IDENTIFICATION OF CpG ISLANDS AND POTENTIAL PROMOTERS IN HUMAN GENOME

By Hehuang Xie

A master project submitted in fulfillment
of the requirements for the Master of
Computer Science degree
in the Graduate College of
The University of Iowa

May 2003

Project supervisor: Assistant Professor Ramon Lawrence

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

With the progress of human genome project, huge amount of biological data has been provided in recent years. Data processing and mining become challenging and unavoidable(need to be solved?) tasks for most of biological scientists. The analysis of human genome sequence data is attracting more and more computational scientists into biological field. This thesis descripts the identifications of important features in human genome, and the applications of findings in biological studies.

Genome, Gene Structure and Gene Expression

Human Genome Composition

In a short description, genome is the nucleotide complements for hereditary information of an organism. For human beings, human genome contains 22 pairs of autosomes, one pair of sex chromosomes: X and Y, and a small mitochondria genome. Each chromosome is a long deoxyribonucleic acid (DNA) molecule with lots of associated proteins. DNA molecule is a continuous string made up of four kinds of nucleotide bases: adenine (A), cytosine (C), guanine (G), and thymine (T). In DNA molecule, four kinds of nucleotide bases form two kinds of pairs: A-T, C-G. The human haploid genome is estimated to contain 3 billions of such nucleotide base pairs.

Eukaryotic Gene Structure and Gene Expression

Functional units controlling hereditary traits in human genome are called genes, which are templates for the production of ribonucleic acids (RNAs). The process to produce RNA based on its template DNA is called gene transcription. In term of function, regulatory regions for gene transcription control, introns and exons can be viewed as main components of a gene. Exons are coding sequences, which are represented by mature messenger RNA (mRNA). Introns are non-coding sequences, which are first transcripted, but removed from messenger RNA in a later RNA splicing event. Regulatory regions for transcription may overlap with or exist in introns and exons region. A DNA with sequence complementary to its corresponding mRNA is named as complementary DNA (cDNA). Proteins are produced in a translation process with mRNAs as instructions. The gene expression flow has been summarized in Figure 1.

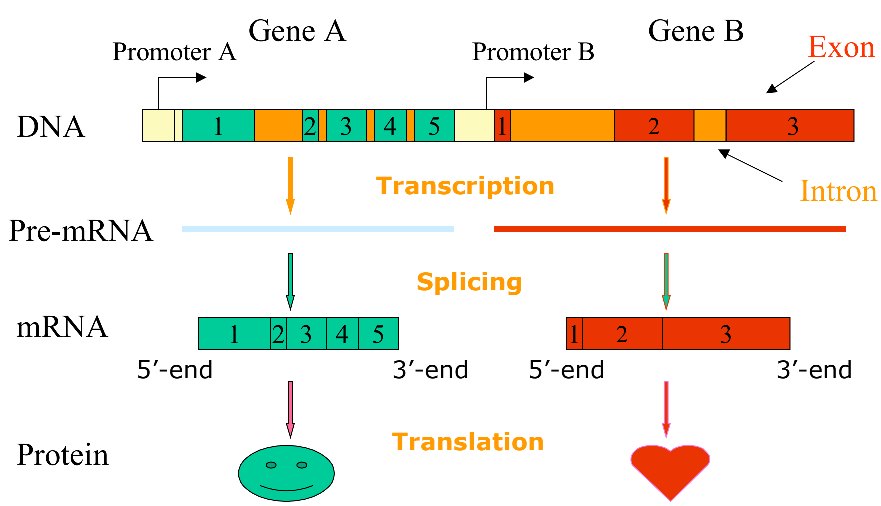

Figure 1: Diagram of gene structure and gene expression

Computational View of Biological Terms

From computer scientist viewpoint, gene structure and gene expression can be represented as Figure 2. Three distinct alphabet sets can be used to represent nucleotide bases for DNA, ribonucleotide bases for RNA and amino acids for protein. DNA, RNA and protein are strings composed of elements from their corresponding sets. Preprocessed mRNA is an exon followed by zero or more intron and exon alternates. mRNA is a concatenate of one or more exons. Three functions involved in gene expression process are: transcription, splicing and translation. Transcription takes DNA as an input, and produces preprocessed mRNA. Splicing converts a preprocessed mRNA to a mature mRNA by removing introns. Hereditary information is finally transmitted and represented on protein level by translation process based on mRNA template.

