IET Signal Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Iterative sequential Monte Carlo algorithm for motif discovery

ISSN 1751-9675
Received on 14th September 2014
Revised on 10th November 2015
Accepted on 29th January 2016
doi: 10.1049/iet-spr.2014.0356
www.ietdl.org

Mohammad Al Bataineh[1] ✉, *Zouhair Al-qudah*[2], *Awad Al-Zaben*[3]

[1]*Telecommunications Engineering Department, Yarmouk University, Irbid, Jordan*
[2]*Department of Electrical and Communication Engineering, Al-Hussein bin Talal University, Ma'an, Jordan*
[3]*Department of Biomedical Systems and Medical Informatics, Yarmouk University, Irbid, Jordan*
✉ *E-mail: mohamadfa@yu.edu.jo*

Abstract: The discovery of motifs in transcription factor binding sites is important in the transcription process, and is crucial for understanding the gene regulatory relationship and evolution history. Identifying weak motifs and reducing the effect of local optima, error propagation and computational complexity are still important, but challenging tasks for motif discovery. This study proposes an iterative sequential Monte Carlo (ISMC) motif discovery algorithm based on the position weight matrix and the Gibbs sampling model to locate conserved motifs in a given set of nucleotide sequences. Three sub-algorithms have been proposed. Algorithm 1 (see Fig. 1) deals with the case of one motif instance of fixed length in each nucleotide sequence. Furthermore, the proposed ISMC algorithm is extended to deal with more complex situations including unique motif of unknown length in Algorithm 2, unique motif with unknown abundance in Algorithm 3 (see Fig. 2) and multiple motifs. Experimental results over both synthetic and real datasets show that the proposed ISMC algorithm outperforms five other widely used motif discovery algorithms in terms of nucleotide and site-level sensitivity, nucleotide and site-level positive prediction value, nucleotide-level performance coefficient, nucleotide-level correlation coefficient and site-level average site performance.

1 Introduction

A motif is a certain conservative sequence pattern that exists in the transcription factor binding sites (TFBSs), which is part of the cis-regulatory regions for a set of co-expressed genes. Furthermore, these co-expressed genes are regulated by similar TFs. Currently, the wide range of motif discovery algorithms can be classified into two categories: consensus-based algorithms and statistical distribution ones.

In consensus-based algorithms, a consensus is formed for each of the possible sets of starting locations, and the best consensus is chosen to describe the motifs in the data [1, 2]. The motif discovery based on consensus [3] is to solve the planted (M, k) motif search (PMS) problem, where each instance of a motif of length M differs from the original at exactly k positions. So, these k positions serve as the constraint for the maximal Hamming distance between the consensus and the motif instances (binding sites). The consensus-based motif discovery algorithms try to enumerate all the strings satisfying the constraints. However, k is difficult to determine. If k is too small, most of the true motifs are missed due to the stringent criteria, while a large k will not only significantly increase the computation complexity, but also augment the false positives. Furthermore, the discrete consensus of the motif which is used in the consensus-based approaches is not accurate enough to represent the different strengths of conservation among different nucleotides. One possible solution to this problem is to use statistical distribution-based algorithms, which always display better performance. Until recently, the research to solve the PMS problem mainly concentrated on pattern-driven algorithms. All the M-length string patterns are taken as candidate motifs, and the string patterns occurring in all input sequences with up to k mutations are the motifs. Typical pattern-driven algorithms use various means to reduce time complexity [4–10]. PairMotif [4] selects multiple pairs of M-mers with relatively large distance from the input sequences to restrict the search space. Compared with recently proposed algorithms, PairMotif requires less storage space and runs faster on most PMS problems. PMS5 [7] computes the common k-neighbours of three M-mers using integer programming formulation. Some other pattern-driven algorithms index the input sequences with a suffix tree to speed up the search of candidate motifs [11–14]. RISOTTO [11] is the fastest algorithm in the family of suffix tree algorithms for the PMS problem.

Statistical distribution-based algorithms describe the statistical distribution of the four possible nucleotides based on a position weight matrix (PWM) at every position in the motif. For motifs of length M, the PWM is a matrix of size $(4 \times M)$, and each column of the PWM represents the distribution of the four nucleotides at the corresponding position in the motif. The PWM is estimated in the various statistical distribution-based algorithms and is used to predict the most likely location of the motif within each sequence. An expectation–maximisation (EM) algorithm [15] is proposed to estimate the PWM and locate the motifs simultaneously. Multiple EM for motif elicitation (MEME) [16], an algorithm based on EM, is capable of finding an unknown number of motifs with an unknown number of occurrences in the sequences. On the basis of MEME, there are some extension algorithms such as projection [3] and Monte Carlo EM motif discovery algorithm (MCEMDA) [17]. Projection projects all l-mers from the input sequences onto many buckets by hashing and then derives the consensus sequences to select some valid buckets. After the effective initialisation step, the EM algorithm is used for refinement. MCEMDA is a modification of the EM algorithm in that the expectation in the E-step is computed numerically through a Monte Carlo simulation. The Gibbs motif sampler and align acid conserved elements (ACE) algorithms [18] are two examples which estimate the PWM and the locations of the motifs in the TFBS genomic sequences based on the Markov chain Monte Carlo (MCMC) model. The Gibbs sampler-based BioProspector technique [19] considers the palindromic patterns in the TFBS sequence. Furthermore, some graph-theoretic methods either based on clustering or heuristic search have also been introduced in the field of motif discovery [20–22]. The cluster refinement algorithm for motif discovery (CRMD) [20] uses entropy-based clustering to find good starting candidate motifs from the input sequences and then employs an

effective greedy refinement to search for optimal motifs from these candidate motifs. VINE [22] is a graph clustering algorithm for motif discovery by finding t-cliques in a t-graph in polynomial time. Generally, the approximate algorithm has speedy runtime and minimal memory consumption. Sometimes, however, they cannot converge to the global optimum [23].

However, statistical distribution-based methods such as Gibbs sampling take samples from TFBS sequences in a probabilistic manner. The procedure to obtain the stable solutions is always very time-consuming. Statistical distribution-based methods are also sensitive to initial parameters, and the local search operations used by them often result in local optima problems. In motif discovery, if the motifs are weakly conserved categories, which is not a rare case, the local optima problems get worst.

As follow-up work of [24], a statistical profile-based approach named deterministic sequential Monte Carlo (DSMC) [25] was proposed. It is a deterministic tree-search method based on the SMC algorithm to discover motifs of an unknown length. These SMC methodologies are a family of statistical inference methods. It has been shown in [26] that the SMC methods provide better performance than the traditional MCMC algorithms, and are more efficient than Gibbs sampling with Dirichlet mixture models. However, the DSMC algorithm still gets involved in the local optima problem. It is partly due to the way of evaluating the conditional posterior distribution of the PWM. The hidden Markov model (HMM) underlying the DSMC can also cause this problem. With HMM, the DSMC algorithm is subject to error propagation, which means that the less conserved motifs discovered at the beginning of the discovery procedure will increase the probability of error location later because of the noise introduced by the less conserved motif candidates [20].

In this paper, we propose an iterative SMC (ISMC) motif discovery algorithm to estimate the PWM and the locations of the motifs. In the ISMC algorithm, a new conditional posterior distribution of the PWM is defined by using a smaller update step size, which can effectively improve the estimation accuracy and thus reduce the problem of local optima in the traditional SMC algorithms. The proposed ISMC algorithm also iteratively estimates the optimum order of sequence for motif discovery based on EM criterion to mitigate error propagation inherent to the HMM-based SMC approach with significantly less computational complexity, so as to obtain global optimum estimation in motif discovery compared with conventional approaches that use statistical expectation. In the traditional SMC motif discovery algorithms, only the PWM is updated in each iteration. However, in our proposed paper, the order of nucleotide sequence used in each iteration is updated as they affect the value of the conditional posterior probability as the PWM as well. In addition, the ISMC algorithm can always find out the optimum motif positions for all of the sequences in this dataset regardless of their initial order. This further verifies the robustness of the ISMC algorithm. Other algorithms are all influenced by initial order of genomic sequences; it indicates that their performances are more likely to be compromised by local optima than the proposed ISMC algorithm. Simulation results show that the accuracy of the proposed ISMC motif discovery algorithm is proven to be superior to five other widely used motif discovery algorithms for datasets without information of motif length and abundance. Moreover, the proposed algorithms offer significant performance improvement over other algorithms in the case of multiple motifs discovery and most importantly provide less computational complexity.

The rest of this paper is organised as follows. In Section 2, the system model is presented for the motif discovery problem. In Section 3, we derive the ISMC motif discovery algorithm for the basic case that a unique motif occurs in each sequence exactly one time. Furthermore, in Section 4, we extend the ISMC algorithm to deal with more complex situations including unique motifs of an unknown length, unique motifs with an unknown abundance and multiple motifs. In Section 5, experimental results are provided on both synthetic and real datasets. Finally, Section 6 concludes this paper.

2 System model

In motif discovery, we are required to discover the common sequence patterns, which are called motif instances, in a given set of nucleotide sequences. At the beginning, for simplicity, it can be assumed that each sequence contains one motif of length M, and each sequence has a fixed length L. In the later part of this paper, this simplified model will be extended to deal with more complicated cases including unique motifs of an unknown length, unique motifs with an unknown abundance and multiple motifs.

Let $\boldsymbol{S}_N = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_t, \ldots, \boldsymbol{s}_N\}$, with $\boldsymbol{s}_t = [s_{t1}, \ldots, s_{tL}]$, be the set of nucleotide sequences of length L, where we wish to find a motif. Let us assume that a motif of length M is present in each one of the sequences at unknown positions. The distribution of the motif can be described by two types of $(4 \times M)$ matrices: a position count matrix (PCM) and a PWM.

A PCM gives the statistics of the four different nucleotide bases appearing at each position of the motif instances. It can be denoted as $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_i, \ldots, \boldsymbol{\phi}_M]$, where $\boldsymbol{\phi}_i = [\phi_{i1}, \ldots, \phi_{i4}]^{\mathrm{T}}$, $i = 1, \ldots, M$, ($[*]^{\mathrm{T}}$ denotes transposition) and ϕ_{ij}, $j = 1, \ldots, 4$ is the number of the nucleotide bases {A,C,G,T} appearing at the ith position of the motif instances. The remaining fragments, which are obtained by removing the motif instances from all sequences, are called background regions. Since background nucleotide is not dependent on position, the distribution of background regions can be described by the count vector (CV) of the background, which is obtained by counting the number of each nucleotide base in the background regions of the sequence set and denoted as $\boldsymbol{\phi}_0 = [\phi_{01}, \ldots, \phi_{04}]^{\mathrm{T}}$. Throughout this paper, we use the mapping {A = 1, C = 2, G = 3, T = 4}, so $\{\phi_{i1}, \ldots, \phi_{i4}\}$ denotes $\{\phi_{i\mathrm{A}}, \ldots, \phi_{i\mathrm{T}}\}$ at the ith position of the motif instances.

