ON THE FUTILITY OF INTERPRETING OVER-REPRESENTATION OF MOTIFS IN GENOMIC SEQUENCES AS FUNCTIONAL SIGNALS

Nikola Stojanovic
Department of Computer Science and Engineering, University of Texas at Arlington
Arlington, TX 76019, USA

Keywords: Transcriptional control signals, DNA motifs, Regulatory modules.

Abstract: Locating signals for the initiation of gene expression in DNA sequences is an important unsolved problem in genetics. Over more than two decades researchers have applied a large variety of sophisticated computational techniques in order to address it, but only with moderate success. In this paper we investigate the reasons for the relatively poor performance of the current models, and outline some possible directions for future work in this field.

1 INTRODUCTION

Eukaryotic gene expression is regulated by a complex network of protein–DNA and protein–protein interactions. The prevailing opinion, corroborated by many studies, is that most of these interactions take place within a few hundred bases upstream from the transcription start site, although this is still somewhat controversial (Nelson et al., 2004). In addition, sites important for the regulation of genes have been found in introns and in downstream sequences, as well as at distant loci, such as the β–globin LCR (Hardison et al., 1997b). Promoter regions in yeast are characterized by multiple occurrences of the same binding motif (van Helden et al., 1998), and this is also the case with many genes from other species. At present, relatively little is known about genetic pathways and the mechanisms of gene co-expression, but this situation is rapidly changing, especially with the advances in microarray technology and protein–protein interaction studies. However, while these advances provide an insight into expression patterns and associations, they do not tell anything about the mechanisms driving them, nor about the sites in DNA responsible for their regulation.

Despite of significant efforts over the last twenty years to computationally predict transcription factor binding signals in promoter and other regions of the genome, this remains an elusive goal. While early approaches relied on a rather naive assumption that the target sites for protein binding must feature information content sufficient for them to be uniquely recognized among all non–sites (Schneider et al., 1986), disillusionment soon followed, as any attempt to isolate functional elements in DNA resulted in an enormous number of false positives. Learning from that experience, and further experimental evidence, the bioinformatics community has widely adopted a view that the motifs for transcription factor binding in functional regions are grouped in regulatory modules, sometimes featuring multiple copies of individual sites. This idea is not new (Ackers et al., 1982; Mehldau and Myers, 1993; Kel et al., 1995), however in the recent years there has been an explosion of computational algorithms designed in an attempt to identify such modules (Hu et al., 2000; GuhaThakurta and Stormo, 2001; Rebeiz et al., 2002; Eskin and Pevzner, 2002; Jegga et al., 2002; Johansson et al., 2003; Sharan et al., 2003; Sinha et al., 2003; Aerts et al., 2004; Donaldson et al., 2005; Kundaje et al., 2005; Pierstorff et al., 2006; Papatsenko, 2007; Schones et al., 2007), to list just a few. Some of the methods also relied on the assumption that multiple copies of the same motif should be a component of these modules (Qin et al., 2003; van Helden, 2004). If a particular motif is over-represented, i.e. if it occurs in a genomic segment or a group of segments more often than expected by chance, it was anticipated that it should indicate a functional signal. Moreover, if multiple motifs in close proximity satisfy this condition, it was presumed to be a strong indication of function. Software developed for the location of such modules generally relied on previous information about the individual binding sites forming the modules. The ap-

Stojanovic N. (2008).
ON THE FUTILITY OF INTERPRETING OVER-REPRESENTATION OF MOTIFS IN GENOMIC SEQUENCES AS FUNCTIONAL SIGNALS.
In *Proceedings of the First International Conference on Bio-inspired Systems and Signal Processing*, pages 464-471
DOI: 10.5220/0001061804640471

proaches were based on the phylogenetic conservation (Jegga et al., 2002; Sharan et al., 2003; Sinha et al., 2004; Dieterich et al., 2004; Donaldson et al., 2005; Pierstorff et al., 2006) of homologous regions or promoters, approximate matching to known motif sequences acquired from databases such as TRANSFAC (Matys et al., 2006) or some combination of both. The search was performed for the statistically significant clusters of motifs (Johansson et al., 2003; Alkema et al., 2004; Kundaje et al., 2005; Schones et al., 2007), and it was often combined with matching them to conserved regions in alignments. This was necessary in order to reduce the search space, but it proved inaccurate. A regulatory module can contain elements that have not been included in the original set, but the elements which did get included were sometimes spurious, at least when binding *in vivo* is concerned. Indeed, over years evaluation studies have been consistently demonstrating that these tools have not been very effective (Fickett and Hatzigeorgiou, 1997; Tompa et al., 2005), despite of the progress in our understanding of the genome, advances in technology and sophistication of the models.

