

An MCMC algorithm for detecting short adjacent repeats shared by multiple sequences

Qiwei Li¹, Xiaodan Fan^{1,*}, Tong Liang² and Shuo-Yen R. Li²¹Department of Statistics and ²Department of Information Engineering, The Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong

Associate Editor: John Quackenbush

ABSTRACT

Motivation: Repeats detection problems are traditionally formulated as string matching or signal processing problems. They cannot readily handle gaps between repeat units and are incapable of detecting repeat patterns shared by multiple sequences. This study detects short adjacent repeats with interunit insertions from multiple sequences. For biological sequences, such studies can shed light on molecular structure, biological function and evolution.

Results: The task of detecting short adjacent repeats is formulated as a statistical inference problem by using a probabilistic generative model. An Markov chain Monte Carlo algorithm is proposed to infer the parameters in a *de novo* fashion. Its applications on synthetic and real biological data show that the new method not only has a competitive edge over existing methods, but also can provide a way to study the structure and the evolution of repeat-containing genes.

Availability: The related C++ source code and datasets are available at <http://ihome.cuhk.edu.hk/%7Eb118998/share/BASARD.zip>.

Contact: xfan@sta.cuhk.edu.hk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 10, 2011; revised on April 14, 2011; accepted on May 2, 2011

1 INTRODUCTION

In the past decades, the identification of repetitive patterns in biological sequences, especially approximate tandem repeats in DNA sequences, has attracted researchers from many fields. Yet, it still remains as a challenging problem. A tandem repeat in DNA is a sequence segment containing two or more contiguous, approximate copies of a pattern of nucleotides (Benson, 1999). In this article, we expand our views by relaxing the constraint of consecutiveness and allowing of short gaps, i.e. interunit insertions, between neighboring units because gaps have been frequently observed in natural repeats. We name this type of repeats with short gaps as *short adjacent repeats*. An example of short adjacent repeat would be $\cdots \overbrace{ATAT} \overbrace{CGATCCGt} \overbrace{ATCCGcc} \overbrace{ATCCcTC} \cdots$, where a gap composed of one background nucleotide *t* is inserted between the second repeat unit and the third repeat unit, and another gap composed of two background nucleotides *cc* is inserted between the third repeat unit and the fourth repeat unit. The consensus of its sequence pattern *ATCCG* has a width of 5 bp and the copy number

of this repeat equals to 4. The lower case highlights the variations from the consensus.

The detection of short adjacent repeats is of considerable significance. Taking tandem repeats as an instance, in recent decades, more and more researches have suggested that tandem repeats play an important role in genetic mapping (Butler *et al.*, 1997; Weber and May, 1989), gene regulation (Du *et al.*, 1997; Lu *et al.*, 1993) and human diseases (Sinden, 1999; Siyanova and Mirkin, 2001; Sutherland and Richards, 1995).

Almost all repeats detection methods formulated it as either a string matching problem or a signal processing problem. To avoid excessive running time, most of the string matching algorithms are composed of two steps (Benson, 1999; Krishnan and Tang, 2004; Sagot and Myers, 1998; Sokol *et al.*, 2007). They first filter out obvious non-repetitive regions of the sequence using some statistical properties or find out repetitive regions using some heuristical methods, and then search locally for a match between a pattern and a subsequence in those candidate repetitive regions. Recently, more and more signal-processing-based methods have been proposed (Buchner and Janjarasjitt, 2003; Gupta *et al.*, 2007; Sharma *et al.*, 2004; Zhou *et al.*, 2009). Signal-processing techniques, such as discrete Fourier transform, short-time periodicity transform, exact periodic subspace decomposition and autoregressive modelling, are used for spectral analysis.

The above formulations often suffer from two constraints: (i) as they largely rely on the periodicity of a short segment, they cannot readily handle gaps between repeat units; (ii) they could be used to detect tandem repeats in one sequence only, in other words, they are blind to the enriched repetitive pattern shared by multiple sequences where the polymorphic nature might contribute to specialization and generation of diversity in biological functions (Larsen *et al.*, 2005). This article attempts to tackle both issues. Specifically, we not only generalize the repeat model by introducing gaps between neighboring repeat units, but also expand the scope of identifying short adjacent repeats to multiple sequences. This is especially helpful in the case where interunit insertions appear to be common and where the sequence pattern shared by multiple sequences from some particular loci shed light on molecular structure, biological function and evolution.

Our full probabilistic model for repeats detection was inspired by Lawrence *et al.* (1993) and Liu *et al.* (1995), which used the sequence motif model to detect the enriched dispersed pattern in multiple sequences. In the scenario of repeats detection, the same pattern is also enriched in a local neighborhood within each sequence. Our model is built to make use of the two levels of

*To whom correspondence should be addressed.

signal enrichment. We implement a full Bayesian approach to infer globally the parameters of the model. After the Markov chain converges, the sampled parameter values give us a whole picture of their joint posterior distribution. The maximum *a posteriori* (MAP) estimate (Gelman *et al.*, 2004) is used as the point estimate of the interested parameters. Moreover, we use a collapsing technique (Liu *et al.*, 1995) to improve computing efficiency and we design well-organized Markov chain Monte Carlo (MCMC) moves to accelerate the convergence rate.

This article is an extended version of our preliminary work appearing in a conference (Li *et al.*, 2010), where a generative model with a binary vector structure was introduced for identifying short adjacent repeats using primitive Bayesian algorithms. A parallel implementation of that preliminary work using evolutionary Monte Carlo algorithm was also presented in a conference (Xu *et al.*, 2010). In this article, we modify the original generative model with a new repeat segment structure and give a full account of the strategy of designing the Bayesian Approach for Short Adjacent Repeat Detection (BASARD) for the first time, together with a discussion of the statistical power of BASARD under different repeat signal strength. This article also provides a new case study on a real biological dataset that can provide a way to study the structure and the evolution of repeat-containing genes and proteins. Last but not the least, our current method can also be applicable for data analysis such as recognizing repetitive pattern in speeches, texts or images.

The rest of this article is arranged as follows. In Section 2, we formulate the probabilistic generative model, explore the parameters structure and deduce the posterior distribution. In Section 3, we present the schematic procedure and demonstrate the MCMC algorithm, BASARD, step by step. In Section 4, we evaluate BASARD via both synthetic and real data. The last section concludes the article and proposes some potential future work.

2 PROBABILISTIC MODEL

2.1 Generative model

In this subsection, we introduce a parametric model for the input data \mathbf{R} and the four kinds of parameters, namely \mathbf{A} , \mathbf{S} , Θ , and Φ . Note that some of parameter definitions and notations are inherited from Liu *et al.* (1995). Supplementary Table S1 lists the key notations used in this article. Although the model is applicable to any 1D sequences of finite alphabet, we will use DNA case to present the model and the algorithm from now on.

