21

Algorithms for CpG Islands Search: New Advantages and Old Problems

Yulia A. Medvedeva
Vavilov Institute of General Genetics, Russian Academy of Sciences, Research Institute for Genetics and Selection of Industrial Microorganisms, Russia

1. Introduction

CpG islands (CGIs) are regions having high GC and CpG content while generally mammalian genomes are CpG-depleted. CGIs are often located in the promoter region of the genes, mostly housekeeping but also tissue-specific. It is widely believed that CpG dinucleotides within promoters CGIs are unmethylated and are targets for specific regulatory protein binding. As a result, CGIs contain special sequence motifs for highly affinitive protein binding (transcription factor binding sites, TFBS). Methylation of cytosine in CpG context within such motifs could decrease the affinity of TF binding, increase the attraction of methyl-binding proteins, affect the histones modification and, therefore, leads to repression of genes transcription. The mechanism of local and global transcription repression via CpG methylation is used in many different normal (development, differentiation, aging, X-chromosome inactivation, imprinting) and pathological processes (cancer and other diseases). However recently it has been reported that a class of normally methylated but active promoters do exist.

Lately evidences of biological relevance of methylated CGIs or CGIs located far from gene promoters appear. Such CGIs could act as regulator for pervasive transcription, which seems to be actual genome feature rather than a side-effect of high-throughput techniques errors. Replication origins are also reported to be associated with CGIs of any location.

As a consequence of specific nucleotide content, CGIs could affect DNA or RNA secondary structures. For example, $G_{2\text{-}3}C_{2\text{-}3}$ motif common within CGIs induces significant local curiosity of DNA. Another motif, G-rich sequence (GRS) in 3′ and 5′ region of RNA, is known to form specific structures, G-quadruplexes, on both end of RNA playing important role in its stability. This motif corresponds to C-rich sequence in DNA, is likely to appear in CGIs.

Classical algorithms for CpG islands search use sliding window (SWM) or running sum (RSM) and several distinct but not independent criteria (GC content, Obs/Exp_{CpG} and length). The thresholds for the criteria are rather arbitrary, unconcerned between species, and demonstrate lack of biological interpretation. SWM algorithms are rather slow, RSM algorithms are faster but tend to split large CGIs into several smaller ones and to omit CGIs with nonuniform distribution of CpG dinucleotides along the sequence. Recently, several different algorithms based on CpG dinucleotides clustering were implemented. Those algorithms have smaller number of parameters and reasonable mathematical basics. The comparison of the algorithms is tricky. Hypermutability of CpG dinucleotides lead to loss of

CGI conservation between species so comparative genomics cannot be applied for estimation of the algorithms effectiveness.
To validate the results of CGI prediction authors use different biological and mathematical properties. One of the most popular quality measures is the fraction of CGIs located near promoters of protein coding genes and avoided overlap with Alu-repeats. This measure couldn't be appropriate at least for two reasons. First, promoters of protein-coding genes are likely to be a small fraction of all promoters as it became clear recently. Second, two classes of promoters (CGI-dependent and CGI-independent) exist and their ratio is unclear. Avoiding of repetitive sequences is more or less reachable for many algorithms, but now authors prefer to remove Alu- repeats and other repetitive DNA sequences in advance.
Prediction of the methylation profile in different tissues in norm and in cancer is another idea for validation. Algorithms of CGI search *per se* fail to predict correctly the distribution of methylated cytosine in the genome. To distinguish between methylated and non-methylated CGI machine-leaning techniques (MLT) are used. Those studies include additional sequence features (di- and trinucleotide distribution, CpG and TpG frequencies, TFBS, repetitive elements and others). Machine-leaning techniques are also applicable for collecting promoter CGIs. The point that GC content and CpG frequency or density of CpG clusters is not enough to describe special types of CGIs, is highly relevant. The main problem of MLT approaches is that resulting model usually has a lot of parameters, sometimes without clear biological meaning. Consistency of the models, build up by different authors in the similar conditions is rather low, so those feastures could hardly be used for CGI validation quality in general case.
A verification problem caused by lack of universal biological properties of CGIs results in an absence of widely accepted definition. It should be mentioned that all algorithms trying to predict CGIs with one particular function (promoter or unethylated CGIs) demonstrate a high false-positive rate, probably due to the complex network of CGIs functions. It's becoming clear that many different functional elements exist within one CGI. Moreover, both methylated and unmethylated, both promoter and non-promoter CGIs seem to be functional. So, one can conclude that contemporary algorithms for CGIs search based only on GC and CpG content or on CpG clustering determine a chimeric class of objects.

2. Algorithms for CpG islands search

Nowadays, most popular algorithms for CpG islands search are still based on criteria established more than twenty years ago (Gardiner-Garden & Frommer, 1987). The DNA segment is considered to be a CpG island if it is not shorter than 200 bp, has GC content no less than 0.5 and the ratio Obs/Exp_{CpG} (1) no less than 0.6.

$$Obs/Exp_{CpG} = N_{CpG}*N/(N_C*N_G), \quad (1)$$

where N_C, N_G and N_{CpG} are numbers of C, G and CpG in the region of length N respectively. Implementations of the basic idea vary in details, mostly in methods for search of the segments having properties mentioned above.

2.1 Sliding window methods

There are several algorithms for CGIs search using sliding window methods (SWM): CpGplot (Rice et al., 2000), CpG Island Searcher (Takai & Jones, 2002), CpG Island Explorer (Wang & Leung, 2004) and CpGProD (Ponger & Mouchiroud, 2002).

CpGplot represent the simplest variant of SWM. GC content and Obs/Exp_{CpG} ratio are calculated over a window of length 100 bp moving along the sequence with 10 bp steps.
CpG Island Searcher (usually referred to as Takai-Jones algorithm) uses a window of 200 bp moving along the sequence with 200 bp steps. It has an additional threshold for minimal CpG dinucleotides in predicted CGI, equal to mathematical expectation of CpG dinucleotides in Bernoulli sequence of given length and nucleotides probabilities, multiplied by Obs/Exp_{CpG} threshold. This feature lets authors exclude "mathematical CGIs" like 300 bp sequence with 150 cytosines and one guanine in CpG context which fits standard CGI criteria. This algorithm also merges two or more CGIs if they are spaced by less than 100 bp. Takai and Jones also suggest using more strict thresholds of 500 bp for CGI length, 0.55 for GC content and 0.65 for Obs/Exp_{CpG} to find out CGIs associated with promoters of known protein-coding genes and to avoid CGIs associated with Alu-repeats.
CpG Island Explorer is a modification of CpG Island Searcher from Takai and Jones. A sliding window of CpG Island Explorer moves more slowly with a step of 10 bp. After merging of close CGIs the resulting CGI is tested ones again to fit the criteria and if it does not, one bp from each side is cutting until final CGI fits the criteria. Takai and Jones believe that CGIs predicted by CpGIE are larger in length. Closely located CGIs are merged more reasonably by CpGIE than by CpGIS.
CpGProD is a program dedicated to the prediction of promoters associated with CpG islands in mammalian genomic sequence. In every sequence found by sliding window and fitted the criteria of CGI the probability to find promoter is estimated as

$$p = \exp(Z) / (1 + \exp(Z)), \quad (2)$$

where Z is linear combination of CGI length, GC content and Obs/Exp_{CpG}. Also the probability of a strand to be a template for transcription is estimated as in (2), where Z is linear combination of AT- and GC-skews which are known properties of the nucleotide sequence around the TSS. Coefficients for Z are estimated from two generalized linear regressions trained with two datasets composed of CGIs obtaining and not obtaining TSS for protein-coding genes or two datasets with different transcription templates in human.

2.2 Running sum methods

Running sum methods (RSM) were developed as an alternative to SWM. RSM try to find segments of DNA having CpG dinucleotides more frequently comparing to the neighboring genomic sequence. RSM work faster comparing to SWM. Initially RSM did not use CGI criteria established in (Gardiner-Garden & Frommer, 1987). Most known methods from this group are CpGreport (newCpGseek) (Rice et al., 2000) and unpublished algorithm of Mikhlem and Hillier which is used in UCSC Genome Browser (http://genome.ucsc.edu) and therefore became *de facto* a standard for CGI search.
CpGreport (or newCpGseek) scores each position in the sequence using a running sum calculated from all positions in the sequence, starting with the first and ending in the last. If there is not a CpG dinucleotide at a position, the score is decremented, if there is one, the score is incremented by a constant value. If the score is higher than a threshold then a putative CGI is declared. Sequence regions scoring above the threshold are searched for recursively. It should be noticed that final CGI from predicted by this algorithm starts and ends with CpG dinucleotide and doesn't necessary reach the initial CGI criteria (Gardiner-Garden & Frommer, 1987). Authors found a lot of rather short CGI with high GC content and CpG frequency and considered such CGI as overprediction (Rice et al., 2000).

UCSC CGI (Algorithm of Mikhlem and Hillier) is based on the RSM but include additional check for CGI to fit the traditional criteria (Gardiner-Garden & Frommer, 1987). Total number of CGIs obtained by UCSC is less than those obtained by CpGplot, as not every frame is tested for fitting the criteria, but only those having score higher than a threshold on the first step. CGIs predicted by the algorithm of Mikhlem and Hillier are often shorter from both ends comparing to those predicted by CpGplot and also starts and ends with CpG dinucleotides.

2.3 CpG clustering methods

Next logical step of CpG searchers development is to implement actual CGI clustering methods (CGCM). There are several such algorithms available: CpGcluster (Hackenberg et al., 2006), CpG clusters (Glass et al., 2007), and CGI HW, an algorithm, developed by H. Wu (Irizarry et al., 2009; H. Wu et al., 2010). These algorithms are based on segmentation of the genome into regions with different frequency of CpG dinucleotides (CGI HW also uses segmentation based on GC content). Unlike methods described above this approach to CGI prediction is data-driven and allows finding CGIs in spices with different average GC-content and CpG frequency.

