8/16/2004

Exploring the characteristics of sequence elements in Proximal Promoters of Human Genes

Minou Bina[a], Phillip Wyss[a], Wenhui Ren[b], Wojciech Szpankowski[b], Elizabeth Thomas[c], Ranjit Randhawa[a], Sreedeepti Reddy[a], Priya M. John[a], Elsie I. Pares-Matos[a], Arnold Stein[d], Hao Xu[a], and Sheryl A. Lazarus[a]

[a]Purdue University, Department of Chemistry, West Lafayette, IN 47907, USA.
[b]Purdue University, Department of Computer Sciences, West Lafayette, IN 47907, USA.
[c]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, One Bungtown Road Cold Spring Harbor, NY 11724
[d]Purdue University, Department of Biological Sciences, West Lafayette, IN 47907, USA.

Keywords: Human genome, gene regulation, transcription factor binding sites, sequence context of human genomic DNA, codes in human DNA

Corresponding author
Minou Bina: Purdue University, Brown Building, Department of Chemistry, West Lafayette, IN 47907, USA. Email Bina@Purdue.edu TEL 765 494 5294 FAX 765 494 0239

Abstract

Central to reconstruction of *cis*-regulatory networks is identification and classification of naturally occurring transcription factor binding sites according to the genes that they control. We have examined salient characteristics of 9-mers that occur in various orders and combinations in the proximal promoters of human genes. In evaluations of a dataset, derived with respect to experimentally defined transcription initiation sites, in some cases we observed a clear correspondence of highly ranked 9-mers with protein binding sites in genomic DNA. Evaluations of the larger dataset, derived with respect to the 5' end of human ESTs, revealed that a subset of the highly ranked 9-mers corresponded to sites for several known transcription factor families (including CREB, ETS, EGR-1, SP1, KLF, MAZ, HIF-1, and STATs) that play important roles in the regulation of vertebrate genes. We identified several highly ranked CpG containing 9-mers, defining sites for interactions with the CREB and ETS family of proteins, and identified potential target genes for these proteins. The results of the studies imply that the CpG-containing transcription factor binding sites regulate the expression of genes with important roles in pathways leading to cell-type-specific gene expression and pathways controlled by the complex networks of signaling systems.

Introduction

The prediction is that the human genome must contain vast amounts of *cis*-regulatory elements that are dispersed in the control regions of genes to direct the spatial and temporal patterns of gene expression in response to external stimuli, metabolic requirements, and developmental programs [1]. It is estimated that the human and mouse genomes include about 22,000 to 30,000 protein-coding genes [2-4]. However, the characteristics of sequence elements in the control regions of so many genes remain largely unknown.

Two regulatory segments at the 5' ends of genes play distinct roles in the control of transcription. One segment contains the basal or core promoters, which support the formation of relatively large multi-protein complexes that direct the RNA polymerase II to begin transcription at correct sites [5]. The other segment corresponds to proximal promoters, which contain the *cis*-elements that interact with sequence-specific DNA binding proteins to control the expression of genes [1]. Evidence suggests that the sequence context of the genomic DNA regions that contain protein binding-sites may determine whether or not these regions could function in regulation of transcription. Since closely spaced transcription factor binding sites can facilitate protein-protein interactions, clustering of protein binding elements is often a hallmark of a subset of the control regions in genomic DNA [6-12].

Numerous studies have examined the context of the DNA segments that control the expression of genes in model systems including yeast [i.e. 13,14]. Large-scale computational studies have highlighted the importance of protein binding elements in proximal promoters to the regulation of vertebrate genes: including immune genes [15] and genes that function during cell cycle [i.e. 16]. Additionally, sequences that are conserved among various species have reinforced the importance of sequence context in regions that serve as control segments in genomic DNA [17].

In this report, we explore the characteristics of sequences in proximal promoters of human genes. Among the sequences that ranked highly in statistical evaluations, the studies identified several 9-mers that produced perfect matches with genomic DNA regions containing functional protein binding sequences. Furthermore, the studies indicated that these elements occurred in proximal promoters of a number of genes with important functions in specific tissues and organs. Taken together, the results predict that the human genome includes relatively long regulatory elements (at least 9 bases) that are shared by many genes. This prediction could be of relevance to the studies of *cis*-regulatory networks, and to the studies of signals in human DNA that control the expression of co-regulated human genes.

Results

Overview of the Strategy

To explore the characteristics of naturally occurring sequences that may contribute to gene regulation, we focused on collecting and analyzing constant length elements (9-bases) from proximal promoters of human genes. Evidence indicates that constant length elements reduces computational burden [18,19]. Generally, short sequence elements (1-5 bases) appear frequently in human DNA. Therefore, elements consisting of 9 base pairs could reduce background noise and enhance the specificity of sequences that may provide recognition sites for regulators of transcription.

To examine the characteristics of 9-base elements, irrespective of their orientation in DNA, we designed the following scheme. First, we computationally generated all possible 9-base elements that may occur in DNA. To make these elements orientation independent, we identified the complementary sequences, thereby producing 131,072 pairs of 9-mers that defined the forward (f) and reverse complement (r) of the sequences (supplementary Table 1).

For data collection, we created a reference table in which the 9-mer-pairs were assigned identifiers (IDs). The IDs kept track of the orientation of the sequences in each 9-mer pair (supplementary Table 1). For example, RF107075f corresponds to TCGCGAGCG and RF107075r defines its reverse complement (CGCTCGCGA), written in 5' to 3' direction by convention. We note that referring to 9-mers as pairs not only allows their identification, irrespective of their orientation in DNA, but also eliminates problems associated with redundancy since, the complementary pairs are considered to represent the same sequence element in the DNA.

Based on the above schemes, we collected two sets of data. To obtain data that were relatively error-free, a small set was derived from proximal promoters with accurately defined transcription initiation sites. A larger set was collected using proximal promoters that were predicted with respect to the 5' end of ESTs [20].

Sequences derived from proximal promoters with experimentally defined transcription start sites

In collecting the first set, we focused on data quality rather than quantity. To ensure accuracy, we conducted extensive manual curation of several hundred sequence-files, and examined published reports to evaluate the experimental data that were used for defining the transcription start sites. Through various screening procedures, we selected 60 files that met the criteria for data quality. For data collection, we analyzed the DNA segments corresponding to proximal promoters for all possible 9-mers. The procedure included recording the number of times each 9-mer-pair matched the elements in the reference table.

