Journal of
Computational and Theoretical Nanoscience
Vol. 11, 1–6, 2014

A Novel Approach to Identify Protein Coding Domains by Sampling Binary Profiles from Genome

Tao Song[1], Xun Wang[2,*], and Yansen Su[1]

[1]*Key Laboratory of Image Processing and Intelligent Control, Department of Control Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China*
[2]*Graduate School of Systems and Information Engineering, University of Tsukuba, Tsukuba, 3050006, Ibaraki, Japan*

In recent works, it was reported that the distributions of specific binary profiles associated with DNA sequences can be used for rapid homology assessment in genome space. In this work, following this line of research, we propose a new and effective approach to identify protein coding domains using binary profiles. In our method, a set of DNA segments having similar algebraic structures are represented by one binary profile. The binary profiles with higher appearance rates in known protein coding domains can be used to find unknown or potential protein coding domains. We test our method on complete sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus gallus. Experiment results show that the binary profile method performs significantly in identifying unknown protein coding domains. By statistic analysis, we conclude that the obtained experimental results are statistically significant.

Keywords: Gene Identification, Binary Profile, Protein Coding Domain Identification, Protein Coding Domain Prediction.

RESEARCH ARTICLE

1. INTRODUCTION

It was stated in Ref. [1] that "Nature is a tinkerer and not an inventor" to emphasize the fact that DNA segments with high similarity hold similar biological functions in distinct species. As usual, after a new genome is sequenced, the first step is to identify its protein coding domains and the putative function of the proteins by aligning with known genes. In the alignment, DNA sequences are compared with some reference functioning DNA segments to detect potential protein coding cites. Experimental methods of DNA sequences alignment, such as DNase foot-printing[2] and gel shift assay,[3] are feasible in laboratories, but are labor-intensive, time consuming and expensive when high-throughput sequencing approaches have grown in recent years (since even a single experiment can generate a huge number of genomic data). With the purpose of using computers to identify genes and protein coding sites, many computational methods for pair-wise sequences alignment have been proposed in Refs. [4–8], as well as some multiple sequences alignment tools were introduced in Refs. [9–11]. Some of these alignment approaches have been widely used as elemental tools in understanding newly sequenced genome. But, in recent decades, available genomic data were growing rapidly at the rate specified by Moore's law, so developing more efficiently computational tools for sequences alignment is still a hot researching field in both bioinformatics and functional genomics.

The idea of the alignment methods is taking advantage of genetic information from other species to identify unknown protein coding domains, but sometimes the internal information in the genome should be paid more attentions, because many individual functions in various organisms have been found and proved to be evolutionary independent.[12–15] Inspired by this biological fact, alignment-free sequence analyzing methods were developed, such as the average mutual information profiles were used as genomic signatures in Refs. [16, 17], k-words statistic method was proposed in Ref. [18], the approach of compression-based classification was discussed in Ref. [19] and fast model based homology protein detection was developed in Ref. [20] and so on. More useful reviews of the alignment-free methods can be found in Ref. [21].

Recently, the method of aligning distribution sequences of $(7, 3, 1)$-difference set in genome space was developed in Ref. [22] for rapid sequence homology assessment. (In the filed of combinatorial design, a difference set corresponds to a specific binary profile, which can be utilized

*Author to whom correspondence should be addressed.

to represent a class of binary sequences with similar complexity and taken as a basic functional unit with specific algebraic structures. Specifically, the $(7, 3, 1)$-difference set corresponds to binary profile 0110100^{23}). It was found that distributions of $(7, 3, 1)$-difference set have a high similarity in homology organisms.[22]

In this work, we propose an approach using binary profiles to identify unknown protein coding domains in genomes. In our method, DNA segments having similar algebraic structures are represented by one binary profile. By sampling specific binary profiles from DNA sequences, we can calculate the appearance rates of the binary profiles in known protein coding domains, and the binary profiles with higher appearance rates are used to predict unknown or potential protein coding domains. This method has following several advantages:

— The notion of algebraic structure becomes close to biological phenomena. DNA segments corresponding to the same binary profile have similar repetitive structures as the binary profile.[22]

— DNA sequences with similar algebraic structures and repetitive subsequences can be represented by one binary profile. This can reduce the cost of time and space to recognize repetitive segments from long DNA sequences.

