A Novel Ant Colony Optimization Based Algorithm for Identifying Gene Regulatory Elements

Wei Liu
Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China
Institute of Information Science and Technology, Yangzhou University , Yangzhou 225127, China
Email: yzliuwei@126.com

Hanwu Chen
Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China
Email: hanwu_chen@163.com

Ling Chen
Institute of Information Science and Technology, Yangzhou University , Yangzhou 225127 , China
National Key Lab of Novel Software Tech, Nanjing University, Nanjing 210093, China
Email:lchen@yzcn.net

Yixin Chen
Department of Computer Science, Washington University in St. Louis, St. Louis, MO 63130,USA

***Abstract*—It is one of the most important tasks in bioinformatics to identify the regulatory elements in gene sequences. Most of the current algorithms for identifying regulatory elements are easily to converge into a local optimum, and have high time complexity. Therefore, we propose a novel optimization algorithm named ACRI(ant-colony-regulatory-identification) for identifying regulatory elements. Based on powerful optimization ability of ant-colony algorithm, the algorithm ACRI can not only improve the quality of results, but also solve the problem at a very high speed. Experimental results show that ACRI can obtain higher quality of solutions using less computational time than other traditional algorithms.**

***Index Terms*—Bioinformatics, Gene regulatory elements, Ant colony optimization**

Ⅰ. INTRODUCTION

A biological system is mainly composed of static and dynamic components. The static components include all genes in the genome, which are the elementary constructional elements of a biological system. With the achievements in the genome sequencing and annotation, special interests have been paid on the gene regulatory elements, the dynamic component of the biological system. Genomic regulatory elements, which are also called DNA motifs, contain abundant biological information reflecting life characteristics, and play an important role in the gene function and structure construction. Now discovering and recognizing gene regulatory elements have become one of the most important approaches in analysis of genome sequences, and have drawn extensive attention in bioinformatics research.

Gene regulatory element identification (also called motif identification) evolves two problems: how to extract motif from biological data or structures, and how to recognize the motif contained in object sequences or structures. Regulatory element identification is a major research area in the study of gene non-coding region. At the transcriptional and post-transcriptional level, gene expression is mainly controlled by some *cis*-regulatory elements which essentially are some shorter DNA sequences. These sequences are often in the upstream region of regulated genes, and are recognized by and combined with the specific DNA-binding protein (transcription factor) so as to regulate DNA metabolism and transcription. Otherwise they could probably be recognized by and combined with the RNA-binding protein, and their combination could influence the processes of RNA modification, localization, translation and degradation. Hence, transcriptional regulatory element analysis and identification is one of the most important tasks for genome behavior understanding and explanation. .

In searching for a known regulatory element or predicting a new one, three problems must be solved: 1) how to describe the regulatory elements, namely, what

This research was supported in part by the Chinese National Natural Science Foundation under grant No. 61070047, Natural Science Foundation of Jiangsu Province under contract BK2008206.

© 2013 ACADEMY PUBLISHER
doi:10.4304/jcp.8.3.613-621

characteristic model would be constructed for regulatory elements? 2) how to define a measurement or scoring function about the probability for a sequence segment being a regulatory element; 3) given the regulatory element model and scoring function, how to detect the regulatory element with the maximal score from sequences to be analyzed, which is just the problem of algorithm design.

In the past two decades, more and more efforts have been dedicated to gene regulatory element identification in DNA sequences. There are mainly two categories of gene regulatory element identifying methods: experimental methods and computational methods. Due to the high time and economic cost, the experimental methods are probably not able to obtain comprehensive results. Therefore, computational methods [1-3] have drawn much more attention because of its effectiveness and high efficiency. However, it is referred to three problems in computational methods: 1) given genome sequences, finding out all known regulatory elements; 2)find out unknown regulatory elements from the upstream of some co-expression or co-regulated genes; 3)find out the unknown gene regulated by a known transcriptional factor. We focus on the second problem which is called sequence-driven regulatory elements identifying. The sequence-driven methods are predicting methods[4] for detecting the common element on the co-regulated gene cluster.

At present, existing algorithms for regulatory elements identifying include: (1) Counting algorithm. It is the most instinctive and simplest exhaustive search method. Since its time complexity is proportional to the exponent of pattern length, this method is only suitable to identify short regulatory elements. (2) EM algorithm. The EM algorithm is an efficient iterative procedure to compute the Maximum Likelihood (ML) estimate in the presence of missing or hidden data. Each iteration of the EM algorithm consists of two steps: the E-step which determines the conditional expectation, and the M-step which maximizes the expectation. Efficiency of the algorithm greatly depends on the initial conditions. With the inappropriate initial setting of the parameters, it will converge to a local optimum instead of the global one. (3) MM(Mixture Model) algorithm. It is an improvement of EM algorithm. The basic idea of MM lies on the conservation of regulatory elements and their corresponding characteristic matrices. During the process of continual iterations, the log-likelihood will be maximal only when both of them are co-adapted. After conserved sequences, sensing matrices or characteristic models are obtained, they are evaluated by their statistical significances. (4) Gibbs Sampling algorithm. It is a special MCMC (Markov Chain Monte Carlo) method to identify motifs of protein sequences proposed by Lawrence[5] et. al. Later Liu[6] et. al. adopted Gibbs sampler into Bayesian model. Their method is used to solve the problem of multiple sequence alignment and achieved admirable results. Now Gibbs Sampling and its improvements have sparked a major increase in the application of regulatory element identification. There is much mature software available on the Internet, such as AlignACE[7],BioProspector[8], Gibbs Motif Sampler[9] etc. The primary principle of Gibbs sampling is to optimize the object function through continually updating regulatory element model and its position in each sequence by a random sampling. The final regulatory element is obtained when the iteration is terminated under a certain condition. At present, some other pieces of popular software is developed such as Consensus[10],MEME[11],ANN-Spec[12],PROJECTION[13],MDScan[14],and YMF[15], which is the most recent one. Recently, some optimization methods are also applied on regulatory element identification, such as statistical analysis, neural network, clustering prediction and word identification etc.