The collected 9-mer-pairs in the small set were ranked according to the number of times that they occurred in proximal promoters of the analyzed genes. Data analyses indicated that 21944 of the collected the 9-mers (88.97%) were found once; 2240 elements were identified twice (9%). The dataset included 344 elements (1.4%), identified 3 times, and 78 elements (0.32%) identified 4 times. The abundance of higher hits was considerably less. For example, two elements (a 9-base element and its complement) were identified 19 times. Another pair was identified 14 times.

Sequences derived from proximal promoters with respect to 5' end of cDNAs

To create a larger dataset, we analyzed the proximal promoter sequences in a listing of predicted 5' ends of human genes [20]. In contrast to the data in the small set, a number of the sequences in the large dataset included regions corresponding to known repetitive DNA elements. To address this problem, initially we masked the repetitive sequences to monitor the occurrences of 9-mers in the unique regions. However, through extensive evaluations, we found that it was more informative to use statistics to analyze the 9-mers in the context of their relative abundance in proximal promoters and in human genomic DNA. We also analyzed the 9-mers in the context of their abundance in known human repetitive DNA sequences.

The frequency of each 9-mer in human genomic DNA was derived from the June 2002 freeze of the human genome [21]. Consequently, the data represented a relatively large sample of genomic DNA (2,809,657,179 base pairs). The results revealed that 9-mers consisting of tracts of A or T were the most abundant sequence elements in human genomic DNA. Tracts of A_9 and T_9 were found 7,109,395 times. The second most abundant sequence corresponded to TGTGTGTGT and ACACACACA, which were found 1,390,980 times. These values can be contrasted with those obtained for CpG-containing elements. For example, in the genomic DNA, TTCGCGCGA and its complement were found only 93 times; TCGCGAGCG and its complement were found 358 times. Remarkably, we noted that all possible 9-mers (4^9) were found not only in total genomic DNA but also in the repetitive DNA sequences. The results of this analysis indicated that masking procedures might not necessarily eliminate potential artifacts arising from repetitive DNA.

Statistical evaluations

To identify the 9-mers that could be of importance to proximal promoters of human genes, we devised a scheme for correlative analyses. The procedure identified and statistically ranked the 9-mers according to their relative abundance in proximal promoters and in total genomic DNA. For correlative analysis, we defined a parameter (Ei) to represent the number of times a given 9-mer was found in proximal promoters of human genes. Another parameter (Gi) defined the number of times the same 9-mer was found in human genomic DNA. Fig. 1 displays a plot of the Gi vs Ei values.

Next, we developed a statistical model for partitioning the 9-mers according to their significance to proximal promoters of human genes. From the work of Regnier and Szpankowski [22], we know that the occurrences of words in a randomly generated text (based on either Bernoulli or Markov model) should be normally distributed around a mean. Based on this knowledge, statistical derivations and analyses were done using the principle of large deviations, often referred to as p-value analyses [23]. As reference, we chose a probabilistic model that assumed that the total genomic DNA could be generated by a memoryless or Markov source. Then for given parameters (beta values, i.e. beta $\approx 10^{-10}$ to 10^{-50}), representing confidence levels, we computed thresholds, through appropriate approximations (detailed in the method section). To visualize the results, the thresholds were computed in the context of slopes of linear lines (Fig. 2). We expected that the 9-mers that occurred equally in proximal promoters and genomic DNA would be distributed in the vicinity of a line with a slope of 1. Lines with slopes higher than 1 would demarcate the elements that occur more frequently than the probabilistic model predicts, and hence would represent statistical significant elements (detailed in the method section).

Thus, in correlative plots, the region between the y-axis and a line with slope 7.0 ($\beta \approx 10^{-50}$) identified a data subset with a relatively high statistical significance (Fig. 2). This subset included 11% of the 9-mers in the large dataset. The region between the y-axis and a line with slope 4.5 (beta $\approx 10^{-20}$ to 10^{-23}) produced a subset that included 17.48% of the elements in the large dataset. Sequences that occurred frequently in the human genomic DNA corresponded to a subset defined with respect to a line with slope 0.4 and the x-axis (Fig. 2).

From the results of the statistical studies, the 9-mers were ranked. The 9-mers that occurred frequently in proximal promoters, as compared to total genomic DNA, received a high ranking (supplementary Table 2). From the ranked values, the 9-mers corresponding to a desired threshold can be obtained. For example, the 9-mers distributed between the y-axis and the line with slope 7.0 ($\beta \approx 10^{-50}$) would correspond to those with ranking of 7 or higher (Fig. 2; supplementary Table 2); the 9-mers distributed between the y-axis and the line with slope 3.0 would correspond to sequences with ranking of 3 or higher; etc.

Data analysis reveals that the highly ranked 9-mers are GC-rich, including CG and tracts of G and C (supplementary Table 2). For example the highest ranked element (TCGCGAGCG) includes 3 CpG, and the next one (TTCGCGCGA) is also rich in CpG (supplementary Table 2). The same trend followed for many of the highly ranked 9-mers, indicating that they probably reflect the characteristics of sequences in CpG islands [24]. The analysis also revealed numerous G-tract/C-tract containing sequences with high rankings. Again, the results were indicative of characteristics of the sequences in CpG islands [24]. However, not all CpG containing sequences received a high-ranking. Consider for example, a CpG containing 9-mer (CGATAGTTT) that was identified only once in the analyzed proximal promoters. While this sequence appeared infrequently in the genomic DNA (15,720 times), it occurred often in human repetitive DNA sequences. Similarly, another 9-mer (ACGCGATCT), containing 2 CpGs, occurred rarely in genomic DNA (2358 times), appeared only twice in the analyzed proximal promoters, but was found frequently in human repetitive DNA. These and similar examples imply that the context of CpG containing sequences is important to their preferential localization in proximal promoters of human genes.

Correlation with Protein-Binding Elements

Next, we asked whether a correlation existed between known protein binding sites and the 9-mers collected from proximal promoters of human genes. We began the correlative analyses using the small dataset. Subsequently, the results were compared to those obtained using the sequences in the large set.