— The genetic information hidden in binary profiles can be dug out and utilized to discover potential protein coding domains.

We test our method on complete sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus gallus, which contain 362, 4199 and 16855 protein coding domains, respectively. In each data experiment, a number of known protein coding domains are initially selected at random and taken as unknown domains, and then we use the binary profiles with higher appearance rates in left known domains (called label binary profiles) to identify the initially chosen domains. Experimental results show that our method performs superior to the average mutual information profiles method and the k-words statistic method on detecting protein coding domains, if fewer protein coding domains are initially chosen to identify and binary profiles with appearance rates at top 0.05% are selected as label binary profiles. For statistical analysis, we do another 30 experiments with the label binary profile being randomly chosen. As results, the accuracy rates of our method decrease greatly and performance coefficients of our method become even worse. This indicates that our method is statistically significant. The binary profile method has been implemented using MATLAB and the original code is available upon request from the authors.

2. THE BINARY PROFILE METHOD

This section is started by introducing the notion of binary profile. With the notion, a set of DNA segments having similar algebraic structures can be represented by one binary profile.

2.1. Binary Profile

Before introducing the notion of binary profile, it is necessary to describe how to split a DNA sequence into four binary subsequences.

DEFINITION 1. Let ω be a DNA sequence of length $n > 0$. For any nucleotide $\alpha \in \{A, T, C, G\}$, we denote the binary subsequence of ω with nucleotide α is $Subs_\alpha(\omega)$, which is obtained by writing 1 on the ith bit of $Subs_\alpha(\omega)$ if α occurring at the ith site of ω; otherwise writing 0 on the ith bit of $Subs_\alpha(\omega)$.

For any DNA sequence, there are four binary subsequences. For example, DNA sequence

$$\omega = ATTACGTCGCTATCGCTAA$$

can be split into four binary subsequences

$$Subs_A(\omega) = 1001000000010000011$$
$$Subs_T(\omega) = 0110001000101000100$$
$$Subs_C(\omega) = 0000100101000101000$$
$$Subs_G(\omega) = 0000010010000010000.$$

For any DNA sequence ω, we can obtain the following proposition.

PROPOSITION 1. *Let ω be a DNA sequence of length $n > 0$. It holds that $Subs_A(\omega) + Subs_T(\omega) + Subs_C(\omega) + Subs_G(\omega) = 11, \ldots, 1$.*

DEFINITION 2. Let υ be a non-zero binary sequence of length $m > 0$. By $Prof_\upsilon$, we denote the set of DNA segments of length m whose binary subsequences include υ. The binary sequence υ is called the binary profile of $Prof_\upsilon$.

It is not hard to find that DNA segments in $Prof_\upsilon$ have similar algebraic structures with its binary profile υ. For example, for binary profile $\upsilon = 110110110$ having three repetitive segments of 110, each DNA segment in $Prof_\upsilon$ consists of three repetitive structures of the form $\alpha\alpha\beta$, where $\alpha \in \{A, T, C, G\}$, $\beta \in \{A, T, C, G\}/\alpha$ and $\alpha\alpha\beta$ can be AAT, TTC, GGC or GGT etc. For any binary profile υ, the size of $Prof_\upsilon$ has a close relation with the number of bits with value 0 in υ.

PROPOSITION 2. *Let υ be a binary profile with length $m > 0$, and the number of bits with value 0 in υ be m_0. There are 4×3^{m_0} DNA segments of length m in $Prof_\upsilon$.*

PROOF Let $\upsilon = \upsilon_1\upsilon_2 \ldots \upsilon_m$ be a binary profile with m_0 bits being value 0. If nucleotide $\alpha \in \{A, T, C, G\}$ corresponds to the bits with value 1 in binary profile υ, then any bit with value 0 corresponds to one of the left three nucleotides in $\{A, T, C, G\}/\alpha$. Hence, there are 4×3^{m_0} DNA segments in $Prof_\upsilon$.

