Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 3854646, 12 pages
https://doi.org/10.1155/2019/3854646

Research Article

Discovery of DNA Motif Utilising an Integrated Strategy Based on Random Projection and Particle Swarm Optimization

Hongwei Ge,[1] **Jinghong Yu**,[1] **Liang Sun**,[1] **Zhen Wang**,[2] **and Yao Yao**[1]

[1]*School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China*
[2]*School of Mathematical Science, Dalian University of Technology, Dalian 116023, China*

Correspondence should be addressed to Hongwei Ge; hwge@dlut.edu.cn

Received 22 September 2018; Accepted 10 April 2019; Published 8 May 2019

Academic Editor: Kalyana C. Veluvolu

During the process of gene expression and regulation, the DNA genetic information can be transferred to protein by means of transcription. The recognition of transcription factor binding sites can help to understand the evolutionary relations among different sequences. Thus, the problem of recognition of transcription factor binding sites, i.e., motif recognition, plays an important role for understanding the biological functions or meanings of sequences. However, when the established search space processes much noise subsequences, many optimization algorithms tend to be trapped into local optimum. In order to solve this problem, a particle swarm optimization and random projection-based algorithm (PSORPS) is proposed for recognizing DNA motifs. First, a random projection strategy is employed to filter the noise subsequences for constructing the objective space. Moreover, the sequence segments distributed in the majority of DNA sequences can be obtained and used for the population initialization of PSO. Then, the motifs of DNA sequences can be automatically searched by using a designed PSO algorithm in the constructed l-mer objective space. Finally, to alleviate the base deviation and further improve the recognition accuracy, the two operators of associated drift and independent drift are performed on the optimization results obtained by PSO. The experiments are conducted on real-world biological datasets, and the experimental results verify the effectiveness of the proposed algorithm.

1. Introduction

In the process of gene expression and regulation, the DNA genetic information can be transferred to protein via transcription [1]. The transcription process starts at transcription factor binding sites (TFBS). Within the biosome, the transcription factor binding sites take effect simultaneously under specific organs or scenes; this leads to the generation of proteins with similar functions. Thus, the discovery of transcription factor binding sites helps to understand the evolutionary and functional relations among different sequences [2, 3]. TFBSs are often characterized by highly similar domain with consensus patterns, which are termed as motif. The objective of motif recognition is to find out such patterns. Formally, the motif recognition problem can be described as follows. Given a set of nucleotide sequences, find out the sequence segments that are mutated from the motifs in each of the sequence [4].

The motifs are often distributed in promoter sequences, and the promoters are mainly constituted by hundreds or thousands of nucleotides. While recognizing the motifs, a certain number of nucleotides are firstly selected from promoters, among which the relative techniques are applied to discover motifs. Insufficient number of nucleotides will lead to losing motif instances. On the other hand, overmuch number of nucleotides will lead to generating too much noise in nucleotides, so that the motifs are difficult to be identified [5]. Thus, 1000-2000 nucleotides starting from the initial transcription position are usually selected as the searching region. Besides, motif recognition task is required to perform in the regions containing a promoter. If the motif conservation is relatively low, it is difficult to recognize the motifs from the noise subsequences. Traditional gene sequencing methods obtain chromosome sequences by biological or chemical experiments [6, 7] and adopt the techniques such as DNA footprinting and chromatin immunoprecipitation.

However, these experiments are generally expensive and time-consuming. Moreover, they cannot be applied to the entire genome efficiently. With the development of gene sequence project and gene expression profile, much efforts turn to recognize the TFBSs by computing-based methods [8–10].

In recent years, many computational technologies have been proposed to discover DNA and protein motifs. In DNA sequences, the promoter is generally the base sequence with length 1000-2000, while the motif is generally the base sequence with length 8-15; thus, identifying motif is more difficult than identifying DNA sequence. Moreover, the motif often mutates, and the biological meanings of the mutations cannot be interpreted. Recently, many computing-based motif recognition algorithms have been validated to be able to predict the actual binding sites efficiently. In order to verify the performance of different motif recognition algorithms, Pevzner and Sze designed a set of challenges [11], i.e., the planted (l, d) motif search problems, generally called PMP problems. The PMP problem is to extract the common substrings that appear in input strings with some mismatches allowed. The existing recognition algorithms for PMP problems can be classified into two categories, i.e., the exact algorithms and the approximation algorithms.

The widely used exact algorithms for PMP include qPMS series [12], iTriplet [13], PairMOtif series [14], RefSelect [15], and RecMotif [16]. All the exact search algorithms require exponential search time. When l and d are quite large, these algorithms require substantial amount of computational efforts. In biological applications, exact optimal solutions are not always needed, and an efficient solution is to explore the pseudo-optimal solutions with less computational efforts incurred. Thus, it is not practicable to recognize the exact optimal solution.

The approximation algorithms generally fall into two categories, i.e., the statistical approaches based on the position weight matrix (PWM) and the enumeration approaches based on the consistency of motif. The statistical approaches use PWM to represent motif and convert the PMP problem into a statistical problem to search the optimal solution. The main drawback of this category of algorithms is that the search is easy to be trapped into local optimum. MEME [17] is a statistical based motif recognition algorithm, which adopts EM algorithm to converge to an optimization solution, and the performance of MEME relies on the initial values. The enumeration methods initialize a motif with a mutation position and then identify the sequences similar to the initialized motif. Then, the conservation of the identified sequences is evaluated, and the most conservative motif instances are combined into the final motif. The algorithms belonging to this category include WINNOWER [18], Projection [19], TreeMotif [20], VPA [21], and CVoting [22].

Recently, population-based swarm intelligence algorithms have been proposed to address PMP problem [23–26]. Representative methods of this kind include PSOGAP [27] and PSO-MAC [28], and PSO-KNN [29]. The PSOGAP enumerates all the l-mers and constructs a similarity matrix for the l-mers. The search space is then converted from discrete space to a semicontinuous space. The fitness function in the optimization process is calculated according to the constructed matrix. The PSO-MAC is achieved by combining PSO and motif alignment clustering (MAC). PSOKNN integrate the PSO with k-nearest neighbors to solve the planted (l, d)-motif finding problem. Before adjusting the trajectory of each individual using PSO, k-nearest neighbors' algorithm is applied to the initial population for performing extra searches. The swarm-based algorithms facilitate research into PMP problems. However, if there is too much noise in the objective space, this kind of algorithms can easily be trapped into local optimum solutions. To address this problem, we propose an automatic motif discovery algorithm for DNA sequences based on random projection and particle swarm optimization (PSORPS). In PSORPS, random projection strategy is used to filter the noise subsequences, and then the preserved subsequences can be used for constructing the objective space, so that the influence of noise subsequences on the recognition of the motif can be reduced. Further, the algorithm can obtain sequence segments distributed in the majority of DNA sequences, and then the obtained sequence segments are taken as candidates for the initial individuals of PSO to improve the search performance. Besides, the operators of independent drift and associated drift alleviate the problem of base deviation and enhance the applicability of the PSO for solving motif recognition problems.

