IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON DATA-ENABLED INTELLIGENCE FOR DIGITAL HEALTH

Received September 15, 2019, accepted October 1, 2019, date of publication October 17, 2019, date of current version October 30, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2947261

DNA Motif Finding Method Without Protection Can Leak User Privacy

XIANG WU[1,2], HUANHUAN WANG[1,2], MINYU SHI[1,2], AMING WANG[1,2], AND KAIJIAN XIA[3]

[1]Institute of Medical informatics, Xuzhou Medical University, Xuzhou 221000, China
[2]Institute of Medical Information and Health Big Data, Xuzhou Medical University, Xuzhou 221000, China
[3]Changshu Affiliated Hospital, Soochow University (Changshu No.1 People's Hospital), Changshu 215500, China

Corresponding author: Xiang Wu (wuxiang@xzhmu.edu.cn)

This work was supported in part by the Natural Science Fund for colleges and universities of Jiangsu Province under Grant 18KJB520049, in part by the industry University-Research-Cooperation Project in Jiangsu Province under Grant BY2018124, and in part by the National Scientific Data Sharing Platform for Population and Health.

ABSTRACT DNA sequence analysis plays an important role in the study of gene regulatory networks. DNA motif finding has become a key discipline in the post-gene era and gradually become a research hotspot by mining key gene sequences corresponding to disease mechanism and important biological functions. However, the research of DNA motif finding is faced with a huge problem of privacy disclosure. DNA motif finding technology cannot manage and use data well under controllable conditions, and the mining process of DNA motif finding itself is prone to reveal private information such as individual traits, characteristics and disease defects. In this paper, we presented an overview of the privacy breaching of DNA motif finding, summarized the main methods and tools of the current DNA motif finding, analyzed its privacy risks, and used two case studies to verify that the DNA motif finding may identify individual privacy information. Finally, we discussed the privacy protection methods for motif finding and proposed the privacy protection solutions.

INDEX TERMS DNA motif finding, privacy disclosure, DNA sequences, privacy protection.

I. INTRODUCTION

A. MOTIVATION

With the completion of the Human Genome Project(HGP) and the rapid development of modern biotechnology, a large number of biological DNA sequence data have emerged [1], [2]. How to effectively analyze and sort out the massive biological sequence data to obtain useful information and knowledge has become a very challenging issue in the field of bioinformatics.

DNA motif finding is an important but unresolved issue in the analysis of DNA sequences to interpret the genome. As a subsequence that coexists in multiple gene data sequences, DNA motif is a repetitive short sequence fragment that is likely to correspond to the core function of an organism [3]–[7]. The identification of key motifs is the core step to understand the mechanism of disease through genetic data, and the conservative motifs obtained by comparison help humans to decode genomic data and find the corresponding relationship between gene loci and physiological functions. It can be said that the efficient and accurate mining and identification of the motif finding provides an opportunity to explore the genetic variation in vivo and organisms. And a large number of genome data mining, comparison and analysis create conditions for a better understanding of the regulatory mechanism of gene expression [8], [9].

However, an individual's private information is easily leaked in the process of discovering motifs because DNA sequences contain a large amount of private information about personal characteristics, functions, illnesses, and personality disorders [10]. In recent years, DNA datasets have caused a serious problem about privacy disclosures, the privacy protection research of motif finding has aroused widespread societal concern. In 2008, Homer *et al.* [11] proved that a person's specific identity could be identified from a set of DNA data. After that, Gymrek *et al.* [12] showed that it was possible to re-identify 50 DNA participants from the 1000 genomes project dataset, which eliminated clear identifying information. The privacy disclosure problem like this happen all the time [13], [14]. There are two main

The associate editor coordinating the review of this manuscript and approving it for publication was Yongtao Hao.

reasons for privacy disclosure in these scenarios. One of the reasons is that existing privacy protection technologies can not effectively protect the privacy of genetic data [15]. The particularity of genomic data determines that it is not possible to be anonymous only by deleting identification information, while the attackers can often re-identify personal information from the publishing DNA mining data using the background information. For example, the information related to Y chromosome genome variation can be obtained from public genetic databases, and human genome project participants can be identified from public population statistics. Another reason for the re-identification attack is the attack on the machine learning model [16]. Most attacks on genetic data use the entire dataset, and even machine learning models trained on genomic data still can reveal the information of data providers, or people with some background knowledge can also access data to provide personal information. The biggest problem in the motif finding mining process is that its mining process is prone to privacy disclosure.

Although relevant gene sequencing platforms have emphasized the importance of privacy protection and taken measures in data processing, there are still many risks such as reidentification, phenotypic judgment and kinship construction. With the decrease of the cost of gene sequencing and the exponential growth of gene data, the privacy leakage problem of motif finding based on gene data is becoming more and more serious [17]–[20].

The problem of privacy security in motif finding needs to be solved urgently, which involves privacy disclosure from data collection to identification. In the stage of data collection, the sampling of genetic data is often complicated, and the process of physical and chemical processing and digitization itself is difficult to carry out confidentiality processing [21], [22]. In the process of data analysis and sharing, the attackers re-identify the individual through the obtained genetic data, which poses an involuntary privacy disclosure threat to gene providers. If there is an unreliable third party in this process, the data privacy will be difficult to be guaranteed [23]–[25]. In addition, there are hidden dangers in the storage and processing environment of gene data needed for motif finding, and the confidentiality of the database and the distributed storage mode of the cloud environment may be difficult to adapt to the actual requirements due to the particularity of the genetic data [26], [27].

At present, neither the algorithms nor the existing platforms have carried out strong protection of privacy. With the genetic data of exponential growth, there may be privacy disclosure problems such as phenotypic information and kinship in the process of data collection, analysis and publishing [28], [29]. This paper mainly summarized the research related to motif finding, from motif finding methods to motif finding tools and platforms, deeply analyzed the possibility of motif finding privacy disclosure, and provided solutions for the possibility of disclosure.

B. CONTRIBUTIONS

The main contributions of this paper are as follows.

(1) Review the current application areas and future prospects, importance and biological significance of motif finding;

(2) Summarize the current motif finding algorithms and tools, and evaluate the current algorithms and tools for motif finding;

(3) A review of the types of motif finding in which there is a risk of privacy leakage, how to disclose it, and the types of genetic data privacy attacks;

(4) Confirm the feasibility of the attack and the necessity of privacy protection through case analysis experiments.

II. PRELIMINARIES

A. MOTIF FINDING

DNA motif finding is defined as recognizing short sequence motifs with high frequency in a biological DNA sequence and repeated subsequences in multiple biological sequences, or a combination of the above two cases. Finally, detect the expressed DNA sequence motifs and mark each occurrence instance in the DNA sequence or protein [30]–[32]. The motif finding process is shown in figure1.

The motif finding algorithm usually does not directly locate the motif instance in the input sequence. Firstly, search the overexpressed motifs in the sequence, then use the scanning sequence of motifs to locate the site indirectly. It should be considered that how to represent the DNA motifs when designed the motif finding algorithm.

In general, multiple sequences of motifs are bound together with the same regulatory protein, and multiple sequences are paired together to form a base matrix, which together reflects the characteristics of the motifs. The motif instance represents one of the sequences, which is a row in the base matrix. However, multiple sequenced motifs contain information that is difficult to reflect, so motif finding algorithms usually transform them into other representations to highlight the information contained in the motifs. The DNA motif finding has three main representation methods.

1) THE CONSENSUS SEQUENCE METHOD [33]

It is a simple abstract description of the motif, but the consensus sequence is not necessary in the biological DNA sequence. DNA motifs appear in biological DNA sequences as instances of a consistent sequence, a new sequence in which the consensus sequence is mutated at certain base positions.