Many approaches were based on gene expression study results, and the postulated co-regulation. Promoter regions of such genes were considered simultaneously, and the programs used Gibbs sampling (Lawrence et al., 1993; Thijs et al., 2002), Bayesian clustering (Qin et al., 2003), Markov Models (Liu et al., 2001), Expectation Maximization (Bailey and Elkan, 1994), Shannon's entropy (Kundaje et al., 2005), simultaneous dyad motif discovery (Eskin and Pevzner, 2002), genetic algorithms (Aerts et al., 2004) and other techniques in order to isolate regulatory modules. Despite of the use of very sophisticated algorithms these methods have not achieved desired accuracy. Why?

2 TARGETING THE OVER–REPRESENTED MOTIFS

The use of motif over-representation for the prediction of transcription factor binding signals can be roughly divided into three categories:

1. Over-representation of single motifs in groups of related functional sequences.
2. Over-representation of motifs from a limited set, such as these recorded in the databases of DNA regulatory elements, in a single region under consideration.
3. Over-representation of phylogenetically conserved blocks in a genomic segment of interest.

Combinations of the above approaches are widely applied. We shall look at each one individually.

2.1 Single Motifs in Groups of Sequences

If the motifs recognized by transcription factor proteins were specific (such as these recognized by restriction enzymes, for instance), the search for systematically present short signals would show promise. It is commonly accepted that a transcription factor binding site consists of 5 to 25 nucleotides, and most experimentally confirmed cores tend to be on the shorter side of that range. Considering just 5 characters, and assuming that most transcriptional regulatory activity indeed happens within about 500 bases upstream of the gene start, under the simplistic model of each DNA base being equally likely, the probability of a chance occurrence of such motif at any single position would be $1/4^5 \approx 0.00098$. Within a window of 500 bases the expected number would thus be around 0.49. Using the Poisson distribution we can estimate the probability of seeing it at least once in any given window to be $1 - e^{-0.49} \approx 0.39$. In consequence, if one would consider a set of just 4 *cis*–regulatory segments of co-expressed genes (determined by microarray experiments, for instance) in order to achieve statistical significance (i.e. a *p*–value of less than 0.05) and 5 to be highly significant (*p*–value < 0.01). This is encouraging, having in mind that the considered motifs are just 5 characters long, and that with 6 and more characters one can achieve statistical significance with regulatory sequences of just 2 co-expressed genes. If several motifs exhibit co-occurrence within a single set of regulatory sequences, that would almost certainly indicate a real signal, or at least a part of it (discarding for the moment the fact that such co-occurrences would also show up at many random places in the genome).

Even genes which are co-expressed under certain conditions may not be regulated in the same way. Their transcription initiation complexes may not be same, or even similar, or they may exhibit a weak similarity sufficient to yield co-expression only under certain circumstances. In addition, in any set of regulatory sequences any given motif may be absent from some, so the requirement that it should be found in all should be relaxed. Regardless of this, one can argue that when a set of regulatory sequences of co-expressed genes is available, one can determine the motifs unlikely to be shared by chance, and reliably identify at least these most common. Further studies can then be performed to identify proteins bound to these motifs, and their co-factors.

Unfortunately, nature does not follow simplistic models. Even as the core promoters lie upstream of the genes, most of their activity depends on the enhancer and other elements, which may be very far from the genes and regulating several of them simultaneously. In some cases, the co-expression pattern may stem from a group of genes affected by the *same* enhancer, rather than several enhancers featuring same motifs. Even when there are separate control elements targeted by same transcription factor proteins, and even if we assume that they would not function across domains, this expands the 500-base window to tens of thousands of bases, where only very long motifs would have a chance of achieving statistical significance.

