Graphical Approach to Weak Motif Recognition in Noisy Data Sets

Loi Sy Ho[1] and Jagath C. Rajapakse[1,2,3]

[1] BioInformatics Research Center, School of Computer Engineering
Nanyang Technological University, Singapore 639798
{slho, asjagath}@ntu.edu.sg
[2] Biological Engineering Division
Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[3] Singapore-MIT Alliance, N2-B2C-15, 50 Nanyang Avenue, Singapore 639798

Abstract. Accurate recognition of motifs in biological sequences has become a central problem in computational biology. Though previous approaches have shown reasonable performances in detecting motifs having clear consensus, they are inapplicable to the recognition of weak motifs in noisy datasets, where only a fraction of the sequences may contain motif instances. This paper presents a graphical approach to deal with the real biological sequences, which are noisy in nature, and find potential weak motifs in the higher eukaryotic datasets. We examine our approach on synthetic datasets embedded with the degenerate motifs and show that it outperforms the earlier techniques. Moreover, the present approach is able to find the wet-lab proven motifs and other unreported significant consensus in real biological datasets.

1 Introduction

The central dogma of molecular biology is that DNA produces RNA, which in turn produces protein. For the regulation of transcription, a set of proteins called transcription factors (TFs) bind to short subsequences in the promoter region and activate transcription machinery. Such subsequences are called transcription factor binding sites (TFBSs) that, since a TF can bind to several sites in the promoter regions of different genes, should have common patterns or motifs. A motif is defined as a representation of a set of subsequences, which are prevalent in a class of biological sequences and share a similar composition of symbols. For instance, the TATA box is a motif at the site of transcription initiation. Motifs such as Shine-Dalgarno sequences (also called Ribosome Binding Sites (RBSs)) are involved in the translational initiation and preserve in most promoter regions of prokaryotic genes. Identification of motifs in DNA sequences provides important clues for the understanding of the proteins, DNA-protein interactions and the gene regulatory networks.

Since not much knowledge is known about most TFs and the variability of their binding sites, the wet-lab experiments to locate related motifs in DNA sequences, such as DNAseI Footprinting Assay and Methylation Interference Assay [10], are both cumbersome and time consuming. Therefore, computational

J.C. Rajapakse, L. Wong, and R. Acharya (Eds.): PRIB 2006, LNBI 4146, pp. 23–31, 2006.

techniques and algorithms, providing efficient and low cost solutions, have been rapidly developed for motif recognition. Based on different assumptions used by these techniques and algorithms they are classified into either probabilistic or deterministic. Probabilistic approaches use a weight matrix to represent a motif and maximize the information content of the alignment of motif instances [1,2,6,11,13]. On the other hand, deterministic approaches exhaustively enumerate or search for motif consensus sequences [4,5,14,17]. Each approach has its own strength and weakness, depending on the task at hand, while a specific type of motif recognition approaches may be more useful than others [7,8,18].

It is observed that, for some TFs, the number of sequences that contain TFBSs with very similar pattern are insufficient to successfully find the motif using existing approaches [3]. Some motif consensus may exactly be present in datasets while others may exist with a small or significant number of de-generations. In practice, the noises are inevitable in datasets due to experimental errors, the failure to retrieve a suitable length of the regions containing the motifs, etc. The problem of weak motif recognition (WMR), that discovers a motif having a significant number of degenerations randomly distributed over its relatively short length, has recently been addressed. The graphical approaches, such as WINNOWER [14], cWINNOWER [12], and MITRA [4] convert the subsequences in the dataset into vertices and use the edges to indicate their relationships among possible instances; the random projection methods, such as PROJECTION [2], Multiprofiler [9], and Planted Motif Search [16], attempt to reduce the sample space by decreasing the motif length or the effective degenerate positions; the other approaches, such as SampleBranching [15] and SP-STAR [14] optimize a target function such as the pair-wise scoring function.

Despite such various attempts, it has been hard to develop an efficient algorithm to deal with the WMR problem. The difficulty is mainly due to two reasons: (1) the large pairwise distance between motif instances of two sequences evades their detection and an instance could be more similar to a random subsequence than to another motif instance, and (2) the time complexity of the detection increases and the accuracy decreases when *corrupted* sequences that do not contain any motif instance are present in the dataset. Therefore, the previous WMR approaches are quite time consuming and vulnerable to noises.

