Discovery and modeling of transcriptional regulatory regions

James W Fickett* and Wyeth W Wasserman†

A complex network of regulatory controls governs the patterns of gene expression. Enabled by the tools of molecular cloning, initial experimental queries into the gene regulatory network elucidated a wide array of transcription factors and their cognate binding sites from hundreds of genes. The recent fusion of genome-scale experimental tools, a more comprehensive gene catalog, and concomitant advances in computational methodology, has extended the range of questions being posed. The potential to further our understanding of the biochemical mechanisms of transcriptional regulation and to accelerate the delineation of regulatory control regions in the human genome is enormous.

Addresses
*Bioinformatics Group, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, Mailstop UW2230, King of Prussia, PA 19406, USA; e-mail: James_W_Fickett@sbphrd.com
†Center for Genomics Research, Karolinska Institute, Stockholm, Sweden
Corresponding author: James W Fickett

Current Opinion in Biotechnology 2000, **11**:19–24

0958-1669/00/$ – see front matter

Abbreviations

PF	phylogenetic footprinting
PWM	position weight matrix
TF	transcription factor
TIC	transcription initiation complex
TRR	transcriptional regulatory region
TSS	transcription start site

Introduction

The sea urchin *Endo 16* gene illustrates the complexity of the gene regulatory network [1••]. The upstream region of this gene contains at least 33 transcription factor (TF; DNA-binding proteins controlling the initiation of transcription) binding sites in five modules (Figure 1).

One may consider the arrangement of TF binding sites to be a control language encoded in the DNA. The core motifs of the language are the DNA binding specificities of individual TFs [2••,3••]. Expression data generated from microarray studies, in conjunction with a comprehensive collection of genes from genomic sequence resources, promises to accelerate the discovery of new motifs. The underlying principle is that a set of genes should be classified tentatively as 'co-regulated' if they share a similar pattern of gene expression and common DNA sequence motifs likely to be bound by TFs [4••]. A second major accelerating force will be comparison of genomic regions from related organisms, or phylogenetic footprinting (PF) [5••]; TF binding sites stand out clearly against a nonconserved background.

Due to a convenient experimental system [6], sufficient data has accumulated concerning gene regulation in muscle [7•] that it may be possible to discover, in this context, the meaning of higher order patterns in the regulatory language. Different genes upregulated in muscle have very different regulatory regions (Figure 2). Nevertheless, it is possible to discriminate, at a practical level of accuracy, between muscle and non-muscle control regions [8••].

To describe current advances in deciphering the regulatory control language we establish a vocabulary concerning the initiation of transcription, describe how PF can be used to enrich signals for analysis, define a formalism for describing TF binding sepcificities and its use, along with PF, in discovering new TFs via genome-scale expression

Figure 1

Regulatory circuitry of the sea urchin *Endo 16* gene (based on [1••]). The basal promoter (Bp) localizes the point of transcription initiation. **(a)** In early development, A alone provides initial upregulation; later it serves as a logical integrator to collect and relay signals from the other modules. **(b)** B upregulates the gene at later stages of development. G is a general enhancer, adding to the effect of A and B. **(c)** CDEF repress the gene in all but the correct tissues.

