Characterization of Eukaryotic Core Promoters Based on Nonlinear Dimensionality Reduction

Xi Yang and Hong Yan

***Abstract*—Characterization and identification of eukaryotic promoter is important for the gene prediction and genome annotation. In this paper, we study the structural characteristics of the core promoters in several eukaryotes through a series of DNA physicochemical properties and adopt a method that combines the alignment and average of multiple promoters and the nonlinear dimensionality reduction technique. The result shows that the eukaryotic core promoters have very special structural characteristics that are coherent between different species and independent of their sequence compositions.**

***Index Terms*— DNA physicochemical properties, eukaryotic promoters, Isomap, structural profiles**

I. Introduction

A promoter is a region of a genomic DNA sequence, which is located near a gene and contains critical elements to control the transcription regulation of the gene. The binding of these response elements with RNA polymerase and transcription factors to initiate the gene transcription constitutes the foundation of gene expression and repression mechanism, and it has been proved that the malfunction of promoter is closely related to quite a few diseases [1]-[3]. Promoter prediction is one of the tasks of genome annotation. Computational methods of promoter prediction mostly rely on the conserved *cis*-acting sequence motifs, such as TATA-boxes, CpG islands and CAAT boxes. However, on the whole, the effectiveness of these prediction methods based purely on the sequence features is quite limited or only workable for some specific groups of promoters. This is because eukaryotic promoters are extremely diverse and so far no ubiquitous sequence pattern that makes sense in all eukaryotic promoters has yet been found. For example, TATA-box, a well-known signature of promoter, is predicted to only appear in a maximum of 20% of mammalian promoters [4],[5]. The enrichment of CpG islands serves as a mark of nucleosome-depletion region and thus a mark of promoter for many vertebrate eukaryotes, but gives little help in characterizing promoters in non-vertebrate eukaryotes [6]. A recent point argued that besides carrying the sequence composition information, the linear DNA molecule also has very distinct physicochemical properties that decide to a large extent its topological structure and binding affinity to RNA polymerase and other protein factors during the formation of transcriptional complex [7],[8]. Over the past twenty years, abundant experiments have been carried out to measure the physicochemical properties of DNA molecules under different conditions, based on which people have summarized a set of empirical physicochemical parameters for various short DNA segments. These parameters have been taken into account in recent studies for the description of promoter, coding and non-coding regions in the genome of specific species, including yeast, Arabidopsis, rice, Plasmodium falciparum, mouse, human, etc [9]-[11]. It was preliminarily observed that irrespective of gene types or species, the regions around transcription start site (TSS) and core promoter (CP) indeed exhibit special structural settings compared with the non-promoter regions [10],[12],[13]. The DNA structural profiles described by these physicochemical parameters are also being used more and more widely in the promoter prediction computation [14],[15], and the prediction performance can be improved by this means.

However, there are also two debates about the use of these structural properties. One is that since the properties are sequence-dependent, it is not clear whether the information in the structural profiles has already been encompassed in the sequence composition or reveals some new aspects of a DNA segment [16]. The other is that the structural profiles of a single DNA sequence are very noisy because they are converted from the original sequence at a dinucleotides or trinucleotides. The noise conceals a lot of useful information in the profiles, while the smoothing of the profiles does reduce the noise but will inevitably lead to the loss of local details [15],[17].

In this paper, we investigate and compare the features in the structural profiles of core promoter regions in several typical eukaryotes. Instead of using a sliding window of specified width to filter noise in the structural profiles of each individual promoter, we align promoters at the TSS for each eukaryote type and get an averaged promoter representative for this eukaryote type. Then we apply a nonlinear dimensionality reduction algorithm – Isomap on the averaged promoter model, which is described by a set of physicochemical parameters, to separate a comprehensive structural profile. The structural profile derived by our method is very different from those in previous studies. Firstly, the avoidance of the sliding window approach can preserve the local details of each single promoter, while the average between individual promoters weakens the local inconsistent structural traits and strengthens the consistent

Manuscript received December 31, 2011; revised January 15, 2012. This work was supported by the Hong Kong Research Grant Council (Project CityU 123809) and City University of Hong Kong (Project 7008094).

