

Chapter 4

In silico Identification of Eukaryotic Promoters

Venkata Rajesh Yella and Manju Bansal

Abstract The identification of promoters is essential for complete annotation of genomes and better understanding of gene regulatory networks. Experimental methods for promoter identification are costly, time-consuming and labor intensive. Hence, *in silico* methods are an attractive alternative. Computational methods for promoter prediction methods are easy, fast and can provide reliable results. A promoter prediction algorithm identifies promoter regions based on the idea that, promoter regions are different from other genomic regions in their features (sequence, context and structure). Promoter prediction algorithms are broadly classified as *ab initio*, hybrid and homology-based, depending on the information used for model design. The different approaches used in promoter prediction are briefly described here.

Keywords Promoter prediction programs · FirstEF · CpGProD · Eponine · PromoterInspector · PromPredict · EP3 · PromH

4.1 Introduction

Recent advances in genome sequencing techniques have provided a wealth of base sequence information, from which the coding and regulatory sequences need to be identified. While experimental as well as *in silico* tools are available for identifying coding sequences, locating regulatory sequences like promoters is a great challenge and the currently available methods are not very efficient. Promoter identification is essential for several reasons: annotating genomic regions for understanding genome architecture and understanding gene regulatory networks. Promoters are identified on the whole genome scale, using experimental techniques like binding assays, ChiP-chip, ChiP-seq, etc, which are costly, labor intensive and time consuming. Hence,

M. Bansal (✉) · V. R. Yella
Molecular Biophysics Unit, Indian Institute of Science,
Bengaluru, Karnataka, India
e-mail: mb@mbu.iisc.ernet.in

V. R. Yella
e-mail: yvrajesh@mbu.iisc.ernet.in

it may not be feasible to characterize all genomes in detail experimentally. Alternatively, computational methods are available to identify promoters, as well as coding regions. There are several Promoter Prediction Programs (PPPs) available, which use different features or statistical models and identify either transcription start sites (TSSs) or promoter regions. In this chapter, we briefly describe the architecture of Eukaryotic promoters and the different kinds of promoter prediction algorithms currently available.

4.2 Eukaryotic Promoter Architecture

A promoter region is generally defined as any genomic DNA where the transcription machinery assembles and initiates transcription. The promoter region consists of protein binding regions along with the transcription start site (TSS). Promoter architecture in Prokaryotes and Eukaryotes differs in complexity. In Prokaryotes, a single RNA polymerase transcribes all types of RNAs and the promoter regions are characterized by the presence of -35 and -10 elements and in some cases the UP element as well. Overall, in the Prokaryotes, the regulatory region is located within 100 base pairs relative to the TSS. In Eukaryotes, promoter structure is more complex, with the complexity increasing from single celled yeast to mammals. Eukaryotes have several different types of RNA polymerases (usually three), with each one responsible for the production of different subsets of RNA. RNA polymerase II is responsible for synthesis of all mRNAs and is well studied compared to other RNA polymerases. Hence, only features corresponding to promoters of genes transcribed by RNA polymerase II are discussed below.

In Eukaryotes, the promoter regions are broadly classified as core promoters, proximal promoters and distal promoters. The core promoter region, where the actual basal transcription machinery assembles, is 30–100 nucleotides in length. These regions are characterized by the presence of sequence motifs such as the TATA box and the Inr element. They may also contain downstream elements like DPE, MTE (in humans) along with the associated TSS (Juven-Gershon et al. 2008; Thomas and Chiang 2006). The proximal promoter regions are the sequences located within 500 base pairs relative to the TSS and contain certain proximal promoter elements, which include the GC box, the CAAT box, *cis*-regulatory modules (CRM) (Lenhard and Sandelin 2012), etc. Distal promoter elements include enhancers, insulators and silencers. The distal promoter region does not have a well-defined length and can extend up to 10 kb from the TSS in upstream as well as downstream regions. Distal promoters interact with transcription activators to increase the rate of transcription. In vertebrates, it is known that 5 % of the genes code for specific transcription activators, which interact with proximal and distal promoter regions.

