BMC Bioinformatics

Research

Open Access

Extracting transcription factor binding sites from unaligned gene sequences with statistical models

Chung-Chin Lu[1], Wei-Hao Yuan[2] and Te-Ming Chen*[1]

Address: [1]Department of Electrical Engineering, National Tsing Hua University, Hsinchu 30013, Taiwan and [2]Alpha Imaging Technology Corp., Jubei City, Hsinchu 302, Taiwan

Email: Chung-Chin Lu - cclu@ee.nthu.edu.tw; Wei-Hao Yuan - whyuan@a-i-t.com.tw; Te-Ming Chen* - dmchen@abel.ee.nthu.edu.tw

* Corresponding author

from Asia Pacific Bioinformatics Network (APBioNet) Seventh International Conference on Bioinformatics (InCoB2008)
Taipei, Taiwan. 20–23 October 2008

Published: 12 December 2008

BMC Bioinformatics 2008, **9**(Suppl 12):S7 doi:10.1186/1471-2105-9-S12-S7

This article is available from: http://www.biomedcentral.com/1471-2105/9/S12/S7

Abstract

Background: Transcription factor binding sites (TFBSs) are crucial in the regulation of gene transcription. Recently, chromatin immunoprecipitation followed by cDNA microarray hybridization (ChIP-chip array) has been used to identify potential regulatory sequences, but the procedure can only map the probable protein-DNA interaction loci within 1–2 kb resolution. To find out the exact binding motifs, it is necessary to build a computational method to examine the ChIP-chip array binding sequences and search for possible motifs representing the transcription factor binding sites.

Results: We developed a program to find out accurate motif sites from a set of unaligned DNA sequences in the yeast genome. Compared with MDscan, the prediction results suggest that, overall, our algorithm outperforms MDscan since the predicted motifs are more consistent with previously known specificities reported in the literature and have better prediction ranks. Our program also outperforms the constraint-less Cosmo program, especially in the elimination of false positives.

Conclusion: In this study, an improved sampling algorithm is proposed to incorporate the binomial probability model to build significant initial candidate motif sets. By investigating the statistical dependence between base positions in TFBSs, the method of dependency graphs and their expanded Bayesian networks is combined. The results show that our program satisfactorily extract transcription factor binding sites from unaligned gene sequences.

Background

Understanding transcription is central to understanding genetic regulatory mechanisms. The transcription of a gene is generally dependent on the presence of specific signals located at upstream regions of the core-promoter. These specific signals derive from their use as binding sites by transcription factors (TFs), and are therefore termed transcription factor binding sites (TFBSs). Recently, chromatin immunoprecipitation followed by cDNA microarray hybridization (ChIP-chip array) has been used to identify potential regulatory sequences, but the procedure can only map the probable protein-DNA interaction loci within 1–2 kilobases resolution [1]. To find out the exact binding motifs, it is necessary to build a computational method to examine the ChIP-chip array binding sequences and search for possible motifs representing the TFBSs (motif discovery).

There are many computational TFBS motif finding tools available [2-4]. The traditional approach for finding TFBSs is to collect and align a set of promoter sequences of co-regulated genes from either the literature or systematic experiments. Numerous computational tools, such as CONSENSUS [5], EM [6], MEME [7] and the Gibbs sampler [8], have utilized the approach to identify short DNA sequence motifs which are statistically over-represented in the promoter sequences.

Other than the alignment-based motif finding algorithms in above, many approaches have tried to extend to the use of evolutionary conservation information such as phylogenetic footprinting or the detection of combinations of binding sites (termed as cis-regulatory modules; CRMs) [2,3]. Phylogenetic footprinting methods [9-11] is an approach that seeks to identify conserved regulatory elements by comparing genomic sequences between related species. However, due to the statistical nature of the approach, e.g., a small amount of closely related species, not all transcription binding sites can be found by using phylogenetic footprinting. Hence, some algorithms have emerged to combine the alignment-based motif prediction with phylogenetic footprinting such as PhyloGibbs [12] and MY sampler [13]. On the other hand, by the detection of CRMs due to the cooperative interactions between TFs, algorithms like those in [14-16] can produce predictions of substantially better specificity than those of isolated sites.

Recently, more effective motif finders, e.g., MDscan [1], ANN-Spec [17], DMOTIFS [18], DME [19] and Cosmo [20], have taken the advantage of a background set, serving as a negative control. The goal of these discriminant motif finders is to search only for motifs that are most discriminating, that is, only those enriched in the foreground set relative to the background set [2]. Although these motif finders have improved the performance of TFBS prediction, it is still a trouble to have a satisfactory solution. How to find out accurate binding motifs may require much attention in the computational biology community. In this study, an improved sampling algorithm is proposed to incorporate the binomial probability model to build significant initial motif sets. By investigating the statistical dependence between base positions in TFBSs, it appears feasible to use statistical models to formulate the structural dependence of a motif in the identification of TFBSs. In light of this observation, the method of dependency graphs and their expanded Bayesian networks [21] is combined and prediction results show that our algorithm is able to find out motifs more consistent with previously known evidence.

Methods

Let *TF* be one of the transcription factors to be investigated. The binding dataset of the transcription factor *TF*, denoted as B_{TF}, consists of the sequences with low binding *p*-value (< 0.001) to the *TF* in the ChIP-chip array data [22]. A sliding window of size w is used to extract segments of length w when sliding through each of the sequences in B_{TF}.

Let S_{TF} be the collection of all extracted segments from B_{TF}, M the number of sequences in the binding dataset B_{TF}, L_i the length of the ith sequence in the binding dataset B_{TF}, and T_{TF} the total number of segments in S_{TF}. Then

$$T_{TF} = \sum_{i=1}^{M} (L_i - w + 1).$$

To discover the binding motifs of the transcription factor *TF*, a number of initial candidate motif sets for *TF* is subsequently built from the collection S_{TF} of extracted segments. Note that the contents of segments, called patterns, in S_{TF} may not be distinct.

