# The state of the art of mammalian promoter recognition

*Thomas Werner*

Date received (in revised form): 9th December 2002

**Thomas Werner**
has a Masters in Chemistry and a PhD in Biochemistry from LMU Munich, as well as a Dr habil. in genetics from Technische Universität Munich. Bioinformatics has been his focus from 1986, particularly functional genomics and transcriptional regulation. He is currently CEO and CSO of Genomatix software GmbH.

## Abstract

The draft sequences of whole genomes are being published at an ever-increasing pace, thus providing access to the human genomic sequence and, more recently, the mouse sequence. Genomes of the invertebrates are also becoming available. Now that the genomic DNA of mammalian species is available, an old problem can be tackled with renewed vigour: mammalian promoter prediction. Gene promoters have proved elusive for more than a decade, despite their pivotal role in gene regulation. Recently, however, several new developments have made it possible to make meaningful large-scale predictions. This paper reviews the methods used for the prediction of mammalian, mostly human, promoters.

***Keywords:*** *promoter finding, promoter analysis, genome analysis*

## INTRODUCTION

The analysis of transcription control, ie the coordination of gene transcription in time and space, is probably of similar importance as proteome analysis. Life in a cell will be understood only when both the fate and actions of proteins as well as how and why they come into existence can be detailed – which is what transcription control is about in the first place. Yet there is still a vast bias in published work towards proteomics rather than transcription control (which will be referred to as regulanomics from here on). For example, a simple PubMed search for 'protein' yields more than ten times the amount of matches as searching for 'transcription'.

Until recently, one of the main reasons for this was the difficulty of finding mammalian promoters in genomic sequences. First of all, a promoter is still not a clearly defined unit. The region upstream of, and containing, the transcription start site (TSS) that is required for the basic events of transcriptional initiation may be referred to as the proximal promoter.[1] The crucial obstacle in finding mammalian promoters is that they usually do not share extensive sequence similarity even when they are functionally correlated, which prevents detection by sequence similarity-based search methods such as BLAST or FastA.[1]

Mammalian promoters can be seen as miniature structures of coding regions with few functional elements (exons) interspersed in a larger sequence of no known function (introns). The promoter 'exons' would be resembled by transcriptional control elements (usually transcription factor (TF) binding sites) while the so-far uncharacterised spacers in between those elements would correspond to promoter 'introns'. TF binding sites are only about a dozen nucleotides in length and even these small stretches are quite variable (Figure 1).

Thus, it becomes clear that overall sequence similarity in promoters is not a general phenomenon, although it does exist in the form of phylogenetic footprints.[2] Promoters contain the transcription start site and therefore always overlap with the first exon of a gene. This would allow promoters to be located by looking upstream of the first exon in genomic sequences. However, mammalian promoters are not readily available by this kind of mRNA analysis via cDNAs. Most cDNAs are truncated at

Thomas Werner,
Genomatix Software GmbH,
Landsberger Strasse 6,
D-80339, München, Germany,
and
Institute of Experimental Genetics,
GSF-Research Center for
Environment and Health GmbH,
Ingolstädter Landstrasse 1,
D-85764 Neuherberg, Germany