RESEARCH ARTICLE

2.2. Distribution Sequence and Appearance Rate

Let us now introduce how to calculate the distribution sequence of a binary profile in a DNA sequence. From the distribution sequence, we can obtain the appearance rate of the binary profile in known protein coding domain(s).

Let $\omega = \omega_1\omega_2\ldots\omega_n$ be a DNA sequence of length $n > 0$, whose binary subsequences is

$$Subs_\alpha(\omega) = Subs_\alpha(\omega_1)Subs_\alpha(\omega_2)\ldots Subs_\alpha(\omega_n)$$

with $\alpha \in \{A, T, C, G\}$, and $\upsilon = \upsilon_1\upsilon_2\ldots\upsilon_m$ be a binary profile of length m with $n \geq m > 0$. We can sample υ from each $Subs_\alpha(\omega)$ by checking the appearance of υ in $Subs_\alpha(\omega)$. Specifically, for any $1 \leq i \leq n - m + 1$, if it satisfies that $Subs_\alpha(\omega_i) = \upsilon_1$, $Subs_\alpha(\omega_{i+1}) = \upsilon_2, \ldots, Subs_\alpha(\omega_{i+m}) = \upsilon_m$, then write 1 on the ith bit of $Rec^\omega_\alpha(\upsilon)$; otherwise write 0, where $Rec^\omega_\alpha(\upsilon)$ is the recording sequence of υ on $Subs_\alpha(\omega)$. As results, four recoding sequences are obtained.

By summing the four recoding binary sequences, the distribution sequence of υ on ω can be achieved, which is

$$Dis_\upsilon(\omega) = \sum_{\alpha\in\{A,T,C,G\}} Rec^\omega_\alpha(\upsilon)$$

The length of each $Rec^\omega_\alpha(\upsilon)$ and $Dis_\upsilon(\omega)$ are both $n-m+1$. By the following proposition, we emphasize the fact that the bits with value 1 in recoding sequences are preserved in the distribution sequence.

Proposition 3. *For any* $1 \leq i \leq n - m + 1$ *and* $\alpha \in \{A, T, C, G\}$, *if* $Rec_\alpha(\omega_i) = 1$, *then it holds* $Dis_\upsilon(\omega_i) = 1$.

A domain $r = [s, s+t]$ in DNA sequence ω means the domain from the sth to $s+t$th bit of ω, where $s \geq 1$ and $s + t \leq n - m + 1$. The appearance rate of υ in domain r is denoted by

$$App^\omega_\upsilon(r) = \sum_{i=s}^{s+t} Dis_\upsilon(\omega_i) \Big/ \sum_{j=1}^{n-m+1} Dis_\upsilon(\omega_j)$$

where $Dis_\upsilon(\omega_k) \in \{0, 1\}$ with $k = 1, 2, \ldots, n + m - 1$.

The notion of appearance rate in one domain can be extended to a set of domains R. Let $R = \{r_1, r_2, \ldots, r_k\}$ be a set of domains with any r_i and r_j being disjoint. The appearance rate of υ in R is

$$App^\omega_\upsilon(R) = \sum_{i=1}^{k} App^\omega_\upsilon(r_i)$$

We will illustrate the notions introduced above by a concrete example. Let $\omega = ATATGCTGTGACGCGC$ and $\upsilon = 1010$. The four binary subsequences of ω are

$$Subs_A(\omega) = 1010000000100000$$
$$Subs_T(\omega) = 0101001010000000$$
$$Subs_C(\omega) = 0000010000010101$$
$$Subs_G(\omega) = 0000100101001010$$

By checking the appearance of binary profile $\upsilon = 1010$, we obtain four recording sequences:

$$Rec^\omega_A(\upsilon) = 1000000000000$$
$$Rec^\omega_T(\upsilon) = 0100001000000$$
$$Rec^\omega_C(\upsilon) = 0000000000010$$
$$Rec^\omega_G(\upsilon) = 0000000100001$$

Hence, the distribution sequence of υ on ω is

$$Dis_\upsilon(\omega) = \sum_{\alpha\in\{A,T,C,G\}} Rec^\omega_\alpha(\upsilon) = 1100001100011$$

The length of ω and υ are 16 and 4, so the length of $Rec^\omega_\alpha(\upsilon)$ and $Dis_\upsilon(\omega)$ are both $16-4+1 = 13$.