CpGcluster has two separate steps: a CpG cluster search and an estimation of the probability to find such a cluster by chance. Distance between neighboring CpG dinucleotides in random sequence is simulated by geometric law with CpG frequency as a parameter. Hackenberg and colleagues (Hackenberg et al., 2006) assume that within functional CpG cluster the distance between neighboring CpGs is smaller than expected in random sequence. Authors show that distances smaller than a median of the theoretical distribution is overrepresented in human genome. The median distance between neighboring CpG (23-53 bp depending of the chromosome) is used as a threshold, so each cluster consists of CpGs located no farther than the threshold. All resulting CGIs start and end with a CpG dinucleotide. Each cluster has a p-value calculated based on negative binomial distribution. Only clusters with p-value less than 1.0e-5 (1.0e-20 in (Hackenberg et al., 2010a)) are considered as CpG islands. Authors find about 200000 CpG islands in human genome (25000 CpG islands using the p-value threshold equal to 1.0e-20). A lot of such CpG islands are shorter than 200 bp. Yet, authors show functionality of some short CGIs and call them CpG islets (Hackenberg et al., 2010a).

CG clusters annotation also has two steps. The location of every CpG dinucleotide is extracted from genomic DNA sequences. Using these positions, every overlapping sequence fragment containing a fixed number of CpGs and having variable length is identified. For each number of CpGs, the frequency of each fragment length is recorded. The threshold for each maximum fragment length is defined as a local minimum in the fragment length histogram, estimated by identifying zero values of the first derivative of a cubic spline fit. Mapping the CpG-dense fragments back to the genomic sequence produces an annotation track there each annotated locus is a conglomeration of one or more overlapping fragments of variable length. As the basis for choosing the optimal track the number of overlapping fragments at a locus normalized by the maximum fragment length is used. A track with maximal fragments overlap per locus is selected based on genomic averages of this metric for different numbers of CpGs per fragment. This approach allows authors to choose the species-specific optimal number of CpGs per fragment for the final annotation.

CGI_HW (Algorithm of H. Wu) assumes that each chromosome is divided into 3 states: Alu repetitive elements, baseline, and CGI. Alu-repetitive elements are removed in advance. Hence, authors characterize the problem as that of a semi-HMM, with a known state for Alu repetitive elements, so they consider the 2-state chain conditional on being in a non-Alu

state. Authors use the number of C, G, and CpG in segment of length L as parameters for the model. Hidden state Y(s) for segment is 1 for CGI and 0 for baseline. Authors assume that Y(s) is a stationary first-order Markov chain. The choice of the state is based on two HMM. One is for GC content to be high or low with assumption of the binomial distribution approximated with the normal density for baseline. The second one is for CpG number with assumption of Poisson distribution for baseline. The length L=16 for the segment was chosen based on the association of CGI with epigenetic marks. The approach summarizes the evidence for CGI status as probability scores. This provides flexibility in the definition of a CGI and facilitates CGI search in different species.

3. Validation problem

Having several methods for CGI prediction one is still unable to select the best one. The main reason is the lack of validation criteria. Su and colleagues (Su et al., 2009) propose cumulative mutual information of CpG dinucleotides as a measure of CGI's quality and show that it's a powerful criterion to avoid CGIs associated with Alu-repeats. Despite the power of this mathematical criterion, most of the authors try using biological features for CGIs validation.

3.1 Sources for biologically relevant validation: DNA methylation and protein binding

Very first work mentioned CG-rich islands (Bird, 1986) considers them as DNA regions where cytosine is unmethylated. Cytosine methylation usually appear in CpG context and increase the probability of its deamination about 10-times (Ehrlich & Wang, 1981), leading to enrichment of TpG and depletion of CpG dinucleotides in DNA. Absence (or decreased level) of cytosine methylation within CGI is usually considered as an origin of CGIs in mammalian genomes (Cross et al., 1994; Eckhardt et al., 2006). Modern research shows that methylated cytosines within CpNpG are also targets for spontaneous deamination (Cooper et al., 2010).

No doubts, that cytosine methylation plays important role in CGI functioning. During early development waves of methylation-demethylation generate tissue-specific genomic methylation profiles. These profiles are stable in somatic cells generations due to replication dependent maintenance methylation system (Brero et al., 2006). About 70-80% of cytosines in CpG context are methylated in differentiated cells (Baylin et al., 1998), recent study shows that cytosine is also methylated within CpHpN context (where H = C, A or T) especially in embryonic stem cells (Baylin et al., 1998). Cytosine methylation influence DNA structure by facilitating Z-from conformation (Behe & Felsenfeld, 1981), it also affect protein binding to DNA, so most transcription factors (TF) usually bind unmethylated DNA.

There is a class of proteins (e.g. MeCP1/2, MBD1-6, SRA, and Kaiso) binding exclusively methylated DNA (Saito & Ishikawa, 2002). MeCP1 protein complex binds methylated cytosine using MBD2 protein (Berger & Bird, 2005) and also includes chromatin remodeling complex NuRD/Mi2. MeCP2 is the key and well-studied member of methyl-binding domain (MBD) protein group (Fatemi & Wade, 2006). Besides methyl-binding domain it contains transcription repression domain (TRD) (Dhasarathy & Wade, 2008) and is involved into DNA methylation establishment with DNMT1 (Kimura & Shiota, 2003). There are evidences that both MeCP2 and MBD1/2 binds not just ^{5m}CpG but more complicated DNA motifs, MeCP2 binds ^{5m}CpG with adjacent $(A/T)_{4+}$, which is not true for MBD1/2 proteins (Klose et al., 2005). MeCP2 binds DNA with higher affinity than MeCP1 complex leading to more stable repression of transcription. For MeCP2 binding single ^{5m}CpG dinucleotide is enough whereas MeCP1 complex needs dense clusters of ^{5m}CpGs (Ng et al., 1999).

Another well-known group of methyl-binding proteins consists of Kaiso and ZBTB4/33. They obtain zinc-finger domain and bind DNA in sequence-specific manner. Data on Kaiso binding site are controversial. Van Roy and McCrea (van Roy & McCrea, 2005) believe that Kaiso binds ^{5m}CG^{5m}CG. Sasai and colleagues (Sasai et al., 2010) assume that ^{5m}CG^{5m}CG motif is a place where two Kaiso molecules bind, one on every strand. The motif also has to be in specific sequence environment. It's also known that Kaiso binds TNGCAGGA motif having non-methylated cytosine, but with 1000-times lower affinity (Daniel et al., 2002). There are some evidences that Kaiso is a global repressor of methylated genes and is essential for early embryonic development. ZBTB4 protein binds CYGCCATC motif as well as M^{5m}CGCYAT (Sasai et al., 2010). It also has been shown that proteins of this group demonstrate affinity to half-methylated DNA (Sasai et al., 2010).
Some other proteins also bind methylated DNA. CpG methylation of the CRE-motif (TGACGTCA) enhances the DNA binding of the C/EBPα (Rishi et al., 2010). UHRF1 and UHRF2 (SET- and Ring finger-associated proteins, SRA) bind hemimethylated CpG and the tail of histone H3 in a highly methylation sensitive manner and help assemble histones and DNA into a nucleosome after replication (Hashimoto et al., 2009).

3.2 Sources for biologically relevant validation: DNA methylation and gene expression

Nowadays there are two main hypotheses explaining DNA methylation origin during evolution. Some authors believe that methylation system arose to inactivate viruses and transposons (Walsh et al., 1998). Despite some evidences in favor of this hypothesis, most of the authors nowadays suppose that main function of DNA methylation is a control of gene expression during development and cell differentiation, most likely by influence on affinity of different protein binding.
Promoter regions of many genes are unmethylated and demonstrate resistance to increasing concentration of methylating agents (Bestor et al., 1992). Yet if promoter region become methylated this usually leads to stable in cell generations and irreversible gene suppression (Razin & Riggs, 1980; Schubeler et al., 2001). However some genes demonstrate rather high expression independently to methylation level of their promoters (Shen et al., 2007) and some promoters need to have methylated cytosine to be activated (Rishi et al., 2010).
Cytosine methylation affects transcription both directly by changing the affinity of TF binding to DNA and indirectly by forming inactive chromatin domains. Both ^{5m}C and T change DNA conformation in core positions of TFBS. For transcription repression in some cases it's enough to have one cytosine methylated, in other cases the level of expression is correlated negatively with methylation level, but is independent on the exact position of cytosine to be methylated. Inhibition of transcription caused by partial DNA methylation can be overpassed by enhancers (Hug et al., 1996), however fully methylated promoters can't be reactivated that way (Schubeler et al., 2001).
The possibility of active demethylation is still under discussion (S. C. Wu & Zhang, 2010). Cytidine deaminase AID could play a role in this process in mammals (Fritz & Papavasiliou, 2010). Recently it has been shown that elongation complex also can participate in demethylation (Okada et al., 2010). Even DNA methyl-transferases DNMT3a/b could force cytosine deamination leading to reparation of T-G mismatch pair into correct C-G pair with GC-biased reparation system (S. C. Wu & Zhang, 2010). Overexpression of MBD3 could also play a role in demethylation (S. E. Brown et al., 2008). Yet active demethylation after implantation of the embryo is very rare occasion (S. C. Wu & Zhang, 2010).

Different tissues and cell types demonstrate specific cytosine methylation patterns (Ushijima et al., 2003), those patterns in the same tissue of different individuals are similar (Lister et al., 2009), but not identical (Bock et al., 2008). Now a lot of regions with tissues-specific methylation profiles (tDMRs) are known (Rakyan et al., 2008; Brunner et al., 2009; Straussman et al., 2009; Xin et al., 2010). DMRs are likely to be involved in gene imprinting (Lopes et al., 2003). Differential activity of imprinted alleles of the gene is dependent on methylation of promoters, enhanserses or silencers of those genes (Li et al., 1993).
Females have one of the X chromosomes inactivated in somatic cells (Gartler & Riggs, 1983). The process of inactivation starts at early embryo stage with Xist activation (S. D. Brown, 1991), which leads to chromatin modification and methylation of promoters of most (Deobagkar & Chandra, 2003) but not all (Zeschnigk et al., 2009) genes. Methylation and gene repression profile of inactivated X chromosome is stable in cell generations.
Defect of normal methylation profile is a distinctive feature for different pathology conditions (Ratt syndrome, psychopathologies (Egger et al., 2004), autoimmune diseases (Richardson, 2007), hypertension (Frey, 2005)). Despite many evidences on epigenetic changes in pathologies, cancer is the most known disease having abnormalities in epigenetics, especially in DNA methylation (Jones & Baylin, 2002; Laird, 2003; Herrera et al., 2008). Tumor cells demonstrate a lot of modifications in epigenetics status: general demethylation of the genome, influencing chromatin structure, increased DNA methyl-transferase activity, and hypermethylation of promoter regions of many genes resulting in their repression. High probability of ^{5m}C to mutate into T brings about a lot of cancer-specific mutations. It's importation to notice, that pathological profiles of methylation often depend on environmental conditions and are inherited (Liu et al., 2008).