E-mail: Werner@genomatix.de

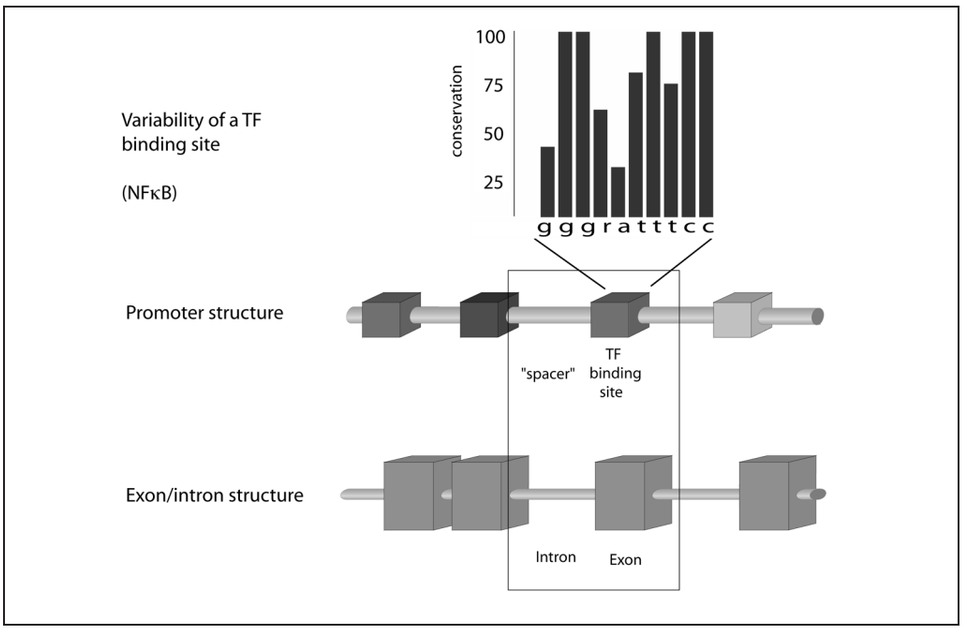


**Figure 1:** The centre shows a schematic promoter structure and below this is a schematic exon/intron structure with a rectangle highlighting the corresponding portions in both structures. On top of the promoter structure a IUPAC (ambiguity code) representation of a NFκB binding site is shown below a bar graph, indicating the different conservation of the individual positions of the binding site

**Incomplete cDNAs**

the 5′-end as they are traditionally amplified starting at their 3′-end, which may result in the promoter being several kbp out of reach (even 5 or 10 base pairs, bp, missing at the 5′-end may cause this if there is a large first intron involved). Therefore, promoter location was almost impossible in the past, making bioinformatics methods very attractive.

Approaches remained largely unsuccessful almost to the end of the twentieth century despite considerable efforts by many groups. Most methods would work only in small regions of DNA and even there with an unacceptable high amount of false positive matches as reviewed in 1997 by Fickett and Hatzigeorgiou.[3] For example, most of the methods came up with one match every 1,000 bp even under favourable conditions (much higher promoter density in the test sets than in the genome). Why did all that early work fair so miserably when it came to specificity (the average was less than 10 per cent in these tests)? One important reason may

**Unspecific methods**

have been some discrepancies in the nature of experimental results available for training. While much has been learnt about various elements that proved to be crucial for promoter function, such as a whole collection of transcription factor binding sites (TF binding sites), in particular the CAAT and the TATA box, publications dealing with functional promoter structures were scarce. The amount of proven promoters for training was also more than limited. The Eukaryotic Promoter Database (EPD) by Philipp Bucher was the only reliable source for a long time, and contained only about 1,200 promoters, including a strong bias towards 'favourite' genes, ie genes studied by more people than other genes.[4] Disappointingly, so-called promoter elements could be found all over virtually every sequence but none was really consistent, meaning that approaches to discriminate promoters from other sequences in general, based on such elements, were almost useless. The only exception was the first promoter

**Promoter structure models**

prediction program published, Dan Prestridge's Promoter Scan, which used frequency profiles for IUPAC sequences representing transcription factor binding sites and reached a specificity or better selectivity of up to 70 per cent.[5] Unfortunately, the sensitivity was very low. Although the program produced few false positives it missed the vast majority of true promoters.[3]

During the 1990s it was noted that promoters contain specific subsets of TF-sites[6–8] and at least in some cases a combination of several binding sites was required for biological function.[6–10] The good news was that this explained why TF-sites could be found everywhere without functionally interfering with normal transcription control. On the other side it also became clear that these arrangements were specific for very small numbers of promoters and could not be generalised in any way. Therefore, it became possible to describe functional groups of promoters in great detail by bioinformatics, but no general search tool could be developed on this basis.[11–13] With the imminent completion of the human genome draft sequence this inability to predict promoters became a real obstacle to genome-wide analysis of gene regulation by bioinformatics. After all, only about 2 per cent of the genomic sequence was found to be coding[14] and a similar amount of sequence can be expected to represent promoters.

**New prediction methods**

Fortunately, starting aptly in the year 2000, a series of new approaches broke through the roadblock imposed by unacceptable high rates of false positives, providing us with easy access to a large amount of genomic promoters.

