Review

Integrating sequence, evolution and functional genomics in regulatory genomics

Martin Vingron*, Alvis Brazma†, Richard Coulson†, Jacques van Helden‡, Thomas Manke*, Kimmo Palin§, Olivier Sand‡ and Esko Ukkonen¶

Addresses: *Computational Molecular Biology, Max-Planck-Institut für molekulare Genetik, Ihnestrasse 73, D-14195 Berlin, Germany. †Microarray Group, European Bioinfomatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. ‡BiGRe - Université Libre de Bruxelles, Campus Plaine, Bvd du Triomphe - CP263, B-1050 Bruxelles, Belgium. §Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. ¶Helsinki Institute for Information Technology, Helsinki University of Technology and University of Helsinki, 00014 Helsinki, Finland.

Correspondence: Martin Vingron. Email: vingron@molgen.mpg.de

Published: 30 January 2009

*Genome **Biology*** 2009, **10:**202 (doi:10.1186/gb-2009-10-1-202)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/1/202

© 2009 BioMed Central Ltd

Abstract

With genome analysis expanding from the study of genes to the study of gene regulation, ‘regulatory genomics’ utilizes sequence information, evolution and functional genomics measurements to unravel how regulatory information is encoded in the genome.

Sequencing and functional genomics have not only led to a better understanding of genes and their expression collectively, but also to a refueling of interest in how transcriptional regulation is encoded in the ‘noncoding’ part of the genome. For many years, the state of the art had been to collect observed transcription factor binding sites (TFBSs) in DNA, use them to build a description of the factor’s binding motif, for example in the form of a positional weight matrix [1,2], and then scan a putative regulatory region for hits to this motif. Promoter-prediction programs were, and are still, used to linearly scan genomic sequence for putative markers of promoters, such as CpG islands and/or TATA box motifs, Inr, and so on [3,4]. After experimental advances, many more elements of this ‘linear’ code for transcription - such as histone positioning, histone modifications and DNA methylation - are currently being studied. Large-scale sequencing has enabled the comparative study of genomes, which in turn helps identify regulatory sequences. Functional genomics and, in particular, gene-expression data, is showing us the consequences of transcriptional activation and has propelled the quest to find regulatory sequences shared between coexpressed groups of genes. This review will attempt to summarize the past few years’ progress in integrating these approaches for the purpose of identifying regulatory sequence elements and their function.

Regulatory feature description with positional weight matrices

Positional weight matrices (PWMs) have for many years been the workhorse of TFBS annotation. A set of experimentally determined binding sites for a transcription factor is aligned and the distribution of bases in each position of the binding site yields the weights in the PWM. There are two major databases for eukaryotic PWMs: TRANSFAC [5] and JASPAR [6]. PWMs and their application are reviewed in [7]. Two operations in conjunction with PWMs are important. First, coming up with the alignment for the PWM may be non-trivial. For example, in a landmark paper, Bucher [8] derived a PWM for the TATA box by extracting the relevant sequence alignment from a set of promoters he had collected [9]. Algorithms like the one used are still being improved and will be briefly summarized in the section Motif discovery.

Second, because the PWM is meant to be a descriptor for a TFBS, a method is required to identify the predicted binding sites. Scanning a sequence with a PWM in a search of predicted binding sites seems simple. Yet because of the notorious lack of information content in the individual binding motif, a search over a long sequence region will inevitably turn up large numbers of probably false-positive results. Computationally, the program MATCH [10] uses predefined cutoffs for determination of binding sites, whereas patser [11], ProfileStats [12], and matrix-scan [13] determine the statistics of PWM matching under different background models. Weight-matrix based, biophysical models of transcription-factor binding [14] constitute an alternative to the *ad hoc* definition of a matching score. They allow the design of learning algorithms [15-17] and can be validated against experimental measurements [18].

When selecting the most appropriate algorithm for predicting binding sites, one needs to distinguish between the two application scenarios of either predicting target genes for a factor with a given PWM, or predicting which factors, represented by their PWMs, might bind upstream of a gene. For the first task, a biophysical approach like TRAP [18], which determines the likelihood that a factor binds somewhere in a, say, promoter region, appears to be most appropriate. After proper statistical normalization, this algorithm can also determine a ranking of which factors might bind in a given sequence region [19]. The program matrix-scan from the RSAT package [11] is most appropriate for determining actual binding sites for a factor, while also supporting the detection of regions enriched in putative binding sites.

The false-positive problem, however, is inherent and to remedy it more information is needed. Besides possibly looking for evolutionarily conserved binding sites, this is usually provided through better motif descriptions, combination of motifs into *cis*-regulatory modules, and by considering the evolutionary conservation of binding sites.