A PWM gives the frequencies of the four different nucleotide bases appearing on every position of all motif instances. It can be presented as $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_i, \ldots, \boldsymbol{\theta}_M]$, where the vector $\boldsymbol{\theta}_i = [\theta_{i1}, \ldots, \theta_{i4}]^{\mathrm{T}}$, $i = 1, \ldots, M$ is the probability distribution of the nucleotides {A,C,G,T} at the ith position of the motif. The PWM ($\boldsymbol{\Theta}$) can be derived from the PCM ($\boldsymbol{\Phi}$) by $\boldsymbol{\Theta} = \boldsymbol{\Phi}/N$, where N is the number of sequences. In the background regions, each nucleotide is assumed to be independent and identically distributed in terms of the background distribution $\boldsymbol{\theta}_0 = [\theta_{01}, \ldots, \theta_{04}]^{\mathrm{T}}$. Generally, in motif discovery, the background distribution $\boldsymbol{\theta}_0$ is prior information; it also can be obtained approximately from the CV of the given dataset by $\boldsymbol{\theta}_0 = \boldsymbol{\phi}_0/\mathrm{NL}$ where NL is the number of nucleotide bases in the whole sequence set. An example of motif discovery is given in Table 1.

Let $\boldsymbol{a}_{y_t, x_t}$ be a sequence fragment of length M in $\boldsymbol{s}_{y_t}$ with the first nucleotide located at position x_t, and $\boldsymbol{a}^c_{y_t, x_t}$ be the remaining fragment in $\boldsymbol{s}_{y_t}$ with $\boldsymbol{a}_{y_t, x_t}$ removed. For example, for the first sequence in Table 1, $\boldsymbol{s}_1 = [\mathrm{ctgaaCGAGTTctcga}]$ and $x_1 = 6$ with $M = 6$, $\boldsymbol{a}_{1,6} = [\mathrm{CGAGTT}]$ and $\boldsymbol{a}^c_{1,6} = [\mathrm{ctgaactcga}]$. We then define a vector $\boldsymbol{n}(\boldsymbol{a}) = [n_1, n_2, n_3, n_4]$ where n_j, $j = 1, \ldots, 4$, denotes the number of each of the four types of nucleotides found in the sequence fragment $\boldsymbol{a}$. Given two vectors $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3, \theta_4]^{\mathrm{T}}$ and $\boldsymbol{n} = [n_1, n_2, n_3, n_4]^{\mathrm{T}}$, we define

$$\boldsymbol{\theta}^{\boldsymbol{n}} \triangleq \prod_{j=1}^{4} \theta_j^{n_j}. \tag{1}$$

Since the nucleotides located in the motif are independent of the other motif nucleotides and non-motif nucleotides, given the PWM ($\boldsymbol{\Theta}$), the background distribution $\boldsymbol{\theta}_0$, and the state at time t, the distribution of the observed sequence $\boldsymbol{s}_t$ with a motif starting at x_t is then given by

$$p(\boldsymbol{s}_t | x_t, \boldsymbol{\Theta}) = \boldsymbol{\theta}_0^{\boldsymbol{n}\left(\boldsymbol{a}^c_{y_t, x_t}\right)} \prod_{m=1}^{M} \boldsymbol{\theta}_m^{\boldsymbol{n}\left(\boldsymbol{a}_{y_t, x_t}(m)\right)}, \tag{2}$$

where the term $\boldsymbol{n}\left(\boldsymbol{a}^c_{y_t, x_t}\right)$ denotes the number of each of the four types of nucleotides found in the background regions $\boldsymbol{a}^c_{y_t, x_t}$, and

Table 1 Example of motif discovery, where (a) lists the sequence set $\boldsymbol{S}_N$ and the PIM, (b) the motif instances and the background regions, (c) the PCMs for motif and background regions and (d) the PWMs for motif and background regions. In the sequences, the segments in lower case are the background regions and the segments in upper case are the motif instances

(a) Sequences and PIM	(b) Motif instances and background regions	(c) PCM ($\boldsymbol{\Phi}$)	(d) PWM ($\boldsymbol{\Theta}$)
$\boldsymbol{S}_1$: ctgaaCGAGTTctcga	$\boldsymbol{a}_{1,6}$: CGAGTT	A(1): 002311	A(1): 0.0 0.0 0.4 0.6 0.2 0.2
$\boldsymbol{S}_2$: cGGTAATtaccgtcgg	$\boldsymbol{a}_{2,2}$: GGTAAT	C(2): 410000	C(2): 0.8 0.2 0.0 0.0 0.0 0.0
$\boldsymbol{S}_3$: agcCCAGGAcgttcac	$\boldsymbol{a}_{3,4}$: CCAGGA	G(3): 140210	G(3): 0.2 0.8 0.0 0.4 0.6 0.6
$\boldsymbol{S}_4$: acCGTAGTagcatagt	$\boldsymbol{a}_{4,3}$: CGTAGT	T(4): 003034	T(4): 0.0 0.0 0.6 0.0 0.2 0.2
$\boldsymbol{S}_5$: tcccgatCGTAGTact	$\boldsymbol{a}_{5,8}$: CGTAGT	⇓	⇓
0000100000000000	$\boldsymbol{a}^c_{1,6}$: ctgaactcga	$\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_6]$	$\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_6]$
0100000000000000	$\boldsymbol{a}^c_{2,2}$: ctaccgtcgg	$\boldsymbol{\phi}_1 = [0, 4, 1, 0]^{\mathrm{T}}$	$\boldsymbol{\theta}_1 = [0, 0.8, 0.2, 0]^{\mathrm{T}}$
0001000000000000	$\boldsymbol{a}^c_{3,4}$: agccgttcac	$\boldsymbol{\phi}_2 = [0, 1, 4, 0]^{\mathrm{T}}$	$\boldsymbol{\theta}_2 = [0, 0.2, 0.8, 0]^{\mathrm{T}}$
0010000000000000	$\boldsymbol{a}^c_{4,3}$: acagcatagt	$\boldsymbol{\phi}_3 = [2, 0, 0, 3]^{\mathrm{T}}$	$\boldsymbol{\theta}_3 = [0.4, 0, 0, 0.6]^{\mathrm{T}}$
0000000100000000	$\boldsymbol{a}^c_{5,8}$: tcccgatact	$\boldsymbol{\phi}_4 = [3, 0, 2, 0]^{\mathrm{T}}$	$\boldsymbol{\theta}_4 = [0.6, 0, 0.4, 0]^{\mathrm{T}}$
—	—	$\boldsymbol{\phi}_5 = [1, 0, 1, 3]^{\mathrm{T}}$	$\boldsymbol{\theta}_5 = [0.2, 0, 0.6, 0.2]^{\mathrm{T}}$
—	—	$\boldsymbol{\phi}_6 = [1, 0, 0, 4]^{\mathrm{T}}$	$\boldsymbol{\theta}_6 = [0.2, 0, 0, 0.8]^{\mathrm{T}}$
—	—	$\boldsymbol{\phi}_0 = [19, 22, 20, 19]^{\mathrm{T}}$	$\boldsymbol{\theta}_0 = [0.2375, 0.275, 0.25, 0.2375]^{\mathrm{T}}$

hence is equal to $\boldsymbol{\phi}_0 = [\phi_{01}, \ldots, \phi_{04}]^{\mathrm{T}}$. On the other hand, the term $\boldsymbol{a}_{y_t,x_t}(m)$ is the mth element of the sequence fragment $\boldsymbol{a}_{y_t,x_t}$ (i.e. $\boldsymbol{a}_{y_t,x_t}(m)$ is a single base rather than a vector). Therefore, $\boldsymbol{n}(\boldsymbol{a}_{y_t,x_t}(m))$ is a (1×4) vector of zeros, except at the position corresponding to the nucleotide $\boldsymbol{a}_{y_t,x_t}(m)$, which is one. For example, according to Table 1, $\boldsymbol{a}_{1,6} = [\mathrm{CGAGTT}]$, $\boldsymbol{a}_{1,6}(2)$ is equal to $[G]$ and hence $\boldsymbol{n}(\boldsymbol{a}_{1,6}(2))$ is equal to [0 0 1 0] assuming the order of bases is A, C, G and then T.

By applying (1) into (2), the distribution of the observed sequence $\boldsymbol{s}_{y_t}$ becomes

$$p\left(\boldsymbol{s}_{y_t}|x_t, \boldsymbol{\Theta}\right) = \underbrace{\prod_{k=1}^{4} \theta_{0k}^{\phi_{0k}}}_{\substack{\text{background}\\ \text{distribution}}} \underbrace{\left[\prod_{m=1}^{M}\left(\prod_{j=1}^{4} \theta_{mj}^{n\left(\boldsymbol{a}_{y_t,x_t}(m)\right)_j}\right)\right]}_{\substack{\text{motif}\\ \text{distribution}}}, \quad (3)$$

where $\boldsymbol{n}(\boldsymbol{a}_{y_t,x_t}(m))_j$ is the jth element in the vector $\boldsymbol{n}(\boldsymbol{a}_{y_t,x_t}(m))$. For example, according to Table 1, since $\boldsymbol{n}(\boldsymbol{a}_{1,6}(2)) = [0 \;\; 0 \;\; 1 \;\; 0]$, then $\boldsymbol{n}(\boldsymbol{a}_{1,6}(2)_3$ is the third element in the vector [0 0 1 0] which is 1. Hence, the distribution of the observed sequence $\boldsymbol{s}_1$ in Table 1 with the motif starting at position $x_t = 6$ is given by $p(\boldsymbol{s}_1|6, \boldsymbol{\Theta})$ which is equal to $(\theta_{01})^{\phi_{01}}(\theta_{02})^{\phi_{02}}(\theta_{03})^{\phi_{03}}(\theta_{04})^{\phi_{04}}(\theta_{12})(\theta_{23})(\theta_{31})(\theta_{43})(\theta_{54})(\theta_{64}) = (0.2375)^{19}(0.2750)^{22}(0.2500)^{20}(0.2375)^{19}(0.8)(0.8)(0.4)(0.4)(0.2)(0.8) = 1.2994 \times 10^{-50}$.

