Comparative Analysis of Regulatory Motif Discovery Tools for Transcription Factor Binding Sites

Wei Wei and Xiao-Dan Yu*

Department of Pathobiology, Center of Computational Biology, Institute of Basic Medical Sciences, Beijing 100850, China.

In the post-genomic era, identification of specific regulatory motifs or transcription factor binding sites (TFBSs) in non-coding DNA sequences, which is essential to elucidate transcriptional regulatory networks, has emerged as an obstacle that frustrates many researchers. Consequently, numerous motif discovery tools and correlated databases have been applied to solving this problem. However, these existing methods, based on different computational algorithms, show diverse motif prediction efficiency in non-coding DNA sequences. Therefore, understanding the similarities and differences of computational algorithms and enriching the motif discovery literatures are important for users to choose the most appropriate one among the online available tools. Moreover, there still lacks credible criterion to assess motif discovery tools and instructions for researchers to choose the best according to their own projects. Thus integration of the related resources might be a good approach to improve accuracy of the application. Recent studies integrate regulatory motif discovery tools with experimental methods to offer a complementary approach for researchers, and also provide a much-needed model for current researches on transcriptional regulatory networks. Here we present a comparative analysis of regulatory motif discovery tools for TFBSs.

Key words: motif, TFBS, non-coding DNA sequence, computational algorithm, motif discovery tool

Introduction

Biological processes in prokaryotic and eukaryotic organisms are guided by genomic information in coding and non-coding DNA sequences. Both kinds of sequences coordinate the construction of transcriptional regulatory networks to perform gene expression with temporal-spatial variations. Compared with the pre-genomic era, which concentrated on deciphering coding DNA sequences and completed the blueprint of the human genome, the post-genomic era puts more emphases on digging the gold mine hidden in non-coding DNA sequences. Currently the identification of specific motifs or transcription factor binding sites (TFBSs) has become one of the key steps in this task.

As we all know, interaction between transcription factors (TFs) and non-coding DNA sequences is a prerequisite for transcription initiation of genes. The function of TFs is to recognize short conserved regions in non-coding DNA sequences, which are called motifs or TFBSs (*1*). However, it is not enough to find motifs or TFBSs in non-coding DNA sequences only depending on experimental methods. For example, systematic evolution of ligands by exponential enrichment (SELEX), serial analysis of gene expression (SAGE), and DNA microarray are only for transcript profiling *in vitro* (*1*, *2*). Chromatin immunoprecipitation (ChIP) can be combined with DNA microarray, namely ChIP-on-chip, to identify protein-DNA interaction *in vivo* (*3*), but it is limited by antibody performance and availability (*4*). For this reason, a wide range of motif discovery tools and databases have been applied to motif or TFBS prediction in biological studies. Unfortunately, 99.9% of their predictions are shown to be futility theorems (*5*).

Motifs or TFBSs are always represented as consensus IUPAC strings, position frequency matrices (PFMs), position weight matrices (PWMs), or position specific scoring matrices (PSSMs) in databases. Commonly, motifs or TFBSs in non-coding DNA sequences are conserved but still tend to be degenerate, which can influence the interaction between TFs and motifs or TFBSs. Therefore, after motif or TFBS data

***Corresponding author.**
E-mail: yuxd@nic.bmi.ac.cn

are collected and aligned from experimental or computational results, relevant consensus IUPAC strings can be constructed by selecting a degeneracy base pair symbol for each position in the alignment (*5*). The motif or TFBS data can also be modeled as PFM by aligning identified sites and counting the frequency of each base pair at each position of the alignment (*6*). Usually, PFM should be converted into PWM or PSSM according to formulas (*5*, *7*). Site scoring of non-coding DNA sequences can be calculated by computing the values for each position in PWM or PSSM model (*5*). Moreover, by using sequence logos, PWM can be displayed with color and height proportional to the base pair frequency and information content for each position by formulas (*8*).

In 1970s, scientists predicted that the pivotal difference between human and chimpanzee was located in non-coding DNA sequences rather than coding DNA sequences (*9*). Since then many essential elements of transcriptional regulatory networks have been identified in non-coding DNA sequences, including promoters, enhancers, insulators, silencers, and locus control regions (*6*). Nowadays, the discovery of motifs is mainly limited in canonical 5′ termini of known genes, where TFs are generally thought to bind in. Nevertheless, recently some researches have shown that only small proportion of motifs or TFBSs lie in immediate upstream sequences of well-characterized protein-coding genes, while the rest of them exist in either introns or 3′ regions (*6*, *10*, *11*).

A number of algorithms to discover motifs have been applied previously, for example, BE95 (*12*), KYD96 (*13*), DB97 (*14*), vHRCV00 (*15*), BJVU98 (*16*), EP20 (*17*), KFQW99 (*18*), and so on. However, many of these algorithms were designed for finding longer or more common motifs rather than for identifying TFBSs (*19*). The price paid for this generality is that many of the cited algorithms are not guaranteed to find globally optimal solutions, since they employ some forms of local search, such as Gibbs sampling, expectation maximization (EM), and phylogenetic algorithms. In this study, we give a brief introduction to the algorithm design and analysis for TFBSs with a focus on problems in comparative motif discovery.

Results and Discussion

Combinatorial approaches

Among the possible algorithmic approaches, combinatorial approaches try to exhaustively explore all the ways that a molecular process could happen. This leads to hard combinatorial problems for which efficient algorithms are required. Thus this kind of algorithms must make use of complex data representations and techniques.

