Finding human promoter groups based on DNA physical properties

Jia Zeng[*]
Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong

Xiao-Qin Cao
School of Creative Media, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

Hongya Zhao
Department of Genitourinary Medical Oncology, M.D. Anderson Cancer Center, The University of Texas, 1515 Holcombe Boulevard, Houston, Texas 77030, USA

Hong Yan[†]
Department of Electronic Engineering, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

(Received 7 June 2009; revised manuscript received 24 August 2009; published 13 October 2009)

DNA rigidity is an important physical property originating from the DNA three-dimensional structure. Although the general DNA rigidity patterns in human promoters have been investigated, their distinct roles in transcription are largely unknown. In this paper, we discover four highly distinct human promoter groups based on similarity of their rigidity profiles. First, we find that all promoter groups conserve relatively rigid DNAs at the canonical *TATA* box [a consensus $TATA(A/T)A(A/T)$ sequence] position, which are important physical signals in binding transcription factors. Second, we find that the genes activated by each group of promoters share significant biological functions based on their gene ontology annotations. Finally, we find that these human promoter groups correlate with the tissue-specific gene expression.

DOI: 10.1103/PhysRevE.80.041917 PACS number(s): 87.14.gk, 87.14.gf, 82.39.Pj

I. INTRODUCTION

In a DNA sequence, the capacity of transcription factors (TFs) to activate gene expression is encoded in the promoter, which commonly refers to the DNA region that is immediately upstream of the transcription start site (TSS) of a gene. Transcription starts from the $5'$ end to the $3'$ end on the DNA sequence. Conventionally, upstream is in the $5'$ direction, whereas downstream is in the $3'$ direction related to a specific site. The promoter is composed of short regulatory elements that function as transcription factor binding sites (TFBSs) for specific TFs, which control and regulate the transcription initiation of the gene. A core promoter is the region with about ±50 base pairs (bps) centered around the TSS at position 0, and a proximal promoter contains several hundred bps immediately upstream of the core promoter [1]. How TFs rapidly find specific TFBSs in the promoter region has been the subject of intense research for decades but largely remains an unsolved problem. An attractive hypothesis is that the DNA three-dimensional structure contains some physical signals for target site selection by TFs.

The DNA three-dimensional structure can be characterized by the local angular parameters (twist, roll, and tilt) as well as the translational parameters (shift, slide, and rise) between two successive base-pair steps. Considering thermodynamic fluctuations, the sequence-dependent DNA physical property, such as rigidity [2], can be theoretically calculated with the statistical-mechanical model from DNA geometry parameters based on experimental data. Typical examples include the trinucleotide model [3] and the tetranucleotide model [4].

Recently, the average rigidity profile of eukaryotic DNA sequences has been extensively examined in several organisms [5–13], and it has been suggested that DNA rigidity influences DNA looping [14], nucleosome positioning [5,10], promoter activities [9,11,12], TF binding [6,7,9,12], protein-DNA recognition [8,13,15], and DNA replication [13]. Although the general DNA rigidity patterns in human promoters have been examined and used for computational promoter prediction [16,17], their distinct roles in transcription are largely unknown. The fact that the average rigidity profile shares a common physical property raises the question whether human promoters can be partitioned into distinct rigidity-based groups that are biologically significant.

To answer this question, we partition human promoters based on the similarity of their DNA rigidity profiles using the graph-based consensus clustering [18]. We discover four distinct rigidity-based groups of promoters, where each promoter group has a highly different average rigidity profile from the general rigidity patterns of all promoters. We have three important observations. First, all groups conserve a relatively rigid DNA region at the canonical *TATA* box [a consensus $TATA(A/T)A(A/T)$ sequence] position, which may be an important physical signal in binding TFs for the assembly of transcriptional machinery. Second, the gene ontology (GO) annotations for the genes regulated by each group of promoters demonstrate that each promoter group activates genes with the high likelihood to share significant biological processes. Finally, based on the normalized mutual information (NMI) measure, we find that these promoter groups correlate with the tissue-specific gene expression.

[*] j.zeng@ieee.org

[†] Also at School of Electrical and Information Engineering, University of Sydney, NSW 2006, Australia.

II. MATERIALS AND METHODS

A. Promoters

The database of transcriptional start sites (DBTSS) [19] provides a good combination of coverage and quality for DNA sequences around the experimentally determined TSSs. We download all 30 964 human promoter sequences from the DBTSS release 6.0 and extract sequence segments from -200 to $+50$ bp relative to the TSS at position 0 because the range $[-200,+50]$ is generally enriched by a large amount of TFBSs characterizing the whole core and part of the proximal promoter region. We remove those sequences having the wildcard N and obtain a total number of 30 946 promoters with 251 bp in length.

B. DNA rigidity

We adopt the tetranucletide model [4] to calculate the rigidity profiles of human promoters. In the tetranucleotide model, slide and shift are the two principal degrees of freedom compared with twist, roll, tilt, and rise. Slide and shift cannot be predicted because they are strongly correlated in neighboring steps, so the conformational energy E_{step} of a dinucleotide step is a function of slide and shift. This function is used in conjunction with the experimental data on the conformations of tetranucleotides to parametrize an energy function E_{junction}, which couples slide and shift in all three steps,

$$E_{\text{junction}} = (\Delta D_y)^2 \times F_{\Delta D_y} + \left(\sum D_y\right)^2 \times F_{\Sigma D_y} + (\Delta D_x)^2 \times F_{\Delta D_x} + \left(\sum D_x\right)^2 \times F_{\Sigma D_x}, \tag{1}$$

where D_y is slide, D_x is shift, $F_{\Delta D_y}$, $F_{\Sigma D_y}$, $F_{\Delta D_x}$, and $F_{\Sigma D_x}$ are the force constants. The energy of an oligomer of N base pairs is

$$E^N_{\text{oligomer}} = \sum_{n=1}^{N-1} E^n_{\text{step}} + \sum_{n=1}^{N-2} E^n_{\text{junction}}. \tag{2}$$