The consensus sequence model directly uses the string consisting of {A, C, G, T} to represent the motifs, this is a relatively simple representation. After the motif instances are aligned, the characters with the highest probability in each column are directly connected to form a sub-sequence as the representation of the DNA motif. The calculation process is as follows.

The consensus sequence model is simple but contains less information. Information such as the location of a specific

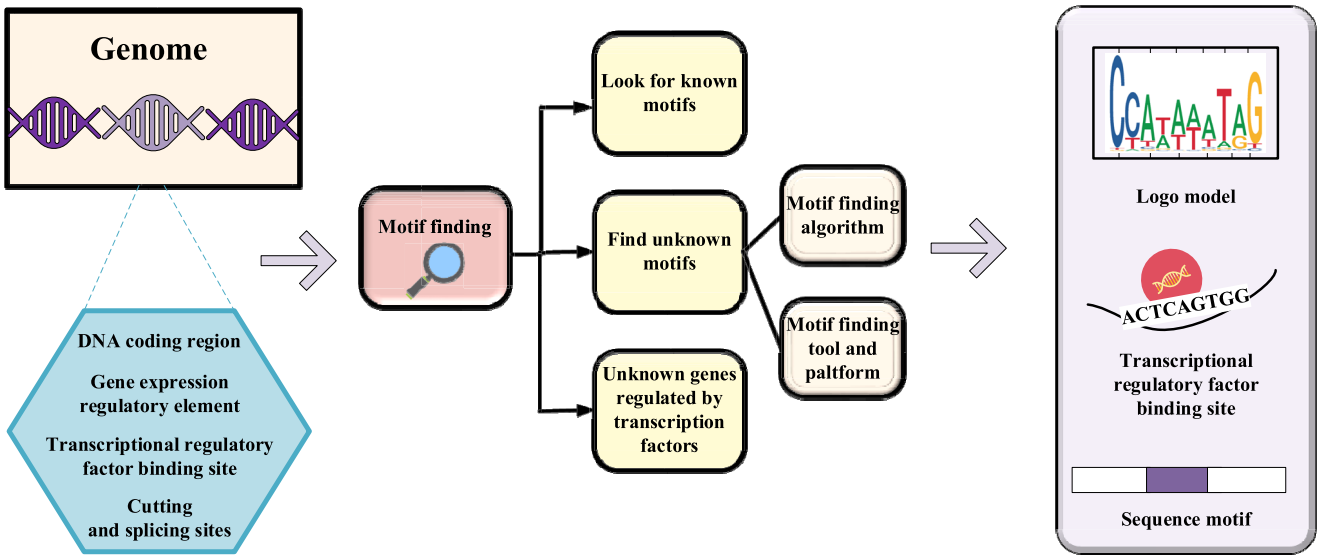


FIGURE 1. The process of motif finding.

motif instance cannot be obtained from a consensus sequence, which only can be used to measure its degree of conservation. Moreover, the model does not consider vacancy and base correlation. The information of weak base in some positions of the motifs will be ignored when constructing the consensus sequence, which increases the difficulty of further evaluation of the motifs.

2) POSITION WEIGHT MATRIX REPRESENTATION [34]

The position weight matrix is in the form of an N row 4 column matrix in which the frequency of each base in {A, C, G, T} is described in the probability matrix of N row 4 columns. According to the different measurement methods of frequency information, there are three kinds of matrix representation methods: Alignment Matrix, Frequency Matrix [35], [36] and Weight Matrix [37], [38]. Among them, the Alignment matrix is the frequency of the occurrence of the corresponding position base, the Frequency matrix is the frequency of the corresponding position base, and the Alignment matrix with the Frequency matrix are the prototype of the position weight matrix. Table 1 shows the three matrix representation methods of a motif.

As you can see from Table 1, matrix representations contains more information than known sequential representations. The method of using the position matrix to represent the motifs can not only show the change at a specific position, but also reflect the probability of different characters at that position and indicate the degree of change.

3) SEQUENTIAL LOGO MODEL [39]

The amount of information in each column of a motif is represented by a color graph, as shown in figure 3, the different color letters are used to represent a sequence fragment, and the height of the letter represents the size of the information content at a certain position on the base, and the information content at a certain position can reflect the conservatism of the motif at the site. All logo models can intuitively show the degree of conservation of the motif and which base play a relatively important role in which positions.

The above various motif representation methods reflect the overexpressed bases and their frequencies of DNA motifs

TABLE 1. Matrix representation of motif.

Alignment matrix				
pos	A	C	G	T
1	0	17	2	1
2	0	6	14	0
3	14	0	6	0
4	6	0	0	14
5	0	14	0	6
6	0	20	0	0
Frequency matrix				
pos	A	C	G	T
1	0.05	0.73	0.13	0.09
2	0.05	0.29	0.61	0.05
3	0.61	0.05	0.29	0.05
4	0.29	0.05	0.05	0.61
5	0.05	0.61	0.05	0.29
6	0.05	0.85	0.05	0.05
Weight matrix				
pos	A	C	G	T
1	-1.25	0.47	-0.36	-0.70
2	-1.25	-0.45	1.17	-1.28
3	1.24	-2.21	0.43	-1.28
4	0.50	-2.21	-1.32	1.21
5	-1.25	0.29	-1.32	0.47
6	-1.25	0.62	-1.32	-1.28

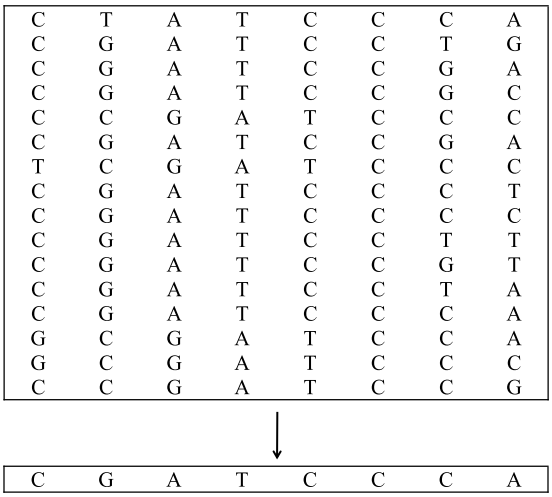


FIGURE 2. The consensus sequence method.

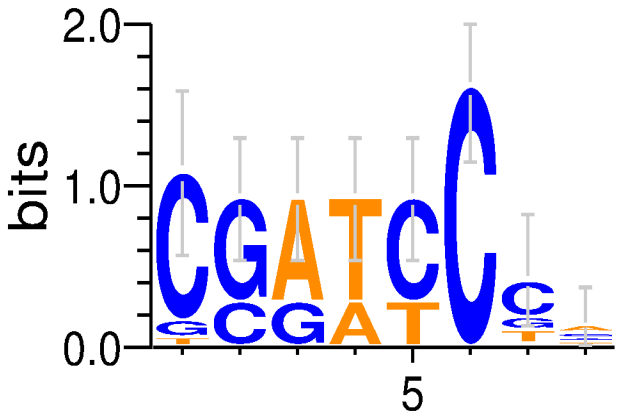


FIGURE 3. Sequence logo model schematic diagram.

in different ways, and give a conservative motif of sites, which facilitate further searching for motif instances in the sequence.

B. SOURCE OF MOTIF FINDING AND RESEARCH DIRECTION

With the rapid development of sequencing technology and the explosive growth of genetic data, more and more shared databases provide a source of genetic data for motif finding. Currently, there are three main types of database storage: public databases, commercial databases, and national databases.