Earlier in [19], Yang and Rajapakse proposed an graphical algorithm (hereinafter known as GWM) with superior running time and performance that can find weak motifs in the datasets where each sequence contains at least one motif instance. However, the robust motif finding algorithm with capabilities of tolerating to a certain amount of noise in datasets is of practical significance. In this paper, we propose a GWM2 approach that extends the previous algorithm to find weak motifs in noisy datasets containing corrupted sequences. Our algorithm shows better robustness to noises and more accuracy than the earlier methods. Moreover, GWM2 is able to find the wet-lab proven motifs and other unreported significant consensus on the real biological datasets. Although the illustration of our method, in this paper, is limited to only DNA sequences, the method is generalizable to other biological sequences such as protein sequences.

2 Method

Suppose that we are interested in finding motifs in m DNA sequences given by the set $\mathbf{D} = \{\mathbf{x}_i : i = 1, 2, ..., m\}$ where the i th sequence $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ has length n_i. The elements of the sequences, $x_{ij} \in \Omega$ for every sequence $\mathbf{x}_i$ and its j th element are drawn from the alphabet of nucleotides $\Omega = \{A, T, G, C\}$. We use Ψ to represent the consensus of the motif that is derived from the alignment of a set of motif instances. Suppose that K is the number of sequences that contain motif instances. If $K = m$, the dataset is called an *exact* dataset, otherwise ($K < m$) a *noisy* dataset. Here, we present an approach to the latter case where each sequence $\mathbf{x}_i$ contains either one or zero motif instance. Let the motif be denoted as a pair (l, d) where l is the length of the motif and d is the maximum degenerate positions allowed to differ a motif instance from the consensus. We look for instances, $\psi_k,\ k = 1, ..., K$ that satisfy $dis(\Psi, \psi_k) \leq d$ where $dis(\cdot, \cdot)$ is a distance measure, say the Hamming distance, between the two subsequences. d can be set to large value, but no more than a threshold d', beyond which random motifs could be found in the same dataset. The d' is restricted by the inequality [2]:

$$4^l(1 - (1 - p)^{n-l+1})^m < 1 \tag{1}$$

where the left hand side gives the expected number of random (l, d') motif occurrences, $n = \max_{i=1}^{m} n_i$, and $p = \sum_{i=0}^{d'} \binom{l}{i} (\frac{3}{4})^i (\frac{1}{4})^{l-i}$ is the probability for two random subsequences having length l to be differed within d' positions.

In graphical representation of the dataset, each subsequence is represented at a vertex [14]. Let vertex v_{ij} represent the subsequence of length l starting at position j of the ith sequence, say $\mathbf{s}_{i,j} = (x_{ij}, x_{ij+1}..., x_{ij+l-1})$. Therefore, K motif instances in the dataset are assigned to certain vertices and are determined from a total of $\sum_{i=1}^{m}(n_i - l + 1)$ number of vertices. For a given (l, d) motif Ψ in the dataset, any two instances of Ψ differ at most $2d$ positions. If the graph is constructed so that any two vertices v_{ij} and v_{pq}, for $1 \leq i \neq p \leq m$, $1 \leq j \leq n_i$, and $1 \leq q \leq n_p$, are linked if $dis(\mathbf{s}_{i,j}, \mathbf{s}_{p,q}) \leq 2d$, the motif instances represented by vertices in the graph are connected to each other and form a clique of size K. Then, the motif recognition problem is equivalent to finding K-cliques in a given graph. Though clique finding in graphs is known as NP-complete problem, in the present context its complexity is significantly lower because of a small ratio of the numbers of edges to the number of vertices of graphs for datasets of nucleotide or amino acid sequences [8]. Our algorithm consists of three steps: graph construction, clique finding, and rescanning.