The appearance rate of $\upsilon = 1010$ in domain $[1, 3]$ is

$$App^\omega_\upsilon(r) = \sum_{i=1}^{3} Dis_\upsilon(\omega_i) \Big/ \sum_{j=1}^{13} Dis_\upsilon(\omega_j) = 33.33\%$$

The appearance rate of $\upsilon = 1010$ in $R = \{r_1 = [1, 3], r_2 = [4, 5], r_3 = [7, 9]\}$ is

$$App^\omega_\upsilon(R) = \sum_{i=1}^{3} App^\omega_\upsilon(r_i) = 66.67\%$$

2.3. Time and Space Complexity

We discuss the efficiency of our method by considering the cost of time and space to calculate appearance rate of a binary profile in one domain.

Let ω be a DNA sequence of length n and υ be a binary profile of length m with $0 < m \leq n$. To split ω into four binary subsequences, we need to check whether a nucleotide is present or not on each of the n bits of ω, and then writing 1 or 0 on the corresponding bits of $Subs_\alpha(\omega)$. This progress costs n checking operations and n writing operations. So, it costs $4 \times 2n = 8n$ steps to split ω into four binary subsequences. To sample υ from ω, it needs $m(n - m + 1)$ steps to check the appearance of υ in each binary subsequence of ω, as well as $n - m + 1$ steps to generate recording sequence. Hence, it costs $4 \times (m + 1)(n - m + 1)$ steps to compute the four recording sequences. We can calculate the distribution sequence of υ on ω by $8n + (4m + 7)(n - m + 1)$ steps. It will cost at most $2(n - m + 1)$ steps to calculate the appearance rate of a binary profile in a given domain. Thus, the cost of time to calculate appearance rate of a binary profile in one domain is $8n + (4m + 7)(n - m + 1) + 2(n - m + 1)$, which is $O(n^2)$.

In the following, we will discuss the space complexity to compute appearance rate of a binary profile in one domain. Initially, we need $O(n)$ bits to store DNA sequence ω, as well as $4n$ bits to store the four binary subsequences of ω. Further more, $4(n - m + 1)$ bits are used to store

RESEARCH ARTICLE

the four recording sequences, $(n-m+1)$ bits to store the distribution sequence, and m bits to store v. Thus, $O(n)+4n+5(n-m+1)+m$ bits are sufficient to compute the appearance rate of any binary profile.

It is worth to point out that the number of binary profiles increases exponentially with respect to the length m, which is $\sum_{i=3}^{m} 2^m - 1$. (The binary profiles with all bits being zero and length less than 3 are ignored in this work, for they have no genetic meaning). In our research, distribution sequences of all binary profiles of length from 3 to 13 are considered.

3. DATA EXPERIMENTS

In the data experiment, a number of protein coding domains are initially chosen at random and taken as non-functioning domains, and then we use the distribution sequence of label binary profiles (binary profiles with higher appearance rates in the left known protein coding domains), to identify those initially chosen protein coding domains. Specifically, if the distribution sequence of one label binary profile has at least one bit with value 1 in the domain, then this domain is said to identified by the label binary profile. According to the description above, the data experiment contains three main steps:

Step 1. Choosing a number of the protein coding domains at random as non-protein coding domains.

Step 2. Selecting label binary profiles with a threshold value of appearance rates in known protein coding domains.

Step 3. Checking whether the protein coding domains chosen in step 1 can be identified by label binary profile(s) selected in step 2.

3.1. An Experiment

We will illustrate the processes above by a concrete data experiment on complete sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1. There are 362 protein coding domains in *Halalkalicoccus jeotgali B*3 *plasmid* 1. Initially, we randomly choose domain [118073, 119122] encoding protein *ABC transporter ATP-binding* as a non-protein coding domain, and then we calculate the appearance rates of binary profiles of length from 3 to 13 in the left 361 protein coding domains. The numbers of binary profiles with different appearance rates are shown in Figure 1, where x axis represents the appearance rates of binary profiles, and y axis is the numbers of the binary profiles.