Evaluation of data derived with respect to experimentally defined transcription start sites

Studies of the smaller dataset included construction of maps to visualize directly the correspondence of the significant 9-mers with respect to experimentally defined protein binding sites in genomic DNA. For the studies we chose a previously described list of transcription factor binding sequences [6]. Since vertebrate transcription factors recognize similar and highly related sequences, the correlative studies include human, mouse, rat, and hamster genomic DNA sequences that have previously been characterized in functional and DNA binding and assays. Several examples highlight the correspondence of 9-mers with protein binding sites.

The first example shows that in some cases, in genomic DNA we could observe a clear correspondence of mapped protein binding sites with occurrences of highly ranked 9-mers. For example, Fig. 3A shows consecutive hits that define a functional segment in the human NF-IL6 gene. In DNA binding assays, when incubated with nuclear extracts from unstimulated human monocyte/macrophages, the segment interacted with SP1 and a short isoform of SP3 [25]. With nuclear extracts from stimulated cells, the segment interacted with SP1, SP2, and a long isoform of SP3 [25,26].

The second example shows that in some cases, the occurrences of the highly ranked 9-mers in genomic DNA corresponded to closely spaced protein-binding elements (Fig. 3B). The example displays the proximal promoter of the rat NMDA receptor subunit type 1 gene and shows the position of two 9-mers that match two adjacent functional sites that interact with SP1 and MAZ [27].

The third example shows hits with sequences in genomic DNA that correspond to in vivo footprints (Fig. 4). Specifically, in a study, dimethyl sulphate footprinting of the mouse phosphoribosyltransferase (aprt) gene identified three "GC-boxes" that were occupied in vivo, most probably by SP1 [28]. We find that two of the vivo footprints correspond to the occurrences of highly ranked 9-mers (Fig. 4).   Two consecutive elements define the first footprint. Another two coincide with the second footprint (Fig. 4).

The fourth example shows hits that correspond to in vitro footprints of protein binding elements (Fig. 5). Specifically, upstream of the TATA signal in the human Fibronectin (FN) gene, a previous study identified 7 cell-type specific DNase I footprints that encompassed the sequences that regulated the expression of this gene [29]. Fig. 5 shows a high correspondence of highly ranked 9-mers with footprint 6 (FP6). Several consecutive hits in this footprint create two overlapping high affinity SP1 sites (Fig. 5). Two additional hits correspond to a sequence defining a site for SP1 and possibly for other factors from the SP1 and KLF family.

FP4 defines a binding site for members of the CREB family (Fig. 5). This site regulates the transcription of FN gene in response to forsklin and follicule-stimulating hormone [29]. Clusters of several 9-base elements match with sequences that correspond to FP7 and FP8 (Fig. 5). In FP7, three overlapping hits create a binding site for the transcription factor AP2. Three consecutive hits are within FP8 (Fig. 5). Two consecutive elements are within FP5 (Fig. 5).

The fifth example shows that multiple hits may create module-like organizations that include several overlapping protein-binding sequences in genomic DNA. The example displays a DNA segment from the proximal promoter of the human cyclin D1 gene encompassing sites for several transcription factors including EGR-1, MAZ, and SP1 (Fig. 6). In functional studies, EGR-1 mediated transcriptional induction of the gene by Angiotensin II [30]. DNA binding assays revealed that the responsive element interacted with EGR-1. The region that includes the responsive element overlaps a site that interacts with SP1 [30].

Evaluation of the dataset derived with respect to 5' end of ESTs

In studies of the larger dataset, the sequences of 9-mers with rankings of 2 higher were compared to a listing of consensus sequences defining potential transcription factor binding sites (supplementary table 3). The sites were grouped according to the structure of the DNA binding domain of the proteins and/or protein families: i.e. E2F family, SP1 family, CREB family, ETS family, EGR-1 family, etc [6].

As the small dataset, in the analyses of the large dataset, the canonical high affinity SP1 site (GGGGCGGGG) received a high ranking (43.28). Similarly, the canonical high affinity site (GCGCGGGCG) for the members of the EGR-1 family also ranked highly (51.86). A potential site (GCCGCCGGC) for binding the members of AP2 family was also among the highly ranked (35.39) sequences. The analyses further identified highly ranked sequences that contained sites for interactions with members of CREB (ranking 40.68) and ETS (ranking 32.77) family of transcription factors. Furthermore, among the highly ranked elements we noted a sequence

(TTACGTGCG) containing a site for members of HIF-1 family (ranking 23.83) and a sequence (TTCCGCGAA) containing a potential site for STATs (ranking 17.75). Several of the highly ranking sequences included an E-box (CACGTG) known to interact with basic-helix-loop-helix family of proteins. We obtained a lower ranking (4.84) for a sequence (GGGGTGGGG) that has been shown to interact with members of SP1 and KLF family. Similarly a binding site for EKLF (AGGGTGGGG) did not rank highly (3.08). Nonetheless, these latter rankings were still among statistically significant sequence elements: $\beta \approx 10^{-27}$ for the SP/KLF site, $\beta \approx 10^{-7}$ for the EKLF site.

Data evaluations also revealed that a number of highly ranked sequences, containing transcription factor binding elements, appeared infrequently in genomic DNA. For example, in the total human genomic DNA sequences that we analyzed, a high ranking 9-mer containing an E-box for basic helix-loop-helix family was found only 1402 times; an element containing a site for CREB family was found 4231 times; a sequence containing a site for ETS family was found 4824 times. A notable exception to the trend is the 9-mer with the high affinity SP1 site (GGGGCGGGG), which appears nearly 48,000 times in the version of the human genomic DNA that we have analyzed. The EKLF binding sequence (AGGGTGGGG), which received a significantly lower ranking, occurred 66,229 times in genomic DNA. A sequence (GGGGTGGGG) for interactions with SP1 and members of the KLF family occurred 141,209 times. The data include many examples of CACCC and CAAT 'boxes' that occurr infrequently in genomic DNA. A subset of the sequences corresponding to sites for ZF87/MAZ/MBP occurred infrequently, while other subsets were abundant in the genomic DNA.