2.1 Graph Construction

Let a selected sequence $\boldsymbol{x}_r$, for $r = 1, \ldots m - K + 1$, be referred as *reference* sequence and suppose that the potential motif instance in the reference sequence is represented by the vertex $v_{r\rho}$ where ρ indicates its starting position. As we are looking for l-length motifs, for each position $\rho = 1, ..., n_r - l + 1$ in the reference sequence, we build a graph $G_\rho = (V_\rho, E_\rho)$ as follows:

1. Set $V_\rho = \{\rho\}$ and $E_\rho = \phi$.
2. For $i = r+1, \ldots, m$, find subsequence $\mathbf{s}_{i,j}$ represented by vertex v_{ij} where $j = 1, 2, \ldots n_i - l + 1$, and if $dis(\mathbf{s}_{r,\rho}, \mathbf{s}_{i,j}) \leq 2d$: $V_\rho = V_\rho \bigcup v_{ij}$.
3. For two different vertices v_{ij} and $v_{pq} \in V_\rho$, if $dis(\mathbf{s}_{i,j}, \mathbf{s}_{p,q}) \leq 2d$: $E_\rho = E_\rho \bigcup e_{v_{ij},v_{pq}}$. As sequence $\mathbf{x}_i$ is assumed to contain at most one motif instance, no edge $e_{v_{ij},v_{ij'}}$, where $j' = 1, 2, \ldots n_i - l + 1$, is added to E_ρ.
4. For each $v_{ij} \in V_\rho$, define a *triangle neighbor* set T_{ij}, which consists of elements p, $r+1 \leq p \leq m$, satisfying $v_{pq} \in V_\rho$ and $e_{v_{ij},v_{pq}} \in E_\rho$ with at least an index q: $1 \leq q \leq n_p$. Remove vertex v_{ij} from V_ρ and its corresponding edges from E_ρ if $|T_{ij}| < K - 2$. This triangle criteria is what Pevzner and Sze called the $k = 2$ case [14].

After constructing the graph G_ρ, if $v_{r\rho}$ represents a real motif instance in the reference sequence $\boldsymbol{x}_r$, the motif instances in other sequences should then be represented by the vertices in the same graph G_ρ. As such, the tenet of our approach is to convert the given dataset into a set of graphs G_ρ where $\rho = 1, ..., n_r - l + 1$, and look for cliques of size K such that each of the vertices in these cliques represents an actual motif instance.

2.2 Clique Finding

If the potential motif instance is represented by the vertex $v_{r\rho}$, the motif instances will be represented by a clique of K vertices in the graph G_ρ. In what follows, we present an iterative approach to search for K-cliques in the graph G_ρ.

1. We define the set $C_k(i,j)$, corresponding to $v_{ij} \in V_\rho$, indicate a set of all possible k-cliques containing k vertices starting from the vertex $v_{r\rho}$ to vertex v_{ij}. Set $C_1(r,\rho) = \{v_{r\rho}\}$.
2. The iterative computation for $C_k(i,j)$ is then:
   (a) Set $C_k(i,j) = \phi$.
   (b) For each $v_{pq} \in V_\rho$, where $r \leq p < i$
   and $e_{v_{ij},v_{pq}} \in E_\rho$, do
   For each *k-1*-clique $c \in C_{k-1}(p,q)$ do
   If $\{cli \bigcup v_{ij}\}$ is a valid k-clique then
   $C_k(i,j) = C_k(i,j) \bigcup \{c \bigcup v_{ij}\}$
   End If
   Repeat
   Repeat
3. By increasing k from 2 to K, if a clique of size K exists in the graph G_ρ, there must exist a non-empty set $C_k(i,j)$ for a vertex $v_{ij} \in V_\rho$ that contains vertices forming a K-clique.

2.3 Rescanning

After obtaining the cliques of size K, motif consensus Ψ could be formed by alignment of the instances corresponding to the vertices of each clique. As the lengths of sequences in the dataset become longer, spurious cliques could appear.

Therefore, an extra step is necessary to rescan the dataset with the motif consensus derived from the earlier steps and save those instances ψ_i satisfying the inequality $dis(\Psi, \psi_i) \leq d$. This guarantees that all the possible motif instances are found exactly in each sequence, including the spurious instances that are preserved as good as the real instances.

