Super Paramagnetic Clustering of DNA Sequences

SUGIARTO RADJIMAN*, HAN LIANYI, WANG JIAN-SHENG
and CHEN YU ZONG

*Department of Computational Science, National University of Singapore, Singapore 117543,
Republic of Singapore*

*(*Author for correspondence, e-mail: sugiarto@cz3.nus.edu.sg)*

Abstract. An unsupervised clustering of 4541 DNA sequences containing active promoter regions from vertebrate and arthropod classes (including their viral genes) was performed. All necessary information was solely gathered a priori from the DNA sequences by measuring frequencies of tri-nucleotides and tetra-nucleotides. We employed Super Paramagnetic Clustering, a novel clustering algorithm based on physical properties of an inhomogeneous granular ferromagnet. This method utilizes Swendsen-Wang cluster Monte Carlo simulations to distinguish clusters by measuring pairs of correlation functions from different resolutions. We identified two strongly separated clusters of human viral genes corresponding to the Epstein-Barr virus and the Herpes Simplex virus type 1. In addition, vertebrate and arthropod sequences were successfully separated into two different classes with merely 9.25% of arthropod sequences being misclassified. From a functional perspective, these sequences have high gene function correlations with sequences from the vertebrate cluster. By tuning a clustering parameter, Super Paramagnetic Clustering was able to classify vertebrate class further into two major clusters, from where a large number of housekeeping genes and tissue-specific genes were found respectively. The indications came from observation of gene expression function and consensus transcription factors which were found grouped together in specific positions of the DNA sequences.

Key words: DNA sequence, promoters, transcription factor binding sites, cluster algorithm, data clustering, statistical physics

Introduction

The study of DNA sequences as building blocks of living multicellular organisms provides opportunities to potentially understand any biological activity in a living system. Residing inside the nucleus of cells, DNA carries genetic information necessary for the organization of cells and controls the inheritance of characteristics. Nevertheless, even though massive amounts of genome-related information is now available due to the increasing number of genes sequenced and the use of new technologies such as DNA microarrays, extracting information from such data is still a challenging task. Novel ways of analyzing a wide variety of genomic data are therefore critical in order to mine useful information about gene function.

Transcription is an important process in the flow of genetic information. During transcription, the initial step of gene expression, the transformation of DNA information into functional molecules, occurs. It works as the central mechanism by which the cell regulates proteins that are to be produced. Viewed from a more detailed scale, we have DNA initiation sites or promoters which work as the simplest regulatory element in the process of transcription. The promoters function as switches and code words that facilitate the identification of the proper transcription initiation site by RNA polymerase. Information encoded ensures that RNA polymerase in cooperation with sigma factor (prokaryotic) or transcription factor (eukaryotic) specifically recognizes the corresponding region of the genome that needs to be decoded. In eukaryotes, frequency of transcription initiation depends heavily on trans-acting factors such as transcription factor (TF). Besides being principally responsible for recognizing promoter regions, TF occasionally performs other actions such as recognizing another factor, enhancer, or RNA polymerase. In summary, promoters and TFs function as key cis-acting and trans-acting regulatory elements in the regulation of gene expression [1]. A comprehensive study of DNA sequences together with promoters and its TFs will therefore help us to understand gene co-regulation on the transcription level.

An abundance of research in the area of DNA sequence analysis, promoter detection and recognition can be found [2–8]. In this work our aim is to analyze and extract the correlated behavior of gene sequences based on their expression patterns with an unsupervised clustering algorithm. For this purpose, only sequences with active promoter regions are selected. These sequences with so-called strong promoters are transcribed frequently. Hence these genes are biologically important in regulating gene expression.