The rigidity parameters of 136 tetranucleotides are calculated from the curvature in the tetranucleotide potential-energy surface in Eq. (2) with respect to slide at the global energy minimum. The parameters range from the lowest $TACA$ $=1.9$ to the highest $AAAC=27.2$ with arbitrary units. Note that higher values correspond to more rigid sequences, and lower values correspond to more flexible sequences. The large difference of rigidity parameters denotes the large change in physical properties. If we sort rigidity parameters of 136 tetranucleotides in ascending order, we find that the rigidity difference is often 0.1 or 0.2 between two successive tetranucleotides. Therefore, generally 1 unit change in the rigidity profile corresponds to a significant change in locally physical properties. The eight most flexible sequences are all composed of alternating pyrimidine-purine sequences, and the top three involve CA/TG or TA/TA flanking steps. The least flexible steps all involve AA/TT and are predominantly purine rich sequences. Although the potential-energy surface tetranucleotide model is a rough approximation for complex statistically derived properties, such as conformational preferences, it agrees well with the experimental data from x-ray crystal structures. More details about the conversion table of rigidity parameters for all tetranucleotides can be found in [[4], Table III].

At each position of the promoter sequence, we calculate the rigidity value based on 6-mers (6-base-long sequences). While the 6-mer may only reflect a local pattern, it is a practical choice to characterize the sequence-dependent rigidity demonstrated in previous work [9–13,15]. The rigidity of the 6-mer is calculated by summing up rigidity parameters of three overlapping component tetranucleotides,

$$r = \sum_{i=1}^{3} t_i, \tag{3}$$

where i is the positional index, and t_i is the rigidity parameter of each component tetranucleotide at position i. We calculate the rigidity of the 6-mer against its start position. For example, the 7-mer $TATAAAA$ has the rigidity value at the first position T,

$$r_{\text{T}} = t_{\text{TATA}} + t_{\text{ATAA}} + t_{\text{TAAA}}, \tag{4}$$

and the rigidity value at the second position A,

$$r_{\text{A}} = t_{\text{ATAA}} + t_{\text{TAAA}} + t_{\text{AAAA}}. \tag{5}$$

If a sequence is L in length, its rigidity profile is $L-5$ in length based on 6-mers. Therefore, it is possible to calculate the rigidity profile $R=[r_1,\ldots,r_{L-5}]$, for any given sequences based on the conversion table of 136 unique tetranucleotide rigidity parameters in [[4], Table III]. We provide a computational tool for DNA rigidity in [20].

A single rigidity profile is rather noisy, so practically we have to smooth each profile within a 100 bp window in order to obtain reliable partitions of human promoters. After smoothing, all rigidity profiles have the average standard deviation 0.68, which is a small value compared with the average distance 3.54 between the mean profiles of two human promoter groups in Fig. 2. However, the smoothing process will destroy local DNA rigidity patterns and, thus, the smoothed profiles cannot show salient local characteristics. Therefore, after partitioning human promoters based on the smoothed profiles, we show the results of the original profiles correspondingly. We average a large set of profiles in order to retain salient local rigidity patterns that are consistent in the majority of profiles. This method has been used to investigate the general rigidity properties in recent studies [9,12,13,15].

C. Graph-based consensus clustering

We adopt the graph-based consensus clustering (GCC) [18] to partition human promoters into different groups based on similarity of their rigidity profiles. Given a prespecified maximum number of groups $K_{\max}$, GCC automatically identifies the true number of groups for the samples according to a validation index called the modified rand index (MRI). Because we do not have much prior knowledge about the true number of groups in human promoters, GCC is a good choice that provides a robust estimate of the number of groups from a large amount of data. The whole process of the GCC consists of three major steps: subspace generation, subspace clustering, and cluster ensemble. First, GCC selects B subsets of rigidity profiles by random sampling controlled by a uniform random variable. The default parameter ensures that the selected subset of profiles will cover around 80% of the whole set. Second, GCC partitions the selected B subsets of profiles by the k-means algorithm. Through subspace clustering, GCC obtains B solutions that predict class labels of the samples. Finally, GCC constructs a consensus matrix by merging the adjacency matrix of B subspace clustering solutions. We set the number of subsets $B=100$ since it is large enough to produce a good subspace generation performance. The GCC tool is available in [21].

D. Gene ontology

GO [22] is a well-accepted standard for gene function categorization. It is a controlled and structured vocabulary including three categories: molecular function, biological process (BP), and cellular component. GO terms are organized in the form of a directed acyclic graph with two semantic relations, such as "is-a" and "part-of," where A *is-a* B means A is a subclass of B, and C *part-of* D means C is always part of D. Each GO term has a unique numerical identifier such as GO: 0 008 150 and a term such as BP. In the GO annotation, the hypergeometric distribution is applied to calculate the p value of the related BP terms to assess the significance of the discovered groups. For example, if the majority of genes in the group have the same biological function, it is unlikely that this happens by chance and the category's p value would be close to zero. When several categories' p values are less than the threshold, it is reasonable to annotate the group with the category that has the minimum p value. In the GO annotation, we ignore the IEA (inferred from electronic annotation) due to their lack of reliability. We use the DAVID GO analysis tool [23] to find significantly enriched GO terms associated with a list of genes.