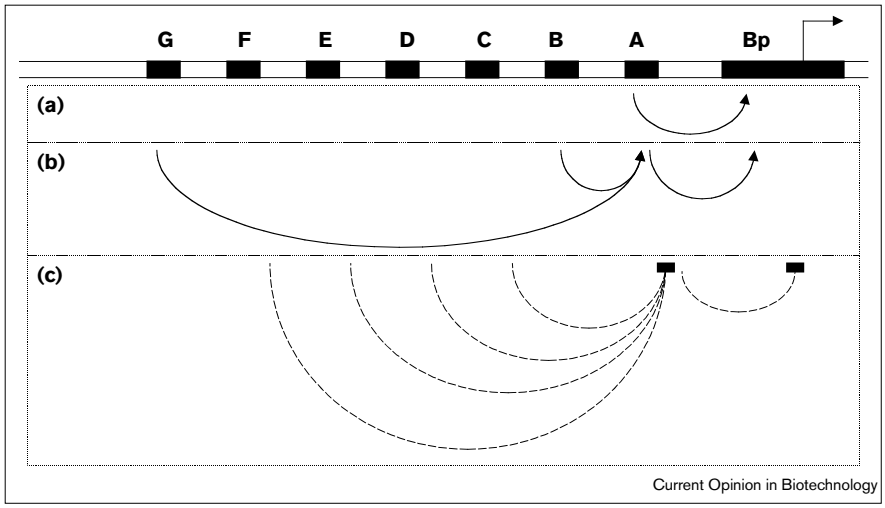
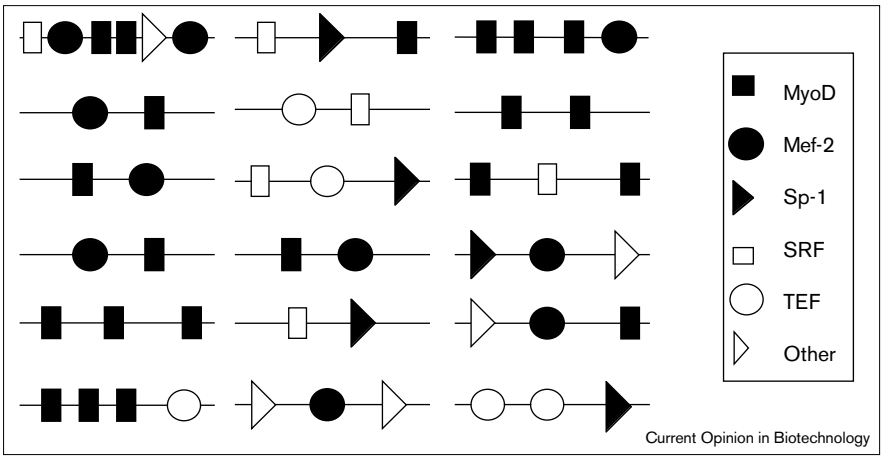


Figure 2



Diversity of modules. Each of the 18 line segments shows schematically the structure of one skeletal-muscle-specific module, in terms of known TF-binding sites, summarized from [7] and listed on the right. Though all these modules upregulate expression of various genes in the same tissue, it is difficult to discern any syntactic similarity.

data, and, finally, summarize early progress on describing the syntax of the language.

Biological background

Control of transcription initiation is a primary component of the control of gene expression and the one to which DNA sequence is most relevant. For recent overviews of the biochemical mechanisms involved see [9–11]. Here we will only define a convenient nomenclature (Figure 3) and mention a few current issues.

By promoter we mean a region of DNA surrounding the transcription start site (TSS) that is able to direct transcription (perhaps at low levels) from the correct TSS. The transcription initiation complex (TIC) is a large complex of proteins, including RNA polymerase II, which assembles on the promoter and initiates mRNA synthesis.

Although the population of TFs varies greatly across temporal and spatial space, it was thought until recently that the TIC was essentially invariant. It has become apparent, however, that some components of the TIC vary in a way that is probably related to the expression context of the affected genes [12,13•].

When several TF binding sites occur in a cluster, and the mutation of any one of them seems to affect the up- or down-regulation of a gene in the same context (as illustrated in the *Endo 16* example above), the cluster is termed a regulatory module. If activation of the module results in higher expression for the gene, it is termed an enhancer, otherwise a repressor. Enhancers, repressors, and promoters are together known as transcriptional regulatory regions (TRRs).

For a wider review of transcription, emphasizing the role of chromatin, see the issue of *Current Opinion in Genetics and Development* introduced by Kadonaga and Grunstein [14•].

Correlation between mRNA expression levels and the corresponding protein levels cannot reliably be determined given the current amount and precision of data [15–17]. Much of the interest in gene expression levels, however, is in quantifying a change in expression level between healthy and diseased tissue, or between the presence and absence of a signaling molecule. It is a reasonable first approximation to assume that when transcription increases, the other effects on protein level remain the same, and the protein level increases as well.