Another problem is in that we may not even be able to detect a true binding signal present in all considered sequences. Transcription factors often feature a notorious lack of specificity, and within any given motif only certain positions, which need not be adjacent, may be important. The true transcription factor binding is determined by a very small number of bases, sometimes as small as 3. The use of position weight matrices (further referred to as PWMs) may be helpful in detecting these, but this method is far from perfect. Their biggest problem is in that they do not take into account the spatial structure of the motifs (such as positioning of the bases critical for binding within the major or minor grove of the DNA helix), which may be crucial in determining whether the specific nucleotide will interact with a protein or not. Alas, even the most recently published work, while allowing for non–contiguous critical residues, still fails to take into account anything but raw sequence information (Chakravarty et al., 2007).

2.2 Modules of Elements Retrieved from Databases

Researchers have spent many years meticulously collecting the experimental data concerning the binding of transcriptional proteins, and compiling the information about the bound motifs in databases such as TRANSFAC (Matys et al., 2006), Jaspar (Vlieghe et al., 2006) or Mapper (Marinescu et al., 2005). The consistency with which certain sequences are bound *in vitro* gives a strong support to the view that the exact nucleotide sequence is important, and that spatial and epigenetic factors may be more instrumental in blocking the sequences which are compositionally similar to the true binding targets, but which should not be used under the particular circumstances. Although it is still somewhat unclear how much of the binding effects *in vitro* would also happen *in vivo* (Jin et al., 2007), one can expect a reasonable correlation.

Concentrating on a motif recorded in a database rather than on any general one that may be repeated dramatically decreases the complexity of the search. In the extreme cases of long motifs with a strong consensus one can perform simple pattern matching and identify the targets uniquely in the genome. However, such motifs are not common, so the promise of this approach lies in the search for database motif clusters, i.e. regulatory modules. This still makes sense: TRANSFAC, the richest of the above mentioned resources, presently contains 7915 transcriptional binding sites, with consensus motifs organized into 398 position weight matrices. They come from different species, however one can use this number in rough calculations. Assuming the average length of a motif represented in a PWM to be around 9 (and for the moment discarding the fact that multiple motifs can match a single consensus) and the same random model as above, the number of possible motifs of this length would be $4^9 = 262144$. Consequently, one could estimate the probability that a motif from a set of 400 would start at any given position in the genome as $400/4^9 \approx 0.0015$. Within a window of 500 bases, putative regulatory region, the expected number of chance occurrences of a motif recorded in the database would roughly be around 0.76. Taking this number as the Poisson λ, one would need as few as 3 motifs recored in a database within a window of 500 bases (presumably serving as the anchoring for a regulatory module) in order to achieve statistical significance (p–value < 0.05). Such considerations have given rise to the creation of many software tools.

The first problem with this approach is that in a large genome such as human, even if we concentrate only on windows upstream of the known or predicted genes that would give us around 30 thousand regions, so with the p–value of 0.05 we would still get around 1500 false positive hits. Of course, for larger modules the p–values would be much lower, but one can hardly expect to locate very large clusters of sites, at least according to the current views on transcriptional regulation. If a module is shared among a few dozen regulatory sequences, and we would want to keep the specificity of the search at 0.5 or better, we would need to have the expected chance groupings at around, say, 50, which would dictate the p–value of 0.0017. Even under the above outlined simplified circumstances this would dictate literally dozens of motifs to participate in the module, forming a common core. Consequently, the poor performance of module searching software comes as no surprise.

The real–world situation is actually much worse: genomic sequences are not random assemblies of 4

letters, the regulatory module locations (moreover, locations that can be taken by individual participating motifs) are not limited to windows of length 500 immediately upstream of the genes, and many variants of a motif may match its consensus (as represented by the PWM). The currently available databases are neither complete nor accurate, and thresholds for matrix matching are set in a very *ad hoc*, heuristic fashion. In consequence, PWMs tend to match large groups of motifs, producing hits literally everywhere. Epigenetics phenomena may act in such fashion as to dramatically reduce the numbers of elements participating in a regulatory module, by making many instances of chance groupings resembling it inaccessible to transcriptional proteins, and many interactions within modules are taking place at the protein, not DNA level, further reducing the number of motifs in the genome that would need to be recognized in order to initiate transcription (and thus the size of the motif cluster corresponding to the module).