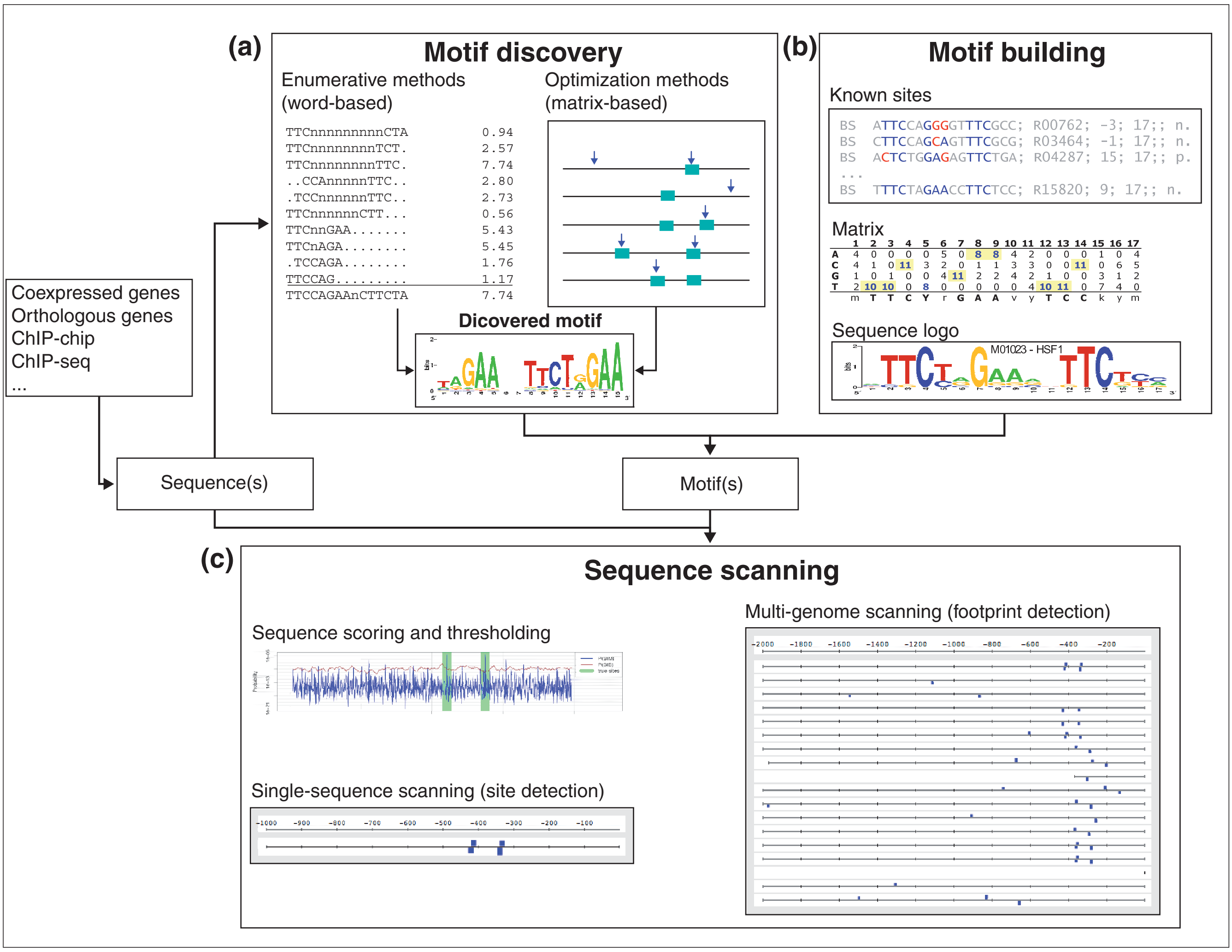
Phylogenetic footprinting recognizes regulatory motifs in evolution

The most powerful remedy for the many false-positive annotations is the presence of evolutionary conservation in noncoding regions across several genomes. Because of the selective pressure exerted on their regulatory function, *cis*-acting elements are likely to be more conserved than the surrounding noncoding sequences. The expression 'phylogenetic footprints' was proposed by Tagle *et al.* [20] to denote anciently conserved *cis*-regulatory elements. Today, phylogenetic footprinting refers to conserved regulatory patterns in orthologous genes, or in regions that are deemed orthologous. Duret and Bucher [21] pointed out the utility of evolutionary sequence conservation for the identification of regulatory elements. With the ever-increasing number of genome sequences, phylogenetic footprint detection seems to be the key approach to deciphering regulatory mechanisms. Its utility is twofold: evolutionary conservation helps in defining a regulatory pattern and helps in reinforcing binding-site predictions. Figure 1 illustrates the general approach from motif discovery to phylogenetic footprinting.