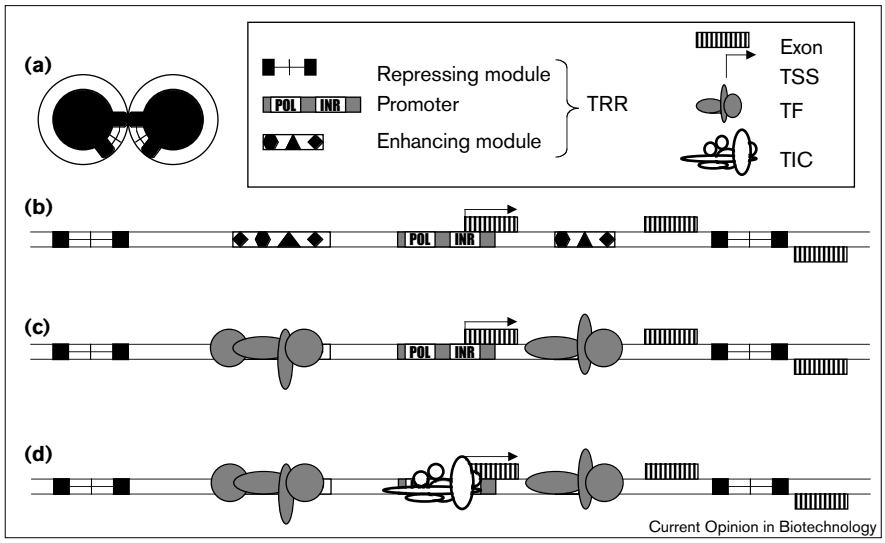
Phylogenetic footprinting to locate transcriptional regulatory regions

The term PF has been suggested, by analogy with DNAase footprinting, "to describe the phylogenetic comparisons that reveal conserved *cis*-elements in the noncoding regions of homologous genes" [18]. One might use the term more generally to cover identification of any functional regions by comparison of orthologous genomic sequence segments between species. PF, followed by experimental verification, is the most efficient method available for the initial identification of most TRRs (CpG islands, which are characterized by their base composition rather than by specific binding sites, may be an exception [19••]).

The optimal evolutionary distance for comparison varies widely with the particular gene [20]. Only 16% of orthologous gene pairs between human and bony fish (separated by about 450 million years) showed 70% identity over at least 50 basepairs (bp) in the noncoding regions [21], yet in some cases the conservation is striking and useful. Large regions known to contain regulatory elements for the *Hoxb-4* gene in the mouse, for example, were narrowed significantly by comparison with pufferfish, and the resulting putative TRRs were verified by transgenesis [22].

Although *Caenorhabditis elegans* and *Caenorhabditis briggsae* diverged only 23–40 million years ago, Thacker *et al.* [23•] found that: "Conservation of DNA sequences is

Figure 3



Modular structures of promoters. **(a)** DNA inaccessible due to chromatin structure. **(b)** DNA remodeled to present regulatory sites. **(c)** Transcriptional activators bind to regulatory sites within modules. **(d)** Polymerase complex recruited to promoter.

confined largely to protein-coding regions and short flanking sequences. Comparative studies have demonstrated that conserved 5′ flanking sequences often constitute *cis*-acting elements that are involved in the regulation of gene expression".

Most human regulatory regions can probably be discovered by comparison of human and rodent sequences. In a dataset of orthologous human and rodent muscle genes, we found that the footprints in the non-coding sequences from these genes cover just 19% of the human non-coding sequence yet contain over 98% of the experimentally-defined, sequence-specific TF binding sites (WW Wasserman *et al.*, unpublished data). An alignment algorithm that finds only 0.14% conserved DNA in randomly paired sequences finds about a third of the noncoding DNA conserved in transcripts and the kilobase upstream (a region likely to overlap considerably with regulatory elements) [24]. New estimates of divergence rates were calculated from 1880 orthologous human–rodent gene pairs by Makalowski and Boguski [25]. A number of examples of human–rodent PF predictions of TRRs have been verified experimentally [19••]. The usefulness of mouse in this regard is one motivating force for plans to sequence the mouse genome [26]. It would appear that PF is one of the most important sources of information in large-scale sequence analysis of the human genome [5••].

As more sequence accumulates, it will be valuable to compare the extent of the evolutionary tree showing conservation of, firstly, a particular sequence element, secondly, the proteins binding that element, and finally, a possibly correlated phenotype [27]. A large collection of homologous sets of vertebrate genes may be found in the database HOVERGEN [28]. Methods for selecting conserved blocks from a multiple alignment are evaluated by Stojanovich *et al.* [29].