Quantitative Measurement of DNA Sequences

The set of gene sequences is taken from The Eukaryotic Promoter Database (EPD), a rigorous genome database containing a collection of non-redundant eukaryotic RNA Polymerase II (POL II) promoters [9]. These promoters have been experimentally determined to be active in higher eukaryotes with an accuracy of ± 5 base pair (bp) for the transcription initiation sites. We assembled $N = 4541$ sequences as a data source by combining 2000 sequences from arthropods (about 99% from *Drosophila* families) with 2541 from vertebrates (about 74% from *Homo sapiens*). The sequences also include a number of viral genes from arthropod and vertebrate. Other divisions, which contributed only a small number of DNA sequences, were removed to eliminate bias in the data. We then aligned all the sequences beforehand relative to a putative Transcription Starting Site (TSS). To obtain some numerical measurements from the data, we counted frequencies of tri-nucleotides and tetra-nucleotides by shifting one nucleotide for each sequence s_i , $i = 1, 2, \dots, N$ from -9999 to 6000 bp regardless of its position. Note that a number of sequences have

shorter length than 16 000, and the mean length is 13 258. Frequencies of higher nucleotides were not used so as to keep the problem feasible. This is required to reduce computational complexities because each combination of nucleotides of length L , which has a one-to-one mapping with the number of features (dimension), increases exponentially as function of length. In addition, by using less stringent conditions, redundancies and noise in feature measurement are removed. Altogether we have a total of 321 features, from where 64 are evaluated from 3-mers, 256 from 4-mers, and one additional feature which measures the length of each DNA sequence. This provides a 4541×321 input matrix of measurement. Data entries are normalized for each column, such that all values fall between zero and one,

$$\hat{m}_{ij} = m_{ij} / \max(m_{ij}), \quad j = \{1, 2, \dots, 321\}.$$

The Euclidean distance $d_{mn} = \|\vec{v}_m - \vec{v}_n\|$ is calculated, and fed into the symmetric distance matrix in our clustering algorithm.

Super Paramagnetic Clustering

The classification was performed with the Super Paramagnetic Clustering (SPC) algorithm to mine information from the dataset. SPC is a novel hierarchical clustering method inspired by physical behavior of the inhomogeneous ferromagnet. The advantages of this method are its robustness against noise and initialization, ability to generate automatic hierarchical structure, and most importantly the fact that no prior knowledge about distribution of data or cluster structure is needed. SPC fits for our problem perfectly since any assumptions and prior information about the DNA sequences are minimized. This method has been successfully tested in various problems, for instance Iris flowers, Landsat data, and yeast gene expression profiles. Another variation of this method called the Coupled Two-Ways Clustering (CTWC) has also been applied to analyze gene microarray data of breast cancer, colon cancer, and leukemia [10, 11–13]. A detailed explanation of the algorithm can be found in papers by Blatt, Wiseman, and Domany [14–17]. Here only a brief introduction to the method is provided.

First, each data $\vec{v}_i$, $i = \{1, 2, \dots, N\}$, is a feature vector that corresponds to a point in $d = 321$ dimensional space. One associates randomly one of q possible Potts spin variables s_i , $i = \{1, 2, \dots, q\}$ to each point $\vec{v}_i$. This procedure is the basic physical framework for Potts model simulation. The next step is to find neighbors for every point $\vec{v}_i$. Instead of using all $\vec{v}_j$, $j = \{1, 2, \dots, N\}$, choose only a selected number of $\vec{v}_j$ as neighbors to limit the number of interactions according to the mutual neighborhood criterion. With this criterion, $\vec{v}_i$ and $\vec{v}_j$ are considered to be neighbors if and only if $\vec{v}_i$ is one of K -nearest neighbors of $\vec{v}_j$, and $\vec{v}_j$ is one of K -nearest neighbors of $\vec{v}_i$. Consequently, computational efficiency increases and simulation running time is greatly reduced. We fixed $K = 10$ for the simulation. Additionally,

we impose the edges from the minimum spanning tree to make sure that a connected graph which links all the data exists. Subsequently, we let each Potts spin variable interact with its neighbors with strength J_{ij} based on a cost function,

$$J_{ij} = \frac{1}{\hat{K}} \exp\left(-\frac{d_{ij}^2}{2a^2}\right).$$

This cost function of distance decreases exponentially fast such that points in high density regions have stronger interactions than those in low density regions. Here the local length scale a is the average of all distances d_{ij} between neighboring pairs $\vec{v}_i$ and $\vec{v}_j$, and $\hat{K}$ is the average number of neighbors. An efficient Swendsen-Wang Monte Carlo algorithm [18, 19] is then implemented to calculate a thermal quantity, the susceptibility χ , which is proportional to the variance of magnetization m