Identification of target genes for naturally occurring protein-binding elements

Next, we asked whether studies of the highly ranked elements, that included known transcription factor binding sites, could give clues into the characteristics of genes regulated by these sites. To study this question, for evaluation we selected examples of highly ranked 9-mers that contained experimentally defined protein-binding sequences. The selected 9-mers included validated recognition sites for members of CREB or ETS families, to highlight the importance of these two families of transcription factors to the regulation of human genes.

In the first example, we examine a highly ranked 9-mer (TGACGTCAC) that includes the palindromic sequence for interactions with members of CREB family. Experimental studies have shown that the palindromic element is highly conserved and regulates gene expression in response to cyclic AMP [31]. The palindromic element, also known as CRE, is subject to hormonal regulation [31]. Among the analyzed files we identified 96 genes as potential downstream targets of the CREB family, through interactions with the highly ranked (40.68) palindromic CRE (supplementary Table 4). These targets include genes that function in the control of cell cycle, and genes for transcription factors that may serve as components of the gene regulatory networks regulated by the CREB family. Notable examples include STAT3, homeo box C10, NFX1, and TAF10 (supplementary Table 4).

Another highly ranked 9-mer (GTGACGTCG) also includes a potential site for interactions with members of CREB family. In a literature survey, we noted that the sequence of the 9-mer matched a functional element in the promoter region of the PCNA gene [32]. Experimental evidence has shown that the element in the PCNA gene is regulated by the members of CREB family, in the control of cell cycle [32]. For regulation through that 9-mer, we identified 23 genes as potential targets of CREB family (supplementary Table 5). The listing includes a gene for a cell cycle related kinase. One of the potential targets is a transcription factor, LIM homeobox 4 (LHX4), with functional attributes that reflect the key role played by CREB family in the genetic networks that control the development of pituitary and neuronal cells (OMIM 602146).

The second example examines a highly ranked 9-mer (CCGGAAGTG) that includes a site for interactions with ETS family members. Examples include an identical element in the promoter region of rpL32, a ribosomal protein [33]. Our analyses identified 119 genes as potential targets for regulation through the 9-mer containing the highly ranked (32.77) ETS site (supplementary Table 6). These targets include ribosomal protein S19, ribosomal protein L27a, and translation initiation factor 3. Other targets highlight the important roles of ETS family in oncogenesis (i.e. the tumor susceptibility gene 101 and the melanoma antigen family F1). Other potential targets are also consistent with the regulatory functions of the ETS family members [34,35]. These targets include several genes encoding membrane proteins and extracellular receptors (i.e. cleft lip and palate associated transmembrane protein, vesicle-associated membrane protein 3), and genes encoding proteins that function in cell motility and organization (i.e. kinesin).

In studies of potential targets for regulation by the ETS family, we noted that the sequence of another highly ranked 9-mer (GCGGAAGTG) matched an element in TRANSFAC that was listed as Fli-1 binding site. This protein also is a member of the ETS family. We identified 68 potential targets for regulation by that 9-mer (supplementary Table 7). The described properties of the target genes were again consistent with the roles played by the ETS family of transcription factors (supplementary Table 7). For example, the list includes genes for proteins implicated in tumor formation and genes associated with membrane functions.

Discussion

To examine the characteristics of the sequences in proximal promoters of human genes, we analyzed two sets of data consisting of 9-mers, collected irrespective of their orientation in DNA. In statistical evaluations of the large dataset, derived from the DNA segments localized with respect to the 5' end of ESTs [20], we found that the proximal promoter sequences of human genes included numerous elements that occurred infrequently in human genomic DNA. In order to distinguish the significant 9-mers, we set the probability thresholds (β values) very low leading to large deviations analyses [23]. Fig. 2 shows a plot that summarizes the results.

In data evaluations, we detected direct correspondences of previously mapped protein binding sites with 9-mers in the smaller dataset, derived from proximal promoters localized with respect to accurately defined transcription start sites (Fig. 4-6). The results revealed examples of 9-mers that overlapped to define high affinity sites for protein-DNA interactions (Fig. 3A), 9-mers that correlated with experimentally localized in vivo and in vitro footprints in genomic DNA (Fig. 4,5), and 9-mers that covered sequences containing tight clusters of protein binding elements (Fig. 6).

As noted in other studies [18,19], it appeared that a relatively small fraction of the data in both sets correlated with potential transcription factor binding elements. In our studies, we could deduce correspondences with known high affinity sites for protein binding. Examples include recognition sites for ETS, CREB, EGR-1, AP2, and SP1 families of transcription factors and E-boxes for basic helix-loop-helix proteins. In many cases, accurate correspondences appeared more difficult to assess. For example, a large fraction of the highly ranked sequences included GGGCGGG. We were confronted with the question of whether the sequences should be considered as SP1 sites, or whether they primarily reflected the characteristics of CpG islands. Similarly, we noted that a number of highly ranked sequences included CACCC. However, it was unclear whether it would be accurate to consider these sequences as sites for CACCC binding proteins.

Data evaluations also revealed that the highly ranked elements in both datasets contained one or more CpG, and in many cases tracts of G or C. The results suggested that the collected elements primarily reflected the characteristics of CpG islands. This finding agrees with the reported association of CpG islands with a significant fraction of human genes [2,3]. However, we noted that not all CpG-containing elements ranked highly in the collected sequences. The results indicate that CpG islands might entail lexical characteristics that do not include all CpG containing sequences.

It has been suggested that protein-binding elements in CpG islands are primarily associated with widely expressed genes [36]. However, our data have revealed that the human genome include CpG containing sequences that are controlled by transcription factor families with complex and elaborate regulatory characteristics. These include the control of cell cycle, cell-growth, cell-type-specific gene expressions, and responses ensued by cellular signaling systems. Therefore, gene regulation through CpG-containing sequences could be more complex that has previously been assumed.

Gene Regulation through ETS sites containing CpG

In mechanisms of gene regulation, there are examples of CpG methylations that cause gene silencing by inhibiting or reducing the binding of transcription factors to the DNA [36]. Examples include the requirement of an unmethylated ETS binding site for activation of an early B-cell-specific gene by Pax-5 [37]; and, a single methylation event abolishing the interactions of an ETS protein with the regulatory region of a gene [38]. Significantly, the ETS proteins are associated with controlling the expression of genes that perform critical roles in mechanisms of cell differentiation, cell migration, cell proliferation, cell-to-matrix and cell-to-cell adhesion, and apoptosis [34,35].