2. Random Projection Strategy

For the (l, d) motif recognition problem, let l-mer set be the set of all possible continuous subsequences with length l. The random projection strategy (RPS) partitions all the l-mers (l-length substring of sequences) subsequences into different groups. Firstly, K base sites are randomly selected to construct a projection site set. Further, RPS keeps the base in the projection site set unchanged and groups the l-mers by using harsh computing. This process can be regarded as projecting l-mer from l dimensional space to K dimensional subspace. Let $p = \{p_i \mid 1 \le i \le K \le l, 1 \le p_i \le l\}$ be the site set with K different sites, and $1 \le p_i < \ldots < p_k \le l$ corresponds to the sites in the l-mer. Let $s = \{s_1, s_2, \ldots, s_t\}$ denote l-mer with length l. In this paper, we define $proj(s \mid p) = \{s_{p1}, s_{p2}, \ldots, S_{pK}\}$ as the projection of l-mer on p. For example, when the site set P is (2, 3, 5, 6, 7, 9), the l-mer set is AGCTTAGACT, TGCATAGTCA, and GCCGTAGCCA; then, $proj_1(s \mid p)$ = (GCTAGC), $proj_2(s \mid p)$ = (CCTAGC). We further digitalize the representation of l-mers: let 1 represent A, 2 represent C, 3 represent G, 4 represent T, and 0 represent nonbase coding. Assign a weight to each of the sites in the l-mer, and let the weight of the j-th site be

$$w(j) = b^{j-1}, \tag{1}$$

where b is taken as 5 for nucleotide sequences. Let X_j be the integer that corresponds to the j-th site of l-mer, and then define the weight of l-mer as follows:

$$F(\mathrm{s}) = \sum_{j \in p} X_j w(j). \tag{2}$$

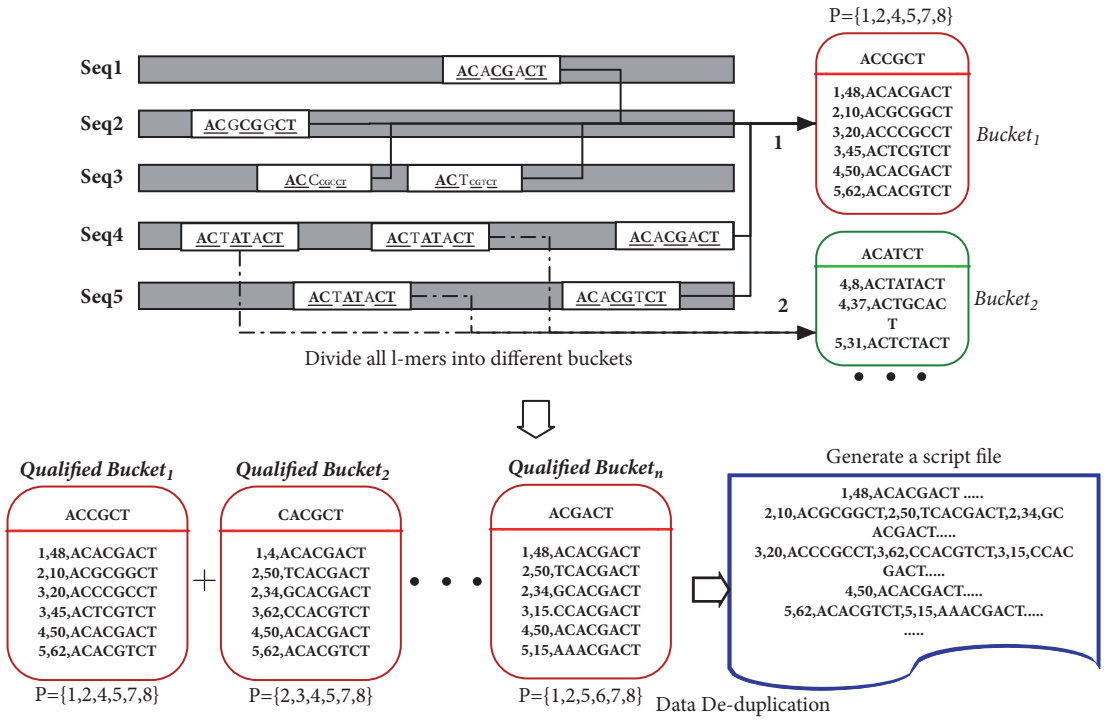


Figure 1: Illustration for the process of random projection strategy.

When conducting harsh computing, we project l-mer to its corresponding bucket according to the harsh value calculated by harsh function. The harsh function is as follows:

$$h_p(s) = \left(\sum_{j \in p} X_j w(j) \right) \mod B, \tag{3}$$

where $X_j \in [1, 4]$, whose value is taken as the j-th base of l-mer, B is the size of harsh table, and the value is usually taken as the number of all the l-mer sets. When the projection is completed, the hash table is destroyed. Use formula (3) to calculate the hash value of the l-mer, and deliver the l-mer into the corresponding bucket according the obtained hash value. In the bucket, the sequence number, the location, and the base sequence of each l-mer are recorded. And then the buckets in which the number of l-mers exceeds the threshold σ are considered to be eligible. The eligible buckets are sorted according to the number of the sequences related to the l-mer in each bucket, and the first n buckets are selected, where n is population size in PSO. All the subsequences in the selected buckets are taken as candidates for the initial individuals of PSO. A qualified script is generated after removing the duplicate subsequences in all the eligible buckets. Finally, PSO is used to automatically search the optimal motif instances from the subsequences in the script. Figure 1 presents an example for illustrating the process of random projection. In the figure, the threshold value σ for the minimum number of l-mers is taken as 4. The number of l-mers in bucket 1 exceeds σ; thus, it is an eligible bucket, while the number of l-mers in bucket 2 is less than σ; thus, it is not an eligible bucket. The qualified script is generated by the above procedures.

3. Motif Recognition Algorithm Based on PSO and RPS

The proposed PSO and PRS-based motif recognition algorithm (PSORPS) works as follows. The positions of the sequences in the qualified script generated by random projection are taken to be solution space of PSO, not the positions of all the l-mers. Under the assumption that each sequence has only one motif instance, the PSO algorithm randomly selects a set of l-mers to be starting positions from the sorted buckets. The particles are then updated by adjusting the trajectory of each individual towards its own best location and towards the best particle of the swarm at each generation. In the process, the fitness value is calculated according to the position-specific frequency matrices (PSFM) and Bayesian scoring function. The detailed descriptions of the algorithm are as follows.