The public databases mainly include 100000 Genomes Project (UK),[1] Personal Genome Project,[2] 1000 Genomes Project,[3] the Cancer Genome Atlas (TCGA).[4] Commercial databases are mainly represented by 23 and Me[5] and Ancestry.[6] National databases have National DNA Index(CODIS)[7] and Interpol[8] et al [40], [41]. Other biological databases include Scansite,[9] Eukaryotic Linear Motif, ELM,[10] Minimotif Miner,[11] Tohoku Medical Megabank.[12]

The research of DNA motif finding is mainly divided into three directions: First, looking for known motif in a given genome sequence. Second, to find an unknown motif in the upstream region of a series of co-expressed or co-regulated genes, that is, new motifs were discovered in the promoter regions of a series of co-expressed genes, and motif was identified by analyzing and extracting DNA sequence features. Third, search for unknown genes regulated by a known transcription factor [42].

III. TOOLS AND ALGORITHM ANALYSIS OF DNA MOTIF FINDING

Generally, the motif characteristics of prokaryotes are obvious and easy to identify. However, the motif of eukaryotes is relatively complex, the length and spatial distribution of motif vary greatly, there is no fixed position, and the binding sites of the same protein factors are also different, which makes it very difficult to identify motif. Therefore, it is almost impossible to design a method that recognizes all motifs, and there are many algorithms and software for different creatures and features of DNA motifs.

There are two types of algorithms for DNA motifs finding. One is the exact algorithms, which is based on strings. For example, Wordup [43], YMF [44], [45], QuickScore [46], MOPAC [47], Consensus [48], Weeder [49], Mitra [50], WINNOWER [51] and so on [52], [53]. And the other is the inexact algorithms, which is based on probability, clustering or heuristic search, including MEME [54], [55], Gibbs Sampling [56], [57], MCL-WMR [58] etc.

A. ALGORITHM ANALYSIS

This section mainly introduces four algorithms, including Weeder (exact algorithm), EXTREME (inexact algorithm), Gibbs Sampler (inexact algorithm) and BIOPROSPECTOR (inexact algorithm).

1) WEEDER

Weeder is based on a greedy algorithm with relatively high time complexity and suitable for finding short sequence DNA motifs.

The suffix tree provides an effective solution to motif finding. To trim the search space, Weeder use suffix-trees

[1] https://www.genomicsengland.co.uk/about-genomics-england/the-100000-genomes-project/
[2] https://my.pgp-hms.org/
[3] http://grch37.ensembl.org/index.html
[4] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[5] https://www.23andme.com/?myg07=true
[6] https://www.ancestry.com/
[7] https://www.fbi.gov/services/laboratory/biometric-analysis/codis
[8] https://www.interpol.int/
[9] http://scansite.mit.edu
[10] http://elm.eu.org
[11] http://sms.engr.uconn.edu
[12] https://www.thermofisher. com/blog/bio-banking/

to hold data and enforce constraints on locations that allow mismatches.

(1) Given a set of sequences $\alpha = \{A, C, G, T\}$, find the motifs with length L that appear in at least k sequences, and no more than n mutations.

(2) Suppose the endpoints of the sequence corresponding to the motif $M = M_1 \ldots M_l$ in the suffix tree are found, in all paths, the word spelling distance M is less than n, and $l < L$. Associate each path with the distance between m of the corresponding string.

(3) If M is valid, we will try to extend it to one symbol.

(4) For each character $a \in \{A, C, G, T\}$, we match a with the next symbol on each path. If a path ends before the node p of the tree, match the character a with the first character on each side of the leaving node p. If it does not match, add previous error1 along the path. Otherwise, the error will remain unchanged. If the new error is greater than n, the path will be discarded.

(5) After checking all paths, the surviving path represents the approximate event $M^{'} = M_1 \ldots M_l$, if the number of times it appears more than k and the length is less than L, expand it. Otherwise continue to use M and the next character of the character set.

Weeder starts with an empty motif at the root of the tree and then recursively expands it. It exhaustively enumerates all short sequences until the required length is reached. The limitation on the length of the motif and the number of mutations allow Weeder to find a motif of any length that satisfies the condition.

2) EXTREME

EXTREME has many similarities with the Expectation Maximization algorithm(EM), and its core is online EM. The online EM algorithm updates the parameters by iteration between the E and M steps, which computes only one observation data set, not the entire data set.

(1) If it does not converge at the end of the transfer, the index A is updated to the midpoint between the current value and a certain value of $A^{'}$, and EXTREME performs another pass in the data set. EXTREME repeats these steps until the required threshold is reached.

(2) EXTREME's seed strategy uses a search-based motif finding algorithm to find motifs that initialize the online EM algorithm. The seed algorithm counts the number of words that appear in the positive and negative sequence sets, and associates the "$S - score$" with each word.$S - score$ is given by:

$$S = \frac{f - f^{'}}{\sqrt{f^{'}}} \tag{1}$$

where f and $f^{'}$ respectively are the number of times a word appears in a positive sequence and a negative sequence. If $f^{'}$ is 0 for a word, modify it to 1 to prevent division by 0.

(3) Each word contains j wildcard letters. Word clustering is converted to a frequency count matrix by counting the number of times each letter appears at each of the aligned positions. The count is normalized by the $S - score$ of each word in the cluster, so that more meaningful words will contribute more to the counting matrix than unimportant words. This seeding strategy can be parallelized, which allow multiple primitives to be discovered simultaneously. Then, hierarchical clustering of discovered motifs can identify a single motif class.

(4) Normally, the online EM algorithm converges after 1∼5 times of calculation, so the time complexity is proportional to the width of the template and the size of the data set. In practice, EXTREME as a whole has a time complexity that is linear with the size of the data set.

A search-based seed strategy combined with an online EM algorithm is effective for new motif finding that is effective in big data sets.

3) GIBBS SAMPLER

The basic principle of motif finding by Gibbs sampler is that the motif finding model and the occurrence position in each sequence are continuously updated by randomly sampling to optimize the objective function, when a certain iteration termination condition is satisfied, the final candidate motif is obtained.

(1) Build the probability models of motif finding and background are established respectively. The motif finding model is represented by a position frequency matrix (PSFM). First, the starting sites of k motifs in each sequence are randomly generated as $B = \{b_i\}$, $i = 1, 2, ..., k$. Candidate motifs are derived from the starting site and the defined length L, and a PSFM is established from these candidate motifs. The background model usually uses an independence model $M_d = \{f_A, f_G, f_C, f_T\}$,which indicates the frequency of occurrence of each base in the background sequence.

(2) Choose a sequence B_i in the input sequence set, and the length is $l_i (i = 1, 2, ..., k)$. Delete the start site belonging to the sequence in B, and recalculate the value modified PSFM. Then calculate all possible candidate motifs according to the motif and background model, namely the scores $Score_m$ and $Score_d$ of $B_i[j, j + L - 1] (j = 1, 2, ..., l_i - L + 1)$, that is, the possibility that the sequence fragment from the jth position to the $j + L - 1$th position in the sequence B_i is a motif and the possibility of being a background sequence is calculated.;

(3) Calculate the ratio R of the two scores, and select a new candidate motif for each j, that is, select a candidate motif with a higher ratio with a larger probability, and add its starting site to B. Candidate motifs are obtained based on the new start site set and motif length, and the score F is calculated by:

$$F = \sum_{i=1}^{L} \sum_{k=1}^{l} p_{ik} \log(\frac{p_{ik}}{q_{ak}}) \tag{2}$$

where $p_{ik} = PSFM(i, a_k)$, and a_k is the k-letter of $A = \{a_1, a_2, ..., a_L\}$. If the score F is greater than the previous score, go to step 2 and continue iteration. otherwise repeat step 3 until the convergence or the number of repetitions reaches the maximum number of iterations. If all sequences are processed, the process ends.

If the score is not improved after many times or the maximum number of iterations is reached, the calculation process is terminated. The Gibbs sampler circulates the variables in turn according to the above steps, and finally obtains a stationary distribution.