Table 1 gives a list of web sites of promoter prediction programs and Table 2 gives a list of web sites for whole gene annotation.

## A NEW GENERATION OF METHODS FOR PROMOTER PREDICTION

There are still many different approaches to attack the problem of promoter recognition and this review will focus on attempts to locate promoters in whole genomes. There was also considerable progress in defining subset-specific or at least associated patterns (eg TF-sites). However, this is a more specialised application and will therefore be mentioned only briefly. A recent review has dealt in more detail with recent advances in pattern finding.[15] Another review giving the history of *in silico* pattern was published by Gary Stormo.[16]

The new generation of promoter predictors appears to be several times better than all previous approaches (see the comparisons[17–20]). This is very encouraging but raises another problem. With the rate of predictions falling well below 1 match in 10,000 bp or even 50,000 bp, determination of specificity becomes an almost impossible task. In order to calculate specificity both the number of true positives (TP) as well as the number of true negatives (TN) need to be known for a sufficiently large test set. For example the formulae used by Larsen *et al*.[21] to calculate sensitivity and specificity are as follows:

$$\text{Sensitivity} = TP/(TP + FN)$$

$$\text{Specificity} = TN/(TN + FP)$$

(where TP = true positives, TN = true negatives, FN = false negatives, FP =

**Table 1:** Web sites of promoter prediction programs

| | |
|---|---|
| PromoterInspector online use (registration page) | http://www.genomatix.de/cgi-bin/promoterinspector/promoterinspector.pl |
| DRAGON Promoter Finder online use | http://sdmc.krdl.org.sg/promoter/promoter1_3/DPFV13.htm |
| Eponine online use | http://servlet.sanger.ac.uk:8080/eponine/ |
| FirstEF online use | http://rulai.cshl.org/tools/FirstEF/ |
| CONPRO online use | http://stl.bioinformatics.med.umich.edu/conpro/ |

**Table 2:** Web sites for whole genome annotation

| | | |
|---|---|---|
| ElDorado | http://www.genomatix.de/free_login.html | Registration site, includes promoter annotation for human and mouse genomes |
| ENSEMBL | http://www.ensembl.org/ | Includes genomes human, mouse, zebrafish, fugu and mosquito |
| UCSC genome browser | http://genome.ucsc.edu/ | Includes human and mouse genomes |
| VISTA genome browser | http://pipeline.lbl.gov/vistabrowser/ | Includes human and mouse genomes |

false positives). This is by no means the only way to calculate characteristics, but all methods require the knowledge of the values for TP, TN, FN and FP.

For the new generation of methods several million base pairs with perfect annotation would be required. Unfortunately, our current knowledge of the human or mouse genome is not enough to allow discrimination between false positives and additional unknown TPs, which also makes the number of TNs inaccessible. Therefore, a less demanding property would be *selectivity*, defined as the ratio of the amount of true positives (total matches that could be correlated with known genes) to additional matches to give at least an idea about the relative performance of the programs. However, as programs work differently with respect to strand orientation and predicted property (region versus transcription start site, TSS) comparison becomes difficult. Since the authors also used different methods and data sets to access and compare their methods, it is virtually impossible to derive a meaningful and fair comparison of performance from the published data at this point.

**Specificity/selectivity**

The best way to compare programs would be to take the genome or a part of it, such as one or more whole chromosomes, and estimate the selectivity and the sensitivity as described above. Of course, this will give no idea of additional matches not conflicting with existing annotation, which might be false positives or real new promoters. Unfortunately, this is also impossible for all the programs based on published results, thus only some published data that explain what kind of problems are associated with the various approaches are referred to.

There are several ways one could categorise the methods. This review takes a genome-oriented approach and differentiates two major classes of approaches: one class of programs that attempts promoter prediction or localisation in whole anonymous genomic sequences and another class that takes advantage of genome annotation or other means to limit the actual search space for promoter finding. Technical criteria such as the basic models used are not applied since only overall results are really important. However, only programs that allow a genome-wide analysis are referred to.