From all the binary profiles, the ones with appearances rates at top 0.05% are chosen as label binary profiles, which are given in Table I. The distribution sequences of the 8 label binary profiles from 117500 to 112000 bit are indicated in Figure 1, where we use bars to represent bits with value 1 of distribution sequences. The binary profiles 100100100101, 010010010110 101001001001, 001011001001, 10001001011 and 100100110100 can identity domain [118073, 119122] (having bit(s) with 1 in domain [118073, 119122]). This implies that the initially chosen domain of protein *ABC transporter ATP-binding* can be identified by the binary profile method.

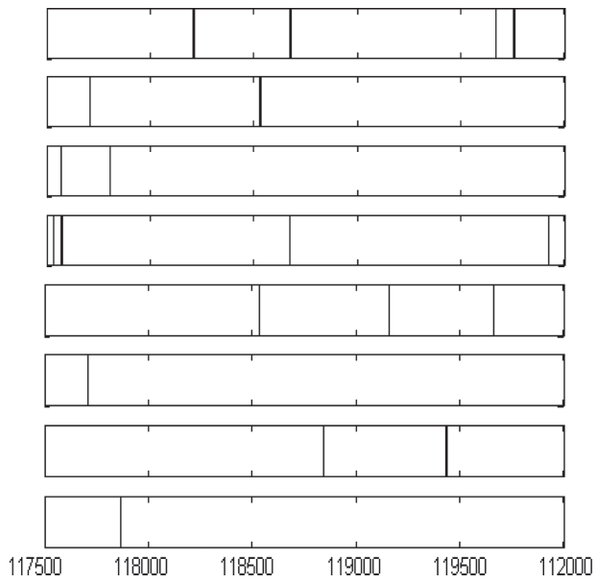

Fig. 1. The number of binary profiles with different appearance rates.

We repeat the experiment for 1000 times. In each experiment, one domain is randomly chosen to identify. In 934 experiments, the randomly chosen protein coding domain can be identified, hence the accuracy rate of the method achieves 93.4%. We do another 30 experiments for statistical analysis, where we randomly choose 8 label binary profiles. The average accuracy of the 30 experiments is 48.6%. The accuracy rates of the 30 experiments are given by blue dots in Figure 3, and the accuracy rate of our method is indicated by the red star in Figure 3. So, we can conclude that our experimental results are statistically significant.

3.2. Experimental Results

We test the binary profile method on complete sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus

Table I. Labeling binary profiles.

Number	Binary profile	Appearance rate (%)	Length
1	100100100101	79.07	12
2	010010010110	77.10	12
3	11010010010	76.89	11
4	101001001001	76.76	12
5	001011001001	76.71	12
6	001001001011	76.64	12
7	10001001011	75.65	11
8	100100110100	75.43	12

Table II. The experimental results on complete sequence of *Halalkalicoccus jeotgali B3 plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus gallus.

H. jeotgali B3 plasmid 1				Escherichia coli ATCC 8739				Gallus gallus			
k	α (%)	AR (BPM) (%)	PC (BPM) (%)	k k	γ γ (%)	AR (BPM) (%)	PC (BPM) (%)	k k	α α (%)	AR (BPM) (%)	PC (BPM) (%)
5	0.05	100	94.4	50	0.05	100	96.7	50	0.05	100	94
5	0.10	100	86.7	50	0.10	100	81.2	50	0.10	100	98
5	0.15	100	60.4	50	0.15	100	70.3	50	0.15	100	96
5	0.20	100	34.2	50	0.20	100	42.1	100	0.05	100	92.6
5	0.25	100	13.6	50	0.25	100	12.6	100	0.10	100	91.4
5	0.30	100	3.4	50	0.30	100	1.4	100	0.15	100	95.6
20	0.05	100	80.6	100	0.05	100	83.6	200	0.05	100	87
20	0.10	100	62.7	100	0.10	100	59.7	200	0.10	100	86
20	0.15	100	40.4	100	0.15	100	39.4	200	0.15	100	81
20	0.20	100	10.2	100	0.20	100	1.2	500	0.05	93	85.7
20	0.25	100	3.3	100	0.25	100	0.3	500	0.10	95	79.3
20	0.30	100	0.12	100	0.30	100	0.019	500	0.15	95	60.3
40	0.05	90	73.5	200	0.05	93	20.5	800	0.05	67	34.2
40	0.10	95	38.7	200	0.10	95	16.7	1200	0.05	43.2	9.3
40	0.15	100	19.4	200	0.15	95	3.4	1600	0.05	62	1.7
80	0.05	73.7	40.9	400	0.05	72	19.9				
80	0.10	76.2	13.3	400	0.10	72	9.3				
80	0.15	80	0.03	400	0.15	80	1.6				
100	0.05	60	20.3	500	0.05	58	11.7				
120	0.05	30	11.5	800	0.05	39	6.5				
140	0.05	21.6	0.66	1000	0.05	18.6	0.3				