Our study focuses on the problem of searching for the binding sites from co-expression gene sequences. The premise of using the computational method to solve the problem is assuming that the genes regulated by the same regulatory element possess the same or similar gene expression mode. The co-expression genes can be obtained by clustering the gene chip data. Our goal is to detect all possible binding sites of transcription factor from the upstream of co-expression genes. Therefore the problem can be defined as an optimization process to search for the conserved sequence segments of certain length from a sequence set. Based on the ant colony optimization, we present a novel method named ACRI(ant-colony-regulatory-identification) for regulatory elements identification. Compared with the existing algorithms, our algorithm ACRI can avoid converging into local optimum and can not only improve the quality of results, but also solve the problem at a very high speed. Experimental results on two groups of standard test data show that our algorithm ACRI can obtain higher quality of solutions using less computational time than other traditional algorithms.

Ⅱ. CONCEPTS AND DEFINITIONS

A. Problem Definition

For convenience in description of the problem, we assume that each regulatory element occurs only once in each sequence. Given the sequence set $X=\{X_1,X_2,\ldots,X_n\}$, where each sequence is composed of four nucleotides: A,T,C and G. The lengths of those sequences are denoted as $l_1,l_2,\ldots,l_n$ respectively. Our goal is to find out the set of conserved sequence segments $M = \{M_1,M_2,\ldots,M_n\}$ consisting of the motif with length w. Hereby, M_i is the substring of X_i with length w, and M_i Ì X_i, (i=1,2,…,n).

In the computational methods mentioned above, the first problem to be solved is how to denote these sequences, namely, to construct a proper feature model for the regulatory elements. In this paper, we use the matrix model, which uses a characteristic matrix to describe the distribution of the regulatory elements. Hence firstly we have to define the characteristic matrix.

© 2013 ACADEMY PUBLISHER

B. Characteristic Matrix

Definition1 Let the length of motif be w and alphabet be $\Sigma = \{A,T,C,G\}$. The characteristic matrix M is a 4*w matrix and its j^{th} element at the i^{th} row is notated as P_j^b, where b is the i^{th} character in the alphabet, and P_j^b denotes the possibility of the i^{th} character appearing at the j^{th} position of the motif.

Example1 Assuming that there are 12 regulatory elements of length 6 shown as follows:

X_1= " A C G C G T "
X_2= " A C G C G T "
X_3= " C C G C G T "
X_4= " T C G C G A "
X_5= " A C G C G T "
X_6= " A C G C G A "
X_7= " A C G C G T "
X_8= " A C G C G A "
X_9= " A C G C G T "
X_{10}= " A C G C G T "
X_{11}= " A C G C G T "
X_{12}= " A C G C G T "

Then we can construct the matrix model as Table Ⅰ:

TABLE I.
SIMPLE MATRIX

	1	2	3	4	5	6
A	10	0	0	0	0	3
T	1	0	0	0	0	9
G	0	0	12	0	12	0
C	1	12	0	12	0	0

Each element in Table Ⅰ denotes the number of the base appeared in this position. For example, the first element at the first row is 10, this means the base "A" appears 10 times in the first position of these regulatory elements. We can easily transform each element in Table Ⅰ into the probability of the base appears in the position and get the characteristic matrix as shown in Table Ⅱ.

TABLE II.
THE CHARACTERISTIC MATRIX

	1	2	3	4	5	6
A	0.8333	0	0	0	0	0.25
T	0.083	0	0	0	0	0.75
G	0	0	1	0	1	0
C	0.083	1	0	1	0	0

If we list the characters which have the maximum probability in each column, we can get a sequence "*ACGCGT*", which is the most potential consensus sequence of the regulatory element. Enlightened by this fact, we use the characteristic matrix as the model reflecting the feature of the regulatory element. Moreover, it also can be used as a tool for regulatory element detecting.

C. Background Model

Suppose the set of DNA sequence is $X=\{X_1,X_2,\ldots,X_n\}$, and the regulatory element appears once and only once in each sequence. Denote the area where the regulatory element located as M which is shown as the shadow part in Figure 1. All the other parts and non-motif areas are regarded as the background.

Definition 2 The background pattern $B=\{p_0^A,p_0^T,p_0^C,p_0^G\}$ denotes the probability of each base appearing in the background area of the sequences, where p_0^b is the probability of character b appearing in the background area.

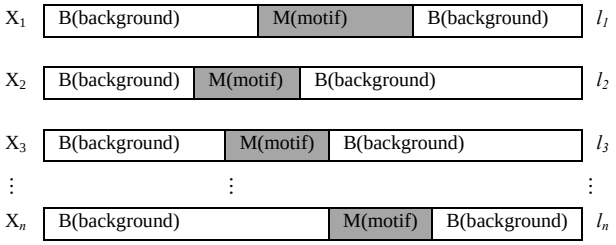

Figure 1. Motif and background

D. Problem Model

Given the background pattern $B=\{p_0^i,\ (i = A,C,G,T)\}$ and the motif characteristic matrix $M=\{p_j^i,\ (i = A,C,G,T, j = 1,2,\ldots,w)\}$, we can get the probability of the motif occurring at the k^{th} position of $X = (X_1,X_2,\ldots,X_n)$ using formula (1).

$$P(X_i|K,M,B) = \prod_{j=0}^{k-1} p_0^{x_j} \prod_{j=k}^{k+w-1} p_{j-k+1}^{x_j} \prod_{j=k+w}^{l_i} p_0^{x_j} \qquad (1)$$

However, in the formula(1), the characteristic matrix M is unknown and the start positions of each motif in each sequence are also unobserved, therefore the above model is actually a hidden model. That is to say, the elements of M and the start positions of the motif are all hidden variables which are just the unknowns we are going to compute.