2.3 Phylogenetic Approaches

Another popular approach to identifying functional signals relies on phylogenetic conservation. Its basis is a very reasonable assumption that a functional constraint prevents mutations in DNA from becoming fixed in population, while sites which are not important are free to independently mutate and fix along separate branches of the evolutionary tree. This hypothesis has been amply confirmed by the study of coding sequences, and within them of the synonymous and non-synonymous substitutions. The encouraging results in the study of genes have led to the assumption that phylogenetic conservation can be exploited in the search for regulatory signals.

For this purpose, many investigators have turned attention to the identification of phylogenetic footprints, both in pair-wise sequence comparisons and multiple alignments. Studies have been performed in order to establish the most informative genetic distance between compared species, which have to be far apart so to minimize the noise coming from random conservation, but close enough to share similar regulatory signals (Hardison et al., 1997a; Miller, 2001), as well as the most informative additional species to place in a multiple alignment (Thomas et al., 2003). Pairwise, within the mammalian scope, sequences which have diverged about 70 million years ago (such as human and mouse) have shown greatest promise, although optimal phylogenetic distance for analysis tends to vary with the genomic locus (Hardison, 2000).

Even under the most favorable circumstances, when the effects of non-specific binding and permissible divergence in regulatory signal consensus, as well as these of inter–species differences, would be minimal, any short signal would not be sufficient to warrant significance, or it would require a multiple alignment of dozens of very close genomic sequences (Stojanovic, 2004). This is becoming feasible with bacterial, but not yet with eukaryotic genomes. Researchers have thus concentrated on the identification of clusters of conserved sites, guided by essentially the same reasoning as outlined in the previous sections. In relatively short segments of DNA it is unlikely that rearrangements would be taking place on a substantial scale, and the positional conservation of regulatory signals would lead to good alignments with short phylogenetic footprints clearly visible.

The probabilistic reasoning applied in this case relied on the strength of the signals (i.e. sequence conservation), the likelihood of seeing such conserved motif by chance, given the phylogenetic distance between the sequences, and, because the later is often difficult to establish, on the empirical determination of the background conservation within the alignment, as its sections which appear to stand out.

The first problem with this approach lies in the quality of the alignment itself: genetic regulatory signals are short and non-specific, and thus not very likely to be precisely positioned, although their relative offsets would probably be small. This, on one hand, may lead to an imprecise definition of motif boundaries, which often shows as only a partial overlap between the footprint and the experimentally confirmed binding site. On the other, the footprint itself may be difficult to identify, as its improperly aligned bases would both lower the signal and increase the neighboring region noise. In protein sequences one can at least partially exploit structural characteristics (such as α–helix signatures) in order to improve the alignment quality, but in DNA the only relatively reliable markers are the exons of genes. If one looks for their immediate upstream promoter regions this may be helpful, but unfortunately the 5' untranslated regions of variable lengths and weaker conservation (with some notable exceptions discussed below) tend to reduce the anchoring strength of the first exon.

Even when the alignment is reliable, the probability of random conservation in even distantly related sequences is too high to lend credibility to any but extremely large groupings of footprints, too large to be plausible anchor sites for the transcriptional complexes. Somewhat surprisingly, such large concentrations of footprints are not uncommon in higher eukaryotic genomes. In fact, many of these are so large that they can hardly be considered as groupings

of individual, discrete elements (Jones and Pevzner, 2006). The most dramatic example are the non-coding ultra-conserved segments, defined as blocks of 200 or more bases with absolute identity among all compared species. Within the human genome there are about 500 such blocks conserved among all sequenced mammals, but sometimes even among all vertebrates. The role of these elements is currently unknown, as many knock-out experiments have repeatedly failed to produce visible effects in model animals. Consequently, some researchers have postulated that the ultra-conservation (as well as conservation of other long non-coding blocks) may be a consequence of a regional repair mechanism of exceptional strength, but so far nobody was able to characterize what that mechanism might be, as well as why it would have been put in place at its target loci.