The first category does not require any kind of *a priori* information about the sequence to be analysed except that it should be a mammalian genome (most programs have been trained on human sequences). The advantage is that such programs can be applied to genomic sequences as they appear, with no need to wait for gene annotation. This also avoids propagation of errors in the annotation process.

## PROGRAMS WORKING ON WHOLE ANONYMOUS GENOMES

The first method of this category, which happens to be also the first representative of the new generation, was PromoterInspector.[17] This method is based on a content analysis of promoter features represented by IUPAC-strings rather than specific transcription elements, and predicts regions containing a promoter. No strand orientation or TSS position is determined, which poses a problem in comparing this with programs that predict promoters directly. The

method was initially reported at a specificity of 43 per cent,[17] and is now claimed to have a specificity of 85 per cent based on an artificial genomic sequence composed from annotated EMBL entries.[22] Full analysis of the human genome was carried out, which is available through the ElDorado system. However, owing to access restriction of the free subscription, the whole set of promoters cannot be accessed at once. The data for chromosome 22 were published and indicate a sensitivity of 45 per cent.[23]

**Whole genome analysis**

The second method in this category was Dragon Promoter Finder, which is based on similar ideas as PromoterInspector but predicts strand-specific TSS.[18] The advance in predictive capabilities is in part attributed to the use of five different promoter models by the authors. Dragon Promoter Finder allows several levels of sensitivity, leaving it up to the user to choose the amount of false positive matches to be tolerated. Dragon Promoter Finder was also tested on whole human chromosome 22 and apparently also works with a high selectivity. The direct comparison of the program with PromoterInspector is complicated by the difference in strand prediction, which the authors compensated for by counting every match of PromoterInspector automatically as false positive (when the promoter was correctly predicted, otherwise one match equals two false positives) to account for the missing strand orientation.

**Gene specific analysis**

The third program, Eponine, belongs half and half to both categories, since it can analyse anonymous sequences such as whole human chromosome 22 in principle, but was applied to a pseudo chromosome, which included only regions around known genes.[19] As their comparison shows, not only does Eponine appear very selective but PromoterInspector also fares much better than on the whole chromosome. This indicates a general bias in favour of programs that was already observed when the short test sets of Fickett and Hatzigeorgiou[3] were replaced by longer genomic regions.[17]

## PROGRAMS WORKING ON A RESTRICTED SEARCH SPACE

There are two more programs taking advantage of existing annotation to restrict the search space to upstream regions of a few kilobases. The FirstExonFinder utilises various discriminatory functions including recognition of the first splice site (intron 1) to predict transcription start sites and has been applied to the 15 kb sequences upstream of known genes on chromosomes 21 and 22, using the approximate position of the gene start and the strand orientation of the genes to restrict the search space.[20] The method apparently works quite well for the known genes on chromosomes 21 and 22. However, it is unclear how the authors came up with the whole genome analysis claiming the existence of about 68,000 genes, since for most of these genes no information about gene start and orientation is available (but is required for the reported specificity to be reached). This was not clear from the original publication but has been clarified in a correction published on the author's web site.[24]

Another method in this category is ConPro, which analyses one gene at a time (at least in the web version) but is not restricted to go after all genes in principle.[25] The methods relies on a consensus formation of five promoter prediction programs previously reviewed by Fickett and Hatzigeorgiou,[3] all of which individually produce large amounts of false positives (~1 in 1,000 bp). By restricting the search space and forming the consensus of the methods, the authors claim to have been able to predict about 14,000 promoters in the genome, 6,400 of which correspond to well-characterised genes. As the authors include only a maximum of 1.5 kb upstream sequences, the relatively low number of true positives is no great surprise as first introns

happen to be several kilobytes on average, which puts many promoters of even slightly truncated genes out of reach for this method.

**CpG island finding**

Hannenhalli and Levy[26] published an approach based on analysis of regions around CpG islands regions for a few selected TF-sites, inferring that the combination of the various parameters indicates promoters. They concluded that generally only CpG island-associated promoters could be detected, as they found the other parameters to have little influence on the overall decision. This is in contrast to most of the other methods discussed here, which were able to predict non-CpG island-correlated promoters, albeit less efficiently. Their method bears some similarity to a previous published CpG island finder and cannot be compared to the general methods of the previous category.[27]

## MISCELLANEOUS PROGRAMS

There is a third category of programs that do not directly attempt to predict promoters but some properties, which might be also useful for promoter prediction and/or analysis. Therefore, those methods are summarised here.