$$\chi = \frac{N}{T}(\langle m^2 \rangle - \langle m \rangle^2),$$

$$m = \frac{qN_{\max} - N}{(q-1)N},$$

with $N_{\max} = \max\{N_1, N_2, \dots, N_q\}$, where N_μ , $\mu = \{1, 2, \dots, q\}$ is the number of spins in state μ . We used $q = 20$ in the simulation. However, the results are insensitive to the choice of q . This is because the number of Potts states q only affects the sharpness of the transition but does not change the clustering results [14, 17]. The susceptibility serves as a detector to monitor changes in phase transitions and major cluster break-up, especially at temperatures where prominent peaks and sudden decreases in susceptibility take place. Concurrently, we also calculate the spin-spin correlation function G_{ij} for neighboring pairs of spins to decide whether two spins belong to the same grain

$$G_{ij} = \frac{(q-1)C_{ij} + 1}{q},$$

where C_{ij} is the probability of $\vec{v}_i$ and $\vec{v}_j$ belonging to the same Swendsen-Wang cluster; it is the total number of times $\vec{v}_i$ and $\vec{v}_j$ are in the same grain divided by the total number of Monte Carlo iterations. Clusters are identified as those connected subgraphs that are built by linking each point via its neighbors with $G_{ij} > 0.5$ and maximum correlation $G_{ij'}$ such that for any point $\vec{v}_i$: $G_{ij'} \geq G_{ij} > \frac{1}{q}$, $j = \{\text{neighbors of } i\}$.

Having a direct affinity with the Ising model, the core parameter for SPC is temperature T . As a control of resolution, its function is to determine details of the clustering result. Running a simulation with higher T means that more detailed classification is achieved. To be more specific, we examine how the system behaves at different temperatures. At low temperature, individual spins have strong interactions with their neighbors; most spins are aligned with high probability. In fact,

at very low temperature near $T = 0$, all the spins are in the same state due to the existence of a connected graph, with a high spin-spin correlation function among neighbors. In this condition all points merge into a single cluster and the system is fully magnetized in the ferromagnetic phase. Increasing T gradually causes the energy of the system to increase, clusters start to melt into several big and small clusters, letting highly correlated data remain in the same group, and at the same time separating clusters with weak interactions into different groups. The system is therefore in the super-paramagnetic phase. We are most interested working at this state, where a good clustering result is expected. Finally, at very high temperature, the interactions are weak even for close neighboring spins, and all the spins flip randomly. Thus there are low spin-spin correlation function among neighbors, causing clusters to break down. In this state the system is in the disordered paramagnetic phase. These three different system states can be detected by continuously monitoring susceptibility density $\frac{\chi T}{N}$ (variance of magnetization). The indication of the first phase transition from ferromagnetic to super-paramagnetic is a pronounced peak in susceptibility density. It happens at the temperature where the susceptibility density reaches its maximum value. On the other hand, the second phase transition from super-paramagnetic to paramagnetic region is indicated by an abrupt decrease in susceptibility density.

Results and Discussion

The system behavior at different temperatures is monitored by plotting susceptibility density. Starting the simulation from initial temperature $T = 0$, we increase temperature gradually till it reaches 0.16, where the value of susceptibility decreases close to zero. For every sampled temperature we did 2×10^5 Monte Carlo sweeps with the Swendsen-Wang algorithm. Focusing only on main clusters in our analysis, a threshold is set and only those connected points with at least 20 members are considered as a cluster. Those clusters with sizes below the cut-off value are left as unclassified points. At low nonzero T we detected two stable clusters of human viral genes Herpes Simplex Virus type 1 (HSV-1) and Epstein-Barr Virus (EBV), near $T \approx 0$ and $T = 0.035$ respectively. The DNA sequences in each cluster share same homology and exhibit more than 50% similarity between -79 and 20 bp [9]. SPC manages to mine 29 (100%) HSV-1 genes and 22 out of 23 ($\approx 96\%$) EBV genes. Taking into account their sizes and threshold value, both clusters are stable because they only dissolve when T reaches 0.06 and 0.05.