To further examine the importance of CpG-containing elements to the regulation of human genes, we identified potential targets for several highly ranked 9-mers that corresponded to experimentally validated sites for interactions with the ETS family. The functional characteristics of the predicted targets imply that CpG containing ETS binding elements may regulate the expression of genes that play important roles in specific cellular tissues and organs (supplementary Table 6 and 7). Consistent with the regulatory roles attributed to ETS proteins, the predicted targets include genes encoding membrane proteins and extracellular receptors, genes encoding proteins that function in cell motility and organization, and genes for proteins implicated in tumor formation. Remarkably, the list also includes genes for proteins that control cell cycle (supplementary Table 6 and 7) suggesting a role for ETS family in this important cellular process.

Gene Regulation through CREB sites containing CpG

Transcription factors of the CREB family correspond to effector molecules that bring about cellular changes in response to discrete sets of instructions [39]. The CREB proteins serve as downstream targets of key signaling molecules including hormones, neurotransmitters, and cytokines. Thus, functional characteristics of CREB family cover a broad spectrum: from the regulation of nervous system, memory, cytokine signaling, and pituitary glands, to the control of spermatogenesis and circadian rhythms [39-41]. Furthermore, a study has shown that methylation of a CREB site can repress the expression of a gene [42].

In this context, we noted that the sequences collected from the proximal promoters of human genes include several highly ranked CpG containing elements, with sites for interactions with CREB proteins (supplementary tables 4,5). A notable example is a palindromic sequence that corresponds to one of the best-characterized CREs, cAMP-responsive elements [31]. The predicted targets of this CRE include genes that function in the control of cell cycle, and genes for transcription factors that may serve as components of the *cis*-regulatory networks regulated by the CREB family. Notable examples include STAT3 and homeo box C10 (supplementary Table 4).

Gene Regulation through AP2 sites

Generally, the sites for AP2 family members appear frequently in CpG islands [43]. Our analysis has identified several highly ranked elements that may be subject to control by AP2 family of transcription factors. Occurrences of AP2 sites in CpG islands further support the notion that these islands are associated with controlling genes that play important roles in cell differentiation and development.

Specifically, evidence has shown that genes associated with AP2 sites may temporally be regulated during embryogenesis, and in pathways leading to the formation of neural crest derivatives and neural, epidermal, and urogenital tissues [44]. Mutations in the AP2β gene cause the human inherited disorder Char syndrome [45]. Studies of mice have shown that the AP2 family members regulate important developmental processes including cranial neural tube closure, and lens development [44]. Furthermore, members of the AP2 family regulate gene expression in responses controlled by signaling systems [46].

Gene Regulation through EGR sites

Our studies have further identified highly ranked elements that include sequences for recognition by members of the EGR-1 family. Family members interact with highly related sequences that include CpG [43]. The results reinforce the importance of naturally occurring sequence elements that regulate the expression of genes with complex characteristics. Specifically, as early growth response genes, members of the EGR-1 family play key roles in pathways controlled by the complex networks of signaling systems. Proteins of the EGR-1 family orchestrate the changes in gene expression that underlie neuronal plasticity [47]. A direct role of EGR-1 in controlling proliferation has been proposed for T-cells, astrocytes, glioma cells, glomerular mesangial cells, and keratinocytes [48].

We conclude that the proximal promoters of human genes include relatively long sequence elements (at least 9 base pairs) that may entail complex regulatory properties. Elements that contain CpG are subject to regulation by methylation [i.e. 36,42]. And, significantly, a subset of CpG containing elements includes sites for transcription factors that control important cellular processes. These cover a broad range: from cell-cycle control and cell-type-specific gene expression, to developmental and responses controlled by signaling pathways. The apparent conservation of these elements in the human genome points to the importance of sequence context of the genomic DNA to the regulation of gene expression. Furthermore, the sequence of the human genome has uncovered a relatively large number of transcription factors with unknown sequence specificities [2,3]. Therefore, it might seem reasonable to expect that the highly ranked sequences that we have discovered could provide good candidates for systematic experimental evaluations.

Materials and Methods

Data sets

To create a dataset of proximal promoters with respect to accurately assigned transcription start sites, from GenBank we retrieved nearly 400 human DNA sequences that contained known promoters. The publications reporting the promoter sequences were examined to evaluate whether the transcription initiation sites were correctly assigned. Files containing ambiguous transcription initiation site were eliminated. The remaining sequences were screened for repetitive DNA. Screening was done using the CENSOR program [49], offered through the web (http://www.girinst.org/Censor_Server.html). From the results, we selected sequence files that did not contain simple and complex repeats between positions -350 to +10 of a gene.

A second set of data consisted of predicted promoter sequences that were previously described [20]. These sequences were identified by aligning full-length cDNA clones from the Mammalian Gene Collection (MGC) to a draft of the human genome [20]. From the set we removed the promoter sequences that were either incomplete or contained ambiguity codes. From the set, we also removed sequences that appeared to be redundant. Another dataset consisted 9-mers corresponding to of total human genomic DNA, derived from the June 2002 freeze of the genome browser at UCSC [21]. Another set consisted of 9-mers derived from "repeat-only" version of the August 2001 freeze, from UCSC. This set included the sequences that were annotated as human repetitive DNA.

Data Collection

The selected GenBank files that contained experimentally defined transcription start site were analyzed for all 9-mers occurring in unique regions in the human DNA, between positions -300 to - 46 with respect to the transcription initiation site: +1, by convention. The 9-base elements were collected in tables created in a database using MySQL. The database was built in a PC, using the LINUX operating system. Data collection, access, retrieval, management, and analyses were done by developing Perl scripts and using Perl modules offered by the developers of Bioperl (http://www.bioperl.org). The scripts kept track of the orientation of each 9-base element with respect to the transcription start site of each gene. For genes that included multiple start sites, we chose the prominent experimentally defined site. The scripts also kept track of the number of times each 9-base element was identified and collected.