Most of early motif finding algorithms, such as Gibbs sampler [8] and MEME [23], have a weakness, where initial candidate motif sets are built by randomly extracting segments from sequences in the binding dataset B_{TF} (i.e. randomly selecting segments from the set S_{TF}). To improve the deficiency, the binomial probability distribution model is firstly utilized in the establishment of a number of initial candidate motif sets in our algorithm.

Then in the process of iterative sampling in our algorithm to expand and/or trim each of the initial candidate motif sets, the method of dependency graphs and their expanded Bayesian networks [21] is used to develop a statistical model for the background motif set identified as the union $S = \cup_{TF} S_{TF}$ of segments extracted from all transcription factor binding datasets.

The basic procedure to find the binding motifs of the transcription factor *TF* is as follows:

1. Build N initial candidate motif sets.

(a) Take N distinct patterns from the set S_{TF} with the most highest significance scores as the candidates by the binomial distribution model (see the Binomial probability distribution model subsection).

(b) Then for each of the significant binding site candidates for the transcription factor , in view of evolution, collect all segments in whose patterns have no more than d Hamming distance matching to the candidate pattern to form an initial candidate motif set.

At this stage, N initial candidate motif sets for the transcription factor TF are built.

2. Iteratively sample through the binding dateset B_{TF} to expand and/or trim each of the N initial candidate motif sets so that their approximate maximum a posteriori (AMAP) scores [1,24] can keep increasing until the N candidate motif sets are invariant in K consecutive iterations (see the Iterative sampling subsection).

(a) In the calculation of AMAP scores in this stage, the background model for the background motif set $S = \cup_{TF}$-S_{TF} is established under the method of dependency graphs and their expanded Bayesian networks (see the Method of dependency graphs and their expanded Bayesian networks subsection).

3. Refine each of the N candidate motif sets by re-examining all the segments already included in the motif set. A segment is removed from the motif set if doing so increases the AMAP score.

A simple flowchart for our algorithm is shown in Figure 1. The following subsections will expatiate on each stage of our algorithm. As an illustration of the dynamics of the PWM and the rank of different candidate motif sets at different stages of our algorithm, a summary of the prediction process for the motif of the transcription factor CBF1 is given in Figure 2.

Initial motif sets building

Our method begins by enumerating those patterns in S_{TF} that appear most often in the binding dataset B_{TF} than in others. What we want to do first is to calculate the appearance probability of a pattern in S_{TF}, which is the probability that the pattern appears no less than n times in the binding dataset B_{TF}. If a pattern b appears more often than other patterns in S_{TF} and its occurrence probability in a generic intergenic region is comparatively low, the calculated significance score of b would be relatively high. We will take patterns with the most highest significance scores as the candidates to build a number of initial candidate motif sets.

Binomial probability distribution model

The probability to observe exactly j occurrences of pattern b in the collection S_{TF} of segments extracted from the binding dataset B_{TF} is estimated by the binomial distribution

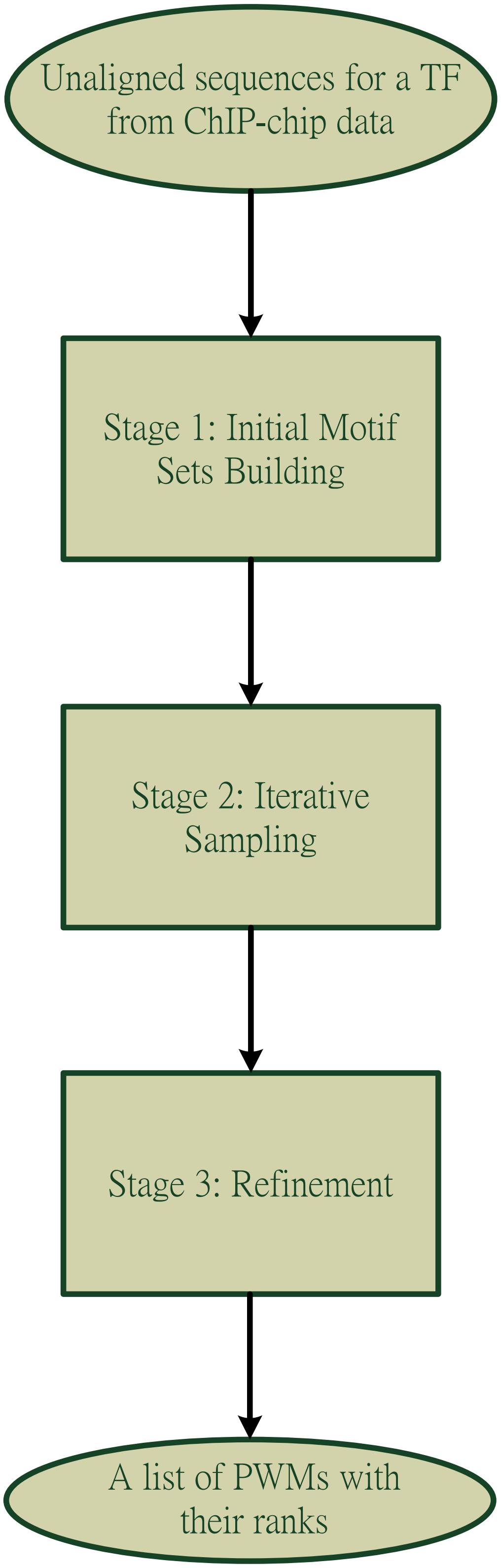
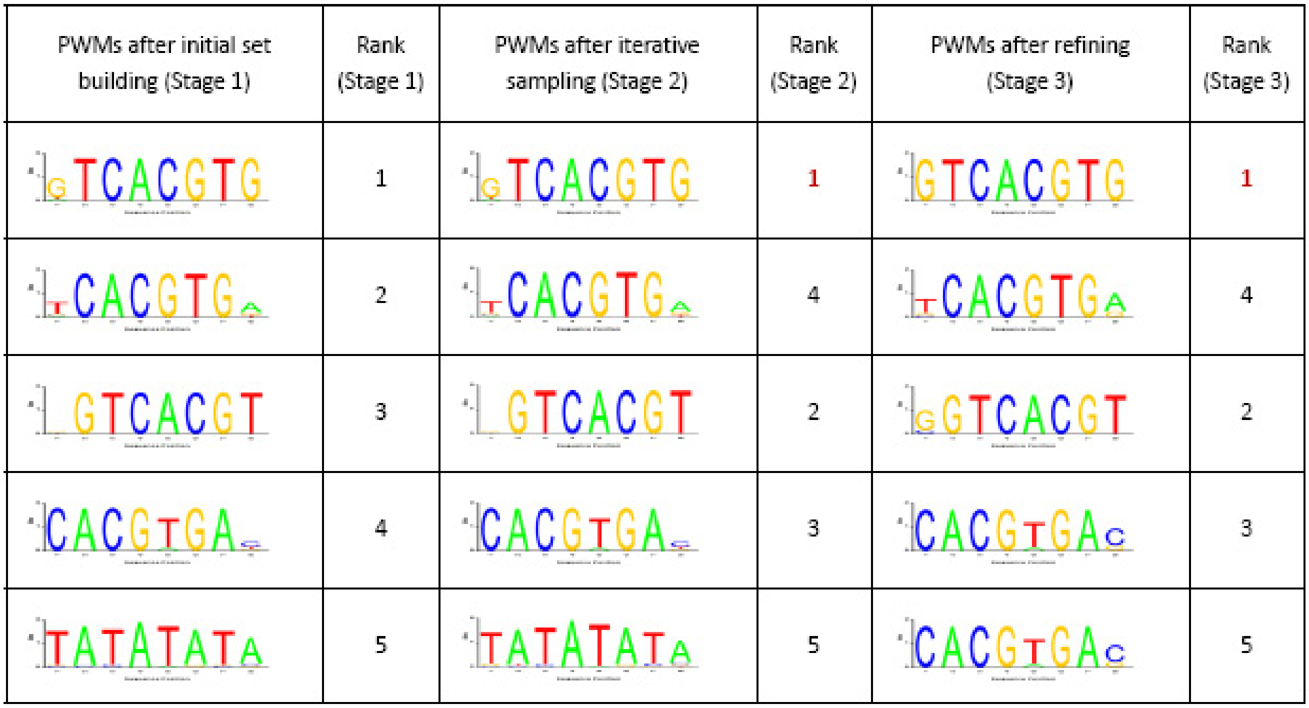