We can enumerate all the subsequences with length w in the sequences of set $X = (X_1,X_2,\ldots,X_n)$ and denote the subsequence of length w which starts from the j^{th} position in sequence X_i as X_{ij}. Thus there are $L_i = (l_i - w + 1)$ subsequences in sequence X_i. Given the subsequence tuple $\{X_{1j_1},X_{2j_2},\ldots,X_{nj_n}\}$, where $j_i = 1,2,\ldots,L_i$, we can compute its characteristic matrix by Definition 1.

Through the analysis mentioned above, it is known that there are $\prod_{i=1}^{n}(L_i - w + 1)$ subsequence tuples. Therefore, $\prod_{i=1}^{n}(L_i - w + 1)$ characteristic matrices can be obtained and we want to choose the best one as the motif characteristic matrix.

© 2013 ACADEMY PUBLISHER

But which one is the "best"? It is referred to the second important problem encountered in finding the binding sites using computing methods. We have to define a measurement to detect the possibility of the sequence segment as a regulatory element, namely, a scoring function is needed.

At present, the most widely used scoring functions are as follows: (1)*Z*-score; (2)χ^2 statistics; (3)Information Content; (4) Consensus Scoring; (5) The log-likelihood. Among the five scoring functions illustrated above, the first two calculate the scores based on the statistical importance of the motif occurrences; and the last three are designed based on the motif conservative. Therefore, the first two are usually applied to identify the regulatory elements based on statistics. And the methods on the basis of sequence alignment usually adopt the last three scoring functions. In our paper, we use information content as the standardized scoring function of the motifs.

E. Information Content

This scoring function describes the pattern conservative, which is deduced from the theorem of information theory. Compared with stochastic sequences, the more the uncertainty of the pattern decreases the higher information content it will have. Therefore the conserved pattern has larger probability to be a regulatory element.

Definition 3 Given a characteristic matrix $M = (p_j^b)$, its information content is defined as:

$$IC(M) = \sum_{j=1}^{n} \sum_{b \in \{A,C,G,T\}} p_j^b * \log(p_j^b / p_0^b) \quad (2)$$

Information content is often used to measure the differences within the set of samples, namely, the conservative of the samples. A higher *IC* value indicates the less difference among the samples and the higher conservative in the set.

From formula (2) we can see that when the probability p_j^b of the base *b* occurring at the j^{th} position is larger than the probability p_0^b of *b* at the background, their ratio will be higher and will make a greater contribution to the *IC* score. In the other words, the higher their ratio is, the more likely the base occurs at the j^{th} position, and consequently the higher conservative it will have.

Compared with other non-coding sequences, the regulatory elements are more conservative and have higher *IC* value. Therefore, from all the tuples of subsequences of length *w*, we must select the one with the highest *IC* score, and use its characteristic matrix as the model of the motif. Nevertheless, since there are $\prod_{i=1}^{n}(L_i - w + 1)$ subsequence tuples, it would cost large amount of time to enumerate them. Since it is an NP-hard problem, many optimization methods are used to solve it. This paper presents an algorithm based on ant-colony optimization to construct the characteristic matrix of the motif.

III. ANT COLONY ALGORITHM OPTIMIZATION

Metaheuristic algorithms such as genetic algorithm, evolutional algorithm, simulated annealing, ant colony optimization, tabu search etc. are algorithms which can solve many combinatorial optimization problems, and have been triumphantly applied to many practical problems[16-23]. Ant colony optimization (ACO) is a new evolution simulation algorithm proposed by Italian researcher M. Dorigo, V. Mahiezzo, A. Colorni etc. ACO has been proved effective in solving complex optimization, especially for the discrete NP-hard combinatorial optimization, such as TSP [17](Traveling Salesman Problem), JSP[18-19] (Job-shop Scheduling Problem), FSP [20-21](Flow shop Scheduling Problem), QAP(Quadratic Assignment Problem),SOP[22] (Sequential Ordering Problem), QOS multicast routing, Dynamic Vehicle Routing Problem [23]etc.

Many real ant species deposit on the ground a substance called pheromone, as they travel to and from a food source. Other ants searching for food can sense the pheromone and have their movements influenced by the strength of the pheromone on the path. Hence the collective behavior actually constructs the positive feedback mechanism: the more ants travel through the path, the more likely the other ants would select it. The pheromone information will direct the future ants to travel on the shortest path. The essences of the optimization process are attributed to the follows: (1) selecting strategy: the path with more pheromone will have more chance to be selected ;(2)updating strategy: pheromone intensity on a path will be reinforced along with the ants traveling through it and decreased over time if the path is not used; (3)coordination strategy: ants can communicate and collaborate with each other via pheromone on the paths. Based on the optimization strategies mentioned above, ant colony algorithm has strong optimization ability and great potential in solving complex combinational optimization problems.

IV. FRAMEWORK OF THE ALGORITHM

At present nearly all regulatory elements identifying algorithms are based on an incomplete search and most of popular software uses local search strategy, such as greedy algorithm、 EM algorithm、Gibbs algorithm and so on. Thus these algorithms are easily to converge into a local optimum and get some results of little biological meaning. In this paper, we propose a novel optimization algorithm named ACRI (ant colony regulatory identification) for regulatory elements identification. Based on powerful optimization ability of ant-colony algorithm, the algorithm ACRI can solve the problem in a very high speed and get higher quality of solutions.