We observe two peaks in the susceptibility density curve (Figure 1). The first sharp peak around $T = 0.05$ signals how the system undergoes major divisions into two separate clusters of arthropod and vertebrate classes with size 1704 and 2548. We find 185 entries (9.25%) of arthropod sequences being misclassified into the vertebrate group. From this group, 36 entries ($\approx 19\%$) come from DNA sequences of polyphemus moth and silkworm, whose function is to regulate the synthesis of

Figure 1. (a) Two peaks were obtained in the measurement of susceptibility density as an indication of major cluster split. (b) The size of four biggest clusters plotted as function of temperature. Notice how the cluster sizes changed drastically right after susceptibility density curve reached maximum position.

chorion-related proteins. Interestingly, the chorion, secreted by the follicle cells, is either a protective membrane which surrounds the eggs of insects and fishes, or the extra-embryonic membrane derived from the trophoblast which surrounds the embryo of amniote vertebrates [20]. The others ($\approx 15\%$) include heat shock protein, ribosomal, histone, and alcohol dehydrogenase family proteins which are also commonly found in the vertebrate cell system.

The first major breaking of clusters suggests that vertebrate and arthropod genes have strong dissimilarities in their genome pattern arrangement. Our study suggests that some of the normalized frequencies of tri-nucleotides are the determining factors which separate vertebrate and arthropod gene sequences. We observed high concentrations of AGG, CCC, CCT and low concentrations of ACG, CAA, CGA,

Figure 2. Distinct levels in normalized tri-nucleotide frequencies of arthropod (solid) and vertebrate (dot) classes that contribute significantly to their separation are shown in 16 out of 64 tri-nucleotides (AAC, AAT, ACG, AGG, ATC, CAA, CAG, CCC, CCT, CGA, CGC, CGG, CGT, GCG, TCG, and TCT).

CGT in the vertebrate class. On the contrary, high concentrations of AAC, ACG, ATC, CGA, TCG and low concentrations of AGG, CCC, CCT were present in the arthropod class (Figure 2). This result agrees well with Abe *et al.*'s result in utilizing the self-organizing map (SOM), an unsupervised neural network algorithm, for analyzing prokaryotic and eukaryotic genomes [21].

The second phase transition near $T = 0.09$ has a lower peak; it signals the breaking of the vertebrate cluster into two separate clusters of size 910 and 1414. The smaller cluster (Cluster VTB1) displayed high positive correlations within

its sequence expression patterns, whereas the bigger cluster (Cluster VTB2) has diverse variations in functionality, and is largely dominated by unclassified chromosomal genes. To further examine correlations inherent in each respective cluster, we searched for consensus Transcription Factor Binding Sites (TFBS) motifs [22, 23] in the region -1000 to 1000 , given the fact that most binding sites tend to cluster together and take place near TSS [24, 25]. All chunks of gene sequences with length 2000 bp are divided into 20 smaller fragments of length 100 bp each. The 1st DNA fragment starts from upstream region -1000 to -901 . We then searched every fragment for consensus octa-nucleotides TFBS and filtered the number of occurrences so that only those consensus motifs with frequency above a threshold value 40 were kept for the analysis.

Our investigation shows that several TFBS motifs have been clustered together in certain positions of DNA sequences. A number of groups of Sp1 TFBS positioned from -200 to 100 bp and TATA box in the upstream region from -100 to TSS were identified in cluster VTB1. Indeed, the TATA box, as the most important cis-acting element for most genes transcribed by RNA polymerase II, is usually centered in the upstream region between -100 and -30 [26]. Similarly we identified Sp1, ETS, NRF-1, CpG and TFBS packed within region -400 to TSS in cluster VTB2 (Table I).

The most important thing to point out from analysis of TFBS consensus motifs is the single occurrence of the TATA box motif in cluster VTB1 and the union of ETS, NRF-1, and CpG in cluster VTB2. Firstly, although the TATA box is commonly found in most eukaryotic genes, it is absent in some genes, particularly housekeeping genes expressed in all tissues and in some tissue-specific genes [1, 27, 28]. This attribute has been revealed in two independent studies by Fitz *et al.* [24] and Murakami *et al.* [29], which indicate TATA box association with promoters of tissue-specific genes. In contrast, ETS, NRF-1, and CpG motifs are predominantly found in promoters of housekeeping genes [24, 30]. Another motif, Clus 1: TCTCGCGA, an unknown TFBS which Fitz *et al.*, also pointed out has a strong relationship with promoters of tissue-specific genes, though not clustered in VTB2, still has very significant occurrence (102 times in total) compared with VTB1 (12 times in total).