Complex features: *cis*-regulatory modules

In 1998, by aligning large genome fragments between mouse and human, Fickett and Wasserman observed that noncoding regions contain conserved fragments of a few kilobases that are enriched in *cis*-acting elements [22,23]. Inspection of aligned regulatory regions typically yields a picture where ungapped conserved elements (describable by a PWM) occur at slightly different spacing in a set of sequences. Such an element is frequently called a *cis*-regulatory module (CRM). An early application of the concept can be found in Gailus-Durner *et al.* [24], where promoters were characterized by their arrangement of binding sites. Again, the problem of identifying CRMs comes in different flavors. In a single sequence the clustering of predicted TFBS may lead to the definition of a CRM. This was shown in the work by Berman *et al.* [25], who identified distal enhancer elements in the genome of *Drosophila melanogaster* by locating a sequence window with a high number of TFBSs. Similar windowed counting methods have been used with good success on the analysis of *Drosophila* early development [26,27]. Improved probabilistic variants search clusters that are significant according to some statistical model [28-33]. A complementary approach uses probabilistic models that consider the likelihoods of binding sites and their distances [34-37].

The problem of finding CRMs from a single sequence is naturally extended to multiple sequences. The resulting methods work on the binding sites that lie on aligned DNA; hence they depend on the correctness of the alignments, which can be suspicious around the patchily conserved CRMs [38]. The multiple sequence methods can also be divided into combinatorial and probabilistic ones [37,39-42]. The enhancer element locator (EEL) [43,44] is a recent CRM prediction tool that takes an orthologous pair of genes from two organisms (say, human and mouse) and searches through the DNA flanking the two genes to locate conserved clusters of TFBSs. The binding sites that might belong to such clusters are computationally determined, using some given collection of PWMs. The clusters are evaluated and ranked using a scoring scheme that gives bonuses and penalties as implied by the underlying biochemically motivated model of CRM structure and evolution. EEL finds the clusters with highest total score using an algorithm that is similar to the Smith-Waterman algorithm for local alignments. The significance analysis of the EEL scores can be based on generalized Karlin-Altschul statistics for local alignment scores [45]. Entire genomes with up to 1 Mbp flanking regions for orthologous pairs of genes can be analyzed by EEL in reasonable time. Some EEL predictions have successfully been verified to have biological function in the mouse using *in situ* hybridizations and transgenic reporter assays [43,46]. The method of



```
TTCnnnnnnnnnCTA      0.94
TTCnnnnnnnnTCT.      2.57
TTCnnnnnnnnTTC.      7.74
..CCAnnnnnTTC..      2.80
.TCCnnnnnnTTC..      2.73
TTCnnnnnnCTT...      0.56
TTCnnGAA.......      5.43
TTCnAGA........      5.45
.TCCAGA........      1.76
TTCCAG.........      1.17
TTCCAGAAnCTTCTA      7.74
```

```
BS  ATTCCAGGGGTTTCGCC; R00762; -3; 17;; n.
BS  CTTCCAGCAGTTTCGCG; R03464; -1; 17;; n.
BS  ACTCTGGAGAGTTCTGA; R04287; 15; 17;; p.
...
BS  TTTCTAGAACCTTCTCC; R15820; 9; 17;; n.
```

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
A	4	0	0	0	0	5	0	8	8	4	2	0	0	0	1	0	4
C	4	1	0	11	3	2	0	1	1	3	3	0	0	11	0	6	5
G	1	0	1	0	0	4	11	2	2	4	2	1	0	0	3	1	2
T	2	10	10	0	8	0	0	0	0	0	4	10	11	0	7	4	0
	m	T	T	C	Y	r	G	A	A	v	y	T	C	C	k	y	m

Figure 1

Illustration of the flow of information in regulatory region annotation. Given a coexpressed group of genes, one can use **(a)** motif discovery or **(b)** search for known motifs in the upstream regions (motif building). Weight matrices can be used to scan sequences in a variety of ways: **(c)** the sites predicted by scanning the mouse *Hspa1b* promoter with the TRANSFAC matrix M01023 (left) and the conservation of those sites in vertebrate promoters (right).

Blanco *et al.* [47] aligns sequences of binding sites in much the same way as EEL, but their alignment is global instead of local. The global alignment is not able to locate novel CRMs from long sequences but requires exact knowledge about the location of the regulatory element. Another recent method [48] combines binding-site annotation and DNA alignment into a single procedure.

Grouping genes into expression clusters

Whereas phylogenetic footprinting aims at delineating conserved patterns in orthologous regulatory elements, another logic says that coexpressed groups of genes from one organism might share regulatory elements, which mediate the coexpression. Clearly, in a first step the coexpressed groups of genes need to be determined, typically from microarray-generated gene-expression data. Many methods for this purpose have been proposed, starting with the work of Eisen *et al.* [49], who used single linkage clustering. The literature on clustering is extensive, and specialized algorithms for gene-expression data have been proposed by Sharan and Shamir [50], and Tamayo *et al.* [51] among others. Graphical methods have also proved useful in identifying clusters and associations [52,53].