In order to quantify this phenomenon, we have looked at the patterns of conservation in mammalian *Hox* gene clusters (Stojanovic and Dewar, 2005), which are well preserved, and home to some of the mentioned ultra-conserved blocks. Interestingly, in *Hox* the highest overall conservation has been observed within the 5' UTR regions of genes, as illustrated in Table 1. While good conservation of untranslated regions is not common genome–wide, it has been observed in several other cases, such as mammalian casein genes (Rijnkels et al., 2003). In summary, this indicates that there is much more to phylogenetic conservation than a simple functional constraint. Before that mechanism is understood, some skepticism concerning the use of sequence conservation as a hallmark of a functional signal is warranted.

3 OVER-REPRESENTATION OF MOTIFS IN GENOMIC ENVIRONMENTS

The over-representation concept itself is problematic. It has been well known, and for a long time now, that genomic sequences, even in large "junk" areas, are not random assemblies of four letters. In order to quantify the genome-wide over-representation of short motifs, we have recently undertaken a systematic study (Singh et al., 2007) in which we have noted a remarkable over-representation of many short motifs throughout the presumably unique human genomic sequences, as well as (to a lesser extent), Markov model generated sequences trained on human chromosomes. As an example, the results counting the average number of repeated occurrences of motifs of lengths 4 through 9 measured in 6 datasets of 100 sequences of length 500 each are shown in Table 2. Our findings clearly indicated that, first, all genomic sequences feature dramatically higher numbers of repeated short motifs than one would expect by chance, and, second, that the differences in numbers of such motifs do not appear to be significant between random intergenic and presumably regulatory sequences upstream of the known genes, despite of the trend that one can notice in the last two columns of Table 2. Repeatedly, chi-square tests performed on these columns and other data could show only mild, but inconclusive, bias. This indicates that something else in addition to the functional signal is at play, but it is somewhat unclear what that might be.

In a series of studies started more than forty years ago (Waring and Britten, 1966) Britten, Davidson and others demonstrated that the nuclear genome of diverse eukaryotes contained a large fraction of repetitive DNA, and recent large–scale genome sequencing has established the ubiquitous existence of repeats. Many of them are of tandem nature, relatively easily recognizable, however the majority are the result of the repeated interspersed insertion of transposable elements, often not capable of further activity (Smit, 1999; Feschotte et al., 2002) — once integrated, these sequences will never transpose again and can be considered molecular fossils. Regardless of their origin and of the mechanisms responsible for their inactivation, it is widely accepted that fossilized transposons, as a whole, do not assume function to the host. Consequently, these inactive copies are progressively eroded by mutations accumulating at a neutral rate until they become unrecognizable. While more recent insertional events can be readily identified due to the high similarity of the copies, characterization of more ancient activity remains a challenge. In the human genome, almost half of the sequence is considered unique, but only a small fraction (about 5% of the total) is thought to be significant, whether coding or not. This leaves an open question about the origin and role of the presumably unique non-functional sequence, which is very likely to originate from ancient transpositions and duplications. Due to its degree of degeneracy, it would remain in the genomic segments under consideration after repeat masking, but it would also introduce a large number of seemingly over-represented motifs.

Therefore, many of the apparent clusters of conserved elements are likely just remnants of transposon insertions. While phylogeny–based approaches are less vulnerable to this effect, it can still be an issue when comparing sequences from species for which good repeat libraries have not yet been compiled. Regardless of the source, the micro-repetitive

Table 1: Fractions of the total number of *Hox* (*A, B, C* and *D* clusters) alignment columns in 7 distinct genomic environments contained in the regions of minimal length of 25 bp, of average conservation with p–value < 0.1 measured against the background conservation of the entire alignment. The intergenic data for *HoxD* have been parenthesized because of the Ensembl gene prediction at the location where many of these regions have been found. Overall, *HoxD* data are not as reliable because only a relatively small amount of high–quality sequence of this cluster was available in all considered species (human, baboon, mouse, rat, cow and pig) at the time of the study.