Terms and Sets:

- DNA = $(Nt)^+$; (Nt Є {A, T, C, G})
- RNA = $(rNt)^+$; (rNT Є {A, U, C, G})
- Protein = $(AA)^+$; (AA Є {20 Amino Acids})
- Pre-mRNA = (Exon)[(Intron)(Exon)]*;
- mRNA = $(Exon)^+$;

Functions:

- Pre-mRNA = F_0(DNA); (F_0: transcription)
- mRNA = F_1(Pre-mRNA); (F_1: Splicing)
- Protein = F_2(mRNA); (F_2: Translation)

Figure 2: Definitions of DNA, RNA and Protein in term of computer science

Promoter and Gene Transcription

Promoter and Transcription Factor

Promoter is a stretch DNA sequence surrounding transcription start site (TSS) of a gene. Conventionally, DNA sequence from –499 to +100 surrounding TSS of a gene is considered as promoter region(Praz, Perier et al. 2002). Gene transcription is regulated by a group of proteins called transcription factors. Promoter contains most of transcription factor binding sites (TFBS) for corresponding gene transcription. The expression level of a particular gene is largely dependent on the interaction of its promoter and transcription factors. The abundances and tissue distribution patterns of transcription factors control their target genes' expression patterns by binding to corresponding TFBS in target genes' promoter region.

Characteristics of Eukaryotic Type II Promoter

Although promoter sequences are various among different genes, a typical eukaryotic type II promoter usually contains several core elements and/or proximal promoter elements for basal level transcription(Hampsey 1998). The most common one is TATA box (Goldberg-Hogness box), which locates about 30 base pairs (bp) upstream from the TSS. Another core element is initiator (Inr), which locates right on TSS. The binding of TATA box binding protein (TBP) to the TATA box or Inr is the first step of the initiation of mRNA transcription. For promoters without TATA box, other mechanisms are used to attract RNA polymerase to promoter. Usually, a general transcription factor called SP1, which binds GC box located from +56 bp to –55 bp around TSS, serves as a mediator between promoter and RNA polymerase(Farnham and Cornwell 1991). One conserve downstream promoter element (DPE) has been found to be located from +28 to +32 relative to TSS. It is the other key element playing a role in core promoters that lack a TATA box motif. Moreover, distinct mechanisms exist for

TATA box dependent promoters and DPE dependent promoters(Kadonaga 2002). Promoters can be classified based on the existences of those elements(Suzuki, Tsunoda et al. 2001).

Promoter Prediction Algorithms

Currently, most of human genomic sequence has been revealed. With expression sequence tag (EST) projects, millions of 3'-end sequences for mRNAs have been deposited onto NCBI EST databases. However, the difficulties in obtaining 5'-end sequences of mRNAs lead to a large number of incomplete gene structures. Therefore, for most genes, promoter is undetermined even partial of gene structure has been revealed. This situation leads to the appearance of promoter prediction as a hot research topic in computational biological field. As early as the beginning of 1980's, attempt has been made to reveal the sequence patterns of phage promoters(Otsuka and Kunisawa 1982). Since then, lots of efforts have been taken to study the sequence patterns of promoters and develop models and algorithms to predict promoters from prokaryotic genomes to eukaryotic genomes(Mulligan, Hawley et al. 1984; Staden 1984; Bucher and Trifonov 1986). Promoter prediction algorithms can be classified into several categories. The first class relies on the recognition of individual transcription elements or consensus sequences(Prestridge 1995; Hutchinson 1996; Scherf, Klingenhoff et al. 2000). The second group is neural networks and markov model based, such as Markov model audic and promoter2.0 (Audic and Claverie 1997; Knudsen 1999). Some methods are taking advantage of highly associated sequences of promoter, such as CpG island(Davuluri, Grosse et al. 2001; Ponger and Mouchiroud 2002). Although most of human genomic DNA has been sequenced, the prediction of human promoters is still far away from successful. In spite of lots of algorithms existed for promoter prediction, it has been well known that most of current promoter prediction algorithms predict much more false positive promoters than real promoters. Even a combination of multiple prediction algorithms couldn't solve this problem(Liu and States 2002). Recently, the prediction algorithm orientated by existent mRNA sequences showed a promising method for less false-positive prediction(Liu and States 2002). However, the arbitrary searching for promoter within 2.5kb range of 5'-end of a cDNA sequence in the algorithm is still likely lead to a positive-positive prediction. It has been shown that human gene structure may follow certain rules, such as the first intron of a gene is twice longer than its second intron(Chen, Gentles et al. 2002). However, a complete study has not been carried out to determine the value of gene structure property in the prediction of its promoter. An attempt to establish a model to predict promoter based on gene structure will be taken in this study.