4) BIOPROSPECTOR

The threshold sampler is executed many times when the BioProspector is running with Gibbs Sampler algorithm. Initialize the motif probability matrix P by random comparison of the input sequences, iterate and randomize the matrix.

(1) BioProspector can use two background file formats: One is a calculated sequence file of independent background model B. The other is pre-computed document describing the background probability of an organism's complete genome, which can use the third-order markov background model to evaluate the probability of generating segment S from independent backgrounds.

(2) Sample the new alignment with two fractional thresholds to deal with the problem that some input sequences contain no copies of the motif and some contain many copies. Introducing two thresholds θ_a, θ_b, the scores in the segment between $[\theta_a, \theta_b]$ can help the program converge more quickly.

(3) For a motif with two blocks, BioProspector uses two probability matrices a_1, a_2 to obtain the two blocks. When these two blocks are palindromic, only one motif probability matrix is needed. Each permutation sequence contributes two fragments to the same matrix.

(4) Motif score distribution is used to measure motif goodness. For each generated sequence set, perform some threshold sampler runs to record the highest motif score. Then, fit the normal distribution to the scores of the program-generated set of identically distributed sequence sets. BioProspector runs the original sequence through a threshold sampler with fractional distribution, and reports motifs that are higher than the standard deviation of the mean of the motif score distribution.

There are two kinds of probability-based algorithms in inexact algorithms. The first is the schema discovery algorithm based on expected maximization of EM algorithm. The second is Gibbs sampling motif finding algorithm. The algorithms based on clustering to solve the problem of motif discovery are as follows: MCL-WMR algorithm [58] and iMCL-WMR algorithm [59]. The improved algorithm not only improves the time efficiency of dealing with large data sets, but also accurately identifies the motifs.

After investigation and analysis, it was found the exact algorithm and the inexact algorithm are two branches of the motif finding algorithm. Accurate algorithms are generally based on exhaustion, while inexact algorithms are based on probabilistic, clustering or heuristic search algorithms. In general, the exact algorithm is often inefficient because it has to be used throughout the entire space, and it consumes a lot of time when running on a large data set, but it can guarantee that the result is a global optimal solution. The inexact algorithm is suitable for searching solutions in large spaces with high efficiency, but easy to fall into local optimum.

In the exact algorithms, Weeder comprehensive evaluation performance is better in many algorithms, but its space complexity is higher, and it only use a single motif as the output result of each operation, which does not meet the actual demand. Yeast Motif Finder (YMF) is an exhaustive algorithm that is generally more practical for short DNA sequences, while for longer sequences, the time consumption often increases exponentially by the size of the calculation increasing.

In the inexact algorithms, the EM algorithm is simple and easy to implement. EXTREME algorithm mentioned in this paper as an improved algorithm of the EM algorithm is very useful for quickly finding multiple motifs in a large data set. However, EM algorithm is slow in calculation, itself or its improved algorithms usually only obtain a local optimal solution, because its essence is an iterative algorithm. Gibbs Sampler is fast, but the accuracy is not high. MEME is a probabilistic statistical method based on the expectation maximization algorithm without a priori knowledge, and use the subsequences appearing in the DNA sequence as the starting point of the algorithm to increase the probability that the algorithm finds the global optimal solution. However, The biggest limitation of this algorithm is the lack of versatility, high computational complexity and long time.

B. DNA MOTIF FINDING TOOLS AND PLATFORMS

There are many kinds of software for motif finding, which have different performance in searching for motifs, and the content of motif finding is not limited to human sequences. The literature [60] compared thirteen motif finding software, including: AlignAEC, ANN-Spec, Consensus, GLAM, The Improbizer, MEME, MITRA, MotifSampler, Oligo/dyad-analysis, QuickScore, SeSiMCMC, Weeder, YMF. In addition, CompleteMOTIFs [61] is a DNA motif finding platform for transcription factor binding experiments. Tmod is the platform that integrate twelve motif finding algorithms.[13] MODSIDE[14] is a motif discovery pipeline and similarity detector. TrawlerWeb[15] is an online motif finding platform.

In fact, the biologists tend to use multiple motif finding tools at the same time in practical applications, and take the first few sequences rather than the optimal sequence given by the software during the experiment. In addition, it is important to select the most appropriate statistics to evaluate the accuracy of the prediction.

[13] http://www.fas.harvard.edu/~junliu/Tmod/
[14] http://modside.org/
[15] http://trawler.erc.monash.edu.a

TABLE 2. Simulation of original research results (100 samples).

Motif	AAAAA	TTTTT	AAAAT	CCTCC	TAAAA	AAATA	ATAAA	AATAA	ATTTT	GAAAA
Counts	689	556	443	393	372	354	349	346	334	323
Motif	CCCAG	AAATT	GGGAG	CCCCC	AAAAG	CTCCC	TATTT	CTGGG	TTTTC	CAGCC
Counts	323	316	316	312	311	311	311	305	295	295

IV. THE ANALYSIS OF PRIVACY LEAKAGE OF MOTIF FINDING

Genetic data is unique and static, and it contains the most important and sensitive personal information. We can obtain the individual's health signs and disease susceptibility by analyzing. The privacy disclosure of motif finding mainly comes from using stage [62], analysis and mining stage [63], storage stage [64], and publishing stage [65]. Genetic data has been proved to be at risk of privacy disclosure in the using, storage and publishing stage. However, whether the privacy information will be leaked in the analysis mining stage (motif finding process) has not been fully studied. Through two case studies, this paper confirmed that the mining process of motif finding is very likely to disclose personal privacy.

Currently, the main ways of privacy disclosure of genetic data are re-identification attack [12], [66], phenotypic inference attack [67], [68] and other attacks [69], [70]. Among them, re-identification attack is the main way of privacy disclosure in the process of motif finding. The risk of re-identification is the most widely studied privacy risk in the transmission and analysis of the human genome, and under this type of attack, unauthorized parties view human genome data that has been published under some kind of protection, and try to re-identify and recover the identity information from the data that has hidden the identity information. Once such an attack is successful, it will cause trouble for the data provider.

The results of gene analysis are likely to contain sensitive information about individual disease susceptibility, phenotypic characteristics and even biological longevity. Most of the output by motif finding algorithm results are the first n term DNA short sequence frequent motif and its occurrence frequency. To a certain extent, this statistical result is de-identified and processed, that is to say, we can not know the individual information only from the result. But if somebody has a certain background knowledge and enough query times, through the statistical results affected by individual data, the motif finding mining process will face a serious risk of privacy disclosure. We propose two attack scenarios to verify that the mining process of motif finding will reveal privacy.

A. CASE STUDY 1

Single nucleotide polymorphism(SNP) is the third generation genetic marker, which is the variation of single nucleotides on the genome. Many phenotypic differences in human body and the susceptibility to drugs or diseases are closely related to SNP. In order to ensure the credibility of sensitive motifs and the realizability of privacy attacks, we used some SNP resources obtained from Ensembl database as the data support of our simulation experiment.

A number of studies [71], [72] have found that many diseases are associated with SNP. It is assumed that a study conducted DNA sequence analysis related to a disease, and the results of motif finding based on the research data is R_b, the collection of motifs are $Motif_m$. When an attacker knows by some means of the addition or deletion of an individual, the result can be queried, and the attacker obtains the result R_a. The new query results show that the set of motifs and their count results also change, if the new results are compared with the results before the addition and deletion, the motif finding results of the change sequence can be obtained, that is, the combination of the motif set $Motif_{n-m}$ and the change count.

Step1: We use the SNP slice sequence data from Ensembl database as the source data to simulate the experimental results. N-gram motif finding algorithm is used in the experiment. The length of the motif is 5 and the hamming distance is 1. The statistical results of motif finding are obtained from 100 simulated data of the source data, including the set of motifs and motif frequency set as shown in table 2.

