

Motif GibbsGA: Sampling Transcription Factor Binding Sites Coupled with PSFM Optimization by Genetic Algorithm

LIU Li-fang^{1,2} JIAO Li-cheng¹

(1 School of Computer Science and Technology, Xidian University, Xi'an 710071, China

2 Institute of Intelligent Information Processing, Xidian University, Xi'an 710071, China

E-mail: xdlilifang@gmail.com

doi:10.4156/jcit.vol5.issue10.18

Abstract

Identification of transcription factor binding sites (TFBSs) or motifs plays an important role in deciphering the mechanisms of gene regulation. Although many experimental and computational methods have been developed, finding TFBSs remains a challenging problem. We propose and develop a novel sampling based motif finding method coupled with PSFM optimization by genetic algorithm, which we call Motif GibbsGA. One significant feature of Motif GibbsGA is the combination of a Gibbs sampling method and a PSFM optimization by genetic algorithm. Based on position-specific frequency matrix (PSFM) motif model, a greedy strategy for choosing the initial parameters of PSFM is employed. Then a Gibbs sampler is build with respect to PSFM model. During the sampling process, PSFM is improved via a genetic algorithm. A post-processing with adaptive adding and removing is used to handle general cases with arbitrary numbers of instances per sequence. So Motif GibbsGA is capable of discovering several different motifs with differing numbers of occurrences in a single dataset. We test our method on the benchmark dataset compiled by Tompa et al. (2005) for assessing computational tools that predict TFBSs. The performance of Motif GibbsGA on this data set compares well to, and in many cases exceeds, the performance of existing tools. This is in part attributed to the significant role played by the genetic algorithm that improved PSFM.

Keywords: Motif Discovery, Gibbs Sampling, Binding Sites, Transcription Factors, Genetic Algorithm

1. Introduction

The rest of paper is organized as follows. In Section 2, we will basic thinking of our proposition. Section 3 gives the specific steps of this algorithm and Section 4 reveals the simulation result of DGAF and compares it with traditional GAF.

In the post-genomic era, identifying regulatory elements is an important step to understanding the mechanisms of gene regulation. Transcription factor binding sites (TFBSs) contribute substantially to the control of gene expression, and because of their biological importance, much experimental effort has expended in identifying them. Over the past few years, there are many computational tools that identify TFBSs as the sub-sequences, or motifs, common to a set of sequences. The difference from each other chiefly in their definition of what constitutes a motif, what constitutes statistical overrepresentation of a motif and the method used to find statistically overrepresented motifs.

Nearly all motif discovery algorithms fall into three general classes: pattern-based, profile-based and combinatorial. In profile-based algorithms, a motif is usually modeled by a $4 \times W$ position-specific frequency matrix (PSFM), where W is the motif's size in base pairs, so that each column of the PSFM represents the distribution of the 4nt at the corresponding position in the motif. One way of the PSFM estimating can be through standard statistical learning theory methods, such as maximum-likelihood estimation (e.g. MEME [1], The Improbizer [2]), the Markov chain Monte Carlo algorithms (e.g. AlignACE [3], BioProspector [4], MotifSampler [5], GLAM [6], DSMC [7]), greedy search (e.g. Consensus [8]), and genetic algorithms (e.g. GAME [9], GALF-P [10]). Another way of finding shared motifs is to compile a library of motifs which previously characterized or randomized, and assess whether any of these motifs are statistically over-represented in the sequences (e.g. Clover [11], SOMBRERO [12]). Library-based method has declared improved performance. However the existing tools are still not effective for discovering motifs [13,14]. For example, as shown in paper [13]'s

evaluation, even the best performing algorithm has sensitivity <9% and precision <30% (sensitivity is percentage of true nucleotides that are predicted and precision is the percentage of predicted nucleotides that are true). To deal with this issue, a class of algorithms called ensemble methods has been proposed (e.g. MotifVoter [15]). Though the existing ensemble methods overall perform better than stand-alone motif finders, the improvement gained is not substantial.