E. Tissue-specific promoters

Human promoters control gene expression in a tissue-dependent manner. Tissue-specific genes are only expressed by activating these tissue-specific promoters, leaving the rest of the tissues unmodified. TIPROD [24] is a database of tissue-specific human promoters, which first finds differentially expressed genes in different pools of tissues or samples based on expressed sequence tags, and then associates each promoter sequence with corresponding genes for a certain tissue type. It holds information about 52 tissues and their gene signatures. To enable the selection of tissue-specific promoters from the database, an index of tissue specificity for each gene is provided. The tissue specificity will be close to one for a gene that is expressed in a tissue at an average level compared with other tissues but significantly higher than one if a gene is specifically expressed in that tissue. We retrieve 4423 promoters with tissue specificity above two as the benchmark. This specificity ensures that all these promoters are highly tissue specific. We find that the number of promoters is very small for some tissues, so we choose the eight categories of tissue-specific promoters with the number of promoters higher than 200. As a result, we obtain a total of 1953 promoters for activating eight categories of tissues: "germcell," "kidney," "lymph node," "muscle," "pancreatic islet," "placenta," "skin," and "testis."

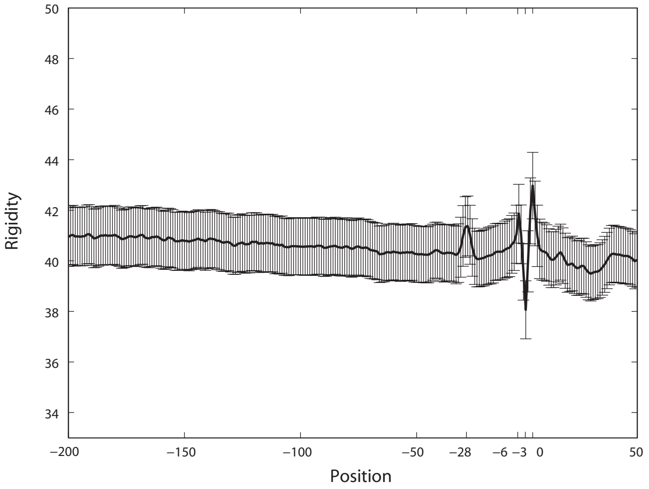

FIG. 1. The average rigidity profile of all 30 946 human promoter sequences as calculated from the tetranucleotide model [4]. The TSS is at position 0. Distinguishable rigidity patterns exist at positions −28, −6, −3, and 0. The region [−200,−50] is slightly more rigid than the region [0,50]. The error bar width is equal to 2 standard deviation of 30 946 smoothed rigidity profiles.

If the human promoter groups based on rigidity is close to the groups of tissue-specific promoters, we have a high confidence that DNA rigidity correlates with tissue-specific gene expression. In a recent comparative study in text mining [25], the NMI is a superior measure to evaluate the closeness of two partitions of data,

$$\mathcal{M}_{\mathrm{NMI}} = \frac{I(P;Q)}{\sqrt{H(P)H(Q)}}, \tag{6}$$

where P and Q are the two partitions of the same data, $I(P;Q)$ is the mutual information between P and Q, and $H(P)$ and $H(Q)$ are the entropies of P and Q, respectively. The $\mathcal{M}_{\mathrm{NMI}}$ value ranges from zero to one, where an $\mathcal{M}_{\mathrm{NMI}}$ value of zero means that P is equal to the almost random partitioning compared with Q, and an $\mathcal{M}_{\mathrm{NMI}}$ value close to one means that P and Q are almost identical partitions. One advantage of the NMI is that it does not require the same number of groups partitioned by P and Q.

III. RESULTS

A. General rigidity patterns

Figure 1 shows the average rigidity profile of all 30 946 human promoter sequences as calculated from the tetranucleotide model. There are three general rigidity patterns [5,9]. First, the TSS at position 0 contains the highly rigid DNAs. There are highly flexible DNAs at position −3 and relatively rigid DNAs at position −6. Thus, the rigidity pattern close to the TSS composes a distinguishable "M" shape. Second,

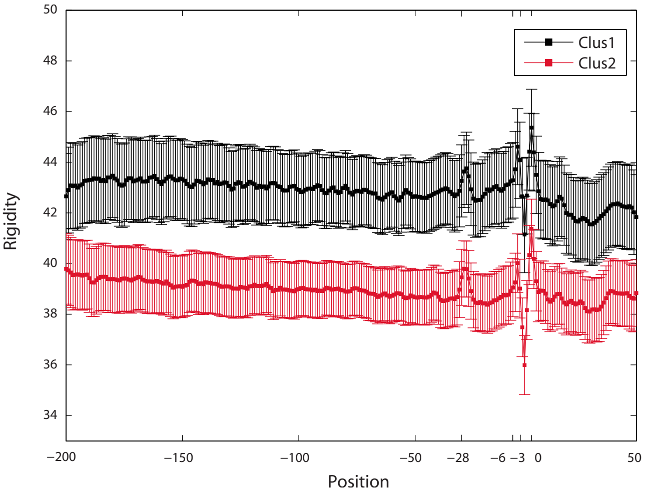

FIG. 2. (Color online) The average rigidity profiles of Clus1 (12 434 promoters) and Clus2 (14 539 promoters). Although they have almost the same fluctuation trend, their rigidity levels differ significantly ($R_{\text{Clus1}}=42.9$ and $R_{\text{Clus2}}=39.0$). The error bar width is equal to 2 standard deviation of 12 434 and 14 539 smoothed rigidity profiles, respectively.

relatively rigid DNAs exist at position −28 corresponding to canonical *TATA* box position. The rigidity value at position −28 is 1.5 higher than its surrounding regions, which denotes a significant physical change as discussed in Sec. II B. Third, the region [−200,−50] is slightly more rigid than the region [0,50]. In our previous study [[12], Fig. 2], however, we found that the region [−200,+50] is highly flexible in a broader range [−1000,+1000]. Therefore, the third general rigidity pattern is only a local property. Indeed, the large-scale rigidity patterns in the range [−2000,+2000] or [−3000,+3000] have recently been used for genome-wide human promoter recognition [26,27].

