

Computational Molecular Biology of Genome Expression and Regulation

Michael Q. Zhang, Ph.D.

Cold Spring Harbor Laboratory, 1 Bungtown Road
Cold Spring Harbor, NY 11724 USA

Abstract. Technological advances in experimental and computational molecular biology have revolutionized the whole fields of biology and medicine. Large-scale sequencing, expression and localization data have provided us with a great opportunity to study biology at the system level. I will introduce some outstanding problems in genome expression and regulation network in which better modern statistical and machine learning technologies are desperately needed.

Recent revolution in genomics has transformed life science. For the first time in history, mankind has been able to sequence the entire human own genome. Bioinformatics, especially computational molecular biology, has played a vital role in extracting knowledge from vast amount of information generated by the high throughput genomics technologies. Today, I am very happy to deliver this key lecture at the First International Conference on Pattern Recognition and Machine Intelligence at the world renowned Indian Statistical Institute (ISI) where such luminaries as Mahalanobis, Bose, Rao and others had worked before. And it is very timely that genomics has attracted new generation of talented young statisticians, reminding us the fact that statistics was essentially conceived from and continuously nurtured by biological problems. Pattern/rule recognition is at the heart of all learning process and hence of all disciplines of sciences, and comparison is the fundamental method: it is the similarities that allow inferring common rules; and it is the differences that allow deriving new rules.

Gene expression, normally referring to the cellular processes that lead to protein production, is controlled and regulated at multiple levels. Cells use this elaborate system of “circuits” and “switches” to decide when, where and by how much each gene should be turned on (activated, expressed) or off (repressed, silenced) in response to environmental clues. Genome expression and regulation refer to coordinated expression and regulation of many genes at large-scales for which advanced computational methods become indispensable. Due to space limitations, I can only highlight some of the pattern recognition problems in transcriptional regulation, which is the most important and best studied.

Currently, there are two general outstanding problems in transcriptional regulation studies: (1) How to find the regulatory regions, in particular, the promoters regions in the genome (throughout most of this lecture, we use promoter to refer to proximal promoters, *e.g.* ~ 1kb DNA at the beginning of each gene); (2) How to identify functional *cis*-regulatory DNA elements within each such region.

1 Introduction

Recent revolution in genomics has transformed life science. For the first time in history, mankind has been able to sequence the entire human genome. Bioinformatics, especially computational molecular biology, has played a vital role in extracting knowledge from vast amounts of information generated by high throughput genomics technologies. Today, I am very happy to deliver this key lecture at the First International Conference on Pattern Recognition and Machine Intelligence at the world renowned Indian Statistical Institute (ISI) where such luminaries as Mahalanobis, Bose, Rao and others have worked before. And it is very timely that genomics has attracted a new generation of talented young statisticians, reminding us of the fact that statistics was essentially conceived from and is continuously nurtured by biological problems. Pattern/rule recognition is at the heart of all learning processes and hence, of all disciplines of sciences, and comparison is the fundamental method: It is the similarities that allow inferring common rules and it is the differences that allow deriving new rules.

Gene expression (normally referring to the cellular processes that lead to protein production) is controlled and regulated at multiple levels. Cells use this elaborate system of “circuits” and “switches” to decide when, where and by how much each gene should be turned on (activated, expressed) or off (repressed, silenced) in response to environmental clues. Genome expression and regulation refer to coordinated expression and regulation of many genes of large-scales for which advanced computational methods become indispensable. Due to space limitations, I can only highlight some pattern recognition problems in transcriptional regulation, which is the most important and best studied. Currently, there are two general outstanding problems in transcriptional regulation studies: (1) how to find the regulatory regions, in particular, the promoters (throughout most of this lecture, we use promoter to refer to proximal promoter, *e.g.* ~ 1kb DNA at the beginning of each gene) regions in the genome; (2) how to identify functional *cis*-regulatory DNA elements within each such region.