2.1.1 Input data \mathbf{R} The input data is composed of N sequences denoted by $R_n, 1 \leq n \leq N$, where $R_n = \{r_{n,1}, r_{n,2}, \dots, r_{n,L_n}\}$. Each residue $r_{n,l}$ is sampled from a finite alphabet χ ($\chi = \{A, T, C, G\}$ for DNA sequences). $L_n, 1 \leq n \leq N$, is the length of the n -th sequence. In this model, we assume that the input sequences are mutually independent. Therefore, the full likelihood function can be written as $P(\mathbf{R}|\mathbf{A}, \mathbf{S}, \Theta, \Phi) = \prod_{n=1}^N P(R_n|a_n, s_n, \Theta, \Phi)$.

2.1.2 Parameters \mathbf{A} : Locations of repeat segments A set of repeat segment starting positions, denoted by \mathbf{A} , can be written as $\mathbf{A} = [a_1 \ a_2 \ \dots \ a_N]^T$, where T denotes matrix transpose and $1 \leq a_n \leq L_n$. Here, we assume that there is only one repeat segment per sequence. The case of multiple repeat segments per sequence is discussed in Section 5.

For the prior distribution of a_n , we use a uniform distribution, which means all positions in each sequence is priorly equally likely to start a repeat segment. We further assume the independence among all elements in \mathbf{A} , i.e. $P(\mathbf{A}) = \prod_{n=1}^N P(a_n) \propto 1$.

2.1.3 Parameters \mathbf{S} : structures of repeat segments A set of repeat segment structures, denoted by \mathbf{S} , can be written as $\mathbf{S} = [s_1^T \ s_2^T \ \dots \ s_N^T]^T$. Here $s_n, 1 \leq n \leq N$, is a base- $(G+1)$ numerical vector of the format

$$s_n = \left[\overbrace{g_{n,1} \ g_{n,2} \ \dots \ g_{n,\Omega_n-1} \ -1 \ \dots \ -1}^{(\Omega-1) \text{ elements in total}} \right],$$

where $\Omega_n, \Omega_n \geq 1$, is the copy number of the n -th repeat segment, and the variable $g_{n,\omega}, 1 \leq \omega \leq \Omega_n - 1$, is the gap length between the ω -th repeat unit and the $(\omega+1)$ -th repeat unit within s_n . G is the maximum allowed gap length and therefore $0 \leq g_{n,\omega} \leq G$. It is required that each s_n should be the same dimension by filling with the trivial value -1 from the Ω_n -th to the $(\Omega-1)$ -th element of each s_n , where Ω is the maximum allowed copy number. If there is only one repeat unit, all elements of this vector s_n will be -1 .

Let

$$z_{n,\omega} = \begin{cases} 1 & \text{if } \omega = 1, \\ (\omega-1)J + \sum_{\kappa=1}^{\omega-1} g_{n,\kappa} + 1 & \text{if } 2 \leq \omega \leq \Omega_n - 1. \end{cases} \quad (1)$$

It indicates that the ω -th repeat unit in the n -th sequence starts at the position $a_n + z_{n,\omega} - 1$ of the sequence R_n . Based on the assumption that each repeat segment is composed of multiple repeat units separated by gaps of random length, we are now able to deal with the case of interunit insertions, using the base- $(G+1)$ numeral vector.

Since longer gaps are usually less likely to occur in nature compared with shorter gaps, we might assume $P(s_n)$ decays exponentially with both the copy number and the total length of gaps, e.g. $P(s_n) \propto \varepsilon_1^{\Omega_n} \varepsilon_2^{\sum_{\omega=1}^{\Omega_n-1} g_{n,\omega}}$, where both of the constants ε_1 and ε_2 are in the range of $(0, 1]$. We also assume that all s_n are mutually independent. Thus, we have $P(\mathbf{S}) = \prod_{n=1}^N P(s_n) \propto \varepsilon_1^{\sum_{n=1}^N \Omega_n} \varepsilon_2^{\sum_{n=1}^N \sum_{\omega=1}^{\Omega_n-1} g_{n,\omega}}$. This prior distribution choice avoids the transdimensional problem while updating the copy number parameters by implicitly requiring $P(g_{n,\omega} = -1) = 1$ for $\Omega_n \leq \omega \leq \Omega - 1$.

2.1.4 Parameters Θ : motif matrix The motif matrix is denoted by Θ . For DNA sequences, it can be written as a $4 \times J$ motif matrix as $\Theta = [\theta_1 \ \theta_2 \ \dots \ \theta_J]$, where $\theta_j = [\theta_{A,j} \ \theta_{T,j} \ \theta_{C,j} \ \theta_{G,j}]^T, 1 \leq j \leq J$, is the relative frequencies of finding each letter $k, k \in \chi$, at the position j among all repeat units. Note that $\sum_k \theta_{k,j} = 1, \forall j$ and $\theta_{k,j} \geq 0, \forall k, j$.

It is assumed that all repeat units are independent and identical samples from this motif matrix Θ . Θ is the parameter of a Product Multinomial (PM) distribution (Liu *et al.*, 1995). We assume Θ follows a Product Dirichlet (PD) prior distribution with the parameter \mathbf{B} (Liu *et al.*, 1995) as $\mathbf{B} = [\beta_1 \ \beta_2 \ \dots \ \beta_J]$, where $\beta_j = [\beta_{A,j} \ \beta_{T,j} \ \beta_{C,j} \ \beta_{G,j}]^T, 1 \leq j \leq J$. Each θ_j is an independent 4D Dirichlet random vector following the distribution $\text{Dirichlet}(\beta_j)$.

Without any prior knowledge, we set all elements in \mathbf{B} equal to 1. As a result, we have $P(\Theta) \propto 1$.

2.1.5 Parameters Φ : Background distribution The background distribution is a multinomial distribution parameterized by Φ . For the DNA sequence case, Φ can be written as $\Phi = [\phi_A \ \phi_T \ \phi_C \ \phi_G]^T$, where $\phi_k, k \in \chi$ is the probabilities of finding each letter k at a non-unit position. Note that $\sum_k \phi_k = 1$ and $\phi_k \geq 0, \forall k$.

We assume that all nucleotides at non-unit positions are independent and identical samples from the background distribution. Similar to Θ , we assume Φ follows a Dirichlet priors with parameter α , where $\alpha = [\alpha_A \ \alpha_T \ \alpha_C \ \alpha_G]^T$. Analogously, we set all elements in α equal to 1. As a result, we have $P(\Phi) \propto 1$.

For ease of presentation, Supplementary Figure S1 displays an example of schematic diagram of our model.

2.2 Parameter structure

Our scheme maintains two evolving parameter groups. The first group containing \mathbf{A} and \mathbf{S} indicates the locations of all repeat units. The second group that comprises Θ and Φ describes the motif pattern and the background distribution, respectively.

Before investigating the relationship between the two parameter groups, we introduce the following function and notation. For a given set of categorical data, e.g. $Y = \{y_1, \dots, y_l, \dots, y_L\}$, where each y_l takes values from the finite alphabet χ , we define the counting function \mathbf{h} (Liu et al., 1995) such that $\mathbf{h}(Y) = [m_A \ m_T \ m_C \ m_G]^T$, where $m_k, k \in \chi$ is the total number of letter k observed in Y . Let $\Psi = \{(n, l) : n = 1, \dots, N; l = 1, \dots, L_n\}$ denote the collection of indices in \mathbf{R} . For any set $W \subseteq \Psi$, we define $\mathbf{R}_W = \{r_{n,l} : (n, l) \in W\}$. We also define $W^C = \Psi \setminus W$ as the set of all elements which are members of Ψ but not members of W .