The rigid region is often correlated with enrichment of *A*/*T* nucleotides. However, this *A*/*T* richness cannot fully account for the physical property of the region, which is also dependent on trinucleotide and tetranucleotide compositions [13]. The TSS is enriched with *TTT* trinucleotides (about 13% promoters), where all tetranucleotides containing *TTT* are rigid, for example, $t_{\text{TTTA}}=18.8$, $t_{\text{TTTC}}=19.3$, $t_{\text{ATTT}}=21.7$, $t_{\text{TTTT}}=23.8$, $t_{\text{TTTG}}=24.3$, $t_{\text{CTTT}}=24.5$, and $t_{\text{GTTT}}=27.2$. The flexible region at position −3 is enriched with CA dinucleotides (about 45% promoters), where most tetranucleotides containing CA are flexible, for example, $t_{\text{TACA}}=1.9$, $t_{\text{TACA}}=3.2$, $t_{\text{TGCA}}=3.8$, $t_{\text{CACG}}=4.1$, $t_{\text{CACT}}=4.8$, and $t_{\text{CACA}}=6.6$. At position −6, the relatively rigid region is CG-rich (about 46% promoters), and *GGGG*/*CCCC* are relatively rigid tetranucleotides. The *TATA* box (about 1% promoters) may partly account for the localized rigid DNA at position −28. Although the *TATA* box contains one of the most flexible flanking steps *TATA* with rigidity parameter $t_{\text{TATA}}=3.6$), its tail contains the highly rigid *AA*/*AA* flanking steps with rigidity parameter $t_{\text{AAAA}}=23.8$. So, the *TATA* box is a half-flexible and half-rigid DNA sequence [12]. We should also note that the *TATA*-less promoters may contribute mainly to the rigid region at position −28. This phenomenon implies that it may be the highly rigid region that plays an important role in assembling the transcriptional machinery in *TATA*-less promoters. To summarize, the nucleotide compositions account for the rigidity patterns in general.

B. Human promoter groups

In our experiments, we take the maximum number of groups in human promoters to be $K_{\text{max}}=30$. In practice, this number is large to cover the possible number of underlying groups in human promoters. After several iterations, the GCC [18] returns two higher MRI values when $K=2$ and $K=4$, which implies that all 30 946 human promoters can be partitioned into either two or four statistically significant groups. When $K=2$, we refer to the discovered groups as Clus1 and Clus2. Interestingly, as far as $K=4$ is concerned, we find that the discovered four groups originate from Clus1 and Clus2, respectively. Specifically, Clus1 can be further partitioned into two groups called Clus3 and Clus4, and Clus2 can be further portioned into two groups called Clus5 and Clus6. The number of groups of more than four provides smaller MRI values. Therefore, human promoters can be broadly partitioned into two groups, and each group can be further partitioned into two smaller groups. To validate the discovered groups, we investigate if Clus3, Clus4, Clus5, and Clus6 can be further partitioned into different subgroups. Based on the same GCC algorithm [18], we find that the subgroups of promoters have no significantly distinct properties (figures not shown here). In addition, the average rigidity profiles of these subgroups often intertwine each other within less than one standard deviation, which indicates that they may not be significantly independent groups.

Figure 2 shows the average rigidity profiles of Clus1 (12 434 promoters) and Clus2 (18 512 promoters). Their profiles are almost the same as that of all the promoters in Fig. 1. For example, they both have the M shapes close to the TSS and a high peak at position −28. While both profiles fluctuate in a similar tendency, their rigidity levels differ significantly. The average rigidity ($R_{\text{Clus1}}=42.9$) of 6-mers in Clus1 is higher than that ($R_{\text{Clus2}}=39.0$) in Clus2. So, Clus2 is more flexible than Clus1 on average. We also observe that both Clus1 and Clus2 have a slightly more rigid region [−200,−50] than the region [0,50], which is consistent with the general rigidity patterns in Fig. 1. It is very interesting to investigate why Clus1 and Clus2 have almost the same rigidity patterns but at different levels.

At the TSS, Clus1 is enriched with *ATTTTT*, while Clus2 is enriched with *AGTTCC*. The higher density of thymine makes Clus1 more rigid than Clus2 at the TSS. Although both Clus1 and Clus2 prefer cytosine and adenine at the positions −1 and 0, Clus1 prefers *TT* but Clus2 prefers CG nearby. They also differ in tetranucleotides at position −6, where Clus1 is thymine rich and Clus2 is CG/GC rich. Indeed, the *TT*/*TT* flanking steps are much more rigid than the GC/CG steps according to the tetranucleotide model. At position −28, Clus1 is enriched with *TTAAAA* corresponding to the *TATA* box, while Clus2 is enriched with *GGAGGG*. Although *TTAAAA* and *GGAGGG* are relatively rigid sequences, *TTAAAA* is more rigid than *GGAGGG*. Thus, the genomic context of Clus1 and Clus2 accounts for their similar rigidity patterns at different levels.