In our previous experiments, we observed that Gibbs sampling strategy tends to become too inefficient to identify the binding sites correctly for long input sequences. The results suggest a need for improving scalability of sampling based discovery algorithm, which is particularly important when motifs are sought from an increasing number of complete genome sequences. One way to tackle this problem is to import stronger global optimization techniques, such as genetic algorithms and others. So-called hybrid algorithms assemble the Gibbs sampling and global optimization algorithms to build a stronger algorithm, and extensive experiments are needed to evaluate the algorithm's performance. In this work, we present the Motif GibbsGA approach - a novel de novo motif identification approach inspired by Leping Li et al. (2007) [16]. Motif GibbsGA is based on Gibbs sampling coupled with PSFM optimization by genetic algorithm. Based on position position-specific frequency (PSFM) motif model, a greedy strategy for choosing the initial parameters of PSFM is employed. Then a Gibbs sampler is build with respect to PSFM model. During the sampling process, PSFM is improved via a genetic algorithm. A post-processing with adaptive adding and removing is used to handle general cases with arbitrary numbers of instances per sequence. We test Motif GibbsGA on the benchmark dataset compiled by Tompa et al. (2005) for assessing computational tools that predict TFBSs. The performance of Motif GibbsGA on this data set compares well to, and in many cases exceeds, the performance of existing tools.

2. Materials and Methods

2.1. Basic matrix models

By far the most common representation of a motif is a position-specific frequency matrix (PSFM) θ_W of length W . Each entry in this matrix gives the probability $p_{b,j}$ of finding a given base b at position j in the binding site, such that $\sum_{i=1}^{|\Sigma|} p_{i,j} = 1$, Σ is a finite alphabet. For a motif instances (substring) $s=(s_1, s_2, \dots, s_W)$ of length W in a sequence, the probability Q_s of x given the motif model θ_W is

$$Q_s = P(s_1 \cdots s_W | \theta_W) = \prod_{i=1}^{i=W} p_{s_i, i} \quad (1)$$

Except for the motif instances, the remainder of the sequences is classified as nonsites, where each base is generated according to a specific background model. The 0-order background model is $B_0 = [q_{A,0}, q_{C,0}, q_{G,0}, q_{T,0}]$, this means that each base is drawn independently from a single discrete distribution. The m -order background model is B_m , this means that the probability of finding a certain nucleotide in a sequence depends on the m previous nucleotides in the sequence. For a segment s in a sequence, the probability P_s of segment s being generated by the background model B_m is given by

$$P_s = P(s_1 \cdots s_W | B_m) = \prod_{i=1}^W P(s_i | s_{i-1}, \dots, s_{i-m}) \cdot \quad (2)$$

Corresponding to PSFM, the position weight matrix (PWM) is defined as M_W , each entry in PWM is $m_{b,j} = \ln(p_{b,j} / q_{b,0})$. Given a segment s , the match score of M_W on s is defined as

$$Score_s = \sum_{i=1}^W m_{s_i, i} \cdot \quad (3)$$

In order to ranking the various motifs found by Motif GibbsGA, the P value of an information content is calculated. The P value is the probability of obtaining an information content greater than or equal to the observed value, given the number of sequences in the alignment and its width. We use the method described in MEME to calculating the P value of information content. The information content of the sequence alignment is defined as

$$I = \sum_{j=1}^W \sum_{i=1}^{|\Sigma|} p_{i,j} \ln(p_{i,j}/q_{i,0}) \quad (4)$$

In what follows, we will show an efficient method (Motif GibbsGA) for discovering motifs based the above-mentioned models.

2.2. Greedy search for choosing starting point

The free parameters for a motif model based on PSFM are the motif length W and the entries in PSFM. One might try using randomly chosen letter frequency matrices as starting points. In this paper, a greedy search strategy is used to choose more intelligent ones.

Since each W -letter segment in the search space may be a candidate of a motif, so we first build a precompiled library of motifs represented by PSFMs, which are constructed from one of the input sequences selected randomly. Each W -letter segment of the selected sequence is associated with a position-specific frequency matrix θ , which constructed as equation (5).

$$\theta = [p_{i,j}], \text{ where } p_{i,j} = \begin{cases} \lambda & \text{if } a_{i,j} = \alpha_i \\ \frac{(1-\lambda)}{(|\Sigma|-1)} & \text{otherwise.} \end{cases} \quad (5)$$