3.3 Sources for biologically relevant validation: CpG islands as promoter regions

The RNA polymerase II core promoter contains DNA motifs directing transcriptional machinery to the transcription start site (TSS). Nowadays four DNA motifs are known to be a part of core promoter: the TATA box, the TFIIB recognition element (BRE), the initiator (Inr), and the downstream promoter element (DPE) (Kutach & Kadonaga, 2000). The TATA box is an A/T-rich sequence, located about 20-30 nucleotides upstream of the TSS, that binds TFIID complex (Burley & Roeder, 1996). The BRE having the consensus SSRCGCC, is located immediately upstream of the TATA element in some promoters and increases the affinity of TFIIB binding (Lagrange et al., 1998). The Inr was originally a motif encompassing the TSS that is sufficient to direct accurate initiation in the absence of a TATA element (Smale, 1997). Inr elements are, however, present in both TATA-containing and TATA-less promoters and play a role in TFIID binding (Chalkley & Verrijzer, 1999). In mammalian promoters, the Inr consensus sequence is $RRA_{+1}NWRR$, where A_{+1} is the TSS (Bucher, 1990). The DPE acts cooperatively with the Inr helping TFIID binding and accuracy of transcription initiation in TATA-less promoters (Burley & Roeder, 1996). The DPE is located about 30 nucleotides downstream of the TSS and contains a common GWCG sequence motif.
Saxonov and colleagues (Saxonov et al., 2006) demonstrate that human genes have two different promoter types: AT-rich and GC-rich (associated with CGIs). They are easily distinguishable not only in AT- or GC content, but also in different motifs overrepresented in each promoter type. One can see that most of core promoter elements are GC-rich and could be a part of a CGI-associated promoter. CGIs are often located in 5' regions of genes, mostly overlapping with TSS (Gardiner-Garden & Frommer, 1987; Davuluri et al., 2001;

Ponger et al., 2001), and participate in regulation of transcription initiation (Rozenberg et al., 2008). Housekeeping genes tend to have CGI promoter more frequently comparing to tissue-specific genes (Zhu et al., 2008). However promoters of tissue-specific genes related to development and embryogenesis are usually located in proximity to CGIs (Robinson et al., 2004).

Many authors believe that CGIs exist since CpG dinucleotides inside them are protected from methylation. The mechanism of such protection is assumed to be protein binding at CGIs boundaries as it has been shown for Sp1 in the promoter of mouse *aprt* gene (Macleod et al., 1994). Later role of Sp1 in CGI boundaries formation has been shown for other genes (Tomatsu et al., 2002). Sp1 is often associated with CGIs as one of the key features (Macleod et al., 1994; Rozenberg et al., 2008). In one of the first works on CGI (Gardiner-Garden & Frommer, 1987) it has been shown that CGIs obtain many G/C-boxes (GGGCGG), which act as a core for Sp1 TFBS (Briggs et al., 1986). Sp1 binds both methylated and unmethylated DNA (Holler et al., 1988). Fan and colleagues (Fan et al., 2007) assume that all proteins with zinc-finger domain can play a role in CpG boundaries formation. Some other proteins, like VEZF1 (Dickson et al., 2010) and CTCF (Filippova et al., 2005; Recillas-Targa et al., 2006), also participate in this process. Naumann (Naumann et al., 2009) shows that loss of such a boundary (in fragile X-chromosome syndrome) leads to spread of methylation and gene inactivation. Moreover CGIs obtaining CTCF binding sites can themselves play a role of insulators forming boundaries of chromatin domains (Filippova et al., 2005).

Other DNA binding proteins with GC-rich binding sites can also decrease the level of DNA methylation (Lin et al., 2000; Recillas-Targa et al., 2006). It's most likely that unmethylated CpG islands form open chromatin structures simplifying the transcription (Choi, 2010). Binding sites for Cfp1 (Thomson et al., 2011), E2F (Weinmann et al., 2002), ETS, NRF-1, BoxA, CRE, E-Box (Rozenberg et al., 2008), p53 (Zemojtel et al., 2009) was found within CGIs.

Besides TFBS other DNA motifs are associated with CGI promoters. GC-skew, a feature of all unidirectional promoters, is stronger for genes starting within CGIs than for genes lacking this property (Polak et al., 2010). Tandem or simple repeats are also found within CGIs (Hutter et al., 2006). Sequence motifs $G_{2\text{-}3}C_{2\text{-}3}$, typical for CGI, induce local DNA curiosity and form G-qudruplexes at 5′ and 3′ ends of RNA molecule**.** G-quadruplexes in DNA restrict methylation of CpG dinucleotides genome-wide (Halder et al., 2010).

3.4 Sources for biologically relevant validation: CpG islands located far from promoter regions

At least 25% of CpG islands are located far from gene promoters (Ponger et al., 2001). Although a lot of such CGIs overlap with repeats, (Graff et al., 1997; Ponger et al., 2001), other CGIs don't (Ponger et al., 2001; Hackenberg et al., 2006). They are often located near 3' gene region (Gardiner-Garden & Frommer, 1987) or within the gene (Hackenberg et al., 2006). Such 3′and intragenic CGIs are subject for natural selection not only on the protein level, but also on the level of nucleic acids, which confirms their functional significance (Medvedeva et al., 2010).

Many of CGIs located far from promoters of protein-coding genes perform important biological functions. For instance, a CGI within intron 10 of *KCNQ1* acting as a promoter of antisense RNA transcript is involved into imprinting regulation of the locus (Smilinich et al., 1999). Imprinting of *MAP3K12* gene is caused by differential methylation of a CGI located in its last exon (Takada et al., 2000). Many CGI around the 3' ends of genes affect its expression in normal tissues (Appanah, Dickerson et al. 2007) and in cancer (Shiraishi et al., 2002).

Intergenic methylation plays an important role in regulation of alternative promoters (Maunakea, Nagarajan et al. 2010), modify chromatin structure (Lorincz, Dickerson et al. 2004) and influence the elongation efficiency (Jacquier, 2009).
Resenly several works show that CGIs located far from known genes in intragenic regions correspond to previously undetected promoters (Carninci et al., 2005; Medvedeva et al., 2010) playing a role during development (Illingworth et al., 2011).
CTCF insulator protein forming a boundary of chromatin active regions (Bell & Felsenfeld, 2000) often binds CCCTC core motif common within CGIs.
CpG islands and mobile elements. There are a lot of repetitive sequences in human genomes having high GC content, so many algorithms find CGI overlapping with repeats (Alu-repeat in human (Graff et al., 1997) and B1-repeat in mouse (Yates et al., 1999)). Cytosines within CGIs associated with Alu-repeats in normal cells are methylated, which in turn represses the expansion of the repeat (Xing et al., 2004). Loss of methylation in Alu-repeats is typical for tumor cells (Xie et al., 2010). Recently absence of methylation in Alu-repeats was shown for germ line (Brohede & Rand, 2006). Ullu and Tschudi (Ullu & Tschudi, 1984) believe that Alu-repeats are possessed pseudogenes of 7SL-RNA, and several Alu families still contain inner promoter of RNA polymerase III (Britten et al., 1988). One can expect that CGIs in Alu-repeats should have different DNA motifs comparing to CGIs in promoters of protein-coding genes transcribed by PolII. Nevertheless, recent studies show that pervasive PolII transcription is also a common feature for pseudogenes and transposons (Frith et al., 2006).
Alu-repeats are source of spreading DNA methylation, so unmethylated CGIs contain TFBS for Sp1 and other proteins to protect themselves from methylation (Caiafa & Zampieri, 2005). Recent studies show that Alu-repeats proximal to CpG islands could themselves form a boundary protecting CpG islands from methylation (Feltus et al., 2003).
Taking into consideration all facts mentioned above, it's obviously too early to exclude Alu- and similar repeats out of attention speaking on CGIs functionality. Most of the authors (Takai & Jones, 2002; H. Wu et al., 2010) try to build an algorithm for CGI search that avoid CGIs around Alu-repeats. There are some differences in GC content, Obs/Exp_{CpG} (Takai & Jones, 2002) or in cumulative mutual information of CpG dinucleotides (Su et al., 2009) between CGIs found near Alu-repeats and around promoters of protein-coding genes. Yet most algorithms excluded *ab initio* all repetitive sequences and therefore all of the CGIs located within them, removing more than a half of CGIs in doing so. The question remains why the same sequences in repetitive elements are of no use while in unique segments are essential.
CpG islands and replication origins. Sequence properties of replication origins in mammals are not studied very well. There are some evidences that CpG islands near 3′ region of the gene (Phi-van & Stratling, 1999) or in other genome regions can play a role of replication origins (Rein et al., 1997; Rein et al., 1999), it's important to know that some CpG should be methylated in those regions for success of replication (Rein et al., 1999).

3.5 Approches for validation

Taking into consideration biological properties mentioned above, DNA methylation is a logically relevant feature for CGI prediction validation. Complicated system of interactions involving CGIs makes it obvious that considering CGI as merely unmethylated region is an oversimplification. As far as DNA methylation plays important role in cell differentiation, the same DNA region can be unmethylated in early stage of development and methylated in later stages (reprogrammed DMR, rDMR), or unmethylated in one tissue and methylated in

another one (tissue-specific DMR, tDMR), or unmethylated in one allele and methylated in another (allele-specific DMR, aDMR) as in case of imprinting or dosage compensation, or demonstrate cross-individual differences in methylation (individual DMR, iDMR). More appropriate way is to associate CGI with DMRs demonstrating absense (or decreased level) of cytosine methylation only in one or few conditions.

Nevertheless even methylated CGIs play a role in transcription regulation, some of them contains TSS of protein-coding (Shen et al., 2007) or non-coding genes (Medvedeva et al., 2010). Recently a mechanism of transcription activation by binding of the C/EBPα transcription factor to the methylated CRE motif (TGACGTCA) was demonstated (Rishi et al., 2010). Thus, the absence of methylation shouldn't be the only criterion for CGIs verification.