CpG island and Gene Transcription

Definition of CpG island

CpG island is a stretch of DNA sequence, which is at least 200 nucleotides long, having C and G contents greater than 0.5 and observed/expected CpG ratio greater than 0.6(Gardiner-Garden and Frommer 1987). The middle 'p' in 'CpG' stands for the phosphate group, which serves as a linkage of C nucleotide to the next G. CpG islands are estimated to be located on ~50% of human gene promoter regions. In 1989, the silencing of retinoblasta (RB) gene, which is a tumor suppressor gene, was found to be associated with the methylation of CpG island on its promoter region(Greger, Passarge et al. 1989). Since then, more and more evidences showed that methylation of CpG island are one of the important mechanisms involved in down regulation of gene expression. Moreover, CpG island methylation has been found to be involved in lots of biological processes, such as cancer oncogenesis, development, and aging process(Ehrlich, Jiang et al. 2002; Kramer, Schultheis et al. 2002; Oakes, Smiraglia et al. 2003). A review issue of *Oncogene* has been provided in

December 2002 on the current advance of CpG methylation study.

Based on the association of CpG islands and promoters, a recent study leaded to a new definition for CpG island which are more likely to be on promoter region(Takai and Jones 2002). New definition for CpG island is that CpG islands are regions of DNA of greater than 500 bp with a G+C equal to or greater than 55% and observed CpG/expected CpG of 0.65. Moreover, it has been shown that CpG island with new definition has excluded most Alu repetitive elements.

Markov Chain and Hidden Markov Model

Traditionally, the approach to identify CpG island can be classified into two groups. One is based on “window slip”. This approach simply scans an input DNA sequence with a desired window size for CpG island. It is relatively easy to be implemented, but doesn’t offer a solution for an accurate segmentation of CpG islands. The other approach is based on Hidden Markov Model (HMM). In HMM, DNA sequence is treated as an output of a state machine, which composed of a series of state transition. The procedure for this approach is that: first develop a matrix for distinguishing in-island subsequence from out-of-island subsequence based on training set of known CpG islands; then apply the matrix on an input DNA sequence to determine the most likely transitions between CpG island and non-CpG island subsequence. Markov chain model and hidden markov model are triplet state machines, and both use a same set of alphabet to represent four kinds of nucleotides existing in a DNA molecule. Transition from one state to next will emit a nucleotide. By running the state machine, a string of sequence will be produced. Markov Chain Model is a finite state machine with triplet: $M = (\Sigma, S, \theta)$, where Σ is an alphabet of $\{A,T,C,G\}$ for nucleotides; S is a finite set of states and S = {A, T, C, G}; and Θ is a finite set of state transition probabilities. HMM is developed from Markov Chain theory. It is also a finite state machine with triplet M= (Σ, S, θ), where Σ is an alphabet of {A,T,C,G}; S is a finite set of states capable of emitting symbols of Σ with S = {A+, T+,C+,G+,A-,T-,C-,G-}; and Θ is a finite set composed of state transition probabilities and emission probabilities. The emission probability of a given state is always 0 or 1 in HMM for CpG island. For instance, for state A, the probability to emit an 'A' is 1, while that of other symbols is 0. The transition probabilities are listed on table 1. (**Richard Durbin. Biological Sequence Analysis. 1999**)

Viterbi Algorithm and Dynamic Programming:

- Initial state:

  $P_0(0) = 1$; and $P_n(0) = 0$ for $n>0$

- Recursive State:

  For i = 1 to L:

  $$P_L(i) = e_L(x_i)(\mathrm{Max}_n\{P_n(i-1)*T_{nL}\})$$

  $P_L(i)$: probability of observation of I in state L

  $e_L(x_i)$: Emission probability of Xi in state L

  T_{nL}: Transition Probability from state n to L

- Termination State:

  $$P(X|p) = \mathrm{Max}_n\{P_n(L)*T_{nf}\})$$

CHAPTER II

PROMOTER PREDICTION WITH HUMAN FULL LENGTH cDNAs

Rational of Promoter Prediction Model

Today's trend in the study of transcription mechanism is to try associate large-scale gene expression data to transcriptional regulatory elements. A long-term goal has been set to establish a gene expression regulatory network. Several promoter and transcription factor databases have been developed for various kinds of purposes. With longest history, Eukaryotic Promoter Database (EPD) contains the most strictly selected promoter sequences with only 276 entries for human promoters associated with supporting experimental data (Praz, Perier et al. 2002). It is a most reliable promoter database, and has been widely used as a good resource for the development of promoter prediction algorithm. However, the limit number of human promoter entries in EPD is far away from practical application in large-scale gene expression study. Compared with EPD, Human Promoter Database constructed in Boston University has 2,004 promoter entries. However, the data set of this database is derived from a simply merge of three different resources, including EPD. Not further verification has been provided, and web interface is not friendly designed for gene expression analysis. A reliable human promoter set is highly desired for currently large-scale gene expression study.

With the progress of full-length cDNA project, more and more cDNA data are available for correctly prediction of human promoters. ESTs and full-length cDNAs have been shown to be valuable for promoter prediction(Liu and States 2002). It has been shown that certain rules may exist for a typical human gene structure(Chen, Gentles et al. 2002). A comprehensive analysis would be necessary to further address the possibility to infer whether a given cDNA fragment is a full-length cDNA sequence based on gene structure. In this study, cDNA sequences from EPD, Chromosome 21 and Chromosome 22 were applied as a training set to establish the promoter prediction model. By mapping these cDNAs to human genome, all corresponding gene structures were obtained and subjected to statistical analysis. An efficient promoter prediction algorithm is developed based on the comprehensive understanding of those gene structures.

Human Genome and Full length cDNA

Human Genomic DNA Sequence

In the middle of 1980s', the advance on DNA sequencing related techniques promoted scientists to start the largest biological project to sequence entire human genome(Hood, Hunkapiller et al. 1987). Human genome project was then officially initiated in 1990 by the Department of Energy and the National Institutes of Health in an effort to determine whole genomic sequence and identify all human genes. By January 2003, 96% of human genome has been complete sequenced, and 3% is under the way, table 2. Human genomic DNA sequence, including draft sequence, is available on web from NCBI (http://www.ncbi.nlm.nih.gov/genome/guide/human/).

Mammalian Gene Collection

Mammalian gene collection (MGC) from National Center for Biotechnology Information (NCBI) is established to eventually provide a complete set of full-length cDNA clones of human and mouse genes. A full-length cDNA clone contains all the coding information for corresponding protein, which is so called as open reading frame (ORF). Based on 5'-end and/or 3'-end sequences of a cDNA clone, cDNA clones are further selected and subjected to high accuracy sequencing. However, in term of mRNA level, full-length cDNA may not contain the transcriptional start site, and not a ‘full length’ template. Currently over ten thousands human cDNA sequences and clones are available from MGC (http://mgc.nci.nih.gov/). Most of those full-length cDNA sequences are believed to be the known cDNAs with longest 5’-end. Moreover, all those sequences and their associated annotations are publicly available to research community.