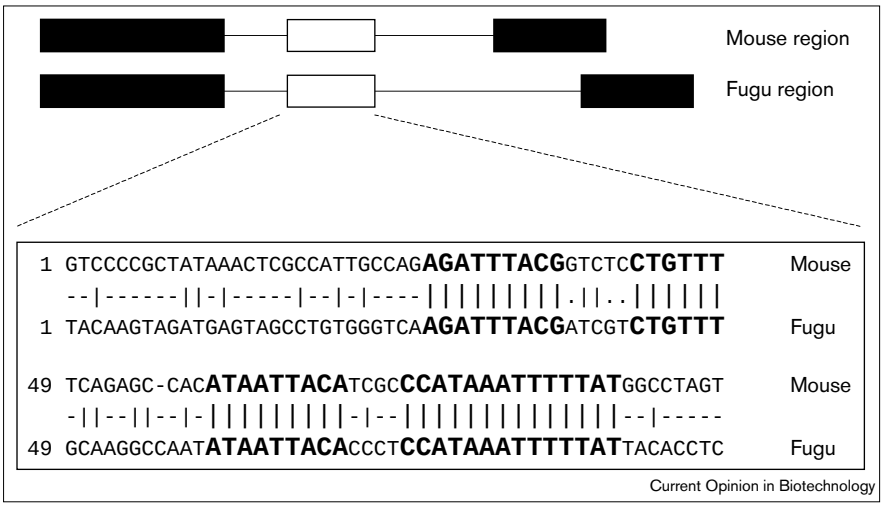
PF requires specialized alignment tools. The probability of spurious matches between arbitrary portions of long genomic segments is high; however, biologically meaningful alignment within a syntenic region is constrained by matches between nearby orthologous elements. Programs designed for very long alignments of syntenic regions include a derivation of SIM in PIPmaker (PIP is percent identity plot) [5••] and MUMmer [30]. Anecdotal evidence suggests that orthologous control elements normally match essentially perfectly (Figure 4). PIP selects blocks over 50 bp, gap-free, and with over 60% identity. The Bayes Block Aligner (BBA) [31••] may be especially well suited to discovery of regulatory elements. It is designed to find gapless blocks and, because the posterior probability of a block is obtained by summing over all alignments containing the block, a diffuse synteny will strongly support a true orthologous block even if there is no one high-quality alignment of the region.

Transcription factors and binding sites

Although the binding specificity of a TF is often expressed by means of a consensus sequence, a position weight matrix (PWM) is usually more appropriate. The PWM assigns a weight to each possible nucleotide at each position within the site, reflecting the frequency with which the given nucleotide occurs at the given position. The score of a particular site is obtained by summing the corresponding weights. This captures more information than a consensus, and has a sound foundation in both statistics (representing likelihood ratios) and thermodynamics (representing binding energies) [2••,32•,33–36].

When an HNF1 (hepatic nuclear factor 1) PWM derived from 21 known sites was used to search the databases, 95% of the sites so discovered were bound by HNF1 *in vitro* [37], suggesting that a PWM developed from a sufficiently

Figure 4



Phylogenetic footprinting for the identification of regulatory sequences. Blocks of contiguous identical bases in alignments of non-coding sequences from distantly related organisms are more likely to contain TF-binding sites. This example is from a comparison of mouse and *Fugu rubripes* (Fugu) Hoxb4 genes [22]. Similar conservation was observed in exons (black boxes) and non-coding sequences (white boxes).

large set of experimentally defined binding sites can accurately predict the *in vitro* binding of a transcription factor. Matches to sites with no *in vivo* function are common. This is not a failure of the computational techniques, but rather reflects biological reality: competition, chromatin structure and other influences are as important as binding affinity [38•]. End-user software for locating matches to a consensus or PWM in genomic DNA has been reviewed [39], and tests for statistical significance developed [40].