Sequence-driven or Sample-driven (SD) algorithms

SD algorithms try to find comparative patterns by comparing the given length strings and looking for local similarities between them. They are based on constructing a local multiple alignment of the given non-coding DNA sequences and then extracting the comparative patterns from the alignment by combining the segments, which is common to most of the non-coding DNA sequences (*20*).

Pattern-driven (PD) algorithms

PD algorithms are based on enumerating candidate patterns in a given length string and inputting substrings with high fitness. The advantage of PD algorithms is that they can search the best comparative patterns in some limited sizes (*20*). Compared with SD algorithms, PD algorithms can be performed intelligently so that patterns are not present in the data that are not generated. For example, if a pattern α is not frequently present in the data, then there will be no frequent refinement that makes α more specific (hitting in even fewer places) in the data either (*20*).

Multiprofiler

This algorithm mainly utilizes multi-profiles that generalize a notion of a profile to detect subtle patterns that might escape detection by standard profiles (*21*). It is designed for finding particularly subtle motifs even in the case when real motifs may be blurred by random ones. The advantage of Multiprofiler is that it takes much less time (*21*). Kravchenko *et al* used Multiprofiler to search and statistically assess putative motifs in promoter regions of co-regulated genes, where the discovered over-represented sites could be totally verified by cell transfection experiments (*22*).

Consensus

This approach determines all possible pairwise alignments of matrices and remains words to create two sequence alignments. It scores the two sequence alignments by using information content, and the highest scoring will be saved (*23*). Each of the two sequence matrices is paired with each word that is not already

in the matrix, and then three sequence matrices are scored for information content, among which the highest will be kept again. This process will continue until each sequence has contributed exactly once to each saved alignment (*24*). In practice, Lenz *et al* scanned the upstream regions of the known *Vibrio cholerae* σ^{54}-regulated genes and obtained a 16 bp motif, which perfectly matches the known σ^{54} binding sites in *V. cholerae* with the consensus sequence "TGGCAC-N_5-TTGCA/T" (*25*). In another study, to prove the hypothesis that IL-2-regulated genes in T1 cells may be influenced by STAT5, Fung *et al* searched for motifs in 5,000 bp upstream regions by using the Consensus approach, and the obtained classic motif "TTCNNNGAA" can be verified by ChIP experiments (*26*).

Teiresias

Teiresias is a two-phase combinatorial approach for general pattern discovery. This algorithm assumes an instance that every motif is present in every sequence, namely, it finds all the maximal patterns with minimum support. Its performance scales quasi-linear sequences with the size of output (*27*). One property that differentiates Teiresias from other algorithms is the type of structural restriction. In this algorithm users are allowed to impose on special patterns to search. For example, only the parameter W needs to be set. It thus becomes possible to discover patterns of arbitrary length as long as preserved positions are not more than W residues away (*28*). In 2005, Kiesler *et al* scanned 23 *Hrp59* target exons by using Teiresias and found the known "GGAGG" core motif. This result was confirmed by ChIP, IP, and RT-PCR experiments, respectively (*29*).

Winnower, SP-STAR, and cWinnower

Winnower first represents motif instances as vertices, then it tries to delete spurious edges and recover motifs with the remaining vertices (*30*). SP-STAR is a local sum of pairwise score improvement algorithm, which considers only the subsequences present in dataset and iteratively updates scores of the motifs (*30*). cWinnower improves its running time by a stronger constraint function (*31*).

MobyDick

In some cases, motifs can be defined as strings whose probability of occurrence greatly exceeds the expectation of background. One problem is to decide which part constitutes the background and natural limits in a motif since large pieces of a motif will show up in a list of improbable strings. MobyDick can resolve this issue perfectly. It is suitable for discovering motifs from a large collection of sequences, for example, all of the upstream regions in the yeast genome or all of the genes regulated during sporulation (*32*). In 2003, based on two clusters of genes gained from microarray experiments, Murphy *et al* scanned 1,000 bp promoter regions of each gene in each cluster and found a motif "T(G/A)TTTAC", which had been previously validated to be bound by a known TF. Moreover, they found a new motif "CTTATCA" that may control gene transcription (*33*).

Smile, Verbumculus, and Weeder

The Smile algorithm takes into consideration the fact that TFBSs may be multiple and present a constrained spatial structure in genomes. Such algorithm is therefore able to identify genomic sequences that are called "structured motifs". A suffix tree is used for finding such motifs (*34*). The inner core of Verbumculus rests on subtly interwoven properties of statistics, pattern matching, and combinatorics on words. Thereby it is more feasible to both detect and visualize such words in a fast and practically useful way (*35*). Weeder permits to extend exhaustive enumeration to signals and does not need to input the exact length of the pattern to be found (*36*).

Mitra

Mitra can be extended to handle insertions and deletions in addition to mismatches in selected sequences. It takes advantage of a new insight, which prunes the patterns that allow for more efficient use of pairwise similarity than in Winnower. For example, unlike previous PD or SD algorithms, Mitra is able to discover composite motifs of a combined length over 30 bp (*37*).

Projection

This algorithm ameliorates the limitations of existing algorithms by using random projections of input. It extends previous projection-based searching techniques to solve a multiple alignment problem that is not effectively addressed by pairwise alignments. It is designed to efficiently solve the problems from the planted-(l, d) motif model, and can do more reliably and substantially difficult instances than previous algorithms (*38*). For t= 20 and n= 600, this algorithm

achieves performance close to the best possible, being limited primarily by statistical considerations (*38*).