A number of databases are available to retrieve or submit microarray data. In particular, the National Center for Biotechnology Information (NCBI) runs the Gene Expression Omnibus (GEO) [54], the European Bioinformatics Institute (EBI) maintains the ArrayExpress (AE) database [55] and Stanford University hosts the Stanford Microarray Database (SMD) [56]. Based on AE, the ArrayExpress Warehouse [55] was established, allowing queries based on a range of gene annotations, including gene symbols, GO terms and disease associations. Coexpressed clusters can be defined by determining which expression profiles in an experiment are significantly correlated with a 'seed gene', which supplies a blueprint for the cluster.

The inclusion of protein links in AE opens up the possibility of identifying human and mouse genes encoding transcription factors. This allows the definition of a group of expression profiles that is not only coherently expressed but is also seeded by a transcription factor. Then, for the transcription factor, all the probe sets significantly correlated to its expression are pooled if they also exhibit significant differential expression for the same experimental factor. Probe sets present on the metazoan Affymetrix microarrays stored in the warehouse have been mapped to ENSEMBL gene entries along with their functional annotations, facilitating the linking of results from protein sequence searches to the expression data stored in the warehouse. An application of this mapping information is the identification of transcription factor expression profiles, initiating the generation of clusters of coexpressed genes.

Motif discovery

Whereas phylogenetic footprinting spots conserved, probably orthologous, patterns, a coexpressed group lets one ask whether the genes in the group are also co-regulated. Motifs that might be responsible for this assumed co-regulation can then be searched for. In one approach, no assumptions about known transcription factors binding in the promoter regions of the genes are made - motifs are searched *de novo*. An early example of such a method was used by Bucher [8]. Today, discovered motifs can *a posteriori* be compared with databases of known motifs, to check if they are likely to be bound by known transcription factors, or whether the motifs are completely novel. The motif-discovery problem can be addressed by various algorithmic approaches, which take as input a set of sequences and return *de novo* predicted motifs. The approaches can roughly be subdivided in two classes: matrix-based or string-based pattern discovery.

Matrix-based pattern discovery algorithms evaluate a large number of possible alignments between fragments of the input sequences, and attempt to return the alignment (summarized as a position-specific scoring matrix) that optimizes some scoring function. Historically, the first approach was developed by Gary Stormo's group: their program Consensus relies on a greedy algorithm, which progressively incorporates sequences to build matrices with maximal information content [11,57]. The Gibbs sampling strategy, initially developed to detect protein domains [58,59], has been adapted to discovery of transcription factor binding motifs in promoters of coexpressed genes [60-62]. More recent versions of the Gibbs sampling [63,64] support background models based on Markov chains, which take into account the higher-order dependencies between adjacent residues in biological sequences. The program MEME implements an expectation-maximization algorithm using multiple starting seeds in order to sample a large number of possible motifs [65,66]. String-based pattern discovery is based on the statistical detection of over-represented oligonucleotides [67-75] or of dyads - that is, spaced pairs of oligonucleotides [76]. Many of these algorithms have been compared and tested by Tompa *et al.* [77] and many have also been adapted to detect phylogenetic footprints in promoters of orthologous genes, with the programs Footprinter [78], PhyloCon [79], PhyME [80], OrthoMEME [81], PhyloGibbs [82] and Footprint-analysis [83]. A problem common to these approaches is the need to specify a theoretical background model, which usually does not fully capture the complexity and heterogeneity of real promoter sequences.

Where to search?

Patterns discovered among the regulatory regions of a coexpressed cluster of genes will typically be PWMs, like those derived from known TFBSs. The catch, however, is that inspection of PWMs for real binding sites has taught us not to expect well defined patterns with high information content. On the other hand, patterns as badly defined as the real ones can be easily extracted from any set of upstream regions if only the region chosen is large enough. This demonstrates an inherent limitation of the ability to identify patterns from co-regulated genes. Nevertheless, in yeast there has been considerable success with *de novo* identification of motifs [84]. Pattern identification in *Drosophila* has profited greatly from the many sequences that are now available in conjunction with fairly well defined enhancer regions made up of clearly discernible regulatory elements. Vertebrate regulatory sequences, however, seem to be much harder to identify. Difficulties stem from the lack of knowledge on how to narrow down the sequence regions in which to look for patterns, and probably from the patchy nature of vertebrate CRMs.

There are several ways of narrowing down the sequence regions to be searched for regulatory patterns. First, systematic identification of complete cDNAs together with new technologies such as cap-analysis gene expression (CAGE) tags have led to highly accurate identification of human and mouse transcription start sites [85]. Therefore, focusing on a promoter sequence is less guesswork today than it was a few years ago. One price to be paid, though, is the increased complexity of promoter definition resulting from the insight that alternative promoters for a gene are more the rule than the exception. It thus seems appropriate to study several promoters per gene. The notion of an enhancer used to be biologically defined, but with more and more complete genomic sequences available, enhancers tend to be identified with highly conserved noncoding regions. Systematic mapping of DNase I hypersensitive sites is also contributing to pinpointing enhancer regions. However, with the identification of transcriptional start sites that are far upstream from the translational start, it is becoming increasingly difficult to distinguish clearly between an enhancer and a promoter.