	500–1000bp 5'	200–500bp 5'	0–200bp 5'	Coding	Introns	0–1000bp 3'	Intergenic
HoxA	0.067	0.315	0.616	0.223	0.066	0.077	0.057
HoxB	0.115	0.342	0.788	0.639	0.071	0.145	0.024
HoxC	0.104	0.202	0.609	0.521	0.089	0.105	0.035
HoxD	0	0	0	0.061	0.026	0.066	(0.027)

Table 2: The mean numbers of repeated patterns of different lengths in different types of nucleotide sequences. Pattern counting has been done over 100 sequences of length 500 in each category.

Pattern Length	Expected Number	Random Synthetic	2^{nd} Order Markov M.	3^{rd} Order Markov M.	5^{th} Order Markov M.	Random Genomic	Upstream Regulatory
4	429.06	425.74	437.99	432.84	432.23	438.97	433.92
5	193.16	189.18	237.83	222.98	222.27	261.64	260.11
6	57.46	55.16	84.33	74.58	75.88	106.62	115.31
7	15.03	14.0	24.5	21.82	23.3	38.66	47.54
8	3.8	3.12	7.05	5.75	6.87	15.72	21.3
9	0.95	0.56	1.94	1.47	1.97	8.57	11.33

structure of genomic sequences of higher eukaryotes makes it very difficult to locate any feature through over-representation, simply because the background is highly non-random.

4 DISCUSSION

So far much of the computational search for genomic regulatory signals have been done using sequence information only, just because it was the most readily available. In the context of sequence analysis looking for statistical over–representation was indeed the most sensible approach. However, while the resulting combinatorial and probabilistic problems are challenging and mathematically interesting, biologically they are questionable. That does not mean that they are of no value whatsoever, only that they are currently not being used in the right way.

Much of the genome study is still in the data collecting phase. We are not yet in a position to build analytical models, and without them the quantification of their effects makes little sense. Over the last few years the scientific community has been increasingly turning attention to epigenetics, and there has recently been a significant increase in the accumulated knowledge about these phenomena. However, the computational community have so far mostly ignored these developments. At this time the study of binding signals in DNA should probably rely more on data mining approaches than on analytical models, although statistical analysis of the data will remain important.

When studying a potential regulatory role of a genomic sequence (or a group of sequences, in cases when co–regulation pattern of a group of genes is suspected), one should take into account, first of all, the specific experimentally confirmed knowledge about the region which can be mined from the literature using currently available technologies. Next, the specific biochemical information about methylation patterns and domain structure should be applied, before raw nucleotide information is considered. At the later stage the prediction and statistical evaluation should be incorporated, but structural data should still be taken into account, when available. Recent studies (Segal et al., 2006; Ioshikhes et al., 2006) have indicated that there may be specific histone proteins positioning codes in DNA, and if further evidence confirms this it would greatly help in the characterization of binding signals for other types of proteins, transcription factors in particular (through easier identification of potentially open chromatin domains). Only at this point one can concentrate on the motif–related considerations, looking for these recorded in databases and these that might be phy-

logenetically conserved. The over–representation *per se* may not be sufficient to provide useful information, but the appearance of similar motifs in areas otherwise postulated to share functionality (based on stronger evidence than just a correlation of expression in microarray experiments) may be indicative enough to warrant confidence.

The true discovery has always been through a well–coordinated combination of computational and experimental approaches. This takes time, although modern technologies are dramatically facilitating such efforts (Jin et al., 2007), so purely computational methods for genome–wide prediction of transcriptional regulatory signals will remain to be of interest. It is only that the methods will have to change in order to be really useful, and not just interesting.

ACKNOWLEDGEMENTS

The author would like to thank Cedric Feschotte of UTA Biology for useful discussions about the nature of DNA repeats, and Subhrangsu Mandal of UTA Biochemistry for the insights concerning epigenetic phenomena. Abanish Singh and David Levine of UTA Computer Science have provided computational infrastructure which generated data leading to our conclusions. This work has been partially supported by NIH grant 1R03LM009033-01A1.

REFERENCES

Ackers, G., A.D.Johnson, and M.A.Shea (1982). Quantitative model for gene regulation by lambda phage repressor. *Proc. Natl. Acad. Sci. USA*, 79:11291133.