EC and MoDEL

The evolutionary computation (EC) approach allows variation of motifs by the measurement of a similarity score. Compared with SD algorithms, which are not always easy to define and rely on the accuracy of PSSM, the EC approach does not rely on any pre-defined or estimated weight matrices (*39*, *40*). MoDEL uses a hybrid strategy consisting of an evolutionary algorithm (global search) and hill-climbing optimizations (local search) according to Brazma's classification (*41*). It addresses a well-known problem: given a set of functionally related sequences, how to choose exactly one occurrence per sequence in a way that all chosen occurrences are maximally similar. Such a set of occurrences will be referred to as ungapped local multiple alignment (*41*).

Probabilistic approaches

Probabilistic or randomized approaches make certain decisions randomly. This concept extends the classical model of deterministic algorithms and has generated many useful and probably efficient algorithms over the last twenty years. Probabilistic approaches are often faster, simpler, or more elegant than their combinatorial counterparts. Probabilistic algorithms that identify gene modules based on motif discovery are highly appropriate for analyzing synthetic lethal genetic interaction datasets, and have great potential in the integrative analysis of heterogeneous datasets (*42*).

EM

The EM algorithm is used to estimate the probability density of a given dataset by employing the Gaussian mixture model. The probability density of a dataset is modeled as the weighted sum of a number of Gaussian distributions. The main advantage of EM is its fast speed, while the disadvantage is that it requires "appropriate" starting values and is difficult to deal with constrained parameters (*43*).

Gibbs Sampler

The Gibbs sampling algorithm is one of the simplest Markov chain Monte Carlo algorithms. By Gibbs sampling, the joint distribution of the parameters will converge to the joint probability of the parameters in the given dataset. Gibbs sampling strategies claim to be fast and sensitive. It generally finds an optimized local alignment model for N sequences in N-linear time, avoiding the problem that the EM algorithm falls into. For example, it requires a relatively large dataset (15 or more sequences) for weakly conserved patterns to reach statistical significance (*44*). In 2000, Petersen *et al* tried to find motifs that are not necessarily 100% conserved in 17 putative promoter regions obtained from microarray experiments by using Gibbs Sampler (*45*). The search was performed in sequences ranging from 6 to 16 bp, where Gibbs Sampler repeatedly found motifs "TTGACT" and "GACTWWHC", both of which had been identified by previous experiments.

MEME

The MEME algorithm extends the EM algorithm for identifying motifs in unaligned sequences. While a drawback of EM is that the maximum it finds is only local (*46*), MEME can either favor motifs that appear exactly once (one-per model) or appear zero or once (zero-or-one-per model) in each sequence in a training set, or give no preference to a number of occurrences (zero-or-more-per model). In 2005, Hall *et al* acquired a set of correlated genes from genomic, transcriptomic, and proteomic analyses. They applied MEME to scan 1,000 bp of the $3'$ end of stop codon, where a 47 bp motif was found in six of the analyzed sequences. Then it was used to search the entire genome and 20 additional genes were identified to have the same motif. This motif was known to be bound to Puf protein, implying that Puf protein may control the transcription of the analyzed genes (*47*).

LOGOS and MotifPrototyper

LOGOS consists of two interacting submodels: HMDM, a model for aligned selected sequences, and HMM, a model for the global distribution of motif instances. HMDM is a hidden Markov-Dirichlet multinomial model that captures rich biological prior knowledge and positional dependence in motif local structure in a principled way. HMM is a standard hidden Markov model, which allows formal and efficient inference of motif locations, and is potentially capable of capturing their dependencies. Model parameters can be fit on training motifs by using a variational EM algorithm within an empirical Bayesian framework (*48*). MotifPrototyper is later used to train the model's parameters and to scan for known regulatory motifs and discover unknown ones (*49*).

Motif Sampler

Motif Sampler uses higher-order Markov models to represent the intergenic motifs in non-coding DNA sequences. It can incorporate higher-order background models to update probabilities of finding a motif at a certain position (*50*). To search for a known TF *Yrrp1* consensus binding site in yeast, Le Crom *et al* used Motif Sampler to search for motifs in the genes regulated by *Yrrp1*, and the result motif "(T/A)CCG(C/T)(G/T)(G/T)(A/T)(A/T)" was confirmed by EMSA experiments (*51*).

AlignACE

AlignACE is based on the Gibbs sampling algorithm, but it differs from Gibbs sampling in the following ways. Firstly, the motif model is changed so that base frequencies for non-site sequences are fixed according to the source genome. Secondly, both strands of input sequences are simultaneously considered at each step of this algorithm. Overlapping sites are not allowed even if these sites are on opposite strands. Thirdly, simultaneous multiple searching is replaced by an approach in which single motif is found and iteratively masked (*52–54*).

ANN-Spec

The objective function for ANN-Spec is designed to find patterns that distinguish the positive dataset from background. It succeeds in identifying the desired patterns specific for the positive dataset. For example, Gibbs sampling and ANN-Spec both work very well when the background is assumed to be random, while ANN-Spec finds patterns with higher specificity and higher correlation coefficients when it is provided with background sequences (*55*, *56*).

BioProspector

BioProspector uses the Markov background to model base dependencies of non-motif bases, which greatly improves the specificity of reported motifs. The parameters of the Markov background model are either estimated from user-specified sequences or precomputed from the whole genome. A new motif scoring function is adopted to allow each input sequence contain zero to multiple copies of the motif. In addition, BioProspector can model gapped motifs with palindromic patterns, which are prevalent motif patterns in prokaryotes (*57*, *58*).