On the other hand, Sp1 is the only motif observed in both clusters. The total number nevertheless is particularly large in the VTB2 cluster. Sp1, a 105 -kDa transcription activator, usually binds with high affinity to promoters that contain CG rich regions (vertebrate class in this case). It is one of the most frequently found transcription activators required for the expression of a large number of regulated genes, and is effectively conserved in the promoters of most housekeeping genes [31–34]. The absence of the signature TATA box, the existence of GC rich regions, and multiple Sp1 binding sites are characteristics of housekeeping genes [7, 35, 36]. All of these characteristics are found in the VTB2 cluster. Comparison of the clustering result with some known housekeeping genes [37, 38] also showed positive correlations; an example of this is the ribosomal gene, a typical housekeeping

Table I. Comparison of TFBS motif occurrences within two vertebrate clusters. Start from leftmost column, the table shows the position of n -th DNA fragment which starts from -1000 to -901 , its motif, and the total number of times the motif occurs

Cluster VTBI					
Sp 1					
8	CCCGCCCC	91	8	CCCCTCCC	73
8	CCCCGCCC	97	9	CCCGCCCC	154
9	GCCCCGCC	143	9	CGCCCCGC	93
9	CCGCCCCC	76	9	CCCCGCCC	146
9	CCGCCCCG	86	9	CCGCCGCC	41
10	GCCGCCGC	85	10	CCGCCGCC	86
TATA					
9	CTATAAAA	48	9	GTATAAAA	54
9	TATAAAAG	71			
Cluster VTBI2					
Sp 1					
6	CCCGCCCC	78	7	CCCGCCCC	117
7	CCGCCCCC	69	7	GCCCCGCC	101
7	CCCCGCCC	141	8	CCCGCCCC	277
8	CCGCCCCC	151	8	CCGCCCCCT	98
8	GCCCCGCC	205	8	CGCCCCGC	109
8	CCGCCCCG	105	8	CCCCGCCT	72
8	GGCCCCGC	108	8	CCCCTCCC	100
8	CGCCCCCT	75	8	CGCCCCCG	74
8	CCCCGCCC	255	9	CCCGCCCC	361
9	CCCCGCCCT	70	9	CCGCCCCC	192
9	GCCGCCGC	134	9	CCGCCCCCT	141
9	CGCCCCGC	159	9	CCCCGCCT	103
9	CCGCCCCG	157	9	CGCCGCCC	65
9	CCGCCGCC	138	9	TCCGCCCC	75
9	CGCGCCGC	78	9	CCCCGCCC	373
9	CGCCGCGC	79	9	CGCCGCCG	127
9	GCGCCGCC	79	10	GCCGCCGC	226
10	CCGCCGCT	59	10	CGCCGCCA	94
10	CCGCCGCC	241	10	CGCCGCCG	156
11	CCCGCCCC	72	11	CCCGCCGC	69
11	GCCGCCGC	104	11	CCGCCGCC	106
12	CCCGCCCC	87	12	GCCGCCGC	79
12	GCCCCGCC	64	12	CCGCCCCC	67
12	CCCGCCCC	73	12	CCCGCCCC	62

(Continued on next page)

Table I. (Continued)

ETS					
9	GCCGGAAG	89	9	CCC GGAAG	62
9	CCGGAAGC	83	9	CGGAAGTG	81
9	CCGGAAGT	99			
NRF-1					
9	CGCCTGCG	90	9	GCCTGCGC	94
9	GCGCCTGC	83			
CpG					
9	GCGCGCGC	56			

gene which is essential for general cell function. Examining both clusters there are 84 entries from the ribosomal related gene in cluster VTB2, while only 23 entries are observed in cluster VTB1.

