W, Barbano PE *et al.*: **A gene expression map for the euchromatic genome of *Drosophila melanogaster*.** *Science* 2004, **306**:655-660.

The authors built custom arrays with probes to individual exons in addition to intergenic sequence, and monitored RNA expression using whole animals during six developmental stages. They found extensive expression of genomic regions that are not coding, and correlation of gene expression within domains syntenic with *D. pseudoobscura*.

67. van Nimwegen E, Zavolan M, Rajewsky N, Siggia ED: **Probabilistic clustering of sequences: inferring new bacterial regulons by comparative genomics.** *Proc Natl Acad Sci USA* 2002, **99**:7323-7328.

68. Chandonia JM, Brenner SE: **Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches.** *Proteins* 2004, in press.

69. Rockman MV, Hahn MW, Soranzo N, Loisel DA, Goldstein DB, Wray GA: **Positive selection on MMP3 regulation has shaped heart disease risk.** *Curr Biol* 2004, **14**:1531-1539.

70. Morgenstern, B: **A space-efficient algorithm for aligning large genomic sequences.** *Bioinformatics* 2000 **16**:948-949.

71. Zhu J, Zhang MQ: **SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*.** *Bioinformatics* 1999, **15**:607-611.

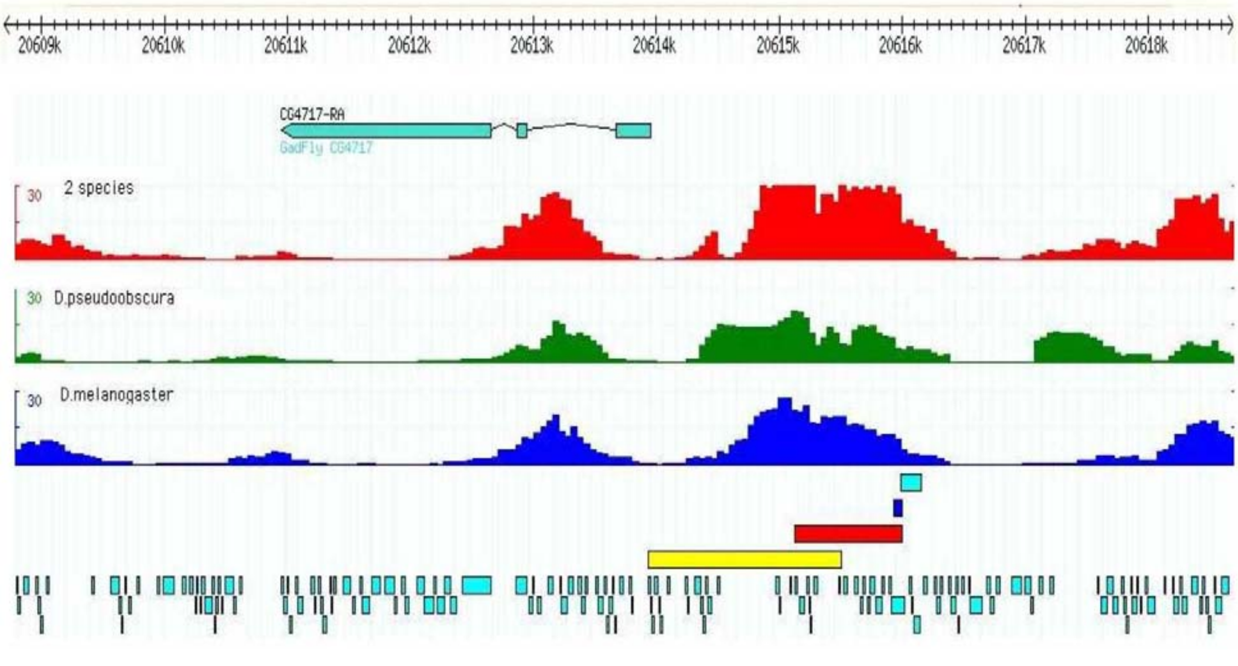

Figure 1

Computational analysis of the regulatory sequence around the *gap* gene knirps from [18*]. The blue (resp green) profiles denote the log likelihood score for regulation by the segmentation genes in *D. melanogaster*, and *D. pseudoobscura*. The red profile is the score based on both species after aligning the genomic sequences. The computation predicted a regulatory module within the first intron which was subsequently verified in [50*]. Prior promoter bashing concentrated on the first few kb of upstream sequence and several of the regions that gave expression are shown as colored bars. They are not commensurate with the module delineation suggested by the computation. The dense array of bars at the bottom, are the blocks of sequence that align between the two species in an order preserving manner, offset for clarity.