3.1. Establishment of Search Space. In the PSORPS, each particle is described by a position vector $\overrightarrow{X}$, which denotes all the possible motif positions in the objective search space obtained after projection. It is worth noting that if the set of l-mers in the script is directly used as the search space, the algorithm cannot achieve a good convergence performance. For example, in Figure 2, P_I denotes the position of a sequence in the qualified script. Before sequence sorting, the l-mers are CTCAC and GTGTA in the positions $P_I = 2$ and $P_I = 3$, respectively. Although they are adjacent, they do not have similar patterns. Therefore, the particles maybe wander blindly in the search space. In order to facilitate converging of PSO, we sort the eligible l-mers in alphabet order using

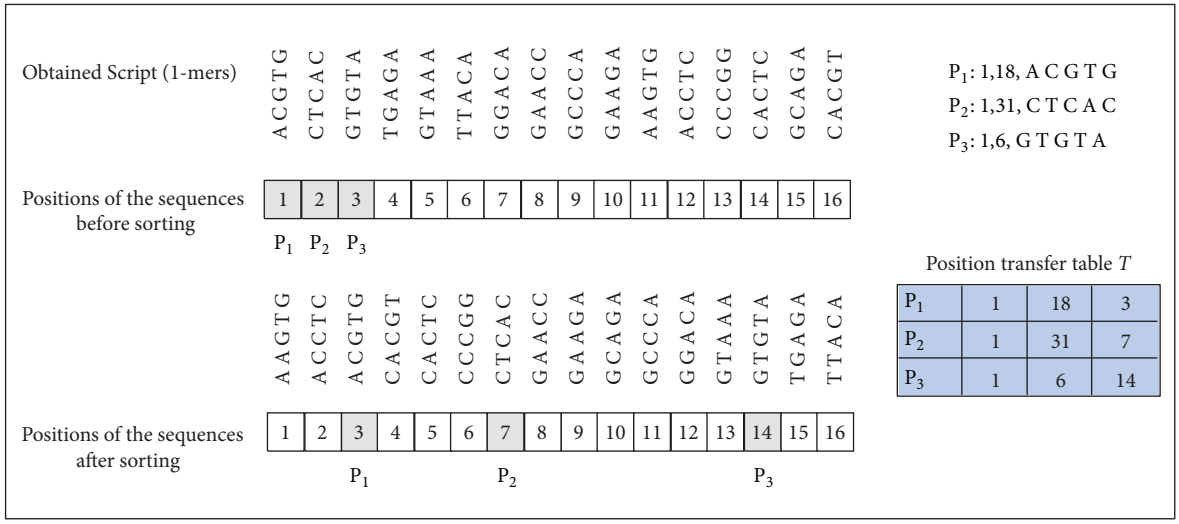


Figure 2: An illustrative example for sorting the l-mers in qualified script.

the continuous ordinal space constructing method proposed in [27]. In this way, an effective l-mer objective search space can be established. Figure 2 presents an illustrative example for the above process. While sorting, a position transfer table T is created to record the position of l-mer in the original sequence, and the positions in the qualified script before sorting and after sorting.

3.2. Initialization and Fitness Computation. Suppose there are t sequences with length N in the DNA set. For each particle in the initial population, randomly select t l-mers from the sorted buckets and ensure that each DNA sequence has a l-mer selected. The selected l-mer set constitutes an initial motif instance set denoted as $L_i = (L_1, L_2, \ldots, L_t)$. Then, the position vector $\overrightarrow{X}_i$ of these motif instances in the constructed continuous ordinal space is taken as the initialization vector of the i-th particle. So the initial position of each particle represents a starting site of the possible motif instance in each sequence.

For evaluating the performance of particles, the position-specific frequency matrices (PSFM) are employed firstly. These matrices give the information on the frequency of each base at matrix. And then the Bayesian scoring function is used to calculate the fitness of particles:

$$\psi\left(\overrightarrow{X_i}\right) = |A|\left(\log\left(\frac{\theta}{1-\theta}\right) - 1 + \sum_{i=1}^{l}\sum_{j=1}^{4} p_{ij}\log\left(\frac{p_{ij}}{q_j}\right)\right), \quad (4)$$

where $|A|$ is the number of motifs to be predicted, $\theta = |A|/L$ is the predicted motif abundance out of L total possible locations, $L = \sum_t (L_t - l + 1)$ is the number of latent motif instances, L_t is the length of the t-th sequence, l is the length of the motif, $\sum_{i=1}^{l}\sum_{j=1}^{4} p_{ij}\log(p_{ij}/q_i)$ measures the amount of information in the sequence, p_{ij} is the frequency that the nucleotide j appears at the i-th nucleotide position, and q_j is the frequency that the nucleotide j appears in the background sequence.

3.3. Velocity and Position Updating. PSO searches the optimal l-mer set with the maximum value of Ψ using the velocity and position updating equations. The searches are performed in the l-mer objective space established in Section 3.1, but not the space including all the l-mers. While the particle updates the global best fitness Ψ_g and its local best fitness Ψ_i, the position vectors $\overrightarrow{O}_g$ and $\overrightarrow{O}_i$ of the motif instances are recorded. The update equation of velocity for particles is as follows:

$$\overrightarrow{V}_i^{k+1} = \alpha\overrightarrow{V}_i^{k} + \rho_1 R_I\left(\overrightarrow{O}_i^{k} - \overrightarrow{X}_i^{k}\right) + \rho_2 R_S\left(\overrightarrow{O}_g^{k} - \overrightarrow{X}_i^{k}\right), \quad (5)$$

where $\overrightarrow{X}_i^k$ is a t-dimensional position vector which represents the position of the i-th particle in the objective space at the k-th iteration, $\overrightarrow{V}_i^k$ is the speed of the i-th particle at the k-th iteration, α is the inertia weight, R_I and R_S are $t * t$ diagonal matrices, and their diagonal elements γ_{jj} are random numbers uniformly distributed in the interval $[0, 1]$. ρ_1 and ρ_2 are two coefficients balancing the individual and social cognition. Let the distance vectors $\overrightarrow{d}_i = \overrightarrow{O}_i^k - \overrightarrow{X}_i^k$ and $\overrightarrow{d}_g^k = \overrightarrow{O}_g^k - \overrightarrow{X}_i^k$; then, the formula $R_i(\overrightarrow{O}_i^k - \overrightarrow{X}_i^k) = R_i\overrightarrow{d}_i$ can be extended as

$$\begin{bmatrix} \gamma_{11} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \gamma_{tt} \end{bmatrix} * \begin{bmatrix} d_{i1} \\ \vdots \\ d_{it} \end{bmatrix} = \begin{bmatrix} \gamma_{11}d_{i1} \\ \vdots \\ \gamma_{tt}d_{it} \end{bmatrix}. \quad (6)$$

The updating formulation for the position of the particles is as follows:

$$\overrightarrow{X}_i^{k+1} = \overrightarrow{X}_i^{k} + \left\lfloor \overrightarrow{V}_i^{k+1} \right\rfloor, \quad (7)$$