A. Coding the Solutions

Suppose the set of the biosequence is $X = (X_1, X_2, \ldots, X_n)$. Here, l_i denotes the length of the sequence X_i and *w* is the length of the conserved sequence fragment where the motif is located. We use an ant to denote a solution. Assuming that each input sequence

© 2013 ACADEMY PUBLISHER

contains just one motif, we code the object motif as an integer vector $J = \{j_1, j_2, \ldots, j_n\}$, where $j_i \in [1, l_i - w + 1]$ denotes the start position of the binding sites in sequence X_i, namely, the start position of the motif. A vector J indicates a tuple consisting n subsequence of length w. Then we can compute its characteristic matrix $M(J)$ and get the information content value $IC(M(J))$.

B. The Digraph Which the Ant-colony Traverse on

In the algorithm ACRI, the artificial ants traverse on a digraph shown in Figure 2. From Figure 2, we can notice that there are n+1 nodes denoted as $X_1, X_2, \ldots, X_n, X_{n+1}$ respectively, where X_i represents the i^{th} (i=1,2,…,n) sequence and X_{n+1} denotes the termination. There are $L_i = l_i - w + 1$ paths linking node from sequence X_i and X_{i+1}. Denote the j^{th} path as C_{ij} which means the start position of the i^{th} sequence is j. The pheromone on path C_{ij} is denoted as τ_{ij}. Each artificial ant starts from node X_1, passes through $X_2, X_3, \ldots$, and then arrives at the termination X_{n+1}. Let the trace of an ant be $C_{1j_1}, C_{2j_2}, \ldots, C_{nj_n}$, then it forms a solution $J = \{j_1, j_2, \ldots, j_n\}$.

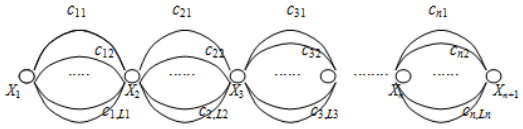

Figure 2. The diagraph which the ant-colony traverse on

C. Probability for Ants' Path Selecting

The probability that ant at node X_i chooses the path C_{ij} linking with X_{i+1} can be defined as:

$$P_{ij}(t) = \frac{[\tau_{ij}(t)]^{\alpha}[h_{ij}(t)]^{\beta}}{\sum_{k=1}^{L_i}[\tau_{ik}(t)]^{\alpha}[h_{ij}(t)]^{\beta}} \tag{3}$$

Here, $\tau_{ij}(t)$ is the pheromone on the edge C_{ij} at time t, $h_{ij}(t)$ is a heuristic function which is defined as the suitability for selecting the j^{th} position of sequence X_i as the start point of the motif. It uses the current optimal motif obtained by the ant for reference.

$$h_{ij}(t) = \frac{P(X_{ij}, X_{i,j+1}, \ldots, X_{i,j+w-1} | M(t))}{P(X_{ij}, X_{i,j+1}, \ldots, X_{i,j+w-1} | M_0)} = \sum_{k=1}^{w} \frac{p_k^{S_{i,j+k-1}}}{p_0^{X_{i,j+k-1}}} \tag{4}$$

The algorithm records the optimal solution obtained so far in the iterations, which is named M_{best}. We use M_{best} as the approximation of the optimal motif so as to set the heuristic function $h_{ij}(t)$. In formula (4), $M(t)$ denotes the characteristic matrix of the historic optimal solution M_{best} obtained by the ant.

D. Fitness Function

For the solution $J = \{j_1, j_2, \ldots, j_n\}$ obtained by the ant, we can get the following subsequence tuple of length w:

$$X_{1j_1}, \ldots, X_{1j_1+w-1}$$
$$X_{2j_2}, \ldots, X_{2j_2+w-1}$$
$$\ldots\ldots\ldots\ldots$$
$$X_{nj_n}, \ldots, X_{nj_n+w-1}$$

By counting the occurrences of each character in each position, we can get the characteristic matrix $M(J)$ and then compute its information content $IC(M(J))$ as the fitness of the solution.

E. Pheromone Update

After each iteration, pheromone τ_{ij} on each edge (i,j) should be updated as follows:

$$\tau_{ij}(t+1) = \rho\,\tau_{ij}(t) + (1-\rho)\sum_{k=1}^{m}\Delta\tau_{ij}^{(k)}(t) \tag{5}$$

Here, $\rho \in (0,1)$ represents the evaporation rate of pheromone, while $\Delta\tau_{ij}^{(k)}$ is the increment of τ_{ij} by the k^{th} ant:

$$\Delta\tau_{ij}^{(k)}(t) = \begin{cases} IC(M_k) & \text{if the kth ant passes edge } C_{ij} \text{ in current tour} \\ 0 & \text{else} \end{cases} \tag{6}$$

where M_k is the characteristic matrix of the solution get by the k^{th} ant, $IC(M_k)$ is the information content of M_k obtained by using formula (2).

F. Framework of the Algorithm

As mentioned above, the framework of the algorithm ACRI is as follows:

Algorithm ACRI (ant-colony-regulatory-identification)
Input: $X_1, X_2, \ldots, X_n$: the set of sequences;
maxnum: maximum number of iterations;
m: number of the ants used;
Output: M_{best}: the characteristic matrix regulating the motifs;
$J_{best} = \{j_1, j_2, \ldots, j_n\}$: the start positions of the motifs;
Begin
1. Initialization, randomly setting the initial characteristic matrix M_{best} and computing the background pattern B.
2. **for** t=1 **to** *maxnum* **do**
3. **for** k=1 **to** m **do**
4. **for** i=1 **to** n **do**
5. ant k selects the edge C_{ij} according to formula (3);
6. local optimization for the start position j of X_i;
7. **end for** i
8. local optimization for the subsequence tuple and get the solution $J = \{j_1, j_2, \ldots, j_n\}$;
9. Compute IC ($M(J)$) by formula (2);
10. **if** IC ($M(J)$) > IC (M_{best}) **then**
11. $M_{best} = M(J)$;
12. $J_{best} = J$;
13. **endif**;
14. **endfor** k
15. Update the pheromone on each edge according to formula(5) and compute $h_{ij}(t+1)$ using the new M_{best};
16. **end for** t
17. Output M_{best}, J_{best};
End