The binding specificity of only a small fraction of transcription factors is known. The TRANSFAC database [3••] contains a significant fraction of published, verified binding sites and may include a PWM when multiple sites are known. The few in-depth studies of a single gene's regulatory apparatus (e.g. [1••,41,42•]) serve in part to highlight how much of the full genome regulatory network remains a mystery. It is hoped that the new microarray technology [43•] (see this issue Epstein and Butow, pp 36–41; Ladunga pp 13–18) or suppression subtractive hybridization (SSH) [44] will provide sets of co-regulated genes on a large scale (see also [45•]). Several groups were able to rediscover sets of sites representing known transcription factor binding specificities, and to suggest new ones, by aligning the upstream regions of genes co-expressed under specific conditions in yeast (see [4••,46] for overviews). The approach used in yeast may have been successful, in part, because yeast regulatory elements (unlike those in multicellular organisms) are almost always within a few hundred bases of the translation start site. In multicellular organisms, the large amount of DNA that must be searched for possible regulatory regions has, in our experience, added enough noise to the data that current multiple alignment algorithms (e.g. [47]) were unable to find common patterns. We found, however, that when PF was used to reduce the sequence space, it was possible to rediscover mammalian control elements *de novo* (WW Wasserman *et al*., unpublished data).

Regulatory modules

As the core promoter contains frequently occurring patterns at fairly well defined positions [33,48••], and also often contains binding sites for specific factors, it was to be hoped that fairly simple computational methods might provide reliable promoter recognition. This turned out not to be the case [49]. The majority of regulatory regions, probably including most core promoters, are to some extent context-specific. It has been known for some time that highly context-specific regulatory modules could sometimes be recognized by means of particular motifs [50,51]. Several groups have developed formalisms to represent multiple patterns occurring within certain spacing constraints; in one application, it was shown that viral long terminal repeats could be located within large sequence collections with high sensitivity and specificity [52].

Recently statistical models have been developed to quantitatively model the clustering of sites often seen in regulatory regions. The first was a Hidden Markov Model (HMM) in which the hidden state signals whether the current nucleotide is or is not within a regulatory region. Within regulatory regions the model expects shorter average spacing between sites, so that clusters of sites are more likely to be modeled as regulatory regions [53]. An advantage of this formalism is that the size of the regulatory region is estimated from the data. To test the significance of a given cluster of sites for a single TF (a generalization to multiple TFs should be possible), one may calculate the probability of finding k sites within a space of X nucleotides, making the neutral assumption of a Poisson distribution and taking sequence heterogeneity into account [54•]. A practical algorithm to estimate the overall probability of TRR function in muscle has been given, using logistic regression to combine PWM scores for a few key TFs [8••].

Current models will become more complex as understanding of regulatory region structure grows. To date no algorithm takes into account the synergistic binding of adjacent TFs that depends on the order and spacing of the binding sites [55,56]. Also, current models tend to assume that all sites of a module are bound, but anecdotal evidence suggests that in many cases a module includes several alternative sites, any one of which may be bound to produce activation. To model this situation it will be important to estimate the overall probability of a transcription factor occupying some one of several closely spaced binding sites [57].

Conclusion

Though individual putative TF binding sites are too abundant to analyze in full, approaches that cluster together sites with some biologically intuitive connection between the predicted TF binding sites and the function of the gene often reduce the output to a manageable size and bring to the fore reasonable hypotheses for testing.

Expression data coupled with phylogenetic footprinting will soon elucidate most of the patterns describing TF binding specificities, increasing the opportunities for unravelling higher order organization in the regulatory language. At the minimum, these advances should result in an understanding of the regulatory modules directing gene expression to many contexts. Ideally, the elucidation of the regulatory language will enable the design of context-specific expression tools for experimental and therapeutic (e.g. [58]) endeavors.

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. Yuh CH, Bolouri H, Davidson EH: **Genomic cis-regulatory logic:**
•• **experimental and computational analysis of a sea urchin gene.** *Science* 1998, **279**:1896-1902.

A decade of work has served to characterize the regulation of this gene as deeply as any. The authors go beyond listing the effects of the individual sites, attempting to lay out what is, in essence, the decision algorithm encoded in the DNA for responding to the more than 15 transcription factors controlling the regulation of this gene.

2. Stormo GD, Fields DS: **Specificity, free energy and information**
•• **content in protein–DNA interactions.** *Trends Biochem Sci* 1998, **23**:109-113.

An excellent overview of the theory and practice underlying mathematical characterization of transcription factor binding sites.