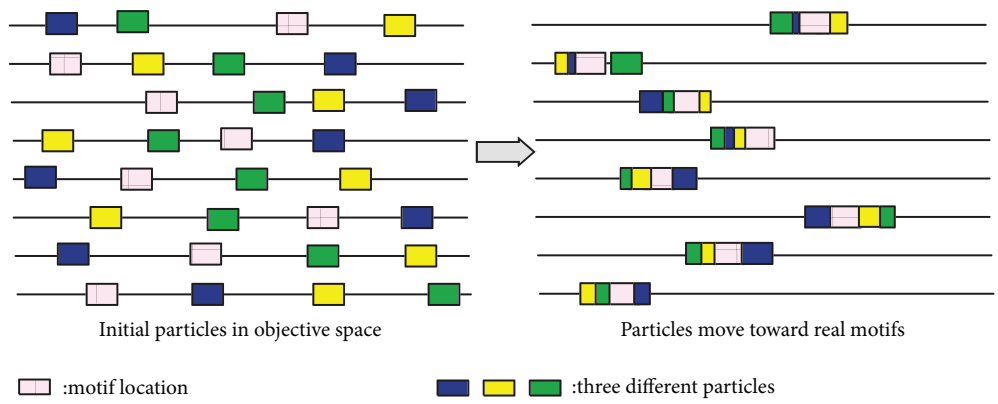


FIGURE 3: Interpretation of particle trajectories in the search space.

where $\overrightarrow{X}_i^{k+1}$ is a nonnegative vector. Since the particles move in the l-mer objective space, the maximum value of the particle $\overrightarrow{X}_i$ in each dimension is the maximum number of l-mers in the same dimension. And a feasible random number is triggered if $\overrightarrow{X}_i$ exceeds the bounds.

Figure 3 interprets an instance of particle trajectories in the search space. In the figure, three particles move from the initial positions to the positions near the real motif instances. Besides, when the particle moves to a new position, each dimension x_{ij} of $\overrightarrow{X}_i^k$ is further moved by five positions on its left and five positions on its right to generate its neighbors. We evaluate the fitness for these new offsprings. And the one that has the highest fitness value will be used to update the position of the particle. In order to avoid premature convergence, the algorithm will restart initialization when the global optimal solution does not improve for 30 successive iterations.

3.4. Associated Drift and Independent Drift for Refinement. When the maximum number of iterations is reached, the algorithm will stop searching and output the optimal solution. However, there may be two problems in the optimization process. One is the base shifting, and the other is the deviation of solution space in the establishment process of the reduced objective space. To overcome the two problems, the associated drift and independent drift are further proposed in this paper. The two operations will be conducted on the optimal solution obtained by the PSORPS for further refinement.

Let $\overrightarrow{X}_{best}$ be the optimal solution obtained by PSORPS. Before implementing the two refinement operations, $\overrightarrow{X}_{best}$ needs to be mapped to the original l-mer space using position transfer table T. Denote the optimal solution after mapping as $\overrightarrow{X}_{best}^{\text{org}}$. In the process of independent drift refinement, an element x_i in $\overrightarrow{X}_{best}^{\text{org}}$ is randomly selected to move step by step in the original l-mer space and all the other elements remain unchanged. Recalculate the fitness value Ψ' of each new obtained solution. If the Ψ' value of new solution $\overrightarrow{X}'^{\text{org}}_{best}$ is greater than the Ψ value of $\overrightarrow{X}_{best}^{\text{org}}$, then replace $\overrightarrow{X}_{best}^{\text{org}}$ with $\overrightarrow{X}'^{\text{org}}_{best}$. If the Ψ' value of new solution $\overrightarrow{X}'^{\text{org}}_{best}$ is smaller than the Ψ value of $\overrightarrow{X}_{best}^{\text{org}}$, then move x_i forward. This process is repeated until the entire sequence is traversed. Figure 4(a) presents an example of independent drift refinement, in which the grey element is randomly selected and it moves step by step in the original l-mer space until a better solution is found or the entire sequence is traversed.

In the associated drift refinement, move the elements of the optimal solution $\overrightarrow{X}_{best}^{\text{org}}$ towards the right in the original *l-mers* space to obtain the new solution $\overrightarrow{X}_{best}^{\text{org}} = \{x_1 + 1, x_2 + 1, \cdots, x_t + 1\}$, or towards the left to obtain the new solution $\overrightarrow{X}'^{\text{org}}_{best} = \{x_1 - 1, x_2 - 1, \cdots, x_t - 1\}$. If the Ψ' value of new position $\overrightarrow{X}'^{\text{org}}_{best}$ is greater than the Ψ value of $\overrightarrow{X}_{best}^{\text{org}}$, then replace $\overrightarrow{X}_{best}^{\text{org}}$ with the new solution. This operator is conducted until there is no improvement on Ψ value. Figure 4(b) presents an example of associated drift refinement. Figure 5 presents the overall flowchart of the PSORPS for motif recognition.

4. Experimental Results and Analysis

4.1. Parameter Settings and Evaluation Metrics

4.1.1. Projection Size K. The parameter of projection size K determines the number of selected l-mers. The appropriate projection size promotes the real motif instances to be preserved in the objective space. If the value of K is too small, it will introduce too much noise l-mers and the formed objective space is also too large, which would affect the accuracy of the prediction. If the value of K is too large, the real motif instances will be lost with a large probability. In DNA sequence, if the projection size is K, then 4^K buckets will be generated, and the total number of l-mers generated by the entire sequence set is $t*(n-l+1)$. Thus, the number of l-mers in each bucket is

$$Q = \frac{t(n-l+1)}{4^K}, \tag{8}$$

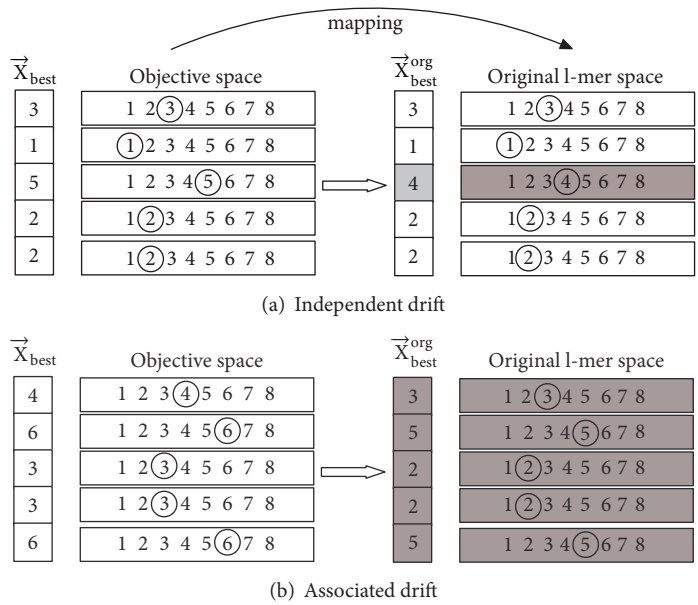


Figure 4: An instance of independent drift and associated drift operators.