The process to estimate the starting locations of the motifs in a set of sequences can be modelled as finding the position indicator matrix (PIM) for the motif instances in the given set of sequences. All elements in PIM are zero, except the ones corresponding to the start locations of motif instances. The **PIM** for the example in Table 1 is

$$\mathbf{PIM} = \begin{bmatrix} 0000010000000000 \\ 0100000000000000 \\ 0001000000000000 \\ 0010000000000000 \\ 0000000100000000 \end{bmatrix}.$$

The **PIM** can also be expressed in the form of position indicator set $I_{\mathrm{S}} = \{(y_t, x_t), t = 1, 2, \ldots, N\}$, where each (y_t, x_t) specifies the location of a '1' in the **PIM**. For example, the '1' in the first row of a **PIM** can be denoted as (1, 6). Let $\boldsymbol{Y}_t = [y_1, y_2, \ldots, y_t]$ and $\boldsymbol{X}_t = [x_1, x_2, \ldots, x_t]$, $t = 1, \ldots, N$, and the motif discovery problem can be presented as finding out $\boldsymbol{X}_t$ based on observations $\boldsymbol{Y}_t$. Obviously, this process can be described by the HMM. With the HMM model, the hidden state is the unknown starting location of the motif instance in a given sequence, and the state space is the set of possible starting locations for the motif instance within the sequence set.

The sequence $\boldsymbol{Y}_t = [y_1, y_2, \ldots, y_t]$ is a permutation of the sequence $[1, 2, \ldots, N]$. At time interval t, the observation is the sequence specified by y_t in the dataset, and we observe all of the nucleotide subsequences with length M in the given sequence $\mathbf{s}_{y_t}$. The hidden state during this interval t is the location index of the first nucleotide of the motif instance in the sequence. For a sequence with a motif of length M, the last $(M-1)$ nucleotides are not valid starting locations. Therefore, the state at step t, $t = 1, \ldots, N$, denoted as x_t, takes its value from the set $\chi = \{1, 2, \ldots, L_M\}$, where $L_M = L_t - (M-1)$ and L_t is the length of $\boldsymbol{s}_{y_t}$. If each sequence has the same length L, then $L_M = L - (M-1)$. Given the starting location (state), the emission probability of sequence $\mathbf{s}_{y_t}$ can be evaluated with the unknown PWM and the previous $(t-1)$ states.

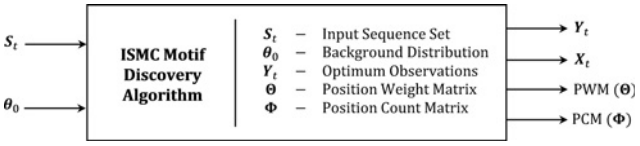

Fig. 1 *Proposed ISMC motif discovery algorithm*

Problem formulation: From the discussion above, the motif discovery problem can be formulated as follows. The state realisations up to time t can be denoted as $\boldsymbol{X}_t \triangleq [x_1, x_2, \ldots, x_t]$, and similarly the observations up to time t as $\boldsymbol{Y}_t \triangleq [y_1, y_2, \ldots, y_t]$. Given the sequence set $\boldsymbol{S}_t$ and the background distribution $\boldsymbol{\theta}_0$, we want to obtain the optimum observations $\boldsymbol{Y}_t$ and the optimum state estimations $\boldsymbol{X}_t$, which are the starting locations of the motif instance in the sequence set $\boldsymbol{S}_t = [\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_t]$, as well as the PWM ($\boldsymbol{\Theta}$) and PCM ($\boldsymbol{\Phi}$), which describe the position-based statistics of the motif. Prior knowledge such as the abundance of motif instances in the dataset, the background frequencies of the nucleotide types, and the probabilities of nucleotides in the motif instances can be easily incorporated in the model.

In the next section, we propose a novel ISMC motif discovery algorithm to estimate the locations of the motifs and the PWM. In the ISMC algorithm, both $\boldsymbol{Y}_t$ and $\boldsymbol{X}_t$ are estimated iteratively to achieve an optimum position indicator set I_{S} for the whole dataset. Furthermore, a new conditional posterior distribution of the PWM is used in evaluating the probability of a candidate subsequence being a motif instance. It uses a smaller update step size at the beginning, thus can effectively mitigate the local optima problem in the conventional SMC algorithms and obtain a more accurate estimation.

3 ISMC motif discovery algorithm

Let $p(\boldsymbol{X}_t|\boldsymbol{X}_{t-1})$ represent the state transition model, and $p(\boldsymbol{Y}_t|\boldsymbol{X}_t)$ represent the measurement model, where $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ represent the state and observation at time t, respectively, and $p(\cdot)$ represents the probability density functions (PDFs). At time t, we want to

IET Signal Process., 2016, Vol. 10, Iss. 5, pp. 504–513

estimate the state realisation $\boldsymbol{X}_t = [x_1, x_2, \ldots, x_t]$ based on the observation $\boldsymbol{Y}_t = [y_1, y_2, \ldots, y_t]$. The optimal solution in terms of any common criterion depends only on the conditional PDF, $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$. Often, direct computation of this conditional PDF is infeasible due to the complexity of the system; therefore, Monte Carlo methods are employed to estimate it. In most cases, however, drawing random samples directly from the conditional PDF $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ is also infeasible. Hence, we employ the importance sampling technique to sample from some trial sampling density $q(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ and properly weigh the samples according to the target distribution.

In this section, the ISMC motif discovery algorithm is derived for the case, where each sequence in the dataset contains exactly one instance of the same motif. In Section 4, the ISMC algorithm will be extended to more complicated situations.

3.1 Iterative SMC

To discover motifs, the conventional SMC algorithms such as DSMC perform the inference on those nucleotide sequences in a sequential order that is specified by a randomly generated $\boldsymbol{Y}_t$. Though $\boldsymbol{X}_t$ can be estimated based on the criterion of maximising the posterior probability $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$, it is good just for given $\boldsymbol{Y}_t$, which cannot guarantee that the estimation of $\boldsymbol{X}_t$ is optimum for the whole dataset. If the initial sequences specified by $\boldsymbol{Y}_t$ have poorly conserved motifs or contain no motifs at all, then the estimation and update of the PWM may be inaccurate, and the inferences made for the following sequences may be poor, this is called error propagation. To overcome this problem, conventional SMC algorithms enumerate all possible sequence orders, i.e. $\boldsymbol{Y}_t$, and decide the optimum $\boldsymbol{X}_t$ based on the results over all possible $\boldsymbol{Y}_t$. This will introduce large computational complexities. Thus, it is possible to find the $\boldsymbol{Y}_t$ associated with optimum $\boldsymbol{X}_t$ based on the EM criterion with significantly less computational complexity. For the EM algorithm, given the nucleotide sequences $\boldsymbol{S}_t = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_t\}$, each sequence consists of two components that model the motif and the non-motif (i.e. the background) positions in the sequence. The starting positions of the motif in each sequence are unknown and represented by $\boldsymbol{X}_t$. The EM algorithm attempts to find the maximum likelihood or maximum a posteriori estimates of $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ in the statistical model, where the model depends on unobserved PWM ($\boldsymbol{\Theta}$). The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximisation (M) step, which computes parameters maximising the expected log-likelihood found on the (E) step. These parameter-estimates are then used to determine the distribution of ($\boldsymbol{\Theta}$) in the next (E) step.

The motif discovery is to find the unknown distribution of PWM ($\boldsymbol{\Theta}$) by estimation

$$\widehat{\boldsymbol{\Theta}} = \arg\max_{\boldsymbol{\Theta}} p(\boldsymbol{Y}_N, \boldsymbol{\Theta}), \tag{4}$$

where $p(\boldsymbol{Y}_N, \boldsymbol{\Theta}) = p(\boldsymbol{Y}_N, \boldsymbol{X}_N)$.

Following the theory of EM [27], (4) is equivalent to

$$\widehat{\boldsymbol{\Theta}}^{(i)} = \arg\max_{\boldsymbol{\Theta}} Q\left(\boldsymbol{\Theta}, \widehat{\boldsymbol{\Theta}}^{(i-1)}\right), \tag{5}$$

where

$$Q\left(\widehat{\boldsymbol{\Theta}}^{(i)}, \widehat{\boldsymbol{\Theta}}^{(i-1)}\right) = \mathrm{E}_{\boldsymbol{X}_N \left| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i-1)} \right.} \left\{ \ln\left[p\left(\boldsymbol{X}_N, \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right)\right]\right\},$$

$$\mathrm{E}_{\boldsymbol{X}_N \left| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i-1)} \right.}[*]$$

is defined as the expectation with respect to $p\left(\boldsymbol{X}_N \middle| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i-1)}\right)$, and i is the index of iteration.

According to [28], this iterative EM process can be approximated by the SMC model. Since $\ln(x)$ is a monotonically increasing function and

$$p\left(\boldsymbol{X}_N, \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right) = p\left(\boldsymbol{X}_N \middle| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right) p\left(\boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right) \propto p\left(\boldsymbol{X}_N \middle| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right),$$

we have

$$Q\left(\widehat{\boldsymbol{\Theta}}^{(i)}, \widehat{\boldsymbol{\Theta}}^{(i-1)}\right) \propto \mathrm{E}_{\boldsymbol{X}_N \left| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i-1)} \right.} \left[p\left(\boldsymbol{X}_N \middle| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right)\right], \tag{6}$$

In each iteration, the SMC algorithm estimates optimum $\boldsymbol{X}_N$ with prior distribution $\widehat{\boldsymbol{\Theta}}^{(i-1)}$ based on the rule of maximum posterior probability with respect to $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$, $t = 1, \ldots, N$. At time $t = N$, all possible extensions from $\boldsymbol{X}_{N-1}$ are considered and their associated weights are obtained. This can be regarded as the 'expectation (E) step' in the EM algorithm. Summing up the weights associated to different ($\boldsymbol{Y}_N$, $\boldsymbol{X}_N$), the one with the largest summation is selected to update prior distribution $\widehat{\boldsymbol{\Theta}}^{(i)}$ for the next iteration, this is called the 'maximisation (M) step'.

Traditionally, only $\boldsymbol{\Theta}$ will be updated in each iteration. However, since $\boldsymbol{Y}_N$ also affects the value $p\left(\boldsymbol{X}_N \middle| \boldsymbol{Y}_N, \widehat{\boldsymbol{\Theta}}^{(i)}\right)$, it should also be updated. Let $\boldsymbol{Y}_N^{(i-1)} = [\boldsymbol{y}_1^{(i-1)}, \ldots, \boldsymbol{y}_T^{(i-1)}]$ be the order of nucleotide sequences used in iteration $(i-1)$ and $\boldsymbol{M}_N^{(i-1)} = [\boldsymbol{\mu}_1^{(i-1)}, \ldots, \boldsymbol{\mu}_N^{(i-1)}]$ be the conservative metrics for $\boldsymbol{Y}_N^{(i-1)}$, $\mu_t^{(i)} = \max\left\{ p\left(\boldsymbol{X}_t \middle| \boldsymbol{Y}_t, \widehat{\boldsymbol{\Theta}}^{(i)}.\right)\right\}$, $t = 1, \ldots, N$

$$\boldsymbol{Y}_N^{(i)} = \mathrm{SORT}_{\boldsymbol{M}_N^{(i)}}\left(\boldsymbol{Y}_N^{(i-1)}\right), \tag{7}$$

where the function $\mathrm{SORT}_x(y)$ denotes generating a new arrangement of the elements in a given sequence y according to the descending order of corresponding values in another sequence x. Therefore, in each iteration, the proposed ISMC algorithm changes the order of the input nucleotide sequences adaptively based on the descending order of the posterior distribution $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$, $t = 1, \ldots, N$ from the previous iteration. In the following section, the new SMC sampling approach will be introduced.