Finding promoter and First Exon (FE) of a multi-exon gene in vertebrate genome

Transcription is the process of pre-mRNA (a gene transcript) synthesis. A typical vertebrate pre-mRNA contains about 9 exons, the intervening sequences (introns) between exons are spliced out during RNA processing to produce a matured RNA (mRNA). Most of the regulatory elements are found in the flanking regions of the FE of the target gene. Finding the FE is therefore the key for locating the transcriptional regulatory regions. Promoter upstream of (and overlapping with) FE functionally directs RNA polymerase II (PolII) to the correct transcriptional start site (TSS, the first base of FE) and the core promoter extending ~35bp on either side of TSS plays a central role in regulating initiation of transcription of pre-mRNA transcripts [35]. As the most important regulatory region, promoter is enriched by many transcription factor binding sites (TFBSs). They form so-called modules, each of which is acting relatively autonomously and responding to a specific set of TFs. Core promoter may be regarded as a general module and is the docking region for the Pre-Initiation Complex (PIC) of largely basal TFs and PolII itself. Core promoter contains one or more of the following *cis*-elements: TFIIB Recognition Element (BRE: ~-37SSRCGCC) and TATA-box (TBP-site: ~-31TATAWAAR) at about -35 upstream of the TSS, Initiator (Inr: -2YYANWYY) at the TSS and Downstream Core Promoter Element (DPE: +28RGWYV). Although these four elements are relatively position specific (with re-

spect to TSS) and they have been used for TSS prediction [46], they are not enough for accurate TSS prediction at the genome-scale because many core promoters may only have one or two such elements and many such putative sites may occur frequently by chance in a large genome. One could use a large-scale promoter finder, such as CpG_Promoter [20, 47] or PromoterInspector [32].

Three general promoter/TSS recognition approaches, briefly described below, may represent the current state-of-the-art; they all are based on some specific statistical pattern learning/prediction procedures. The first is called Dragon Promoter Finder (DPF) [2, 3]. Its algorithm uses sensors for three functional regions (promoters, exons and introns) and an Artificial Neural Network (AAN) for integrating signals. The second is called Eponine [14]. Its algorithm uses a hybrid of Relevance Vector Machine (RVM) [41] and Monte Carlo sampling from extremely large model space of possible motif weight matrices and Gaussian position distributions. The third is called First Exon Finder (FirstEF) [13]. Its algorithm uses two-level discriminant analysis: At the first level filtering, it computes a Quadratic Discriminant Analysis (QDA) score for the splice donor site from several 3'-sensors and another QDA score for the promoter from several 5'-sensors; at the second level, it integrates these flanking region sensors with additional exon-sensors using yet another QDA to arrive at the *a posteriori* probability $p(\text{FirstExon}|\text{data})$. It has been demonstrated recently that addition of ortholog comparison with other evolutionarily related species can further improve the prediction accuracy [44]. FirstEF not only can provide promoter/TSS predictions, but also predict the 5' splice site (donor site) of the first intron, which also often contains many regulatory *cis*-elements.

Currently, promoter prediction has been hampered by very limited training data and poor understanding of molecular details of regulation mechanisms. The performance of even the best prediction programs are still far from satisfactory [4], leaving ample room for further improvements. Because of high false-positives when predicting promoters in the whole genome, it should always locate the beginning (ATG) of protein coding regions first [48]. Multiple comparisons of evolutionarily related genomic DNA sequences can be very useful for finding conserved promoters. Some open problems are: (1) identification of alternative promoters [21]; (2) identification of non-CpG island related promoters [13]; (3) tissue/developmental specific classification and lineage relationship [38]; (4) epigenetic controls [16]; (5) coupling to RNA processing [27]; (6) good cross-species promoter alignment algorithms [31, 40]; (7) promoter evolution [43]; (8) gene regulation networks [23] and dynamics [26].

Identifying TFBSs in vertebrate promoters

Once approximate regulatory regions, such as promoters, are located, the next task is to identify *cis*-regulatory elements (largely TFBSs) within such regions. A single TFBS pattern (also called motif) can be characterized by either IUPAC consensus (as given above for the core promoter motifs) or position weight matrix (PWM), although more complicated models, such as WAM [45], HMM [24], ANN [30], VOBN [6], etc., are also possible, but less popular. Here I will focus on PWM model as it is the most useful and is directly related to protein-DNA binding affinity measurements [7]. There are many different PWM definitions, all of which are derived from frequency weight matrixes (FWM).