X. Yang is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (phone: 852-2784-4263; e-mail: yangxi_anne@ yahoo.com.cn).

H. Yan is with the Department of Electronic Engineering, City University of Hong Kong, Hong Kong (e-mail: h.yan@ cityu.edu.hk).

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

traits. Secondly, the correlation between the original physicochemical parameters is captured in the first principal dimensionality. The structural profile under this collective dimensionality can reflect the coordinated structural variation defined by multiple physicochemical properties rather than a single one. The result shows that eukaryotic core promoters have very special structural characteristics, which are independent of their DNA sequence content.

II. MATERIALS AND METHODS

A. Core-promoter datasets

We choose the promoters of rice, fruit fly, chicken, mouse, rat and human from Eukaryote Promoter Database (EPD), the numbers of which are 13046, 1926, 72, 196, 119 and 1871 respectively. The reason why we choose EPD is that the promoter sequences in EPD are all experimentally determined, so they can reflect the true nature of promoters. We extract [-100, +50] relative to the TSS from each sequence as the core promoter.

B. Physicochemical properties of DNA

We use fourteen most commonly used ditrinucleotide and trinucleotide physicochemical properties to describe the core promoters (Table 1) [10],[14],[15]. Each promoter sequence is converted to a string of numerical values based on these ditrinucleotide and trinucleotides properties and thus is described by a 151 by 14 matrix. Within each group, a varying number of promoters (N=5, 10, 20, 30, 50) are randomly selected and aligned at transcription start site (TSS), and then averaged. By this means, we obtain an averaged CP representative for each eukaryote type. Then, we used the nonlinear dimensionality reduction algorithm-Isomap to extract a principal structural feature from the averaged CP.

TABLE 1
DNA PHYSICOCHEMICAL PROPERTIES

Num.	Physical property	Min	Max
1	A-philicity	0.13	1.04
2	B-DNA twist	30.6°	43.2°
3	DNA bendability	-0.280	+0.194
4	DNA-bending stiffness	20 nm	130 nm
5	DNA denaturation	64.35 cal/mol	135.38 cal/mol
6	Duplex disrupt energy	0.9 kcal	3.1 kcal
7	Duplex free energy	-2.1 kcal/mol	-0.9 kcal/mol
8	GC content	0	3
9	Nucleosome positioing	-36%	+45%
10	Propeller twist	-18.66°	-8.11°
11	Protein-DNA twist	31.5°	37.8°
12	Protein-induced deformability	1.6	12.1
13	Stacking energy	-14.59 kcal	-3.82 kcal
14	Z-DNA stabilizing energy	5.9 kcal/mol	0.7 kcal/mol

C. Isometric feature mapping (Isomap)

Isomap is an extension of classical multidimensional scaling (MDS) by substituting the straight-line Euclidean distance with the geodesic distances to measure the pairwise distance between data points. It is one of the most widely used algorithms in the manifold learning field. The goal of manifold learning is to find the underlying low-dimensional manifold that the sample points in the high-dimensional space actually lie on and construct an approximation to the true geometry of the data manifold. The Isomap algorithm achieves this goal in three steps [18]:

1) Build a neighborhood graph. This step determines which points are neighbors on the manifold M. If the distance $d_X(i, j)$ between two points i, j in the input space X satisfies the criteria of K-nearest neighbors or ε-radius, they are regarded as neighbors. A weighted graph G over all the data points is defined by this means.
2) Calculate shortest paths. The shortest path distances $d_G(i, j)$ in the graph G defined above are calculated and used to approximate the true geodesic distances $d_M(i, j)$ between all pairs. This can be done by various graph analysis algorithms. Floyd's algorithm, for example, iteratively improves the estimate on the shortest paths by comparing all the possible paths between all point pairs through the graph, until the optimal value is obtained. Graph G is firstly initialized by

$$d_G(i, j) = \begin{cases} d_X(i, j) & \text{i,j are linked by an edge} \\ \infty & \text{otherwise} \end{cases}$$