**Phylogenetic comparisons**

The program rVISTA takes advantage of phylogenetic conservation of functional binding sites in regulatory regions between human and mouse sequences (any other pair of genomic sequences would work as well as long as there is a clear phylogenetic link of the corresponding genes[28]). The authors show that in their test case, a cytokine cluster, they can successfully weed out most of the unspecific matches and retain functional binding sites. This approach can be applied to any phylogenetic pair of regulatory regions, not just promoters, which is why this approach can be regarded as a logical next step, once promoters were actually located by another method. Another method that can be used through the web is the TraFaC program by Jegga *et al.*[29] The authors follow the idea that similar regions (promoters as well as enhancers) may contain a similar composition of TF binding sites, not necessarily in a conserved order. Results are depicted as regulograms to be interpreted by the user.

Levitsky *et al.*[30] published a system that attempts to calculate various properties of genomic DNA, among them the potential to form nucleosomal complexes, which they claim to be elevated in tissue-specifically expressed promoters. Given the enormous amount of chromatin remodelling involved in transcriptional activation, this finding should be taken with care. A potential for nucleosomal structures *per se* may well too weak to predict promoters. However, again after promoters have been located this might be a good tool to assess properties of such promoters.

## CONCLUSIONS

Although several new methods have been published in a relatively short period of time, it is difficult to assess their respective value for the user, as this depends very much on the individual problems to be solved. To obtain an initial annotation of whole genomes PromoterInspector and Dragon Promoter Finder should be the first choice. However, it is important that such promoters are put into the genomic context afterwards. Promoters of known species such as human or mouse should always be considered in the genomic annotation context. Currently, ElDorado is the most complete system offering such combined information, but the popular genome browsers can be expected to follow this lead within the next two years.

The programs in the second category require *a priori* knowledge about genes but offer a few benefits in return. They can afford to be more sensitive owing to the more restricted search space and some yield additional information about the promoter, so they might be a good choice to find promoters of known genes, which are missed by the general approaches.

Although this review focuses on promoter recognition by *in silico* methods, the growing amount of experimentally

**Experimental TSS determination**

determined TSS deserves mentioning. In particular, oligo-capping has been used to analyse thousands of cDNAs[31] in order to map the true 5′-end of the mRNA which would be located right inside the promoter as promoters extend also 3′ of the TSS.[1] Two problems with that approach remain. The first is that because of technical restrictions, not all mapped 5′-ends are the real TSS of the genes since the whole procedure has an overall efficiency of about 70–80 per cent. This means that the experimental approach yields about as many false positives as the *in silico* methods. Therefore, an experimentally mapped TSS is not always indicating a promoter. However if such a mapped TSS is located within a reasonable range (less than 2 kb) with a predicted and/or a mapped gene start from gene finding, or cDNA mapping, chances are much better that this will represent a real promoter. This already highlights the second problem. As oligo-capping identifies only short, presumably 5′, regions of mRNAs correctly, mapping of the results onto the genome is not a trivial task. Precise mapping definitely requires more sophisticated tools than BLAST and only carefully mapped data will be useful.

**Promoter/gene prediction**

The final question that remains is whether *in silico* promoter recognition was just a short interlude that would be superseded by genome annotation based on experimental results. This may well be the case for the human genome as enormous efforts have been made to complete the annotation. However, this is still continuing two years after the first draft was published and is not expected to be finished for another two years. In the mean time, *in silico* promoter recognition will remain as a valuable additional source of information. As the genomes of more and more mammalian and other species become sequenced, the amount of time for annotation can be expected to be reduced. However, complete *in silico* promoter annotation of a mammalian genome can be done ahead of any other annotation in a matter of days on relatively inexpensive computer systems. These approaches will also help to define non-coding RNA genes, suggest alternative transcription start sites and aid in the definition of new genes that bear no resemblance to any known genes. Therefore, it is expected that such *in silico* approaches will remain important tools for the analysis of new genomic sequences. Of course, gene prediction (ie largely coding region prediction) and promoter prediction can complement each other as they usually do not rely on the same features, as noted above.[23]