Unaligned sequences for a TF from ChIP-chip data
Stage 1: Initial Motif Sets Building
Stage 2: Iterative Sampling
Stage 3: Refinement
A list of PWMs with their ranks

Figure 1
A flowchart of our algorithm.



PWMs after initial set building (Stage 1)	Rank (Stage 1)	PWMs after iterative sampling (Stage 2)	Rank (Stage 2)	PWMs after refining (Stage 3)	Rank (Stage 3)
GTCACGTG	1	GTCACGTG	1	GTCACGTG	1
TCACGTGA	2	TCACGTGA	4	TCACGTGA	4
GTCACGT	3	GTCACGT	2	GGTCACGT	2
CACGTGA	4	CACGTGA	3	CACGTGAC	3
TATATATA	5	TATATATA	5	CACGTGAC	5

Figure 2
An illustration of the dynamics of the PWM and the rank of five candidate motif sets at different stages of our algorithm in the motif prediction of the transcription factor CBF1.

$$P_{TF}(occ(b)=j)=\binom{T_{TF}}{j}\times(f(b))^{j}\times(1-f(b))^{T_{TF}-j},$$

where $occ(b)$ is the occurrence times of pattern b in S_{TF} and $f(b)$ is the probability that pattern b occurs in the intergenic region and is estimated as the relative frequency of pattern b in the union $S = \cup_{TF}S_{TF}$ of segments extracted from all transcription factor binding datasets. The probability to observe n or more occurrences of the pattern b in S_{TF} is

$$P_{TF}(occ(b)\geq n)=1-\sum_{j=0}^{n-1}P_{TF}(occ(b)=j). \qquad (1)$$

We define the significance score $sig_{TF}(b)$ of a pattern b to *TF* as

$$sig_{TF}(b) = -\log_{10}(P_{TF}(occ(b)\geq n)).$$

The less probable pattern b in S appears more than n times in S_{TF}, the more probable will it be a binding site candidate for the transcription factor *TF*. We will take N distinct patterns with the most highest significance scores as the candidates.

For each of the N significant binding site candidates for the transcription factor *TF*, in view of evolution, collect all segments in S_{TF} whose patterns have no more than d Hamming distance matching to the candidate pattern to form an initial candidate motif set. Thus N initial candidate motif sets for the transcription factor *TF* are built at the end of this stage. As an example, the PWM and the rank of five initial candidate motif sets for the motif prediction of the transcription factor CBF1 are shown in Figure 2.

Iterative sampling

In this stage, a sampling method is used to expand and/or trim each of the N initial candidate motif sets M_1, M_2,.., M_N. For our purpose, a false motif set M_{N+1} is created by randomly selecting e_0 (e_0 is equal to the maximum size of the N initial candidate motif sets) segments from the collection S_{TF} such that $M_i \cap M_{N+1} = \emptyset$, for all i = 1, 2,..., N. In addition, let the collection $S = \cup_{TF}S_{TF}$ of segments extracted from all transcription factor binding datasets represent the intergenic background and here be denoted as M_{BG}.

Approximate maximum a posteriori (AMAP) measure

The score $amap_{M_i}$ of the approximate maximum a posteriori (AMAP) measure of the candidate motif set M_i is defined as [1,24]

$$amap_{M_i} = \frac{1}{w}\left\{ \sum_{s=0}^{w-1} \sum_{j \in A,T,G,C} p_{s,j}\log(p_{s,j}) - \frac{1}{n_i}\sum_{m \in M_i}\log(P(m \mid M_{BG})) \right\},$$

where $p_{s,\,j}$ is the frequency of nucleotide j at base position s in the candidate motif set M_i(which can be retrieved from the position specific scoring matrix (PSSM) of M_i), n_i is the number of segments in M_i, and $P(m|M_{BG})$ is the probability of the pattern of segment m in the motif set M_i under an expanded Bayesian network (EBN) model [21] developed from the background motif set M_{BG} (EBN model will be discussed shortly).