A similar procedure was used for collecting the 9-base elements from the dataset derived from the sequences reported by Trinklein et al. [20]. The data that we collected corresponded to the sequences between positions -550 to -46 with respect to predicted the transcription start sites.

Correlative Analysis

For data evaluation, genomic DNA sequences from GenBank were examined for occurrences of transcription factor binding sites and occurrences of 9-base elements that were collected from the proximal promoters. Data for experimentally defined factor binding sites have previously been described [6]. For correlative mapping analyses of the small dataset, we created a file that was formatted for compatibility with the map program of the GCG sequence analysis package from Accelrys (http://www.accelrys.com/about/gcg.html), as previously described [43].

Statistical evaluations

In the analysis, G denotes total genomic DNA, E denotes the proximal promoter segments, and F denotes the regions that do not correspond to proximal promoters. The reference tables of 9-mers ($w_1, \ldots, w_M$), consisting of the large dataset and human genomic DNA, were used to identify the elements that matched the sequences in the E and G regions. For $1 \leq i \leq M$ we denote by E_i, F_i and G_i, respectively, the numbers of times the *ith* 9-mer w_i occurs in regions E, F and G. L_E corresponds to the total length of the proximal promoters used for data collection, and L_G to the total length of the analyzed genomic. L_F corresponds to L_G - L_E. Since L_E is significantly shorter than L_G, as an approximation we assume $|L_E| \langle\langle\ |L_F| \approx |L_G|$.

Actually, the quantities of interest are:

$$e_i = \frac{E_i}{|L_E|},$$
$$f_i = \frac{F_i}{|L_F|},$$
$$g_i = \frac{G_i}{|L_G|}.$$

We shall make two additional approximations that are well justified in our experiments:

$$f_i \approx g_i,$$
$$f_i \approx p(w_i)$$

where the last approximation is a reflection of the fact that with $|L_G|$ very large, the ratio f_i is well approximated within the margin of error by the probability $p(w_i)$ of 9-mer w_i occurrence.

We aim at finding a threshold slope α_{th} such that we can assign statistical significance to the fact that $e_i > \alpha f_i$ (or $e_i > \alpha g_i$). Therefore, we need to compare empirical results to a *reference model* in order to decide whether a given experiment carries any useful information. Our reference model is a probabilistic model in which we assume that the whole genome G is generated by a memoryless or Markov source. The quantities e_i and f_i become random variables (but as described above, we simplify the analysis by assuming that $f_i = p(w_i)$ is a constant). Then, we decide that for a given *beta* (e.g. $\beta \approx 10^{-10} – 10^{-50}$) the event $e_i > \alpha f_i$ is *statistically significant* if the probability of $e_i > \alpha f_i$ is smaller than β, that is,

$$P(e_i > \alpha f_i) < \beta.$$

This means that the chance that randomness will generate the event $e_i > \alpha f_i$ is very small.

Since in our experiments, the mean value of e_i equal to $p(w_i) \approx f_i$, we expect a_i to be likely bigger than f_i. Therefore we must set β very small (e.g. $\beta \approx 10^{-10} – 10^{-50}$) and use large deviations [23] to compute the α_{th} threshold.

Now we compute the slope α. As discussed above, we assume that the genome is generated randomly and that E_i is the random variable representing the number of w_i occurrence in the proximal promoter regions. From [22] it is known that E_i values would be normally distributed around the mean $\mathbf{E}[E_i] = L_E p(w_i)$. More precisely, when E_i deviates no more than $O(\sqrt{L_E p(w_i)})$, it is proved that

$$E_i \sim \mathrm{N}\ (L_E P(w_i), L_E \sigma^2(w_i))$$

where we denote by $N(\mu, \sigma^2)$ the normal distribution with mean μ and variance σ^2. However, when E_i deviates $O(\sqrt{L_E p(w_i)})$ another probabilistic law governs E_i behavior, namely the large deviations law. Previously Regnier and Szpankowski [22] proved that

$$P(E_i > (1+\delta) L_E p(w_i)) \sim \frac{1}{\sqrt{2\pi L_E}} \ \exp(-L_E I(\delta)) \qquad (1)$$

where $I(\delta)$ is a complicated function of δ that depends on some moment generating functions (cf. also see reference [50]).

To compute the threshold slope $\alpha = 1 + \delta > 1$, we need to estimate δ from the following equation

$$P(e_i > (1+\delta) p(w_i)) < \beta$$

which translates into $P(E_i > (1+\delta) L_E p(w_i)) < \beta$. Clearly, this is within the large deviations domain. We need to apply (1), but numerical computations of the large deviation function $I(\delta)$ are rather cumbersome. Therefore, we shall propose two good approximations. Observe first that we only need a good bound on the large deviation probability.

First, observe that E_i is (ignoring overlapping of 9-mers) a sum of Bernoulli independent random variables. If this would have been the case, then the following bound can be found (cf. for example, see reference [50])

$$P\left(E_i > (1+\delta)L_E p(w_i)\right) \le \exp\left(-L_E p(w_i)\delta^2 / 3\right) \tag{2}$$

But if we want to be rigorous and take into account overlapping of 9-mers, we must relax a little bit the above. Appealing to Azuma's inequality (cf., in refence [23]) we conclude that

$$P\left(E_i > (1+\delta)L_E p(w_i)\right) \le \exp\left(-L_E p(w_i)\delta^2 / 2\right) \tag{3}$$

Thus, by (3) we obtain the following estimate for the threshold $\alpha_{th} = 1 + \delta$

$$a_{th} = 1 + \sqrt{\frac{2\ln \beta^{-1}}{L_E p(w_i).}} \tag{4}$$

In practice, the threshold computed from (2) and (3) is a good compromise leading to

$$1 + \sqrt{\frac{2\ln \beta^{-1}}{L_E p(w_i)}} \le \alpha_{th} \le 1 + \sqrt{\frac{3\ln \beta^{-1}}{L_E p(w_i)}}\,.$$

For example, for $\beta=10^{-40}$ (a good choice for our case) we find that $5.7 \le \alpha_{th} \le 6.7$. In large deviation we can take very small β leading even to higher slope.