CRMs and coexpressed groups

Even when one knows where to look for regulatory modules, identification of a CRM that might be responsible for co-regulation of a group of genes is hard, and methods development is a very active field of research. In general, available methods put the emphasis either on *de novo* pattern discovery or on a dictionary-based approach. A dictionary of patterns may contain PWMs from the existing databases, patterns identified through systematic phylogenetic footprinting [86,87], or the output from *de novo* pattern identification. The dictionary serves to search for associations between sequence motifs and gene clusters. Often, this association is formalized as a statistical over-representation: that is, that the genes in the clusters contain a certain motif more often than expected.

The prototypic approach to over-representation is the use of the hypergeometric distribution to quantify the probability that within a large set two subsets have an overlap exceeding a certain size. Hughes *et al.* [62] apply this to the predicted target genes containing a certain motif and clusters of genes. Clearly, if the overlap between the two sets is large, this hints at a biological role for the motif. Dieterich *et al.* [88] analyze human cell-cycle data and quantify the occurrence of motifs upstream of genes that peak in particular phases of the cell cycle. Many more procedures try to solve the problem of finding novel CRMs similar to a dictionary of CRMs [89-92]. Good results have been achieved in flies, and to some extent in mammals [23,27,90,93,94]. Bussemaker [95] has pioneered linear models to explain expression data in terms of the occurrence of motifs in the upstream region of the genes. A recent paper by Bulyk and co-workers [96] relates over-represented CRMs to their target genes in human myogenic differentiation. Other algorithms identify common CRMs from a set of co-regulated genes without assuming a given dictionary [33,97-101].

Some methods find CRMs by distinguishing the regulatory DNA regions from neutrally evolved sequences and from sequences conserved for a reason other than transcriptional regulation [102-105]. These methods are universal in the sense that they do not need prior transcription factor binding information. The output provides the putative regulatory sequences but gives no clue of the transcription factors binding them. The tissue specificity of the regulators can be tested in the wet lab or predicted using other computational methods [106,107].

The availability of gene-expression data in conjunction with chromatin immunoprecipitation and DNA microarray (ChIP-chip) and sequence data has led to many attempts to construct entire gene regulatory networks, rather than predicting particular regulatory connections. These efforts are known as ‘reverse engineering’ of networks, but are beyond the scope of this review.

What are the eventual chances of producing systematic genomic annotation of regulatory elements in the human genome? Impressive progress is clearly being made. The ORegAnno database [108] collects binding-site information for many organisms and the Encode project [85] has provided a plethora of regulatory information, such as transcription start sites, DNase I hypersensitive sites, histone-modification data and more, much of which is available in the Ensembl Regulatory Build [109]. The extension of Encode to the entire human genome and to model organisms will provide significantly more of this kind of information. However, the obvious discrepancy between the large number of transcription factors and the comparatively small number of PWMs in the databases makes it clear that the best we can hope for at the moment is an annotation with regulatory elements. The actual binding factor may remain unknown, although the sequence element can be used as a predictor for expression in a certain tissue or under certain conditions. Extensive transcription factor binding experiments such as ChIP-chip, ChIP followed by DNA sequencing (ChIP-seq), protein-binding arrays, and DNA adenine methyltransferase tagging of TFBSs (DamID) will hopefully help in establishing better links between experimental data and computationally derived data. Likewise, the predicted CRMs constitute a huge resource of hypotheses waiting to be tested experimentally.

Acknowledgements

Support from EU FP6 NoE BioSapiens (contract number LSHG-CT-2003-503265) is gratefully acknowledged.

References

1. Staden R: **Methods for calculating the probabilities of finding patterns in sequences.** *Comput Appl Biosci* 1989, **5:**89-96.
2. Stormo GD: **DNA binding sites: representation and discovery**. *Bioinformatics* 2000, **16:**16-23.
3. Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL: **Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment.** *Genome Biol* 2006, **7(Suppl 1):**S3.1-13.
4. Smale ST, Kadonaga JT: **The RNA polymerase II core promoter.** *Annu Rev Biochem* 2003, **72:**449-479.
5. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Münch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles.** *Nucleic Acids Res* 2003, **31:** 374-378.
6. Vlieghe D, Sandelin A, De Bleser PJ, Vleminckx K, HertzWasserman WW, van Roy F, Lenhard B: **A new generation of JASPAR, the open-access repository for transcription factor binding site profiles.** *Nucleic Acids Res* 2006, **34(Database issue):**D95-D97.

7. Bulyk ML: **Computational prediction of transcription-factor binding site locations.** *Genome Biol* 2003, **5:**201.

8. 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.

9. Bucher P: **[Description of eukaryotic promoters in the EPD database].** *Mol Biol (Mosk)* 1997, **31:**616-625. In Russian.

10. Kel AE, Gossling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E: **MATCH: A tool for searching transcription factor binding sites in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3576-3579.

11. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15:**563-577.