The first part of the AMAP score is a negative entropy, which is higher if there are more similar patterns in the candidate motif set M_i. A motif set M_i with all identical patterns has the maximum negative entropy 0, whereas equal nucleotide frequencies at every position in the PSSM of M_i has the minimum negative entropy. And a segment m in the candidate motif set M_i which has a pattern much different from the background motif model built from M_{BG} would have lower appearance probability $P(m|M_{BG})$ and hence increases the score $amap_{M_i}$ of the AMAP measure of M_i.

Sampling strategy

In each iteration, there are two steps for the sampler, the S-step and the M-step.

In the S-step, the sampler samples a site by randomly selecting a sequence from B_{TF} and then randomly picking up a site in the selected sequence to extract a segment m_s of length w. For $1 \le i \le N$, if the sampled segment m_s appears in M_i, segment m_s will be removed from M_i if the AMAP score $amap_{M_i}$ of the candidate motif set M_i increases after its removal; otherwise, segment m_s will be kept in M_i. Note that the PSSM of the motif model M_i should be retrained if the sampled segment m_s is removed from M_i.

Which one of the N + 1 motif sets would be the best motif set for the sampled segment m_s will depend on the appendant score app_{M_i} that the segment m_s is derived from M_i [24,25]

$$app_{M_i} = \log\left(\frac{P(n_i)}{1-P(n_i)} \frac{P(m_s|M_i)}{P(m_s|M_{BG})} \right), 1 \le i \le N+1, \qquad (2)$$

where n_i is the size of current motif set M_i, $P(n_i)$ equals $\frac{n_i}{T_{TF}}$, $P(m_s|M_i)$ and $P(m_s|M_{BG})$ are the probabilities of the content of the sampled segment m_s under the PSSM model developed from the current motif set M_i and under an EBN model developed from the background M_{BG}, respectively. The sampled segment m_s will be considered to append into the motif set M_i with the highest appendant score app_{M_i}. If $app_{M_{N+1}}$ is the highest score, then the sampled segment m_s is appended into the false motif set M_{N+1} unless m_s is already there and the current iteration stops here. If for some i, $1 \le i \le N$, app_{M_i} is the highest score, the sampled segment m_s will be further checked in the M-step to see if we really want to append m_s into M_i unless we have processed m_s for M_i at the beginning of this S-step as in above and the current iteration stops here.

In the M-step, the sampler has to decide whether the newly sampled segment m_s should be appended into the candidate motif set M_i or not. The AMAP measure again will be used to evaluate our decision. The sampled segment m_s is appended into the candidate motif set M_i if and only if the score $amap_{M_i}$ of the motif model M_i is increased once the sampled segment m_s is appended to M_i. Note that the PSSM of the motif model M_i should be retrained after the sampled segment b_s is appended to M_i. Now the M-step is done and the current iteration stops here.

The sampler will iteratively sample through the binding dataset B_{TF} to expand and/or trim the N candidate motif sets M_1, M_2,.., M_N so that their AMAP scores $amap_{M_i}$ will keep increasing. The N candidate motif sets will tend to be invariant after a (larger) number of iterations. The stopping criterion of the sampling process is that all the N candidate motif sets are invariant in K consecutive iterations. The parameter K is usually set to be 1% of the size of S_{TF}.

Alternative sampling strategy

There is an alternative sampling strategy as follows.

In the S-step, the new sampler also randomly samples a site from a sequence in B_{TF} to extract a segment $m_s$of length w. For $1 \leq i \leq N$, if the pattern of the sampled segment m_s appears in M_i, all the segments in M_i whose pattern is the same as that of m_s will be removed if the AMAP score $amap_{M_i}$ of the motif set M_i increases after their removal. Otherwise, these segments will be kept in M_i.

Also in the S-step, if $app_{M_{N+1}}$ is the highest among all app_{M_i}, $1 \leq i \leq N + 1$, then all segments in the set S_{TF} having the same pattern as that of the sampled segment m_s will be appended into the false motif set M_{N+1} unless these segments are already there and the current iteration stops here. If app_{M_i} is the highest for some i, $1 \leq i \leq N$, the sampled segment m_s will be further checked in the M-step to see if we really want to append those segments in the set S_{TF} having the same pattern as that of the sampled segment m_s into M_i unless we have already processed those segments for M_i at the beginning of this S-step as in above and the current iteration stops here.

In the M-step, all the segments in the set S_{TF} having the same pattern as that of the sampled segment m_s are decided to append to the candidate motif set M_i if and only if the AMAP score $amap_{M_i}$ of M_i increases after these segaments are appended into M_i.

Method of dependency graphs and their expanded Bayesian networks

Considering the binding mechanism of transcription factors to specific DNA sites (motifs), there must be distinctive features for the specific motif regions from other intergenic regions which represent the background DNA sequence. Hence, it is conceivable that we can use a statistical model to capture the feature of a specific DAN site (motif) or a generic DNA intergenic region (background). Since the size of a candidate motif set M_i is often small, a PSSM model is commonly used for M_i instead of any other more sophisticated statistical model. However, the size of the background motif set M_{BG} is usually large enough to be equipped with a more sophisticated one.

As reported in [21], a dependency graph model is used to fully capture the intrinsic interdependency between base positions in a motif or region. The establishment of dependency between two positions is based on a χ^2-test from known sample data. An edge is established between two nodes (a node represents a base position) in the graph if the two corresponding base positions of the motif or region are dependent. After all dependent edges have being established completely, a dependency graph for the motif or region is constructed. An example of a dependency graph with 7 nodes is shown in Figure 3.