Database of Human Transcriptional Start Sites

Another attractive full-length cDNA resource is Database of Human Transcriptional Start Sites (DBTSS) in Japan, which provides 7,889 gene entries(Suzuki, Yamashita et al. 2002). By using oligo-capping technique, full-length cDNA clones in DBTSS are assumed to have the transcriptional start sites. Therefore, compared to MGC, DBTSS are believed to be even better for promoter identification, since it may provide longer 5’-end sequences and correct TSS, which are essential to determine promoter region for a particular gene. It has been shown that 4,802 sequences from DBTSS extend reference sequences in NCBI to 5’-end. However, 4,194 sequences didn’t provide longer 5’-end sequence. It has been estimated that around 60% of human genes have a coding first exon(Davuluri, Grosse et al. 2001). According to personal discussion with Dr. Tom Bair, most of translations start sites (ATG) were found in first exon in MGC. It implies that a number of genes in MGC may not contain TSS although they have full-length coding sequence. The other interesting result is that the average length ratio between first intron and second intron for MGC genes is above 4.0 (not for sure yet, Tom Bair). Therefore, although MGC and DBTSS provide much more entries than EPD, extra work is still required to predict the true transcription start sites before corresponding promoter sequences can be extracted from human genomic sequence.

Softwares used and/or implemented

UICluster

UICluster was developed in Thomas Casavant's lab in the University of Iowa. It takes a file with cDNA sequences in fasta format as an input, and grouped cDNAs together based on sequence similarity. As default value, all sequences share 40 bp identical substring of sequence are considered to be derived from one gene.

Basic Local Alignment Tool

BLAT was developed by W. James Kent from department of biology and center for molecular biology of RNA, University of California-Santa Cruz in 2002. It is extremely powerful for align large number of DNA entries against nucleotide database. Blat has been used to align human full length cDNAs against human genomic sequence.

Results and Interpretation

Clustering full length cDNAs

A total number of 14,584 MGC clones' sequences were downloaded from MGC (http://mgc.nci.nih.gov/index_html). 4,802 cDNAs' sequences, which were shown to have longer 5'-end than reference sequences have been downloaded from DBTSS (http://dbtss.hgc.jp/samp_home.html). Two sets of cDNAs sequences were concocted and subjected to UICluster. Sequences sharing more than 100 bp identical substring were clustered together. For 19,386 input sequences, 11,485 clusters were obtained. Sequences with longest 5'-end were selected from each cluster.

Model Test and Application

Genes from EPD, chromosome 21 and chromosome 22 will be divided into two groups. One group will be used for generating models, while the other group will be saved as a verification set. The model will be test by verification the predicted result in a verification subset of training set. An estimation of prediction efficiency will be provided, such as the ratio of false positive prediction.

The follow-up phase will be the characterization of those predicted potential promoter regions. For each of the potential promoter region, the existence information of promoter elements will be obtained, such as TATA box, GC box, Inr etc. The methylation status of CpG Island has been known as a suppression mechanism for a number of genes involved in cancer development. Whether a potential promoter is localized on CpG island or not will also be determined.

Those potential promoter regions with associated characterization information will be provided for a parallel project to construct a human specialized promoter database for microarray data mining.

CHAPTER III

CpG ISLAND IDENTIFICATION IN HUMAN GENOME WITH HIDDEN HARKOVE MODLE

Softwares used and/or implemented

A CpG island finder has been implement with window shifting algorithm. The inputs for CPG island finder are: Human Genomic Sequence (Contig sequence); CpG island length threshold; GC content threshold; observed CpG/expected CpG ratio threshold. The outputs of the program are: CpG island's location; CpG island length; GC content; observed CpG/expected CpG ratio; CpG island sequence. This program has been developed and tested for various kinds of nucleotide sequence inputs.

Results and Interpretation

One side effect of CpG island methylation is that the methylated C tends to mutate into T. This mutation leads to a low CG content in genome. Based on human genomic sequence to date, the average CG content in human genome is about 40.9%. The chromosome with the highest CG content is chromosome 19 with 48.3% CG content, while chromosome with the lowest CG content is chromosome 4 with 38.2% CG content. The nucleotide compositions and CG content of human genome has been summarized on table 3.

CHAPTER IV

FUTURE DIRECTION

The ultimate goal of my study is to develop a comprehensive gene expression analysis platform to provide a maximal integration of informational and software resources. To achieve this goal, the first step will be the understanding and generation of biological mechanisms on gene expression. The study of genome, gene structure and the regulation of gene expression is the cornerstone for the design of any analysis tool. The research descripted in this thesis provide the fundamental data for further human gene promoter prediction and gene expression study.