Then, for each $k = 1, 2, \mathrm{K}$, N in turn, $d_G(i, j)$ is replaced by the minimal value in $\{d_G(i, j),\ d_G(i, k) + d_G(k, j)\}$. Finally, the collection $D_G = \{d_G(i, j)\}$ represents the shortest paths between all point pairs in graph G.
3) Construct d-dimensional embedding. The classical MDS is applied to construct an embedding of geodesic distance data $D_G = \{d_G(i, j)\}$ into a d-dimensional Euclidean space Y. The vectors y_i in Y are those that can minimize the cost function

$$E = \left\| \tau(D_G) - \tau(D_Y) \right\|_{L^2}$$

in which $\| \ \|_{L^2}$ is the L^2 matrix norm and τ is the inner product operator. The number of dimensionality d is decided by the top d eigenvectors in the matrix $\tau(D_G)$.

Here, Isomap is used to reduce the dimensionality of the averaged CP vectors so that each averaged CP will be described by the underlying principal dimensionality of the 151 sample points.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

III. RESULTS AND DISCUSSION

A. Physicochemical property profiles of CP

For each single core promoter sequence (N=1), its structural profiles in terms of the fourteen original physicochemical properties are quite rough and of high noise caused by direct coding of the DNA sequence (Figure1). By applying the nonlinear dimensionality reduction technique, we obtain a low-dimensional representation of the CP structure. The top three principal dimensionalities account for 98.5% variances, and the first principal dimensionality alone represents 92.2% variances. So we use the first principal dimensionality instead of the fourteen original physicochemical properties to characterize the nature of the CP. The structural variation of the CP under the first principal dimensionality is marked as $Y(n)$, n=-100,…+50. However, $Y(n)$ is also very noisy (Figure1). One reason is the high noise in the fourteen original physicochemical property profiles, and the other is the very limited number of data points (only 151) involved in the dimensionality reduction calculation.

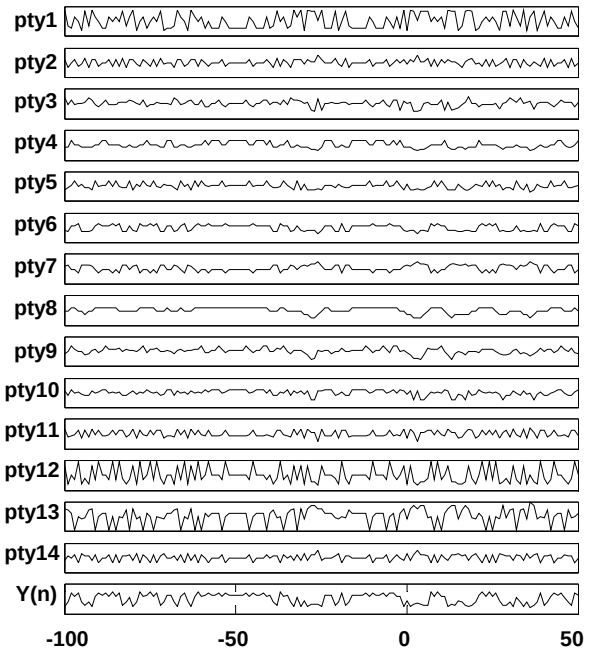

Fig. 1. Profiles of fourteen physicochemical properties and *Y(n)* for a human CP (EPD entry: EP07056). Refer to Table 1 for the fourteen physicochemical properties (pty1-14).

B. Structural feature found in the averaged CP

Since the structural information in single CP is too obscure, we turn to the investigation of the common feature of certain CP group. We randomly select different numbers of promoters (N=5, 10, 20, 30, 50) within each eukaryote type, align them at transcription start site (TSS) and obtain fourteen averaged physicochemical property profiles. Isomap is then used to extract the first principal dimensionality from the averaged CP. To distinguish the structural signal obtained by this means from that extracted from a single CP, we mark it as $Y_{N\_ave}(n)$, n=-100,…+50. It can be found that a significant feature located around [-40, -20] becomes more and more evident as the number of promoters involved increases and this rule makes sense for all the eukaryotes in our study (Figure2). Therefore, it is concluded that the "valley" in the [-40, -20] region of the structural profile is a mark of the eukaryotic core promoters. This hallmark signal is the result of accumulative effect, because the original signal in each single CP is too weak to be observed. The average of a group of promoters makes the local inconsistent structural traits cancel each other out while intensifies the local consistent traits.