Along with the transcription factor binding elements, mammalian promoter regions also contain CpG islands. In humans, it is known that 60 % of promoters belong to the CpG island-containing class. Figure 4.1 shows a schematic representation of different promoter elements and their activators in Eukaryotes. Recent studies have

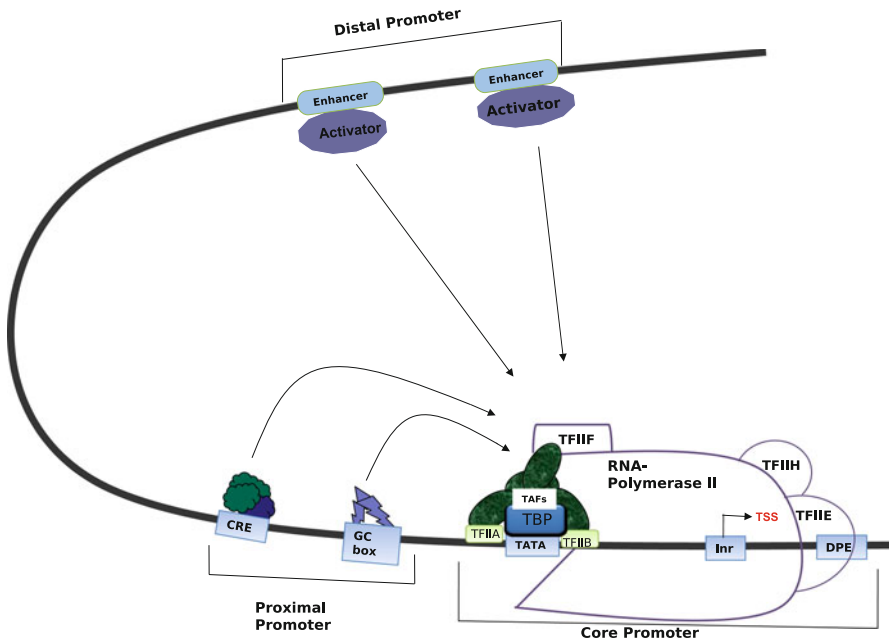


Fig. 4.1 A schematic representation of Eukaryotic RNA polymerase II promoter elements and basal transcription machinery. Promoter regions are divided into three classes, namely, *core promoters*, *proximal promoters* and *distal promoters*. Core promoter elements bind to basal transcription factors like TFIID. Proximal and distal promoter elements bind to transcription activators and increase the rate of transcription

shown that in Eukaryotes, especially in humans, each promoter is associated with many TSSs, which are spread over 50–100 nucleotides (referred to as transcriptionally active regions) (Carninci et al. 2006). Promoters can also be bidirectional (Xu et al. 2009). For detailed reviews on Eukaryotic promoters refer to Juven-Gershon et al. (2008), Lenhard and Sandelin (2012), Sandelin et al. (2007), Thomas and Chiang (2006). Recent understanding of vertebrate promoters is that though promoters differ in their motif content (with most of them lacking a consensus motifs), GC content (with lower Eukaryotes being AT rich and mammals being GC rich), some properties such as nucleosome free region and epigenetic features around TSSs are quite common (Valen and Sandelin 2011).

4.3 Experimental Methods of Promoter Identification

Experimental methods for promoter identification and characterization generally identify TSSs or DNA sequences that bind to proteins such as TFs and RNAPII (Lenhard and Sandelin 2012; Sandelin et al. 2007). Earlier methods such as nuclease protection and primer extension carry out promoter identification on a gene-by-gene

basis and cannot be used for whole genome promoter identification. Current high-throughput methods measure either products from transcription (mRNA) or promoter activity in whole genome. They provide a snapshot of all transcribed regions or DNA-protein interactions in the genome for given experimental conditions. Recent advancements in promoter region identification consist of sequencing methods and hybridization methods (Sandelin et al. 2007). Sequencing methods such as RACE, 5'-tag sequencing and 5'-3' paired-end sequencing provide information about the mRNA or cDNA sequences. All these methods use reverse transcription to get cDNA. Then the cDNA is fragmented and the fragments amplified and sequenced from the 5'-end. The sequenced fragments are mapped to the genomic DNA sequence to get information about TSS location. Hybridization methods, instead of sequencing, use short oligonucleotides to hybridize with target DNA. Two widely used methods are tiling arrays and ChiP-chip, which characterize TSSs and promoter elements respectively. Oligonucleotide tiling arrays are designed with parts of contiguous regions of sequenced genome or some times even whole genomes. They can provide information about the whole transcriptome along with the location of TSSs. The ChiP-chip method is an application of tiling arrays to identify protein bound regions of genomic DNA. ChiP-chip method uses chromatin immunoprecipitation (ChiP) to isolate DNA-bound promoter-associated proteins and then bound DNA is identified using tiling arrays (Sandelin et al. 2007).