2.4 Algorithmic Complexity

For exact datasets where $K = m$, the motif recognition problem is efficiently solved by Yang and Rajapakse [19] in $O(nmA^2)$, where $A = n\sum_{i=0}^{2d}\binom{l}{i}(3/4)^i (1/4)^{l-i}$ is the random number instances of a motif (l, d) existing in a sequence with length n. The present approach GWM2 is a direct extension of our previous algorithm GWM for noise datasets, where $K \leq m$, hence requiring on the order of $\binom{K}{m}nkA^2$ computations. If in the graph G_ρ most vertices are spurious or unrelated and have been included in the $C_k(i, j)$ repeatedly, it could cost memory and time for maintaining such sets of cliques. However, as indicated in [14], when the size of cliques becomes larger, less spurious vertices are included; most $C_k(i, j)$ become mostly empty as k increases to K. Therefore, as will be shown in the next section, the running time of our approach in most cases in the experiments is reasonably small.

3 Experiments and Results

This section presents our experiments to evaluate the GWM2 approach on synthetic datasets and real biological datasets for TFBSs recognition, and compare its performance with the earlier methods. In case of real biological datasets, which are extracted from both prokaryotic and eukaryotic organisms, some sequences are exact while the others are noisy.

3.1 Synthetic Data

The techniques of motif recognition in our experiments were evaluated based on two standard performance measures defined as follows: performance coefficient (PC), $PC = |\psi \cap \hat{\psi}|/|\psi \cup \hat{\psi}|$, where ψ is the set of the known motif instances and $\hat{\psi}$ is the set of motif instances predicted [14], and success rate (SR) [15] is the ratio of the number of successes to the total number of trials. Because we use the consensus presentation for the motifs found, SR is used for evaluation of our algorithm.

Exact Data. The exact datasets are those used in [14]: there are 10 datasets, each of which consists a total of 20 DNA sequences of length 600 bp and generated with identical and independent nucleotides distributions. The results of the former approaches were referenced in [15,16].

Table 1 shows the performance measure and running time. It can be seen that the probabilistic approaches might perform faster than the GWM2 approach, but they could not guarantee to find precisely the embedded motifs. Compared with

Table 1. Comparison of the performance and running time by different approaches on the datasets used in the Challenge Problem [14] for finding $(l = 15, d = 4)$ motifs

Algorithm	SR	PC	Running time
GibbsDNA	-	0.32	40 s
ProfileBranching	-	0.57	80 s
CONSENSUS	-	0.20	40 s
MEME	-	0.14	5 s
MITRA	100%	-	5 m
PROJECTION	100%	-	2 m
MULTIPROFILER	99.7%	-	1 m
PMS	100%	-	217 s
PatternBranching	99.7%	-	3 s
GWM [19]	100%	-	21 s
GWM2	100%	-	64 s

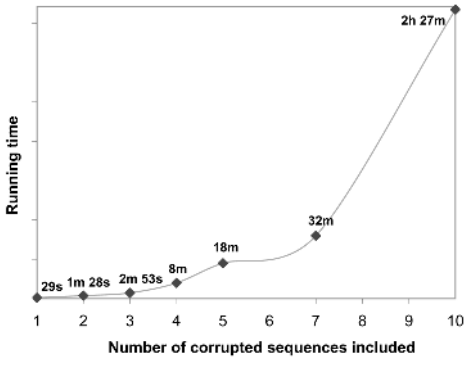

Fig. 1. Results of GWM2 on noisy datasets. Each dataset has 20 sequences that contain $(15, 4)$ motif instances and m' corrupted sequences without containing any motif instances.

GWM [19], while both archive 100% success rate, GWM2 has a slower running time. Since GWM2 was designed to address the motif recognition problem in noisy datasets which contain corrupted sequences, it has to handle more complex characteristics of the given problem, and was not optimized to recognize the motifs in exact datasets in the fastest possible way. However, if we allowed only one motif to be recognized in the dataset, the running time of GWM2 decreased to an average of 26 seconds at $SR = 100\%$. All performances and the running times reported were averaged over the datasets.

Noisy Data. To show the tolerance to noise, we further evaluate the GWM2 approach on the noisy datasets by artificially introducing noisy sequences to the dataset. The noisy datasets were generated that consist of $m = 20$ sequences having motif instances and m' corrupted sequences. The sequences were chosen from the previous exact datasets and mixed randomly.