Step2: We add a SNP simulation sequence from the same site of homologous population to the original dataset, and then identify the motifs again. This process is similar to an attack process that an attacker has known that a piece of data is added and then query the result again. The result of motif finding R_a is obtained, including the set of motifs $Motif_n$ and motif frequency set C_a as shown in table 3.

Step3: Compare R_b with R_a, we obtain the motif finding result of the individual by $motif_n - motif_m$,,$C_a - C_b$ and the motif finding results contain the motif corresponding to the sensitive SNP fragments involved in the original study as shown in table 4.

If the motif 'AAAAA' is the key SNP site in the study, and it has been confirmed to be related to a disease by subsequent studies, the attacker obtains the disease-related SNP, which links the individual to the disease susceptibility and health status to a great extent. Through the re-identification of the individual, the privacy attack is successfully achieved by

TABLE 3. Simulates the result of an increase in SNP sequence (101 samples).

Motif	AAAAA	TTTTT	AAAAT	CCTCC	TAAAA	AAATA	ATAAA	AATAA	ATTTT	GAAAA
Counts	701	561	447	395	377	356	351	348	339	327
Motif	CCCAG	AAATT	GGGAG	AAAAG	CCCCC	TATTT	CTCCC	CTGGG	TTTTC	CAGCC
Counts	323	320	316	312	313	311	311	305	295	296

TABLE 4. Recognition results of individual motifs from simulated data.

Motif	AAAAA	TTTTT	AAAAT	CCTCC	TAAAA	AAATA	ATAAA	AATAA	ATTTT	GAAAA
Counts	12	5	5	5	5	5	2	2	5	4
Motif	CCCAG	AAATT	GGGAG	CCCCC	AAAAG	CTCCC	TATTT	CTGGG	TTTTC	GATGA
Counts	0	4	0	1	1	0	0	0	4	1

TABLE 5. Overall sample motif finding results (100 samples).

Motif	AAAAAA	TTTTTT	AATAAA	CCTCCC	AAAAAT	AAATAA	AAAATA	ATAAAA
Counts	343	234	179	175	167	160	140	138
Motif	TAAAAA	ATTTTT	TTAAAA	...	TACGGC	CGTACG *	GTACGA	CGTACC
Counts	136	131	128	...	1	1	1	1

TABLE 6. Simulate a patient population motif finding results (100 samples).

Motif	AAAAAA	TTTTTT	AATAAA	CCTCCC	AAAAAT	AAATAA	AAAATA	ATAAAA
Counts	343	234	179	175	167	160	140	138
Motif	TAAAAA	ATTTTT	TTAAAA	...	TACGGC	CGTACG *	GTACGA	CGTACC
Counts	136	131	128	...	1	1	1	1

obtaining the disease susceptibility information through the published results of the motif finding process

B. CASE STUDY 2

For this attack scenario, assuming that we have a wide range of motif finding data, such as a collection of data A from all disease patients at a biological data analysis organization and the data sets of patients with each disease $a, b, c, ...$, and the motif finding results. In addition, we learn about an individual I through background knowledge is in the dataset A, and we have the results of the motif analysis of someone. From the results of motif finding, if it is found that some specific motifs of individuals will also appear in the results of group motif finding of a, then we may think that individual I comes from a disease population to a large extent.

We verify the realizability of a background knowledge attack based on statistical query results.

Step1: To build a large sample. We use the SNP data sequence downloaded from the Ensembl database as the source data to randomly generate 100 simulated sequences as the large sample data set. Then used n-gram motif finding algorithm to identify the motifs, the length of the motif is 6, the hamming distance is 1, and the results are shown in Table 5.

Step2: Using 30 pieces of data to simulate a small set a with a certain diseased, the same method as step1 was used to obtain the motif finding results as shown in table 6.

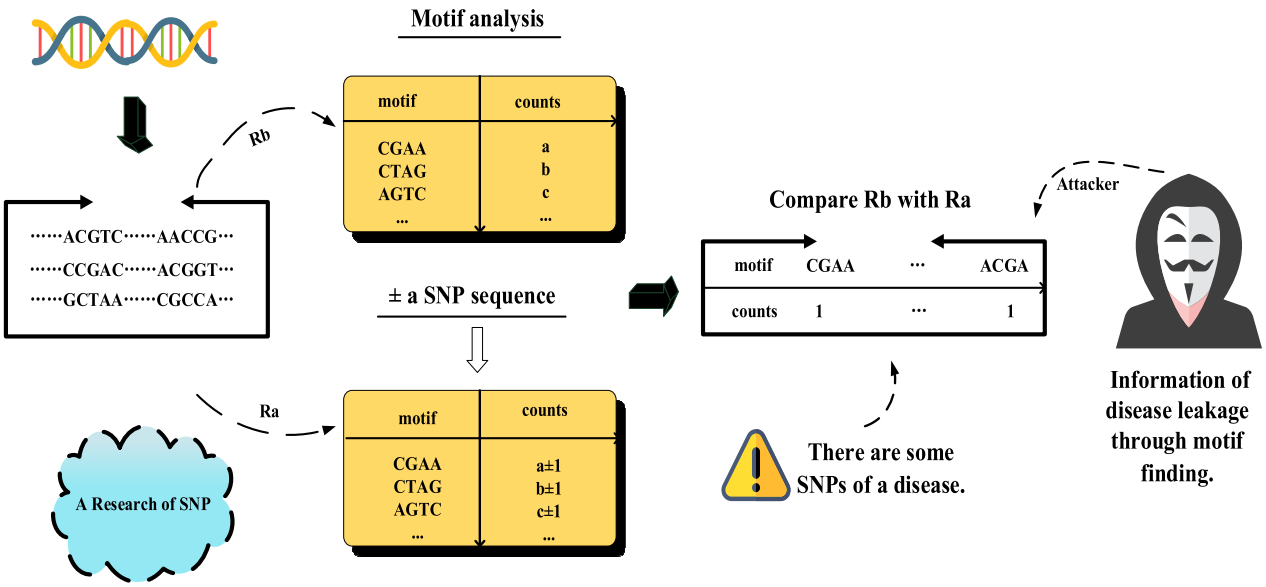


FIGURE 4. The scenario attack of motif finding (a).

TABLE 7. The known motif finding results of an individual.

Motif	AAAAAA	GAGATG	AGAAAA	AAAGAA	TCGTAT	...	ACGTAG	GTACGA *
Counts	8	3	3	3	3	…	1	1

Step3: The motif finding of a single sequence data is used as the motif finding result of an individual, and then the three sets of motif finding results are compared and analyzed. Firstly, find out the specific motifs in the individual, that is to say, the only motifs that the person has in the whole large sample. Then search for this motif in the recognition results of population *a* motif, compared with table 5, 6, 7, it can be seen that the motif 'GTACGA' (marked with *) was only found once in the entire sample set, and in the population *a* and the individual also appeared once.

Step4: As shown in the tables, the motif frequency of 'GTACGA' appears equally in a large sample and in an individual, and the same frequency occurs in *a* group, then it can be proved that the individual is located in *a* group. In other words, if the frequency of a single motif in a group is less than the number of times in an individual, it can be considered that the individual is not in this group.

The case studies show that motif finding mining process is prone to reveal private information.