12. Rahmann S, Muller T, Vingron M: **On the power of profiles for transcription factor binding site detection.** *Stat Appl Genet Mol Biol* 2003, **2:**Article7.

13. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J: **Using RSAT to scan genome sequences for transcription factor binding sites and *cis*-regulatory modules.** *Nat Protoc* 2008, **3:**1578-1588.

14. Gerland U, Moroz JD, Hwa T: **Physical constraints and functional characteristics of transcription factor-DNA interaction.** *Proc Natl Acad Sci USA* 2002, **99:**12015-12020.

15. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.** *Bioinformatics* 2006, **22:**e141-149.

16. Kinney JB, Tkacik G, Callan CG Jr: **Precise physical models of protein-DNA interaction from high-throughput data.** *Proc Natl Acad Sci USA* 2007, **104:**501-506.

17. Djordjevic M, Sengupta AM, Shraiman BI: **A biophysical approach to transcription factor binding site discovery.** *Genome Res* 2003, **13:**2381-2390.

18. Roider HG, Kanhere A, Manke T, Vingron M: **Predicting transcription factor affinities to DNA from a biophysical model.** *Bioinformatics* 2007, **23:**134-141.

19. Manke T, Roider HG, Vingron M: **Statistical modeling of transcription factor binding affinities predicts regulatory interactions.** *PLoS Comput Biol* 2008, **4:**e1000039.

20. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **Embryonic epsilon and gamma 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.

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

22. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE: **Human-mouse genome comparisons to locate regulatory sites.** *Nat Genet* 2000, **26:**225-228.

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

24. Gailus-Durner V, Scherf M, Werner T: **Experimental data of a single promoter can be used for *in silico* detection of genes with related regulation in the absence of sequence similarity.** *Mamm Genome* 2001, **12:**67-72.

25. Berman BP, Nibu Y, Pfeiffer BD, Tomancak P, Celniker SE, Levine M, Rubin GM, Eisen MB: **Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome**. *Proc Natl Acad Sci USA* 2002, **99:**757-762.

26. 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.

27. Halfon MS, Grad Y, Church GM, Michelson AM: **Computation-based discovery of related transcriptional regulatory modules and motifs using an experimentally validated combinatorial model.** *Genome Res* 2002, **12:**1019-1028.

28. Wagner A: **Genes regulated cooperatively by one or more transcription factors and their identification in whole eukaryotic genomes.** *Bioinformatics* 1999, **15:**776-784.

29. Rebeiz M, Reeves NL, Posakony JW: **SCORE: a computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation.** *Proc Natl Acad Sci USA* 2002, **99:**9888-9893.

30. Johansson O, Alkema W, Wasserman WW, Lagergren J: **Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.** *Bioinformatics* 2003, **19(Suppl 1):**i169-i176.

31. Bailey TL, Noble WS: **Searching for statistically significant regulatory modules.** *Bioinformatics* 2003, **19(Suppl 2):**ii16-ii25.

32. Sharan R, Ovcharenko I, Ben-Hur A, Karp RM: **CREME: a framework for identifying cis-regulatory modules in human-mouse conserved segments.** *Bioinformatics* 2003, **19(Suppl 1):**i283-i291.

33. Kreiman G: **Identification of sparsely distributed clusters of cis-regulatory elements in sets of co-expressed genes.** *Nucleic Acids Res* 2004, **32:**2889-2900.

34. Frith MC, Li MC, Weng Z: **Cluster-Buster: finding dense clusters of motifs in DNA sequences.** *Nucleic Acids Res* 2003, **31:**3666-3668.

35. Segal E, Sharan R: **A discriminative model for identifying spatial *cis*-regulatory modules.** *J Comput Biol* 2005, **12:**822-834.

36. 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, **3:**30.

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

38. Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM: **rVista for comparative sequence-based discovery of functional transcription factor binding sites.** *Genome Res* 2002, **12:**832-839.

39. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of *cis*-regulatory modules.** *Bioinformatics* 2003, **19(Suppl 2):**ii5-ii14.

40. Donaldson IJ, Chapman M, Gottgens B: **TFBScluster: a resource for the characterization of transcriptional regulatory networks.** *Bioinformatics* 2005, **21:**3058-3059.

41. Philippakis AA, He FS, Bulyk ML: **Modulefinder: a tool for computational discovery of cis regulatory modules.** *Pac Symp Biocomput* 2005, 519-530.

42. Blanchette M, Bataille AR, Chen X, Poitras C, Laganière J, Lefèbvre C, Deblois G, Giguère V, Ferretti V, Bergeron D, Coulombe B, Robert F: **Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression.** *Genome Res* 2006, **16:**656-668.

43. Hallikas O, Palin K, Sinjushina N, Rautiainen R, Partanen J, Ukkonen E, Taipale J: **Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity.** *Cell* 2006, **124:**47-59.

44. Palin K, Taipale J, Ukkonen E: **Locating potential enhancer elements by comparative genomics using the EEL software.** *Nat Protoc* 2006, **1:**368-374.