gallus, which contain 362, 4199 and 16855 protein coding domains, respectively.

The performances of the binary profile method on complete sequence of *Halalkalicoccus jeotgali B3 plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus gallus are given in Table I, where k is the number of initially chosen protein coding domains for identification, and γ is the threshold value of selecting label binary profiles (if $\gamma = 0.05\%$, then the binary profiles with appearance rates in known protein coding domains at top 0.05% will be selected as label binary profiles). Let K be the set of initially selected protein coding domains, and W be the set of predicted protein coding domains by the label binary profiles. The accuracy rate (AR) of identifying protein coding domains is $|K \cap W|/|K|$. By $|K \cap W|/|K \cup W|$, we denote the performance coefficient (PC) to evaluate the performance of the binary profile method.

As shown in Table II, in data experiment on complete sequence of *Halalkalicoccus jeotgali B3 plasmid* 1, when k are assigned with small values (5 and 20), the binary profile method can significantly identify the randomly chosen protein coding domains. The performance coefficients of our method are well when γ is 0.05%. With the increment of γ from 0.05% to 0.30%, although the accuracy rates are still 100%, the performance coefficients decrease rapidly. This is due to the fact when γ is assigned with large value, the number of label binary profiles increases,

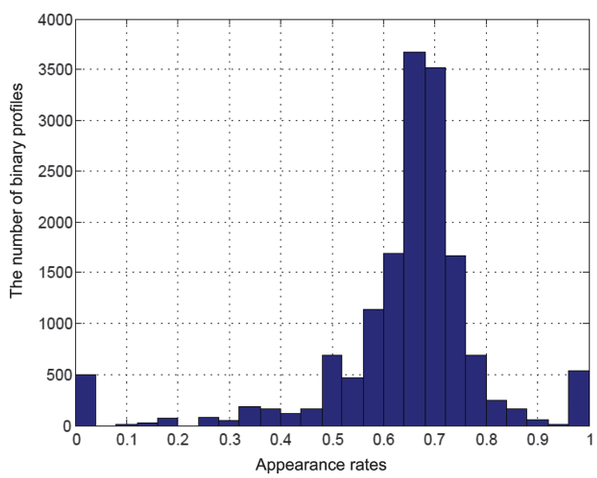

Fig. 2. The partial distribution sequences of labeling binary profiles in Table I.

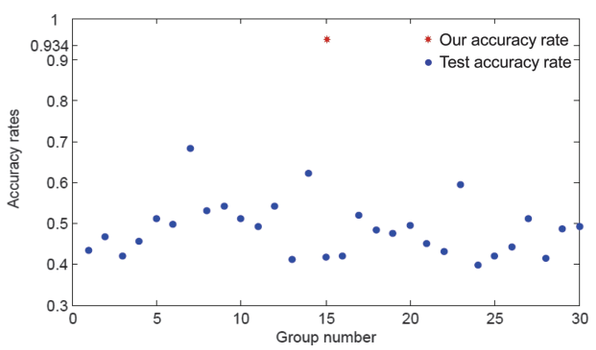

Fig. 3. Accuracy rates of the 30 experiments for statistical analysis.