V. DISCUSSION

The motif finding process is the key to correctly understanding the process of gene expression regulation. Motif finding is not a simple motif repetition - there are nucleotide mutations, insertions and deletions in DNA sequences, and it is not practical to identify incomplete conserved motif instances in sequences based on fixed motif, which makes it challenging to accurately identify motifs outside of motif size. The more serious problem is that the DNA motif finding mining process has potential privacy leakage risk. The case studies in this paper proves the realistic feasibility of the attack and provided the necessary theoretical basis for the next step to identify the privacy protection of the motif. How to perform motif finding efficiently, safely and accurately, while taking into account the availability and security of data, becomes the next problem to be solved. For the leakage of the motif finding algorithm, traditional data mining methods with privacy protection always have some defects, which make the processed data less available. The currently more suitable method is differential privacy technology [73], [74].

Differential privacy [75] is based on a rigorous mathematical model that appends moderate noise to a dataset to mask sensitive information. The main idea of differential privacy is to ensure that the presence or absence of a single record does not affect the outcome. Therefore, if all records have been mastered except for one, an adversary cannot decide whether this record exists in the dataset. In a word, differential privacy has effectively prevented sensitive information from leaking. Reference [73] proposed a high-utility motif finding algorithm based on ε-differential privacy, which is known as a rigorous definition of privacy with meaningful privacy guarantees in the presence of arbitrary external information. This paper is proved that differential privacy technology can guarantee the data availability and protect the privacy security of DNA data sharing. Therefore, we will further analyze DNA motif finding technology and use differential privacy technology to solve DNA motif finding problems in different scenarios.

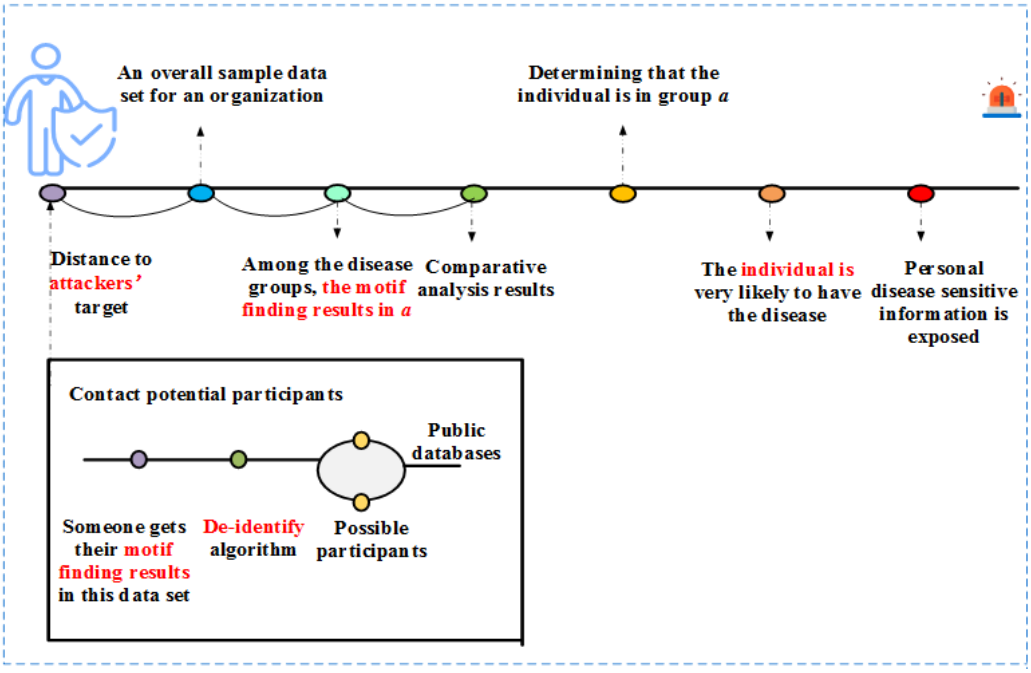


FIGURE 5. The scenario attack of motif Finding (b).

VI. CONCLUSION

Genetic data is a unique data type, which raises clear and inevitable interdependent privacy issues. The identified conserved motif helps humans decode genome data and find the relationship between gene locus and physiological function, which is a key step to understand the functional mechanism of organisms. The ability to uniquely identify an individual, predict health-related issues, and even learn about familial history, makes handle and share genetic data a unique challenge for security and medical experts alike. In this study, we summarized the current motif finding algorithms and tools, reviewed the types of privacy disclosure risks and attack methods in motif finding, and verified the feasibility of the attacks and the necessity of privacy protection through case studies. Finally, we gave the privacy protection strategy to solve the privacy disclosure problem of motif finding. Our current work is an important stepping stone to the development of motif finding research for the intention to share genetic data. Our next work will further confirm the privacy disclosure scenario of motif finding, study the privacy protection method, and continuously improve the model to improve the identification accuracy and the availability of results, so as to provide a biological research technology platform with reliability, security and authority.

REFERENCES

[1] J. Xing, Y. Zhang, K. Han, A. H. Salem, S. K. Sen, C. D. Huff, Q. Zhou, E. F. Kirkness, S. Levy, M. A. Batzer, and L. B. Jorde, "Mobile elements create structural variation: Analysis of a complete human genome," *Genome Res.*, vol. 19, no. 9, pp. 1516–1526, 2009.

[2] S. Choudhury, J. R. Fishman, M. L. Mcgowan, and E. T. Juengst, "Big data, open science and the brain: Lessons learned from genomics," *Frontiers Hum. Neurosci.*, vol. 8, p. 239, May 2014.

[3] N. K. Lee, X. Li, and D. Wang, "A comprehensive survey on genetic algorithms for DNA motif prediction," *Inf. Sci.*, vol. 466, pp. 25–43, Oct. 2018.

[4] R. Al-Ouran, R. Schmidt, A. Naik, J. Jones, F. Drews, D. Juedes, L. Elnitski, and L. Welch, "Discovering gene regulatory elements using coverage-based heuristics," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1290–1300, Jul./Aug. 2018.

[5] S. Peng, M. Cheng, K. Huang, Y. Cui, Z. Zhang, R. Guo, X. Zhang, S. Yang, X. Liao, Y. Lu, Q. Zou, and B. Shi, "Efficient computation of motif discovery on intel many integrated core (MIC) architecture," *BMC Bioinf.*, vol. 19, no. 9, 2018, Art. no. 282.

[6] K.-C. Wong, "DNA motif recognition modeling from protein sequences," *iScience*, vol. 7, pp. 198–211, Sep. 2018.

[7] S. Patra and A. Mohapatra, "Motif discovery in biological network using expansion tree," *J. Bioinf. Comput. Biol.*, vol. 16, no. 6, 2018, Art. no. 1850024.

[8] A. Painsky and S. Rosset, "Optimal set cover formulation for exclusive row biclustering of gene expression," *J. Comput. Sci. Technol.*, vol. 3, pp. 423–435, May 2014.

[9] T. C. Glenn, "Field guide to next-generation DNA sequencers," *Mol. Ecol. Resour.*, vol. 11, no. 5, pp. 759–769, 2011.

[10] B. Malin, "Protecting dna sequence anonymity with generalization lattices," School Comput. Sci., Inst. Softw. Res. Int., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-ISRI-04-134, 2004.

[11] N. Homer, S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig, "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays," *PLoS Genetics*, vol. 4, no. 8, 2008, Art. no. e1000167.

[12] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

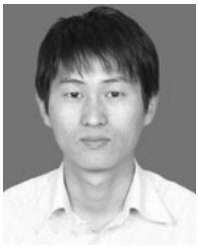
[13] M. Angrist, "Genetic privacy needs a more nuanced approach," *Nature*, vol. 494, no. 7435, p. 7, 2013.

[14] X. Shi and X. Wu, "An overview of human genetic privacy," *Ann. New York Acad. Sci.*, vol. 1387, p. 61, Jan. 2016.

[15] B. A. Malin, "An evaluation of the current state of genomic data privacy protection technology and a roadmap for the future," *J. Amer. Med. Inform. Assoc.*, vol. 12, no. 1, pp. 28–34, 2005.