45. Karlin S, Altschul SF: **Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes.** *Proc Natl Acad Sci USA* 1990, **87:**2264-2268.

46. Vokes SA, Ji H, McCuine S, Tenzen T, Giles S, Zhong S, Longabaugh WJ, Davidson EH, Wong WH, McMahon AP: **Genomic characterization of Gli-activator targets in sonic hedgehog-mediated neural patterning.** *Development* 2007, **134:**1977-1989.

47. Blanco E, Messeguer X, Smith TF, Guigo R: **Transcription factor map alignment of promoter regions.** *PLoS Comput Biol* 2006, **2:**e49.

48. Bais AS, Grossmann S, Vingron M: **Simultaneous alignment and annotation of cis-regulatory regions.** *Bioinformatics* 2007, **23:**e44-49.

49. Eisen M, Spellman P, Brown P, Botstein D: **Cluster analysis and display of genome-wide expression patterns** *Proc Natl Acad Sci* USA 1998, **95:**14863-14868.

50. Sharan R, Maron-Katz A, Shamir R: **CLICK and EXPANDER: a system for clustering and visualizing gene expression data.** *Bioinformatics* 2003, **19:**1787-1799.

51. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation.** *Proc Natl Acad Sci USA* 1999, **96:**2907-2912.

52. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97:**10101-10106.

53. Fellenberg K, Hauser NC, Brors B, Neutzner A, Hoheisel JD, Vingron M: **Correspondence analysis applied to microarray data.** *Proc Natl Acad Sci USA* 2001, **98:**10781-10786.

54. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles - database and tools update.** *Nucleic Acids Res* 2007, **35(Database issue):**D760-D765.

55. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A:

ArrayExpress - a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 2007, **35(Database issue):**D747-D750.

56. Demeter J, Beauheim C, Gollub J, Hernandez-Boussard T, Jin H, Maier D, Matese JC, Nitzberg M, Wymore F, Zachariah ZK, Brown PO, Sherlock G, Ball CA: **The Stanford Microarray Database: implementation of new analysis tools and open source release of software.** *Nucleic Acids Res* 2007, **35(Database issue):**D766-D770.
57. Hertz GZ, Hartzell GW, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *CABIOS* 1990, **6:**81-92.
58. 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.
59. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4:**1618-1632.
60. Roth FP, Hughes JD, Estep PW, Church GM: **Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation.** *Nat Biotechnol* 1998, **16:** 939-945.
61. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM: **Systematic determination of genetic network architecture.** *Nat Genet* 1999, **22:**281-285.
62. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296:**1205-1214.
63. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P, Moreau Y: **A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling.** *Bioinformatics* 2001, **17:**1113-1122.
64. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Pac Symp Biocomput* 2001, 127-138.
65. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc Int Conf Intell Syst Mol Biol* 1994, **2:**28-36.
66. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** *Proc Int Conf Intell Syst Mol Biol* 1995, **3:**21-29.
67. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281:**827-842.
68. Schbath S, Prum B, de Turckheim E: **Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences.** *J Comput Biol* 1995, **2:**417-437.
69. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements in silico on a genomic scale.** *Genome Res* 1998, **8:** 1202-1215.
70. Brazma A, Jonassen I, Eidhammer I, Gilbert D: **Approaches to the automatic discovery of patterns in biosequences.** *J Comput Biol* 1998, **5:**279-305.
71. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9:**211-223.
72. Blanchette M, Schwikowski B, Tompa M: **An exact algorithm to identify motifs in orthologous sequences from multiple species.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**37-45.
73. Tompa M: **An exact method for finding short motifs in sequences, with application to the ribosome binding site problem.** *Proc Int Conf Intell Syst Mol Biol* 1999, 262-271.
74. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using a probabilistic segmentation model.** *Proc Int Conf Intell Syst Mol Biol* 2000, **8:**67-74.
75. Vanet A, Marsan L, Sagot MF: **Promoter sequences and algorithmical methods for identifying them.** *Res Microbiol* 1999, **150:**779-799.
76. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28:**1808-1818.
77. Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Régnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-144.
78. Blanchette M, Tompa M: **FootPrinter: A program designed for phylogenetic footprinting.** *Nucleic Acids Res* 2003, **31:**3840-3842.
79. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19:**2369-2380.
80. Sinha S, Blanchette M, Tompa M: **PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences.** *BMC Bioinformatics* 2004, **5:**170.
81. Prakash A, Blanchette M, Sinha S, Tompa M: **Motif discovery in heterogeneous sequence data.** *Pac Symp Biocomput* 2004, **9:**348-359.
82. Siddharthan R, Siggia ED, van Nimwegen E: **PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny.** *PLoS Comput Biol* 2005, **1:**e67.
83. Janky R, van Helden J: **Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution.** *BMC Bioinformatics* 2008, **9:**37.
84. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*.** *BMC Bioinformatics* 2006, **7:**113.
85. Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.** *Nature* 2007, **447:**799-816.
86. Xie X, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M: **Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals.** *Nature* 2005, **434:**338-345.
87. Xie X, Mikkelsen TS, Gnirke A, Lindblad-Toh K, Kellis M, Lander ES: **Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites.** *Proc Natl Acad Sci USA* 2007, **104:**7145-7150.
88. Dieterich C, Rahmann S, Vingron M: **Functional inference from non-random distributions of conserved predicted transcription factor binding sites.** *Bioinformatics* 2004, **20(Suppl 1):**i109-i115.
89. Crowley EM, Roeder K, Bina M: **A statistical model for locating regulatory regions in genomic DNA.** *J Mol Biol* 1997, **268:**8-14.
90. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res* 2001, **11:**1559-1566.
91. Aerts S, Thijs G, Coessens B, Staes M, Moreau Y, De Moor B: **Toucan: deciphering the cis-regulatory logic of coregulated genes.** *Nucleic Acids Res* 2003, **31:**1753-1764.
92. Ovcharenko I, Nobrega MA: **Identifying synonymous regulatory elements in vertebrate genomes.** *Nucleic Acids Res* 2005, **33(Web Server issue):**W403-W407.
93. 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.
94. Donaldson IJ, Chapman M, Kinston S, Landry JR, Knezevic K, Piltz S, Buckley N, Green AR, Gottgens B: **Genome-wide identification of cis-regulatory sequences controlling blood and endothelial development.** *Hum Mol Genet* 2005, **14:**595-601.
95. Bussemaker HJ, Foat BC, Ward LD: **Predictive modeling of genome-wide mRNA expression: from modules to molecules.** *Annu Rev Biophys Biomol Struct* 2007, **36:**329-347.
96. Warner JB, Philippakis AA, Jaeger SA, He FS, Lin J, Bulyk ML: **Systematic identification of mammalian regulatory motifs' target genes and functions.** *Nat Methods* 2008, **5:**347-353.
97. Zhou Q, Wong WH: **CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modeling.** *Proc Natl Acad Sci USA* 2004, **101:**12114-12119.
98. Thompson W, Palumbo MJ, Wasserman WW, Liu JS, Lawrence CE: **Decoding human regulatory circuits.** *Genome Res* 2004, **14:** 1967-1974.
99. Gupta M, Liu JS: **De novo cis-regulatory module elicitation for eukaryotic genomes.** *Proc Natl Acad Sci USA* 2005, **102:**7079-7084.
100. Zhu Z, Shendure J, Church GM: **Discovering functional transcription-factor combinations in the human cell cycle.** *Genome Res* 2005, **15:**848-855.
101. Donaldson IJ, Gottgens B: **CoMoDis: composite motif discovery in mammalian genomes.** *Nucleic Acids Res* 2007, **35:**e1.
102. 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.