The three classical methods for promoter motif discovery are all based on multiple sequence alignment [49]: (1) CONSENSUS based on a greedy algorithm [37]; (2) MEME based on Expectation-maximization (EM) of likelihood for a mixture model [1]; (3) Gibbs sampling based on a simple Monte Carlo Markov Chain model [22]. In the mixture model, it is assumed that in the motif region, the base-pairs are generated with probabilities specified by $P(x, B_x)$ (x is the position within the motif and B_x is the base-pair at x) for which the matrix elements of FWM are the maximum likelihood estimator; outside the motif region, the base-pairs are generated according to a uniform random background model $P_0(B)$ which can be estimated by the composition of B (If B were a word of length k , the background model would be a Markov model of order $k-1$). The mixing coefficient and motif starting positions will be the model parameters to be optimized by maximizing the Log-likelihood function. All of these three methods have since been further improved with more functionalities and user-friendliness. Better initial seeding may be done by word-based motif-finding methods [5].

The above motif-finding methods are used when the motif is known to be enriched in a given set of sequences. To increase specificity and sensitivity, it is better to construct two input sequence sets: One is the positive (foreground) and the other is the negative (background). Then the interesting problem is to find motif(s) that can maximally discriminate/classify the positive set from the negative set. For example, the positive set may be the genes that are co-regulated or bound by a TF and the nega-

tive set contains the genes that are not regulated or bound by the TF. If the consensus pattern (word or spaced words) are good enough for motif description, a very fast Discriminate Word Enumerator (DWE) algorithm [38] can be used in which all possible words are efficiently enumerated and ranked by the p -values derived from hyper-geometric function (Fisher exact test). The first discriminant matrix method ANN-Spec [42] is based on a perceptron (a single layer ANN) and uses a Gibbs sampling to optimize parameters (matrix elements) for maximum specificity (differential binding of the positive set vs. the negative set) through local multiple sequence alignment. Since the positives and the negatives are usually not linearly separable, the simple perceptron maybe generalized by nonlinear models using SVM [29] or Boosting approaches [19]. More recently, a novel matrix-centric approach – Discriminate Matrix Enumerator (DME) [36] has also been developed, which allows to exhaustively and efficiently enumerate and rank all possible motifs (satisfying user specified minimum information-content requirement) in the entire (discretized) matrix space (hence guaranteeing global optimality). This binary classification problem may be generalized to multi-classification problems [33].

If one has a continuous quality score for each gene (such as fold-change in expression microarray data or binding probability in ChIP-chip data), one can further generalize the classification/discrimination problem to a regression one. The first successful application of linear regression for motif finding algorithm in yeast is REDUCE [10], using MobyDick [9] to build the initial word motifs. A similar method Motif_Regressor [11], but using MDscan [25] as a feature extraction tool, can improve the sensitivity and specificity due to matrix-based motifs. Recently, nonlinear regression methods, such as, MARS_Motif [12] based on Multiple Adaptive Regression Splines [17], have also been developed, that can model synergistic motifs with a *cis*-regulatory module (CRM). Regression methods are very powerful. They can either be used for selecting functional motifs or for predicting mRNA expression levels.

Some open problems are: (1) identification of distal enhancers/silencers [28, 8]; (2) identification of tissue/developmental specific CRMs [23]; (3) higher order structural constraints [34]; (5) TFBS evolution [18].

Future directions

I have only touched upon one special (albeit an important) area of genome expression and regulation. Even for protein-coding gene transcription, there are also many other regulatory steps (such as: promoter escape, pausing, elongation and termination in addition to chromatin remodeling and initiation), let alone those for many other RNA genes [15]. There are yet many steps of post-transcription control and regulation, such as, Capping, RNA splicing, polyadenylation, RNA transport, in the nucleus; and various translational regulation and post-translational modifications [27, 50]. The future challenge will include integration of data coming from various levels, especially how DNA, RNA (including miRNAs, or ncRNA in general) and protein are interrelated in the gene regulation networks.

Acknowledgements:

My Lab is supported by grants from NIH and NSF. I would like to thank present and past members who have contributed to various methods discussed in this text.