In accordance with [18], in this experiment we restricted to find the best motif per each run. Running times for GWM2 were averaged over five random datasets. As seen from figure 1, while our approach still archived 100% success rate, its running times were strongly effected by the number of the corrupted sequences in the dataset. This is because the probability of the motif could reach a threshold that causes many pairwise similarities to occur by chance [2,8]. It may further require a preprocessing step that handles the variability of the data to filter corrupted sequences. Fortunately, our approach is considered sufficiently fast for common applications.

3.2 Real Biological Data

We tested our approach on the following biological datasets: DHFR, preproinsulin, and *c-fos*, which consist of upstream regions of eukaryotic genes [9]. These biological datasets were also analyzed in [2,9,15]. For all experiments, we set $l = 20$ and $d = 4$. The number of the sequences assumed to contain the number of motif instances that was initially set to the number of the sequences in the dataset ($K = m$), then was decreased until the motifs were found or $K < m/2$. Once a motif was found in the dataset, it was likely that if the location of the motif was shifted to left or right several positions, other preserved motifs might also be found. Hence, for multiple shifted versions of the motif, only one with the lowest total distance score was selected. Table 2 lists the motifs that match the referenced known motifs with underlined letters corresponding to the matching areas. As seen, GWM2 successfully recognized the reference motifs. Moreover, in many circumstances (results not shown), even the motifs found by GWM2 do not accord with the motifs identified by wet-labs, they actually match to those reported in [4]. It indicates that our approach is able to find the potentially significant motifs.

Table 2. Performance of GWM2 on eukaryotic promoter sequences, using parameters $l = 20$ and $d = 4$. The motifs that match the motifs found by wet-lab experiments [2,9] are listed with underlined letters indicating the matching areas.

Dataset (seqs/bases)	K	Best motifs by GWM2	Experimentally defined motifs
preproinsulin	4	GC<u>AGACCCAGCA</u>CCAGGGAA	AGACCCAGCA
(4/7689)		GAAATTGCAG<u>CCTCAGCCCC</u>	CCTCAGCCCC
		AGG<u>CCCTAATGGGCCA</u>GGCG	CCCTAATGGGCCA
DHFR (4/800)	3	TGCA<u>ATTTCGCGCCA</u>AACTT	ATTTCnnGCCAAACT
c-*fos* (6/4710)	5	<u>CCATATTAGGACATCTG</u>CGT	CCATATTAGGACATCTG

4 Discussion

As more high throughput sequence techniques are being available, recognizing meaningful but weak signals or sites in biological sequences becomes more pressing. However, solving the problem of WMR usually involves with two difficulties: (1) the large pairwise distance between the motif instances cause false pairwise

distances likely to occur at random elsewhere in the dataset that possibly obscures the true motifs, and (2) the increased running time with the increase of the motif length and the noises (the presence of corrupted sequences in the dataset). Therefore, despite various attempts, the existing computational techniques are far from achieving satisfactory results [18,7]. This paper has proposed a graphical approach named as GWM2 to recognize weak motifs in datasets that bear noise. Through experiments, our approach GWM2 has tolerated well to noises, where a fraction of the sequences may not contain any motif instances, while the running time is comparable if not faster than the former methods. GWM2 has been applied with real biological datasets that share the common TFBSs and showed good performance. Moreoever, as three steps in the present method were designed independently of a sequence alphabet, GWM2 is generalizable to other biological sequences such as protein sequences.

One limitation of our approach may be how to determine the motif length l and the degenerate positions d. Fortunately, in most cases of real biological dataset, prior information about the potential motif length is usually provided. Therefore, we could fix the motif length beforehand while varying the value of d. Even if no prior information is available, the motif could be recognized by a trial and error approach with a range of different values of l.