3. Heinemeyer T, Chen X, Karas H, Kel AE, Kel OV, Liebich I, Meinhardt T,
•• Reuter I, Schacherer F, Wingender E: **Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms.** *Nucleic Acids Res* 1999, **27**:318-322.

The largest existing collection of known transcription factor binding specificities.

4. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM:
•• **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22**:281-285.

Clusters of genes derived from expression data are shown to correlate with functional groupings. Transcription factor binding specificities, some old and some new, are derived and shown to be specific to the clusters and near translation starts. This paper also provides an overview of other similar analyses.

5. Ansari-Lari MA, Oeltjen JC, Schwartz S, Zhang Z, Muzny DM, Lu J,
•• Gorrell JH, Chinault AC, Belmont JW, Miller W, Gibbs RA: **Comparative sequence analysis of a gene-rich cluster at human chromosome 12p13 and its syntenic region in mouse chromosome 6.** *Genome Res* 1998, **8**:29-40.

A large-scale analysis that shows both the utility and the difficulties of phylogenetic footprinting.

6. Taylor SM, Jones PA: **Multiple new phenotypes induced in 10T1/2 and 3T3 cells treated with 5-azacytidine.** *Cell* 1979, **17**:771-779.

7. Lopez LL, Wasserman WW, Fickett JW: **Muscle-specific regulation**
• **of transcription: a catalog of regulatory elements.** http://www.cbil.upenn.edu/MTIR/HomePage.html

No longer up to date, but still one of the best available collections of data about the regulation of genes in one particular context.

8. Wasserman WW, Fickett JW: **Identification of regulatory regions**
•• **which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278**:167-181.

The authors describe the first practical algorithm for discriminating those regulatory regions active in a particular context.

9. Tjian R: **Molecular machines that control genes.** *Sci Am* 1995, **272**:54-61.

10. Blackwood EM, Kadonaga JT: **Going the distance: a current view of enhancer action.** *Science* 1998, **281**:61-63.

11. Hampsey M, Reinberg D: **RNA polymerase II as a control panel for multiple coactivator complexes.** *Curr Opin Genet Dev* 1999, **9**:132-139.

12. Rabenstein MD, Zhou S, Lis JT, Tjian R: **TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family.** *Proc Natl Acad Sci USA* 1999, **96**:4791-4796.

13. Holstege FCP, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Bren MR,
• Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95**:717-728.

Various components of the transcription initiation complex are knocked out or mutated, and the effect on expression of large numbers of genes measured. The study shows that several proteins formerly thought to play a generic role in transcription initiation probably govern the expression of specific classes of genes.

14. Kadonaga JT, Grunstein M: **Chromosomes and expression**
• **mechanisms: chromatin: the packaging is the message.** *Curr Opin Genet Dev* 1999, **9**:129-131.

An excellent overview of the role of chromatin in controlling gene expression.

15. Kawamoto S, Matsumoto Y, Mizuno K, Okubo K, Matsubara K: **Expression profiles of active genes in human and mouse livers.** *Gene* 1996, **174**:151-158.

16. Anderson L, Seilhamer J: **A comparison of selected mRNA and protein abundances in liver.** *Electrophoresis* 1997, **18**:533-537.

17. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast.** *Mol Cell Biol* 1999, **19**:1720-1730.

18. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **Embryonic ε and γ globin genes of a prosimian primate (*Galago crassicaudatus*) nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.** *J Mol Biol* 1988, **203**:439-455.

19. Hardison RC, Oeltjen J, Miller W: **Long human-mouse sequence**
•• **alignments reveal novel regulatory elements: a reason to sequence the mouse genome.** *Genome Res* 1997, **8**:959-966.

A lucid discussion of principles, practice, and specific examples, giving a good overview of the utility of phylogenetic footprinting.

20. Koop BF, Richards JE, Durfee TD, Bansberg J, Wells J, Gilliam AC, Chen H-L, Clausell A, Tucker PW, Blattner FR: **Analysis and comparison of the mouse and human immunoglobulin heavy chain J_h-Cμ-Cδ locus.** *Mol Phylogenet Evol* 1996, **5**:33-49.

21. Duret L, Bucher P: **Searching for regulatory elements in human noncoding sequences.** *Curr Opin Struct Biol* 1997, **7**:399-406.