As reported in [21], although the dependency graph can fully capture the intrinsic interdependency between base positions in a motif or region, it is difficult, if not impossible, to perform statistical inference based on the dependency graph. To resolve the dilemma, the dependency graph is expanded to form a Bayesian network (which is a directed acyclic graph that facilitates statistical reasoning) by allowing a base position in the dependency graph to appear more than once in the Bayesian network as nominally distinct nodes. Figure 4 shows an example of an expanded Bayesian network of the dependency graph in Figure 3. For the detailed procedure of constructing an expanded Bayesian network (EBN) from a dependency graph, please see [21]. In this paper, we use EBNs to model the background motif set M_{BG}.

Continued with the same example of the motif prediction of the transcription factor CBF1, the PWM and the rank of the five candidate motif sets at the end of the iterative sampling stage are also shown in Figure 2, together with the final results at the end of the refinement stage.

Results and discussion

Data

In order to search for the transcription factor binding sites that regulate gene expressions, we collected binding pro-

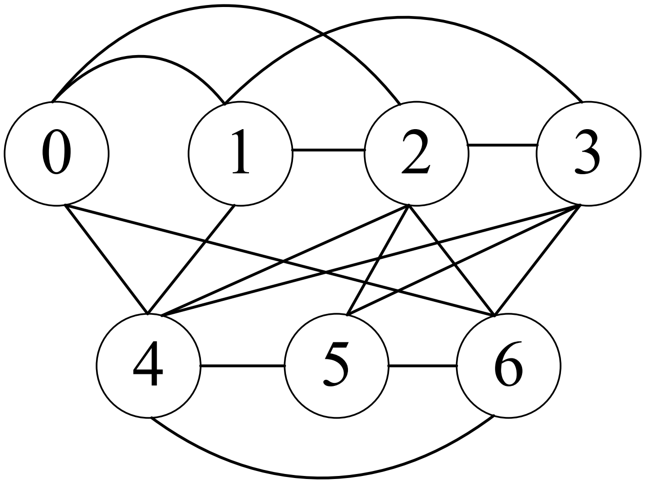


Figure 3
An example of dependency graph.

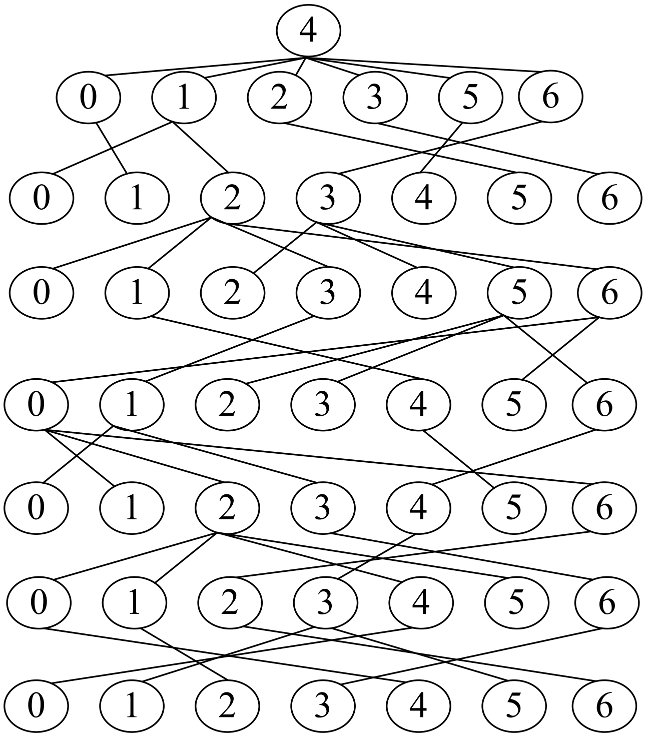


Figure 4
An example of expanded Bayesian network.

motor sequences from the cDNA microarray hybridization (ChIp-chip array) of yeast genome [22]. Each of the binding sequences may contain some unknown motifs that are implanted at unknown positions. These data represent the binding affinity of a target transcription factor to the promoter region of a gene *in vivo*. The experiment protocol assigns a binding *p*-value to each binding promoter sequence of the corresponding transcription factor. A sequence with binding *p*-value less than 0.001 is considered to be bound by the corresponding transcription factor. The threshold of 0.001 is set up to reduce the false positive identification in yeast genome-wide screening.

We collected the ChIp-chip array sequence data from the "Motif discovery results – Discovered motifs, version 24" at [26]. For a transcription factor *TF* to be investigated, we collected all sequences with binding *p*-value less than the threshold 0.001 to *TF* into the binding dataset B_{TF}. There are 65 binding datasets B_{TF} being able to be collected from Harbison's website.

Accuracy measurement and comparison

To evaluate the performance of our program, we collected known specificities from many famous websites, such as YPD, SCPD, Transfac and from the literature with experimental evidence [27] to compare with the discovered specificities predicted by our program.

Among the 65 binding datasets B_{TF} collected from Harbison's website, we chose 36 transcription factor binding datasets which have known specificities with experimental evidence to evaluate the performance of our program. The results of our program for the 36 transcription factor binding datasets are listed in Figure 5. It is deserved to be mentioned that the specificity reported for transcription factor PHO2 in Harbrison *et al*'s website is "GTGCGsyGCG", while the predicted result of our program is "ATTATC". In this case, the newly found motif by our program is more consistent with the results reported by Barbaric *et al* [28] that PHO2 binds to an AT-rich region than the specificity reported in Harbrison *et al*'s website.

In this study, we compared our program with two online programs, MDscan [29] and Cosmo [30]. MDscan is a famous program that can be used to examine the ChIP-array selected sequences and search for DNA sequence motifs representing the protein-DNA interaction sites. It takes the advantage of combining two widely adopted motif search strategies, word enumeration and position-specific weight matrix updating, and incorporates the ChIP enrichment information to accelerate the search and enhance its success rate. The comparison of MDscan with our program is shown in Table 1. Also reported in Table 1 is the performance of our algorithm when the the PSSM model [8] instead of the EBN model [21] is used to model the background motif set M_{BG} in the calculation of the AMAP scores of the *N* candidate motif sets and the appendant scores of the *N* + 1 motif sets. In Table 1, for each transcription factor, the number in each 'Rank' column indicates the rank of the predicted motif which is most consistent with the known evidence from the top ten predicted candidate motifs.