103. Taylor J, Tyekucheva S, King DC, Hardison RC, Miller W, Chiaromonte F: **ESPERR: learning strong and weak signals in genomic sequence alignments to identify functional elements.** *Genome Res* 2006, **16:**1596-1604.
104. Pierstorff N, Bergman CM, Wiehe T: **Identifying cis-regulatory modules by combining comparative and compositional analysis of DNA.** *Bioinformatics* 2006, **22:**2858-2864.
105. Prabhakar S, Poulin F, Shoukry M, Afzal V, Rubin EM, Couronne O, Pennacchio LA: **Close sequence comparisons are sufficient to identify human cis-regulatory elements.** *Genome Res* 2006, **16:**855-863.
106. Pennacchio LA, Ahituv N, Moses AM, Prabhakar S, Nobrega MA, Shoukry M, Minovitsky S, Dubchak I, Holt A, Lewis KD, Plajzer-Frick I, Akiyama J, De Val S, Afzal V, Black BL, Couronne O, Eisen MB, Visel A, Rubin EM: **In vivo enhancer analysis of human conserved non-coding sequences.** *Nature* 2006, **444:**499-502.
107. Pennacchio LA, Loots GG, Nobrega MA, Ovcharenko I: **Predicting tissue-specific enhancers in the human genome.** *Genome Res* 2007, **17:**201-211.
108. Griffith OL, Montgomery SB, Bernier B, Chu B, Kasaian K, Aerts S, Mahony S, Sleumer MC, Bilenky M, Haeussler M, Griffith M, Gallo SM, Giardine B, Hooghe B, Van Loo P, Blanco E, Ticoll A, Lithwick S, Portales-Casamar E, Donaldson IJ, Robertson G, Wadelius C, De Bleser P, Vlieghe D, Halfon MS, Wasserman W, Hardison R, Bergman CM, Jones SJ; Open Regulatory Annotation Consortium: **ORegAnno: an open-access community-driven resource for regulatory annotation.** *Nucleic Acids Res* 2008, **36(Database issue):**D107-D113.
109. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, *et al.*: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36(Database issue):**D707-D714.