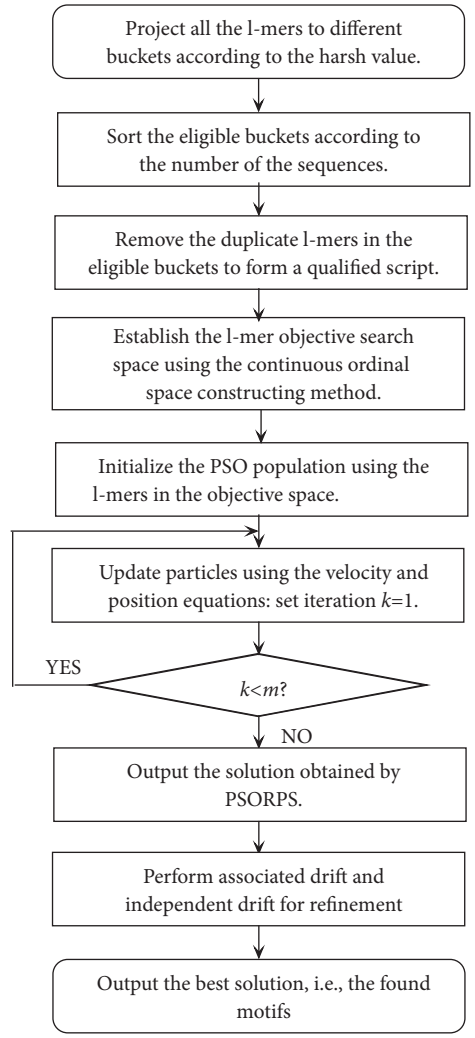


Figure 5: Flowchart of the proposed PSORPS for motif recognition.

where t is the number of sequences, n is the length of sequences, and l is the length of l-mer. In the implementation, Q is usually set to be less than 1, and 0.9 is taken in this paper. Thus, K can be obtained by

$$K = \left\lceil \log_4 \frac{t\,(n-l+1)}{0.9} \right\rceil. \quad (9)$$

In the planted (l, d) motif recognition problem, each motif instance with length l is obtained by muting d elements of the motif. Thus, the probability for an eligible bucket containing the motif instance is

$$p\,(l,d,K) = \frac{C_{l-d}^{K}}{C_{l}^{K}}. \quad (10)$$

Suppose that the number of motifs in the entire DNA sequences is r; then, the Bernoulli experiment should be conducted for r times. Let $B_{r,p(l,d,K)}(\sigma)$ denote the probability that the number of motif instances deposited to the eligible buckets in each projection does not exceed the threshold σ; then, the following formula can be obtained:

$$B_{r,p(l,d,K)}\,(\sigma) = \sum_{i=0}^{\sigma-1} C_r^i\, p\,(l,d,K)^i \left(1 - p\,(l,d,K)\right)^{r-i}. \quad (11)$$

So when the projection is conducted for m times, the probability that the number of motif instances being projected to the eligible buckets is no less than σ is

$$q = 1 - \left[B_{r,p(l,d,K)}\,(\sigma)\right]^m. \quad (12)$$

According to [18], $q \geq 0.95$; then, m can be taken as

$$m = \left\lfloor \frac{\log\,(1-q)}{\log\left(B_{r,p(l,d,K)}\,(\sigma)\right)} \right\rfloor. \quad (13)$$

4.1.2. Threshold Value for the Minimum Number of l-Mers in Eligible Bucket. In the implementation, the threshold value σ for the minimum number of l-mers in the eligible bucket is set according to the size of the DNA sequence. Reference [18] shows that if the test problem contains 4 to 20 motifs, and the length of the sequence ranges from 600 to 1000, σ is usually taken as 4.

4.1.3. Parameters for PSO. In the implementation, the two cognition coefficients are taken as $\rho_1 = 1.0$ and $\rho_2 = 1.0$, inertia weight α is taken as 0.8 through empirical study, the maximum number of iterations is taken as 1000, and the population size in PSO is taken as 100.

4.1.4. Evaluation Metrics. In the experiment, we select nucleotide level performance coefficient (nPC) to evaluate the performance of the algorithm, which is commonly used to test the performance of the motif recognition algorithms [6, 12]. The nPC is proposed by Pevzner and Sze [11] and it is defined by

$$nPC = \frac{nTP}{nTP + nFN + nFP}, \quad (14)$$

where the elements in (14) are defined as follows:

(i) nTP is the number of nucleotide positions in both known sites and predicted sites

(ii) nFN is the number of nucleotide positions in known sites but not in predicted sites

(iii) nTN is the number of nucleotide positions in neither known sites nor predicted sites

(iv) nFP is the number of nucleotide positions not in known sites but in predicted sites.

4.2. Experiments on CRP Dataset. The *Escherichia coli* cyclic AMP receptor protein (CRP) dataset consists of 18 sequences. In the dataset, the length of each sequence is 105, and the length of transcription factor binding sites is 22. To compare the performance of the proposed algorithm, we select motif consensus-based random projection algorithm (Projection) [19], position weight matrix-based statistical algorithm (MEME) [17], and particle swarm optimization-based intelligent algorithm (PSOKNN) [29] for comparisons. Table 1 presents the comparison results. The contents include the known motif locations, the locations of the predicted motif, and the deviations between the known motifs and the predicted motifs. It can be seen from Table 1 that the Projection, MEME, and PSOKNN fail to find the motif location for some instances, and the PSORPS predicted motif locations are the closest to the ground-truth ones. The results indicate that the PSORPS can effectively predict the locations of the motifs. To present the results intuitively, the motif logo graph [30] introduced by Schneider and Stephens is depicted in Figure 6. The X-axis presents the position of the nucleotide; at each position, four letters (A, C, G, and T) are distributed with their frequencies symbolized by the letter heights, and a higher letter represents its occurrence in higher frequency. It can be seen that the location has the deviation with three bases for the cole1 data in the motif logo graph obtained by using Projection algorithm and has the deviation with two bases when using the PSOKNN. The motif logo graph obtained by PSORPS is more similar to the real motif logo graph. It can also be seen that all the algorithms can predict the majority of motif instances. However, the PSORPS obtains more accurate results.

4.3. Experiments on Eukaryotes Gene Dataset. In Eukaryotes gene dataset, the five large scale DNA sequences in Table 2 are selected, which include the Preproinsulin, c-fos, Metallothionein, dihydrofolate reductase (DHFR), and Yeast ECB [14]. The details of the species information, motif instances, length of sequence, and the starting position of the motif instances are presented in Table 2. As can be seen from Table 2, the sequence length in the dataset ranges from 250 to 8351. Each experiment is run independently 10 times, and the nPC index is used to evaluate the performance.

Table 1: Prediction accuracy of different algorithms on CRP dataset. Table 1 is reproduced from a preliminary work in BIBM2017 [31] by Ge et al. under the Creative Commons Attribution License/public domain.