As shown in Table 1, our approach with EBN background model outperforms the other two methods. Our approach with EBN background model gives 30 out of the 36 most predicted motifs for the corresponding 36 transcription factors with the 1st rank, while MDscan and our approach with PSSM background model give only 20 out of 36 and 15 out of 36 most predicted motifs with the 1st rank, respectively. Moreover, MDscan fails in discovering a motif for three transcription factor binding datasets, while our approach in this study is still able to predict a motif consistent with the known evidence.

Cosmo (constrained search for motifs) is a general purpose algorithm for conserved motif detection that allows the search to be supervised by specifying a set of constraints that the PWM of the unknown motif must satisfy. Such constraints may be formulated derived from prior biological knowledge about the structure of the transcription factor, such as the length of the motif intervals. Although Cosmo is based on the same two-component

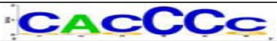
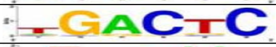
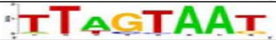
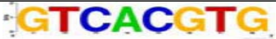
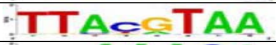
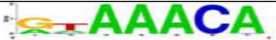
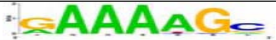
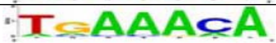
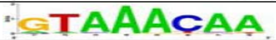
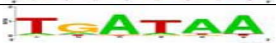
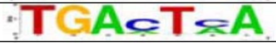
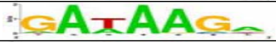
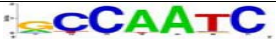
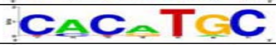
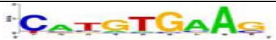
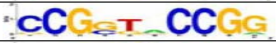
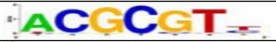
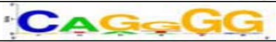
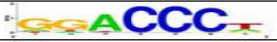
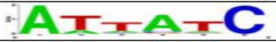
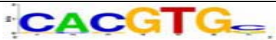
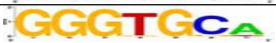
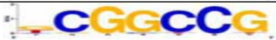
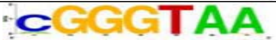
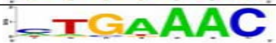
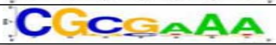
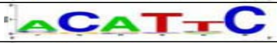
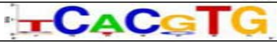
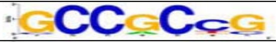
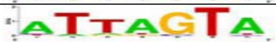
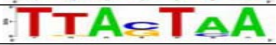
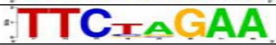
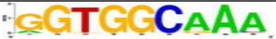
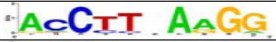
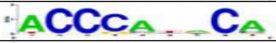
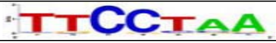
TF	Known specificity	Specificity source	Motif PWM from our program	Rank	Match
AFT2	...AAAGTG**CACCC**ATT...	YPD		1	TP=5, FN=9, FP=1
BAS1	**TGACTC**	SCPD		1	TP=6, FN=0, FP=0
CAD1	**TTACTAA**	YPD		1	TP=6, FN=1, FP=2
CBF1	**RTCACRTG**A	SCPD		1	TP=8, FN=1, FP=0
CIN5	**TTACRTAA**	YPD		1	TP=8, FN=0, FP=0
FKH2	G**GTAAACA**A	Tfac		1	TP=7, FN=2, FP=0
DAL82	**GAAAA**TTGCGTT	Dorrington RA		2	TP=5, FN=7, FP=2
DIG1	**TGAAAC**	SGD		2	TP=6, FN=0, FP=2
FKH1	G**GTAAACAA**	Tfac		1	TP=8, FN=1, FP=0
GAT1	**GATAA**	YPD		1	TP=5, FN=0, FP=1
GCN4	AR**TGACTCW**	Tfac		1	TP=7, FN=2, FP=0
GLN3	**GATAAGA**TAAG	YPD		1	TP=7, FN=4, FP=0
HAP4	YCN**NCCAATN**ANM	Tfac		1	TP=7, FN=6, FP=0
INO2	ATTT**CACATGC**	Tfac		1	TP=7, FN=4, FP=0
INO4	**CATGTGAA**AT	YPD		2	TP=8, FN=2, FP=1
LEU3	YG**CCGGTACCGG**YK	SCPD		1	TP=10,FN=4, FP=0
MBP1	**ACGCGT**	YPD		1	TP=6, FN=0, FP=1
MSN2	**MAGGGG**	Tfac		2	TP=6, FN=0, FP=0
NRG1	**CCCT**	Park SH		1	TP=4, FN=0, FP=3
PHO2	**ATTA**	Barbaric S		1	TP=4, FN=0, FP=2
PHO4	**CACGTK**NG	Tfac		1	TP=6, FN=2, FP=1
RCS1	AAMT**GGGTGCA**KT	Tfac		1	TP=7, FN=6, FP=0
RDS1	**KCGGCCG**	SGD		1	TP=7, FN=0, FP=0
REB1	**CGGGTRR**	SGD		1	TP=7, FN=0, FP=0
STE12	A**TGAAAC**	Tfac		1	TP=6, FN=1, FP=1
SWI4	**CNCGAAA**	SCPD		3	TP=7, FN=0, FP=0
TEC1	**CATTC**Y	YPD		1	TP=5, FN=1, FP=1
TYE7	**CANNTG**	YPD		1	TP=6, FN=0, FP=1
UME6	W**GCCGCCG**W	Tfac		1	TP=7, FN=2, FP=0
YAP1	**TTASTM**A	Nguyen DT		1	TP=6, FN=1, FP=1
YAP7	**TTACTAA**	YPD		1	TP=7, FN=0, FP=0
HSF1	**TTCTAGAA**NNTTCT	Tfac		1	TP=8, FN=6, FP=0
RPN4	**GGTGGCAAA**	Tfac		1	TP=9, FN=0, FP=0
ZAP1	**ACCCTAAAGG**T	Tfac		1	TP=8, FN=3, FP=2
RAP1	WRM**ACCCATACA**YY	Tfac		1	TP=9, FN=5, FP=0
MCM1	W**TTCCYAA**WNNGGTAA	Tfac		2	TP=7, FN=8, FP=0

Figure 5

Predicted results of our program compared with known evidence. The letter symbols used in the 'Known specificity' column have the following mapping: aA: a tT: t gG: g cC: c wW: at rR: ag mM: ac kK: tg yY: tc sS: gc dD: atg hH: atc vV: agc bB: tgc nN: atgc

Table 1: Comparison of MDscan and our program.