MDscan and Motif Regressor

MDscan mainly examines ChIP-on-chip selected sequences. It combines the advantages of two widely adopted motif search strategies, word enumeration and PSSM, and incorporates ChIP enrichment information to accelerate the searching and enhance its success rate. Motif Regressor uses linear regression analysis to select motifs whose sequence matching scores are significantly correlated with ChIP-on-chip enrichment or downstream gene expression values. Ranking motifs by linear regression p-value, Motif Regressor automatically picks the best one with optimal width (*59–61*).

Improbizer

Improbizer searches for motifs that occur with improbable frequency by using a variation of the EM algorithm. It works by finding the patterns that occur more frequently than they should occur by chance. The simple way to estimate how frequently a particular nucleotide should occur by chance is to put one quarter to the power of the number of nucleotides in the sequence. Optionally, Improbizer constructs a Gaussian model of motif placement, so that motifs occurring in similar positions in the input sequences are more likely to be found (*62*).

SeSiMCMC

SeSiMCMC is a tool for multiple local alignment of a set of non-coding DNA sequences, which is based on a modification of the Gibbs sampling algorithm. Its primary objective is to create a computationally efficient tool that uses user-defined motif symmetry and evaluates motif length from dataset. Sequence fragments in a training set can have arbitrary orientation, and there is a probability for a sequence to contain no sites (*63*).

GMS-MP

GMS-MP performs significantly better than standard PWM-based Gibbs sampling methods. Compared with the Bayesian network approach, GMS-MP has a simpler model, easier prescribing prior, and much faster computation. The step of sampling pairwise correlations takes up only about 3% of the total computing time, which is much faster than the Bayesian network. This method also does not suffer any problems with over-fitting, which is likely to occur due to the employment of a rather conservative prior distribution on model pattern (*64*).

Phylogenetic footprinting approaches

Phylogenetic footprinting approaches discover regulatory elements in a set of orthologous regulatory regions from multiple species by identifying the best conserved motifs in those orthologous regions (*65*).

PhyloCon

Phylogenetic-Consensus (PhyloCon) takes into account both conserved orthologous genes and co-regulated genes within a species. The key idea of PhyloCon is to compare aligned sequence profiles from orthologous genes or co-regulated genes rather than unaligned sequences. PhyloCon integrates the knowledge of co-regulated genes in single species with sequence conservation across multiple species to improve the performance of motif discovery. An advantage of PhyloCon is that it reports motifs of varying lengths, instead of requiring the motif length to be input (*66*, *67*).

EMnEM and OrthoMeme

Expectation-maximization on evolutionary mixtures (EMnEM) considers special motifs that are generated from ancestral sequences. The ancestral sequences are made of two component mixtures of motifs and background, each with their own evolutionary model. The value of varying evolutionary models has been realized in other contexts as well, and such models have been successfully trained by using EM. Normally, MEME often scores better than EMnEM with a substitution model, except for higher evolutionary distances, where EMnEM takes the head (*68*). OrthoMeme is the first algorithm to deal with heterogeous data sources in a truly integrated manner by using all the data from onset of analysis (*69*).

PhyME

PhyME integrates two different axes of information content in evaluating the significance of candidate motifs. One axis is the overrepresentation that depends on the number of occurrences of motifs in each species. The other axis is the level of conservation of each motif instance across species. An important feature of PhyME is that it allows motifs to occur in evolutionarily conserved as well as unconserved regions in orthologous sequences. PhyME treats the two kinds of occurrences differently when it scores a motif (*70*).

FootPrinter

The unique character of FootPrinter is that it takes input as a set of unaligned homologous sequences from various species and elicits a phylogenetic tree relating to these species. It then searches for short regions of the sequences that are highly conserved according to a parsimony criterion. The regions identified will be good candidates for regulatory elements (*71*).

CompareProspector

CompareProspector identifies regulatory elements by using information content from both intraspecies pattern enrichment and interspecies sequence conservation. This distinguishes it from other phylogenetic footprinting programs that use orthologous sequences of a single gene from multiple species to identify regulatory elements (*44*).

Conclusion

In the last decade, computational identification of motifs or TFBSs by analyzing non-coding DNA sequences has emerged as a major new technology for elucidating transcriptional regulatory networks. Combinatorial algorithms assume a discrete model and search for motifs with a high rate of occurrences in non-coding DNA sequences. One major drawback of combinatorial algorithms is that they are sometimes difficult to understand and many "hidden" details make them hard to implement. Probabilistic algorithms often run faster than their corresponding combinatorial algorithms. Moreover, many probabilistic algorithms are easier to implement and describe than combinatorial algorithms of comparable performance. However, these algorithms may miss lots of useful information when searching in non-coding DNA sequences. Phylogenetic footprinting assumes that functional sequences tend to be conserved through evolution. Motifs or TFBSs can thus be identified by looking for conservation of small regions within multiple alignments of non-coding DNA sequences.

Up to date, more than 120 motif discovery tools have been applied in biological researches. All the time the main challenge of motif discovery tools has been the application of effective algorithms that can treat all the intrinsic complexities associated with the nature of motifs or TFBSs. However, there still exist some considerations that we should bear in mind when thinking of computational approaches to tackle biological problems. One is the issue of futility the-

orem, which means we still do not have any good methods other than traditional molecular biology to find out whether our predictions of individual motif or TFBS have any relationships with a clear function *in vivo*. Another is that pattern discovery methods are severely restricted by the signal-to-noise problem, because the information content of motifs is strictly limited by its intrinsic nature. Additionally, some algorithms that work well for yeast might not work for human due to the complexity of DNA structure. Therefore, all observed patterns must be carefully considered.