© 2013 ACADEMY PUBLISHER

G. The Strategy of Local Optimization for Single Sequence

Line 6 in the algorithm ACRI performs local optimization to adjust the start position j of X_i, that is, the ant will locally search for a better start position in the neighbor of j. Since this local optimization adjusts the start position j only in sequence X_i, we call it "local optimization for single sequence". The detail of the strategy is as follows:

Suppose the i^{th} sequence is $X_i = \{x_{i1}, x_{i2}, ..., x_{i,l_i}\}$, and the ant selects j as the starting position, j ? $[1, l_i \quad w + 1]$, namely, subsequence $x_{i,j}, x_{i,j+1}, ..., x_{i,j+w-1}$ is considered as the motif. In the local optimization, we test the subsequences $x_{i,j-1}, x_{i,j}, ..., x_{i,j+w-2}$ starting from j-1, and $x_{i,j+1}, x_{i,j+2}, ..., x_{i,j+w}$ starting from j+1, and compare them with the subsequence $x_{i,j}, x_{i,j+1}, ..., x_{i,j+w-1}$ starting from position j.

Using the historic optimal solution M_{best}, we compute $P(X_i | j, M_{best}, B), P(X_i | j - 1, M_{best}, B), P(X_i | j + 1, M, B)$ by formula (1), which respectively denotes the probabilities of the motif appears at position j, j-1 and j+1 in sequence X_i. We choose the one with the highest probability as the start position in X_i.

H. The Strategy of Local Optimization for Subsequence Tuple

Line 8 in the algorithm ACRI performs local optimization on the solution $J = \{j_1, j_2, ..., j_n\}$ obtained in current iteration. Since this local optimization adjusts the start positions on the sequences in the tuple by searching in the neighbor of the solution, we call it "local optimization for subsequence tuple". The detail of the strategy is as follows:

For the solution $J = \{j_1, j_2, ..., j_n\}$, we compute and compare the information content values of another two solutions $J^- = \{j_1-1, j_2-1, ..., j_n-1\}$ and $J^+ = \{j_1+1, j_2+1, ..., j_n+1\}$ respectively. We select the one of the highest IC among J、J^-、J^+ as the solution obtained by ant k in current iteration.

We can compute $IC(M(J^-))$ and $IC(M(J^+))$ on the basis of $IC(M(J))$. It's not necessary recalculate them by the formula (2). Actually from (2) we can see that:

$$IC(M(J)) = \sum_{j=1}^{n} \sum_{b \in \{A,C,G,T\}} p_j^b * \log(p_j^b / p_0^b) = \sum_{j=1}^{n} IC(M(J_j))$$

Here, $IC(M(J_j))$ is the information content of the element on the j^{th} column of $M(J)$, namely:

$$IC(M(J_j)) = \sum_{b \in \{A,C,G,T\}} p_j^b * \log(p_j^b / p_0^b)$$

It is obvious that $IC(M(J))$ is just the summation of all the IC values of its columns and the IC value of each column can be calculated independently.

For the solution $J = \{j_1, j_2, ..., j_n\}$, it contains a group of subsequences as follows:

$$\begin{matrix} x_{1,j_1} & x_{1,j_1+1} & x_{1,j_1+2} & \cdots & x_{1,j_1+w-2} & x_{1,j_1+w-1} \\ x_{2,j_2} & x_{2,j_2+1} & x_{2,j_2+2} & \cdots & x_{1,j_1+w-2} & x_{2,j_2+w-1} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n,j_n} & x_{n,j_n+1} & x_{n,j_n+2} & \cdots & x_{1,j_1+w-2} & x_{n,j_n+w-1} \end{matrix}$$

Denote the j^{th} column vector in J as J_j, and $J_0 = (x_{1,j_1-1}, x_{2,j_2-1}, ..., x_{n,j_n-1})^T$, $J_n = (X_{1,j_1+w}, X_{2,j_2+w}, ..., X_{n,j_n+w})^T$.

Because $IC[M(J)] = \sum_{j=1}^{n} IC(M(J_j))$, we have:

$$IC(M(J^-)) = \sum_{j=0}^{w-1} IC(M(J_j)) = IC(M(J)) + IC(M(J_0)) - IC(M(J_w))$$

$$IC(M(J^+)) = \sum_{j=2}^{w+1} IC(M(J_j)) = IC(M(J)) + IC(M(J_{w+1})) - IC(M(J_1))$$

Therefore, we just need to select the highest one among $IC(M(J_0)) - IC(M(J_w))$, 0 and $IC(M(J_{w+1})) - IC(M(J_1))$.

V. EXPERIMENTAL RESULTS AND ANALYSIS

We test our algorithm ACRI by a set of experiments to evaluate its efficiency. All the experiments were conducted on a 3.0GHzPentium4 with 1GB memory. All codes were complied using Microsoft Visual C++ 6.0.

In the experiments, all test data are from the standard databases. These test data include five transcriptional factors of Saccharomyces cerevisiae, and 18 gene sequences contain *E. coli* transcription factor binding sites. Both of them are used as the standard test data to evaluate the performance of the algorithm.

A. Analysis on the Quality of the Results

1) Analysis for transcription factor binding sites of Saccharomyces cerevisiae

In the experiments, we test our algorithm using Saccharomyces cerevisiae from the uniform database SCPD[24] （http://rulai.cshl.edu/SCPD/）. The database contains more plentiful information about regulatory elements and transcriptional factors compared with other congener databases such as TRANSFAC etc. We select five groups of transcription factor binding sites as tested data shown in table Ⅲ. We download some promoter sequences of length 550, each of which contains the transcription factor binding site from SCPD.