4.4 *In silico* Methods for Promoter Identification

The computational methods for identification of promoter regions are mostly based on the basic premise that promoter regions have distinct sequences when compared to other genomic regions. Promoter Prediction Programs (PPPs) use experimentally identified promoter regions aligned with respect to TSSs, or transcription factor binding site information from databases (TRANSFAC (Wingender et al. 2000), EPD (Schmid et al. 2004) and DBTSS (Suzuki et al. 2002)) as a training dataset, to derive principles that differentiate promoters from non-promoter regions. PPPs can be broadly classified into three types based on the information used for promoter characterization. They are *ab initio*, hybrid and homology based algorithms.

Ab initio or *de novo* methods use only DNA sequence information for promoter identification. *Ab initio* methods are further classified (as shown in Fig. 4.2) as search-by-signal, search-by-content and search-by-structure algorithms based on features used for modeling (Zeng et al. 2009). Some current algorithms integrate two or more features for efficient promoter prediction.

Hybrid methods use sequence information with other accessory information such as epigenetic features, nucleosome occupancy and gene expression data. Homology based PPPs use orthologous gene information to identify promoter elements. Here, we will focus on *ab initio* PPPs in detail and also provide an introduction to other methods. Detailed information on the history, feature selection, model design and performance assessment of these PPPs is available in several excellent reviews (Abeel

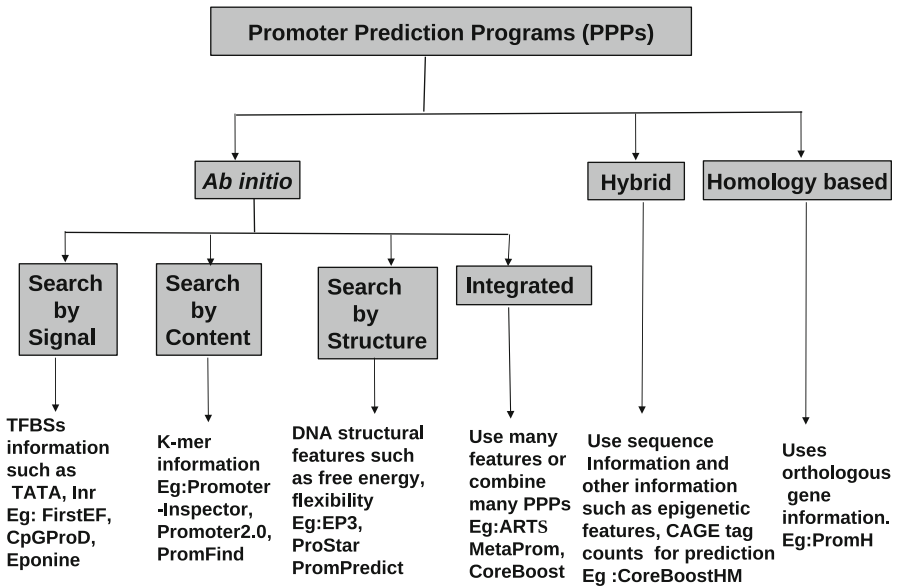


Fig. 4.2 Classification of Promoter Prediction Programs (PPPs) based on the information used for prediction

et al. 2009; Bajic et al. 2004; Bajic et al. 2006; Fickett and Hatzigeorgiou 1997; Ohler and Niemann 2001; Pedersen 1999; Zeng et al. 2009; Zeng 2011).

4.4.1 *Ab initio* Methods

Ab initio algorithms use only DNA sequence information to predict promoter regions. They identify either putative TSSs or promoter regions or in some cases, both. *Ab initio* methods may use three different kinds of features: biological signals such as core promoter elements, TFBSs or sequence context information like oligonucleotide composition or DNA structural features. Along with feature selection, they use different statistical and machine learning methods such as weight matrices (Bucher 1990), artificial neural networks (Reese 2001; Wang and Ungar 2007), Markov chains (Audic and Claverie 1997), quadratic discriminant analysis (Davuluri and Grosse 2001), genetic algorithms (Levitsky and Katokhin 2003), principle component analysis (Li et al. 2008) and kernel methods which employ support vector machines (Abeel et al. 2008b; Gangal and Sharma 2005), etc.