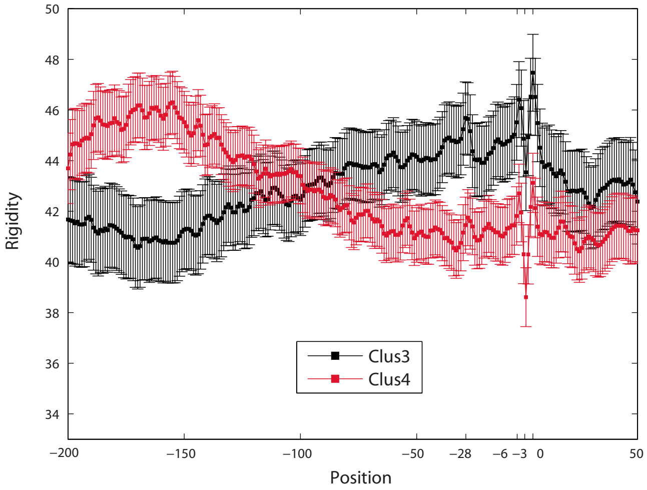

FIG. 3. (Color online) The average rigidity profile of Clus3 (6410 promoters) and Clus4 (6024 promoters). Clus3 increases in the upstream region of the TSS until the peak at the TSS, while Clus4 decreases in the upstream region of the TSS until the valley at position −3. The tendency of two profiles is approximately the reverse in the upstream region of the TSS but retains almost the same "M" shape patterns around the TSS at different levels. The error bar width is equal to 2 standard deviation of 6410 and 6024 smoothed rigidity profiles, respectively.

Clus1 can be further partitioned into Clus3 (6410 promoters) and Clus4 (6024 promoters), whose average rigidity profiles are shown in Fig. 3. Obviously, both profiles are quite different from that in Fig. 1 and seem to be symmetrical with each other. Clus3 has a much more flexible [−200,50] region, whereas Clus4 has a much more rigid region [0,50]. After the position −100, Clus3 and Clus4 have almost the same fluctuation tendency but locate at different rigidity levels akin to Clus1 and Clus2 in Fig. 2. In particular, the rigidity patterns at the positions −28, −3, 0, and downstream of the TSS in Clus3 and Clus4 are almost the same with those in Figs. 1 and 2. For example, both Clus3 and Clus4 conserve the relatively rigid DNA at position −28 and retain the same M shape rigidity profile around the TSS.

Clus3 is enriched with relatively flexible *CCTTCC* and rigid *TTAAAA* at the positions −170 and −28. Thus, Clus3 increases steadily in the upstream region of the TSS. On the other hand, Clus4 has the relatively rigid *ATTTTT* and flexible *ATAAAG* at the positions −170 and −28. So, Clus4 decreases steadily in the upstream region of the TSS. Another big difference is that Clus3 is thymine dominant but Clus4 is CT rich around the TSS, so that Clus3 locates at a higher rigidity level than Clus4.

Clus2 can be further partitioned into Clus5 (8199 promoters) and Clus6 (10 313 promoters), whose average rigidity profiles are shown in Fig. 4. Both profiles have no significant difference in the upstream and downstream regions of the TSS. For example, they both conserve the relatively rigid region at position −28 similar to Clus2. We see that Clus5 and Clus6 are quite different around the TSS. For example, Clus5 has only a highly rigid peak at the TSS while Clus6 has only a highly flexible valley at position −3. Interestingly enough, Clus5 and Clus6 do not retain the M shape rigidity pattern in Fig. 1 around the TSS. However, since Clus5 is highly rigid at the TSS whereas Clus6 is highly flexible immediately upstream of the TSS, the combined profile of Clus5 and Clus6 will restore the M shape rigidity pattern as that of Clus2. In this sense, Clus5 and Clus6 share similar physical properties with Clus2.

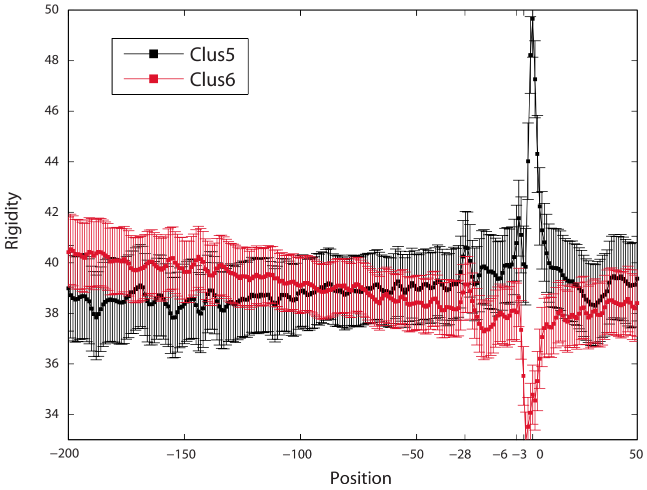

FIG. 4. (Color online) The average rigidity profiles of Clus5 (8199 promoters) and Clus6 (10 313 promoters). Clus5 is highly rigid at the TSS, while Clus6 is highly flexible at position −3. Both Clus5 and Clus6 are enriched with cytosine and guanine at the positions −6 and −28 and prefer cytosine and adenine at the positions −1 and 0, respectively. The error bar width is equal to 2 standard deviation of 6468 and 8071 smoothed rigidity profiles, respectively.

Clus5 (34.02% promoters in Clus5) prefers thymine but Clus6 (59.06% promoters in Clus6) prefers guanine or cytosine at positions +1, +3, and +4. At position −3, Clus5 is enriched with *CATT* ($t=12.7$) and *CAGT* ($t=11.9$), which are more rigid than the corresponding steps *CAGT* ($t=11.9$) and *CACT* ($t=4.8$) in Clus6. At the TSS, Clus5 is enriched with *AGTTTC*, where *TTTC* ($t=19.3$) is much more rigid than the corresponding step *TGCC* ($t=12.1$) of the enriched motif *AGTGCC* in Clus6.