[16] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. USENIX Secur. Symp., UNIX Secur. Symp.*, 2014, pp. 17–32.
[17] C. Dwork, A. Smith, T. Steinke, and J. Ullman, "Exposed! A survey of attacks on private data," *Annu. Rev. Statist. Appl.*, vol. 4, no. 1, pp. 61–84, 2017.
[18] B. Malin, "Re-identification of familial database records," in *Proc. AMIA Annu. Symp., AMIA Symp.*, 2006, p. 524.
[19] S. S. Shringarpure and C. D. Bustamante, "Privacy risks from genomic data-sharing beacons," *Amer. J. Hum. Genet.*, vol. 97, no. 5, pp. 631–646, 2015.
[20] R. Wang, Y. F. Li, X. F. Wang, H. Tang, and X. Zhou, "Learning your identity and disease from research papers: Information leaks in genome wide association study," in *Proc. Conf. Comput. Commun. Secur. (CCS)*, Chicago, IL, USA, Nov. 2009, pp. 534–544.
[21] T. Takai-Igarashi *et al.*, "Security controls in an integrated biobank to protect privacy in data sharing: Rationale and study design," *BMC Med. Inform. Decis. Making*, vol. 17, no. 1, p. 100, 2017.
[22] T. Haeusermann, M. Fadda, A. Blasimme, B. G. Tzovaras, and E. Vayena, "Genes wide open: Data sharing and the social gradient of genomic privacy," *AJOB Empirical Bioethics*, vol. 9, no. 4, pp. 207–221, 2018.
[23] M. Thomas, "Sociogenetic risks—Ancestry DNA testing, third-party identity, and protection of privacy," *New England J. Med.*, vol. 379, no. 5, pp. 410–412, 2018.
[24] Y. Bregman-Eschet, "Genetic databases and biobanks: Who controls our genetic privacy," *Santa Clara Comput. High Technol. Law J.*, vol. 23, p. 1, Feb. 2006.
[25] Z. Wan, Y. Vorobeychik, M. Kantarcioglu, and B. Malin, "Controlling the signal: Practical privacy protection of genomic data sharing through Beacon services," *BMC Med. Genomics*, vol. 10, no. S2, 2017, Art. no. 39.
[26] X. Lei, X. Zhu, H. Chi, and S. Jiang, "Cloud-assisted privacy-preserving genetic paternity test," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Nov. 2015, pp. 1–6.
[27] A. B. Carter, "Considerations for genomic data privacy and security when working in the cloud," *J. Mol. Diagnostics*, vol. 21, no. 4, pp. 542–552, 2019.
[28] D. Roden, J. M. Pulley, M. A. Basford, G. R. Bernard, E. W. Clayton, J. R. Balser, and D. R. Masys, "Development of a large-scale de-identified DNA biobank to enable personalized medicine," *Clin. Pharmacol. Therapeutics*, vol. 84, no. 3, pp. 362–369, 2008.
[29] A. Harmanci and M. Gerstein, "Analysis of sensitive information leakage in functional genomics signal profiles through genomic deletions," *Nature Commun.*, vol. 9, no. 1, 2018, Art. no. 2453.
[30] N. T. L. Tran and C.-H. Huang, "A survey of motif finding Web tools for detecting binding site motifs in ChIP-Seq data," *Biol. Direct*, vol. 9, no. 1, p. 4, 2014.
[31] E. Zaslavsky and M. Singh, "A combinatorial optimization approach for diverse motif finding applications," *Algorithms Mol. Biol.*, vol. 1, Aug. 2006, Art. no. 13.
[32] A. Maiti and A. Mukherjee, "On the Monte–Carlo expectation maximization for finding motifs in DNA sequences," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 2, pp. 677–686, Mar. 2015.
[33] L. Liljas, "Consensus sequences," *Brenners Encyclopedia Genet.*, vol. 1, pp. 163–164, 2013.
[34] U. Ohler and H. Niemann, "Identification and analysis of eukaryotic promoters: recent computational approaches," *Trends Genet.*, vol. 17, no. 2, pp. 56–60, 2001.
[35] D. E. Schones, P. Sumazin, and M. Q. Zhang, "Similarity of position frequency matrices for transcription factor binding sites," *Bioinformatics*, vol. 21, no. 3, pp. 307–313, 2005.
[36] U. J. Pape, S. Rahmann, and M. Vingron, "Natural similarity measures between position frequency matrices with an application to clustering," *Bioinformatics*, vol. 24, no. 3, pp. 350–357, 2008.
[37] H. G. Zhang and X. Z. Hu, "An improved method for predicting structure class of 27-class protein folds using increment of diversity," *J. Inner Mongolia Univ.*, vol. 40, no. 3, pp. 285–290, 2009.
[38] G. D. Stormo and G. W. Hartzell, "Identifying protein-binding sites from unaligned DNA fragments," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 4, pp. 1183–1187, 1989.
[39] T. D. Schneider and R. M. Stephens, "Sequence logos: A new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, no. 20, pp. 6097–6100, 1990.
[40] O. Devuyst, "The 1000 genomes project: Welcome to a new world," *Peritoneal Dialysis Int.*, vol. 35, no. 7, pp. 676–677, 2015.
[41] R. Akbani *et al.*, "A pan-cancer proteomic perspective on the cancer genome atlas," *Nature Commun.*, vol. 5, no. 1, 2014, Art. no. 3887.
[42] J. H. Hung and Z. P. Weng, "Motif finding," *Cold Spring Harbor Protocols*, vol. 2017, no. 2, pp. 92–97, 2017.
[43] G. Pesole, N. Prunella, S. Liuni, M. Attimonelli, and C. Saccone, "WORDUP: An efficient algorithm for discovering statistically significant patterns in DNA sequences," *Nucleic Acids Res.*, vol. 20, no. 11, pp. 2871–2875, 1992.
[44] S. Sinha and M. Tompa, "Discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Res.*, vol. 30, no. 24, pp. 5549–5560, 2002.
[45] S. Sinha and M. Tompa, "YMF: A program for discovery of novel transcription factor binding sites by statistical overrepresentation," *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3586–3588, 2003.
[46] Y. Li, L. H. U, M. L. Yiu, and Z. Gong, "Quick-motif: An efficient and scalable framework for exact motif discovery," in *Proc. IEEE Int. Conf. Data Eng.*, Apr. 2015, pp. 579–590.
[47] S. Chaturvedi, V. K. Bhartiya, and S. Negi, "Charge calculation studies done on a single walled carbon nanotube using MOPAC," *Indian J. Phys.*, vol. 92, no. 4, pp. 479–485, 2018.
[48] G. Z. Hertz and G. D. Stormo, "Identifying DNA and protein patterns with statistically significant alignments of multiple sequences," *Bioinformatics*, vol. 15, no. 7, pp. 563–577, 1999.
[49] G. Pavesi, P. Mereghetti, G. Mauri, and G. Pesole, "Weeder Web: Discovery of transcription factor binding sites in a set of sequences from co-regulated genes," *Nucleic Acids Res.*, vol. 32, no. 1, pp. 199–203, 2004.
[50] E. Eskin and P. A. Pevzner, "Finding composite regulatory patterns in DNA sequences," in *Proc. 10th Annu. Int. Conf. Intell. Syst. Mol. Biol.*, Edmonton, AB, Canada, Aug. 2002, pp. 354–363.
[51] S. Liang, M. P. Samanta, and B. A. Biegel, "cWINNOWER algorithm for finding fuzzy dna motifs," *J. Bioinf. Comput. Biol.*, vol. 2, no. 1, pp. 47–60, 2008.
[52] G. C. G. Wei and M. A. A. Tanner, "A Monte Carlo implementation of the em algorithm and the poor man's data augmentation algorithms," *J. Amer. Stat. Assoc.*, vol. 85, no. 411, pp. 699–704, 1990.
[53] J. van Helden, A. F. Rios, and J. Collado-Vides, "Discovering regulatory elements in non-coding sequences by analysis of spaced dyads," *Nucleic Acids Res.*, vol. 28, no. 8, pp. 1808–1818, 2000.
[54] T. L. Bailey and C. Elkan, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers," in *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 1994, pp. 28–36.
[55] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li, "MEME: Discovering and analyzing DNA and protein sequence motifs," *Nucleic Acids Res.*, vol. 34, pp. W369–W373, Jul. 2006.
[56] C. E. Lawrence, S. F. Altschul, M. S. Boguski, J. S. Liu, A. F. Neuwald, and J. C. Wootton, "Detecting subtle sequence signals: A Gibbs sampling strategy for multiple alignment," *Science*, vol. 262, no. 5131, p. 208, 1993.
[57] L. J. Palmer, W. O. C. M. Cookson, A. L. James, A. W. Musk, and P. R. Burton, "Gibbs sampling–based segregation analysis of asthma-associated quantitative traits in a population-based sample of nuclear families," *Genetic Epidemiol.*, vol. 20, no. 3, pp. 356–372, 2001.
[58] C. Boucher and J. King, "Fast motif recognition via application of statistical thresholds," *BMC Bioinf.*, vol. 11, no. 1, 2010, Art. no. S11.
[59] C. Boucher, D. G. Brown, and P. Church, "A graph clustering approach to weak motif recognition," in *Proc. Int. Workshop Algorithms Bioinf.*, in Lecture Notes in Computer Science, vol. 4645, 2007, pp. 60–149.
[60] M. Tompa *et al.*, "Assessing computational tools for the discovery of transcription factor binding sites," *Nature Biotechnol.*, vol. 23, no. 1, pp. 137–144, 2005.
[61] L. Kuttippurathu, M. Hsing, Y. Liu, B. Schmidt, D. L. Maskell, K. Lee, A. He, W. T. Pu, and S. W. Kong, "CompleteMOTIFs: DNA motif discovery platform for transcription factor binding experiments," *Bioinformatics*, vol. 27, no. 5, pp. 715–717, 2011.
[62] A. P. Schwab, H. S. Luu, J. Wang, and J. K. Park, "Genomic privacy," *Clin. Chem.*, vol. 64, no. 12, pp. 1696–1703, 2018.
[63] C. Patsakis, A. Zigomitros, and A. Solanas, "Privacy-aware genome mining: Server-assisted protocols for private set intersection and pattern matching," in *Proc. IEEE Int. Symp. Comput.-Based Med. Syst.*, Jun. 2015, pp. 276–279.
[64] J. Gillott, "Genetic databases: Socio-ethical issues in the collection and use of DNA," *Mod. Law Rev.*, vol. 68, no. 4, pp. 705–708, 2010.