Given \mathbf{A} and \mathbf{S} , the observation of all repeat units' indices can be denoted as set $U = \{(n, a_n + z_{n,\omega} - 1 + j - 1)\}$, where $n = 1, \dots, N, \omega = 1, \dots, \Omega_{n-1}, j = 1, \dots, J$, and $z_{n,\omega}$ is described by Equation (1). We also write the set of the residues occupied in the j -th positions of all repeat units as $\mathbf{R}_{U(j)}$, and the corresponding set of indices is $U(j) = \{(n, a_n + z_{n,\omega} - 1 + j - 1)\}$, where $n = 1, \dots, N, \omega = 1, \dots, \Omega_{n-1}$. As U gives the indices of all repeat units, the observations of all nucleotides at non-unit positions can be denoted as \mathbf{R}_{U^C} .

We can get the sufficient statistics by applying the counting function \mathbf{h} on each $\mathbf{R}_{U(j)}$ and \mathbf{R}_{U^C} . We denote the results as a $4 \times J$ matrix $\mathbf{h}(\mathbf{R}_U) = [\mathbf{h}(\mathbf{R}_{U(1)}) \ \mathbf{h}(\mathbf{R}_{U(2)}) \ \dots \ \mathbf{h}(\mathbf{R}_{U(J)})]$, where $\mathbf{h}(\mathbf{R}_{U(j)}) = [m_{A,j} \ m_{T,j} \ m_{C,j} \ m_{G,j}]^T$ and a 4×1 vector $\mathbf{h}(\mathbf{R}_{U^C}) = [w_A \ w_T \ w_C \ w_G]^T$, respectively. Note that the summation of all elements in $\mathbf{h}(\mathbf{R}_{U(j)})$ equals to the total number of the observed repeat units, i.e. $\sum_{n=1}^N \Omega_n$, for all j and the summation of all elements in $\mathbf{h}(\mathbf{R}_{U^C})$ equals to the total number of nucleotides within the background area, i.e. $\sum_{n=1}^N L_n - J \sum_{n=1}^N \Omega_n$. As the same as in Liu et al. (1995), $\mathbf{h}(\mathbf{R}_U)$ follows a PM distribution with parameters Θ , and each $\mathbf{h}(\mathbf{R}_{U(j)})$ follows the multinomial distribution, i.e. $\mathbf{h}(\mathbf{R}_{U(j)}) \sim \text{Multinomial}(\sum_{n=1}^N \Omega_n; \theta_j)$. In addition, $\mathbf{h}(\mathbf{R}_{U^C})$ follows a multinomial distribution with parameters Φ , i.e. $\mathbf{h}(\mathbf{R}_{U^C}) \sim \text{Multinomial}(\sum_{n=1}^N L_n - J \sum_{n=1}^N \Omega_n; \Phi)$.

For the Bayesian inference of Θ and Φ , we use conjugate priors, which are the PD distribution and the Dirichlet distribution, respectively. Therefore, the posterior distribution of Θ is $PD(\mathbf{B} + \mathbf{h}(\mathbf{R}_U))$ and the posterior distribution of Φ is Dirichlet($\alpha + \mathbf{h}(\mathbf{R}_{U^C})$).

2.3 Posterior distribution

In this subsection, we first deduce the full posterior distribution. Then, a collapsing technique is used to make the proposed algorithm introduced in the next section more efficient.

2.3.1 The full posterior distribution Given all the parameters \mathbf{A} , \mathbf{S} , Θ and Φ , the likelihood of observing the given data \mathbf{R} can be written as

$$P(\mathbf{R}|\mathbf{A}, \mathbf{S}, \Theta, \Phi) = \Phi^{\mathbf{h}(\mathbf{R}_{U^C})} \prod_{j=1}^J \theta_j^{\mathbf{h}(\mathbf{R}_{U(j)})}, \quad (2)$$

where we define the vector power of a vector as the product of all elements after taking corresponding power, i.e. $\theta_j^{\mathbf{h}(\mathbf{R}_{U(j)})} = \theta_{A,j}^{m_{A,j}} \theta_{T,j}^{m_{T,j}} \theta_{C,j}^{m_{C,j}} \theta_{G,j}^{m_{G,j}}$ and $\Phi^{\mathbf{h}(\mathbf{R}_{U^C})} = \phi_A^{w_A} \phi_T^{w_T} \phi_C^{w_C} \phi_G^{w_G}$. With mutually independent priors, the joint posterior distribution of \mathbf{A} , \mathbf{S} , Θ and Φ can be written as $P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R}) \propto P(\mathbf{R}|\mathbf{A}, \mathbf{S}, \Theta, \Phi)P(\mathbf{A})P(\mathbf{S})P(\Theta)P(\Phi)$. According to the complete-data likelihood given by Equation (2) and the prior distributions of each kind of parameters described previously, we can obtain the full posterior distribution as

$$P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R}) \propto \varepsilon_1^{\sum_{n=1}^N \Omega_n} \varepsilon_2^{\sum_{n=1}^N \sum_{\omega=1}^{\Omega_n-1} g_{n,\omega}} \Phi^{\mathbf{h}(\mathbf{R}_{U^C})} \prod_{j=1}^J \theta_j^{\mathbf{h}(\mathbf{R}_{U(j)})}. \quad (3)$$

2.3.2 The collapsed posterior distribution The algorithm based on the above joint posterior distribution can be inefficient because of the high dimensionality of the parameter space. One solution is to integrate out the nuisance parameters. As \mathbf{A} and \mathbf{S} are mainly concerned, Θ and Φ are thus not parameters of interest, or at least, it is not difficult to estimate approximate Θ and Φ conditional on \mathbf{A} and \mathbf{S} . Using the collapsing technique (Liu et al., 1995), we can actually integrate out Θ and Φ in order to make the proposed algorithm more efficient in terms of computing.

Note that $P(\mathbf{A}, \mathbf{S}|\mathbf{R}) = \int \int P(\mathbf{A}, \mathbf{S}, \Theta, \Phi|\mathbf{R}) d\Theta d\Phi$, our choices of the Dirichlet priors for Θ and Φ enable us to integrate out both Θ and Φ . The resulting collapsed posterior distribution can be written as

$$P(\mathbf{A}, \mathbf{S}|\mathbf{R}) \propto \varepsilon_1^{\sum_{n=1}^N \Omega_n} \varepsilon_2^{\sum_{n=1}^N \sum_{\omega=1}^{\Omega_n-1} g_{n,\omega}} \frac{\Gamma(|\alpha|) \Gamma(\mathbf{h}(\mathbf{R}_{U^C}) + \alpha)}{\Gamma(\alpha/\Gamma(|\mathbf{h}(\mathbf{R}_{U^C})| + |\alpha|))} \prod_{j=1}^J \frac{\Gamma(|\beta_j|) \Gamma(\mathbf{h}(\mathbf{R}_{U(j)}) + \beta_j)}{\Gamma(\beta_j/\Gamma(|\mathbf{h}(\mathbf{R}_{U(j)})| + |\beta_j|))}, \quad (4)$$

where we define the absolute value of a vector as the summation of all elements within the vector.