3.2 SMC sampling

The DSMC algorithm for the motif discovery problem was discussed in [25, 26]. In the conventional SMC algorithms, a statistical expectation function with respect to different $\boldsymbol{\Theta}$ is used in the state transition probability distribution. The expectation function contributes to the stability of the algorithms, but it also introduces an estimation error. In this section, based on a different state transition and parameter update model, we will derive a new SMC sampling method from the theory of sequential importance sampling for the Monte Carlo process. In this new SMC method, the state transition probability is evaluated based on a new conditional posterior distribution of motif instance, which can provide a more accurate estimation rather than a statistical expectation.

At time t, the SMC motif discovery algorithm makes an inference of the starting locations of the motifs $\boldsymbol{X}_t = [x_1, \ldots, x_t]$ based on the observations $\boldsymbol{Y}_t = [y_1, \ldots, y_t]$. It can be assumed that we have several sets of a sequence sample and its associated weight $\left\{\left(\boldsymbol{X}_{t-1}^{(k)}, \boldsymbol{w}_{t-1}^{(k)}\right), k = 1, \ldots, P\right\}$ at time $(t-1)$. Each set corresponds to a particle, thus P is the number of particles. Since the first nucleotide of the motifs can locate at any position of the observed sequence, we shall consider all values in the finite set of possible locations χ as possible extensions for each particle at time t.

Thus at time t, for each sequence sample $\boldsymbol{X}_{t-1}^{(k)}$, $k = 1, \ldots, P$, all possible $\boldsymbol{L}_M$ extensions are considered and their associated weights

are evaluated. On the basis of the $K \times L_M$ weights achieved, we choose the P extensions with the largest weights to update the particles and obtain $\left\{\left(\widehat{\boldsymbol{X}}_t^{(k)}, \widehat{\boldsymbol{w}}_t^{(k)}\right), k = 1, \ldots, P\right\}$, while the optimum estimation of $\boldsymbol{X}_t$ can be obtained based on the criterion of maximum posterior probability with respect to $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$. So the SMC algorithm is deterministic and optimum for the given dataset and $\boldsymbol{Y}_t$. With those $K \times L_M$ sets of extended samples for sequence $\boldsymbol{X}_t$ and associated weight $\left\{\left(\boldsymbol{X}_t^{(k)}, \boldsymbol{w}_t^{(k)}\right), k = 1, \ldots, P \cdot L_M\right\}$, the posterior distribution of $\boldsymbol{X}_t$ can be expressed by

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_t) = \frac{1}{W_t}\sum_{k=1}^{P\cdot L_M} \boldsymbol{w}_t^{(k)} \cdot \delta(\boldsymbol{X}_t - \boldsymbol{X}_t^{(k)}), \tag{8}$$

where

$$W_t = \sum_{k=1}^{P\cdot L_M} \boldsymbol{w}_t^{(k)}, \quad \text{and} \quad \delta(x) = \begin{cases} 1, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases},$$

$$\boldsymbol{w}_t^{(k)} = \frac{p(\boldsymbol{X}_t|\boldsymbol{Y}_t)}{q(\boldsymbol{X}_t|\boldsymbol{Y}_t)}. \tag{9}$$

$q(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ is called the importance density of the samples in $\boldsymbol{X}_t^{(k)}$.

At time t, the samples constituting an approximation to $p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})$ are available, and one could approximate $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ with a new set of samples.

If the importance density is factorised as

$$q(\boldsymbol{X}_t|\boldsymbol{Y}_t) = q(x_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t, \boldsymbol{\Theta})q(\boldsymbol{X}_{t-1}, \boldsymbol{\Theta}|\boldsymbol{Y}_{t-1}), \tag{10}$$

then the samples $\boldsymbol{X}_t^{(k)} \sim q(\boldsymbol{X}_t|\boldsymbol{Y}_t)$ can be obtained by augmenting each of the existing samples $\boldsymbol{X}_{t-1}^{(k)} \sim q(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})$ with the new states $x_t^{(k)} \sim q(x_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t)$. The sign '$\sim$' means 'distributed according to'. With Bayes' theorem

$$\begin{aligned} p(\boldsymbol{X}_t|\boldsymbol{Y}_t) &= \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{Y}_{t-1}, \boldsymbol{\Theta})p(\boldsymbol{X}_t, \boldsymbol{\Theta}|\boldsymbol{Y}_{t-1})}{p(y_t|\boldsymbol{Y}_{t-1})}, \\ &= \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{Y}_{t-1}, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_{t-1})}{p(y_t|\boldsymbol{Y}_{t-1})} \\ &\quad \cdot p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1}), \\ &= \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}{p(y_t|\boldsymbol{Y}_{t-1})} \cdot p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1}), \\ &\propto p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1}). \end{aligned} \tag{11}$$

Then, by substituting (10) and (11) into (9), the weight can be updated with

$$\begin{aligned} \boldsymbol{w}_t^{(k)} &= \frac{p(\boldsymbol{X}_{t-1}|\boldsymbol{Y}_{t-1})p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}{q(x_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t, \boldsymbol{\Theta})q(\boldsymbol{X}_{t-1}, \boldsymbol{\Theta}|\boldsymbol{Y}_{t-1})}, \\ &= \boldsymbol{w}_{t-1}^{(k)} \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}{q(x_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t, \boldsymbol{\Theta})}. \end{aligned} \tag{12}$$

Let the importance density $q(x_t|\boldsymbol{X}_{t-1}, \boldsymbol{Y}_t, \boldsymbol{\Theta}) = q(x_t|\boldsymbol{X}_{t-1}, y_t, \boldsymbol{\Theta})$, then

$$\boldsymbol{w}_t^{(k)} = \boldsymbol{w}_{t-1}^{(k)} \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}{q(x_t|\boldsymbol{X}_{t-1}, y_t, \boldsymbol{\Theta})}. \tag{13}$$

The optimal importance density function $q(x_t|\boldsymbol{X}_{t-1}, y_t, \boldsymbol{\Theta})$ that minimises the variance of the weight $\boldsymbol{w}_t^{(k)}$ is proven to be [29, 30]

$$\begin{aligned} q(x_t|\boldsymbol{X}_{t-1}, y_t, \boldsymbol{\Theta}) &= p(x_t|\boldsymbol{X}_{t-1}, y_t, \boldsymbol{\Theta}), \\ &= \frac{p(y_t|\boldsymbol{X}_t, \boldsymbol{\Theta})p(x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}{p(y_t, x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1})}. \end{aligned} \tag{14}$$

With (13) and (14), it can easily be derived that

$$\boldsymbol{w}_t^{(k)} = \boldsymbol{w}_{t-1}^{(k)} \cdot p(y_t, x_t, \boldsymbol{\Theta}|\boldsymbol{X}_{t-1}). \tag{15}$$

From (8) and (15), the posterior distribution of $\boldsymbol{X}_t$ can be approximated as

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_t) = \frac{1}{W_t}\sum_{k=1}^{P} \boldsymbol{w}_t^{(k,x_t)}\delta(\boldsymbol{X}_t - \widehat{\boldsymbol{X}}_t^{(k,x_t)}), \tag{16}$$

where $\widehat{\boldsymbol{X}}_t^{(k,x_t)} = \left[\boldsymbol{X}_{t-1}^{(k)}, x_t\right]$ denotes the sample sequence obtained by appending extension $x_t \in \{1, \ldots, L_M\}$ to $\boldsymbol{X}_{t-1}^{(k)}$, where x_t is the possible position of motif instance, and $W_t = \sum_{k=1}^{P\cdot L_M} \boldsymbol{w}_t^{(k)}$ with

$$\begin{aligned} \boldsymbol{w}_t^{(k,x_t)} &= \sum_{k=1}^{P\cdot L_M} \boldsymbol{w}_t^{(k)}\delta\left(\boldsymbol{X}_t^{(k)} - \widehat{\boldsymbol{X}}_t^{(k,x_t)}\right), \\ &= \boldsymbol{w}_{t-1}^{(k)} \cdot p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right). \end{aligned} \tag{17}$$

Substituting (17) into (16), we have

$$\begin{aligned} p(\boldsymbol{X}_t|\boldsymbol{Y}_t) = \frac{1}{W_t}\sum_{k=1}^{P} &\boldsymbol{w}_{t-1}^{(k)} \\ &\cdot p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right)\delta\left(\boldsymbol{X}_{t-1} - \boldsymbol{X}_{t-1}^{(k)}\right), \end{aligned} \tag{18}$$

where $\boldsymbol{X}_{t-1} = [x_1, x_2, \ldots, x_{t-1}]$ denotes the subsequence of $\boldsymbol{X}_t$ without $\boldsymbol{x}_t$.

As a suboptimal approach, (18) can be simplified to

$$p(\boldsymbol{X}_t|\boldsymbol{Y}_t) = \frac{1}{W_t}\sum_{k=1}^{P} \boldsymbol{w}_{t-1}^{(k)} \cdot p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right). \tag{19}$$

As shown in (18) and (19), $p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right)$ plays an important role in evaluating the $p(\boldsymbol{X}_t|\boldsymbol{Y}_t)$, as it determines the performance of the SMC motif discovery algorithm. Evaluating $p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right)$ can be computationally complex. In the next section, the approach to evaluate $p\left(y_t, x_t, \boldsymbol{\Theta}\middle|\boldsymbol{X}_{t-1}^{(k)}\right)$ and the new model to update the prior distribution $\boldsymbol{\Theta}$ will be derived.

3.3 Distribution of the motif instance

In conventional approaches for motif discovery, the Dirichlet distribution [20] is used in evaluating the probability of the motif instance based on the prior distribution of PWM ($\boldsymbol{\Theta}$). Moreover, it is assumed that nucleotides in the motif instance are independent of each other. This model works well in most cases [31], and it can significantly simplify the motif discovery algorithm. However, the Dirichlet distribution is computationally complex, and it gives poor performance when accurate prior distribution of PWM ($\boldsymbol{\Theta}$) is not available [24, 25].