We attempt to classify the data further by increasing temperature parameter T until it reaches 0.11 (Figure 3 shows the hierarchical organization of cluster break-ups). This is when the number of unclassified points has gone above 38% and the system has reached the paramagnetic region. Beyond this temperature all clusters simply disintegrate, because strong correlations that characterize members of individual clusters have disappeared. Since each spin does random flips, and we have small spin-spin correlation functions with neighbors ($G_{ij} \approx 1/q$), hence

Figure 3. Dendrogram of clusters based on simulation with different temperature parameter.

creating many solitary clusters. Between temperature 0.11 and 0.12 the number of unclassified data in main arthropod cluster (ATP1) increases drastically, causing this cluster to break up rapidly from 1 into 19 clusters within a short temperature range. This phenomenon shows that the arthropod cluster only has one phase transition from ferromagnetic to paramagnetic and has no super-paramagnetic region. It merges naturally as a single large cluster and therefore quickly disintegrates into many small grains after reaching a critical temperature. Domany *et al.*, encounter a similar phenomenon when applying the SPC method to two simulated datasets generated by uniform and Gaussian distributions. SPC demonstrates its superiority by not imposing a partition on the data when there are no natural classes present [16].

Details of Clustering

CLUSTER HSV-1

With size 29 this cluster is completely formed by DNA sequences from human viral genome HSV-1, the pathogen that causes cold sore disease. The genes consist of sequences which are involved in the lytic cycle of HSV-1 infection from immediate early (alpha), delayed early (beta), intermediate late (beta/gamma), and late phases (gamma).

CLUSTER EBV

This cluster comprises of 22 members from EBV viral genes. The EBV virus, which also falls under the herpesvirus family, has two types of genome that express proteins in either the productive lytic or the latent cycle of infection. During latent infection only a few viral genes are expressed [39]. As a result, the number of genes established in latent infection are much smaller (3 genes) than those in lytic phases (19 genes).

CLUSTER VTB1

This cluster consists of DNA sequences from different species such as *Gallus gallus* (chicken), *Mus musculus* (house mouse), *Rattus norvegicus* (rat), and *Homo sapiens* (human). The sequences can be categorized into: (a) small nuclear RNA, (b) structural protein, (c) storage, transport proteins and apoproteins, (d) enzymes, (e) hormones, growth factors, regulatory proteins, (f) proteins related to stress, antibody or pathogen defense, (g) all transposable elements and retroviruses, especially for long terminal repeats, (h) all human viral genes from adenoviruses, and (i) a small number of unclassified chromosomal genes with various functions. The rest ($\approx 4\%$) are chromosomal genes regulating synthesis of structural proteins, storage and transport proteins, enzymes, antibodies, and hormones from misclassified arthropods.

CLUSTER VTB2

This cluster has many genes involved in basic functions needed for the sustenance of the cell. About 99% of the genes come from human. The sequences are dominated by chromosomal genes of unclassified function, which cover a total of 1293 entries or approximately 92% of the cluster density. Some of their functions which are constitutively expressed include those associated with ribosomal proteins, apolipoproteins, molecules transport, molecules binding, different families of globins, inhibitors, protease, and zinc-finger proteins.

CLUSTER ATP1

This is the main cluster of the arthropod class, with almost 100% of the cluster's members formed by DNA sequences from fruit fly (*Drosophila*) families. It has similar structure composition in gene expression function as VTB1 and VTB2 such as structural proteins, storage and transport protein, enzymes, hormones and growth factors, and chromosomal genes of unclassified function. The members have combinations of a huge number of genes for general metabolism and a fair amount of genes for specific functionality.

CLUSTER ATP2

This cluster contains mostly DNA sequences from the fruit fly (*Drosophila melanogaster*). It has 110 sequences of unclassified function, with diverse functionality such as those responsible for synthesis of histone, protein kinase, serine/threonine, and other proteins. Others include genes regulating oxidoreductase in both adult and larva forms of the fruit fly; this is a type of alcohol dehydrogenase protein useful for catalysis of oxidation-reduction. These genes are absent in the main arthropod cluster and explain the reason behind ATP2 separation from the ATP1 cluster.

Summary

We performed data clustering on a set of DNA sequences with active promoter regions. The first step was converting long nucleotide sequences into numerical values by evaluating all combinations of tri and tetra-nucleotide frequencies. The features of measurement were used to calculate a dissimilarity matrix which served as input data for the clustering algorithm. In this work, we employed the SPC method which is inspired by a statistical physics model of a disordered ferromagnet. With SPC, we were able to mine some important clusters and capture correlations lying inside the clusters. Besides successfully separating arthropod and vertebrate classes, we found two human viral genome clusters: EBV and HSV-1. Their members were gene sequences which expressed proteins in lytic and latent cycles of infection.