Where $\alpha_i \in \Sigma, \lambda > 1/|\Sigma|$, then a PWM $m_{i,j} = \ln(p_{i,j}/q_{i,0})$ is also defined. Next each W -letter segment in the search space is matched to one of the motifs in the library based on the match score's P value. A W -letter segment is matched to one of the motifs with the smallest match score's P value. Following Bailey, and Gribskov (1998) [17], we obtain the match score's P value by calculating the cumulative density function. Let $M^{(k)}(x)$ be the match score probability density function for the motif PWM matrix if it consisted of only its first k columns, then the density for the matrix consisting of the first $k+1$ columns is

$$M^{(k+1)}(x) = \sum_{j=1}^{|\Sigma|} M^{(k)}(x - m_{j,k+1}) q_{j,0}. \quad (6)$$

To start the induction, set $M^{(0)}(0)=1$ and $M^{(0)}(x)=0$ for $x>0$. After W iterations, $M^{(W)}(x)$ contains the probability density for matching the motif with a random W -letter segment, from which the P value of the match score of a motif M_w on segment s is

$$P(\text{Score}_s) = \sum_{x \geq \text{Score}_s} M^{(W)}(x). \quad (7)$$

Finally every W -letter segments have matched to a motif in the library, but only one of the alignment matrices, which has the smallest P value of an information content, is saved as the start point.

2.3. Site sampler

Every possible segment of width W within a sequence is considered as a possible instance of the motif. Site sampler finds one occurrence per sequence of the motif in the dataset. Based on the probabilities P_s and Q_s , the weight $A_s = Q_s/P_s$ is assigned to segment s , and with each segment within a sequence so weighted. Thus, a motif instance can be randomly drawn from all possible segments with probability proportional to A_s . The implementation of our site sampler algorithm is based on the original Gibbs sampling algorithm previously described by Lawrence *et al.* [18]. The terminating condition of our site sampler is either after a user-specified maximum number of iterations (by default 500) or until the change in θ_w (Euclidean distance) falls below a user-specified threshold (by default 10^{-6}).

2.4. PSFM optimization by genetic algorithm

PSFM optimization is applied during the site sampling process, and substantially increased the sensitivity/specificity of a poorly estimated PSFM, and further improved the quality of a good PSFM. We use a genetic algorithm inspired by Leping Li *et al.* (2007) [16] for PSFM optimization. The method is described as follows.

1. Initialization: We made 100 ‘clones’ of the starting PSFM, which estimated by site sampler, to form a ‘population’. Different population sizes can be used.
2. Mutation: All except the raw PSFM were subject to mutation. For each PSFM to be mutated, n (the number of columns subject to mutation) was randomly generated from the following distribution, $P(n=k) = 1/2^k, k=2, \dots, W$, and $P(n=1) = 1 - \sum_{k=2}^W 1/2^k$. Next, n columns were randomly selected with equal probability for the W columns. For each column chosen, one of the four bases was then randomly selected and a small value, d , was added to or subtracted from $p_{i,j}$ with equal probability. Mutation was performed independently for each selected column. After mutation, we set negative values to 0 and standardized each column. We typically set the value of d to 0.05 for the first 100 iterations of site sampler and to 0.02 for the remaining iterations of site sampler. Convert all the PSFMs to PWMs.
3. Substring assignment: After mutation, a W -letter segment in the search space is matched to a PWM if and only if their match score’s P value little than the threshold P_{th} (such as 10^{-3} or 10^{-4}).
4. Fitness evaluation: The P value of the information content is used as the fitness score. The alignment matrix with the smallest P value is kept down for the remainder of site sampler iterations.

2.5. The description of Motif GibbsGA

Site sampler finds one occurrence per sequence of the motif in the dataset. To further extend Motif GibbsGA, we use a simple scan procedure to extract additional motif sites within a set of sequences. Site sampler obtained an optimum PWM_{opt}. Start from this PWM_{opt} configuration, our scan algorithm cycles through all remaining potential motif sites, and selects any additional sites that give the statistical significance of the match score (the threshold P value, P_{th} , is given by user). In other word, if $P(\text{Score}_s)$ is the P value of the match score of a motif M_w on segment s , then we accept this addition if and only if $P(\text{Score}_s) < P_{th}$.

In order to finding more than one motif in a set of sequences, Motif GibbsGA uses an iterative-masking approach: the binding sites of a discovered motif are masked out of the sequence dataset and then Motif GibbsGA is re-applied to this masked dataset to find additional motifs. The pseudo-codes of Motif GibbsGA is shown in Table 1.