Materials and Methods

Web-based resources for non-coding DNA sequence datasets

The non-coding DNA sequence dataset perspectives in web-based resources give the tools for biologists to work with relational experimental researches in their application development. The relational dataset tools include views, wizards, editors, and other features that make it easy for users to predict and test the experimental elements of their applications (partially in Table 1).

Web-based resources for regulatory motif or TFBS datasets

The relational motif or TFBS datasets help biologists create and manipulate the data definitions for their own projects, in terms of relational dataset schemas. Users can access relational motif or TFBS datasets under the analysis perspective, which allows users to browse or import dataset schemas in the servers view, create and work with dataset schemas in the data definition view, and change dataset schemas in the table editor. Users can also export data definitions to another dataset installed either locally or remotely (partially in Table 2).

Table 1 Selected web-based resources for promoter databases

Database	Explanation	URL
EPD	Eukaryotic promoter database	http://www.epd.isb-sib.ch/
DBTSS	Database of transcriptional start sites (human)	http://dbtss_old.hgc.jp/hg17/
SCPD	*Saccharomyces cerevisiae* promoter database	http://rulai.cshl.edu/SCPD/
DCPD	*Drosophila* core promoter database	http://www-biology.ucsd.edu/labs/Kadonaga/DCPD.html
PlantProm DB	Plant promoter database	http://mendel.cs.rhul.ac.uk/mendel.php?topic=plantprom
CSHLmpd	Cold Spring Harbor Laboratory mammalian promoter database	http://rulai.cshl.edu/CSHLmpd2/
TRED	Transcriptional regulatory element database	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home

Table 2 Selected web-based resources for regulatory motifs or TFBSs

Database	Explanation	URL
JASPAR	A collection of transcription factor DNA-binding preferences	http://mordor.cgb.ki.se/cgi-bin/jaspar2005/jaspar_db.pl
TRANSFAC	Database on eukaryotic transcription factors, their genomic binding sites and DNA-binding profiles	http://www.gene-regulation.com/pub/databases.html#transfac
TRRD	Transcription regulatory regions database	http://wwwmgs.bionet.nsc.ru/mgs/gnw/
RegulonDB	A computational model of mechanisms of transcriptional regulation	http://regulondb.ccg.unam.mx/html/What_is_RegulonDB.jsp
TFD	Transcription factor databases	http://www.ifti.org/

Web-based resources for motif or TFBS discovery algorithms

Emphases are placed on the development of general design algorithms and data structures that are particularly suited for biological problems. Applications in a variety of areas such as genetic information systems, computer graphics, alignments, and computer aided designs are performed (partially in Table 3).

Table 3 Selected web-based resources for motif discovery tools

Algorithm	Motif model	Match model	Ref.
AlignACE	matrix	PWM	*52*
ANN-Spec	matrix	PWM	*55*
BioOptimizer	–	PWM	*72*
BioProspector	matrix, dyad	PWM	*57*
CAGER	–	–	*73*
Cis-analyst	–	PWM	*74*
CisModule	–	PWM	*75*
Cister	–	PWM	*76*
Clover	–	PWM	*77*
ClusterScan	–	PWM	*78*
CoBind	matrix, dyad	PWM	*79*
COMET	–	–	*80*
CompareProspector	–	–	*57*
ConsecID	–	PWM	*81*
Consensus	matrix	PWM	*24*
ConSite	–	PWM	*82*
COOP	–	reg.exp	*83*
cWinnower	string	mismatch	*31*
DMotifs	string	reg.exp	*84*
DMS	–	PWM	*85*
Dyad analysis	string, dyad	oligos	*15*
EC	string	fitness	*39*
EMnEM	–	–	*68*
FastM	–	PWM	*18*
FootPrinter	–	mismatch	*71*
FrameWorker	–	PWM	*86*
GANN	–	flexible	*87*
Gibbs sampler	matrix	PWM	*44*
Gibbs recursive	matrix	PWM	*88*
GLAM	string	–	*89*
GMS-MP	GWM	HMM	*64*
HMDM	–	DM	*90*
Improbizer	–	PWM	*62*
LOGOS	HMDM	DM	*48*
MAPPER	–	HMM	*91*
MCAST	–	PWM	*92*
MDScan	matrix	PWM	*59*
MEME	matrix	PWM	*46*
MERMAID	string	PWM	*93*
MISAE	–	mismatch	*94*

Algorithm	Motif model	Match model	Ref.
Mitra	string, dyad	mismatch	*48*
Mitra-dyad	–	mismatch	*17*
MITRA-PSSM	matrix	PWM	*95*
MM	–	PWM	*96*
MobyDick	string	mismatch	*32*
MoDEL	string	PWM	*41*
ModelGenerator	–	PWM	*97*
ModelInspector	–	PWM	*97*
Modulescanner	–	PWM	*98*
ModuleSearcher	–	PWM	*98*
MotifLocator	–	PWM	*98*
MotifPrototyper	–	DM	*49*
Motif regressor	–	PWM	*41*
Motif sampler	–	PWM	*50*
MSCAN	–	PWM	*99*
MultiProfiler	string	mismatch	*21*
NestedMICA	–	PWM	*100*
NONPAR	–	mixture	*101*
Oligo-analysis	string	oligos	*102*
OrthoMEME	–	PWM	*69*
Pattern-assembly	–	–	*103*
PhyloCon	–	PWM	*66*
PhyME	–	–	*70*
Pratt2	–	reg.exp	*104*
Projection	string	PWM	*38*
ProMapper	–	DM	*105*
PromoterInsp	–	oligos	*106*
QuickScore	string	IUPAC	*107*
REDUCE	–	PWM	*108*
SCORE	–	–	*109*
SeSiMCMC	–	PWM	*63*
SMILE	string, mult	mismatch	*34*
SOMBERO	–	PWM	*110*
Splash	–	reg.exp	*111*
Stubb	–	PWM	*42*
Teiresias	string	reg.exp	*27*
TFBScluster	–	PWM	*112*
Verbumculus	string	mismatch	*35*
Weeder	string	mismatch	*36*
Winnower	string	mismatch	*30*
YMF	string	reg.exp	*113*