Our approach could be further adapted to find (l, d) motifs with large l and d values. Recently proposed techniques [2,16], that find long motifs with acceptable performance, try to find motifs (l', d') with $l' \ll l$ and $d' \ll d$ ($d' \ll l'$) by using probabilistic sampling techniques. In effect, they change the longer motifs recognition to the shorter ones, then recover the original motifs. However, we believe that a better way to improve the present approach for recognizing weak motifs in the large datasets is to reveal the potential motif by using only a small number of sequences and subsequently validate these motifs with the remaining sequences. For instance, instead of having to find K-cliques, where K is large, we can find k-cliques with $k \ll K$ and recover the potential motifs. Each potential motif will be evaluated against the dataset and if in the dataset we find no less than K number of subsequences having Hamming distance within d different positions from this potential motif, then it is recognized as a valid motif. We plan to further explore this possibility.

References

1. Bailey T. and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *2nd ISMB*, 1994, 33-54.
2. Buhler J. and M. Tompa, "Finding motifs using random projections", *J Comput Biol*, 2002, **9(2)**, 225-242.
3. Chin F., H. Leung, S.Yiu, T. Lam, R. Rosenfeld, W. Tsang, D. Smith and Y. Jiang, "Finding Motifs for Insufficient Number of Sequences with Strong Binding to Transcription Factor", *RECOMB 2004*, San Diego, USA, 125-132.
4. Eskin E. and P. Pevzner, "Finding composite regulatory patterns in DNA sequences", *Bioinformatics*, 2002, **18 Suppl 1**, S354-S363.

5. Helden J., B. Andre, and J. Collado-Vides, "Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies", *J Mol Biol.*, 1998.
6. Hertz G. and G. Stormo G., "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences", *Bioinformatics*, 1999, **15(7-8)**, 563-77.
7. Hu J., B. Li, and D. Kihara, "Limitations and Potentials of Current Motif Discovery Algorithms", *Nucleic Acids Res.*, 2005, **33(15)**, 48994913.
8. Jensen K., M. Styczynski, I. Rigoutsos, and G. Stephanopoulos, "A generic motif discovery algorithm for sequential data", *Bioinformatics*, 2005, **in press**.
9. Keich U. and P.A. Pevzner, "Finding motifs in the twilight zone", *Bioinformatics*, 2002, **18(10)**, 1374-81.
10. Latchman S., Eukaryotic Transcription Factors, *Academic Press*, 2003.
11. Lawrence C., S. Altschul, M. Boguski, J. Liu, A. Neuwland, and J. Wootton, "Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment", *Science*, 1993, **262**, 208-214.
12. Liang S., M. Samanta and B. A. Biegel, "cWINNOWER Algorithm for Finding Fuzzy DNA Motifs", *Journal of Bioinformatics and Computational Biology*, 2004, **2(1)**, 47-60.
13. Liu S., A. Neuwald, and C. Lawrence, "Bayesian models for multiple local sequence alignment and Gibbs sampling strategies", *J. Amer. Statist. Assoc.*, 1995, **90**, 1157-1170.
14. Pevzner P. and S. Sze., "Combinatorial approaches to finding subtle signals in DNA sequences", *Intelligent Systems for Molecular Biology*, 2000, 269-278.
15. Price A., S. Ramabhadran S., and P. Pevzner, "Finding subtle motifs by branching from sample strings", *Bioinformatics*, 2003, **19 Suppl 2**, II149-II155.
16. Rajasekaran S., S. Balla, and C. Huang, "Exact Algorithm for Planted Motif Challenge Problems", *3rd Asia-Pacific Bioinformatics Conference*, 2003, 249-259.
17. Sinha S. and M. Tompa, "A statistical method for finding transcription factor binding sites", *Proc Int Conf Intell Syst Mol Biol*, 2000, **8**, 344-54.
18. Tompa M., N. Li, T. Bailey , G. Church , B. De Moor, E. Eskin, A. Favorov, M. Frith, Y. Fu, W. Kent, V. Makeev, A. Mironov, W. Noble, G. Pavesi, G. Pesole, M. Regnier, N. Simonis, S. Sinha, G. Thijs, J. van Helden, M. Vandenbogaert, Z. Weng, C. Workman, C. Ye, and Z. Zhu, "Assessing Computational Tools for the Discovery of Transcription Factor Binding Sites", *Nature Biotechnology*, 2005, **23(1)**, 137 - 144.
19. Yang X. and J. Rajapakse, "Graphical approach to weak motif recognition", *Genome Informatics*, 2004, **15(2)**, 52-62.