Table 1. Transition Probabilities for CpG island prediction with HMM model

	A+	C+	G+	T+	A-	C-	G-	T-
A	0.180	0.274	0.426	0.120	0.300	0.205	0.285	0.210
C	0.171	0.368	0.274	0.188	0.322	0.298	0.078	0.302
G	0.161	0.339	0.375	0.125	0.248	0.246	0.298	0.208
T	0.079	0.355	0.384	0.182	0.177	0.239	0.292	0.292

Table 2: Current Status of Human Genome Sequencing Process (Jan 5, 2003)*

Chromosome	Total Clones	Draft Clones	Finished Clones	Percent Finished
1	2202	71	2089	94.9%
2	1966	14	1948	99.1%
3	1740	82	1630	93.7%
4	1629	47	1572	96.5%
5	1775	54	1700	95.8%
6	1795	2	1790	99.7%
7	1526	1	1514	99.2%
8	1239	79	1133	91.4%
9	999	22	963	96.4%
10	1132	6	1117	98.7%
11	1147	43	1088	94.9%
12	1140	74	1044	91.6%
13	854	0	854	100.0%
14	655	1	642	98.0%
15	710	68	618	87.0%
16	725	44	672	92.7%
17	692	132	541	78.2%
18	600	6	585	97.5%
19	861	9	852	99.0%
20	632	0	632	100.0%
21	473	0	472	99.8%
22	527	0	527	100.0%
X	1588	37	1508	95.0%
Y	200	0	200	100.0%
Total	26807	792	25691	95.8%

*Adapted from http://www.ncbi.nlm.nih.gov/genome/seq/ (Jan 5, 2003)

Table 3: Nucleotide Compositions and CG Content of Human Genome*

Chromosome	Con tig	TotalNT	aCount	tCount	cCount	gCount	nCount	cgContent
chromosome 1	111	221782893	64532501	64666282	46259567	46251685	72858	0.417125283
chromosome 2	67	237637456	70911807	71105627	47775263	47824440	20319	0.402292234
chromosome 3	107	194846173	58727854	58669259	38668317	38688705	90775	0.397015865
chromosome 4	57	188402715	58140436	58164039	36016604	36033636	48000	0.382426761
chromosome 5	58	177705559	53637443	53771935	35091021	35138115	67045	0.395199432
chromosome 6	17	175762617	52891826	52814604	35013076	35043110	1	0.398584108
chromosome 7	14	153794793	45567291	45630891	31314481	31282130	0	0.407013851
chromosome 8	52	142788062	42769319	42690950	28651630	28615675	60028	0.401065076
chromosome 9	53	117013362	34301875	34365726	24168748	24160480	16533	0.413023155
chromosome 10	29	131098977	38267257	38314243	27245588	27256413	15476	0.415731703
chromosome 11	43	133239679	38918604	38909875	27663099	27681230	66798	0.415374229
chromosome 12	74	129362603	38253072	38315714	26329402	26367790	96623	0.407360325
chromosome 13	7	95228136	29246468	29330772	18326525	18324371	0	0.384874655
chromosome 14	7	88182284	25971017	26166493	17998763	18042993	3018	0.408718785
chromosome 15	33	83582680	24129759	24090964	17669778	17656733	35446	0.422653485
chromosome 16	37	80889146	22271234	22359413	18065334	18147858	45162	0.447689137
chromosome 17	49	80734148	21862814	21988386	18398790	18403479	80679	0.45584514
chromosome 18	14	74619305	22451800	22471114	14827882	14850569	17940	0.397731539
chromosome 19	17	56446152	14562199	14586541	13625730	13653808	17873	0.483284281
chromosome 20	7	59424940	16503719	16705836	13087575	13127810	0	0.441151224
chromosome 21	5	33917895	10059003	9994559	6937434	6926717	179	0.408756233
chromosome 22	11	33821705	8846873	8800702	8090307	8083806	17	0.478216962
chromosome X	69	147274156	44549100	44643242	29008458	29036066	37288	0.39412566
chromosome Y	6	22660226	6859095	6948801	4410543	4441787	0	0.390654974
Summary	944	2860215662	844232366	845505968	584643915	585039406	792058	0.408949345

* Based on human genomic sequence released on Jan 5, 2003, NCBI

REFERENCES

Audic, S. and J. M. Claverie (1997). "Detection of eukaryotic promoters using Markov transition matrices." Comput Chem **21**(4): 223-7.