3 COMPUTING METHODS

At the beginning of this section, we present the basic and the improved schematic procedures, respectively. Then, we describe BASARD based on the improved schematic procedure step by step.

3.1 Schematic procedure

Given a set of N sequences \mathbf{R} , our objective is to identify the location and the structure of the most probable repeat segment within each sequence. These N repeat segments are obtained by locating \mathbf{A} and adjusting \mathbf{S} that maximize the posterior probability. The MCMC algorithm makes a full tour of the target distribution $P(\mathbf{A}, \mathbf{S}, \Theta, \Phi | \mathbf{R})$ or $P(\mathbf{A}, \mathbf{S} | \mathbf{R})$. When it converges, we output the MAP of \mathbf{A} and \mathbf{S} .

We can use Metropolis-in-Gibbs scheme (Gelman *et al.*, 2004) on Equation (3) to iteratively sample one set of parameters given all the other sets of parameters as well as the observed data: (i) sample and update \mathbf{A} conditional on \mathbf{S} , Θ , Φ and \mathbf{R} ; (ii) sample and update \mathbf{S} conditional on \mathbf{A} , Θ , Φ and \mathbf{R} ; (iii) sample and update Θ and Φ conditional on \mathbf{A} , \mathbf{S} and \mathbf{R} . However, this standard Metropolis-in-Gibbs scheme based on the full posterior distribution is often too time consuming. To improve the computing efficiency, we work on the collapsed posterior distribution described in Equation (4). The improved schematic procedure of the MCMC algorithm, namely BASARD is described as follows.

BASARD proceeds through iterations after initialization, each of which updates a_n and s_n one sequence after another from 1 to N . When stochastically updating a_n and s_n for one sequence, we pretend that the repeat segments of the remaining $(N-1)$ sequences have been known. Specifically, when the n -th sequence R_n is selected, we use the given information, $\mathbf{A}_{[-n]}$ and $\mathbf{S}_{[-n]}$, to estimate the 'motif matrix' $\hat{\Theta}_n$ and 'background distribution' $\hat{\Phi}_n$ so as to determine new a_n and s_n sequentially. Here, $\mathbf{A}_{[-n]}$ and $\mathbf{S}_{[-n]}$ denote the set of locations and structures, respectively, of all repeat segments excluding the n -th sequence. Intuitively, the more accurate the estimated $\hat{\Theta}_n$ and $\hat{\Phi}_n$ constructed in the predictive update step, the more accurate the determination of a_n and s_n in the following sampling steps, and vice versa.

3.2 Initialization and predictive update step

The first step is to initialize \mathbf{A} and \mathbf{S} . As each s_n is a $1 \times \Omega$ base- $(G+1)$ numeral vector, we first choose a possible copy number in the range of $[1, \Omega]$, and then we assign a random gap number g in the range of $[0, G]$ for each pair of adjacent repeat units. \mathbf{A} is a $N \times 1$ vector and the initial value of each a_n is randomly chosen from its possible values in the range of $[1, L_n - J\Omega_n - \sum_{\omega=1}^{\Omega_n-1} g_{\omega} + 1]$.

Suppose that we are proceeding through the i -th iteration. Let $\hat{\theta}_{n,j}^{(i)}$ denote the j -th column vector in $\hat{\Theta}_n^{(i)}$, then each estimated $\hat{\theta}_{n,j}^{(i)}$ is calculated as,

$$\hat{\theta}_{n,j}^{(i)} = \frac{\beta_j + \mathbf{h}(\mathbf{R}_{U_{[-n]}^{(i-1)}}(j))}{|\beta_j| + \sum_{\kappa=1, \kappa \neq n}^N \Omega_{\kappa}^{(i-1)}}, \quad (5)$$

where $U_{[-n]}$ denotes all repeat units indices excluding those repeat units within R_n . Also, the estimated $\hat{\Phi}_n$ is calculated as,

$$\hat{\Phi}_n^{(i)} = \frac{\alpha + \mathbf{h}(\mathbf{R}_{U_{[-n]}^C}^{(i-1)})}{|\alpha| + \sum_{\kappa=1, \kappa \neq n}^N L_{\kappa} - J\Omega_{\kappa}^{(i-1)}}, \quad (6)$$

where $U_{[-n]}^C$ denotes all background nucleotides indices excluding those nucleotides within R_n .

3.3 Gibbs sampling step for a_n

Conditional on the repeat segment structure s_n , we consider every possible repeat segment X in R_n . Using the results obtained in the predictive update step and according to Equation (3), we can calculate the probability of generating those matching repeat units within X according to the current 'motif matrix' $\hat{\Theta}_n^{(i)}$ and the probability of generating all letters within X according to the 'background distribution' $\hat{\Phi}_n^{(i)}$, respectively. The ratio of these two probabilities is assigned as the sampling weight to each X , of which starting position is from 1 to $L_n - \Omega_n^{(i-1)}J - \sum_{\omega=1}^{\Omega_n^{(i-1)}-1} g_{n,\omega} + 1$. To sum up,

we use Gibbs sampling to update new $a_n^{(i)} = a'_n$ as follows:

$$P(a'_n | s_n^{(i-1)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R}) \propto \prod_{j=1}^J \prod_{\omega=1}^{\Omega_n^{(i-1)}-1} \frac{\hat{\theta}_{n,a'_n+z_{n,\omega}+j-1}^{(i)}}{\hat{\theta}_{n,a'_n+z_{n,\omega}+j-1}^{(i)}}. \quad (7)$$

3.4 Metropolis-Hastings sampling step for s_n

The base- $(G+1)$ numeral vector s_n , of which dimension is $\Omega-1$, allows $1 + \sum_{\omega=1}^{\Omega-1} (G+1)^{\omega}$ states, which is usually a very large number. Thus, it is difficult to compute the normalization constant of $P(s_n | a_n, \hat{\Theta}_n, \hat{\Phi}_n, \mathbf{R})$. We will use the Metropolis-Hastings sampling to update $s_n^{(i)}$. In our case, the Hastings ratio can be written as

$$\lambda = \frac{P(s'_n | a_n^{(i)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R}) P(s_n^{(i-1)}; s'_n)}{P(s_n^{(i-1)} | a_n^{(i)}, \hat{\Theta}_n^{(i)}, \hat{\Phi}_n^{(i)}, \mathbf{R}) P(s'_n; s_n^{(i-1)})}, \quad (8)$$

where $P(s'_n; s_n^{(i-1)})$ is the proposal density, which specifies the probability of proposing a move to s'_n given the previous state $s_n^{(i-1)}$ and $P(s_n^{(i-1)}; s'_n)$ is the flipped case. The move is accepted, $s_n^{(i)} = s'_n$, with the probability $\min(1, \lambda)$; otherwise, the move is rejected, $s_n^{(i)} = s_n^{(i-1)}$.