CRP (22, 6)	Known motif locations	Projection (Error)	MEME (Error)	PSOKNN (Error)	PSORPS (Error)
cole1	17, 61	64 (3)	61	63 (2)	61
Ecoarabop	17, 55	58 (3)	55	57 (2)	55
ecobglr1	76	79 (3)	76	78 (2)	76
Ecocrp	63	66 (3)	63	65 (2)	63
Ecocya	50	53 (3)	13 (-37)	52 (2)	50
Ecodeop	7,60	10 (3)	7	9 (2)	7
Ecogale	42	45 (3)	42	44 (2)	24 (-18)
Ecoilvbpr	39	42 (3)	39	41 (2)	39
Ecolac	9,80	12 (3)	9	11 (2)	9
Ecomale	14	17 (3)	14	16 (2)	14
Ecomalk	61	64 (3)	61	63 (2)	61
Ecomalt	41	44 (3)	41	43 (2)	41
Ecomap	48	51 (3)	48	50 (2)	48
Ecotnaa	71	74 (3)	71	73 (2)	71
ecouxul	17	20 (3)	75 (58)	19 (2)	17
pbr-p4	53	56 (3)	53	55 (2)	53
trn9cat	1,84	8 (7)	27 (26)	75 (-9)	5 (4)
Tdc	78	81 (3)	76 (-2)	80 (2)	78

Table 1 is reproduced from Ge et al. (2017BIBM).

Table 2: Ground-truth motifs and sequence characteristics of 5 DNA sequences in Eukaryotes gene data set.

DNA	Ground-truth motif	Species (length)	Starting position of motif
Preproinsulin (15,2)	CAGCCTCAGCCCCCA	AOTUS (2113)	386
	CAACCTCAGCCCCCT	CAVIA (1472)	38
	CAGCCTCAGCCCCCA	HUMAN (4992)	2062
	CAGCTTCAGCCCCTC	RAT (2852)	822
DHFR (11,2)	TTTTCGTGGGA	MELANOGASTER (1954)	752
	ATTTCGCGCCA	CRIGR (1521)	825
	ATTTCGCGCCA	HUMAN (1133)	387
	ATTTCGCGCCA	MOUSE (1235)	942
c-fos (9,2)	CCAAACTGG	CHICK (600)	274
	CTACATTTG	FUGRU (600)	142
	CCATATTAG	HUMAN (600)	420
	CCATATTAG	MOUSE (600)	244
	TCAAATATG	TEFL (600)	557
Metallothionein (15,2)	CTCTGCGCCCGGCCC	BOVIN (2749)	466
	CTCTGCACCCCGCCC	CRIGR (1380)	1151
	CTCTGCACTCCGCCC	RAT (8351)	7013
	TTCTGCACACGGCAC	TROUT (1147)	992
Yeast ECB (16,3)	TTTCCCTTTTAGGAAA	CDC46 (250)	47
	TTTCCTTATAAGGAAA	CDC47 (300)	38
	TTTCCAGATCAGGAAA	CDC6 (300)	58
	TTACCCGTTTAGGAAA	CLN3 (1050)	96
	TTTCCCGTTTAGGAAA	SWI4 (550)	48

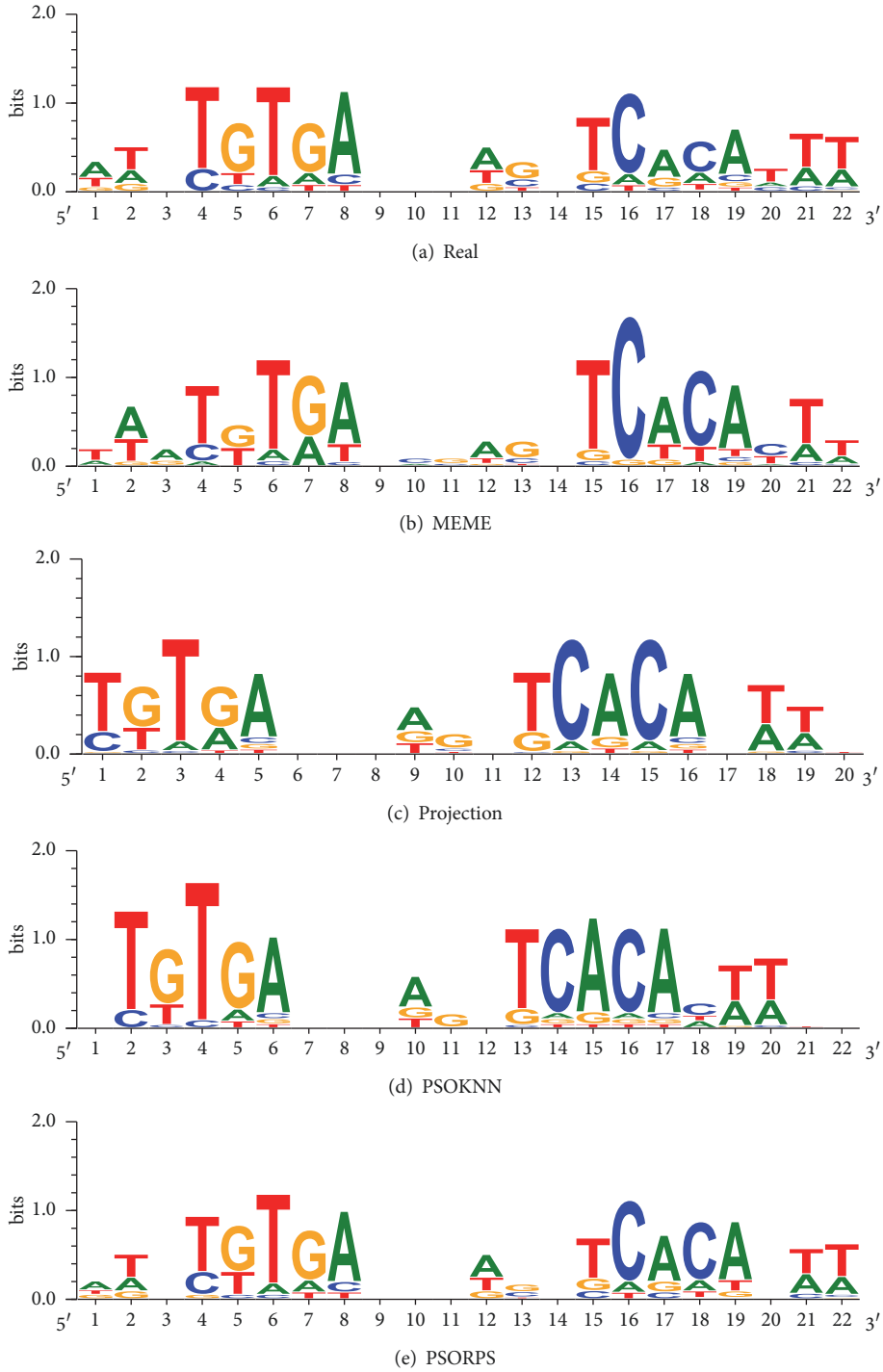


Figure 6: Logo graphs of predicted motif of different algorithms on CRP dataset.