In conclusion, human promoters can be broadly partitioned into two rigidity-based groups Clus1 and Clus2 in Fig. 2. Furthermore, Clus1 can be partitioned into two groups Clus3 and Clus4 (Fig. 3), and Clus2 can be partitioned into two groups Clus5 and Clus6 (Fig. 4). Important observations are as follows. First, Clus1 and Clus2 have almost the same rigidity patterns but at different rigidity levels. Second, Clus3 and Clus4 have highly flexible and highly rigid regions [−200,−50], respectively. Third, Clus5 (Clus6) contains only highly rigid (flexible) DNAs around the TSS. Finally, all groups conserve the localized rigid DNAs at position −28 but at different levels.

C. Biological functions

To show the validity of the resulting promoter groups, we use the GO to perform functional analysis of genes activated by each group of promoters. Because Clus1 and Clus2 are two broad groups, we only examine the GO annotations of

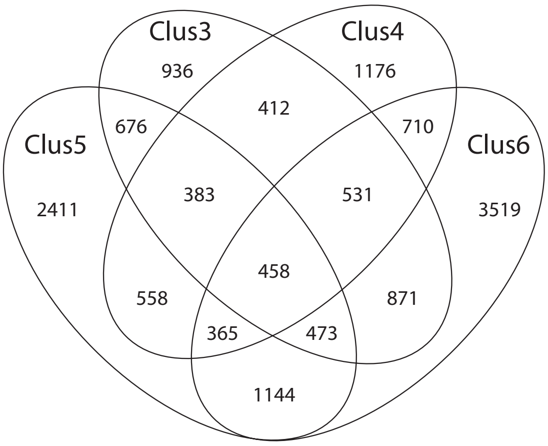

FIG. 5. The Venn diagram of the number of activated genes by different promoter groups.

Clus3, Clus4, Clus5, and Clus6. Each promoter sequence in the DBTSS is associated with an Entrez ID denoting the activated gene. Each gene can be activated by multiple alternative promoters [17], and 30 946 promoters in the DBTSS activate a total number of 14 623 unique genes. The repetitive genes activated by each promoter group are not considered in the GO annotation because they have the same biological function. As a result, 6410 promoters in Clus3 activate 4740 different genes, 6024 promoters in Clus4 activate 4593 different genes, 6468 promoters in Clus5 activate 6468 different genes, and 8071 promoters in Clus6 activate 8071 different genes. Since promoters in different groups may activate the same genes, there is an overlap among the four corresponding gene groups. This overlap will influence the GO annotation of each gene group because the overlapped genes tend to be annotated with the same GO term. Therefore, it is necessary to remove those overlapped genes in order to obtain the proper GO annotation. Finally, we obtain 936, 1176, 2411, and 3519 nonoverlapped genes for groups Clus3, Clus4, Clus5, and Clus6, respectively. The total number of annotated genes is 8042, which accounts for 55% of 14 623 genes in the DBTSS. Since the majority of the genes are included in the GO analysis, we believe that the final GO annotation can reflect the biological functions of four gene groups activated by different promoter groups. Figure 5 shows the Venn diagram of the number of activated genes by Clus3, Clus4, Clus5, and Clus6. We see that coactivated genes by different promoter groups account for about the half of total number of genes. Besides the GO annotations of genes activated by only one promoter group, we are also interested in the GO annotations for genes activated by multiple promoter groups. For example, it would be interesting to find the function of genes coactivated by promoters from Clus3 and Clus4. When the specific gene has more than one GO term, we select the top three GO terms that have the minimum p values.

Table I lists the top three significant GO terms in BP category associated with genes activated by Clus3, Clus4, Clus5, Clus6, and their various combinations in Fig. 5. Obviously, the four groups of genes activated by Clus3, Clus4, Clus5, and Clus6 are associated with quite different GO terms representing different biological processes. For example, Clus3 involves response to external stimulus and Clus5 is for cellular biosynthetic process. Some combinations of gene groups share similar functions. For example, both Clus4-6 and Clus3-4-6 involve the protein modification process. This observation is reasonable because the genes are activated by similar combinations of promoter groups. From Table I, we also observe that genes activated by Clus3, Clus4, Clus5, and Clus6 have a higher likelihood (lower p value) of being biologically significant compared with their combinations, which demonstrates that the discovered promoter groups based on DNA rigidity may also be biologically significant in terms of their roles in gene expression regulation. The genes activated by Clus4-5-6 and Clus3-4-5-6 also have lower p values, which implies that some basic and common biological processes may be involved in all clusters.

It is also important to investigate the role of DNA rigidity in tissue-specific promoters. To measure the distance between the discovered promoter groups and the tissue-specific promoter groups, we select the same 1953 promoters from Clus3, Clus4, Clus5, and Clus6 as those in the benchmark tissue-specific promoter sets. We have two different partitions of these 1953 genes. The first partition leads to four promoter groups based on rigidity profiles, and the other partition results in eight groups of tissue-specific promoters according to the TIPROD database [24]. The NMI value between these two partitions is 0.42 from the rigidity-based groups to the tissue-specific groups. To show how significant this NMI value is, we perform 1000 random partitions of all 1953 promoters into four groups and obtain the average NMI value 0.018 compared with the tissue-specific promoter groups. Obviously, the discovered groups have a significantly higher NMI value than random groups. So, the discovered rigidity-based groups are significantly closer to the tissue-specific promoter groups. This observation demonstrates that DNA rigidity may also correlate with the tissue-specific promoters.