In order to make the Markov chain ergodic and fast convergent, we design five types of moves: rear insertion, rear deletion, partial shift, front insertion and front deletion as shown in the Supplementary Figure S10. We denote Q_{κ} , $1 \leq \kappa \leq 5$, as the probability to propose the corresponding type of move on each MCMC step. For ease of presentation, the first two types are clustered into a group named rear indel and the last two types are clustered into a group named front indel.

3.4.1 Rear indel moves The first category of moves is to insert a repeat unit from one of the $(G+1)$ possible positions behind the last repeat unit with proposing probability Q_1 or to delete the last repeat unit with proposing probability Q_2 . According to Equation (1), $a_n^{(i)} + z_{n,\Omega_n^{(i-1)}-1} + J - 2$ is the ending position of the last repeat unit within R_n . Hence, for the attempt of rear insertion in the current iteration, we first choose a possible position v in the range of $[a_n^{(i)} + z_{n,\Omega_n^{(i-1)}-1} + J - 1, a_n^{(i)} + z_{n,\Omega_n^{(i-1)}-1} + J - 1 + G]$, and then calculate the Hastings ratio as follows:

$$\lambda = \varepsilon_1 \varepsilon_2^{g'} \frac{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1}}^{(i)}}{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1}}^{(i-1)}} \frac{Q_2}{Q_1 / (G+1)}, \quad (9)$$

where $g' = v - (a_n^{(i)} + z_{n,\Omega_n^{(i-1)}-1} + J - 1)$.

For the rear deletion case, the last repeat unit within s_n is located at $v = a_n^{(i)} + z_{n,\Omega_n^{(i-1)}-1} - 1$, therefore the Hastings ratio is

$$\lambda = \frac{1}{\varepsilon_1 \varepsilon_2^{g_{n,\Omega_n^{(i-1)}-1}}} \frac{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1}}^{(i)}}{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1}}^{(i-1)}} \frac{Q_1 / (G+1)}{Q_2}. \quad (10)$$

3.4.2 Partial shift moves This type of moves is to make a selected subsegment, i.e. from the b -th repeat unit ($b \in [2, \Omega_n]$) to the last repeat unit, shifted left or right by up to a possible number with proposing probability Q_3 . We randomly choose the order b in the range of $[2, \Omega_n]$, and then we randomly choose the shift number μ among all its possible values in the range of $[-(z_{n,b}^{(i-1)} - z_{n,b-1}^{(i-1)} - J), G - (z_{n,b}^{(i-1)} - z_{n,b-1}^{(i-1)} - J)]$, where the negative number indicates left shifts and the positive number means right shift. The Hastings ratio in this case is

$$\lambda = \frac{\prod_{j=1}^J \prod_{\omega=b-1}^{\Omega_n^{(i-1)}-1} \frac{\hat{\theta}_{r_{n,a_n+z_{n,\omega}+j-1+\mu}}^{(i)}}{\hat{\theta}_{r_{n,a_n+z_{n,\omega}+j-1+\mu}}^{(i-1)}}}{\prod_{j=1}^J \prod_{\omega=b-1}^{\Omega_n^{(i-1)}-1} \frac{\hat{\theta}_{r_{n,a_n+z_{n,\omega}+j-1}}^{(i)}}{\hat{\theta}_{r_{n,a_n+z_{n,\omega}+j-1}}^{(i-1)}}}. \quad (11)$$

3.4.3 Front indel moves Similar to the rear indel moves, we design the insertion moves, with proposing probability Q_4 , as adding a repeat unit from a possible position v in the range of $[a_n^{(i)} - G - J, a_n^{(i)} - J]$ in front of the first repeat unit. If the move is accepted, we renew $a_n^{(i)}$ with v at the same time. Also, we design the deletion move, with proposing probability Q_5 , as removing the first repeat unit within s_n while renewing $a_n^{(i)}$ with $a_n^{(i)} + z_{n,2}^{(i-1)} - 1$ if $\lambda \geq 1$. The corresponding Hastings ratios for the insertion case and the deletion case are

$$\lambda = \varepsilon_1 \varepsilon_2^{g'} \frac{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1},j}^{(i)}}{\prod_{j=1}^J \hat{\theta}_{r_{n,v+j-1},j}^{(i)}} \frac{Q_5}{Q_4/(G+1)}, \quad (12)$$

and

$$\lambda = \frac{1}{\varepsilon_1 \varepsilon_2^{g_{n,1}^{(i-1)}}} \frac{\prod_{j=1}^J \hat{\phi}_{r_{n,a_n^{(i)}+j-1}}^{(i)}}{\prod_{j=1}^J \hat{\theta}_{r_{n,a_n^{(i)}+j-1},j}^{(i)}} \frac{Q_4/(G+1)}{Q_5}, \quad (13)$$

respectively, where $g' = v - \left(a_n^{(i)} - G - J\right)$.

For ease of presentation, Supplementary Figure S2 shows an example of full state transition diagram for all s_n , $1 \leq n \leq N$.

3.5 Phase shifts

Although the collapsed sampler seems to work well in the MCMC algorithm, it may face the phase problems (Lawrence *et al.*, 1993), which gets BASARD stuck in a local optimum. The solution is to compare the current \mathbf{A} with sets shifted left and right (Lawrence *et al.*, 1993). We randomly choose a reasonable number μ and denote a $1 \times N$ vector $\boldsymbol{\mu}$ with all elements equal to μ . Then, the Hastings ratio can be written as

$$\lambda = \frac{P(\mathbf{A} + \boldsymbol{\mu} | \mathbf{R}, \mathbf{S})}{P(\mathbf{A} | \mathbf{R}, \mathbf{S})} \propto \frac{P(\mathbf{R} | \mathbf{A} + \boldsymbol{\mu}, \mathbf{S})}{P(\mathbf{R} | \mathbf{A}, \mathbf{S})}. \quad (14)$$

We relocate the starting position of each repeat segment to position $a_n + \mu$ if the move is accepted.

4 RESULTS AND DISCUSSION

We design two experiments to evaluate our algorithm. The first experiment on synthetic data explores the convergence property and the statistical performance of BASARD. This experiment also answers the following question numerically: (i) how does the repeat signal strength, in terms of the degeneracy degree of motif model and the copy number of repeat segments, affect the power of BASARD and other methods? (ii) how does the choice of prior distribution affect the estimation accuracy of BASARD? (iii) how to choose the pattern width within a range of plausible ones?

The other experiment on real data demonstrates the superiority of BASARD over the existing methods to identify a short adjacent repeat in exon III of the dopamine receptor D4 (*DRD4*) gene from different mammalian species. All experiments are conducted on a PC with 2.40G CPU and 4.00G memory.