Another important topic that gains more importance is comparative genomics. As discussed above, such comparative approaches are also applicable to promoters.[28] However, since, in particular, first non-coding exons and intron sequences are often much less conserved than coding exons, *in silico* promoter recognition will be important to ensure that promoter sequences are being compared, since promoters also show very limited sequence conservation in general.[1]

In summary, the field of *in silico* promoter recognition has seen some dramatic improvements during recent years. Further improvements, especially in the sensitivity of methods, can be expected in the near future, placing those tools among the most powerful instruments for *in silico* analysis of genome sequences. Combinations with gene prediction methods have already been successfully applied.[17,25] Therefore, a more general improvement of genome annotation by promoter recognition can be expected.

Even a perfect method for promoter recognition would not put the groups working on promoter recognition out of a job. There are several major tasks still awaiting solutions in the field of *in silico* analysis of gene regulation. For example, there is nothing like an enhancer finder that could be applied to genomic sequences and the whole field of epigenetic gene regulation is completely off limits for bioinformatics analysis at

present (eg methylation, histone acetylation/deacetylation, chromatin remodelling in general, for reviews see Harju *et al*.[32] and Horn and Peterson[33]). Fortunately, there is considerable progress in the experimental analysis of these phenomena, which might soon provide the basis for bioinformatics approaches as well.

**Chromatin structure**

Further improvements in DNA structure prediction, especially some simple chromatin structural elements such as nucleosomal or solenoid structures, can be expected to boost bioinformatics. At the moment, the three-dimensional structures involved in gene regulation are completely ignored in bioinformatics approaches, as they do not yet yield to theoretical analysis in a useful way. However, this field is also making progress, which might open new opportunities for bioinformatics.[34]

The analysis of transcriptional regulation on a genomic level will remain the crucial factor in understanding life on a molecular basis and will help put the wealth of knowledge gained from proteomics into the right context. *In silico* promoter recognition will remain an important contributor to this goal.

## *References*


1. Werner, T. (1999), 'Models for prediction and recognition of eukaryotic promoters', *Mamm. Genome*, Vol. 10(2), pp. 168–175.
2. Wasserman, W. W. and Fickett, J. W. (1998), 'Identification of regulatory regions which confer muscle-specific gene expression', *J. Mol. Biol.*, Vol. 278, pp. 167–181.
3. Fickett, J. W. and Hatzigeorgiou, A. C. (1997), 'Eukaryotic promoter recognition', *Genome Res.*, Vol. 7, pp. 861–878.
4. Praz, V. *et al.* (2002), 'The Eukaryotic Promoter Database, EPD: New entry types and links to gene expression data', *Nucleic Acids Res.*, Vol. 30, pp. 322–324.
5. Prestridge, D. S. (1995), 'Predicting Pol II promotor sequences using transcription factor binding sites', *J. Mol. Biol.*, Vol. 249, pp. 923–932.
6. Kel, A. *et al.* (1999), 'Recognition of NFATp/AP-1 composite elements within genes induced upon the activation of immune cells', *J. Mol. Biol.*, Vol. 288, pp. 353–376.
7. Klingenhoff, A. *et al.* (1999), 'Functional promoter modules can be detected by formal models independent of overall nucleoside sequence similarity', *Bioinformatics*, Vol. 15, pp. 180–186.
8. Frech, K. *et al.* (1997), '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.*, Vol. 270, pp. 674–687.
9. Fickett, J. W. (1996), 'Coordinate positioning of MEF2 and myogenin binding sites', *Gene*, Vol. 172, pp. GC19–GC32.
10. GuhaThakurta, D. and Stormo, G. D. (2001), 'Identifying target sites for cooperatively binding factors', *Bioinformatics*, Vol. 17(7), pp. 608–621.
11. Frech, K. *et al.* (1996), 'Common modular structure of lentivirus LTRs', *Virology*, Vol. 224, pp. 256–267.
12. Kel, A. E. *et al.* (2001), 'Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors', *J. Mol. Biol.*, Vol. 309(1), pp. 99–120.
13. Frech, K. *et al.* (1998), 'Muscle actin genes: A first step towards computational classification of tissue specific promoters', *In Silico Biol.*, Vol. 1(1), pp. 29–38 and Vol. 1(4), pp. 372–380.
14. Dunham, I. *et al.* (1999), 'The DNA sequence of human chromosome 22', *Nature*, Vol. 402, pp. 489–495.
15. Ohler, U. and Niemann, H. (2001), 'Identification and analysis of eukaryotic promoters: Recent computational approaches', *Trends Genet.*, Vol. 17(2), pp. 56–60.
16. Stormo, G. D. (2000), 'DNA binding sites: representation and discovery', *Bioinformatics*, Vol. 16(1), pp. 16–123.
17. Scherf, M. *et al.* (2000), 'Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach', *J. Mol. Biol.*, Vol. 297(3), pp. 599–606.
18. Bajic, V. B. *et al.* (2002), 'Dragon Promoter Finder: Recognition of vertebrate RNA polymerase II promoters', *Bioinformatics*, Vol. 18(1), pp. 198–199.
19. Down, T. A. and Hubbard, T. J. (2002), 'Computational detection and location of transcription start sites in mammalian genomic DNA', *Genome Res.*, Vol. 12(3), pp. 458–461.
20. Davuluri, R. V. *et al.* (2001), 'Computational identification of promoters and first exons in the human genome', *Nat Genet.*, Vol. 29(4), pp. 412–417.
21. Larsen, N. I. *et al.* (1995), 'Analysis of eukaryotic promoter sequences reveals a systematically occurring CT-signal', *Nucleic Acids Res.*, Vol. 23, pp. 1223–1230.