C. The influence of TATA-box

The [-40, -20] segment relative to TSS is also the area in which TATA-box often appears, so we investigate if the hallmark in the structural profile revealed above is practically caused by the TATA-box. The pattern of TATA-box follows the human promoter elements definition given by Narang *et al.*, which includes eighteen hexanucleotide types altogether [19].The human promoters from EPD are scanned and classified into TATA-box group and no-TATA-box group. The scanning is done by the online server GPMiner [20]. We then select fifty CPs (N=50) from the two groups respectively and calculate the $Y_{50\_ave}(n)$ for both CP collections. The result shows that the same feature appears in the structural profiles of both groups (Figure 3), implying that the structural or physical characteristic of the core promoter is intrinsic and independent of TATA-box. In other words, the upstream region in the immediate vicinity of TSS has very special texture, which is not influenced by the sequence composition of the core promoter. On the other hand, there is another remarkable feature appearing at the TSS in the structural profile for the TATA-box-free group. Based on this phenomenon, we speculate that if TATA-box does not exist at the upstream region very near to TSS, the TSS itself will take special structural or physical settings to ensure the precise transcriptional initiation.

IV. CONCLUSION

The eukaryotic core promoters have very special structural characteristics that are coherent between different species and independent of their sequence compositions. Average over certain number of promoters can accumulate the local consistent physicochemical traits, while the nonlinear dimensionality reduction technique has been proved to be an effective method to extract the hallmark structural signal from the averaged core promoter models. The hallmark located at the region of [-40, -20] is to a great extent the intrinsic nature for eukaryotic core promoter rather than a sign for the TATA-box, besides the feature in the region of [-40, -20], the TSS is also observed to have an extreme value in the structural profile.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