Resently a lot of works dedicated to prediction of DNA methylation status in different normal tissues ((Bock et al., 2008; Zhao & Han, 2009) and refs in them) and cancer (Feltus et al., 2006) appeared. Various machine leaning techniques (support-vector machine (Bhasin et al., 2005; Das et al., 2006), alternative decision trees (Carson et al., 2008), discriminant analysis (Feltus et al., 2003)) were used to distinguish between methylated and unmethylated regions. Authors use GC content, different di- and tri nucleotides (Das et al., 2006; Fang et al., 2006), Alu-repeat location (Das et al., 2006; Fang et al., 2006), TpG fraction, TFBS, repeats, predicted DNA structures (Bock et al., 2006) and other DNA patterns and properties (Bhasin et al., 2005; Bock et al., 2007; Oakes et al., 2007; Carson et al., 2008; Ehrich et al., 2008) as parameters for those studies. Results obtained by different authors are incomparable, as in every case the model is built on distinct set of tissues and usually not in a genome-wide manner. Features demonstrating high selectivity in one work don't do the same in other works. The consistency of features is low, so one can conclude that those models are overlearned.

Promoter proximity is another traditional key feature for CGI validation. The most popular criterion is a fraction of predicted CGIs located near promoter regions of protein coding genes. As a negative set Alu-repeats are usually used. SWM with higher thresholds for length, GC content and Obs/Exp_{CpG} (Takai & Jones, 2003; Han & Zhao, 2009) and clustering algorithms (Glass et al., 2007; Hackenberg et al., 2010a; H. Wu et al., 2010) show best results. Takai-Jones algorithm predicts 40% of CGIs to be located near promoters of RefSeq genes, CpGcluster can reach the amount of 50% of all CGIs to be near promoter regions (with p-value = 1.0e-20). Wu and colleagues (H. Wu et al., 2010) believe that CGHW predicts more CGI to be located near promoters of RefSeq genes comparing to UCSC CGI and CG clusters.

Despite the fact that about half of CGIs are located near TSS of protein-coding genes the rest are not. Lately various evidences of pervasive transcription appear (Carninci et al., 2005). New high-throughput techniques (CAGE, SAGE, ets) identify at least ten times more transcriptionally active regions comparing to number of protein-coding genes. Most of those regions contain TSS for ncRNA of different types. CGIs located far from TSS of protein-coding genes can act as their promoters. Nowadays discovery of new protein-coding genes is rare occasion. Nevertheless our knowledge about ncRNA genes is extremely uncomplete. On the other side, one shouldn't forget that mammalian genomes have not only CGI-dependent promoters, but also TATA-dependent ones (Saxonov et al., 2006). The proportion of both types is still unclear. Therefore fraction of CGIs associated with protein-coding genes promoters is not an appropriate measure.

Other genomic features, like insulators, replication origins, recombination hot-spots, are also co-located with CGIs and make the whole picture more complicated. It's also becoming clear

that CGI is not functionally equipotential throughout the length. CGI is not only a region with high GC content and CpG frequency. Even in very early works on CGIs (G/C)-box was mentioned as its structure element. Currently, it's obvious that not only Sp1 but also a lot of different TFs bind DNA within CGIs, so a huge fraction of them contains TFBS and their clusters. Also, at least some CGIs have boundary regions containing binding sites for Sp1, CTCF, VEZF1 or other TFs. Recently it was shown that G-quadruplex could also form a boundary of CGIs. It should be emphasized that quality of biologically relevant feature prediction is higher, if the method uses not only CGI prediction but includes other sequence properties. Therefore the concept of complex CGI definition based not only on GC or CpG content but also on other features like TFBS, repeats or DNA structure elements looks promising.

4. Unsolved problems and perspectives

Despite the huge amount of works in the area commonly accepted definition of CpG islands still doesn't exist. Most likely such situation is a result of difficulties with biological verification of predictions (Segal, 2006). Authors of SWMs and to lower extend of clustering algorithms choose the parameters arbitrarily complicating biological interpretations. Authors of machine-learning techniques usually find too many distinguishing parameters important in their models, which are not important in modeling of similar processes in other cases.

Specifically it should be emphasized that all attempts to construct CGI prediction algorithm based on simple DNA sequence properties (GC content, Obs/Exp_{CpG}, distance between neibouring CpG dinucleotides) having in mind prediction of complex biological feature (promoter regions, unmethylated regions and so on) bring about a high level of false positive predictions. For example, in case of promoter CGI prediction at least one third of CGIs are located far from promoters. It admits of no doubt that existing CGI searchers find a chimeric class of DNA segments, which don't have single common function. A collection of DNA motifs relevant to different biological functions could result into more adequate CGI definition. For instance, GC-skew and known core promoter elements could help to find CGI or regions within them related to TSS.

Speaking on another feature of CGIs, namely lack of DNA methylation, it should be mentioned that new high-throughput techniques show that not all CpG within CGIs are unmethylated in normal cells, as previously believed. Nowadays it became clear that not only CpGs but also CpNpGs are subject to methylation (Lister et al., 2009). Such a motif also should be included in CGI prediction model (Hackenberg et al., 2010b).

The ability of a CGI searcher to predict DMRs but not unmethylated regions seems more appropriate for quality evaluation. (Dai et al., 2008; Rakyan et al., 2008; Previti et al., 2009). Unfortunately now we are still lack of high-quality and high-resolution data on genome-wide DNA methylation in different tissues, states of developmet and conditions. High-throughput techniques, like MeDIP, MeDIP-seq (Down et al., 2008), MethylCap-Seq (Brinkman et al., 2010), bisulphyte conversion based methods (RRBS (Eckhardt et al., 2006) and Methyl-seq (Lister et al., 2009)), let us hope for a complete map of DMRs in the nearest future, which will help with CGI validation.

There is a lot of evidences that methylated cytosine also could play important functional role as sites for methyl-binding proteins. We still haven't enougth relaibale data on motif preferences for all such proteins but we expect ChIP-seq (Mardis, 2007) technique to help

with the issue. There are proofs showing that it's premature to exclude Alu- and other repetitive mostly methylated sequences out of considereation speaking on CGI functions.
To resolve mentioned problems it is necessary to figure out as many biological functions associating with CGIs as possible and to find out structure elements within CGI relating to those functions or to separate CGI on several different functional groups. Such approach should result in more precise and biologically adequate CGIs definition and, therefore construction of relevant algorithm with low false positive and negative rates which in turn will improve our knowledge in genetic and epigenetic regulation of genome functioning.

5. Comparison of different algorithms

A lot of comparisons between algorithms for CGI search have been performed. This work is focused on study of various genome features potentially relates to CGIs. Three algorithms for CpG islands search participate in the comparison: UCSC CGI, CpGcluster (with p-value threshold of clusters equal to 1.0e-10, 1.0e-15, and 1.0e-20) and CGHW (the algorithm implemented by Wu and colleagues). I prefer to focus on the algorithms of a "new wave" and UCSC CGI as a reference because the last one is the most widespread now.
ENCODE regions of human genome (version hg18) were used for the study. All annotations were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/.
Standard sensitivity (3) and specificity (4) measures for prediction quality were used.

$$Sn = L_{TP} / (L_{FP}+L_{FN}), \quad (3)$$

$$Sp = L_{TN} / (L_{FP}+L_{TN}), \quad (4)$$

where L_{TP} – total length (bp) of overlap of CGIs with tested annotation, L_{FP} – total length (bp) of CGIs not overlapping with tested annotation, L_{FN} - total length (bp) of tested annotation not overlapping with CGIs, L_{TN} - total length (bp) of ENCODE regions not overlapping neither with tested annotation no with CGIs.

5.1 Basic statistics

As a first step I collected the summary of statictical properties of CGIs predicted by different algorithms. CGI HW covers more then 2.2 % of total length of all ENCODE regions. CpGcluster (p-value 1.0 e-20 as recommended in (Hackenberg et al., 2010a)) demonstrate the smallest genome coverage of 0.6%. CpGcluster predicts shorter CGIs with higher average GC-content and Obs/Exp_{CpG} value comparing to other algorithms. UCSC CGI obtains the largest average number of CpGs per one CGI.

	UCSC	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
#CGI	507	1124	1093	633	418
CGI total length	396722	685514	303160	222603	172676
avarage length	782	610	277	352	413
avarage GC content	0.66	0.64	0.7	0.71	0.72
avarage #CpG per CGI	71	48	29	38	46
avarage Obs/Exp_{CpG}	0.86	0.74	0.91	0.92	0.92
ENCODE fraction	0.0132	0.0229	0.0101	0.0074	0.0058

Table 1. Basic statistics for different CGIs.

In general one could see that CGI HW finds more "relaxed" CGIs comparing to UCSC CGI (with lower GC-content, Obs/Exp_{CpG} value and CpG frequency), whereas CpGcluster finds more "strict" CGIs comparing to UCSC CGI.

5.2 Regulatory potential

TSS prediction. It's widely accepted that a large fraction of CGIs is found around TSS of protein-coding genes. Recent studies show that total amount of TSS is about 10-times higher than the amount of protein-coding genes, so it seems more appropriate to test the CGI searchers for their ability to find TSS of any type. Several experimental techniques are able to detect any type of TSSs. Cap analysis gene expression (CAGE) is one of the most known techniques to produce a snapshot of the 5' ends of the total cellular RNA transcribed by PolII. A collection of CAGE-tags (encodeRikenCagePlus and encodeRikenCageMinus tables from UCSC) was used as a representative set of PolII TSS.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
CGI fraction	0.0136	0.0090	0.0130	0.0152	0.0164
CAGE fraction	0.7274	0.7909	0.4632	0.4331	0.3903
Sn	0.0136	0.0091	0.0128	0.0149	0.0158
Sp	0.9869	0.9773	0.9900	0.9927	0.9943

Table 2. CAGE-tags clusters within different CGIs.

Table 2 shows that CGI HW has the lowest sensitivity, although they obtain the highest fraction of CAGE-tags clusters. CpGcluster20 demonstrates the highest selectivity and specificity but obtain only 39% of CAGE-tags clusters. UCSC CGI has the intermediate values of Sn and Sp.

TFBS prediction. Although TFBS prediction is a classical problem for computational molecular biology, prediction of one single but highly reliable TFBS still remains tricky. I used TFBS conserved in the human/mouse/rat alignment based on Transfac Matrix Database (tfbsConsSites and tfbsConsFactors tables from UCSC). Keeping in mind that using of conserved TFBS leads to omission of all types of species-specific regulation regions, conserved TFBS are more likely to be functional comparing to other predicted TFBS.