TF	Rank (EBN model)	Rank (PSSM model)	Rank (MDscan)
AFT2	1	1	1
BAS1	1	5	2
CAD1	1	5	4
CBF1	1	2	1
CIN5	1	2	1
FKH2	1	1	3
DAL82	2	1	N*
DIG1	2	3	1
FKH1	1	2	1
GAT1	1	2	2
GCN4	1	1	1
RPN4	1	1	1
GLN3	1	1	1
HAP4	1	3	1
INO2	1	2	1
INO4	2	2	1
LEU3	1	1	1
MBP1	1	1	2
MSN2	3	3	4
NRG1	1	2	1
PHO2	1	1	2
PHO4	1	2	1
RCS1	1	1	5
RAP1	1	3	2
RDS1	1	1	1
REB1	1	3	4
STE12	1	2	1
SWI4	3	2	4
TEC1	1	4	1
TYE7	1	1	4
UME6	1	1	3
YAP1	1	5	1
YAP7	1	2	1
HSF1	1	1	1
AZF1	1	1	N
MCM1	2	2	N

* N means that the program predicts no motif.

multinomial mixture model used in MEME, it employs the likelihood principle instead of the *E*-value criterion in MEME. In addition, three model types (OOPS, ZOOPS, or TCM) can data-adaptively be selected in Cosmo to achieve better performance. Since there is no prior knowledge used in our program, we compared it to the constraint-less version of the Cosmo program. On the other hand, since the Cosmo program reports only one motif PWM for a dataset, instead of a list of ranked candidate motif PWMs as in MDscan, we adopted only the rank 1 results of our program in this comparison. To evaluate the performance of both programs, we used the statistics proposed by Tompa et al. [4]. For a (computational) tool at the site level, the performance statistics on a dataset are defined as follows:

$$\begin{aligned}\text{Sensitivity} &: \quad Sn = TP/(TP+FN)\\ \text{Positive predictive value} &: \quad PPV = TP/(TP+FP)\\ \text{Average site performance} &: \quad ASP = (Sn+PPV)/2\end{aligned}$$

where *TP* is the number of known sites overlapped by predicted sites, *FN* is the number of known sites not overlapped by predicted sites, and *FP* is the number of predicted sites not overlapped by known sites. To summarize the performance of a given tool over a collection C of datasets, we compute the "combined" statistics as though *C* were one large dataset by adding *TP*, *FP* and *FN* respectively over the datasets in C. Then the combined statistics of our program are *Sn* = 0.6698, *PPV* = 0.8206, and *ASP* = 0.7452, while those of Cosmo are *Sn* = 0.6573, *PPV* = 0.5134 and *ASP* = 0.5854. For the detailed Cosmo prediction results and the comparison of the two programs, please see Figure S1 (see Additional file 1). The comparison shows that our program can offer better performance than Cosmo, especially in the elimination of false positives.

The parameters used in our program include the sliding window size *w* used to extract segments from binding datasets B_{TF} to form S_{TF}, the Hamming distance *d* used to collect segments from S_{TF} to establish initial candidate motif sets, the number of most dependent edges used to form a dependency graph for the background motif model and the number of parents used in the construction of an expanded Bayesian network from the dependency graph [21]. The parameters used in our program to give the best predicted motifs for each of the 36 transcription factors are listed in Table S1 (see Additional file 2). Comparing the performance of the two sampling strategies discussed in the Method section, as shown in Figure S2 (see Additional file 3), we found that the alternative sampler is faster and has almost identical best predicted motifs with those by the primary sampler, except that transcription factors GCN4, HAP4 and PHO4 have the best predicted motifs one nucleotide position shift from those by the primary sampler. In addition, the alternative sampler is slightly better than the primary sampler in the sense that the best predicted motif for the transcription factor DIG1 promotes its rank from the 2nd place by the primary sampler to the 1st place by the alternative sampler.

Conclusion

In this study, we employed the binomial probability model to establish a number of initial candidate motif sets, and used the method of dependence graphs and their expanded Bayesian networks to model the background motif set as a control to predict TFBSs (motifs) from a set of unaligned DNA sequences. The prediction results suggest that, overall, our algorithm outperforms MDscan since the predicted motifs are more consistent with previously known specificities reported in the literature and have better prediction ranks. And when compared with the constraint-less Cosmo program, our algorithm has a slightly higher combined sensitivity *Sn*, a much higher positive predictive value *PPV* and a higher average site performance *ASP*. However, the performance of our algorithm is not much better if the length of possible binding sites are too long (more than 12 bps). Further research is needed to discover long motifs.

Furthermore, variable spacing within binding sites is legitimate for some transcription factors while this study focuses on ungapped motif discovery. Programs such as BIPAD [31] and spaced dyad [32] have investigated into such a bipartitie sequence element discovery problem. Therefore another direction for our future research is to investigate into gapped motifs.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

CL and WY developed and implemented the method. All authors participated in discussions and writing of the paper.