IV. DISCUSSION

In Fig. 2, Clus1 and Clus2 have almost the same rigidity patterns but at different levels, so it is not the specific rigidity pattern but the different rigidity level that differentiates these two groups. This observation supports the hypothesis that the proportion of flexible region in the whole fragment influences promoter activity [9]. Since Clus2 is much more flexible than Clus1, the promoters in Clus2 may be more active than those in Clus1. According to the cap analysis of gene expression, mammalian promoters can be broadly classified into *TATA*-rich and CpG-rich [a large concentration of cytosine (C) and guanine (G) pairs] promoters, and different tissues and families of genes differentially use distinct types of promoters [28]. Our results also support this classification. Clus1 (*TATA*-rich) represents a minority of the set of human promoters (46.1%), and it may commonly associate with tissue-specific genes and high conservation across species. On the other hand, Clus2 (CpG-rich) represents the majority of the human promoters (53.9%), and it seems to be particu-

TABLE I. GO functional analysis of the gene groups in biological process. The number in square brackets gives the power of ten multiplying the preceding number.

Groups	GO terms	Name	p value
Clus3	9605/6952/9611	Response to external stimulus/defense response/response to wounding	6.7[−8]/2.9[−7]/4.5[−7]
Clus4	6952/7156/48731	Defense response/homophilic cell adhesion/system development	2.5[−6]/5.1[−6]/3.9[−5]
Clus5	44249/6412/8152	Cellular biosynthetic process/translation/metabolic process	4.3[−9]/9.7[−9]/1.1[−8]
Clus6	32502/16043/22402	Developmental process/cellular component organization and biogenesis/cell cycle process	9.4[−14]/1.6[−10]/1.2[−9]
Clus3-4	51276/6325/6323	Chromosome organization/chromatin organization/DNA packaging	1.3[−3]/2.1[−3]/2.6[−3]
Clus3-5	43170/44238/44237	Macromolecule metabolic process/primary metabolic process/cellular metabolic process	5.4[−7]/1.8[−6]/4.8[−6]
Clus3-6	15031/45184/33036	Protein transport/establishment of protein localization/ macromolecule localization	8.9[−6]/1.5[−5]/1.7[−5]
Clus4-5	9161/9156/9124	Ribonucleoside monophosphate metabolic process/ ribonucleoside monophosphate biosynthetic process/ nucleoside monophosphate biosynthetic process	1.3[−4]/1.3[−4]/1.7[−4]
Clus4-6	43412/6464/43283	Biopolymer modification/protein modification process/ biopolymer metabolic process	8.2[−4]/8.5[−4]/3.5[−3]
Clus5-6	7242/6468/7265	Intracellular signaling cascade/protein amino acid phosphorylation/Ras protein signal transduction	1.3[−4]/1.7[−4]/2.4[−4]
Clus3-4-5	9966/16043/16192	Regulation of signal transduction/cellular component organization and biogenesis/vesicle-mediated transport	7.6[−6]/1.8[−5]/2.2[−5]
Clus3-4-6	6464/43412/43687	Protein modification process/biopolymer modification/ post-translational protein modification	2.7[−5]/6.2[−5]/1.7[−4]
Clus3-5-6	16568/32502/16043	Chromatin modification/developmental process/cellular component organization and biogenesis	1.1[−5]/1.1[−5]/2.7[−5]
Clus4-5-6	16043/32502/6810	Cellular component organization and biogenesis/ developmental process/transport	9.5[−9]/2.1[−5]/3.6[−4]
Clus3-4-5-6	7010/30036/30029	Cytoskeleton organization and biogenesis/actin filament-based processactin cytoskeleton organization and biogenesis/actin filament-based process	3.7[−8]/1.5[−7]/4.5[−7]

larly rapidly evolving in mammals and is more active in transcription.

One interesting finding is that all groups conserve the relatively rigid DNAs at the canonical *TATA* box position −28, where Clus3 and Clus4 are *TA*-rich but Clus5 and Clus6 are GC rich (this rigid region is not very salient in Clus4 compared with that in Clus3, Clus5, and Clus6). Fukue *et al.* [9] also observed that both *TATA*-containing and *TATA*-less promoters contain highly rigid DNAs at the −28 position. Based on the experiments on synthetic DNA fragments, they suggested that rigid DNAs around −28 have some positive influence on transcription. Furthermore, Tirosh *et al.* [11] compared rigidity profiles among 11 yeast species and found similar rigid DNAs conserved in the *TATA*-less promoters, which could assist in the assembly of the transcriptional machinery. Our results reveal that these rigid DNAs in *TATA*-less promoters are caused by GC-rich tetranucleotides. Because the nucleotide bases vary considerably in all four groups at position −28, we speculate that it is the relatively rigid DNAs rather than specific nucleotides that facilitate complex protein-DNA interactions to initiate transcription. Proteins find a specific target site along a DNA sequence through three distinct search mechanisms [29,30], i.e., the sliding mechanism, the intersegment transfer mechanism, and the hopping/jumping mechanism. Lower salt concentration can make DNAs more rigid in order for short-range sliding of proteins along DNAs [2,13]. So, TFs may slide along the rigid DNAs at position −28 to accurately locate the target site through the indirect readout mechanism. On the other hand, it has been shown that increased variability by the targeted mutation of the *TATA* box in gene expression can be beneficial after an acute change in environmental conditions [31]. In contrast to the *TATA*-less promoters, the *TATA*-containing promoters are more likely to enable a rapid individual cell response and increased cell-cell variability through transcriptional bursting, which provides a clear benefit confronted with an environmental stress. We speculate that such difference between the *TATA*-rich and *TATA*-less promoters may relate to their difference in the overall rigidity levels as shown in Fig. 2. Although *TATA*-rich Clus1 and *TATA*-less Clus2 have almost the same local rigidity patterns, they are significantly different in terms of the average rigidity value, which may lead to different stress responses through the stochastic process of transcriptional bursting.