These algorithms search for biological signal features of core promoter elements, for example, the TATA box, initiator element (Inr), DPE (Downstream promoter Element), specific TFBSs and CpG islands (in mammals). Generally, these algorithms either predict core promoter elements or, in some cases, give the TSS position

along with the distance between the binding site and the TSS. These models first derive consensus signals from experimentally identified TSSs or promoter elements. They then use different statistical methods like weight matrices, artificial neural networks and discriminant models to discriminate between promoter regions and their neighbouring sequences. Typical examples of this class of PPPs include PWMs (Bucher 1990), NNPP (Reese 2001), CpGProD (Ponger and Mouchiroud 2002), CpG-promoter (Ioshikhes and Zhang 2000), FirstEF (Davuluri and Grosse 2001) and Eponine (Down and Hubbard 2002). Search-by-signal PPPs are considered to be first generation methods. Earlier published PPPs did not use CpG-islands and their prediction efficiency was low, where as recent improved algorithms to predict promoters in mammalian genomes include use of CpG islands (Ioshikhes and Zhang 2000; Ponger and Mouchiroud 2002).

1. **FirstEF**: FirstEF (Davuluri and Grosse 2001), which uses CpG islands, is not a pure promoter prediction program. It identifies first exons along with putative promoter regions (Bucher 1990). The developers of this PPP observed that CpG distribution in the vicinity of TSSs is bimodal, so there are two classes of first exons that exist, such as CpG containing and non-CpG containing ones. It uses a probabilistic model to identify potential first exons (splice donor sites) for both classes of promoter regions. It considers upstream promoter region and downstream splice donor sites (GT) and checks whether the intermediate region is an exon or not. The algorithm is optimized to find potential first donor sites along with CpG-related and non-CpG-related promoter regions.
2. **CpGProD**: CpGProD (CpG Island Promoter Detection) uses CpG islands to identify mammalian promoter regions in large genomic sequences (Pedersen 1998). Although it is strictly dedicated to this particular promoter class, which corresponds to 50 % of the genes in humans, it exhibits a higher sensitivity and specificity than the other tools used for promoter prediction.
3. **Eponine**: Eponine (Down and Hubbard 2002) is one of the best algorithms and uses sequence motif signals for locating the TSS. It combines weight matrices with discrete probability distributions of differently positioned constraints. The Eponine DNA weight matrix model for any signal is represented by the following equation.

$$\phi(i; S) = \log \sum_{j=-\infty}^{+\infty} P(j).W(a + i + j; S) \quad (4.1)$$

$P(j)$ is a discrete probability distribution; $W(x;S)$ is the weight matrix score, aligning the first column to position x on sequence S ; a is the center position of the distribution, relative to the TSS; and i is the position of the true TSS. These PWM models were chosen for a set of four constraint elements in 599 mammalian promoter regions. They are

- i. a diffuse preference for CpG enrichment downstream of the TSS.
- ii. a TATAAA motif with focused distribution centered at position -30 relative to the TSS.

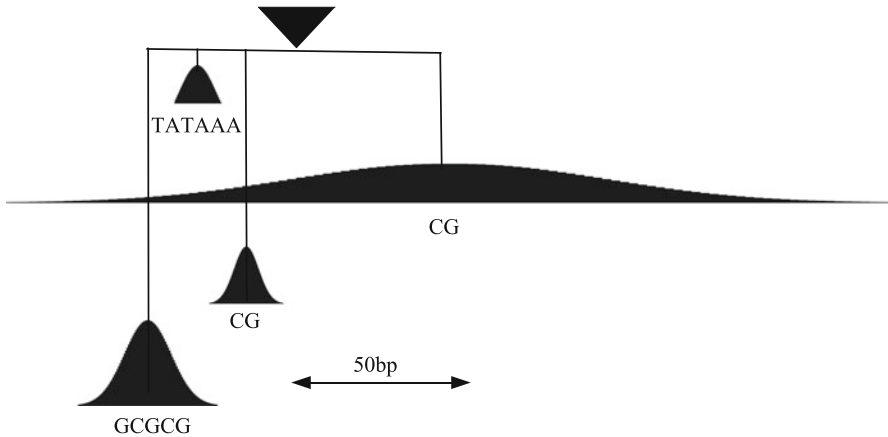


Fig. 4.3 A schematic representation of the Eponine core promoter model, showing four constraint element distributions, which were used for a weight-matrix consensus. (Down and Hubbard 2002)

iii & iv. two GC-rich matrices (GCGCG and GC) closely flanking the TATA box and positioned upstream and downstream respectively (Fig. 4.3).