Acknowledgements

We thank Mr. Maximilian Häußler and Dr. Saurabh Sinha for providing their theses, Prof. Finn Drabløs for his precious documents, Dr. Zhiping Weng and Prof. Michael Q. Zhang for their helpful websites, and Dr. Jinkuk Choi for his useful critical review. Especially we thank Dr. Aiguo Li of the National Cancer Institute in NIH for her helpful advice on the manuscript.

Authors' contributions

WW carried out the study, and YX supervised the research. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Roulet, E., *et al.* 2002. High-throughput SELEX-SAGE method for quantitative modeling of transcription-factor binding sites. *Nat. Biotechnol.* 20: 831-835.
2. van Steensel, B. 2005. Mapping of genetic and epigenetic regulatory networks using microarrays. *Nat. Genet.* 37: S18-24.
3. Cam, H., *et al.* 2004. A common set of gene regulatory networks links metabolism and growth inhibition. *Mol. Cell* 16: 399-411.
4. Blais, A. and Dynlacht, B.D. 2005. Constructing transcriptional regulatory networks. *Genes Dev.* 19: 1499-1511.
5. Wasserman, W.W. and Sandelin, A. 2004. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 5: 276-287.
6. Vavouri, T. and Elgar, G. 2005. Prediction of *cis*-regulatory elements using binding site matrices—the successes, the failures and the reasons for both. *Curr. Opin. Genet. Dev.* 15: 395-402.
7. Staden, R. 1984. Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12: 505-519.
8. Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 18: 6097-6100.
9. King, M.C. and Wilson, A.C. 1975. Evolutions at two levels in humans and chimpanzees. *Science* 188: 107-116.
10. Cawley, S., *et al.* 2004. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116: 499-509.
11. Impey, S., *et al.* 2004. Defining the CREB regulon: a genome-wide analysis of transcription factor regulatory regions. *Cell* 119: 1041-1054.
12. Tompa, M., *et al.* 2005. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23: 137-144.
13. Kirchhamer, C.V., *et al.* 1996. Modular *cis*-regulatory organization of developmentally expressed genes: two genes transcribed territorially in the sea urchin embryo, and additional examples. *Proc. Natl. Acad. Sci. USA* 93: 9322-9328.
14. Pennacchio, L.A. and Rubin, E.M. 2001. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* 2: 100-109.
15. van Helden, J., *et al.* 2000. Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucleic Acids Res.* 28: 1808-1818.
16. Brazma, A., *et al.* 1998. Predicting gene regulatory elements *in silico* on a genomic scale. *Genome Res.* 8: 1202-1215.
17. Eskin, E., *et al.* 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 18: S354-363.
18. Klingenhoff, A., *et al.* 1999. Functional promoter modules can be detected by formal models independent of overall nucleotide sequence similarity. *Bioinformatics* 15: 180-186.
19. Gusfield, D. 1997. *Algorithms on Strings, Trees and Sequences.* Cambridge University Press, Cambridge, UK.
20. Ohler, U. and Niemann, H. 2001. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet.* 17: 56-60.
21. Keich, U. and Pevzner, P.A. 2002. Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics* 18: 1382-1390.
22. Kravchenko, J.E., *et al.* 2005. Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin. *Nature* 436: 735-739.
23. Stormo, G.D. and Hartzell, G.W.III. 1989. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. USA* 86: 1183-1187.
24. Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15: 563-577.
25. Lenz, D.H., *et al.* 2004. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. *Cell* 118: 69-82.
26. Fung, M.M., *et al.* 2005. IL-2- and STAT5-regulated cytokine gene expression in cells expressing the Tax