Let the prior distribution of the motif be given by the PCM ($\boldsymbol{\Phi}$) and the PWM ($\boldsymbol{\Theta}$). Assuming independent priors, then the motif instance follows the product Dirichlet distribution given by

$$\boldsymbol{\Theta} \sim p(\boldsymbol{\Theta}|\boldsymbol{\Phi}) = \prod_{m=1}^{M} \boldsymbol{\theta}_m^{\boldsymbol{\phi}_m}, \tag{20}$$

where $\boldsymbol{\theta}_m^{\boldsymbol{\phi}_m} = \prod_{j=1}^{4} \boldsymbol{\theta}_{mj}^{\boldsymbol{\phi}_{mj}}$, $m = 1, \ldots, M$.

Since the PCM ($\boldsymbol{\Phi}$) and PWM ($\boldsymbol{\Theta}$) columns are related by $\boldsymbol{\theta}_m = \boldsymbol{\phi}_m/N$, then the conditional posterior distribution of the motif

Algorithm 1

Input: The nucleotide sequence set $\boldsymbol{S}_N$.

Output: The locations of the best set of motif instances, $(\boldsymbol{Y}_N, \boldsymbol{X}_N)$.

Initialisation: • $\boldsymbol{X}_N^{(0)} = [1,2,\ldots,N]$;

• Use the first γ sequences to enumerate the $L_M^{(\gamma)}$ possible particles, evaluate the weights and initial prior distributions associated with the particles, where γ is the largest integer for $L_M^{(\gamma)} < P$.

For $i = 1,2,\ldots,l_{max}$, l_{max} is the maximum iteration time **do**

  For each $\boldsymbol{Y}_t \in \boldsymbol{Y}_N, t = 1,2,\ldots,N$ **do**

    For $k = 1,2,\ldots,P$ **do**

      • Enumerate all possible extensions. $\boldsymbol{X}_t^{(k,h_t)} = \left[\boldsymbol{X}_{t-1}^{(k)}, x_t\right], x_t = 1,\ldots,L_M$

      • $\forall \boldsymbol{X}_t^{(k,x_t)}$, compute the weight $\boldsymbol{w}_t^{(k,x_t)}$ according to (14) and (21).

    • Select and preserve P distinct sample streams $\left\{\hat{\boldsymbol{X}}_t^{(k)}, k = 1,\ldots,P\right\}$ with the highest importance weights $\left\{\hat{\boldsymbol{w}}_t^{(k)}, k = 1,\ldots,P\right\}$ from the set $\left\{\left(\boldsymbol{X}_t^{(k)}, \boldsymbol{w}_t^{(k)}\right), k = 1,\ldots,P \cdot L_M\right\}$.

    • $\forall k$, update the PCM $\boldsymbol{\phi}_t^{(k)} = U\left(\boldsymbol{\phi}_{t-1}^{(k)}, y_t, x_t^{(k)}\right)$ according to (22).

  • Obtain $\boldsymbol{Y}_N^{(i)}$ based on $\boldsymbol{Y}_N^{(i-1)}$ according to (7).

  If $\boldsymbol{X}_t{}^{(i)} = \boldsymbol{X}_t{}^{(i-1)}, t = 1,2,\ldots,N$, i is the iteration index, the iteration will be terminated.

  Else $i = i + 1$.

Fig. 2 *ISMC motif discovery algorithm for case of one fixed-length motif instance in each sequence*

instance can be formulated using (2) as

$$p\left(\boldsymbol{y}_t,\ x_t,\ \boldsymbol{\Theta}|\boldsymbol{X}_{t-1}^{(k)}\right) = \boldsymbol{\theta}_0^{n\left(a^c_{y_t,x_t}\right)} \prod_{m=1}^{M} \left(\left[\frac{\boldsymbol{\phi}_m}{N}\right]_t\right)^{n\left(a_{y_t,x_t}(m)\right)}, \qquad (21)$$

where the term $[\boldsymbol{\phi}_m/N]_t$ is equivalent to $\boldsymbol{\theta}_m$ (which is the mth column of the PWM $\boldsymbol{\Theta}$) estimated at time t. $\boldsymbol{\phi}_m$ is the mth column of the PCM ($\boldsymbol{\Phi}$) and N is the number of nucleotide sequences considered at time t. The PCM ($\boldsymbol{\Phi}$) is updated as $\boldsymbol{\Phi}_t = \boldsymbol{\Phi}_{t-1}^{(k)} + \boldsymbol{n}\left(\boldsymbol{a}_{y_t,x_t}\right)$, where $\boldsymbol{\Phi}_{t-1}^{(k)}$ is the PCM for $\left(\boldsymbol{X}_{t-1}^{(k)},\ \boldsymbol{w}_{t-1}^{(k)}\right)$ at time $(t-1)$. The term $\boldsymbol{n}\left(\boldsymbol{a}_{y_t,x_t}\right)$ is a $(4 \times M)$ binary matrix with the 1's corresponding to the presence of a certain nucleotide in the sequence fragment $\boldsymbol{a}_{y_t,x_t}$ at a certain position, and 0's corresponding to their absence. For example, for $\boldsymbol{a}_{1,6} = [\text{CGAGTT}]$ in Table 1, the corresponding $\boldsymbol{n}(\boldsymbol{a}_{1,6})$ matrix is

$$\begin{matrix} \text{A} \\ \text{C} \\ \text{G} \\ \text{T} \end{matrix} \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{bmatrix}.$$

The PCM update can be defined as a function with

$$\boldsymbol{\Phi}_t = U\left(\boldsymbol{\Phi}_{t-1}^{(k)},\ y_t,\ x_t\right). \qquad (22)$$

When $x_t = 0$, $\boldsymbol{\Phi}_t = U\left(\boldsymbol{\Phi}_{t-1}^{(k)},\ y_t,\ 0\right) = \boldsymbol{\Phi}_{t-1}^{(k)}$.

Different from the Dirichlet distribution used in evaluating the conventional posterior distribution for $\boldsymbol{\Theta}$, a constant factor $1/N$ instead of $1/\sum_{k=1}^{4} \phi_{ik}$ is used in (21). This factor defines an appropriate step size for updating the distribution of $\boldsymbol{\Theta}$. Particularly at the initial phase of motif discovery, when accurate prior distribution of PCM ($\boldsymbol{\Phi}$) and PWM ($\boldsymbol{\Theta}$) are not available, $1/N \ll 1/\sum_{k=1}^{4} \phi_{ik}$, and the step size or the increase of $[\boldsymbol{\phi}_m/N]_t$ at each update is much smaller than those conventional approaches. This reduces the fluctuation in the estimation of motif location and the probability of trapping in the local optima. In addition, it enables the proposed posterior distribution $p\left(y_t,\ x_t,\ \boldsymbol{\Theta}|\boldsymbol{X}_{t-1}^{(k)}.\right)$ to offer better estimation resolution than conventional approaches.

The efficiency of the ISMC algorithm will significantly decrease when most of the samples have very small weights [24]. Though it rarely happens, it is necessary to take measures to prevent this degeneracy. The effective sample size $\hat{P}_{\text{eff}} = 1/\sum_{k=1}^{P}\left(\boldsymbol{w}_t^{(k)}\right)^2$ can be used to detect the degeneracy of the sample weights. When the effective sample size is less than the threshold, e.g. $\hat{P}_{\text{eff}} \leq P/20$, re-sampling should be performed to re-initiate the weights of samples with $\boldsymbol{w}_t^{(k)} = P^{-1}$, $k = 1,\ \ldots,\ P$ [24]. The steps of the ISMC motif discovery algorithm for the case of one motif instance with fixed length in each sequence are described in Algorithm 1 (Fig. 1).

4 Extensions of ISMC for general motif discovery situations

In the previous section, the ISMC algorithm finds an optimal set of motif instances from the nucleotide sequences, where a unique motif occurs in each sequence exactly one time. In this section, we extend the ISMC algorithm to deal with more complicated situations including unique motifs of an unknown length, unique motifs with an unknown abundance and multiple motifs.

4.1 Motif length selection

In the previous section, the length of the motif is assumed to be a given parameter, and the ISMC algorithm simply locates a motif of a given length. However, in practical situation, the motif length is generally unknown when detecting an unknown motif in a dataset. ISMC algorithm can be easily adapted to estimate the unknown motif length.

First, the range for the motif length can be approximately estimated as $M \in [M_{\min},\ M_{\max}]$. Then, every motif length value in this range is applied to Algorithm 2 to discover the motifs in the target dataset. A score evaluated based on the estimated results can be obtained corresponding to each length value. The motif length that gets the highest score is chosen as the estimated motif length. A possible criterion is the conditional distribution $p(\boldsymbol{Y}_N|\boldsymbol{X}_N)$ [25]

$$p\left(\boldsymbol{Y}_N|\boldsymbol{X}_N\right) = \prod_{t=1}^{N} \boldsymbol{\theta}_0^{a^c_{y_t,x_t}} \prod_{m=1}^{M} \boldsymbol{\theta}_m^{a_{y_t,x_t}(m)}, \qquad (23)$$

where $\boldsymbol{\theta}_m$ is the mth column of the estimated PWM.

The corresponding logarithm score is a better choice to reduce the computation complexity. On the other hand, it is easy to understand that the longer the sequence, the smaller the probability that it happens. A motif of a longer length is generally having a smaller

score. By eliminating the influence of the length, a more reliable normalised score function is proposed as

$$\overline{S}_M = \frac{1}{M} S_M = \frac{1}{M} \sum_{t=1}^{N} \left[\boldsymbol{a}^c_{y_t, x_t} \log \boldsymbol{\theta}_0 + \sum_{m=1}^{M} \boldsymbol{a}_{y_t, x_t}(m) \log \boldsymbol{\theta}_m \right]. \quad (24)$$

The steps of the motif discovery algorithm with an unknown motif length are described in Algorithm 2. In the next section, the ISMC algorithm will be extended to motif discovery scenario without prior motif abundance information.

Algorithm 2: ISMC algorithm for the discovery of a motif with an unknown length

Input: The nucleotide sequence set $\boldsymbol{S}_N$.

Output: The locations of the best set of motif instances ($\boldsymbol{Y}_N$, $\boldsymbol{X}_N$), optimum motif length $\hat{M}$.

For each $M \in [M_{\min}, M_{\max}]$ **do**

- Obtain ($\boldsymbol{Y}_N$, $\boldsymbol{X}_N$) using **Algorithm 1** (see Fig. 1).
- Compute $\overline{\boldsymbol{S}}_M$ according to (24).

- Obtain the estimated motif length by $\hat{M} = \arg\max_D \{\overline{S}_M | M \in [M_{\min}, M_{\max}].\}$.
- Obtain ($\boldsymbol{Y}_N$, $\boldsymbol{X}_N$) with $\hat{M}$ using **Algorithm 1** (see Fig. 1).

4.2 Motif discovery without prior abundance information

To discover a motif in datasets with an unknown instance number of the same motif, an important issue is choosing an appropriate number of predicted motif instances. Predicting too many motif instances may lead to many false instances, while predicting too few motif instances may miss many true instances. To address the issue of the unknown number of motif instances, the ISMC algorithm must be capable of estimating the number of motif instances adaptively according to the criterion of maximum posterior probability.