[65] M. D. Sorani, J. K. Yue, S. Sharma, G. T. Manley, A. R. Ferguson, S. R. Cooper, K. Dams-O'Connor, W. A. Gordon, H. F. Lingsma, A. I. R. Maas, D. K. Menon, D. J. Morabito, P. Mukherjee, D. O. Okonkwo, A. M. Puccio, A. B. Valadka, and E. L. Yuh, "Genetic data sharing and privacy," *Neuroinformatics*, vol. 13, no. 1, pp. 1–6, 2014.
[66] C. Lippert *et al.*, "Identification of individuals by trait prediction using whole-genome sequencing data," *Proc Nat. Acad. Sci. USA*, vol. 114, no. 38, p. 10166, 2017.
[67] J. Gitschier, "Inferential genotyping of Y chromosomes in latter-day saints founders and comparison to Utah samples in the HapMap project," *Amer. J. Hum. Genet.*, vol. 84, no. 2, pp. 251–258, 2009.
[68] J. Kaye, P. Boddington, J. de Vries, N. Hawkins, and K. Melham, "Ethical implications of the use of whole genome methods in medical research," *Eur. J. Hum. Genet.*, vol. 18, no. 4, pp. 398–403, 2010.
[69] J. B. Gutierrez, M. Frith, and K. Nakai, "A genetic algorithm for motif finding based on statistical significance," in *Bioinformatics and Biomedical Engineering*. Cham, Switzerland: Springer, 2015.
[70] M. Naveed, E. Ayday, E. W. Clayton, J. Fellay, C. A. Gunter, J.-P. Hubaux, B. A. Malin, and X. Wang, "Privacy in the genomic era," *ACM Comput. Surv.*, vol. 48, no. 1, 2015, Art. no. 6.
[71] J. Chen, F. Sun, J. Fu, and H. Zhang, "Association of *TBX*20 gene polymorphism with congenital heart disease in Han Chinese neonates," *Pediatric Cardiol.*, vol. 36, no. 4, pp. 737–742, 2015.
[72] S.-H. Jo, S.-G. Kim, Y. J. Choi, N.-R. Joo, G.-Y. Cho, S.-R. Choi, E.-J. Kim, H.-S. Kim, H.-J. Kim, and C.-Y. Rhim, "KLOTHO gene polymorphism is associated with coronary artery stenosis but not with coronary calcification in a Korean population," *Int. Heart J.*, vol. 50, no. 1, pp. 23–32, 2009.
[73] X. Wu, Y. Wei, Y. Mao, and L. Wang, "A differential privacy DNA motif finding method based on closed frequent patterns," *Cluster Comput.*, vol. 21, pp. 1–13, Jan. 2018.
[74] J. L. Raisaro, G. Choi, S. Pradervand, R. Colsenet, N. Jacquemont, N. Rosat, V. Mooser, and J.-P. Hubaux, "Protecting privacy and security of genomic data in i2b2 with homomorphic encryption and differential privacy," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1413–1426, Sep./Oct. 2018.
[75] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, nos. 3–4, pp. 211–407, 2014.



XIANG WU received the B.Eng. degree in information engineering and the M.S. and Ph.D. degrees in communication and information system from the China University of Mining and Technology, Xuzhou, China, in 2007, 2010, and 2014, respectively, where he is currently pursuing the Ph.D. degree with the School of Safety Engineering. He is currently an Associate Professor with Xuzhou Medical University. His research interests include privacy protection and information security.

HUANHUAN WANG received the master's degree in computer technology from the Anhui University of Technology, in 2015. She is currently pursuing the Ph.D. degree with the China University of Mining and Technology. She is currently an Experimenter with Xuzhou Medical University. Her research interests include medical privacy protection, and machine learning approaches in medical informatics.

MINYU SHI received the B.S. degree in management from Shanxi Medical University, Taiyuan, China, in 2018. She is currently pursuing the M.S. degree in medical informatics with Xuzhou Medical University, Xuzhou, China. Her research interest includes medical privacy protection.

AMING WANG received the M.S. degree in computer science from Wuhan University, Wuhan, China, in 1993, and the Ph.D. degree in computer science from the China University of Mining and Technology, Xuzhou, China, in 2004. His research interests mainly include information security and image processing.

KAIJIAN XIA is currently a Senior Engineer with the Changshu Hospital, Soochow University, and an Associate Professor with Xuzhou Medical University, China. He has authored or coauthored more than 50 research articles in international/national journals. His current research interests include medical information, medical image processing, deep learning, transfer learning, computational intelligence, and their applications in smart medicine. He is also a member of the Information Professional Committee of the Chinese Hospital Association and the Program Vice-Chair of the CyberLife 2019 Procedure Committee. He is also an Associate Editor for the *Journal of medical imaging and health informatics* and the Lead Guest Editor for the IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS and the *Journal of Grid Computing*.