Bucher, P. and E. N. Trifonov (1986). "Compilation and analysis of eukaryotic POL II promoter sequences." Nucleic Acids Res **14**(24): 10009-26.

Chen, C., A. J. Gentles, et al. (2002). "Genes, pseudogenes, and Alu sequence organization across human chromosomes 21 and 22." Proc Natl Acad Sci U S A **99**(5): 2930-5.

Davuluri, R. V., I. Grosse, et al. (2001). "Computational identification of promoters and first exons in the human genome." Nat Genet **29**(4): 412-7.

Ehrlich, M., G. Jiang, et al. (2002). "Hypomethylation and hypermethylation of DNA in Wilms tumors." Oncogene **21**(43): 6694-702.

Farnham, P. J. and M. M. Cornwell (1991). "Sp1 activation of RNA polymerase II transcription complexes involves a heat-labile DNA-binding component." Gene Expr **1**(2): 137-48.

Gardiner-Garden, M. and M. Frommer (1987). "CpG islands in vertebrate genomes." J Mol Biol **196**(2): 261-82.

Greger, V., E. Passarge, et al. (1989). "Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma." Hum Genet **83**(2): 155-8.

Hampsey, M. (1998). "Molecular genetics of the RNA polymerase II general transcriptional machinery." Microbiol Mol Biol Rev **62**(2): 465-503.

Hood, L. E., M. W. Hunkapiller, et al. (1987). "Automated DNA sequencing and analysis of the human genome." Genomics **1**(3): 201-12.

Hutchinson, G. B. (1996). "The prediction of vertebrate promoter regions using differential hexamer frequency analysis." Comput Appl Biosci **12**(5): 391-8.

Kadonaga, J. T. (2002). "The DPE, a core promoter element for transcription by RNA polymerase II." Exp Mol Med **34**(4): 259-64.

Knudsen, S. (1999). "Promoter2.0: for the recognition of PolII promoter sequences." Bioinformatics **15**(5): 356-61.

Kramer, A., B. Schultheis, et al. (2002). "Alterations of the cyclin D1/pRb/p16(INK4A) pathway in multiple myeloma." Leukemia **16**(9): 1844-51.

Liu, R. and D. J. States (2002). "Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling." Genome Res **12**(3): 462-9.

Mulligan, M. E., D. K. Hawley, et al. (1984). "Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity." Nucleic Acids Res **12**(1 Pt 2): 789-800.

Oakes, C. C., D. J. Smiraglia, et al. (2003). "Aging results in hypermethylation of ribosomal DNA in sperm and liver of male rats." Proc Natl Acad Sci U S A **6**: 6.

Otsuka, J. and T. Kunisawa (1982). "Characteristic base sequence patterns of promoter and terminator sites in phi X174 and fd phage DNAs." J Theor Biol **97**(3): 415-36.

Ponger, L. and D. Mouchiroud (2002). "CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences." Bioinformatics **18**(4): 631-3.

Praz, V., R. Perier, et al. (2002). "The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data." Nucleic Acids Res **30**(1): 322-4.

Prestridge, D. S. (1995). "Predicting Pol II promoter sequences using transcription factor binding sites." J Mol Biol **249**(5): 923-32.

Scherf, M., A. Klingenhoff, et al. (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.

Staden, R. (1984). "Computer methods to locate signals in nucleic acid sequences." Nucleic Acids Res **12**(1 Pt 2): 505-19.

Suzuki, Y., T. Tsunoda, et al. (2001). "Identification and characterization of the potential promoter regions of 1031 kinds of human genes." Genome Res **11**(5): 677-84.

Suzuki, Y., R. Yamashita, et al. (2002). "DBTSS: DataBase of human Transcriptional Start Sites and full-length cDNAs." Nucleic Acids Res **30**(1): 328-31.

Takai, D. and P. A. Jones (2002). "Comprehensive analysis of CpG islands in human chromosomes 21 and 22." Proc Natl Acad Sci U S A **99**(6): 3740-5.