Another important result was the separation of the vertebrate class into two large clusters. We deduced that these two clusters and dominated by housekeeping and tissue-specific genes, respectively, by conducting a rigorous analysis of consensus octa-nucleotides of transcription factor binding sites.

Acknowledgements

We thank Ung Choon Yong for fruitful discussions and for giving a biological perspective to the result. This work was supported in part by Faculty Research Grants of the National University of Singapore. Detailed data file and classification results are available upon request.

References

1. Butler, J.E.F. and Kadonaga, J.T.: The RNA Polymerase II Core Promoter: A Key Component in The Regulation of Gene Expression, *Genes and Development* **16** (2002), 2583–2592.
2. Burden, S., Lin, Y.X. and Zhang, R.: Improving Promoter Prediction for the NNPP2.2 Algorithm: A Case Study Using E-Coli DNA Sequences, *Bioinformatics* **21** (2004), 601–607.
3. Chou, K.C.: Prediction of Protein Signal Sequences and Their Cleavage Sites, *Proteins: Structure, Function, and Genetics* **42** (2001), 136–139.
4. Davuluri, R.V., Suzuki, Y., Sugano, S. and Zhang, M.Q.: CART Classification of Human 5' UTR Sequences, *Genome Research* **10** (2000), 1807–1816.
5. Ma, X.T., Qian, M.P. and Tang, H.X.: Predicting Polymerase II Core Promoters by Cooperating Transcription Binding Sites in Eukaryotic Genes, *Acta Biochimica et Biophysica Sinica* **36** (2004), 250–258.
6. Rhee, J.W., Yeo, S.S., Na, K.R. and Kim, S.K.: New Training Method for Computational Identification of Promoter, *Genome Informatics* **13** (2002), 293–294.
7. Trinklein, N.D., Aldred, S.J.F., Saldanha, A.J. and Myers, R.M.: Identification and Functional Analysis of Human Transcriptional Promoters, *Genome Research* **13** (2003), 308–312.
8. Zhu, Z., Pilpel, Y. and Church, G.M.: Computational Identification of Transcription Factor Binding Sites via a Transcription-Factor-Centric Clustering (TFCC) Algorithm, *J. Molecular Biology* **318** (2002), 71–81.
9. Schmid, C.D., Praz, V., Delorenzi, M., Périer, R. and Bucher, P.: The Eukaryotic Promoter Database EPD: The Impact of in Silico Primer Extension, *Nucleic Acids Res.* **32** (2004), D82–D85.
10. Domany, E.: Cluster Analysis of Gene Expression Data, *J. Statistical Physics* **110** (2003), 1117–1139.
11. Getz, G., Levine, E., Domany, E. and Zhang, M.Q.: Super-paramagnetic Clustering of Yeast Gene Expression Profiles, *Physica A* **279** (2000), 457–464.
12. Getz, G., Levine, E. and Domany, E.: Coupled Two-Way Clustering Analysis of Gene Microarray Data, *PNAS* **97** (2000), 12079–12084.
13. Getz, G., Gal, H., Notterman, D.A. and Domany, E.: Coupled Two-Way Clustering Analysis of Breast Cancer and Colon Cancer Gene Expression Data, *Bioinformatics* **19** (2003), 1079–1089.
14. Blatt, M., Wiseman, S. and Domany, E.: Super-Paramagnetic Clustering of Data, *Physical Review Letters* **76** (1996), 3251–3255.