4.1 Evaluation with synthetic data

4.1.1 Generating synthetic datasets We use synthetic data to carry out the experiment so that we can use the known ground truth to evaluate our algorithm. The synthetic DNA sequence set \mathbf{R} are sampled according to the generative model. It contains N sequences of the same length L , each of which contains only one repeat segment. The parameter values of \mathbf{S} and \mathbf{A} are independently drawn from the following prior distributions: $P(\Omega_n) \propto \varepsilon_1^{\Omega_n}, \Omega_{\min} \leq$

$\Omega_n \leq \Omega_{\max}$; $P(g_{n,\omega}) \propto \varepsilon_2^{g_{n,\omega}}, 0 \leq g_{n,\omega} \leq G$; and $P(a_n) \propto 1, 1 \leq a_n \leq L - J\Omega_n - \sum_{\omega=1}^{\Omega_n-1} g_{n,\omega} + 1$. Here, we fix $\varepsilon_1 = 0.5$, $\varepsilon_2 = 0.5$ and $G = 2$. All repeat units are generated using one of the six experimental motif models listed in the Supplementary Table S2. The six motif models are selected from JASPAR CORE Vertebrata database (Bryne *et al.*, 2008) such that they represent motifs of different pattern width and degeneracy as measured by average entropy. The average entropy is defined as the average information content (Schneider and Stephens, 1990) across all positions. Each motif model is assigned one of the six abbreviations: 6L, 6M, 6H, 12L, 12M and 12H, where the prefix number indicates its pattern width and the suffix signs L, M and H denote low, medium and high degeneracy degree, respectively. The background nucleotides are sampled with $\Phi = [0.25 \ 0.25 \ 0.25 \ 0.25]^T$.

4.1.2 Convergence diagnosis We used the synthetic data to check the convergence of our algorithm. Multiple independent chains starting from randomly sampled parameter values were run. The joint posterior probability was evaluated at each iteration. The joint posterior probability evaluated at the true parameter values was used as a reference. Details of this convergence diagnosis are given in the Supplementary Material. It shows that all chains converged fast to the reference value, which indicates that our algorithm has satisfactory convergence property.

4.1.3 Statistical performance To assess the accuracy of BASARD under different repeat signal strength, 12 categories of datasets, where $N=6$ and $L=2000$, were synthesized. For convenience, we denote L as the copy number ranges in [10, 15] and S as the copy number ranges in [5, 10]. Combined with the previous defined abbreviation of motif model, the 12 naturally combined categories can be represented as 12L-L, 12M-L, 12H-L, 12L-S, 12M-S, 12H-S, 6L-L, 6M-L, 6H-L, 6L-S, 6M-S and 6H-S. Taking 12M-L as an example, the prefix sign 12M indicates that all repeat units are generated by the motif model with abbreviation 12M and the suffix sign L means that the copy number $\Omega_n \in [10, 15], \forall n$, in the dataset.

To quantify the accuracy of BASARD, we use the performance metric F -score as it considers both the *precision* and the *recall*. *Precision* is defined as the number of actual repeat units that are correctly estimated, divided by the number of all estimated repeat units. *Recall* is defined as the number of actual repeat units that are correctly estimated, divided by the number of all actual repeat units. F -score, which takes values between 0 and 1, is defined as $2 * \text{precision} * \text{recall} / (\text{precision} + \text{recall})$. The larger it is, the more accurate the result is.

To demonstrate the superiority of BASARD, we compared it with Tandem Repeats Finder [TRF, Benson (1999)] and Gibbs Motif Sampler [GMS, Thompson *et al.* (2003)]. TRF is one of efficient and popular tools for detecting tandem repeats (Du *et al.*, 2007) and GMS is a classic motif discovery software. To compare them fairly, we evaluated the results from all three methods at the nucleotide level where each nucleotide was labelled as either an element of repeat units or an element of background area.

For each of the 12 categories, we independently generated 100 datasets. For each of the 100 datasets, we ran all the three methods and computed their individual *precision*, *recall* and F -score. The detailed settings and results of BASARD, TRF and GMS are

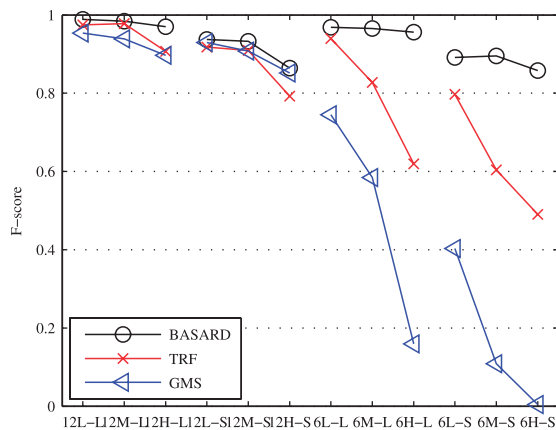


Fig. 1. The average F -score achieved by BASARD, TRF and GMS under different repeat signal strength.

summarized in the Supplementary Tables S4 and S5. The average F -score under different repeat signal strength are displayed in Figure 1. It shows that the performance of BASARD generally outperforms TRF and GMS in terms of estimation accuracy. When the repeat signal strength is powerful enough like 12L-L, 12M-L and 12H-L, the statistical performance insignificantly differ from each other. Weakening repeat signal strength leads to greater disparity between BASARD and the other methods. It also shows that both TRF and GMS are more sensitive to the degeneracy degree of motif model than BASARD.

With the result from this experiment, we conclude as follows: (i) BASARD performs better than TRF because TRF uses a window-based method to search local enriched pattern for each sequence without making use of the fact that the enriched pattern is also shared by multiple sequences. (ii) BASARD consistently outperforms GMS because BASARD considers that repeat units cluster round a local neighborhood within each DNA sequence. This is mainly because GMS is not designed for repeats detection, while BASARD is a generalization of GMS to detect repeats. Neglect of local clustering information results in lower efficiency and higher false positive rate, e.g. GMS usually reports a remote pseudo motif instance.

In summary, in the scenario of short adjacent repeats detection, we not only use the sequence motif model to detect the enriched pattern in multiple sequences, but also consider such pattern is also enriched in a local neighborhood within each sequence. It is the use of the two levels of signal enrichment that makes BASARD perform the best.

4.1.4 Sensitivity to priors and pattern width selection We again used simulation on synthetic datasets to check the sensitivity of BASARD to different prior distributions and discuss on the pattern width selection problem. The details are given in the Supplementary Material. The results show that the statistical power of BASARD is reasonably insensitive to key prior choices. The experiments also suggest that the information per parameter (Lawrence *et al.*, 1993) can be used to select the pattern width.

4.2 Experiment on real data

4.2.1 Introduction to the real data Schoots and Tol (2003) suggested functional implications in the size variation of the tandem

repeat in exon III in the human *DRD4* gene. One explanation is that the length of the tandem repeat modulates the level of expression of *DRD4*, which is associated with cognitive function (Previc, 1999).

Larsen *et al.* (2005) and Mogensen *et al.* (2006) used TRF to identify a tandem repeat composed of 18 bp basic units in exon III of *DRD4* from different mammals. However, TRF is not perfect for such multiple sequence case because the information that all input sequences share the same motif pattern cannot be utilized. In addition, the output results have to be manually amended so as to keep the consensus pattern and period size consistent with each other.

We used BASARD to detect short adjacent repeats in exon III of the *DRD4* genes from different mammalian species. We retrieved 24 public available DNA sequences and their corresponding amino acid sequences of the *DRD4* exon III from GenBank. The detailed information is given in the Supplementary Material.