Both GO and tissue-specific promoters support that the discovered human promoter groups differ significantly in biological functions. For example, Clus3 is mainly associated with response to external stimulus, Clus4 is associated with defense response, Clus5 is related to cellular biosynthesis process, and Clus6 is involved in developmental process. Also the partition based on similarity of rigidity profiles is closer to the partition based on tissue-specific gene expression, which implies that DNA rigidity may be an important characteristic in tissue-specific promoters.

V. CONCLUSIONS

We have found that human promoters can be broadly partitioned into two groups Clus1 and Clus2, which have almost the same rigidity patterns but at different rigidity levels. Furthermore, Clus1 can be partitioned into two groups Clus3 and Clus4. Clus2 can also be partitioned into two groups Clus5 and Clus6, which are quite different around the TSS. All groups conserve the relatively rigid DNAs at the canonical *TATA* box position, which may be important physical signals for binding TFs by the indirect readout and sliding mechanisms. Based on the GO annotations and tissue-specific promoters, we demonstrate that the discovered promoter groups differ significantly in biological functions.

The canonical core promoter elements consist of the *TATA* box, initiator (Inr), downstream core promoter element, TFIIB recognition element, and the motif 10 element. The synergic combinations of core promoter elements are important in transcription. In future work, we shall investigate the relationship between DNA rigidity patterns and the synergy of core promoter elements in order to better understand how DNA physical properties influence DNA transcription.

ACKNOWLEDGMENTS

This work was partially supported by the Hong Kong Research Grant Council (Project CityU No. 123408) and the City University of Hong Kong (Project No. 9610034).

[1] S. T. Smale and J. T. Kadonaga, Annu. Rev. Biochem. **72**, 449 (2003).

[2] N. L. Goddard, G. Bonnet, O. Krichevsky, and A. Libchaber, Phys. Rev. Lett. **85**, 2400 (2000).

[3] I. Brukner, R. Sanchez, D. Suck, and S. Pongor, EMBO J. **14**, 1812 (1995).

[4] M. J. Packer, M. P. Dauncey, and C. A. Hunter, J. Mol. Biol. **295**, 85 (2000).

[5] A. G. Pedersen, P. Baldi, Y. Chauvin, and B. Soren, J. Mol. Biol. **281**, 663 (1998).

[6] D. B. Starr, B. C. Hoopes, and D. K. Hawley, J. Mol. Biol. **250**, 434 (1995).

[7] A. Grove, A. Galeone, L. Mayol, and P. E. Geiduschek, J. Mol. Biol. **260**, 120 (1996).

[8] D. M. Gowers and S. E. Halford, EMBO J. **22**, 1410 (2003).

[9] Y. Fukue, N. Sumida, J. I. Tanase, and T. Ohyama, Nucleic Acids Res. **33**, 3821 (2005).

[10] K. Florquin, Y. Saeys, S. Degroeve, P. Rouze, and Y. V. Peer, Nucleic Acids Res. **33**, 4255 (2005).

[11] I. Tirosh, J. Berman, and N. Barkai, Trends Genet. **23**, 318 (2007).

[12] X.-Q. Cao, J. Zeng, and H. Yan, Phys. Rev. E **77**, 041908 (2008).

[13] X.-Q. Cao, J. Zeng, and H. Yan, Phys. Biol. **5**, 036012 (2008).

[14] K. S. Matthews, Microbiol. Mol. Biol. Rev. **56**, 123 (1992).

[15] X.-Q. Cao, J. Zeng, and H. Yan, Phys. Biol. **6**, 036012 (2009).

[16] J. Zeng, X.-Y. Zhao, X.-Q. Cao, and H. Yan, IEEE/ACM Trans. Comput. Biol. Bioinf. (to be published).

[17] J. Zeng, S. Zhu, and H. Yan, Briefings Bioinf. **10**, 498 (2009).

[18] Z. Yu, H. Wong, and H. Wang, Bioinformatics **23**, 2888 (2007).

[19] R. Yamashita, Y. Suzuki, H. Wakaguri, K. Tsuritani, K. Nakai, and S. Sugano, Nucleic Acids Res. **34**, D86 (2006).

[20] http://www.comp.hkbu.edu.hk/jiazeng/Software/Flexibility/flexibility.html

[21] http://bioinformatics.oxfordjournals.org/cgi/content/full/btm463/DC1

[22] http://www.geneontology.org/

[23] http://david.abcc.ncifcrf.gov/tools.jsp

[24] X. Chen, K. H. J. M. Wu, A. Kel, and E. Wingender, Nucleic Acids Res. **34**, D104 (2006).

[25] S. Zhong and J. Ghosh, Knowledge Inf. Syst. **8**, 374 (2005).

[26] T. Abeel, Y. Saeys, E. Bonnet, P. Rouze, and Y. Van de Peer, Genome Res. **18**, 310 (2008).

[27] J. R. Goni, A. Perez, D. Torrents, and M. Orozco, Genome Biol. **8**, R263 (2007).

[28] P. Carninci *et al.*, Nat. Genet. **38**, 626 (2006).

[29] T. Hu and B. I. Shklovskii, Phys. Rev. E **74**, 021903 (2006).

[30] T. Hu and B. I. Shklovskii, Phys. Rev. E **76**, 051909 (2007).

[31] W. J. Blake, G. Balazsi, M. A. Kohanski, F. J. Isaacs, K. F. Murphy, Y. Kuang, C. R. Cantor, D. R. Walt, and J. J. Collins, Mol. Cell **24**, 853 (2006).