15. Blatt, M., Wiseman, S. and Domany, E.: Clustering Data through an Analogy to the Potts Model, *Advances in Neural Information Processing System* **8** (1996), 416–422.
16. Blatt, M., Wiseman, S. and Domany, E.: Data Clustering Using a Model Granular Magnet, *Neural Computation* **9** (1997), 1805–1842.
17. Blatt, M., Wiseman, S. and Domany, E.: Super-paramagnetic Clustering of Data, *Physical Review E* **57** (1998), 3767–3783.
18. Swendsen, R.H. and Wang, J.-S.: Nonuniversal Critical Dynamics in Monte Carlo Simulation, *Physical Review Letters* **58** (1987), 86–88.
19. Wang, J.-S. and Swendsen, R.H.: Cluster Monte Carlo Algorithms, *Physica A* **167** (1990), 565–579.
20. Bairoch, A., Apweiler, R., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M.J., Natale D.A., O'Donovan, C., Redaschi, N. and Yeh, L.S.: The Universal Protein Resource (UniProt), *Nucleic Acids Res.*, **33** (2005), D154–D159.
21. Abe, T., Kanaya, S., Kinouchi, M., Ichiba, Y., Kozuki, T. and Ikemura, T.: Informatics for Unveiling Hidden Genome Signatures, *Genome Research* **13** (2003), 693–702.
22. Faisst, S. and Meyer, S.: Compilation of Vertebrate-Encoded Transcription Factors, *Nucleic Acids Res.* **20** (1992), 3–26.
23. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Matys, V., Michael, H., Ohnhäuser, R., Prüß, M., Schacherer, F., Thiele, S. and Urbach, S.: The TRANSFAC System on Gene Expression Regulation, *Nucleic Acids Res.* **29** (2001), 281–283.
24. FitzGerald, P.C., Shlyakhenko, A., Mir, A.A. and Vinson, C.: Clustering of DNA Sequences in Human Promoters, *Genome Research* **14** (2004), 1562–1574.
25. Mariño-Ramírez, L., Spouge, J.L., Kanga, G.C. and Landsman, D.: Statistical Analysis of Over-Represented Words on Human Promoter Sequences, *Nucleic Acids Res.* **32** (2004), 949–958.
26. Berg, J.M., Tymoczko, J.L. and Stryer, L.: *Biochemistry*, 5th ed. W.H. Freeman and Company, New York 2002.
27. Latchman, D.S.: *Eukaryotic Transcription Factors*, 4th eds. Academic Press, London, 2004.
28. Weis, L. and Reinberg, D.: Transcription by RNA Polymerase II Initiator Directed Formation of Transcription Competent Complexes, *FASEB Journal* **6** (1992), 3300–3309.
29. Murakami, K., Kojima, T. and Sakaki, Y.: Detection of Tissue Specific Genes by Putative Regulatory Motifs in Human Promoter Sequences, *Genome Informatics* **14** (2003), 408–409.
30. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H.: CpG Islands as Gene Markers in The Human Genome, *Genomics* **13** (1992), 1095–1107.
31. Courey, A. J. and Tjian, R.: Analysis of Sp1 *in vivo* Reveals Multiple Transcriptional Domains, Including a Novel Glutamine-Rich Activation Motif, *Cell* **55** (1988), 887–898.
32. Dynan, W.S. and Tjian, R.: The Promoter-Specific Transcription Factor Sp1 Binds to Upstream Sequences in The SV40 Early Promoter, *Cell* **35** (1983), 79–87.
33. Kadonaga, J.T., Carner, K.R., Masiarz, F.R. and Tjian, R.: Isolation of cDNA Encoding Transcription Factor Sp1 and Functional Analysis of The DNA Binding Domain, *Cell* **51** (1987), 1079–1090.
34. Philipsen, S. and Suske, G.: A Tale of Three Fingers: The Family of Mammalian Sp/XKLF Transcription Factors, *Nucleic Acids Res.* **27** (1999), 2991–3000.
35. Nakamura, Y., Miura, K., Fujino, Y., Iwao, H., Ogita, S. and Yamanaka, S.: Evolution, Structure, and Expression of GNPI/Oscillin Orthologous Genes, *Genomics* **68** (2000), 179–186.
36. Yamabe, Y., Shimamoto, A., Goto, M., Yokota, J., Sugawara, M. and Furuichi, Y.: Sp1-Mediated Transcription of the Werner Helicase Gene Is Modulated by Rb and p53, *Molecular and Cellular Biology* **18** (1998), 6191–6200.
37. Eisenberg, E. and Levanon, E.Y.: Human Housekeeping Genes are Compact, *Trends in Genetics* **19** (2003), 362–365.

38. Hsiao, L.-L., Dangond, F., Yoshida, T., et al.: A Compendium of Gene Expression in Normal Human Tissues, *Physiol. Genomics* **7** (2001), 97–104.
39. Laux, G., Perricaudet, M. and Farrell, P.J.: A Spliced Epstein-Barr Virus Gene Expressed in Immortalized Lymphocytes is Created by Circularization of The Linear Viral Genome, *EMBO Journal* **7** (1988), 769–774.