To derive an efficient model, the data was trained using a relevant vector machine (RVM) algorithm with a Monte Carlo sampling process.

4.4.1.1 Search-by-content Algorithms

Search-by-content algorithms are considered to be more advanced compared to earlier approaches, as they achieve greater sensitivity and specificity. These algorithms are inspired by linguistics. The basic principle underlying all search-by-content methods is that promoter and non-promoter regions differ in their grammar and can be differentiated using certain threshold values. Context features are generally oligonucleotides represented by a set of k -tuples (or k -mers). Promoters and non-promoter regions are different in their tuple statistics. This characteristic statistical property of oligonucleotide composition can be used to discriminate promoter from non-promoter regions. Typical examples of PPPs, which use this feature, include PromFind (Hutchinson 1996), Promoter2.0 (Knudsen 1999), PromoterInspector (Scherf et al. 2000) and PCAHPR (Li et al. 2008). These classes of algorithms were shown to be more discriminative compared to search by signal algorithms. All these PPPs may differ in their statistical models but discriminate promoters from non-promoters using k -mer ($k = 2, 3, \dots, 6$) frequencies.

1. **PromoterInspector**: PromoterInspector uses discriminant functions to identify promoters and was considered the best PPP at one time (Scherf et al. 2000). This was trained using a brute-force algorithm to discover a set of sequence motifs overrepresented in promoter regions. Their models introduce IUPAC words by incorporating wildcards in multiple positions of an oligomer, except at the

start and end of words (AGCNGCA, AGCNNGCA). Using a certain threshold, it classifies IUPAC words into promoter related and non-promoter related candidates. From these pre-derived threshold values, PromoterInspector scans target the genome through a sliding window to identify promoter regions. The predictions are not strand-specific and do not provide information about the TSS. This tool was developed for mammalian genomes.

4.4.1.2 Search-by-property Algorithms

It is known that DNA structural features play a role in DNA-protein recognition (Pedersen 1998). The biological significance of different DNA structural properties in promoter regions is described in the accompanying chapter 13. These structural features are more conserved compared to sequence features. Search-by-property based algorithms use DNA structural features such as flexibility/bendability, curvature, base stacking and free energy to predict promoter regions. These algorithms are more recent compared to the methods described above and are based on one or more structural features to derive principles of learning. Generally, these kinds of models use simple statistical methods (Abeel et al. 2009); Rangannan and Bansal 2010) or advanced machine-learning approaches such as support vector machines (Abeel et al. 2008b) and are applicable across genomes, though genome based cut-offs may have to be specified. McPromoter (Ohler 2000), Prostar (Goni et al. 2007), EP3 (Abeel et al. 2008a), PromPredict (Rangannan and Bansal 2010) and ProSOM (Abeel et al. 2008b) are examples of these types of methods. Some of these algorithms (Abeel et al. 2008b) cluster sequences using structural profiles and use these clusters to classify unknown sequence into different promoter classes. Others use derived threshold property values to distinguish promoters from non-promoter regions (Abeel et al. 2009; Rangannan and Bansal 2010). If a given genomic sequence has a feature score in a defined window which is greater or smaller (depending on the property) than the pre-derived threshold, then it is classified as a promoter. These algorithms generally identify promoter regions rather than giving TSS positions.

1. **PromPredict:** PromPredict (Rangannan and Bansal 2010) uses the dinucleotide free energy values obtained from differential melting stability of DNA duplex as a predictor of promoters (SantaLucia 1998). The idea behind using DNA duplex stability is that promoter regions should be less stable than neighbouring regions for easy melting at the time of transcription initiation. Compared to other structural features, stability (or base stacking) is found to be the most prevalent feature in the promoter region (Abeel et al. 2008a). Although it was developed for bacterial promoter prediction, it also works well for Eukaryotes (Morey et al. 2011). The program takes an input genome or a fragment of a sequence along with a defined window (100 or 50) and gives the start and end of predicted promoter regions as well as least stable nucleotide position. PromPredict can be applied to any genome and also to fragments of genomic sequences, independent of their size or GC composition.