protein of HTLV-1. *Oncogene* 24: 4624-4633.
27. Rigoutsos, I. and Floratos, A. 1998. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. *Bioinformatics* 14: 55-67.
28. Jensen, K.L., *et al.* 2006. A generic motif discovery algorithm for sequential data. *Bioinformatics* 22: 21-28.
29. Kiesler, E., *et al.* 2005. Hrp59, an hnRNP M protein in *Chironomus* and *Drosophila*, binds to exonic splicing enhancers and is required for expression of a subset of mRNAs. *J. Cell Biol.* 168: 1013-1025.
30. Pevzner, P.A. and Sze, S.H. 2000. Combinatorial approaches to finding subtle signals in DNA sequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8: 269-278.
31. Liang, S., *et al.* 2004. cWINNOWER algorithm for finding fuzzy DNA motifs. *J. Bioinform. Comput. Biol.* 2: 47-60.
32. Bussemaker, H.J., *et al.* 2000. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis. *Proc. Natl. Acad. Sci. USA* 97: 10096-10100.
33. Murphy, C.T., *et al.* 2003. Genes that act downstream of DAF-16 to influence the lifespan of *Caenorhabditis elegans*. *Nature* 424: 277-283.
34. Marsan, L. and Sagot, M.F. 2000. Algorithms for extracting structured motifs using a suffix tree with an application to promoter and regulatory site consensus identification. *J. Comput. Biol.* 7: 345-362.
35. Apostolico, A., *et al.* 2004. Verbumculus and the discovery of unusual words. *J. Comput. Sci. Technol.* 19: 22-41.
36. Pavesi, G., *et al.* 2001. An algorithm for finding signals of unknown length in DNA sequences. *Bioinformatics* 17: S207-214.
37. Eskin, E. and Pevzner, P.A. 2002. Finding composite regulatory patterns in DNA sequences. *Bioinformatics* 18: S354-363.
38. Buhler, J. and Tompa, M. 2002. Finding motifs using random projections. *J. Comput. Biol.* 9: 225-242.
39. Fogel, G.B., *et al.* 2004. Discovery of sequence motifs related to coexpression of genes using evolutionary computation. *Nucleic Acids Res.* 32: 3826-3835.
40. Gertz, J., *et al.* 2005. Discovery, validation, and genetic dissection of transcription factor binding sites by comparative and functional genomics. *Genome Res.* 15: 1145-1152.
41. Hernandez, D., *et al.* 2004. MoDEL: an efficient strategy for ungapped local multiple alignment. *Comput. Biol. Chem.* 28: 119-128.
42. Sinha, S., *et al.* 2003. A probabilistic method to detect regulatory modules. *Bioinformatics* 19: i292-301.
43. Moon, T.K. 1996. The expectation-maximization algorithm. *IEEE Signal Proc. Mag.* 13: 47-60.
44. Lawrence, C.E., *et al.* 1993. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* 262: 208-214.
45. Petersen, M., *et al.* 2000. *Arabidopsis* MAP kinase 4 negatively regulates systemic acquired resistance. *Cell* 103: 1111-1120.
46. Bailey, T.L. and Gribskov, M. 1998. Methods and statistics for combining motif match scores. *J. Comput. Biol.* 5: 211-221.
47. Hall, N., *et al.* 2005. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* 307: 82-86.
48. Xing, E.P., *et al.* 2004. LOGOS: a modular Bayesian model for *de novo* motif detection. *J. Bioinform. Comput. Biol.* 2: 127-154.
49. Xing, E.P. and Karp, R.M. 2004. MotifPrototyper: a Bayesian profile model for motif families. *Proc. Natl. Acad. Sci. USA* 101: 10523-10528.
50. Thijs, G., *et al.* 2001. A higher-order background model improves the detection of regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113-1122.
51. Le Crom, S., *et al.* 2002. New insights into the pleiotropic drug resistance network from genome-wide characterization of the YRR1 transcription factor regulation system. *Mol. Cell. Biol.* 22: 2642-2649.
52. Roth, F.P., *et al.* 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16: 939-945.
53. Wade, J.T., *et al.* 2004. The transcription factor Ifh1 is a key regulator of yeast ribosomal protein genes. *Nature* 432: 1054-1058.
54. Wade, J.T., *et al.* 2005. Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia coli* genome and identifies unconventional target sites. *Genes Dev.* 19: 2619-2630.
55. Workman, C.T. and Stormo, G.D. 2000. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac. Symp. Biocomput.* 467-478.
56. GuhaThakurta, D., *et al.* 2002. Identification of a novel *cis*-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Res.* 12: 701-712.
57. Liu, X., *et al.* 2001. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* 127-138.
58. Mukherjee, S., *et al.* 2004. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* 36: 1331-1339.
59. Liu, X.S, *et al.* 2002. An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.* 20: 835-839.
60. Carroll, J.S., *et al.* 2005. Chromosome-wide mapping of estrogen receptor binding reveals long-range regulation requiring the forkhead protein FoxA1. *Cell* 122: 33-43.