Additional material

Additional file 1

Figure S1 – Predicted results of the constraint-less Cosmo program and the comparison with our program.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S12-S7-S1.pdf]

Additional file 2

Table S1 – Parameters used in our program to give the best predicted motifs for the 36 transcription factors.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S12-S7-S2.pdf]

Additional file 3

Figure S2 – Comparison of the predicted results with the primary and alternative samplers.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-S12-S7-S3.pdf]

Acknowledgements

We thank Yi-Sian Liao for helping the execution of our prediction program. We also thank the reviewers for their valuable suggestions which improve the presentation of this paper. This work was supported by the National Science Council, Taiwan, under Contracts NSC 93-3112-B-007-004 and NSC95-3114-P-002-005-Y.

This article has been published as part of *BMC Bioinformatics* Volume 9 Supplement 12, 2008: Asia Pacific Bioinformatics Network (APBioNet) Seventh International Conference on Bioinformatics (InCoB2008). The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/9?issue=S12.

References

1. Liu XS, Brutlag DL, Liu JS: **An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitaion microarray experiments.** *Nat Biotechnol* 2002, **20:**835-839.
2. Zhang MQ: **Computational analyses of eukaryotic promoters.** *BMC Bioinformatics* 2007, **8(Suppl 6):**.
3. Wasserman WW, Sandelin A: **Applied bioinformatics for the identification of regulatory elements.** *Nat Rev Genet* 2004, **5:**276-287.
4. Tompa M, Li N, Bailey TL, Church GM, Moor BD, Eskin E, Favorov AV, Frith MC, Fu Y, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng Z, Workman C, Ye C, Zhu Z: **Assessing computational tools for the discovery of transcription factor binding sites.** *Nat Biotechnol* 2005, **23:**137-144.
5. Hertz GZ, George W, Hartzell I, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6:**81-92.
6. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7:**41-51.
7. Bailey TL, Elkan C: **The value of prior knowledge in discovering motifs with MEME.** In *Proceedings of the Third International Comference on Intelligent Systems for Molecular Biology Menlo Park, CA: AAAI Press*; 1995:21-29.
8. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262:**208-214.
9. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT: **Embryonic ε and γ globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints.** *J Mol Biol* 1988, **203:**439-455.
10. Blanchette M, Schwikowski B, Tompa M: **Algorithms for phylogenetic footprinting.** *J Comput Biol* 2002, **9:**211-223.
11. Lenhard B, Sandelin A, Mendoza L, Engstrom1 P, Jareborg N, Wasserman WW: **Identification of conserved regulatory elements by comparative genome analysis.** *J Biol* 2003, **2:**13.
12. Siddhartan R, Siggia ED, van Nimwegen E: **PhyloGibbs: A Gibbs sampling motif finder that incorporates phylogeny.** *PLoS Comput Biol* 2005, **1:**e67.
13. Andersson SA, Lagergren J: **Motif Yggdrasil: Sampling sequence motifs from a tree mixture model.** *J Comput Biol* 2007, **14(5):**682-697.
14. Wasserman WW, Fickett JW: **Identification of regulatory regions which confer muscle-specific gene expression.** *J Mol Biol* 1998, **278:**167-181.
15. Johansson O, Alkema W, Wasserman WW, Lagergren J: **Identification of functional clusters of transcription factor binding motifs in genome sequences: the MSCAN algorithm.** *Bioinformatics* 2003, **19(Suppl 1):**i169-i176.
16. Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B: **Computational detection of cis-regulatory modules.** *Bioinformatics* 2003, **19(suppl 2):**ii5-ii14.
17. Workman CT, Stormo GD: **ANN-Spec: A method for discovering transcription factor binding sites with improved specificity.** *Pac Symp Biocomput* 2002, **5:**467-478.
18. Sinha S: **Discriminative motifs.** *J Comput Biol.* 2003, **10:**599-615.
19. Smith AD, Sumazin P, Zhang MQ: **Identifying tissue-selective transcription factor binding sites in vertebrate promoters.** *Proc Natl Acad Sci USA* 2005, **102:**1560-1565.

20. Bembom O, Keles S, van der Laan MJ: **Supervised detection of conserved motifs in DNA sequences with Cosmo.** *Stat Appl Genet Mol Biol* 2007, **6:**8.
21. Chen TM, Lu CC, Li WH: **Prediction of splice sites with dependency graphs and their expanded bayesian networks.** *Bioinformatics* 2004, **21:**471-482.
22. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Hannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, KT T, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431:**99-104.
23. Bailey TL, Elkan C: **Unsupervised learning of multiple motif in biopolymers using expectation maximization.** *Machine Learning* 1995, **21:**51-80.
24. Liu J, Neuwald AF, Larence CE: **Bayesian models for multiple local sequence alignment and Gibbs sampling strategies.** *J Am Stat Assoc* 1995, **90:**1156-1170.
25. Neuwald AF, Liu JS, Lawrence CE: **Gibbs motif sampling: Detection of bacterial outer membrane protein repeats.** *Protein Sci* 1995, **4:**1618-1632.
26. **Motif discovery results – Discovered motifs, version 24** [http://fraenkel.mit.edu/Harbison/release_v24/final_set/Final_Motifs/]
27. MacIsaac KD, Wang T, Gordeon DB, Gifford DK, Stormo GD, Fraenkel E: **An improved map of conserved regulatory sites for Saccharomyces cerevisiae.** *BMC Bioinformatics* 2006, **7:**113.
28. Barbaric S, Munsterkotter M, Svaren J, Horz W: **The homeodomain protein Pho2 and the basic-helix-loop-helix protein Pho4 bind DNA cooperatively at the yeast PHO5 promoter.** *Nucleic Acids Res* 1996, **24:**4479-4486.
29. **MDscan: A fast and accurate motif finding algorithm with aApplications to chromatin immunoprecipitation microarray experiments** [http://ai.stanford.edu/~xsliu/MDscan/]
30. **Cosmo – Constrained search for motifs in DNA sequences** [http://cosmoweb.berkeley.edu/]
31. Bi C, Rogan PK: **BIPAD: A web server for modeling bipartite sequence elements.** *BMC Bioinformatics* 2006, **7:**76.
32. van Helden J, Rios AF, J CV: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acid Res* 2000, **28:**1808-1818.