Table 3 demonstrates that CpGcluster predicts CGI with fewer different TFs and lower sensitivity comparing to USCS CGI and CGI HW. The highest fraction of total TFBS length is covered by CGI HW, the very same algorithm shows the highest sensitivity and the lowest specificity. It's not obvious what fraction of the CGIs one should expect to be covered by TFBS but CpGcluster20 demonstrates the largest coverage (about 19 %).

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
#TF	167	167	161	154	153
CGI fraction	0.1834	0.1347	0.1896	0.1915	0.1917
TFBS fraction	0.0860	0.1098	0.0688	0.0509	0.0393
Sn	0.0676	0.0696	0.0567	0.0443	0.0355
Sp	0.9889	0.9796	0.9916	0.9938	0.9952

Table 3. Conserved TFBS within different CGIs.

As it's difficult to estimate the expected coverage of TFBS, I compared the coverage of CGIs with the coverage of their adjacent regions of 100 bp. Results in Table 4 show that all adjacent to CGI regions contain conserved TFBS.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
#TF	157	167	166	162	151
CGI fraction	0.0564	0.0648	0.0871	0.0859	0.0820
TFBS fraction	0.0069	0.0177	0.0231	0.0132	0.0083
Sn	0.0063	0.0143	0.0189	0.0117	0.0077
Sp	0.9967	0.9928	0.9931	0.9960	0.9974
TFBS ratio	12.38	6.21	2.98	3.86	4.72

Table 4. Conserved TFBS within +/- 100 bp around different CGIs.

Last row of the Table 4 demonstrates the reduction of coverage in CGI adjacent regions comparing to CGI bodies. The adjacent regions of UCSC CGI and CGI HW contain more then 12 and 6 times less TFBS comparing to CGI body respectively. One should expect some TFBS around CGI which can function as CGI's boundaries. One the other hand, if we believe that CGI itself is the regulatory region, expected amount of TFBS in the adjacent regions should be dramatically lower comparing to CGI body, which is not the case for CpGcluster.

Insulators. CTCF is well known as a DNA binding protein acting both as transcriptional factor and insulator protein. To test which CGI prediction algorithm finds more CTCF binding sites I used data on CTCF binding (oregano and oreganoAttr tables from UCSC). One can see that CGI HW shows the highest sensitivity in CTCF binding prediction. It's also important to mention that CGIs from CGI HW contain more than 25% of all CTCF sites. CpGcluster10 shows the second best result, and the quality of prediction decreases in case of CpGcluster15 and CpGcluster20.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
CGI fraction	0.0809	0.0658	0.0503	0.0569	0.0478
CTCF fraction	0.1395	0.2517	0.1871	0.1224	0.0680
Sn	0.0267	0.0434	0.0305	0.0241	0.0157
Sp	0.9872	0.9806	0.9916	0.9939	0.9953

Table 5. CTCF binding sites within different CGIs.

DNase sensitivity regions. DNase sensitivity regions are often considered as regions of open chromatin which correspond to regulatory regions of all types. To test what algorithm predicts CGI more often associated with DNase sensitivity regions I use joined data for several tissues available in UCSC (table wgEncodeRegDnaseClustered). All CGIs demonstrate rather good association with DNase sensitivity regions, at least one third of their length is located in sensitive area. UCSC CGI shows highest sensitivity and rather good spesifisity. Vast fraction of CpGcluster CGIs are also associated with DNase sensitivity regions; althougth sensivity of the algorithm is not very good.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
CGI fraction	0.6047	0.3221	0.4312	0.5872	0.6040
DNase fraction	0.0768	0.0707	0.0418	0.0418	0.0334
Sn	0.0789	0.0655	0.0413	0.0424	0.0338
Sp	0.9942	0.9827	0.9936	0.9966	0.9975

Table 6. DNase sensitivity regions within CGIs predicted by different algorithms and quality of prediction.

Differently methylated regions. Data on regions differently methylated during development was downloaded from the UCSC (table rdmr). Table 7 shows that CGI HW predicts CGI located near over 43% of all rDMRs. This algorithm demonstrates also the best sensitivity in this case. It shoud be menthioned that CpGcluster20 has the lowest sensitivity and those CGIs are located near only 7% of rDMRs.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
#rDMR fraction	0.2500	0.4310	0.2241	0.1293	0.0776
CGI fraction	0.0170	0.0262	0.0179	0.0161	0.0137
rDMR fraction	0.0534	0.1424	0.0432	0.0284	0.0187
Sn	0.0132	0.0231	0.0130	0.0105	0.0080
Sp	0.9869	0.9776	0.9900	0.9927	0.9943

Table 7. rDMRs within different CGIs.

Replication origins. To figure out if there is any preference for replication origins to be found by one of CGI searchers data from encodeUvaDnaRepOriginsNSGM table were used. Only CGI HW and CpGcluster10 find 5 and 2 replication origins within or around (+/- 100 bp) CGI respectively. Other algorithms (and CpGcluster with more strict parameters) are unable to find any replication origins.

Polymorphic loci. Data from SNP130 were used for study of polymorphic loci within different CGIs. CGI from CGI HW contains the highest fraction of SNPs and demonstrates highest sensitivity, so one should expect more interindividual variants within those CGIs.

	UCSC CGI	CGI HW	CpGcluster10	CpGcluster15	CpGcluster20
CGI fraction	0.0072	0.0082	0.0080	0.0073	0.0066
SNP fraction	0.0140	0.0276	0.0120	0.0080	0.0056
Sn	0.0048	0.0064	0.0049	0.0038	0.0031
Sp	0.9868	0.9771	0.9899	0.9926	0.9942

Table 8. SNPs within different CGIs.

6. Conclusions

In summary, no one algorithm for CGI search predicts all biologically relevant features with appropriate accuracy. In all cases a lot of both false positives and false negatives appear.

All algorithms participating in competition have its strong sides. CpGcluster (p-value = 1.0e-15 and p-value = 1.0e-20) demonstrate the highest specificity in TSS prediction. Although such CGIs obtain the smallest fraction of CAGE-tags, this may be not a disadvantage as we don't know for sure the proportion of GC- and AT-rich promoters. The largest fraction of CGIs length is covered by TFBS in case of CGIs predicted by CpGcluster, on the other hand the largest part of their adjacent regions is also covered by TFBS. This brought me to conclusion that CpGcluster finds "cropped" promoter CGIs, espessially in case of p-value = 1.0e-20.
On the contrary CGI HW demonstrates the best sensitivity in CTCF binding sites and rDMR prediction. CGI from CGI HW are associated with at least some of origins of repliacation, thereas other algoritms (with recommended parameters) don't. They are also more prone to find diversities between humans. Also those CGIs find the highest fraction of TSS. So, CGI HW finds regions with broad regulatory potential. However all those features are related to DNA methylation, which allow me to assume that CGI HW finds DMR-associated CGIs.
UCSC CGI demonstrates moderate behavior. This algorithm has intermediate sensitivity both in TSS and rDMR prediction. Those CGIs have the highest decrease of TFBS in CGI adjasent regions and the highest sensitivity to DNase. It looks like UCSC finds CGI around promoter and also includes regulation regions, so those are promoter region CGIs.
It's quite clear that CGI is a complex object, which doesn't correspond to any single biological feature. It seems more appropriate to segregate a class of interconnected biological features: differential DNA methylation, active transcription at least in one cell type or development stage and replication. CGI HW algorithm made the first step in this direction, whereas CpGcluster (with high threshold for p-value) moves to the opposite direction and finds specific narrow class of promoters. Traditional UCSC approach still stands ground demonstrating comparable or in some points even higher quality. Hence the CpG island problem is still far from final solution.

7. Acknowledgments

Author is very grateful to N. Oparina, V. Makeev, I. Artamonova and A. Favorov for fruitful discussions on the topic of this article. This study was partially supported by RFBR grant 11-04-02016-a and by the state contract P1376 of the Federal Special Program "Scientific and educational human resources of innovative Russia" for 2009 – 2013.

8. References

Baylin, S. B., J. G. Herman, J. R. Graff, P. M. Vertino and J. P. Issa (1998). Alterations in DNA methylation: a fundamental aspect of neoplasia. *Adv Cancer Res*, Vol.72, 1998), pp. 141-96

Behe, M. and G. Felsenfeld (1981). Effects of methylation on a synthetic polynucleotide: the B--Z transition in poly(dG-m5dC).poly(dG-m5dC). *Proc Natl Acad Sci U S A*, Vol.78, No.3, (Mar, 1981), pp. 1619-23

Bell, A. C. and G. Felsenfeld (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, Vol.405, No.6785, (May 25, 2000), pp. 482-5

Berger, J. and A. Bird (2005). Role of MBD2 in gene regulation and tumorigenesis. *Biochem Soc Trans*, Vol.33, No.Pt 6, (Dec, 2005), pp. 1537-40

Bestor, T. H., G. Gundersen, A. B. Kolsto and H. Prydz (1992). CpG islands in mammalian gene promoters are inherently resistant to de novo methylation. *Genet Anal Tech Appl*, Vol.9, No.2, (Apr, 1992), pp. 48-53

Bhasin, M., H. Zhang, E. L. Reinherz and P. A. Reche (2005). Prediction of methylated CpGs in DNA sequences using a support vector machine. *FEBS Lett*, Vol.579, No.20, (Aug 15, 2005), pp. 4302-8

Bird, A. P. (1986). CpG-rich islands and the function of DNA methylation. *Nature*, Vol.321, No.6067, (May 15-21, 1986), pp. 209-13

Bock, C., M. Paulsen, S. Tierling, T. Mikeska, T. Lengauer and J. Walter (2006). CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet*, Vol.2, No.3, (Mar, 2006), pp. e26

Bock, C., J. Walter, M. Paulsen and T. Lengauer (2007). CpG island mapping by epigenome prediction. *PLoS Comput Biol*, Vol.3, No.6, (Jun, 2007), pp. e110

Bock, C., J. Walter, M. Paulsen and T. Lengauer (2008). Inter-individual variation of DNA methylation and its implications for large-scale epigenome mapping. *Nucleic Acids Res*, Vol.36, No.10, (Jun, 2008), pp. e55

Brero, A., H. Leonhardt and M. C. Cardoso (2006). Replication and translation of epigenetic information. *Curr Top Microbiol Immunol*, Vol.301, 2006), pp. 21-44

Briggs, M. R., J. T. Kadonaga, S. P. Bell and R. Tjian (1986). Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. *Science*, Vol.234, No.4772, (Oct 3, 1986), pp. 47-52