22. Aparicio S, Morrison A, Gould A, Gilthorpe J, Chaudhuri C, Rigby P, Krumlauf R, Brenner S: **Detecting conserved regulatory elements with the model genome of the Japanese puffer fish, *Fugu rubripes*.** *Proc Natl Acad Sci USA* 1995, **92**:1684-1688.

23. Thacker C, Marra MA, Jones A, Baillie DL, Rose AM: **Functional**
• **genomics in *Caenorhabditis elegans*: an approach involving comparisons of sequences from related nematodes.** *Genome Res* 1999, **9**:348-359.

As the genome of *C. elegans* is completed, and that of *C. briggsae* will be completed in the near future, the worm will be a very important case for

learning to apply phylogenetic footprinting on a large scale in multicellular organisms. This paper provides a good overview of the current situation.

24. Jareborg N, Birney E, Durbin R: **Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs.** *Genome Res* 1999, **9**:815-824.

25. Makalowski W, Boguski MS: **Evolutionary parameters of the transcribed mammalian genome: an analysis of 2829 orthologous rodent and human sequences.** *Proc Nat Acad Sci USA* 1998, **95**:9407-9412.

26. Battey J, Jordan E, Cox D, Dove W: **An action plan for mouse genomics.** *Nat Genet* 1999, **21**:73-75. [Updated at http://www.nih.gov/news/pr/oct99/nhgri-05.htm]

27. Gumucio DL, Shelton DA, Zhu W, Millinoff D, Gray T, Bock JH, Slightom JL, Goodman M: **Evolutionary strategies for the elucidation of *cis* and *trans* factors that regulate the developmental switching programs of the β-like globin genes.** *Mol Phylogenet Evol* 1996, **5**:18-32.

28. Duret L, Mouchiroud D, Gouy M: **HOVERGEN, a database of homologous vertebrate genes.** *Nucleic Acids Res* 1994, **22**:2360-2365. [ftp://pbil.univ-lyon1.fr/pub/hovergen/]

29. Stojanovich N, Florea L, Riemer C, Gumucio D, Slightom J, Goodman M, Miller W, Hardison R: **Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions.** *Nucleic Acids Res* 1999, **27**:3899-3910.

30. Delcher AL, Kasif S, Fleischmann RD, Peterson J, White O, Salzberg SL: **Alignment of whole genomes.** *Nucleic Acids Res* 1999, **27**:2369-2376.

31. •• Zhu J, Liu JS, Lawrence CE: **Bayesian adaptive sequence alignment algorithms.** *Bioinformatics* 1998, **14**:25-39.
The authors present a profoundly original alignment algorithm that may be most appropriate for discerning short conserved blocks, such as those found in regulatory regions. The original algorithm, intended for protein sequences, is adapted to DNA sequences on http://www.wadsworth.org/resnres/bioinfo/.

32. • Roulet E, Fisch I, Junier T, Bucher P, Mermod N: **Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA.** *In Silico Biol* 1998, **1**:21-28.
An overview of position weight matrices that emphasizes the limitations of the technology by exploring its use in one particularly difficult case.

33. Bucher P: **Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences.** *J Mol Biol* 1990, **212**:563-578.

34. Barrick D, Villanueba K, Childs J, Kalil R, Schneider TD, Lawrence CE, Gold L, Stormo GD: **Quantitative analysis of ribosome binding sites in *E. coli*.** *Nucleic Acids Res* 1994, **22**:1287-1295.

35. Fickett JW: **Quantitative discrimination of MEF2 sites.** *Mol Cell Biol* 1996, **16**:437-441.

36. Fields DS, He Y-Y, Al-Uzri AY, Stormo GD: **Quantitative specificity of the Mnt repressor.** *J Mol Biol* 1997, **271**:178-194.

37. Tronche F, Ringeisen F, Blumenfeld M, Yaniv M, Pontoglio M: **Analysis of the distribution of binding sites for a tissue-specific transcription factor in the vertebrate genome.** *J Mol Biol* 1997, **266**:231-245.

38. • Audic S, Claverie J-M: **Visualizing the competitive recognition of TATA-boxes in vertebrate promoters.** *Trends Genet* 1998, **14**:10-11.
This study makes a clear case for local competition of sites being at least as important as binding affinity.