Acknowledgments

The research of Wenhui Ren and Wojciech Szpankowski was in part supported by NSF Grant CCR-0208709 and the NIH Grant R01 GM068959-01. We thank you Dr. R. M. Myers for allowing us to use the putative promoter sequences collected in his laboratory.

References

[1] B. Lemon, R. Tjian, Orchestrated response: a symphony of transcription factors for gene control, Genes Dev. 14 (2000) 2551-2569.

[2] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome, Nature (London) 409 (2001) 860-921.

[3] J.C. Venter, et al., The sequence of the human genome, Science 291 (2001) 1304-1351.

[4] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520-562.

[5] S.T. Smale, J.T. Kadonaga, The RNA polymerase II core promoter, Annu.Rev. Biochem. 72 (1999) 449-479.

[6] E.M. Crowley, K. Roeder, M. Bina, A statistical model for locating regulatory regions in genomic DNA, J. Mol. Biol. 268 (1997) 8-14.

[7] C.H. Yuh, H. Bolouri, E.H. Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene, Science 279 (1998) 1896-1902.

[8] E.M. Crowley, A Bayesian method for finding regulatory segments in DNA, Biopolymers 58 (2001) 165-174.

[9] B.P. Berman, Y. Nibu, B.D. Pfeiffer, P. Tomancak, S.E. Celniker, M. Levine, G.M. Rubin, M.B. Eisen, Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome, Proc. Natl. Acad. Sci. U.S.A. 99 (2002) 757-762.

[10] M.C. Frith, J.L. Spouge, U. Hansen, Z. Weng, Z, Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences, Nucleic Acids Res. 30 (2002) 3214–3224.

[11] A.G. Jegga, S.P. Sherwood, J.W. Carman, A.T. Pinski, J.L. Phillips, J.P. Pestian, B.J. Aronow, Detection and visualization of compositionally similar Cis-regulatory element clusters in orthologous and coordinately controlled genes, Genome Res. 12 (2002) 1408-1417.

[12] W.W. Wasserman, J.W. Fickett, Identification of regulatory regions which confer muscle-specific gene expression, J. Mol. Biol. 278 (1998) 67-181.

[13] H.J. Bussemaker, H. Li, E.D. Siggia, Regulatory element detection using a probabilistic segmentation model, Proc. Int. Conf. Intell. Syst. Mol. Biol. 8 (2000) 67-74.

[14] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, E.S. Lander, Sequencing and comparison of yeast species to identify genes and regulatory elements, Nature 423 (2003) 241-254.

[15] R. Liu, R.C. McEachin, D.J. States, Computationally identifying novel NF-kappa B-regulated immune genes in the human genome, Genome Res.13 (2003) 654-661.

[16] A.E. Kel, O.V. Kel-Margoulis, P.J. Farnham, S.M. Bartley, E. Wingender, M.Q. Zhang, Computer-assisted identification of cell cycle-related genes: new targets for E2F transcription factors, J. Mol. Biol.309 (2001) 99-120.

[17] R. Hardison, W. Miller, Use of long sequence alignments to study the evolution and regulation of mammalian globin gene clusters, Mol. Biol. Evol. 10 (1993) 73-102.

[18] G.B. Hutchinson, The prediction of vertebrate promoter regions using differential hexamer frequency analysis, Comput. Appl. Biosci. 12 (1996) 391-398.

[19] L. Marino-Ramirez, J.L. Spouge, G.C. Kanga, D. Landsman, Statistical analysis of over-represented words in human promoter sequences, Nucleic Acids Res. 32 (2004) 949-958.

[20] N.D. Trinklein, S.J, Aldred, A.J. Saldanha, R.M. Myers, Identification and functional analysis of human transcriptional promoters, Genome Res.13 (2003), 308-12.

[21] W.J. Kent, C.W. Sugnet, T.S. Furey, K.M. Roskin, T.H. Pringle, A.M. Zahler, D. Haussler, The human genome browser at UCSC, Genome Res. 12 (2002) 996-1006.

[22] M. Regnier, W. Szpankowski, On pattern frequency occurrences in a Markovian sequence algorithmica, 22 (1998) 631-649.

[23] W. Szpankowski, Average Case Analysis of Algorithms on Sequences, Wiley, NewYork, 2001.

[24] M. Gardiner-Garden, M. Frommer, CpG islands in vertebrate genomes, J. Mol. Biol, 196 (1987) 261-282.

[25] Y. Yang, E.I. Pares-Matos, V.M. Tesmer, C. Dai, S. Ashworth, J. Huai, M. Bina, Organization of the promoter region of the human NF-IL6 gene, Biochim. Biophys. Acta 1577 (2002) 102-108.

[26] E.I. Pares-Matos, Regulation of Gene Expression by DNA-protein and Protein-protein Interactions, Ph.D. Thesis, Purdue University, 2000.

[27] S. Okamoto, K. Sherman, G. Bai, S.A. Lipton, Effect of the ubiquitous transcription factors, SP1 and MAZ, on NMDA receptor subunit type 1 (NR1) expression during neuronal differentiation, Brain Res. Mol. Brain Res. 107 (2002) 89-96.

[28] D. Macleod, J. Charlton, J. Mullins, A.P. Bird, Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island, Genes Dev. 8 (1994) 2282-2292.

[29] V.A. Bernath, A.F. Muro, A.D. Vitullo, M.A. Bley, J.L. Baranao, A.R. Kornblihtt, Cyclic AMP inhibits fibronectin gene expression in a newly developed granulosa cell line by a mechanism that suppresses cAMP-responsive element-dependent transcriptional activation, Biol. Chem. 265 (1990) 18219-18226.

[30] L. Guillemot, A. Levy, M. Raymondjean, B. Rothhut, Angiotensin II-induced transcriptional activation of the cyclin D1 gene is mediated by Egr-1 in CHO-AT(1A) cells, J. Biol. Chem. 276 (2001) 39394-39403.

[31] M.R. Montminy, G.A. Gonzalez, K.K. Yamamoto, Regulation of cAMP-inducible genes by CREB, Trends Neurosci. 13 (1990) 184-188.

[32] B.H. Lee, M.B. Mathews, Transcriptional coactivator cAMP response element binding protein mediates induction of the human proliferating cell nuclear antigen promoter by the adenovirus E1A oncoprotein, Proc. Natl. Acad. Sci. U.S.A. 94 (1997) 4481-4486.