Aerts, S., Van Loo, P., Moreau, Y., and De Moor, B. (2004). A genetic algorithm for the detection of new cis–regulatory modules in sets of co-regulated genes. *Bioinformatics.*, 20:1974–1976.

Alkema, W., Johansson, O., Lagergren, J., and Wasserman, W. (2004). MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.*, 32:W195–W198.

Bailey, T. and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36. AAAI Press.

Chakravarty, A., Carlson, J. M., Khetani, R. S., DeZiel, C. E., and Gross, R. H. (2007). SPACER: identification of *cis*–regulatory elements with non–contiguous critical residues. *Bioinformatics*, 23:1029–1031.

Dieterich, C., Rahmann, S., and Vingron, M. (2004). Functional inference from non-random distributions of conserved predicted transcription factor binding sites. *Bioinformatics.*, 20:i109–i115.

Donaldson, I. J., Chapman, M., and Gottgens, B. (2005). TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics*, 21:3058–3059.

Eskin, E. and Pevzner, P. A. (2002). Finding composite regulatory patterns in DNA sequences. *Bioinformatics*, 18(S1):S354–S363.

Feschotte, C., Jiang, N., and Wessler, S. (2002). Plant transposable elements: where genetics meets genomics. *Nat. Rev. Genet.*, 3:329–341.

Fickett, J. and Hatzigeorgiou, A. (1997). Eukaryotic promoter recognition. *Genome Res.*, 7:861–878.

GuhaThakurta, D. and Stormo, G. (2001). Identifying target sites for cooperatively binding factors. *Bioinformatics*, 17:608–621.

Hardison, R., Oeltjen, J., and Miller, W. (1997a). Long human–mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, 7:959–966.

Hardison, R., Slightom, J., Gumucio, D., Goodman, M., Stojanovic, N., and Miller, W. (1997b). Locus control regions of mamalian β–globin gene clusters: combining phylogenetic analyses and experimental results to gain functional insights. *Gene*, 205:73–94.

Hardison, R. C. (2000). Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet.*, 16:369–372.

Hu, Y., Sandmeyer, S., McLaughlin, C., and Kibler, D. (2000). Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics*, 16:222–232.

Ioshikhes, I. P., Albert, I., Zanton, S. J., and Pugh, B. F. (2006). Nucleosome positions predicted through comparative genomicsenetic. *Nature Genetics*, 38:1210–1215.

Jegga, A., Sherwood, S., Carman, J., Pinski, A., Phillips, J., Pestian, J., and Aronow, B. (2002). Detection and visualization of compositionally similar *cis*–regulatory element clusters in orthologous and coordinatelly controlled genes. *Genome Res.*, 12:1408–1417.

Jin, V. X., O'Geen, H., Iyengar, S., Green, R., and Farnham, P. J. (2007). Identification of an OCT4 and SRY regulatory module using integrated computational and experimental genomics approaches. *Genome Res.*, 17:807–817.

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

Jones, N. C. and Pevzner, P. A. (2006). Comparative genomics reveals unusually long motifs in mammalian genomes. *Bioinformatics*, 22:e236–e242.

Kel, O., Romaschenko, A., Kel, A., Wingender, E., and Kolchanov, N. (1995). A compilation of composite regulatory elements affecting gene transcription in vertebrates. *Nucleic Acids Res.*, 23:4097–4103.

Kundaje, A., Middendorf, M., Gao, F., Wiggins, C., and Leslie, C. (2005). Combining sequence and time series expression data to learn transcriptional modules. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2:194–202.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs Sampling strategy for multiple alignment. *Science*, 262:208–214.

Liu, X., Brutlag, D., and Liu, J. (2001). Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. In *Pac. Symp. Biocomput.*, pages 127–138.

Marinescu, V., Kohane, I., and Riva, A. (2005). The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res.*, 33D:D91–D97.

Matys, V., Kel–Margoulis, O.V., Fricke, E. *et al.* (2006). TRANSFAC®and its module TRANSCompel®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.*, 34:D108–D110.

Mehldau, G. and Myers, G. (1993). A system for pattern matching applications on biosequences. *Comput. Appl. Biosci.*, 9:299–314.