Table 1. Pseudo-code of Motif GibbsGA

<p>Procedure Motif GibbsGA</p> <p>Begin</p> <p>S: dataset of sequences; W_{min}: the minimum width of the motif; W_{max}: the maximum width of the motif; W_{step}: increase step for motif width.</p> <p>For $W = W_{min}$ to W_{max} by W_{step} Do</p> <p style="padding-left: 20px;">Greedy search for choosing starting point (θ_W) from dataset S;</p> <p style="padding-left: 20px;">While ($t < Iteration_{max}$ and Euclidean distance ($\theta_W^t, \theta_W^{t-1}$) $< 10^{-6}$) Do</p> <p style="padding-left: 40px;">Run site sampler;</p> <p style="padding-left: 40px;">If ($t \% 20 == 0$)</p> <p style="padding-left: 60px;">PSFM optimization by genetic algorithm;</p> <p style="padding-left: 40px;">End If</p> <p style="padding-left: 20px;">End While</p> <p style="padding-left: 20px;">Extract additional motif sites if and only if $P(Score_s) < P_{th}$.</p> <p style="padding-left: 20px;">Print the motif instances, calculate the P value of the information content;</p> <p style="padding-left: 20px;">“Erase” occurrences of best motif found above;</p> <p>End For</p> <p>End</p>
--

3. Results

The datasets are available as a benchmark at the assessment web site <http://bio.cs.washington.edu/assessment/>.

These datasets comprised of eukaryotic binding sites belonging to 52 transcription factors representing four different species, 6 belonging to fly, 26 belonging to human, 12 belonging to mouse and 8 belonging to yeast. The binding sites of each transcription factor were presented in three different background models, ‘real’, ‘generic’, and ‘Markov’. For each of the different type of background model, the known positions of the binding sites were kept unchanged. Four additional datasets of type ‘Markov’ containing no planted binding sites were included as negative controls. Using these datasets Tompa *et al.* evaluated the performance of 13 different computational tools for *de novo* prediction of regulatory elements. We evaluated Motif GibbsGA on these datasets according to the performance measures described in Tompa *et al.* and compared it to the 13 tools evaluated in that paper. The comparative results of this evaluation are described as below.