At each iteration, the motif instances detected at the previous iteration are removed from the sequences where they are found, and the remaining fragments of the same sequences are concatenated together to form new sequences. The state space will also be updated according to the new sequences. By keeping the location index mapping tables between the original sequences and those new ones, we can determine start locations of those possible motif instances in the original sequences. Hence, eliminating those impossible ones, which are composed of widely separated segments in the original sequences.

In the ith iteration, the similarity between a candidate motif instance with a start location $x_t = j$ and the motif instances with state realisation $\boldsymbol{X}_N^{(i-1)}$, discovered in iteration $(i-1)$, is measured by the posterior probability $p\left(\boldsymbol{X}_N^{(i-1)}, \boldsymbol{Y}_N^{(i-1)} | x_t^{(i)}, y_t^{(i)}\right)$, which is

$$p\left(\boldsymbol{X}_N^{(i-1)}, \boldsymbol{Y}_N^{(i-1)} | x_t^{(i)}, y_t^{(i)}\right) = \left(\boldsymbol{\theta}_0^{(i-1)}\right)^{a^c_{y_t, x_t}} \prod_{m=1}^{M} \left(\boldsymbol{\theta}_m^{(i-1)}\right)^{a_{y_t, x_t}(m)}, \quad (25)$$

where $\boldsymbol{\theta}_0^{(i-1)}$ is the background distribution and $\boldsymbol{\theta}_m^{(i-1)}$ is the probability distribution of the nucleotides at the mth position of the PWM.

To decide whether the candidate x_t is a motif, we define a threshold Z_1. Since there is no prior knowledge of the instances to be included or excluded, the thresholds should take into account all the possible subsequences of a certain distribution. Therefore, the threshold that is used to discover the motif is the expected value of a random subsequence generated from a certain distribution with respect to the current statistics $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots, \boldsymbol{\phi}_i, \ldots, \boldsymbol{\phi}_M]$, where $\boldsymbol{\phi}_i = [\phi_{i1}, \ldots, \phi_{i4}]^{\mathrm{T}}$, i = 1, …, M

$$\begin{aligned} z_1(\boldsymbol{\Theta}) &= E\left\{ p\left(\boldsymbol{X}_N^{(i-1)}, \boldsymbol{Y}_N^{(i-1)} \middle| \boldsymbol{a}_p\right) | \boldsymbol{\Theta}^{(i)} \right\}, \\ &= \sum_{p=1}^{4^M} p\left(\boldsymbol{X}_N^{(i-1)}, \boldsymbol{Y}_N^{(i-1)} \middle| \boldsymbol{a}_p\right) p\left(\boldsymbol{a}_p | \boldsymbol{\Theta}^{(i)}\right), \\ &= \sum_{p=1}^{4^M} \left(\boldsymbol{\theta}_0^{(i-1)}\right)^{\bar{\boldsymbol{n}}_0^{(i-1)}} \prod_{m=1}^{M} \left(\boldsymbol{\theta}_m^{(i-1)}\right)^{n\left(a_p(m)\right)} \\ &\quad \times \prod_{m=1}^{M} \left(\boldsymbol{\theta}_m^{(i)}\right)^{n\left(a_p(m)\right)} \prod_{l=1}^{M} \left(\boldsymbol{\theta}_l^{(i)}\right)^{[\phi_l]_{t-1}}, \\ &\propto \left(\boldsymbol{\theta}_0^{(i-1)}\right)^{\bar{\boldsymbol{n}}_0^{(i-1)}} \prod_{l=1}^{M} \left(\boldsymbol{\theta}_l^{(i)}\right)^{[\phi_l]_{t-1}} \\ &\quad \cdot \sum_{p=1}^{4^M} \prod_{m=1}^{M} \left(\boldsymbol{\theta}_m^{(i-1)} \cdot \boldsymbol{\theta}_m^{(i)}\right)^{n\left(a_p(m)\right)}. \end{aligned} \quad (26)$$

The steps of the ISMC motif discovery algorithm for the case of an unknown abundance are described in Algorithm 3 (Fig. 2).

where the distribution of $\boldsymbol{\Theta}^{(i)}$ is obtained by (17), $\boldsymbol{a}_A$ is an artificial sequence of length M, $\boldsymbol{a}_p(m) \in \boldsymbol{B}$, where $\boldsymbol{B} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\}$, and $\boldsymbol{n}(\boldsymbol{a}_p(m))$ is a binary vector with a single '1' corresponding to the base $\boldsymbol{a}_p(m)$ and zeros elsewhere (e.g. $\boldsymbol{n}(A) = [1\ 0\ 0\ 0]$, $\boldsymbol{n}(C) = [0\ 1\ 0\ 0]$, $\boldsymbol{n}(G) = [0\ 0\ 1\ 0]$ and $\boldsymbol{n}(T) = [0\ 0\ 0\ 1]$). The term $\bar{\boldsymbol{n}}_0^{(i-1)}$ is the vector containing an average number of each of the four types of nucleotides found in the background sequences from iteration $(i-1)$.

Since (26) only involves the complete enumeration over the Cartesian product of the sets

$$\left\{ \left(\boldsymbol{\theta}_m^{(i-1)} \cdot \boldsymbol{\theta}_m^{(i)}\right)^{n(a_p(m))} \middle| \boldsymbol{a}_p(m) \in B \right\},$$

with

$$\overbrace{\sum_{b_1 \in B} \sum_{b_2 \in B} \cdots \sum_{b_M \in B}}^{M} \prod_{j=1}^{M} \theta_{j, b_j} = \prod_{j=1}^{M} \sum_{b_j \in B} \theta_{j, b_j}$$

[20], (26) can be further simplified to

$$\begin{aligned} z_1(\boldsymbol{\Theta}) &\propto \left(\boldsymbol{\theta}_0^{(i-1)}\right)^{\bar{\boldsymbol{n}}_0^{(i-1)}} \prod_{l=1}^{M} \left(\boldsymbol{\theta}_l^{(i)}\right)^{[\phi_l]_{t-1}} \\ &\quad \cdot \prod_{m=1}^{M} \sum_{y_p(m) \in B} \left(\boldsymbol{\theta}_m^{(i-1)} \cdot \boldsymbol{\theta}_m^{(i)}\right)^{n(a_p(m))}, \end{aligned} \quad (27)$$

and

$$Z_1 \propto \int z_1(\boldsymbol{\Theta}) p\left(\boldsymbol{\Theta} | \sigma_{t-1}\right) \mathrm{d}\boldsymbol{\Theta} = E_{\Theta | \Phi_{t-1}} \left\{ z_1(\boldsymbol{\Theta}) \right\}. \quad (28)$$

If $p(\boldsymbol{X}_t | \boldsymbol{Y}_t) > Z_1$, the sequence y_t contains a motif at x_t; otherwise, it can be decided that there is no motif in y_t, update the statistics $\boldsymbol{\Phi}_t^{(k)} = U\left(\boldsymbol{\Phi}_{t-1}^{(k)}, y_t, x_t = 0\right)$ according to (22). The ISMC algorithm for discovering motifs in the datasets without prior motif abundance information is described in Algorithm 3 (see Fig. 2). In the following section, the ISMC algorithm will be used to discover multiple motifs.

4.3 Discovery of multiple motifs

The ISMC motif discovery algorithm can be used to discover multiple motifs in a given dataset. Similar to the method [Algorithm 3 (see Fig. 2)] introduced in Section 4.2, the ISMC algorithm searches the sequences in the dataset iteratively to locate the instances for each motif. After several iterations, e.g. $l_{\max}$, the

IET Signal Process., 2016, Vol. 10, Iss. 5, pp. 504–513

Algorithm 3

Input: The nucleotide sequence set $\boldsymbol{S}_N$.

Output: The locations of the best set of motif instances, $I_s = \{(y_j, x_j), j = 1, \ldots, z\}$.

Initialisation: • Obtain the locations of the tentative set of motif instances $(\boldsymbol{Y}_N^{(0)}, \boldsymbol{X}_N^{(0)})$ and motif length M with **Algorithm 2**, based on the assumption of one motif instance per sequence;

• Obtain the prior distribution $\boldsymbol{\theta}_m^{(0)}$, $m = 1, \ldots, M$, using the output of **Algorithm 2**, $z = 0$.

For $i = 1,2, \ldots, l_{max}$, l_{max} is the maximum iteration time **do**

  For each $Y_t \in \boldsymbol{Y}_N, t = 1,2, \ldots, N$ **do**

    For $k = 1, 2, \ldots, P$ **do**

      • Enumerate all possible sample extensions. $\boldsymbol{X}_t^{(k,x_t)} = \left[\boldsymbol{X}_{t-1}^{(k)}, x_t\right], x_t = 1, \ldots, L_M$.

      • $\forall \boldsymbol{X}_t^{(k,x_t)}$, compute the weight $\boldsymbol{w}_t^{(k,x_t)}$ according to (14) and $z_1(\Theta^{(k)})$ according to (27).

    • compute Z_1 according to (28).

    If $p(X_t|Y_t) > Z_1$

      • Decide that there is a motif instance at x_t, put (y_t, x_t) into M_s, $z = z + 1$.

      • Select P distinct sample streams $\{\widehat{\boldsymbol{X}}_t^{(k)}, k = 1, \ldots, P\}$ with the highest importance $\{\widehat{w}_t^{(k)}, k = 1, \ldots, P\}$ from the set $\{(\boldsymbol{X}_t^{(k)}, \boldsymbol{w}_t^{(k)}), k = 1, \ldots, P \cdot L_M\}$.

      • $\forall k$, update the PCM $\boldsymbol{\Phi}_t^k = U(\boldsymbol{\Phi}_{t-1}^{(k)}, x_t^{(k)}, y_t)$ according to (22).

      • Remove the motif instance $\boldsymbol{a}_{y_t,x_t}$ and concatenate the remaining fragments to form a new sequence.

    Else Decide that there is no motif in sequence v_t, update the PCM $\boldsymbol{\Phi}_t^{(k)} = U(\boldsymbol{\Phi}_{t-1}^{(k)}, y_t, 0)$.

  • Obtain $\boldsymbol{Y}_N^{(i)}$ based on $\boldsymbol{Y}_N^{(i-1)}$ according to (7).

  If $\boldsymbol{X}_t{}^{(i)} = \boldsymbol{X}_t{}^{(i-1)}, t = 1,2, \ldots, N$, i is the iteration index, the iteration will be terminated.

  Else $i = i + 1$.