References

1. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol.* (1994) 2:28-36.
2. Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V. Dragon Promoter Finder: Recognition of vertebrate RNA polymerase II promoters. *Bioinformatics.* (2002) 18(1):198-199.
3. Bajic VB, Brusic V. Computational detection of vertebrate RNA polymerase II promoters. *Methods Enzymol.* (2003) 370:237-250.
4. Bajic VB, Tan SL, Suzuki Y, Sagano S. Promoter prediction analysis on the whole human genome. *Nat Biotechnol.* (2004) 22(11):1467-1473.
5. Barash Y, Bejerano G, Friedman N. A simple hyper-geometric approach for discovering putative transcription factor binding sites. In: Gascuel O, Moret BME (eds): *Algorithms in Bioinformatics. Proc First Intl Wksp, #2149 LNCS.* (2001) 278-293.
6. Ben-Gal I, Shani A, Gohr A, Grau J, Arviv S, Shmilovici A, Posch S, Grosse I. Identification of transcription factor binding sites with variable-order Bayesian networks. *Bioinformatics.* (2005) 21(11):2657-2666.

7. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J Mol Biol.* (1987) 191(4):723-750.
8. Boffelli D, Nobrega MA, Rubin EM. Comparative genomics at the vertebrate extremes. *Nat Rev Genet.* (2004) 5(6):456-465.
9. Bussemaker HJ, Li H, Siggia ED. Building a dictionary for genomes: Identification of presumptive regulatory sites by statistical analysis. *Proc Natl Acad Sci U S A.* (2000) 97(18):10096-10100.
10. Bussemaker HJ, Li H, Siggia ED. Regulatory element detection using correlation with expression. *Nat Genet.* (2001) 27(2):167-171.
11. Conlon EM, Liu XS, Lieb JD, Liu JS. Integrating regulatory motif discovery and genome-wide expression analysis. *Proc Natl Acad Sci U S A.* (2003) 100(6):3339-3344.
12. Das D, Banerjee N, Zhang MQ. Interacting models of cooperative gene regulation. *Proc Natl Acad Sci U S A.* (2004) 101(46):16234-16239.
13. Davuluri RV, Grosse I, Zhang MQ. Computational identification of promoters and first exons in the human genome. *Nat Genet.* (2001) 29(4):412-417. Erratum: *Nat Genet.* (2002) 32(3):459.
14. Down TA, Hubbard TJ. Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.* (2002) 12(3):458-461.
15. Eddy SR. Computational genomics of noncoding RNA genes. *Cell.* (2002) 109(2):137-140.
16. Fazzari MJ, Grealley JM. Epigenomics: Beyond CpG islands. *Nat Rev Genet.* (2004) 5(6):446-455.
17. Friedman MJ. Multivariate adaptive regression splines. *Ann Stat.* (1991) 19:1-67.
18. Gasch AP, Moses AM, Chiang DY, Fraser HB, Berardini M, Eisen MB. Conservation and evolution of cis-regulatory systems in ascomycete fungi. *PloS Biol.* (2004) 2(12):e398.
19. Hong P, Liu XS, Zhou Q, Lu X, Liu JS, Wong WH. A boosting approach for motif modeling using ChIP-chip data. *Bioinformatics.* (2005) 21(11):2636-2643.
20. Ioshikhes IP, Zhang MQ. Large-scale human promoter mapping using CpG islands. *Nat Genet.* (2000) 26(1):61-63.
21. Kim TH, Barrera LO, Zheng M, Qu C, Singer MA, Richmond TA, Wu Y, Green RD, Ren B. A high-resolution map of active promoters in the human genome. *Nature.* (2005) [e-pub ahead of print].

22. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment. *Science*. (1993) 262(5131):208-214.
23. Levine M, Davidson EH. Gene regulatory networks for development. *Proc Natl Acad Sci U S A*. (2005) 102(14):4936-4942.
24. Li W, Meyer CA, Liu XS. A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*. (2005) 21 Suppl 1:i274-i282.
25. Liu XS, Brutlag DL, Liu JS. An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotechnol*. (2002) 20(8):835-839.
26. Lucchetta EM, Lee JH, Fu LA, Patel NH, Ismagilov RF. Dynamics of *Drosophila* embryonic patterning network perturbed in space and time using microfluidics. *Nature*. (2005) 434(7037):1134-1138.
27. Maniatis T, Reed R. An extensive network of coupling among gene expression machines. *Nature*. (2002) 416(6880):499-506.
28. Nobrega MA, Ovcharenko I, Afzal V, Rubin EM. Scanning human gene deserts for long-range enhancers. *Science*. (2003) 302(5644):413.
29. Pavlidis P, Furey TS, Liberto M, Haussler D, Grundy WN. Promoter region-based classification of genes. *Pac Symp Biocomput*. (2001) 151-163.
30. Pedersen AG, Engelbrecht J. Investigations of *Escherichia coli* promoter sequences with artificial neural networks: New signals discovered upstream of the transcriptional startpoint. *Proc Int Conf Intell Syst Mol Biol*. (1995) 3:292-299.
31. Prakash A, Tompa M. Statistics of local multiple alignments. *Bioinformatics* (2005) 21 Suppl 1:i344-i350.
32. Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel contact analysis approach. *J Mol Biol*. (2000) 297(3):599-606.
33. Segal E, Barash Y, Simon I, Friedman N, Koller D. From promoter sequence to expression: A probabilistic framework. *Proc 6th Intl Conf Res Comp Mol Biol*. (2002) 263-272.
34. Siggers TW, Silkov A, Honig B. Structural alignment of protein-DNA interfaces: Insights into the determinants of binding specificity. *J Mol Biol*. (2005) 345(5):1027-1045.

35. Smale ST, Kadonaga JT. The RNA Polymerase II core promoter. *Annu Rev Biochem.* (2003) 72:449-479.
36. Smith AD, Sumazin P, Zhang MQ. Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A.* (2005) 102(5):1560-1565.
37. Stormo GD, Hartzell GW 3rd. Identifying protein-building sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A.* (1989) 86(4):1183-1187.
38. Sumazin P, Chen G, Hata N, Smith AD, Zhang T, Zhang MQ. DWE: Discriminating word enumerator. *Bioinformatics.* (2005) 21(1):31-38.
39. Taatjes DJ, Marr MT, Tjian R. Regulatory diversity among metazoan co-activator complexes. *Nat Rev Mol Cell Biol.* (2004) 5(5):403-410.
40. Tharakaraman K, Marino-Ramirez L, Sheetlin S, Landsman D, Spouge JL. Alignments anchored on genomic landmarks can aid in the identification of regulatory elements. *Bioinformatics.* (2005) 21 Suppl 1:i440-i448.
41. Tipping ME. Space Bayesian learning and the relevance vector machine. *J Machine Learning Res.* (2001) 1:211-244.
42. Workman CT, Stormo GD. ANN-Spec: A method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput.* (2000) 467-478.
43. Wray GA. Transcriptional regulation and the evolution of development. *Int J Dev Biol.* (2003) 47(7-8):675-684.
44. Xuan Z, Zhao F, Wang JH, Chen GX, Zhang MQ. Genome-wide promoter extraction and analysis in human, mouse and rat. *Genome Biol.* (2005) In Press.
45. Zhang MQ, Marr TG. A weight array method for splicing signal analysis. *Comput Appl Biosci.* (1993) 9(5):499-509.
46. Zhang MQ. Identification of human gene core promoters *in silico*. *Genome Res.* (1998) 8(3):319-326.
47. Zhang MQ. Discriminant analysis and its application in DNA sequence motif recognition. *Brief Bioinform.* (2000) 1(4):331-342.
48. Zhang MQ. Computational prediction of eukaryotic protein-coding genes. *Nat Rev Genet.* (2002) 3(9):698-709.
49. Zhang MQ. Computational methods for promoter recognition. In: Jiang T, Xu Y, Zhang MQ, (eds.): *Current Topics in Computational Molecular Biology*, MIT Press Cambridge, Massachusetts (2002) 249-268.

50. Zhang MQ. Inferring gene regulatory networks. In: Lengauer, T. (ed.) Bioinformatics – from Genome to Therapies. Wiley-VCH. (2005) Submitted.