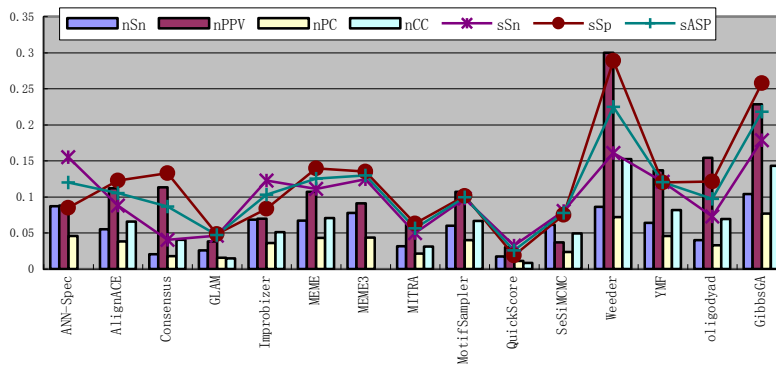


Figure 1. Combined measures of correctness over all 56 data sets

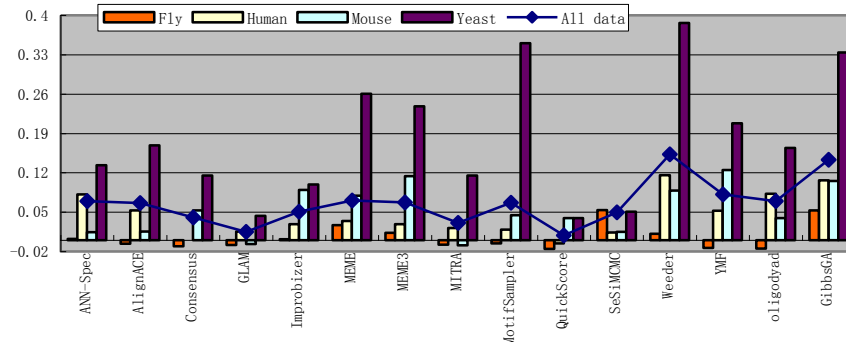


Figure 2. Correlation coefficient (nCC) by species

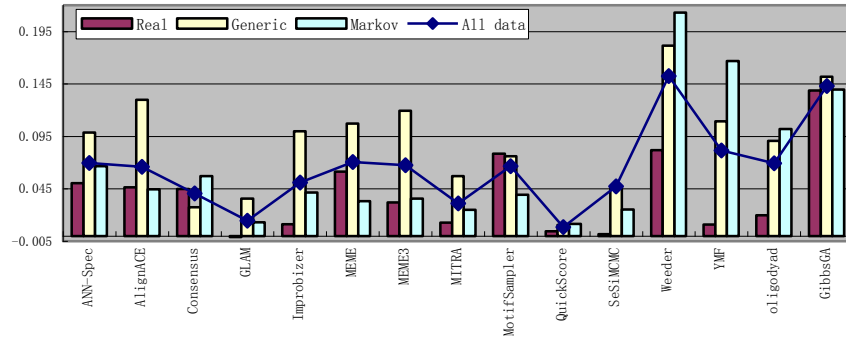


Figure 3. Correlation coefficient (nCC) by data set type

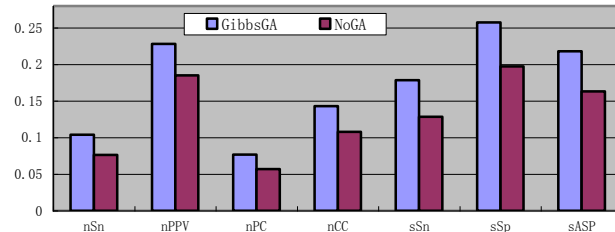


Figure 4. Comparison of using or not using genetic algorithm

For Motif GibbsGA, we set $W_{min}=6$, $W_{max}=12$, $W_{step}=1$, $Iteration_{max}=1000$ and $P_{th}=10^{-4}$, use the 3-order background model. Figure 1 shows the comparative results for the seven statistics, nucleotide-level sensitivity (nSn), nucleotide-level positive predictive value ($nPPV$), nucleotide-level performance coefficient (nPC), nucleotide-level correlation coefficient (nCC), site-level sensitivity (sSn), site-level positive predictive value ($sPPV$) and site-level average site performance ($sASP$), which summarized over all 56 datasets, regardless of species or background model. The Motif GibbsGA outperforms the other tools in all indicators except for Weeder. Figure 2 gives the breakdown of the performance, high-lighted by the correlation coefficient, nCC , of each tool on the datasets of the different species (regardless of the background model). From Figure 2, we can see that Motif GibbsGA does better than most of the other tools in all species. Figure 3 breaks down the datasets according to the different background models, ‘real’, ‘generic’ and ‘Markov’ (regardless of the species). Motif GibbsGA performs better than most of the other tools in prediction motifs in all the three background models.

From Figure 3, we can see that Motif GibbsGA does better than all the other Gibbs or Gibbs-like based method (ANN-Sped, AlignACE, GLAM, MotifSampler and SeSiMCMC), this

is in part attributed to the significant role played by the genetic algorithm that improved PSFM. Gibbs sampling is a heuristic search algorithm, the performances of this method are subject to potential suboptimal solutions in the search space. One way to tackle this problem is to import stronger global optimization techniques, such as genetic algorithms and others. In order to see the effect of genetic algorithm, we delete the PSFM optimization step, and run the program on the same datasets. Figure 4 gives the result, and compared to that of Motif GibbsGA.

4. Discussion

We have developed a novel *de novo* motif identification approach based on Gibbs sampling coupled with PSFM optimization by genetic algorithm, which we implemented in our software Motif GibbsGA. We test Motif GibbsGA on the benchmark dataset compiled by Tompa *et al.* (2005) for assessing computational tools that predict TFBSs. The performance of Motif GibbsGA on this data set compares well to, and in many cases exceeds, the performance of existing tools. From the result, we can see that here are considerable differences in results from these different programs, which is due to the existence of a large number of possible solutions. Optimization algorithms such as ANN-Sped, AlignACE, GLAM, MotifSampler, SeSiMCME and MEME can locally optimize their motif discovery results, but the inherent multimodality of the solution space restricts these local optimization procedures from exploring many different solutions. The Motif GibbsGA framework allows a greater flexibility of movement around the solution space by applying an evolutionary process to an entire population of possible solutions.

Despite considerable effort to date, it remains a complex challenge for computational biologists to convincingly predict regulatory elements in DNA sequences. Further efforts will be put in for several issues, the most important ones of which are the correlation between motif positions [19], motif positional information [20] and synergistic relationships between transcription factors [21]. As the complexity of these models increased, the need for sophisticated algorithms for finding optimal solutions to these models will become increasingly important.