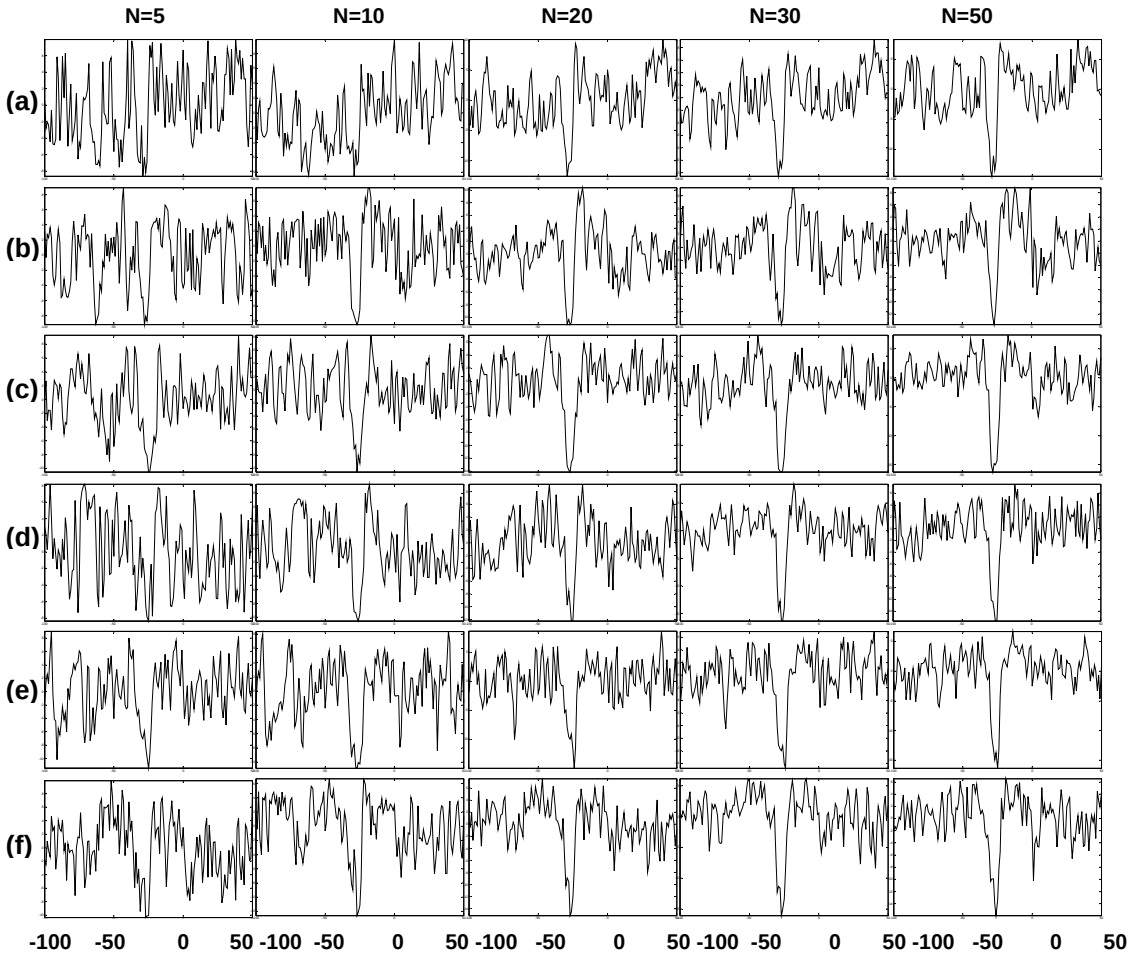

Fig. 2. The profiles of $Y_{N\_ave}(n)$ (N=5, 10, 20, 30, 50) for (a) rice, (b) fruitfly, (c) chicken, (d) mouse, (e) rat and (f) human

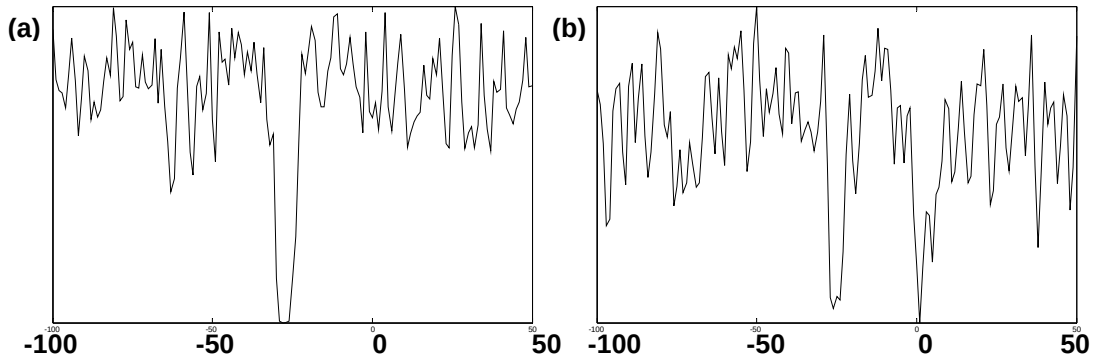

Fig. 3. The profiles of $Y_{N\_ave}(n)$ (N=50) for two human promoter groups: (a) TATA-box group and (b) no-TATA-box group

REFERENCES

[1] A. Carmine, S. Buervenich, D. Galter, E. G. Jönsson, G. C. Sedvall, L. Farde, J. P. Gustavsson, H. Bergman, K. V. Chowdari, V. L. Nimgaonkar, M. Anvret, O. Sydow, and L. Olson, "NURR1 promoter polymorphisms: Parkinson's disease, schizophrenia, and personality traits," *Am. J, Med. Genet. B Neuropsychiatr. Genet.*, vol. 120B, no. 1, pp. 51-57, Jul. 2003.

[2] M. Tanaka, P. Chang, Y. Li, D. Li, M. Overman, D. M. Maru, S. Sethi, J. Phillips, G. L. Bland, J. L. Abbruzzese, and C. Eng, "Association of CHFR promoter methylation with disease recurrence in locally advanced colon cancer," *Clin. Cancer Res.*, vol. 17, no. 13, pp. 4531-4540, Jul. 2011.

[3] K. Kingo, S. Kõks, H. Silm, and E. Vasar, "IL-10 promoter polymorphisms influence disease severity and course in psoriasis," *Genes Immun.*, vol. 4, no. 6, pp. 455-457, Sep. 2003.