39. Frech K, Quandt K, Werner T: **Finding protein-binding sites in DNA sequences: the next generation.** *Trends Biochem Sci* 1997, **22**:103-104.

40. Claverie JM, Audic S: **The statistical significance of nucleotide position-weight matrix matches.** *Comput Appl Biosci* 1996, **12**:431-439.

41. Stamatoyannopoulos G, Nienuis AW: **Hemoglobin switching.** In *The Molecular Basis of Blood Diseases*. Edited by Stamatoyannopoulos G, Nienhuis AW, Majerus PW, Varmus H. Philadelphia: WB Saunders; 1994:107-155.

42. • Qin W, Khuchua Z, Cheng J, Boero J, Payne RM, Strauss AW: **Molecular characterization of the creatine kinases and some historical perspectives.** *Mol Cell Biochem* 1998, **184**:153-167.
The regulation of creatine kinase has been extensively studied by many groups. The complexity of the regulation makes a fine case study for understanding the sorts of overlaps and interactions that must be taken into account in computational modeling.

43. • Lander ES: **Array of hope.** *Nat Genet* 1999, **21**:3-4.
A succinct overview of the promise and limitations of cDNA microarray expression data.

44. Yang GP, Ross DT, Kuang WW, Brown PO, Weigel RJ: **Combining SSH and cDNA microarrays for rapid identification of differentially expressed genes.** *Nucleic Acids Res* 1999, **27**:1517-1523.

45. • Mironov AA, Koonin EV, Roytberg MA, Gelfand MS: **Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes.** *Nucleic Acids Res* 1999, **27**:2981-2989.
A group of similarly regulated genes in one bacterium is extended by iterated use of orthology with another species, operon structure, and similarity of operon promoter patterns. In the process, the key transcription factor binding specificities are refined.

46. Zhang MQ: **Large-scale gene expression data analysis: a new challenge to computational biologists.** *Genome Res* 1999, **9**:681-688.

47. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**:208-214.

48. •• Perier RC, Junier T, Bonnard C, Bucher P: **The eukoryotic promoter database (EPD): recent developments.** *Nucleic Acids Res* 1999, **27**:307-309.
The EPD is a carefully curated set of verified and annotated transcription start sites, cross referenced to the sequence databases.

49. Fickett JW, Hatzigeorgiou AG: **Eukaryotic promoter recognition.** *Genome Res* 1997, **7**:861-878.

50. Claverie JM, Sauvaget I: **Assessing the biological significance of primary structure consensus patterns using sequence databanks. I. Heat-shock and glucocorticoid control elements in eukaryotic promoters.** *Comput Appl Biosci* 1985, **1**:95-104.

51. Fondrat C, Kalogeropoulos A: **Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerivisiae*: application to chromosome III.** *Comput Appl Biosci* 1996, **12**:363-374.

52. Frech K, Danescu-Mayer J, Werner T: **A novel method to develop highly specific models for regulatory units detects a new LTR in GenBank which contains a functional promoter.** *J Mol Biol* 1997, **270**:674-687.

53. Crowley EM, Roeder K, Bina M: **A statistical model for locating regulatory regions in genomic DNA.** *J Mol Biol* 1997, **269**:8-14.

54. • Wagner A: **A computational "genome walk" technique to identify regulatory interactions in gene networks.** *Pac Symp Biocomput* 1998, 264-278.
A good introduction to the issues involved in trying to characterize, in a mathematically rigorous way, the clustering seen in most transcriptional modules.

55. Kel OV, Romaschenko AG, Kel AE, Wingender E, Kolchanov NA: **A compilation of composite regulatory elements affecting gene transcription in vertebrates.** *Nucleic Acids Res* 1995, **23**:4097-4103.

56. Fickett JW: **Coordinate positioning of MEF2 and myogenin binding sites.** *Gene* 1996, **172**:GC19-GC32.

57. Saroff HA, Kiefer JE: **Analysis of the binding of ligands to large numbers of sites: the binding of tryptophan to the 11 sites of the trp RNA-binding attenuation protein.** *Anal Biochem* 1997, **247**:138-142

58. Somia NV, Kafri T, Verma IM: **Piecing together more efficient gene expression.** *Nat Biotechnol* 1999, **17**:224-225.