The nPC values of the PSORPS and the compared algorithms are presented in Figure 7. On the Yeast ECB data, all the algorithms obtain the nPC value 1.0. The main reason is that the Yesast ECB data possesses the shorter sequence length and higher sequence conservation compared with the other four datasets. On the Metallothionein data, DHFR data, and Preproinsulin data, the PSORPS obtains the nPC values 0.857, 0.517, and 0.504, respectively. On the c-fos data, all the algorithms obtain low nPC, mainly because the the c-fos data has less sequence conservation. Figure 8

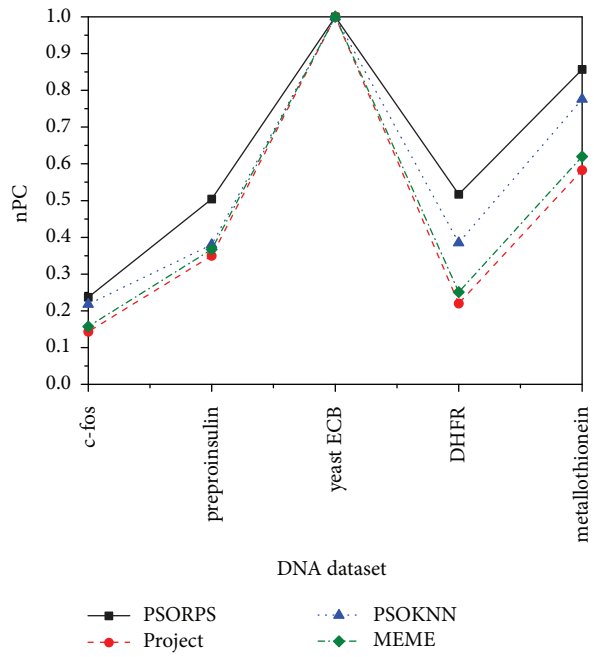


Figure 7: Prediction accuracy of different algorithms on real biological data.

presents the motif logo of the PSORPS and the compared algorithms. It can be seen that the motifs obtained by the PSORPS are identical to the ground-truth motifs on the Yeast ECB and Metallothionein data. On the Preproinsulin and DHFR data, the PSORPS approximates the starting site of the ground-truth motifs. In short, the results demonstrate that the PSORPS obtains better results than the compared algorithms.

5. Conclusion

Traditional gene sequencing methods and exact motif recognition algorithms are generally expensive and time-consuming. Swarm-based intelligence computation algorithms have been proposed to solve motif recognition problems. However, they usually have difficulties in achieving satisfactory results, especially when the objective search space possesses much noise subsequences. To address this problem, this paper proposes a motif discovery algorithm for DNA sequences based on random projection and particle swarm optimization (PSORPS). The random projection strategy is adopted to filter the DNA sequences and then the obtained eligible l-mers are used for initial population of the PSO. Then, the PSO automatically search the motifs of DNA sequences in the constructed l-mer objective space, and the particles are evaluated by Bayesian scoring function. To further improve the motif recognition accuracy, two local search operators of associated drift independent drift and independent drift are, respectively, designed to alleviate the problem of base deviation. The experiments are conducted on the Escherichia coli cyclic AMP receptor protein dataset and the Eukaryotes gene dataset. The comparative results indicate the proposed algorithm provide an effective way for DNA motif discovery.

Data Availability

The code used in this paper is released, which is written in matlab and available at https://github.com/MenSanYan/DNA_motif_discovery.

Disclosure

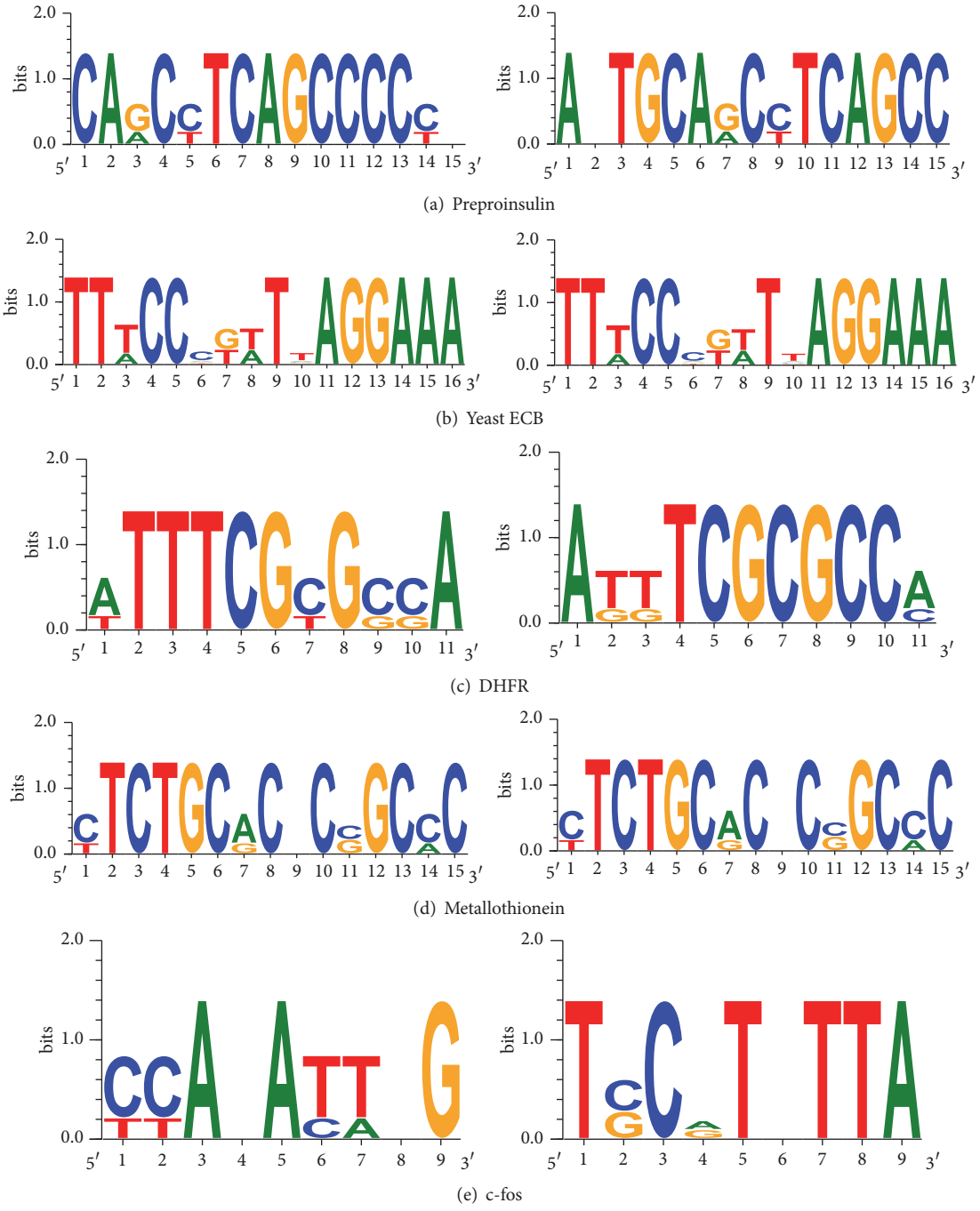
An earlier version of this manuscript was presented in 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM2017.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61572104, 61402076) and the project



(a) Preproinsulin

(b) Yeast ECB

(c) DHFR

(d) Metallothionein

(e) c-fos

FIGURE 8: Real motif logo and predicted motif logo.

of the Key Laboratory of Symbolic Computation and Knowledge Engineering in Jilin University (93K172017K03).