[4] S. J. Cooper, N. D. Trinklein, E. D. Anton, L. Nguyen, and R. M. Myers, "Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome," *Genome Res.*, vol. 16, no. 1, pp. 1-10, Jan. 2006.

[5] N. I. Gershenzou, and I. P. Ioshikhes, "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis," *Bioinformatics*, vol. 21, no. 8, pp. 1295-1300, Apr. 2005.

[6] Y. Y. Yamamoto, T. Yoshitsugu, T. Sakurai, M. Seki, K. Shinozaki, and J. Obokata, "Heterogeneity of Arabidopsis core promoters revealed by high-density TSS analysis," *Plant J.*, vol. 60, no. 2, pp. 350-362, Jun. 2009.

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

[7] A. G. Pedersen, P. Baldi, Y. Chauvin, and S. Brunak, "DNA structure in human RNA polymerase II promoters," *J. Mol. Biol.*, vol. 281, no. 4, pp. 663-673, Aug. 1998.

[8] J. Zeng, X. Q. Cao, H. Zhao, and H. Yan, "Finding human promoter groups based on DNA physical properties," *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, vol. 80, no. 4 Pt 1, pp. 041917, Oct. 2008.

[9] X. Q. Cao, J. Zeng, and H. Yan, "Structural properties of replication origins in yeast DNA sequences," *Phys Biol.*, vol. 5, no. 3, pp. 036012, Sep. 2008.

[10] C. Morey, S. Mookherjee, G. Rajasekaran, and M. Bansal, "DNA free energy-based promoter prediction and comparative analysis of Arabidopsis and rice genomes," *Plant Physiol.*, vol. 156, no. 3, pp. 1300-1315, Jul. 2011.

[11] P. Meysman, T. H. Dang, K. Laukens, R. De Smet, Y. Wu, K. Marchal, and K. Engelen, "Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli.," *Nucleic Acids Res.*, vol. 39, no. 2, pp. 1-11, Jan. 2011.

[12] K. Florquin, Y. Saeys, S. Degroeve, P. Rouzé, and Y. Van de Peer, "Large-scale structural analysis of the core promoter in mammalian and plant genomes," *Nucleic Acids Res.*, vol. 33, no. 13, pp. 4255-4264, Jul. 2005.

[13] U. Ohler, H. Niemann, G. Liao, and G. M. Rubin, "Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition," *Bioinformatics*, vol.17, Suppl. 1, pp. S199-S206, Apr. 2001.

[14] K. Brick, J. Watanabe, and E. Pizzi, "Core promoters are predicted by their distinct physicochemical properties in the genome of Plasmodium falciparum," *Genome Biol.*, vol. 9, no. 12, pp. R178, Dec. 2008.

[15] T. Abeel, Y. Saeys, E. Bonnet, P. Rouzé, and Y. Van de Peer, "Generic eukaryotic core promoter prediction using structural features of DNA," *Genome Res.*, vol. 18, no. 2, pp. 310-323, Dec. 2008.

[16] P. Baldi, Y. Chauvin, S. Brunak, J. Gorodkin, and A. G. Pedersen, "Computational applications of DNA structural scales," *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 6, pp. 35-42, 1998.

[17] L. Kelbauskas, J. Yodh, N. Woodbury, and D. Lohr, "Intrinsic promoter nucleosome stability/dynamics variations support a novel targeting mechanism," *Biochemistry*, vol. 48, no. 20, pp. 4217-4219, May 2009.

[18] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319-2323, Dec. 2000.

[19] V. Narang, W. K. Sung, and A. Mittal, "Computational modeling of oligonucleotides positional densities for human promoter prediction," *Artf. Intell. Med.*, vol. 35, no. 1-2, pp. 107-119, Sep.-Oct. 2005.

[20] T. Lee, W. Chang, J. B. Hsu, T. Chang, and D. Shien, "GPMiner: an integrated system for mining combinatorial cis-regulatory elements in mammalian gene group," *BMC Genet.*, vol. 13(Suppl 1), no. S3, pp. 1-12. Available:http://gpminer.mbc.nctu.edu.tw/index.php

ISBN: 978-988-19251-1-4
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)