Miller, W. (2001). Comparison of genomic DNA sequences: solved and unsolved problems. *Bioinformatics*, 17:391–397.

Nelson, C., Hersh, B., and Carroll, S. B. (2004). The regulatory content of intergenic DNA shapes genome architecture. *Genome Biol.*, 5:R25.

Papatsenko, D. (2007). ClusterDraw web server: a tool to identify and visualize clusters of binding motifs for transcription factors. *Bioinformatics*, 23:1032–1034.

Pierstorff, N., Bergman, C. M., and Wiehe, T. (2006). Identifying *cis*–regulatory modules by combining comparative and compositional analysis of DNA. *Bioinformatics*, 22:2858–2864.

Qin, Z., McCue, L., Thompson, W., Mayerhofer, L., Lawrence, C., and Liu, J. (2003). Identification of coregulated genes through Bayesian clustering of predicted regulatory binding sites. *Nature Biotechnology*, 21(4):435–439.

Rebeiz, M., Reeves, N. L., and Posakony, J. W. (2002). SCORE: A computational approach to the identification of *cis*–regulatory modules and target genes in whole–genome sequence data. *Proc. Natl. Acad. Sci. USA*, 99(15):9888–9893.

Rijnkels, M., Elnitski, L., Miller, W., and Rosen, J. M. (2003). Multispecies comparative analysis of a mammalian–specific genomic domain encoding secretory proteins. *Genomics*, 82:417–432.

Schneider, T. D., Stormo, G. D., Gold, L., and Ehrenfeucht, A. (1986). Information content of binding sites on nucleotide sequences. *J. Mol. Biol.*, 188:415–431.

Schones, D. E., Smith, A. D., and Zhang, M. Q. (2007). Statistical significance of cis-regulatory modules. *BMC Bioinformatics*, 8:19.

Segal, E., Fondufe–Mittendorf, Y., Chen, L., Thastrom, A., Field, Y., Moore, I. K., Wang, J.-P. Z., and Widom, J. (2006). A genomic code for nucleosome positioning. *Nature*, 442:772–778.

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

Singh, A., Feschotte, C., and Stojanovic, N. (2007). A study of the repetitive structure and distribution of short motifs in human genomic sequences. *Int. J. Bioinformatics Research and Applications*, 3:523–535.

Sinha, S., Schroeder, M., Unnerstall, U., Gaul, U., and Siggia, E. (2004). Cross–species comparison significantly improves genome–wide prediction of *cis*–regulatory modules in *drosophila*. *BMC Bioinformatics.*, 5:129.

Sinha, S., vanNimwegen, E., and Siggia, E. (2003). A probabilistic method to detect regulatory modules. *Bioinformatics*, 19:i292–i301.

Smit, A. (1999). Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, 9:657–663.

Stojanovic, N. (2004). Computational methods for the analysis of differential conservation in groups of similar DNA sequences. *Genome Informatics*, 15:21–30.

Stojanovic, N. and Dewar, K. (2005). A probabilistic approach to the assessment of phylogenetic conservation in mammalian *Hox* gene clusters. In *Proceedings of the BIOINFO 2005, International Joint Conference of InCoB, AASBi and KSBI*, pages 118–123.

Thijs, G., Marchal, K., Lescot, M., Rombauts, S., De Moor, B., Rouze, P., and Moreau, Y. (2002). A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J. Comput. Biol.*, 9(2):447–464.

Thomas, J.W., Touchman, J.W., Blakesley, R.W. *et al.* (2003). Comparative analysis of multi-species sequences from targeted genomic regions. *Nature*, 424:788–793.

Tompa, M., Li, N., and Bailey, T.L. *et al.* (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*, 23(1):137–144.

van Helden, J. (2004). Metrics for comparing regulatory sequences on the basis of pattern counts. *Bioinformatics*, 20:399–406.

van Helden, J., Andre, B., and Collado-Vides, J. (1998). Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.*, 281:827–842.

Vlieghe, D., Sandelin, A., De Bleser, P. J., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of JASPAR, the open–access repository for transcription factor binding site profiles. *Nucleic Acids Res.*, 34:D95–D97.

Waring, M. and Britten, R. (1966). Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science*, 154:791–794.