TABLE III.

THE FIVE TRANSCRIPTIONAL FACTORS OF SACCHAROMYCES CEREVISIAE

TF	Size	Length	Consensus Sequence
GAL4	6	17	CGGNNNNNNNNNNCCG
RAP1	16	7	RMACCCA
REB1	9	7	YYACCCG
MCB	6	6	WCGCGW
PDR3	7	8	TCCGYGGA

© 2013 ACADEMY PUBLISHER

We use logo creation model to visualize the experimental results. In the model, each logo consists of stacks of symbols, one stack for each position in the sequence. The overall height of the stack indicates the sequence conservation at that position, while the height of symbols within the stack indicates the relative frequency of each amino or nucleic acid at that position. In general, a sequence logo provides a richer and more precise description of, for example, a binding site, than a consensus sequence. We test these data with our algorithm ACRI and create sequences logos for our results through the weblogo (http://weblogo.berkeley.edu/logo.cgi) as shown in Figure 3-7. All the results are identical to the results by DNA footprinting method. This indicates that our algorithm is effective.

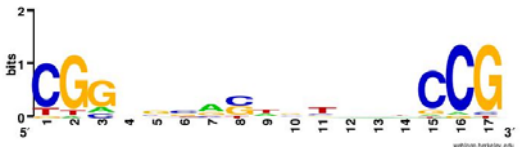

Figure 3. The running results about GAL4 through algorithm ACRI

Figure 4. The running results about RAP1 through algorithm ACRI

Figure 5. The running results about REB1 through algorithm ACRI

Figure 6. The running results about MCB through algorithm ACRI

Figure 7. The running results about PDR3 through algorithm ACRI

2) Analysis for CRP binding sites of Escherichia coli

Most of the existing software for identifying the regulatory elements uses the CRP binding sites[25] of Escherichia coli as test data. This data set consists of 18 sequences of length 105. Twenty-three of those CRP binding sites have been recognized by DNA footprinting method. Usually, the binding sites detected coincide with half of the sequence in the true motif model, the binding sites identification can be considered as a successful one. The 18 sequences of the CRP binding sites for escherichia coli are shown in Table Ⅳ. Generally, most of the popular computing methods set the length of the CRP binding sites as 18 to 25, so our algorithm sets it as 22. We consider that the binding site detecting is successful if and only if the difference between the obtained CRP binding sites and the known ones is less than 10.

TABLE IV. THE 18 SEQUENCES OF THE CAP BINDING SITES FOR ESCHERICHIA COLI

Name	Sequence
CE1CG	TAATGTTTGTGCTGGTTTTTGTGGCATCGGGCGAGAATAGCGCGTGGTGTGAA AGACTGTTTTTTTGATCGTTTTCACAAAAATGGAAGTCCACAGTCTTGACAG
ECOARABOP	GACAAAAACGCGTAACAAAAGTGTCTATAATCACGGCAGAAAAGTCCACATTG ATTATTTGCACGGCGTCACACTTTGCTATGCCATAGCATTTTTATCCATAAG
ECOBGLR1	ACAAATCCCAATAACTTAATTATTGGGATTTGTTATATATAACTTTATAAATT CCTAAAATTACACAAAGTTAATAACTGTGAGCATGGTCATATTTTTATCAAT
ECOCRP	CACAAAGCGAAAGCTATGCTAAAACAGTCAGGATGCTACAGTAATACATTGAT GTACTGCATGTATGCAAAGGACGTCACATTACCGTGCAGTACAGTTGATAGC
ECOCYA	ACGGTGCTACACTTGTATGTAGCGCATCTTTCTTTACGGTCAATCAGCAAGGT GTTAAATTGATCACGTTTTAGACCATTTTTTCGTCGTGAAACTAAAAAAACC
ECODEOP2	AGTGAATTATTTGAACCAGATCGCATTACAGTGATGCAAACTTGTAAGTAGAT TTCCTTAATTGTGATGTGTATCGAAGTGTGTTGCGGAGTAGATGTTAGAATA
ECOGALE	GCGCATAAAAAACGGCTAAATTCTTGTGTAAACGATTCCACTAATTTATTCCA TGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCATAAGCC
ECOILVBPR	GCTCCGGCGGGGTTTTTTGTTATCTGCAATTCAGTACAAAACGTGATCAACCC CTCAATTTTCCCTTTGCTGAAAAATTTTCCATTGTCTCCCCTGTAAAGCTGT
ECOLAC	AACGCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTT TATGCTTCCGGCTCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCAC
ECOMALBA	ACATTACCGCCAATTCTGTAACAGAGATCACACAAAGCGACGGTGGGGCGTAG GGGCAAGGAGGATGGAAAGAGGTTGCCGTATAAAGAAACTAGAGTCCGTTTA
ECOMALBA2	GGAGGAGGCGGGAGGATGAGAACACGGCTTCTGTGAACTAAACCGAGGTCATG TAAGGAATTTCGTGATGTTGCTTGCAAAAATCGTGGCGATTTTATGTGCGCA
ECOMALT	GATCAGCGTCGTTTTAGGTGAGTTGTTAATAAAGATTTGGAATTGTGACACAG TGCAAATTCAGACACATAAAAAAACGTCATCGCTTGCATTAGAAAGGTTTCT
ECOOMPA	GCTGACAAAAAAGATTAAACATACCTTATACAAGACTTTTTTTTCATATGCCT GACGGAGTTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTACATCGCC

TABLE V.