22. Werner, T. (2002), 'Finding and decrypting of promoters contributes to the elucidation of gene function', *In Silico Biol.*, Vol. 2, pp. 1–7.

23. Scherf, M. *et al.* (2001), 'First pass annotation of promoters on human chromosome 22', *Genome Res.*, Vol. 11(3), pp. 333–340.

24. URL: http://rulai.cshl.org/tools/FirstEF/Cc/cc.html

25. Liu, R. and States, D. J. (2002), 'Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling', *Genome Res.*, Vol. 12(3), pp. 462–469.

26. Hannenhalli, S. and Levy, S. (2001), 'Promoter prediction in the human genome', *Bioinformatics*, Vol. 17, Suppl. 1, pp. S90–S96.

27. Ioshikhes, I. P. and Zhang, M. Q. (2000), 'Large-scale human promoter mapping using CpG islands', *Nat Genet.*, Vol. 26(1), pp. 61–63.

28. Loots, G. G. *et al.* (2002), 'rVista for comparative sequence-based discovery of functional transcription factor binding sites', *Genome Res.*, Vol. 12(5), pp. 832–839.

29. Jegga, A. G. *et al.* (2002), 'Detection and visualization of compositionally similar *cis*-regulatory element clusters in orthologous and coordinately controlled genes', *Genome Res.*, Vol. 12(9), pp. 1408–1417.

30. Levitsky, V. G. *et al.* (2001), 'Nucleosome formation potential of eukaryotic DNA: calculation and promoters analysis', *Bioinformatics*, Vol. 17(11), pp. 998–1010.

31. Suzuki, Y. *et al.* (2001), 'Diverse transcriptional initiation revealed by fine, large-scale mapping of mRNA start sites', *EMBO Rep.*, Vol. 2(5), pp. 388–393.

32. Harju, S. *et al.* (2002), 'Chromatin structure and control of beta-like globin gene switching', *Exp. Biol. Med. (Maywood)*, Vol. 227(9), pp. 683–700.

33. Horn, P. J. and Peterson, C. L. (2002), 'Molecular biology. Chromatin higher order folding–wrapping up transcription', *Science*, Vol. 297(5588), pp. 1824–1827.

34. Lafontaine, I. and Lavery, R. (1999), 'Collective variable modelling of nucleic acids', *Current Opin. Struct. Biol.*, Vol. 9(2), pp. 170–176.