Fig. 3 *ISMC discovery algorithm for motifs with an unknown abundance*

motif instances discovered for the same motif are removed from all sequences, and the remnant fragments of each sequence are concatenated to form a new sequence. Thus, we get an updated dataset to discover new motif. Before running new iterations of the ISMC algorithm over the modified dataset, the weights and the prior distributions associated with the particles should be initiated according to the new sequences, and the state space should be updated in terms of the possible starting locations of the new motif. These parameters will be preserved and updated during the next $l_{\max}$ iterations, and then they will be re-initiated to discover another motif. This procedure will be performed until the ISMC algorithm cannot discover new motifs anymore. This re-initiation and detection approach enables us to search for motifs one by one in the descending order of conservative.

5 Experimental results

In this section, the performance of the proposed ISMC motif discovery algorithms is evaluated in different conditions over synthetic and real datasets. The results are compared with those of CRMD [20], MEME [32], DSMC [25], BioProspector [19] and AlignACE [18, 33]. In the performance comparisons, a modified performance coefficient defined in [34] is used. It can be described as follows

$$\psi(\boldsymbol{A}_{\mathrm{T}}, \boldsymbol{A}_{\mathrm{T}}^*) = \frac{\sum_{t=1}^{N} \left|\boldsymbol{a}_{y_t,x_t} \otimes \boldsymbol{a}_{y_t}^*\right|}{\sum_{t=1}^{N} \left|\boldsymbol{a}_{y_t,x_t} \oplus \boldsymbol{a}_{y_t}^*\right|}, \tag{29}$$

where $\boldsymbol{a}_{y_t,x_t}$ is the estimated motif instance in sequence $\boldsymbol{s}_{y_t}$ and $\boldsymbol{a}_{y_t}^*$ is the verified motif instance in sequence $\boldsymbol{s}_{y_t}$, thus $\boldsymbol{A}_N^* \triangleq \left\{\boldsymbol{a}_{y_1}^*, \ldots, \boldsymbol{a}_{y_N}^*\right\}$ is the verified motif instance set associated with the sequence set $\boldsymbol{S}_N$ and $\boldsymbol{A}_N \triangleq \left\{\boldsymbol{a}_{y_1,x_1}, \ldots, \boldsymbol{a}_{y_N,x_N}\right\}$ is the set of estimated motif instances in $\boldsymbol{S}_N$. The notation $|c_1 \otimes c_2|$ denotes the number of nucleotides in the overlapped part for both segments c_1 and c_2 and $|c_1 \oplus c_2|$ denotes the total number of nucleotides in both c_1 and c_2 with the overlapped nucleotides counted only once. In the following, the performance of ISMC motif discovery algorithm will be evaluated in different conditions.

5.1 Motif of known length occurring in each sequence exactly one time

The synthetic datasets for performance comparisons are generated as follows. In the background parts of the nucleotide sequences, each of the four bases (nucleotides) happens with a probability (frequency) of 25% at each position [25]. While in the motif instances, the dominant nucleotide at each position appeared with the probability of 70%, and each of the remaining three types of nucleotides appears at each position with a probability of 10%. About 20 motif datasets are generated for each motif lengths from 17 to 22. In each dataset, there are 50 sequences of 200 nucleotides (nt) long.

A motif instance of a specified length is inserted in each of the sequences. Since the motif length is known and there is one occasion of a motif instance per sequence, the ISMC algorithm discovers the motif instances according to Algorithm 1 (see Fig. 1). The performance of the ISMC algorithm, the CRMD, the DSMC algorithm [25], BioProspector, MEME and AlignACE over the synthetic datasets against different motif lengths is given in Figs. 3 and 4. The performance of each algorithm is evaluated

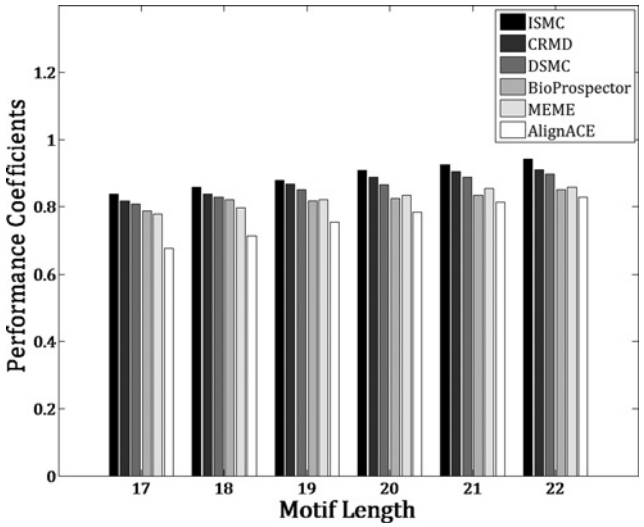

Fig. 4 *Accuracy comparison using synthesised dataset with motifs of different lengths*

Table 2 Performance comparisons of motif discovery algorithms using the CRP dataset

	ISMC	CRMD	DSMC	BioProspector	MEME	AlignACE
performance coefficient	0.89	0.86	0.84	0.79	0.55	0.66
predicted length	22	22	23	22	20	24
normalised score	−22.4	−22.9	−23.3	−24.5	−28.1	−26.4

with the performance coefficient defined in (29). It can be observed from Fig. 4 that the ISMC algorithm outperforms all other algorithms at all motif lengths. In the regions of short motif lengths, due to the low conservation and short sequence lengths, AlignACE algorithm is more likely to include error insertions due to its sequence alignment strategy, thus gets poor performance. While the ISMC algorithm performs well at all motif lengths, its performance improvement over other algorithms increases along with the motif length.

5.2 Unique motif of unknown length

In this section, the ISMC algorithm is applied to the cyclic-AMP receptor protein (CRP) binding sites dataset according to Algorithm 2. The CRP binding sites [35] dataset contains 18 sequences of 105 nt from *Escherichia coli*. There are 23 instances of the same motif that were determined by biological experiments in this dataset [36]. Each motif instance is of 22 nt long. The performance of the ISMC algorithm, the CRMD algorithm, the DSMC algorithm, BioProspector, MEME and AlignACE over CRP dataset will be compared under the assumption that there is only one motif instance in each sequence [37]. All of the algorithms search the dataset using motif lengths between 18 and 26. The results found by each algorithm using a different motif length are transformed into normalised motif logarithm score with (24) so that they can be analytically compared. The length that results in best performance will be used as the predicted length.

As shown in Table 2, only the ISMC, the CRMD and BioProspector obtain the accurate estimation of motif length, but the ISMC algorithm is the best of all these algorithms in terms of performance coefficient. The normalised logarithm score indicates how conserved the predicted CRP motif instances are. The ISMC algorithm got the highest normalised logarithm score, and offered the most conserved prediction results compared with other algorithms.

On the other hand, the ISMC algorithm can always find out the optimum motif positions for all of the sequences in this dataset regardless of their initial order. This proves the robustness of the ISMC algorithm. Other algorithms are all influenced by the initial order of the genomic sequences; it indicates that their performances are more likely to be compromised by local optima than the proposed ISMC algorithm.

5.3 Unique motif of unknown abundance

In this section, the performance of the ISMC algorithm is compared with those of the DSMC algorithm, the CRMD algorithm, MEME and AlignACE over the Tompa dataset, which contains 56 datasets of eukaryotes including yeast, fly, mouse and human. The background sequences for Tompa dataset are promoter sequences or the synthetic sequences generated with a third-ordered Markov chain model. There are also four negative control datasets, which do not contain any motifs. In the Tompa dataset, the motifs are poorly conserved. The performances of 13 motif discovery algorithms are assessed over Tompa dataset in [38].

Since the abundance of the motifs is not known, the ISMC algorithm discovers the motifs on the Tompa dataset according to Algorithm 3 (see Fig. 2). Its performance is compared with those of the DSMC algorithm, the CRMD algorithm, MEME and AlignACE in the same way as [38]. In [38], a different set of metrics is defined and used for performance evaluation. The metrics *n*Sn and *s*Sn represent the nucleotide and site-level sensitivity, *n*PPV and *s*PPV indicate the nucleotide and site-level positive prediction value, *n*PC is the nucleotide-level performance coefficient, while *n*CC and *s*ASP indicate the nucleotide-level correlation coefficient and site-level average site performance, respectively.

In Fig. 5, it can be observed that the ISMC algorithm has the best performances on all of the metrics. The DSMC algorithm has a good score on *n*PPV, but its performance is worst than MEME and AlignACE on all other metrics. As metrics of sensitivity, *n*Sn and *s*Sn are performance coefficients on the nucleotide and site-level without considering the false positives, the good performances of the ISMC on these two metrics indicate that the prediction accuracy of the ISMC algorithm on the position of the discovered motif instance is better than the other algorithms. The highest score on *n*PC shows that there are also less false positives predicted by the ISMC algorithm than by the CRMD, the DSMC, MEME and AlignACE. While the biggest values on nucleotide and site-level positive prediction value, *n*PPV and *s*PPV, means that the ratio of the true motif instances in the prediction results of the ISMC algorithm is more than the other algorithms.

5.4 Multiple motifs

In this section, the performance of the ISMC on the assessment dataset with multiple motifs is studied. The assessment dataset consists of genomic sequences from ERE, CRP, E2F, CREB, MEF2, MYOD, SRF and TBP dataset [20]. The ERE dataset consists of the genomic sequences with estrogen responsive elements. The CRP dataset was introduced in the previous Section 5.2, it includes 18 sequences of 105 nt, which contain CRP binding sites from *E. coli*. The sequences in E2F dataset are from mammalian species, they contain the binding sites for the TFs of E2F family. The datasets of CREB, MEF2, MYOD, SRF and TBP contain the annotated binding sites for different regulatory elements [20]; they were generated with the ABS database [39].