COMPARISON OF THE RESULTS BETWEEN ACRI AND TRADITIONAL ALGORITHMS

No.	Binding sites	Gibbs Sampler	difference	AlignACE	difference	MEME	difference	ACRI	difference
1	17,61	59	-2	63	2	61	0	63	2
2	17,65	53	-2	57	2	55	0	57	2
3	76	74	-2	78	2	76	0	78	2
4	63	59	-4	65	2	63	0	65	2
5	50	11	-39	52	2	13	-37	52	2
6	7,60	5	-2	9	2	7	0	9	2
7	42	40	-2	26	-16	42	0	44	2
8	39	37	-2	41	2	39	0	41	2
9	9,80	7	-2	11	2	9	0	11	2
10	14	12	-2	16	2	14	0	16	2
11	61	59	-2	63	2	35	-16	63	2
12	41	47	6	43	2	34	-7	43	2
13	48	46	-2	50	2	48	0	50	2
14	71	69	-2	73	2	71	0	73	2
15	17	15	-2	19	2	75	58	19	2
16	53	49	-4	55	2	6	-47	55	2
17	1,84	25	24	68	-16	27	26	95	4
18	78	74	-4	80	2	16	-2	78	0

Table Ⅴ shows the experimental results of our algorithm ACRI in comparison with other traditional algorithms. It can be observed from table Ⅴ that there are 5 mistakes made by MEME[11] (http://meme.nbcr.net/meme4_4_0/intro.html), which are the 5th, the 11th, the 15th, the 16th and the 17th sequences. The differences between the five sequences and the known ones are more than 10 (shown as the number in the shadow in Table Ⅴ). Similarly, Each of algorithms Gibbs Sampler [9] (http://bayesweb.wadsworth.org/gibbs/gibbs.html) and AlignACE[7] made 2 mistakes. Because the similarity of the 17th sequence's binding site is lower than others, three software mentioned above can not find its binding site. However our algorithm ACRI has successfully found all binding sites of these 18 sequences. Especially for the

© 2013 ACADEMY PUBLISHER

17th sequence, although its corresponding difference is lager than others, our algorithm can thoroughly find it. The main reason for its success is that our algorithm ACRI uses powerful optimization ability of ant colony algorithm and the local search method so that the motif we found has much higher information content, which can be seen from Table Ⅵ. Thus it can be seen that the results of our algorithm ACRI is more precise than other traditional algorithms, and it is really very accurate and effective to solve regulatory elements identification problems.

TABLE VI.

COMPARISON OF THE COMPUTATION INFORMATION CONTENT WITH DIFFERENT SOFTWARE

Software	Information Content
ACRI	10.273
MEME	9.508
AlignACE	9.752
Gibbs Sampler	9.229

B. The Running Speed Analysis of Our Algorithm

To validity the speed of our algorithm ACRI, we use the promoter binding sites of RAP1 as the test data since it has the largest volume of data. In the experiment, the initial characteristic matrix M_{best} is composed of some random numbers and contains 20 groups of initial Motifs. We test our algorithm ACRI and the greedy-based algorithm Consensus on these initial motifs to compare their computation time. Table Ⅶ shows the comparison of the computation time of the two algorithms. From table Ⅶ, we can see that the speed of algorithm ACRI is 8-12 times faster than greedy algorithms, which indicates that the efficiency of our algorithm is superior over the others.

TABLE VII.

COMPARISON OF THE COMPUTATION TIME BETWEEN ACRI AND GREEDY ALGORITHM

No.	Iteration number	ACRI(ms)	Greedy algorithm(ms)
1	40	189	1875
2	22	107	1031
3	26	103	1078
4	22	96	1062
5	20	87	875
6	23	101	1000
7	25	103	1235
8	32	169	1640
9	28	133	1313
10	28	125	1250
11	22	94	984
12	24	119	1313
13	30	140	1484
14	24	107	1141
15	25	130	1328
16	22	101	1016
17	26	107	1109
18	25	107	1109
19	28	124	1235
20	29	122	1281

We also compared the running time of our algorithm with other algorithms such as AlignACE, MEME, Gibbs Sampler etc. Table VIII shows the comparison of the computation time of ACRI and other algorithms. From table VIII, we can see that the speed of ACRI is much faster than all of other algorithms. For instance, algorithm ACRI is eight times faster than Gibbs Sampler. This indicates our algorithm is more efficient.

TABLE VIII.

COMPARISON OF THE COMPUTATION TIME BETWEEN ACRI AND TRADITIONAL ALGORITHMS

No.	Iteration number	AlignACE (ms)	MEME (ms)	Gibbs Sampler(ms)	ACRI (ms)
1	40	1565	1912	1890	179
2	22	948	1231	1048	101
3	26	897	1304	1101	99
4	22	843	1239	1072	92
5	20	527	987	903	83
6	23	723	1105	1021	97
7	25	1001	1407	1395	91
8	32	1214	1842	1640	155
9	28	998	1479	1298	119
10	28	998	1250	1257	111
11	22	623	1084	992	80
12	24	1223	1512	1279	105
13	30	1100	1620	1348	126
14	24	841	1148	1037	93
15	25	1232	1491	1385	116
16	22	704	1230	1054	87
17	26	811	1156	1063	93
18	25	797	1009	982	91

Since most of current regulatory elements identification algorithms use local search approach which neither guarantees the optimal results nor reduces time complexity. Our algorithm ACRI has higher optimization ability due to the powerful optimization ability of ant-colony algorithm. Algorithm ACRI can not only avoid converging to local optimal, but also greatly improve the efficiency.

VI. CONCLUSION

Most of current regulatory elements identification algorithms are easily to converge into a local optimum, and have high time complexity. Based on the ant colony optimization, we propose a novel algorithm named ACRI for regulatory elements identification. Due to the powerful optimization ability of ant-colony algorithm, the algorithm ACRI can not only improve the quality of results, but also solve the problem in a very high speed. Experimental results show that ACRI have not only higher quality of solutions but also less computational complexity compared with other traditional algorithms.