5. Acknowledgment

This work was partially supported by the NNSF of China under Grant Nos.60705004 the Fundamental Research Funds for the Central Universities under Grants No. JY K50510030004.

6. Reference

- [1] Bailey,T.L and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning*, 21, 51-80
- [2] Ao,W., Gaudet,J., Kent,W.J., Muttumu,S. and Mango,S.E. (2004) Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, 305, 1743-1746.
- [3] Hughes,J.D., Estep,P.W., Tavazoie,S., and Church,G.M. (2000) Computational identification of *cis*-regulatory elements associated with functionally coherent groups of genes in *Saccharomyces cerevisiae*. *J.Mol.Biol.*, 296, 1205-1214
- [4] Liu,X., Brutlag,D.L. and Liu,J.S. (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.*, 6, 127-138.
- [5] Thijs,G., Marchal,K., Lescot,M., Rombauts,S., De Moor,B., Rouze,P. and Moreau,Y. (2002) A Gibbs sampling methods to detect overrepresented motifs in the upstream regions of co-expressed genes. *J. Comput. Biol.*, 9, 447-464.
- [6] Frith,M.C., Hansen,U., Spouge,J.L. and Weng,Z. (2004) Finding functional sequence elements by multiple local alignment. *Nucleic Acids Research*, 32, 189-200.
- [7] Liang,K.C., Wang,X.D., and Anastassiou,D. (2008) A profile-based deterministic sequential Monte Carlo algorithm for motif discovery. *Bioinformatics*, 24, 46-55.
- [8] Hertz,G. and Stormo,G. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, 15, 563-577.

- [9] Zhi Wei., and Shane T.Jensen. (2006) GAME: detecting *cis*-regulatory elements using a genetic algorithm. *Bioinformatics*, 22, 1577-1584.
- [10] Chan,T.M., Leung,K.S., and Lee,K.H. (2008) TFBS identification based on genetic algorithm with combined representations and adaptive post-processing. *Bioinformatics*, 24, 341-349.
- [11] Martin C. Frith., Yutao Fu., and Liqun Yu., *et al.* (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Research*, 32, 1372-1381.
- [12] Mahony,S., Hendrix,D., Golden,A., Smith,T.J. and Rokhsar,D.S. (2005) Transcription factor binding site identification using the self-organizing map. *Bioinformatics*, 21, 1807-1814.
- [13] Tompa,M., Li,N., Bailey,T.L., Churc,G.M, De Moor,B., Eskin,E. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology*. 23, 137-144.
- [14] Jianjun Hu, Bin Li, Daisuke Kihara. (2005) Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Research*, 33, 4899-4913.
- [15] Edward Wijaya., Siu-Ming Yiu. and Ngo Thanh Son1. *et al.* (2008) MotifVoter: a novel ensemble method for fine-grained integration of generic motif finders. *Bioinformatics*, 24, 2288 - 2295.
- [16] Leping Li., Yu Liang. and Robert L. Bass. (2007) GAPWM: a genetic algorithm method for optimizing a position weight matrix. *Bioinformatics*, 23, 1188 - 1194.
- [17] Timothy L. Bailey and Michael Gribskov. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, 14, 48-54.
- [18] Lawrence, C.E., Altschul, S.F., Bogouski, M.S., Liu, J.S., Neuwald, A.F., and Wooten, J.C. (1993) Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment. *Science*, 262, 208-214.
- [19] Paulo G. S. da Fonseca., Christian Gautier, Katia S. Guimaraes and Marie-France Sagot. (2008) Efficient representation and P-value computation for high-order Markov motifs. *Bioinformatics*, 24, i160-i166.
- [20] Ana C Casimiro1, Susana Vinga, Ana T Freitas and Arlindo L Oliveira. (2008) An analysis of the positional distribution of DNA motifs in promoter regions and its biological relevance. *BMC Bioinformatics*, 2008, 9:89.
- [21] Li Shen, Jie Liu and Wei Wang (2008) GBNet: Deciphering regulatory rules in the co-regulated genes using a Gibbs sampler enhanced Bayesian network approach. *BMC Bioinformatics*, 2008, 9:395.

Li-fang LIU, born in 1972, Ph.D., associate professor. Her current research interests include bioinformatics, machine learning, pattern recognition and intelligent computation.
E-mail:qixiaogang@gmail.com