61. Ben, Y.S., *et al.* 2005. Defining a centromere-like element in *Bacillus subtilis* by identifying the binding sites for the chromosome-anchoring protein RacA. *Mol. Cell* 17: 773-782.
62. Ao, W., *et al.* 2004. Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science* 305: 1743-1746.
63. Favorov, A.V., *et al.* 2005. A Gibbs sampler for identification of symmetrically structured, spaced DNA motifs with improved estimation of the signal length. *Bioinformatics* 21: 2240-2245.
64. Zhou, Q. and Liu, J.S. 2004. Modeling within-motif dependence for transcription factor binding site predictions. *Bioinformatics* 20: 909-916.
65. Blanchette, M., *et al.* 2002. Algorithms for phylogenetic footprinting. *J. Comput. Biol.* 9: 211-223.
66. Wang, T. and Stormo, G.D. 2003. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* 19: 2369-2380.
67. Hu, Y., *et al.* 2004. RNA interference of achaete-scute homolog 1 in mouse prostate neuroendocrine cells reveals its gene targets and DNA binding sites. *Proc. Natl. Acad. Sci. USA* 101: 5559-5564.
68. Moses, A.M., *et al.* 2004. Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac. Symp. Biocomput.* 324-335.
69. Prakash, A., *et al.* 2004. Motif discovery in heterogeneous sequence data. *Pac. Symp. Biocomput.* 348-359.
70. Sinha, S., *et al.* 2004. PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics* 5: 170.
71. Blanchette, M. and Tompa, M. 2003. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.* 31: 3840-3842.
72. Jensen, S.T. and Liu, J.S. 2004. BioOptimizer: a Bayesian scoring function approach to motif discovery. *Bioinformatics* 20: 1557-1564.
73. Ruan, J. and Zhang, W. 2005. CAGER: classification analysis of gene expression regulation using multiple information sources. *BMC Bioinformatics* 6: 114.
74. Berman, B.P., *et al.* 2002. Exploiting transcription factor binding site clustering to identify *cis*-regulatory modules involved in pattern formation in the *Drosophila* genome. *Proc. Natl. Acad. Sci. USA* 99: 757-762.
75. Zhou, Q. and Wong, W.H. 2004. CisModule: *de novo* discovery of *cis*-regulatory modules by hierarchical mixture modeling. *Proc. Natl. Acad. Sci. USA* 101: 12114-12119.
76. Frith, M.C., *et al.* 2001. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* 17: 878-889.
77. Frith, M.C., *et al.* 2004. Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res.* 32: 1372-1381.
78. Kel-Margoulis, O.V., *et al.* 2002. Automatic annotation of genomic regulatory sequences by searching for composite clusters. *Pac. Symp. Biocomput.* 187-198.
79. GuhaThakurta, D. and Stormo, G.D. 2001. Identifying target sites for cooperatively binding factors. *Bioinformatics* 17: 608-621.
80. Frith, M.C., *et al.* 2002. Statistical significance of clusters of motifs represented by position specific scoring matrices in nucleotide sequences. *Nucleic Acids Res.* 30: 3214-3224.
81. Sharan, R., *et al.* 2003. CREME: a framework for identifying *cis*-regulatory modules in human-mouse conserved segments. *Bioinformatics* 19: i283-291.
82. Sandelin, A., *et al.* 2004. ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic Acids Res.* 32: W249-252.
83. Bortoluzzi, S., *et al.* 2005. A multistep bioinformatic approach detects putative regulatory elements in gene promoters. *BMC Bioinformatics* 6: 121.
84. Sinha, S. 2003. Discriminative motifs. *J. Comput. Biol.* 10: 599-615.
85. Hu, Y.J., *et al.* 2000. Combinatorial motif analysis and hypothesis generation on a genomic scale. *Bioinformatics* 16: 222-232.
86. Cartharius, K., *et al.* 2005. MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics* 21: 2933-2942.
87. Beiko, R.G. and Charlebois, R.L. 2005. GANN: genetic algorithm neural networks for the detection of conserved combinations of features in DNA. *BMC Bioinformatics* 6: 36.
88. Thompson, W., *et al.* 2003. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.* 31: 3580-3585.
89. Frith, M.C., *et al.* 2004. Finding functional sequence elements by multiple local alignment. *Nucleic Acids Res.* 32: 189-200.
90. Xing, E.P., *et al.* 2002. A hierarchical Bayesian Markovian model for motifs in biopolymer sequences. In *Advances in Neural Information Processing Systems* (eds. Becher, S., *et al.*). Vol. 16. MIT Press, Cambridge, USA.
91. Marinescu, V.D., *et al.* 2005. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics* 6: 79.
92. Bailey, T.L. and Noble, W.S. 2003. Searching for statistically significant regulatory modules. *Bioinformatics* 19: ii16-25.
93. Hu, Y.J. 2003. Finding subtle motifs with variable gaps in unaligned DNA sequences. *Comput. Methods Programs Biomed.* 70: 11-20.
94. Sun, Z., *et al.* 2004. MISAE: a new approach for regulatory motif extraction. *Proc. IEEE Comput. Syst. Bioinform. Conf.* 173-181.

95. Leung, H.C. and Chin, F.Y. 2005. Finding exact optimal motifs in matrix representation by partitioning. *Bioinformatics* 21: ii86-92.
96. Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 28-36.
97. Frech, K. and Werner, T. 1997. Specific modelling of regulatory units in DNA sequences. *Pac. Symp. Biocomput.* 151-162.
98. Aerts, S., *et al.* 2003. Computational detection of *cis*-regulatory modules. *Bioinformatics* 19: ii5-14.
99. Alkema, W.B., *et al.* 2004. MSCAN: identification of functional clusters of transcription factor binding sites. *Nucleic Acids Res.* 32: W195-198.
100. Down, T.A. and Hubbard, T.J. 2005. NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic Acids Res.* 33: 1445-1453.
101. King, O.D. and Roth, F.P. 2003. A non-parametric model for transcription factor binding sites. *Nucleic Acids Res.* 31: e116.
102. van Helden, J., *et al.* 1998. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281: 827-842.
103. Narasimhan, C., *et al.* 2003. Background rareness-based iterative multiple sequence alignment algorithm for regulatory element detection. *Bioinformatics* 19: 1952-1963.
104. Jonassen, I. 1997. Efficient discovery of conserved patterns using a pattern graph. *Comput. Appl. Biosci.* 13: 509-522.
105. Pudimat, R., *et al.* 2005. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics* 21: 3082-3088.
106. Scherf, M., *et al.* 2000. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. *J. Mol. Biol.* 297: 599-606.
107. Boeva, V., *et al.* 2006. Short fuzzy tandem repeats in genomic sequences, identification, and possible role in regulation of gene expression. *Bioinformatics* 22: 676-684.
108. Bussemaker, H.J., *et al.* 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27: 167-171.
109. Rebeiz, M., *et al.* 2002. SCORE: a computational approach to the identification of *cis*-regulatory modules and target genes in whole-genome sequence data. Site clustering over random expectation. *Proc. Natl. Acad. Sci. USA* 99: 9888-9893.
110. Mahony, S., *et al.* 2005. Transcription factor binding site identification using the self-organizing map. *Bioinformatics* 21: 1807-1814.
111. Hart, R.K., *et al.* 2000. Systematic and fully automated identification of protein sequence patterns. *J. Comput. Biol.* 7: 585-600.
112. Donaldson, I.J., *et al.* 2005. TFBScluster: a resource for the characterization of transcriptional regulatory networks. *Bioinformatics* 21: 3058-3059.
113. Sinha, S. and Tompa, M. 2003. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* 31: 3586-3588.