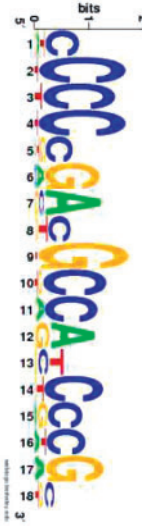
4.2.2 Results Supplementary Table S6 summarizes the settings of BASARD and TRF that were implemented in this experiment. Important characteristics of those repeats detected by BASARD and TRF are summarized in Table 1.

Both methods found that a repeat with pattern width 18 bp was shared by all sequences. However, we have to use cyclic alignment on the consensus patterns reported by TRF in order to really see their similarity. Both BASARD and TRF pointed that the copy number varied from about 3 in the bear family to about 20 in the horse species. Different copy number variants were also detected even in the same species such as sperm whale, minke whale, domestic horse, donkey and onager. The average entropy of the motif model is 1.0958, reflecting significant sequence degeneracy across repeat units. As the same as Larsen *et al.* (2005) and Mogensen *et al.* (2006) observed, we also found the sequence motif is GC rich. For the cases where TRF reported more copy numbers than BASARD did, it is mainly due to the deletion inside those repeat units. Such deletion case is not yet modelled by our current method.

One interesting discovery from this case study is that the short adjacent repeat identified by BASARD could be classified to two types of repeat segment structure, namely Type I and Type II. Asiatic black bear, polar bear and domestic cat are of Type I where there is no gap in repeat segments. All the other species shared structure Type II where these are an extra repeat unit in the end of tandem repeat segment. This observation probably implies the closer evolutionary relationship between species who shared the same repeat segment structure. As a matter of fact, the clade IV of the phylogenetic tree shown in Murphy *et al.* (2001) verifies this conjecture. *Ursus* (e.g. bear) and *Felis* (e.g. cat) that belong to carnivora lineage had repeat segment structure of Type I while *Ceratotherium* (e.g. whale, dolphin and porpoise) and *Equus* (e.g. horse, donkey and zebra) that belong to perissodactyla lineage had repeat segment structure of Type II. It might indicate that this shared extra repeat unit was gained or lost before the specialization of the studied species had differentiated into these two different lineages.

We also ran BASARD on the corresponding amino acid sequence set to detect the protein repeat motif. Supplementary Tables S6 and S7 summarize the settings of BASARD and the important characteristics of the detected amino acid repeats. As a comparison, Larsen *et al.* (2005) mentioned that they were unable to detect repeat motifs in the deduced amino acid sequences from the two types of bear and otter. However, BASARD reported 1 repeat unit for each of

Table 1. The comparison of the repeats detected by BASARD and TRF

Species	GenBank accession number	BASARD				TRF		
		Motif model	Location	Copy number	Structure	Consensus pattern	Location	Copy number
Asiatic black bear	AB069664		134–187	3	Type I	CGCCCCCGAGGCCGTTG	131–197	3.7
Polar bear	AY611807		89–142	3	Type I	CGCCCCCGAGGCCGTTG	86–152	3.7
Gray seal	DQ071548		105–217	6	Type II	CCGACCCCCGAGGCCATC	101–193	5.2
Common raccoon	AB069663		134–246	6	Type II	CCCCGAGGCCGTCGCGAC	120–216	5.6
European otter	DQ029098		116–192	4	Type II	CCCCCGAGGCCATCCAGA	116–167	2.8
Domestic cat	AB069665		134–187	3	Type I	CGCCCCCGACGCCGTCG	131–183	2.9
Sperm whale	AY615863		97–209	6	Type II	CCGCCCCCGACGCCATC	93–188	5.3
Sperm whale	AY615864		74–204	7	Type II	GCCCCCGACGCCACCCC	40–183	8.2
Sperm whale	AY615865		88–236	8	Type II	CCGCCCCCGACGCCATC	84–215	7.3
Minke whale	AY615866		56–258	11	Type II	CCGGCCCCGACGGCAGC	52–237	10.3
Minke whale	AY615867		56–348	16	Type II	CCGGCCCCGACGCCATC	52–327	15.3
Domestic cow	AB069666		134–264	7	Type II	CGCCCCCGCCCCGACGC	93–228	7.8
White beaked dolphin	AB069666		56–276	12	Type II	CCGGCCCCGACGCCATC	52–255	11.3
Harbor porpoise	AY615862		56–294	13	Type II	CCGGCCCCGACGCCATC	52–273	12.3
Domestic horse	AB080626		161–489	18	Type II	CCGCCCCCGACGCCACC	91–468	21.3
Domestic horse	AB080627		161–471	17	Type II	CCGCCCCCGACGCCACC	91–450	20.3
Wild horse	AB080628		161–489	18	Type II	CCGCCCCCGACGCCACC	91–468	21.3
Donkey	AB080629		161–471	17	Type II	CCGCCCCCGACGCCACC	91–450	20.3
Donkey	AB080630		161–381	12	Type II	CCGCCCCCGACGCCATC	91–360	15.3
Onager	AB080631		161–471	17	Type II	CCGCCCCCGACGCCACC	91–450	20.3
Onager	AB080632		161–435	15	Type II	CCGCCCCCGACGCCATC	91–414	18.3
Plains zebra	AB080633		161–399	13	Type II	CCGCCCCCGACGCCATC	91–378	16.3
Gravys zebra	AB080634		161–489	18	Type II	CCGCCCCCGACGCCACC	91–468	21.3
Mountain zebra	AB080635		161–489	18	Type II	CCGCCCCCGACGCCACC	91–468	21.3

the bear sequence and 3 repeat units for otter. We also found the copy number of amino acid sequences was usually one less than the copy number of DNA sequences, especially for those which had repeat segment structure Type II in the corresponding nucleotide sequence. The reason is that the shared extra repeat unit had a 5 bp gap, not a number divisible by length of codons 3. It was not surprising that the copy number of two sperm whale, AY615863 and AY615865, was only 1 after examining their nucleotide and amino acid sequences, because there is an open reading frame shift occurring during the translation process.

5 CONCLUSION

In this article, we expand the views of tandem repeats by introducing short adjacent repeats. In order to handle gaps between neighboring repeat units, we design a base- $(G + 1)$ numeral vector data structure. Also, we expand the scope of identifying short adjacent repeats to multiple DNA sequences, by relaxing the implicit assumption of a single DNA sequence in existing methods. This is helpful for analyzing the relationship among input sequences, e.g. DNA sequences of different species in the course of evolution.

To detect short adjacent repeats, we introduce a full probabilistic generative model to model repeats in multiple sequences. In this article, we only considered the case where each sequence contains only one repeat segment. If there are indeed more than one repeat segment in one sequence, BASARD is likely to report the most probable one. It is also possible that no repeat segment exists in some sequence. In this case, BASARD might report a repeat segment made up of only one repeat unit and its starting position might

vary randomly. In order to help exclude this kind of false positive, BASARD also reports the significance level of each estimated repeat unit, i.e. the P -value of such repeat unit under the background distribution. If the P -value is larger than a prespecified threshold, it is recommended to deny the corresponding estimated repeat units and conclude that the corresponding sequence actually does not contain the repeat pattern. Similar to the motif discovery problem in Liu *et al.* (1995), our algorithm can be extended to allow multiple repeat segments per sequence by introducing an indicator vector for each sequence, $\xi_n = [\xi_{n,1} \ \xi_{n,2} \ \dots \ \xi_{n,L_n}]$, $1 \leq n \leq N$. The binary variable $\xi_{n,l} = 1$ indicates that a repeat segment occurs from position l within the n -th sequence; otherwise $\xi_{n,l} = 0$.