Brinkman, A. B., F. Simmer, K. Ma, A. Kaan, J. Zhu and H. G. Stunnenberg (2010). Whole-genome DNA methylation profiling using MethylCap-seq. *Methods*, Vol.52, No.3, (Nov, 2010), pp. 232-6

Britten, R. J., W. F. Baron, D. B. Stout and E. H. Davidson (1988). Sources and evolution of human Alu repeated sequences. *Proc Natl Acad Sci U S A*, Vol.85, No.13, (Jul, 1988), pp. 4770-4

Brohede, J. and K. N. Rand (2006). Evolutionary evidence suggests that CpG island-associated Alus are frequently unmethylated in human germline. *Hum Genet*, Vol.119, No.4, (May, 2006), pp. 457-8

Brown, S. D. (1991). XIST and the mapping of the X chromosome inactivation centre. *Bioessays*, Vol.13, No.11, (Nov, 1991), pp. 607-12

Brown, S. E., M. J. Suderman, M. Hallett and M. Szyf (2008). DNA demethylation induced by the methyl-CpG-binding domain protein MBD3. *Gene*, Vol.420, No.2, (Sep 1, 2008), pp. 99-106

Brunner, A. L., D. S. Johnson, S. W. Kim, A. Valouev, T. E. Reddy, N. F. Neff, E. Anton, C. Medina, L. Nguyen, E. Chiao, et al. (2009). Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res*, Vol.19, No.6, (Jun, 2009), pp. 1044-56

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J Mol Biol*, Vol.212, No.4, (Apr 20, 1990), pp. 563-78

Burley, S. K. and R. G. Roeder (1996). Biochemistry and structural biology of transcription factor IID (TFIID). *Annu Rev Biochem*, Vol.65, 1996), pp. 769-99

Caiafa, P. and M. Zampieri (2005). DNA methylation and chromatin structure: the puzzling CpG islands. *J Cell Biochem*, Vol.94, No.2, (Feb 1, 2005), pp. 257-65

Carninci, P., T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, et al. (2005). The transcriptional landscape of the mammalian genome. *Science*, Vol.309, No.5740, (Sep 2, 2005), pp. 1559-63

Carson, M. B., R. Langlois and H. Lu (2008). Mining knowledge for the methylation status of CpG islands using alternating decision trees. *Conf Proc IEEE Eng Med Biol Soc*, Vol.2008, 2008), pp. 3787-90

Chalkley, G. E. and C. P. Verrijzer (1999). DNA binding site selection by RNA polymerase II TAFs: a TAF(II)250-TAF(II)150 complex recognizes the initiator. *EMBO J*, Vol.18, No.17, (Sep 1, 1999), pp. 4835-45

Choi, J. K. (2010). Contrasting chromatin organization of CpG islands and exons in the human genome. *Genome Biol*, Vol.11, No.7, 2010), pp. R70

Cooper, D. N., M. Mort, P. D. Stenson, E. V. Ball and N. A. Chuzhanova (2010). Methylation-mediated deamination of 5-methylcytosine appears to give rise to mutations causing human inherited disease in CpNpG trinucleotides, as well as in CpG dinucleotides. *Hum Genomics*, Vol.4, No.6, (Aug 1, 2010), pp. 406-10

Cross, S. H., J. A. Charlton, X. Nan and A. P. Bird (1994). Purification of CpG islands using a methylated DNA binding column. *Nat Genet*, Vol.6, No.3, (Mar, 1994), pp. 236-44

Dai, W., J. M. Teodoridis, J. Graham, C. Zeller, T. H. Huang, P. Yan, J. K. Vass, R. Brown and J. Paul (2008). Methylation Linear Discriminant Analysis (MLDA) for identifying differentially methylated CpG islands. *BMC Bioinformatics*, Vol.9, pp. 337

Daniel, J. M., C. M. Spring, H. C. Crawford, A. B. Reynolds and A. Baig (2002). The p120(ctn)-binding partner Kaiso is a bi-modal DNA-binding protein that recognizes both a sequence-specific consensus and methylated CpG dinucleotides. *Nucleic Acids Res*, Vol.30, No.13, (Jul 1, 2002), pp. 2911-9

Das, R., N. Dimitrova, Z. Xuan, R. A. Rollins, F. Haghighi, J. R. Edwards, J. Ju, T. H. Bestor and M. Q. Zhang (2006). Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A*, Vol.103, No.28, (Jul 11, 2006), pp. 10713-6

Davuluri, R. V., I. Grosse and M. Q. Zhang (2001). Computational identification of promoters and first exons in the human genome. *Nat Genet*, Vol.29, No.4, (Dec, 2001), pp. 412-7

Deobagkar, D. D. and H. S. Chandra (2003). The inactive X chromosome in the human female is enriched in 5-methylcytosine to an unusual degree and appears to contain more of this modified nucleotide than the remainder of the genome. *J Genet*, Vol.82, No.1-2, (Apr-Aug, 2003), pp. 13-6

Dhasarathy, A. and P. A. Wade (2008). The MBD protein family-reading an epigenetic mark? *Mutat Res*, Vol.647, No.1-2, (Dec 1, 2008), pp. 39-43

Dickson, J., H. Gowher, R. Strogantsev, M. Gaszner, A. Hair, G. Felsenfeld and A. G. West (2010). VEZF1 elements mediate protection from DNA methylation. *PLoS Genet*, Vol.6, No.1, (Jan, 2010), pp. e1000804

Down, T. A., V. K. Rakyan, D. J. Turner, P. Flicek, H. Li, E. Kulesha, S. Graf, N. Johnson, J. Herrero, E. M. Tomazou, et al. (2008). A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. *Nat Biotechnol*, Vol.26, No.7, (Jul, 2008), pp. 779-85

Eckhardt, F., J. Lewin, R. Cortese, V. K. Rakyan, J. Attwood, M. Burger, J. Burton, T. V. Cox, R. Davies, T. A. Down, et al. (2006). DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet*, Vol.38, No.12, (Dec, 2006), pp. 1378-85

Egger, G., G. Liang, A. Aparicio and P. A. Jones (2004). Epigenetics in human disease and prospects for epigenetic therapy. *Nature*, Vol.429, No.6990, (May 27, 2004), pp. 457-63

Ehrich, M., J. Turner, P. Gibbs, L. Lipton, M. Giovanneti, C. Cantor and D. van den Boom (2008). Cytosine methylation profiling of cancer cell lines. *Proc Natl Acad Sci U S A*, Vol.105, No.12, (Mar 25, 2008), pp. 4844-9

Ehrlich, M. and R. Y. Wang (1981). 5-Methylcytosine in eukaryotic DNA. *Science*, Vol.212, No.4501, (Jun 19, 1981), pp. 1350-7

Fan, S., F. Fang, X. Zhang and M. Q. Zhang (2007). Putative zinc finger protein binding sites are over-represented in the boundaries of methylation-resistant CpG islands in the human genome. *PLoS One*, Vol.2, No.11, pp. e1184

Fang, F., S. Fan, X. Zhang and M. Q. Zhang (2006). Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, Vol.22, No.18, (Sep 15, 2006), pp. 2204-9

Fatemi, M. and P. A. Wade (2006). MBD family proteins: reading the epigenetic code. *J Cell Sci*, Vol.119, No.Pt 15, (Aug 1, 2006), pp. 3033-7

Feltus, F. A., E. K. Lee, J. F. Costello, C. Plass and P. M. Vertino (2003). Predicting aberrant CpG island methylation. *Proc Natl Acad Sci U S A*, Vol.100, No.21, (Oct 14, 2003), pp. 12253-8

Feltus, F. A., E. K. Lee, J. F. Costello, C. Plass and P. M. Vertino (2006). DNA motifs associated with aberrant CpG island methylation. *Genomics*, Vol.87, No.5, (May, 2006), pp. 572-9

Filippova, G. N., M. K. Cheng, J. M. Moore, J. P. Truong, Y. J. Hu, D. K. Nguyen, K. D. Tsuchiya and C. M. Disteche (2005). Boundaries between chromosomal domains of X inactivation and escape bind CTCF and lack CpG methylation during early development. *Dev Cell*, Vol.8, No.1, (Jan, 2005), pp. 31-42

Frey, F. J. (2005). Methylation of CpG islands: potential relevance for hypertension and kidney diseases. *Nephrol Dial Transplant*, Vol.20, No.5, (May, 2005), pp. 868-9

Frith, M. C., L. G. Wilming, A. Forrest, H. Kawaji, S. L. Tan, C. Wahlestedt, V. B. Bajic, C. Kai, J. Kawai, P. Carninci, et al. (2006). Pseudo-messenger RNA: phantoms of the transcriptome. *PLoS Genet*, Vol.2, No.4, (Apr, 2006), pp. e23

Fritz, E. L. and F. N. Papavasiliou (2010). Cytidine deaminases: AIDing DNA demethylation? *Genes Dev*, Vol.24, No.19, (Oct 1, 2010), pp. 2107-14

Gardiner-Garden, M. and M. Frommer (1987). CpG islands in vertebrate genomes. *J Mol Biol*, Vol.196, No.2, (Jul 20, 1987), pp. 261-82

Gartler, S. M. and A. D. Riggs (1983). Mammalian X-chromosome inactivation. *Annu Rev Genet*, Vol.17, pp. 155-90

Glass, J. L., R. F. Thompson, B. Khulan, M. E. Figueroa, E. N. Olivier, E. J. Oakley, G. Van Zant, E. E. Bouhassira, A. Melnick, A. Golden, et al. (2007). CG dinucleotide clustering is a species-specific property of the genome. *Nucleic Acids Res*, Vol.35, No.20, pp. 6798-807

Graff, J. R., J. G. Herman, S. Myohanen, S. B. Baylin and P. M. Vertino (1997). Mapping patterns of CpG island methylation in normal and neoplastic cells implicates both upstream and downstream regions in de novo methylation. *J Biol Chem*, Vol.272, No.35, (Aug 29, 1997), pp. 22322-9

Hackenberg, M., G. Barturen, P. Carpena, P. L. Luque-Escamilla, C. Previti and J. L. Oliver (2010a). Prediction of CpG-island function: CpG clustering vs. sliding-window methods. *BMC Genomics*, Vol.11, pp. 327

Hackenberg, M., P. Carpena, P. Bernaola-Galvan, G. Barturen, A. M. Alganza and J. L. Oliver (2010b). WordCluster: detecting clusters of DNA words and genomic elements. *Algorithms Mol Biol*, Vol.6, pp. 2