[33] T. Yoganathan, N.K. Bhat, B.H. Sells. A positive regulator of the ribosomal protein gene, beta factor, belongs to the ETS oncoprotein family. Biochem J. 287 (1992) 349-353.

[34] T. Hsu, M. Trojanowska, D.K. Watson, Ets proteins in biological control and cancer, J. Cell Biochem. 91 (2004) 896-903.

[35] E. Lelièvre, F. Lionneton, F. Soncin, B. Vandenbunder, The Ets family contains transcriptional activators and repressors involved in angiogenesis, Int. J. Biochem. Cell Biol. 33 (2001) 391-407.

[36] P.A. Jones, P.W. Laird, Cancer epigenetics comes of age, Nat. Genet. 21 (1999) 163-167.

[37] H. Maier, J. Colbert, D. Fitzsimmons, D. R. Clark, J. Hagman, Activation of the early B-cell-specific mb-1 (Ig-alpha) gene by Pax-5 is dependent on an unmethylated Ets binding site, Mol. Cell Biol. 23 (2003) 1946-1960.

[38] K. Gaston, M. Fried, CpG methylation and the binding of YY1 and ETS proteins to the Surf-1/Surf-2 bidirectional promoter, Gene. 157 (1995) 257-259.

[39] B.E. Lonze, D.D. Ginty, Function and regulation of CREB family transcription factors in the nervous system, Neuron. 35 (2002) 605-623.

[40] U. Otten, P. Marz, K. Heese, C. Hock, D. Kunz, S. Rose-John, Cytokines and neurotrophins interact in normal and diseased states, Ann N.Y. Acad. Sci. 917 (2000) 322-330.

[41] D. De Cesare, G.M. Fimia, P. Sassone-Corsi, CREM, a master-switch of the transcriptional cascade in male germ cells, J. Endocrinol. Invest. 23 (2000) 592-596.

[42] R.C. Iannello, J. Young, S. Sumarsono, M.J. Tymms, H.H. Dahl, J. Gould, M. Hedger, I. Kola, Regulation of Pdha-2 expression is mediated by proximal promoter sequences and CpG methylation, Mol. Cell Biol. 17 (1997) 612-619.

[43] Bina, M. and Crowley, E. Sequence patterns defining the 5′ boundary of human genes, Biopolymers 59 (2001) 347-355.

[44] K. Hilger-Eversheim, M. Moser, H. Schorle, R. Buettner, Regulatory roles of AP-2 transcription factors in vertebrate development, apoptosis and cell-cycle control, Gene 260 (2000) 1-12.

[45] Zhao, F., Weismann, C.G., Satoda, M., Pierpont, M.E., Sweeney, E., Thompson, E.M. and Gelb, B.D. Novel TFAP2B mutations that cause Char syndrome provide a genotype-phenotype correlation, Am. J. Hum. Genet. 69 (2001) 695-703.

[46] M. Imagawa, R. Chiu, M. Karin, Transcription factor AP-2 mediates induction by two different signal-transduction pathways: protein kinase C and cAMP, Cell 51 (1987) 251-260.

[47] K.J. O'Donovan, W.G. Tourtellotte, J. Millbrandt, J.M. Baraban, The EGR family of transcription-regulatory factors: progress at the interface of molecular and systems neuroscience, Trends Neurosci. 22 (1999) 167-173.

[48] G. Thiel, G. Cibelli, Regulation of life and death by the zinc finger transcription factor Egr-1, J. Cell Physiol. 193 (2002) 287-292.

[49] J. Jurka, P. Klonowski, V. Dagman, P. Pelton, CENSOR - a program for identification and elimination of repetitive elements from DNA sequences, Comput. Chem. 20 (1996) 119-122.

[50] P. Jacquet, P., W. Szpankowski, Autocorrelation on words and its applications, Analysis of suffix trees by string-ruler approach, J. Combinatorial Theory Ser. A, 66 (1994) 237-269.

Figure Captions.

Fig. 1. Correlation analysis. The number of the occurrences of each 9-mer in human genomic DNA (Gi) was plotted against the number of the occurrences of the same 9-mer in proximal promoters of human genes (Ei). The plot does not include Gi values greater than 550,000.

Fig. 2. Statistical evaluations. The plot provides the normalized occurrence of each 9-mer in genomic DNA (Gi / L_G) vs the normalized occurrence of the same 9-mer in proximal promoters of human genes (Ei / L_E). Normalizations were done with respect to total length of the genomic DNA (L_G) and proximal promoter (L_E) sequences that were analyzed for data collection. The lines partition the sequences according to their statistical significance. The plot does not include normalized Ei and Gi values greater than 0.0002.

Fig. 3. Examples of occurrences of 9-base elements defining protein-binding sites. (A) The upstream region of the human NF-IL6 gene (Accession no. AF350408) was analyzed for occurrences of the data with respect to an experimentally defined site that interacts with several members of the SP1 family [25,26]. (B) The upstream region of the rat N-methyl-D-aspartate receptor 1 gene (Accession no. AY157515) was analyzed for data occurrences with respect to experimentally defined sites that interact with SP1 and MAZ [27]. The sequences defined by overlapping or consecutive 9-base elements are shown in bold.

Fig. 4. An example of correlation of occurrences of 9-base elements with in vivo footprints. The promoter region of mouse adenine phosphoribosyltransferase gene (Accession no. M11310) was analyzed for occurrences of the collected data. The regions defined by the 9-base elements are

shown in bold. The underline marks the position of in vivo footprints (FP) produced by dimethyl sulfate [28].

Fig. 5. An example of correlation of occurrences of 9-base elements with footprints produced by DNase I. The promoter region of the human fibronectin (FN) gene (Accession no. M15801) was analyzed for occurrences of the data. The regions defined by the 9-base elements are shown in bold. The underlined regions mark the footprints (FP) produced DNase I [29].

Fig. 6. An example of an extended region correlating with occurrences of 9-base elements. The analysis examined the promoter region of human cyclin D1 gene (Accession no. AF511593). The regions defined by the 9-base elements are shown in bold. The underline marks the region defined by DNA binding and functional assays [30].