We introduce a Bayesian approach to detect short adjacent repeats in a *de novo* fashion. To improve computing efficiency, we use a collapsing technique to reduce the dimension of the parameter space. After the MCMC chain converges, the sampled parameter values give us a whole picture of their joint posterior distribution. In the end, we demonstrate the effectiveness of BASARD through experiments on both synthetic data and real data.

The Bayesian model and its Metropolis-in-Gibbs sampling strategy for identifying short adjacent repeats in multiple sequences can find many applications, not only in areas of biology, but also in other fields. Our work takes the initial step to enable this repeat identification across multiple sequences and serves as a call for participation. Many interesting and important directions are worth exploring. For example, our work is limited by not allowing intraunit insertions and deletions. We might refer to motif discovery work by Gupta and Liu (2003) to deal with this kind of variations. Same as most existing purely likelihood-based methods, the length-bias

phenomenon (Hoh *et al.*, 2002) may affect BASARD's results. This problem is less serious for evolutionary study because the sequences are usually of similar length, but further studies may be needed when sequence lengths are substantially different from each other. Another question is how to further improve the convergence rate and to escape from local optima. Xu *et al.* (2010) used parallel computing techniques, such as evolutionary Monte Carlo method, to deal with this computing problem and achieved some encouraging results.

Funding: Research Grants Council of the Hong Kong SAR (Project no. CUHK 400709); CUHK direct grant (Project no. CUHK 2060362).

Conflict of Interest: none declared.

REFERENCES

- Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.
- Bryne, J.C. *et al.* (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res.*, **36**, D102–D106.
- Buchner, M. and Janjarsjitt, S. (2003) Detection and visualization of tandem repeats in DNA sequences. *IEEE Trans. Signal Process.*, **51**, 2280–2287.
- Butler, J.M. *et al.* (1997) STRBase: a short tandem repeat DNA internet-accessible database. In *Proceeding of the 8th International Symposium on Human Identification*. Promega, Scottsdale, Arizona, USA, pp. 38–47.
- Du, J. *et al.* (1997) Analysis of immunoglobulin Sgamma3 recombination breakpoints by PCR: implications for the mechanism of isotype switching. *Nucleic Acids Res.*, **25**, 3066–3073.
- Du, L. *et al.* (2007) OMWSA: detection of DNA repeats using moving window spectral analysis. *Bioinformatics*, **23**, 631–633.
- Gelman, A. *et al.* (2004) *Bayesian Data Analysis*. Chapman & Hall/CRC, New York, USA.
- Gupta, M. and Liu, J.S. (2003) Discovery of conserved sequence patterns using a stochastic dictionary model. *J. Am. Stat. Assoc.*, **98**, 55–66.
- Gupta, R. *et al.* (2007) A novel signal processing measure to identify exact and inexact tandem repeat patterns in DNA sequences. *EURASIP J. Bioinform. Syst. Biol.*, p. Article ID 43596.
- Hoh, J. *et al.* (2002) The p53MH algorithm and its application in detecting p53-responsive genes. *Proc. Natl Acad. Sci.*, **99**, 8467–8472.
- Krishnan, A. and Tang, F. (2004) Exhaustive whole-genome tandem repeats search. *Bioinformatics*, **20**, 2702–2710.
- Larsen, S.A. *et al.* (2005) Identification and characterization of tandem repeats in exon III of dopamine receptor D4 (DRD4) genes from different mammalian species. *DNA Cell Biol.*, **24**, 795–804.
- Lawrence, C.E. *et al.* (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*, **262**, 208–214.
- Li, Q. *et al.* (2010) Bayesian approach for identifying short adjacent repeats in multiple DNA sequences. In *Proceedings of the 2010 International Conference on Bioinformatics and Computational Biology (BIOCOMP'10)*, Vol. 1, Las Vegas, Nevada, USA, pp. 255–261.
- Liu, J.S. *et al.* (1995) Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. *J. Am. Stat. Assoc.*, **90**, 1156–1170.
- Lu, Q. *et al.* (1993) (CT)_n (GA)_n repeats and heat shock elements have distinct roles in chromatin structure and transcriptional activation of the *Drosophila* hsp26 gene. *Mol. Cell. Biol.*, **13**, 2802–2814.
- Mogensen, L. *et al.* (2006) Identification and characterization of a tandem repeat in exon III of the dopamine receptor D4 (DRD4) gene in cetaceans. *J. Heredity*, **97**, 279–284.
- Murphy, W.J. *et al.* (2001) Molecular phylogenetics and the origins of placental mammals. *Nature*, **409**, 614–618.
- Previc, F.H. (1999) Dopamine and the origins of human intelligence. *Brain Cognit.*, **41**, 299–350.
- Sagot, M.F. and Myers, E.W. (1998) Identifying satellites and periodic repetitions in biological sequences. *J. Comput. Biol.*, **5**, 539–553.
- Schneider, T.D. and Stephens, R. (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, **18**, 6097–6100.
- Schoots, O. and Tol, H.H.M.V. (2003) The human dopamine D4 receptor repeat sequences modulate expression. *Pharmacogenomics J.*, **3**, 343–348.
- Sharma, D. *et al.* (2004) Spectral Repeat Finder (SRF): identification of repetitive sequences using fourier transformation. *Bioinformatics*, **20**, 1405–1412.
- Sinden, R.R. (1999) Biological implications of the DNA structures associated with disease-causing triplet repeats. *Am. J. Hum. Genet.*, **64**, 346–353.
- Siyanova, E.Y. and Mirkin, S.M. (2001) Expansion of trinucleotide repeats. *Mol. Biol.*, **35**, 168–182.
- Sokol, D. *et al.* (2007) Tandem repeats over the edit distance. *Bioinformatics*, **23**, e30–e35.
- Sutherland, G.R. and Richards, R.I. (1995) Simple tandem DNA repeats and human genetic disease. *Proc. Natl Acad. Sci.*, **92**, 3636–3641.
- Thompson, W. *et al.* (2003) Gibbs recursive sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.
- Weber, J.L. and May, P.E. (1989) Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *Am. J. Hum. Genet.*, **44**, 388–396.
- Xu, J. *et al.* (2010) An evolutionary Monte Carlo algorithm for identifying short adjacent repeats in multiple sequences. In *Proceeding of the 2010 International Conference on Bioinformatics and Biomedicine (BIBM'10)*. Hong Kong, pp. 643–648.
- Zhou, H. *et al.* (2009) Detection of tandem repeats in DNA sequences based on parametric spectral estimation. *IEEE Trans. Informat. Technol. Biomed.*, **13**, 747–755.