Table 1

Scoring method	Known sites	Random sites
Continuous	28.0	21.7 +/−0.7
Binary (stringent, 407* max)	184	113 +/−8
Binary (permissive, 407 max)	267	194 +/−9

Interspecies conservation for a set of 407 experimentally footprinted binding sites upstream of 194 genes in *S. cerivisiae*. Noncoding sequence upstream of the gene was obtained for all the sensu-stricto species [37,35•] and aligned according to the method presented by Morgenstern [70]. Each site was then scored by three different methods† for its overlap with sequence that aligned with the other species (i.e. was conserved). The sites were then randomized in position, and the overlap rescored in a consistent way. The similarity between the second and third columns shows that much of the conservation that is interpreted as evidence for functionality is to be expected by chance. The last two rows show that only about half of the sites would be recovered if one demanded conservation. *407 refers to the maximum possible score. †For the continuous scoring method, the score is the sum over all bases in the site, of the number of species aligned with *S. cerivisiae*. For the stringent scoring method, a site was counted as conserved if at least 75% of its bases were aligned in at least 3 other species. For the permissive method, at least 2 other species had to show 75% sequence conservation**[Au Q2: Sorry, I'm still unsure what was meant. Is this right?]**. If fewer than the minimum number of species were available, then they all had to be aligned for conservation. Imposing 50% base conservation increases the recovery of known sites and the number of random sites by 10%. The randomization was done so as to preserve the distribution in position relative to transcription start. The binding sites were taken from the study by Zhu and Zhang [71] and filtered for overlaps.

Table 2

Factor	Motif	Predicted [35•]	Predicted [41]
ABF1	RTCRYnnnnnACG	RTCRYknnnnACGR	WRTCAnnnnADACGDM
UME6	TCGGCGGCTA	TSGGCGGCTAWW	TCGGCGGCTA, TGGGCGGCTA
CBF1	RTCACRTG	RTCACGTGV	RTCACGTG, ATCACGTGA
REB1	TTACCCGG	RTTACCCGRM	TACCCGG, GTTACCCG, TATTACCCG, TACCCGGC
MCM1a	TTWCCCnWWWRGGAAA	TTCCnaAttnGGAAA	TTTCCnnnnnnGGAAA
SWI6(MCB)	ACGCG	WCGCGTCGCGt	ACGCGTTT, ACGCGTCGCG, ACGCGTCA
PHO4	CACGTG	RTCACGTGV	CACGTGMT, GTCACGTG, AGCACGTG, TCACGTGC
SWI4(SCB)	TTTTCGCG	WTTTCGCGTT	TTTCGCGT, TTCGCGTT
DAL81	GATAAG	–	AGATAAGA, GATAAGGA
RPN4	TTTTGCCACC	TTTTGCCACCG	TTTGCCACC
MSN2	CCCCT	hRCCCYTWDt	CCGCCCCT,ATCCCCCT, CCCCTCAT, GCCCCTTC, CCCCTTCC
PDR1	CCGCGG	YCCGSGGS	CCCGCGGC, CCGCGGA
MSE(NDT80)	TTTTGTG	TTTTGTGTCRC	–
STE12	RTGAAACA	–	–
DIG1	RTGAAACA	–	–
MET4	TGGCAAATG	CGGTGGCAAAA	CGGTGGCAAA
HAP4	TnRTTGGT	–	TTGTTGCT, TGATTAGT
SMP1	ACTACTAWWWWTAG	–	–
ACE2(SWI5)	GCTGGT(KGCTGR)	–	TTGCTGAC, GGCTGGGC, TTGCTGTT
YAP1	TTACTAA	–	–
CIN5	TTACTAA	–	–
RME1	GAACCTCAA	–	CAACCTCA, CCTCAATG, ACCTCATC
HAC1	CAGCGTG	–	GTCAGCGT, ACAGCGAG, AGAGCGTG, TCAGCGTC
GCR1	GGAAG	–	GGGAAGGG, GGGGAAGG, GGGAAGAG, AATGGAAG, GGAAGCCC

Recovery of known yeast binding motifs from genome-wide interspecies comparisons. Predictions in column 3 are calculated using the methodology set out by Kellis *et al.* (Table 2 [35•]), compared with predictions from a single genome in column 4 [41]. Only the top and bottom 12 entries from [35•] are shown. The results are comparable.