Hackenberg, M., C. Previti, P. L. Luque-Escamilla, P. Carpena, J. Martinez-Aroza and J. L. Oliver (2006). CpGcluster: a distance-based algorithm for CpG-island detection. *BMC Bioinformatics*, Vol.7, pp. 446

Halder, R., K. Halder, P. Sharma, G. Garg, S. Sengupta and S. Chowdhury (2010). Guanine quadruplex DNA structure restricts methylation of CpG dinucleotides genome-wide. *Mol Biosyst*, Vol.6, No.12, (Dec 8, 2010), pp. 2439-47

Han, L. and Z. Zhao (2009). CpG islands or CpG clusters: how to identify functional GC-rich regions in a genome? *BMC Bioinformatics*, Vol.10, pp. 65

Hashimoto, H., J. R. Horton, X. Zhang and X. Cheng (2009). UHRF1, a modular multi-domain protein, regulates replication-coupled crosstalk between DNA methylation and histone modifications. *Epigenetics*, Vol.4, No.1, (Jan, 2009), pp. 8-14

Herrera, L. A., D. Prada, M. A. Andonegui and A. Duenas-Gonzalez (2008). The epigenetic origin of aneuploidy. *Curr Genomics*, Vol.9, No.1, (Mar, 2008), pp. 43-50

Holler, M., G. Westin, J. Jiricny and W. Schaffner (1988). Sp1 transcription factor binds DNA and activates transcription even when the binding site is CpG methylated. *Genes Dev*, Vol.2, No.9, (Sep, 1988), pp. 1127-35

Hug, M., J. Silke, O. Georgiev, S. Rusconi, W. Schaffner and K. Matsuo (1996). Transcriptional repression by methylation: cooperativity between a CpG cluster in the promoter and remote CpG-rich regions. *FEBS Lett*, Vol.379, No.3, (Feb 5, 1996), pp. 251-4

Hutter, B., V. Helms and M. Paulsen (2006). Tandem repeats in the CpG islands of imprinted genes. *Genomics*, Vol.88, No.3, (Sep, 2006), pp. 323-32

Illingworth, R. S., U. Gruenewald-Schneider, S. Webb, A. R. Kerr, K. D. James, D. J. Turner, C. Smith, D. J. Harrison, R. Andrews and A. P. Bird (2010) Orphan CpG islands identify numerous conserved promoters in the mammalian genome. *PLoS Genet*, Vol.6, No.9, (Sep 23, 2010), e1001134

Irizarry, R. A., H. Wu and A. P. Feinberg (2009). A species-generalized probabilistic model-based definition of CpG islands. *Mamm Genome*, Vol.20, No.9-10, (Sep-Oct, 2009), pp. 674-80

Jacquier, A. (2009). The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat Rev Genet*, Vol.10, No.12, (Dec, 2009), pp. 833-44

Jones, P. A. and S. B. Baylin (2002). The fundamental role of epigenetic events in cancer. *Nat Rev Genet*, Vol.3, No.6, (Jun, 2002), pp. 415-28

Kimura, H. and K. Shiota (2003). Methyl-CpG-binding protein, MeCP2, is a target molecule for maintenance DNA methyltransferase, Dnmt1. *J Biol Chem*, Vol.278, No.7, (Feb 14, 2003), pp. 4806-12

Klose, R. J., S. A. Sarraf, L. Schmiedeberg, S. M. McDermott, I. Stancheva and A. P. Bird (2005). DNA binding selectivity of MeCP2 due to a requirement for A/T sequences adjacent to methyl-CpG. *Mol Cell*, Vol.19, No.5, (Sep 2, 2005), pp. 667-78

Kutach, A. K. and J. T. Kadonaga (2000). The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. *Mol Cell Biol*, Vol.20, No.13, (Jul, 2000), pp. 4754-64

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg and R. H. Ebright (1998). New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. *Genes Dev*, Vol.12, No.1, (Jan 1, 1998), pp. 34-44

Laird, P. W. (2003). The power and the promise of DNA methylation markers. *Nat Rev Cancer*, Vol.3, No.4, (Apr, 2003), pp. 253-66

Li, E., C. Beard and R. Jaenisch (1993). Role for DNA methylation in genomic imprinting. *Nature*, Vol.366, No.6453, (Nov 25, 1993), pp. 362-5

Lin, I. G., T. J. Tomzynski, Q. Ou and C. L. Hsieh (2000). Modulation of DNA binding protein affinity directly affects target site demethylation. *Mol Cell Biol*, Vol.20, No.7, (Apr, 2000), pp. 2343-9

Lister, R., M. Pelizzola, R. H. Dowen, R. D. Hawkins, G. Hon, J. Tonti-Filippini, J. R. Nery, L. Lee, Z. Ye, Q. M. Ngo, et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, Vol.462, No.7271, (Nov 19, 2009), pp. 315-22

Liu, L., Y. Li and T. O. Tollefsbol (2008). Gene-environment interactions and epigenetic basis of human diseases. *Curr Issues Mol Biol*, Vol.10, No.1-2, pp. 25-36

Lopes, S., A. Lewis, P. Hajkova, W. Dean, J. Oswald, T. Forne, A. Murrell, M. Constancia, M. Bartolomei, J. Walter, et al. (2003). Epigenetic modifications in an imprinting cluster are controlled by a hierarchy of DMRs suggesting long-range chromatin interactions. *Hum Mol Genet*, Vol.12, No.3, (Feb 1, 2003), pp. 295-305

Macleod, D., J. Charlton, J. Mullins and A. P. Bird (1994). Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes Dev*, Vol.8, No.19, (Oct 1, 1994), pp. 2282-92

Mardis, E. R. (2007). ChIP-seq: welcome to the new frontier. *Nat Methods*, Vol.4, No.8, (Aug, 2007), pp. 613-4

Medvedeva, Y. A., M. V. Fridman, N. J. Oparina, D. B. Malko, E. O. Ermakova, I. V. Kulakovskiy, A. Heinzel and V. J. Makeev (2010). Intergenic, gene terminal, and intragenic CpG islands in the human genome. *BMC Genomics*, Vol.11, No.1, (Jan 19, 2010), pp. 48

Naumann, A., N. Hochstein, S. Weber, E. Fanning and W. Doerfler (2009). A distinct DNA-methylation boundary in the 5'- upstream sequence of the FMR1 promoter binds nuclear proteins and is lost in fragile X syndrome. *Am J Hum Genet*, Vol.85, No.5, (Nov, 2009), pp. 606-16

Ng, H. H., Y. Zhang, B. Hendrich, C. A. Johnson, B. M. Turner, H. Erdjument-Bromage, P. Tempst, D. Reinberg and A. Bird (1999). MBD2 is a transcriptional repressor belonging to the MeCP1 histone deacetylase complex. *Nat Genet*, Vol.23, No.1, (Sep, 1999), pp. 58-61

Oakes, C. C., S. La Salle, D. J. Smiraglia, B. Robaire and J. M. Trasler (2007). A unique configuration of genome-wide DNA methylation patterns in the testis. *Proc Natl Acad Sci U S A*, Vol.104, No.1, (Jan 2, 2007), pp. 228-33

Okada, Y., K. Yamagata, K. Hong, T. Wakayama and Y. Zhang (2010). A role for the elongator complex in zygotic paternal genome demethylation. *Nature*, Vol.463, No.7280, (Jan 28, 2010), pp. 554-8

Phi-van, L. and W. H. Stratling (1999). An origin of bidirectional DNA replication is located within a CpG island at the 3" end of the chicken lysozyme gene. *Nucleic Acids Res*, Vol.27, No.15, (Aug 1, 1999), pp. 3009-17

Polak, P., R. Querfurth and P. F. Arndt (2010). The evolution of transcription-associated biases of mutations across vertebrates. *BMC Evol Biol*, Vol.10, pp. 187

Ponger, L., L. Duret and D. Mouchiroud (2001). Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res*, Vol.11, No.11, (Nov, 2001), pp. 1854-60

Ponger, L. and D. Mouchiroud (2002). CpGProD: identifying CpG islands associated with transcription start sites in large genomic mammalian sequences. *Bioinformatics*, Vol.18, No.4, (Apr, 2002), pp. 631-3

Previti, C., O. Harari, I. Zwir and C. del Val (2009). Profile analysis and prediction of tissue-specific CpG island methylation classes. *BMC Bioinformatics*, Vol.10, pp. 116

Rakyan, V. K., T. A. Down, N. P. Thorne, P. Flicek, E. Kulesha, S. Graf, E. M. Tomazou, L. Backdahl, N. Johnson, M. Herberth, et al. (2008). An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs). *Genome Res*, Vol.18, No.9, (Sep, 2008), pp. 1518-29

Razin, A. and A. D. Riggs (1980). DNA methylation and gene function. *Science*, Vol.210, No.4470, (Nov 7, 1980), pp. 604-10

Recillas-Targa, F., I. A. De La Rosa-Velazquez, E. Soto-Reyes and L. Benitez-Bribiesca (2006). Epigenetic boundaries of tumour suppressor gene promoters: the CTCF connection and its role in carcinogenesis. *J Cell Mol Med*, Vol.10, No.3, (Jul-Sep, 2006), pp. 554-68

Rein, T., T. Kobayashi, M. Malott, M. Leffak and M. L. DePamphilis (1999). DNA methylation at mammalian replication origins. *J Biol Chem*, Vol.274, No.36, (Sep 3, 1999), pp. 25792-800

Rein, T., H. Zorbas and M. L. DePamphilis (1997). Active mammalian replication origins are associated with a high-density cluster of mCpG dinucleotides. *Mol Cell Biol*, Vol.17, No.1, (Jan, 1997), pp. 416-26

Rice, P., I. Longden and A. Bleasby (2000). EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*, Vol.16, No.6, (Jun, 2000), pp. 276-7

Richardson, B. (2007). Primer: epigenetics of autoimmunity. *Nat Clin Pract Rheumatol*, Vol.3, No.9, (Sep, 2007), pp. 521-7

Rishi, V., P. Bhattacharya, R. Chatterjee, J. Rozenberg, J. Zhao, K. Glass, P. Fitzgerald and C. Vinson (2010). CpG methylation of half-CRE sequences creates C/EBPalpha binding sites that activate some tissue-specific genes. *Proc Natl Acad Sci U S A*, Vol.107, No.47, (Nov 23, 2010), pp. 20311-6