Since these eight datasets include binding sites and motifs for eight different TFs, the assessment dataset contains multiple motifs. For a motif, there may be more than one instance in a genomic sequence, the abundance of each motif is assumed to be a known parameter. Therefore, this comprehensive assessment dataset is able to provide convincing assessment results for the case of multiple motifs discovery. The performance comparison of

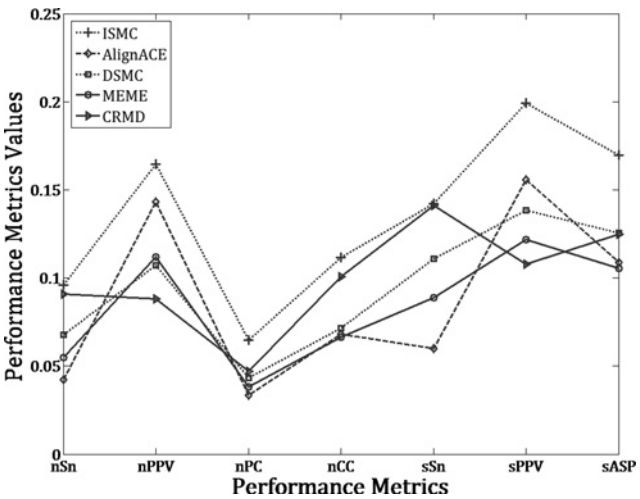

Fig. 5 *Performance comparison using Tompa dataset without motifs abundance information*

IET Signal Process., 2016, Vol. 10, Iss. 5, pp. 504–513

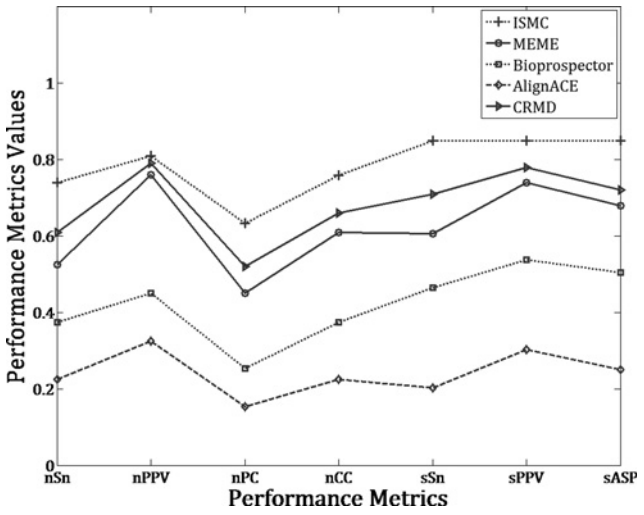

Fig. 6 *Performance comparison using real datasets with multiple motifs*

the ISMC, the CRMD, MEME, AlignACE and BioProspector is based on the metrics used in Section 5.3 [38].

As shown in Fig. 6, compared with the CRMD, MEME, AlignACE and BioProspector, the ISMC algorithm has the best performance with respect to all metrics. It predicted less false positives and has higher positive predictive values, indicating that more of its predictions are the true motifs in the binding sites.

6 Conclusions

In this paper, we propose a novel ISMC solution to the HMM of the motif discovery problem. On the basis of the simulations results, the ISMC algorithm not only outperforms the compared algorithms for datasets without information of motif length and abundance, but also offers significant performance improvement over other algorithms in the case of multiple motifs discovery.

The proposed ISMC algorithm is also competitive in terms of computational complexity. For a computer with an Intel Core 2 Dual E6300 central processing unit, in motif discovery with length and abundance information on a synthetic test dataset with 50 sequences of 200 nucleotides long, the ISMC algorithm takes about 58 s to obtain the solution, whereas the CRMD, the DSMC and the Gibbs sampling algorithm needs about 2 and 4 min, respectively. Research on processing time of different algorithms in condition that motif length and abundance are not known is currently underway.

Though the ISMC algorithms presented in this paper focus on improving the performance of conventional SMC models in motif discovery, and the nucleotides in the motif instance are assumed to be independent to each other, high-order Markov model can be adopted in the proposed ISMC algorithms to cope with the situation that the nucleotides in the motif instance are correlated. The ISMC algorithms can also use the location correlation information of motif instances such as cis-regulatory module [40] to obtain better performance in motif discovery.

7 References

1 Hertz, G., Stormo, G.: 'Identifying DNA and protein patterns with statistically significant alignments of multiple sequences', *Bioinformatics*, 1999, **15**, (7), p. 563
2 Schneider, T.D.: 'Consensus sequence zen', *Appl. Bioinf.*, 2002, **1**, (3), p. 111
3 Buhler, J., Tompa, M.: 'Finding motifs using random projections', *J. Comput. Biol.*, 2002, **9**, (2), pp. 225–242
4 Yu, Q., Huo, H., Zhang, Y., *et al.*: 'PairMotif: a new pattern-driven algorithm for planted (L, D) DNA motif search', 2012
5 Chin, F.Y., Leung, H.C.: 'Voting algorithms for discovering long motifs the research was supported in parts by the Rgc Grant Hku 7135/04e', *Inst. Infocomm Res. (Singapore)*, 2005, **17**, p. 21
6 Davila, J., Balla, S., Rajasekaran, S.: 'Fast and practical algorithms for planted (L, D) motif search', *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2007, **4**, (4), pp. 544–552
7 Dinh, H., Rajasekaran, S., Kundeti, V.K.: 'PMS5: an efficient exact algorithm for the (ℓ, D)-motif finding problem', *BMC Bioinf.*, 2011, **12**, (1), p. 410
8 Ho, E.S., Jakubowski, C.D., Gunderson, S.I.: 'Itriplet, a rule-based nucleic acid sequence motif finder', *Algorithms Mol. Biol.*, 2009, **4**, (14)
9 Dinh, H., Rajasekaran, S., Davila, J.: 'Qpms7: a fast algorithm for finding (ℓ, D)-motifs in DNA and protein sequences', *PloS One*, 2012, **7**, (7), p. e41425
10 Nicolae, M., Rajasekaran, S.: 'Efficient sequential and parallel algorithms for planted motif search', *BMC Bioinf.*, 2014, **15**, (1), p. 34
11 Pisanti, N., Carvalho, A.M., Marsan, L., *et al.*: 'RISOTTO: fast extraction of motifs with mismatches'. Latin 2006: Theoretical Informatics, 2006
12 Floratou, A., Tata, S., Patel, J.M.: 'Efficient and accurate discovery of patterns in sequence data sets', *IEEE Trans. Knowl. Data Eng.*, 2011, **23**, (8), pp. 1154–1168
13 Sagot, M.-F.: 'Spelling approximate repeated or common motifs using a suffix tree'. Latin'98: Theoretical Informatics, 1998
14 Pavesi, G., Mauri, G., Pesole, G.: 'An algorithm for finding signals of unknown length in DNA sequences', *Bioinformatics*, 2001, **17**, (suppl 1), pp. S207–S214
15 Lawrence, C., Reilly, A.: 'An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences', *Proteins: Struct. Funct. Bioinf.*, 1990, **7**, (1), pp. 41–51
16 Bailey, T.L., Williams, N., Misleh, C., *et al.*: 'MEME: discovering and analyzing DNA and protein sequence motifs', *Nucleic Acids Res.*, 2006, **34**, (suppl 2), pp. W369–W373
17 Bi, C.: 'A Monte Carlo EM algorithm for de novo motif discovery in biomolecular sequences', *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2009, **6**, (3), pp. 370–386
18 Hughes, J., Estep, P., Tavazoie, S., *et al.*: 'Computational identification of cis-regulatory elements associated with groups of functionally related genes in saccharomyces cerevisiae1', *J. Mol. Biol.*, 2000, **296**, (5), pp. 1205–1214
19 Liu, X., Brutlag, D., Liu, J.: 'Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes'. The Sixth Pacific Symp. on Biocomputing 2001, 2001
20 Li, G., Chan, T., Leung, K., *et al.*: 'A cluster refinement algorithm for motif discovery', *IEEE/ACM Trans. Comput. Biol. Bioinf.*, 2010, **7**, (4), pp. 654–668
21 Van Dongen, S.M.: 'Graph clustering by flow simulation', 2001
22 Huang, C.-W., Lee, W.-S., Hsieh, S.-Y.: 'An improved heuristic algorithm for finding motif signals in DNA sequences', *IEEE/ACM Trans. Comput. Biol. Bioinf. (TCBB)*, 2011, **8**, (4), pp. 959–975
23 Sun, C., Huo, H., Yu, Q., *et al.*: 'An affinity propagation-based DNA motif discovery algorithm', *BioMed Res. Int.*, 2015, **2015**, pp. 1–10
24 Liang, K., Wang, X., Anastassiou, D.: 'A sequential Monte Carlo method for motif discovery', *IEEE Trans. Signal Process.*, 2008, **56**, (9), pp. 4496–4507
25 Liang, K., Wang, X., Anastassiou, D.: 'A profile-based deterministic sequential Monte Carlo algorithm for motif discovery', *Bioinformatics*, 2008, **24**, (1), p. 46
26 Fearnhead, P.: 'Particle filters for mixture models with an unknown number of components', *Stat. Comput.*, 2004, **14**, (1), pp. 11–21
27 Dempster, A., Laird, N., Rubin, D.: 'Maximum likelihood from incomplete data via the EM algorithm', *J. R. Stat. Soc. B (Methodol.)*, 1977, **39**, (1), pp. 1–38
28 Cappé, O.: 'Online sequential Monte Carlo Em algorithm'. IEEE 15th Workshop on Statistical Signal Processing, 2009
29 Doucet, A., Johansen, A.M.: 'A tutorial on particle filtering and smoothing: fifteen years later', in Dan Crisan, Boris Rozovskii (Eds.): 'Handbook of nonlinear filtering' (Oxford University Press, 2010)
30 Sanjeev Arulampalam, M., Maskell, S., Gordon, N., *et al.*: 'A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking', *IEEE Trans. Signal Process.*, 2002, **50**, (2), pp. 174–188
31 Benos, P., Bulyk, M., Stormo, G.: 'Additivity in protein-DNA interactions: how good an approximation is it?', *Nucleic Acids Res.*, 2002, **30**, (20), p. 4442
32 Bailey, T., Elkan, C.: 'Unsupervised learning of multiple motifs in biopolymers using expectation maximization', *Mach. Learn.*, 1995, **21**, (1), pp. 51–80
33 Roth1JT, F., Hughes, J., Estep, P., *et al.*: 'Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome MRNA quantitation', *Nat. Biotechnol.*, 1998, **16**, p. 939
34 Pevzner, P., Sze, S.: 'Combinatorial approaches to finding subtle signals in DNA sequences'. The Eighth Int. Conf. on Intelligent Systems for Molecular Biology, 2000
35 Stormo, G., Hartzell, G.: 'Identifying protein-binding sites from unaligned DNA fragments', *Proc. Natl. Acad. Sci. USA*, 1989, **86**, (4), p. 1183
36 Liu, J., Gupta, M., Liu, X., *et al.*: 'Statistical models for biological sequence motif discovery', *Case Stud. Bayesian Stat.*, 2004, **6**, pp. 3–23
37 Jensen, S., Liu, J.: 'Biooptimizer: a Bayesian scoring function approach to motif discovery', *Bioinformatics*, 2004, **20**, (10), p. 1271
38 Tompa, M., Li, N., Bailey, T., *et al.*: 'Assessing computational tools for the discovery of transcription factor binding sites', *Nat. Biotechnol.*, 2005, **23**, (1), pp. 137–144
39 Blanco, E., Farré, D., Alba, M., *et al.*: 'AB: a database of annotated regulatory binding sites from orthologous promoters', *Nucleic Acids Res.*, 2006, **34**, (suppl 1), p. D63
40 de-Leon, S.B.T., Davidson, E.H.: 'Gene regulation: gene control network in development', *Annu. Rev. Biophys. Biomol. Struct.*, 2007, **36**, pp. 191–212