2. **EP3:EP3** (Abeel et al. 2008a) is similar to PromPredict; it uses a base-stacking property to distinguish promoter regions from other regions. For a given sequence of DNA, it calculates inverted base-stacking values over a window size of 400 base pairs in non-overlapping fashion and calls a region as promoter when the structural feature value crosses the threshold score, which is genome specific.

4.4.1.3 Integrated Algorithms

For *ab initio* promoter prediction, it is important to choose the most discriminatory features along with the discriminative model (statistical model). Some programs integrate different features to achieve better prediction (Zeng et al. 2010). ARTS (Sonnenburg et al. 2006), CoreBoost (Zhao et al. 2007), PromoterExplorer (Xie et al. 2006) and SCS (Zeng et al. 2010) are a few examples of such new-generation algorithms. which use two or more features to predict promoters. PPPs, such as MetaProm (Wang and Ungar 2007), integrate many algorithms to predict promoters. The integrated algorithms are generally better discriminators of promoter regions, compared to the algorithms described earlier.

4.4.2 Hybrid Methods

Hybrid PPPs have been developed very recently. Along with the intrinsic features of promoter sequences, they use experimental information such as gene expression and histone modification data (Wang et al. 2012). CoreBoost_HM (Wang et al. 2009) and a method using ChIP-seq Pol-II enrichment data (Gupta et al. 2010) belong to the class of hybrid PPPs. CoreBoost_HM integrates specific histone modification profiles and DNA sequence features (core promoter elements, TFBSs, flexibility) to predict human Pol II promoters. Similarly another recent method integrates gene expression data from Chip-seq and CAGE methods (average and maximum tag counts per million) as well as DNA sequence features (10 sequence composition variables and 22 property variables) to predict promoter regions in humans. Both these methods have outperformed earlier methods in terms of sensitivity and specificity.

4.4.3 Homology Based

The idea behind using DNA sequence homology for promoter prediction is that, like coding regions, regulatory regions are also evolutionarily under selective pressure and are free of mutations, whereas non-regulatory, non-coding regions can accumulate mutations. Phylogenetic foot printing (Fickett and Wasserman 2000) is one of the methods used in this type of PPP. These methods are only applicable to identify promoter regions of orthologous genes. PromH (Solovyev and Shahmuradov 2003)

is one PPP which uses orthologous gene information to predict promoter regions. PromH checks the conservation of TATA boxes in the upstream region, the conservation of nucleotide sequences around the TSS and the conservation of regulatory motifs in the upstream and downstream regions of the TSS and then uses a discriminator function to identify conserved promoter regions in pairs of orthologous genes. The program was developed specifically for testing human and rodent orthologous pairs. These kinds of algorithms are not applicable to whole genome promoter identification.

4.5 Conclusions and Future Perspectives

In silico identification of promoters is a great challenge in computational biology. A large number of promoter prediction programs are available and they differ in terms of the feature used for discriminating promoter regions from the large mass of genome sequence information. Search-by-structure or integrated algorithms appear to be promising as they are applicable to different model systems, whereas hybrid algorithms are generally efficient but are restricted to the systems for which accessory experimental information is available (such as epigenetic features and CAGE tag counts). With the rapid development of high-throughput technologies, which provide genome wide information about transcription, our understanding of promoter features is changing.

Current notion about vertebrate promoters is that while promoter regions differ in their GC and motif content, some common properties are present, such as the nucleosome free region near the TSS and epigenetic features. So, future algorithms can use this information along with other features to design new PPPs. There is always scope for the development of better algorithms based on new features and high throughput data. Most of the current PPPs are focused on promoter regions of protein coding genes. Now, with the increasing importance of non-coding RNAs in gene regulation, it is essential to analyze them. New algorithms are needed to identify promoter regions of these non-coding genes. Promoter prediction is required even if we have experimental promoter data, as we need statistical models to understand and explain promoter architecture. Up and down regulation of genes and interaction between genes is carried out through the inherent features of promoter regions. So, promoter identification and its characterization as weak or strong can serve as an important input for better understanding of systems biology of diverse organisms.

Acknowledgement MB is a recipient of the J. C. Bose National Fellowship of DST, India. We thank Rajasekaran for assistance in the preparation of Fig. 4.1.