Robinson, P. N., U. Bohme, R. Lopez, S. Mundlos and P. Nurnberg (2004). Gene-Ontology analysis reveals association of tissue-specific 5' CpG-island genes with development and embryogenesis. *Hum Mol Genet*, Vol.13, No.17, (Sep 1, 2004), pp. 1969-78

Rozenberg, J. M., A. Shlyakhtenko, K. Glass, V. Rishi, M. V. Myakishev, P. C. FitzGerald and C. Vinson (2008). All and only CpG containing sequences are enriched in promoters abundantly bound by RNA polymerase II in multiple tissues. *BMC Genomics*, Vol.9, No.1, pp. 67

Saito, M. and F. Ishikawa (2002). The mCpG-binding domain of human MBD3 does not bind to mCpG but interacts with NuRD/Mi2 components HDAC1 and MTA2. *J Biol Chem*, Vol.277, No.38, (Sep 20, 2002), pp. 35434-9

Sasai, N., M. Nakao and P. A. Defossez (2010). Sequence-specific recognition of methylated DNA by human zinc-finger proteins. *Nucleic Acids Res*, Vol.38, No.15, (Aug, 2010), pp. 5015-22

Saxonov, S., P. Berg and D. L. Brutlag (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc Natl Acad Sci U S A*, Vol.103, No.5, (Jan 31, 2006), pp. 1412-7

Schubeler, D., M. C. Lorincz and M. Groudine (2001). Targeting silence: the use of site-specific recombination to introduce in vitro methylated DNA into the genome. *Sci STKE*, Vol.2001, No.83, (May 22, 2001), pp. pl1

Segal, M. R. (2006). Validation in genomics: CpG island methylation revisited. *Stat Appl Genet Mol Biol*, Vol.5, Article29

Shen, L., Y. Kondo, Y. Guo, J. Zhang, L. Zhang, S. Ahmed, J. Shu, X. Chen, R. A. Waterland and J. P. Issa (2007). Genome-wide profiling of DNA methylation reveals a class of normally methylated CpG island promoters. *PLoS Genet*, Vol.3, No.10, (Oct, 2007), pp. 2023-36

Shiraishi, M., A. Sekiguchi, M. J. Terry, A. J. Oates, Y. Miyamoto, Y. H. Chuu, M. Munakata and T. Sekiya (2002). A comprehensive catalog of CpG islands methylated in human lung adenocarcinomas for the identification of tumor suppressor genes. *Oncogene*, Vol.21, No.23, (May 23, 2002), pp. 3804-13

Smale, S. T. (1997). Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta*, Vol.1351, No.1-2, (Mar 20, 1997), pp. 73-88

Smilinich, N. J., C. D. Day, G. V. Fitzpatrick, G. M. Caldwell, A. C. Lossie, P. R. Cooper, A. C. Smallwood, J. A. Joyce, P. N. Schofield, W. Reik, et al. (1999). A maternally methylated CpG island in KvLQT1 is associated with an antisense paternal transcript and loss of imprinting in Beckwith-Wiedemann syndrome. *Proc Natl Acad Sci U S A*, Vol.96, No.14, (Jul 6, 1999), pp. 8064-9

Straussman, R., D. Nejman, D. Roberts, I. Steinfeld, B. Blum, N. Benvenisty, I. Simon, Z. Yakhini and H. Cedar (2009). Developmental programming of CpG island methylation profiles in the human genome. *Nat Struct Mol Biol*, Vol.16, No.5, (May, 2009), pp. 564-71

Su, J., Y. Zhang, J. Lv, H. Liu, X. Tang, F. Wang, Y. Qi, Y. Feng and X. Li (2009). CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. *Nucleic Acids Res*, Vol.38, No.1, (Jan, 2009), pp. e6

Takada, S., M. Tevendale, J. Baker, P. Georgiades, E. Campbell, T. Freeman, M. H. Johnson, M. Paulsen and A. C. Ferguson-Smith (2000). Delta-like and gtl2 are reciprocally expressed, differentially methylated linked imprinted genes on mouse chromosome 12. *Curr Biol*, Vol.10, No.18, (Sep 21, 2000), pp. 1135-8

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*, Vol.99, No.6, (Mar 19, 2002), pp. 3740-5

Takai, D. and P. A. Jones (2003). The CpG island searcher: a new WWW resource. *In Silico Biol*, Vol.3, No.3, pp. 235-40

Thomson, J. P., P. J. Skene, J. Selfridge, T. Clouaire, J. Guy, S. Webb, A. R. Kerr, A. Deaton, R. Andrews, K. D. James, et al. CpG islands influence chromatin structure via the CpG-binding protein Cfp1. *Nature*, Vol.464, No.7291, (Apr 15), pp. 1082-6

Tomatsu, S., K. O. Orii, M. R. Islam, G. N. Shah, J. H. Grubb, K. Sukegawa, Y. Suzuki, T. Orii, N. Kondo and W. S. Sly (2002). Methylation patterns of the human beta-glucuronidase gene locus: boundaries of methylation and general implications for frequent point mutations at CpG dinucleotides. *Genomics*, Vol.79, No.3, (Mar, 2002), pp. 363-75

Ullu, E. and C. Tschudi (1984). Alu sequences are processed 7SL RNA genes. *Nature*, Vol.312, No.5990, (Nov 8-14, 1984), pp. 171-2

Ushijima, T., N. Watanabe, E. Okochi, A. Kaneda, T. Sugimura and K. Miyamoto (2003). Fidelity of the methylation pattern and its variation in the genome. *Genome Res*, Vol.13, No.5, (May, 2003), pp. 868-74

van Roy, F. M. and P. D. McCrea (2005). A role for Kaiso-p120ctn complexes in cancer? *Nat Rev Cancer*, Vol.5, No.12, (Dec, 2005), pp. 956-64

Walsh, C. P., J. R. Chaillet and T. H. Bestor (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nat Genet*, Vol.20, No.2, (Oct, 1998), pp. 116-7

Wang, Y. and F. C. Leung (2004). An evaluation of new criteria for CpG islands in the human genome as gene markers. *Bioinformatics*, Vol.20, No.7, (May 1, 2004), pp. 1170-7

Weinmann, A. S., P. S. Yan, M. J. Oberley, T. H. Huang and P. J. Farnham (2002). Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev*, Vol.16, No.2, (Jan 15, 2002), pp. 235-44

Wu, H., B. Caffo, H. A. Jaffee, R. A. Irizarry and A. P. Feinberg (2010). Redefining CpG islands using hidden Markov models. *Biostatistics*, Vol.11, No.3, (Jul, 2010), pp. 499-514

Wu, S. C. and Y. Zhang (2010). Active DNA demethylation: many roads lead to Rome. *Nat Rev Mol Cell Biol*, Vol.11, No.9, (Sep, 2010), pp. 607-20

Xie, H., M. Wang, F. Bonaldo Mde, V. Rajaram, W. Stellpflug, C. Smith, K. Arndt, S. Goldman, T. Tomita and M. B. Soares (2010). Epigenomic analysis of Alu repeats in human ependymomas. *Proc Natl Acad Sci U S A*, Vol.107, No.15, (Apr 13, 2010), pp. 6952-7

Xin, Y., B. Chanrion, M. M. Liu, H. Galfalvy, R. Costa, B. Ilievski, G. Rosoklija, V. Arango, A. J. Dwork, J. J. Mann, et al. (2010). Genome-wide divergence of DNA methylation marks in cerebral and cerebellar cortices. *PLoS One*, Vol.5, No.6, pp. e11357

Xing, J., D. J. Hedges, K. Han, H. Wang, R. Cordaux and M. A. Batzer (2004). Alu element mutation spectra: molecular clocks and the effect of DNA methylation. *J Mol Biol*, Vol.344, No.3, (Nov 26, 2004), pp. 675-82

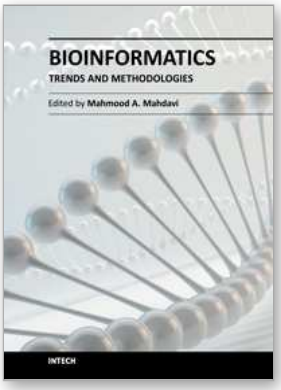
Yates, P. A., R. W. Burman, P. Mummaneni, S. Krussel and M. S. Turker (1999). Tandem B1 elements located in a mouse methylation center provide a target for de novo DNA methylation. *J Biol Chem*, Vol.274, No.51, (Dec 17, 1999), pp. 36357-61

Zemojtel, T., S. M. Kielbasa, P. F. Arndt, H. R. Chung and M. Vingron (2009). Methylation and deamination of CpGs generate p53-binding sites on a genomic scale. *Trends Genet*, Vol.25, No.2, (Feb, 2009), pp. 63-6

Zeschnigk, M., M. Martin, G. Betzl, A. Kalbe, C. Sirsch, K. Buiting, S. Gross, E. Fritzilas, B. Frey, S. Rahmann, et al. (2009). Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. *Hum Mol Genet*, Vol.18, No.8, (Apr 15, 2009), pp. 1439-48

Zhao, Z. and L. Han (2009). CpG islands: algorithms and applications in methylation studies. *Biochem Biophys Res Commun*, Vol.382, No.4, (May 15, 2009), pp. 643-5

Zhu, J., F. He, S. Hu and J. Yu (2008). On the nature of human housekeeping genes. *Trends Genet*, Vol.24, No.10, (Oct, 2008), pp. 481-4



Bioinformatics - Trends and Methodologies
Edited by Dr. Mahmood A. Mahdavi

ISBN 978-953-307-282-1
Hard cover, 722 pages
Publisher InTech
Published online 02, November, 2011
Published in print edition November, 2011

Bioinformatics - Trends and Methodologies is a collection of different views on most recent topics and basic concepts in bioinformatics. This book suits young researchers who seek basic fundamentals of bioinformatic skills such as data mining, data integration, sequence analysis and gene expression analysis as well as scientists who are interested in current research in computational biology and bioinformatics including next generation sequencing, transcriptional analysis and drug design. Because of the rapid development of new technologies in molecular biology, new bioinformatic techniques emerge accordingly to keep the pace of in silico development of life science. This book focuses partly on such new techniques and their applications in biomedical science. These techniques maybe useful in identification of some diseases and cellular disorders and narrow down the number of experiments required for medical diagnostic.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821