References

[1] O. A. Stasyuk, D. Jakubec, J. Vondrášek, and P. Hobza, "Noncovalent interactions in specific recognition motifs of protein-DNA complexes," *Journal of Chemical Theory and Computation*, vol. 13, no. 2, pp. 877–885, 2017.

[2] W. A. Al-Zyoud, R. M. G. Hynson, L. A. Ganuelas et al., "Binding of transcription factor GabR to DNA requires recognition of DNA shape at a location distinct from its cognate binding site," *Nucleic Acids Research*, vol. 44, no. 3, pp. 1411–1420, 2016.

[3] E. Czeizler, T. Hirvola, and K. Karhu, "A graph-theoretical approach for motif discovery in protein sequences," *IEEE Transactions on Computational Biology and Bioinformatics*, vol. 14, no. 1, pp. 121–130, 2017.

[4] F. Zambelli, G. Pesole, and G. Pavesi, "Motif discovery and transcription factor binding sites before and after the next-generation sequencing era," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 225–237, 2013.

[5] K. Chowdhury, S. Kumar, T. Sharma et al., "Presence of a consensus DNA motif at nearby DNA sequence of the mutation susceptible CG nucleotides," *Gene*, vol. 639, pp. 85–95, 2018.

[6] T. Q. Nguyen, K. W. Lim, and A. T. Phan, "A dual-specific targeting approach based on the simultaneous recognition of duplex and quadruplex motifs," *Scientific Reports*, vol. 7, no. 1, article 11969, 2017.

[7] P. Radecki, M. Ledda, and S. Aviran, "Automated recognition of RNA structure motifs by their SHAPE data signatures," *Gene*, vol. 9, no. 6, p. 300, 2018.

[8] D. L. González-Álvarez, M. A. Vega-Rodríguez, and Á. Rubio-Largo, "Multiobjective optimization algorithms for motif discovery in DNA sequences," *Genetic Programming and Evolvable Machines*, vol. 16, no. 2, pp. 167–209, 2015.

[9] D. Jakubec, R. A. Laskowski, and J. Vondrasek, "Sequence-specific recognition of DNA by proteins: binding motifs discovered using a novel statistical/computational analysis," *PLoS ONE*, vol. 11, no. 7, Article ID e0158704, 2016.

[10] N. K. Lee, F. L. Azizan, Y. S. Wong, and N. Omar, "DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery," *Biotechnology & Biotechnological Equipment*, vol. 32, no. 3, pp. 759–768, 2018.

[11] P. A. Pevzner and S. H. Sze, "Combinatorial approaches to finding subtle signals in DNA sequences," in *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology*, pp. 269–278, San Diego, Calif, USA, 2000.

[12] M. Nicolae and S. Rajasekaran, "qPMS9: an efficient algorithm for quorum planted motif search," *Scientific Reports*, vol. 5, no. 1, p. 7813, 2015.

[13] E. S. Ho, C. D. Jakubowski, and S. I. Gunderson, "iTriplet, a rule-based nucleic acid sequence motif finder," *Algorithms for Molecular Biology*, vol. 4, no. 1, pp. 1–14, 2009.

[14] Q. Yu, H. Huo, Y. Zhang, H. Guo, and H. Guo, "PairMotif+: a fast and effective algorithm for De Novo motif discovery in DNA sequences," *International Journal of Biological Sciences*, vol. 9, no. 4, pp. 412–424, 2013.

[15] Q. Yu, H. Huo, R. Zhao, D. Feng, J. S. Vitter, and J. Huan, "Reference sequence selection for motif searches," in *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM '15)*, pp. 569–574, IEEE, Washington, DC, USA, November 2015.

[16] H. Q. Sun, M. Y. H. Low, W. J. Hsu, and J. C. Rajapakse, "RecMotif: a novel fast algorithm for weak motif discovery," *BMC Bioinformatics*, vol. 11, no. 11, 2010.

[17] T. L. Bailey, M. Boden, F. A. Buske et al., "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. 2, pp. W202–W208, 2009.

[18] X. Yang and J. C. Rajapakse, "Graphical approach to weak motif recognition.," *Genome informatics. International Conference on Genome Informatics*, vol. 15, no. 2, pp. 52–62, 2004.

[19] J. Buhler and M. Tompa, "Finding motifs using random projections," *Journal of Computational Biology*, vol. 9, no. 2, pp. 225–242, 2002.

[20] H. Q. Sun, M. Y. H. Low, W. J. Hsu, C. W. Tan, and J. C. Rajapakse, "Tree-structured algorithm for long weak motif discovery," *Bioinformatics*, vol. 27, no. 19, Article ID btr459, pp. 2641–2647, 2011.

[21] M. M. Abbass and H. M. Bahig, "An efficient algorithm to identify DNA motifs," *Mathematics in Computer Science*, vol. 7, no. 4, pp. 387–399, 2013.

[22] Y. Xu, J. Yang, Y. Zhao, and Y. Shang, "An improved voting algorithm for planted (l, d) motif search," *Information Sciences*, vol. 237, pp. 305–312, 2013.

[23] D. L. González-Álvarez, M. A. Vega-Rodríguez, and Á. Rubio-Largo, "Convergence analysis of some multiobjective evolutionary algorithms when discovering motifs," *Soft Computing*, vol. 18, no. 5, pp. 853–869, 2014.

[24] Z. Cui and Y. Zhang, "Swarm intelligence in bioinformatics: Methods and implementations for discovering patterns of multiple sequences," *Journal of Nanoscience and Nanotechnology*, vol. 14, no. 2, pp. 1746–1757, 2014.

[25] D. L. Gonzalez-Alvarez, *Hybrid multiobjective artificial bee colony with differential evolution applied to motif finding*, Springer, Berlin, Germany, 2013.

[26] J. Serrà and J. L. Arcos, "Particle swarm optimization for time series motif discovery," *Knowledge-Based Systems*, vol. 92, pp. 127–137, 2016.

[27] C. Lei and J. Ruan, "A particle swarm optimization-based algorithm for finding gapped motifs," *BioData Mining*, vol. 3, no. 1, 2010.

[28] C. Bi, "Tackling the challenging motif problem through hybrid particle swarm optimized alignment clustering," in *Proceedings of the 2011 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology, CIBCB 2011*, pp. 84–91, Paris, France, April 2011.

[29] U. S. Reddy, M. Arock, and A. V. Reddy, "A particle swarm optimization solution for challenging planted(l, d)-Motif problem," in *Proceedings of the 2013 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, Singapore, 2013.

[30] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Research*, vol. 18, no. 20, pp. 6097–6100, 1990.

[31] H. Ge, L. Sun, Y. Yao, and J. Yu, "An automatic motif recognition algorithm in DNA sequences based on particle swarm optimization and random projection," in *Proceedings of the 2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017*, pp. 2241–2243, USA, November 2017.