References

- Abeel T, Saeys Y, Bonnet E, Rouze P, Van de Peer Y (2008a) Generic eukaryotic core promoter prediction using structural features of DNA. *Genome Res* 18(2):310–323
- Abeel T, Saeys Y, Rouze P, Van de Peer Y (2008b) ProSOM: core promoter prediction based on unsupervised clustering of DNA physical profiles. *Bioinformatics* 24(13):24–31
- Abeel T, Van de Peer Y, Saeys Y (2009) Toward a gold standard for promoter prediction evaluation. *Bioinformatics* 25(12):i313–i320
- Audic S, Claverie JM (1997) Detection of eukaryotic promoters using Markov transition matrices. *Comput Chem* 21(4):223–227
- Bajic VB, Seah SH (2003) Dragon gene start finder: an advanced system for finding approximate locations of the start of gene transcriptional units. *Genome Res* 13(8):1923–1929
- Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V (2002) Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* 18(1):198–199
- Bajic VB, Tan SL, Suzuki Y, Sugano S (2004) Promoter prediction analysis on the whole human genome. *Nat Biotechnol* 22(11):1467–1473
- Bajic VB, Brent MR, Brown RH, Frankish A, Harrow J, Ohler U, Solovyev VV, Tan SL (2006) Performance assessment of promoter predictions on ENCODE regions in the EGASP experiment. *Genome Biol* 7(Suppl 1):1–13
- Bucher P (1990) Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol* 212(4):563–578
- Carninci P, Sandelin A, Lenhard B, Katayama S, Shimokawa K, Ponjavic J, Semple CA, Taylor MS, Engstrom PG, Frith MC, Forrest AR, Alkema WB, Tan SL, Plessy C, Kodzius R, Ravasi T, Kasukawa T, Fukuda S, Kanamori-Katayama M, Kitazume Y, Kawaji H, Kai C, Nakamura M, Konno H, Nakano K, Mottagui-Tabar S, Arner P, Chesni A, Gustincich S, Persichetti F, Suzuki H, Grimmond SM, Wells CA, Orlando V, Wahlestedt C, Liu ET, Harbers M, Kawai J, Bajic VB, Hume DA, Hayashizaki Y (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* 38(6):626–635
- Davuluri RV, Grosse I, Zhang MQ (2001) Computational identification of promoters and first exons in the human genome. *Nat Genet* 29(4):412–417
- Down TA, Hubbard TJ (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res* 12(3):458–461
- Fickett JW, Hatzigeorgiou AG (1997) Eukaryotic promoter recognition. *Genome Res* 7(9):861–878
- Fickett JW, Wasserman WW (2000) Discovery and modeling of transcriptional regulatory regions. *Curr Opin Biotechnol* 11(1):19–24
- Gangal R, Sharma P (2005) Human pol II promoter prediction: time series descriptors and machine learning. *Nucleic Acids Res* 33(4):1332–1336
- Goni JR, Perez A, Torrents D, Orozco M (2007) Determining promoter location based on DNA structure first-principles calculations. *Genome Biol* 8(12):R263
- Gupta R, Wikramasinghe P, Bhattacharyya A, Perez FA, Pal S, Davuluri RV (2010) Annotation of gene promoters by integrative data-mining of ChIP-seq Pol-II enrichment data. *BMC Bioinformatics* 11(Suppl 1):S65
- Hutchinson GB (1996) The prediction of vertebrate promoter regions using differential hexamer frequency analysis. *Comput Appl Biosci* 12(5):391–398
- Ioshikhes IP, Zhang MQ (2000) Large-scale human promoter mapping using CpG islands. *Nat Genet* 26(1):61–63
- Juven-Gershon T, Hsu JY, Theisen JW, Kadonaga JT (2008) The RNA polymerase II core promoter—the gateway to transcription. *Curr Opin Cell Biol* 20(3):253–259
- Knudsen S (1999) Promoter2.0: for the recognition of PolII promoter sequences. *Bioinformatics* 15(5):356–361
- Lenhard B, Sandelin A, Carninci P (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat Rev Genet* 13(4):233–245