ACKNOWLEDGEMENTS

This research was supported in part by the Chinese National Natural Science Foundation under grant Nos. 61070047、61070133 and 61003180,Natural Science Foundation of Jiangsu Province under contracts BK2010318, BK21010134, and Natural Science Foundation of Education Department of Jiangsu Province under contract 12KJB520019.

© 2013 ACADEMY PUBLISHER

REFERENCES

[1] J.van Helden,B.Ander,and L.Collado-Vides. Extracting regulatory sites from upstream region of yeast genes by computational analysis of oligonucleotide frequencies, J.Mol.Biol., vol.17,pp.520-525,2001.

[2] Stormo G. DNA Binding Sites: Representation and Discovery. Bioinformatics, 16:16-23,2000.

[3] Bulyk ML:Computational prediction of transcription-factor binding site locations.Genome Biol, 5:201,2003.

[4] Qiu P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network.Biochem Biophys Res Commun 309:495-501,2003.

[5] Lawrence C,Reilly A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. Proteins,7:41-51,1990.

[6] Liu J,Neuwald A,Lawrence C. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies.J Am.Stat. Assoc., 90(432):1156-1170,1995.

[7] Andrew Moore. Very Fast EM-based Mixture Model Clustering Using Multiresolution KD-trees.Advances in Neural Information Processing Systems,1999.

[8] Liu X,Brutlag D,Liu J. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.In Proc.Pacific Symposium on Biocomputing,6:127-138,2001.

[9] Thijs G,Moreau Y,Rombauts S,et al. A Gibbs sampling method to detect over-represented motifs in the upstream regions of co-expressed genes.J.Comput.Biol., 9(2):447-464,2002.

[10] G.Z.Hertz,G.D.Stormo.Identifying DNA and protein patterns with statistically significant alignments of multiple sequences,Bioinformatics 15:563–577,1999.

[11] T.L.Bailey,C.Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers,Proc.Int.Conf.Intell.Syst.Mol.Biol.,2:28-36,1994.

[12] C.T.Workman,G.D.Stormo.ANN-Spec:a method for discovering transcription factor binding sites with improved specificity,Pac.Symp.Biocomput.,467–478,2000.

[13] J.Buhler,M.Tompa.Finding motifs using random projections,J.Comput.Biol.,9:225–242,2002.

[14] X.S.Liu,D.L.Brutlag,J.S.Liu. An algorithm for finding protein–DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments, Nat.Biotechnol, 20: 835–839,2002.

[15] S.Sinha,M.Tompa.YMF:a program for discovery of novel transcription factor binding sites by statistical overrepresentation,Nucleic Acids Res,31:3586–3588,2003.

[16] Zhibin Liu, Ling Zhang, Xiangsong Meng. A Novel Hybrid Stochastic Searching Algorithm Based on ACO and PSO: A Case Study of LDR Optimal Design. Journal of Software, 6(1):56~63,2011.

[17] Krzysztof Socha, Marco Dorigo. Ant colony optimization for continuous domains. European Journal of Operational Research, 185(3): 1155~1173,2008.

[18] D.D.Duc, H.Q. Dinh, H.H. Xuan. On the Pheromone Update Rules of Ant Colony Optimization Approaches for the Job Shop Scheduling Problem. Intelligent Agents and Multi-Agent Systems,5357:153~160,2008.

[19] Jingyao Li, Shudong Sun, Yuan Huang. Adaptive Hybrid ant colony optimization for solving Dual Resource Constrained Job Shop Scheduling Problem. Journal of Software, 6(4):584~594,2011.

[20] Wolfram Wiesemann, Thomas Stützle. Iterated Ants: An Experimental Study for the Quadratic Assignment Problem. ANTS Workshop, 179~190,2006.

[21] V.Maniezzo,A.Carbonaro. An ANTS heuristic for the frequency assignment problem. Future Generation Computer Systems, 16(8):927~935,2000.

[22] Julia Handl, Joshua D. Knowles, Marco Dorigo. Ant-Based Clustering and Topographic Mapping. Artificial Life, 12(1): 35~62,2006.

[23] Jiangqing Wang, Rongbo Zhu. Efficient Intelligent Optimized Algorithm for Dynamic Vehicle Routing Problem. Journal of Software, 6(11):2201~2208,2011.

[24] Zhu J, Zhang MQ:SCPD:a promoter database of the yeast Saccharomyces cerevisiae. Bioinformatics,15:607-611,1999.

[25] Stormo GD and Hartzell GW 3rd. Identifying protein-binding sites from unaligned DNA fragments. Proc Natl Acad Sci U S A 86(4):1183-1187,1989.

Wei Liu was born in Jiangyin, Jiangsu Province, P.R.China, in July 1, 1982. She received B. Sc degree and M. Sc degree in computer science from Yangzhou University, P.R. China in 2004 and 2007 respectively. In 2010,she received Ph.D degree in the department of computer science from Nanjing University of Aeronautics and Astronautics. She is currently a lecturer and Master's Supervisor in the Institute of Information Science and Technology, Yangzhou University, Yangzhou, P.R.China.

Her research interest includes data mining, bioinformatics and parallel processing. She has published more than 30 papers in journals and conferences.

Hanwu Chen was born in 1955, Ph.D., Professor, Ph.D. supervisor. His current research interests include quantum computing, information theory.

Ling Chen was born in Baoying, 1951. Professor, Ph.D. supervisor. He is a member of IEEE and ACM. His research interest includes data mining, bioinformatics and parallel processing.

Yixin Chen was born in Yangzhou,1979, Ph.D., Associate Professor. He is a member of IEEE and AAAI. His research interest includes the general areas of nonlinear optimization, artificial intelligence, data mining, machine learning, cloud computing and game theory.

© 2013 ACADEMY PUBLISHER