- Levitsky VG, Katokhin AV (2003) Recognition of eukaryotic promoters using a genetic algorithm based on iterative discriminant analysis. *In Silico Biol* 3(1-2):81–87
- Li X, Zeng J, Yan H (2008) PCA-HPR: a principle component analysis model for human promoter recognition. *Bioinformation* 2(9):373–378
- Morey C, Mookherjee S, Rajasekaran G, Bansal M (2011) DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes. *Plant Physiol* 156(3):1300–1315
- Ohler U (2000) Promoter prediction on a genomic scale—the Adh experience. *Genome Res* 10(4):539–542
- Ohler U, Niemann H (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet* 17(2):56–60
- Ohler U, Liao GC, Niemann H, Rubin GM (2002) Computational analysis of core promoters in the Drosophila genome. *Genome Biol* 3(12):RESEARCH0087
- Pedersen AG, Baldi P, Chauvin Y, Brunak S (1998) DNA structure in human RNA polymerase II promoters. *J Mol Biol* 281(4):663–673
- Pedersen AG, Baldi P, Chauvin Y, Brunak S (1999) The biology of eukaryotic promoter prediction—a review. *Comput Chem* 23(3–4):191–207
- Ponger L, Mouchiroud D (2002) CpGProd: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics* 18(4):631–633
- Prestridge DS (1995) Predicting Pol II promoter sequences using transcription factor binding sites. *J Mol Biol* 249(5):923–932
- Rangannan V, Bansal M (2010) High-quality annotation of promoter regions for 913 bacterial genomes. *Bioinformatics* 26(24):3043–3050
- Reese MG (2001) Application of a time-delay neural network to promoter annotation in the Drosophila melanogaster genome. *Comput Chem* 26(1):51–56
- Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA (2007) Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet* 8(6):424–436
- SantaLucia J (1998) A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 95(4):1460–1465
- Scherf M, Klingenhoff A, Werner T (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J Mol Biol* 297(3):599–606
- Schmid CD, Praz V, Delorenzi M, Perier R, Bucher P (2004) The Eukaryotic Promoter Database EPD: the impact of in silico primer extension. *Nucleic Acids Res* 32(Database issue):D82–D85
- Solovyev VV, Shahmuradov IA (2003) PromH: Promoters identification using orthologous genomic sequences. *Nucleic Acids Res* 31(13):3540–3545
- Sonnenburg S, Zien A, Ratsch G (2006) ARTS: accurate recognition of transcription starts in human. *Bioinformatics* 22(14):e472–e480
- Suzuki Y, Yamashita R, Nakai K, Sugano S (2002) DBTSS: dataBase of human Transcriptional Start Sites and full-length cDNAs. *Nucleic Acids Res* 30(1):328–331
- Thomas MC, Chiang CM (2006) The general transcription machinery and general cofactors. *Crit Rev Biochem Mol Biol* 41(3):105–178
- Valen E, Sandelin A (2011) Genomic and chromatin signals underlying transcription start-site selection. *Trends Genet* 27(11):475–485
- Wang J, Ungar LH, Tseng H, Hannenhalli S (2007) MetaProm: a neural network based meta-predictor for alternative human promoter prediction. *BMC Genomics* 8:374
- Wang J, Ma C, Zhou D, Zhang L, Zhou Y (2012) Accurately predicting transcription start sites using logitlinear model and local oligonucleotide frequencies. In: *Bio-Inspired Computing and Applications*, pp 107–114
- Wang X, Xuan Z, Zhao X, Li Y, Zhang MQ (2009) High-resolution human core-promoter prediction with CoreBoost_HM. *Genome Res* 19(2):266–275
- Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 28(1):316–319

- Xie X, Wu S, Lam KM, Yan H (2006) PromoterExplorer: an effective promoter identification method based on the AdaBoost algorithm. *Bioinformatics* 22(22):2722–2728
- Xu Z, Wei W, Gagneur J, Perocchi F, Clauder-Munster S, Camblong J, Guffanti E, Stutz F, Huber W, Steinmetz LM (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature* 457(7232):1033–1037
- Zeng J, Zhu S, Yan H (2009) Towards accurate human promoter recognition: a review of currently used sequence features and classification methods. *Brief Bioinformatics* 10(5): 498–508
- Zeng J, Zhao XY, Cao XQ, Yan H (2010) SCS: signal, context, and structure features for genome-wide human promoter recognition. *IEEE/ACM Trans Comput Biol Bioinform* 7(3):550–562
- Zhang MQ (2011) Computational promoter prediction in a vertebrate genome. In: *Handbook of Statistical Bioinformatics*, pp 73–85
- Zhao X, Xuan Z, Zhang MQ (2007) Boosting with stumps for predicting transcription start sites. *Genome Biol* 8(2):R17