RESEARCH ARTICLE

hence the number of predicted protein coding domains will greatly increase. In this case, many protein coding domains that are NOT protein coding domains are identified by label binary profiles. As a result, the value of $|K \cup W|$ becomes rather large, so the performance coefficient rapidly decreases. When k becomes large (from 40 to 140), that is, more protein coding domains are considered as unknown in the experiments, the accuracy rates will become unacceptable. In the experiment on the genome of Escherichia coli ATCC 8739, we get similar performances of the binary profile method. Specifically, if k is associated with small values, the accuracy rates and performance coefficients are both significant. But, when k becomes large, the accuracy rates and performance coefficients will decrease rapidly.

The accuracy rates of the average mutual information profiles method on complete sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1 and genome of Escherichia coli ATCC 8739 are 96.5% and 89.6%, and the performance coefficients are 87% and 91%, respectively. The performance of the binary profile method is superior to the average mutual information profiles method and the k-words statistic method when k are assigned with small values and $\gamma = 0.05\%$. But, the binary profile method performs worse when k becomes large.

Experimental results on genome of Gallus gallus show that the binary profile method performs significantly on accuracy rates to identify unknown protein coding domains with k being from 50 to 200. But, when k becomes large, the binary profile method can achieve well accuracy rates, but the performance coefficients will become even worse. The accuracy rate of the average mutual information profiles method to detect protein coding domains on genome of Gallus gallus is 92.5% and performance coefficient is 81%, while the accuracy rate of the k-words statistic method is 87.1%, and the performance coefficient is 76%. The performance of the binary profile method is superior to the average mutual information profiles method and the k-words statistic method when k is less than 200 and γ is 0.05%.

RESEARCH ARTICLE

4. CONCLUSION

In this work, we have proposed an efficient approach, called binary profile method, to identify protein coding domains in genomes. We test the method on the complete DNA sequence of *Halalkalicoccus jeotgali B*3 *plasmid* 1, genome of Escherichia coli ATCC 8739 and genome of Gallus gallus. The performances of the binary profile method to identify protein coding domains are superior to the the average mutual information profiles method and the k-words statistic method when k are assigned small values and $\gamma = 0.05\%$.

For further research, it remains open that if different splitting strategies can improve the performances of the binary profile method. Another interesting problem is whether binary profiles can be used as a potential attempt for handling huge computational cost problems of recognizing DNA segments.

References

1. F. Jacob, *Science* 4295, 1161 (**1977**).
2. D. J. Galas and A. Schmitz, *Nucleic Acids Research* 5, 3157 (**1978**).
3. M. M. Garner and A. Revzin, *Nucleic Acids Research* 9, 3047 (**1981**).
4. B. Conrad, P. T. Henri, F. Ekobena, C. Timoléon, and J. Kofané, *J. Comput. Theor. Nanosci.* 8, 2220 (**2011**).
5. S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, *Nucleic Acids Research* 25, 3389 (**1997**).
6. R. J. George and J. Heringa, *Journal of Molecular Biology* 316, 839 (**2002**).
7. W. Wang and T. Wang, *J. Comput. Theor. Nanosci.* 8, 2266 (**2011**).
8. S. F. Altschul, W. Gish, W. Miller, E. M. Myers, and D. J. Lipman, *Journal of Molecular Biology* 215, 403 (**1990**).
9. B. Morgenster, *Bioinformatics* 15, 211 (**1999**).
10. H. Fan, R. Wu, B. Liao, and X. Lu, *J. Comput. Theor. Nanosci.* 9, 1558 (**2012**).
11. E. Birney, *IBM Journal of Research and Development* 45, 449 (**2001**).
12. E. L. Braun, A. L. Halpern, M. A. Nelson, and D. O. Natvig, *Genome Research* 10, 416 (**2000**).
13. P. R. Macdonald, P. S. Coelho, T. Roemer, S. Agarwal, et al., *Nature* 402, 413 (**1999**).
14. A. Wagner, *Trends Genetic* 17, 237 (**2001**).
15. S. Wong, G. Butler, and K. H. Wolfe, *Proceedings of the National Academy of Sciences* 9, 9272 (**2002**).
16. M. Bauer, S. M Schuster, and K. Sayood, *BMC Bioinformatics* 9, 1 (**2008**).
17. Z. Cui, D. Liu, J. Zeng, and Z. Shi, *J. Comput. Theor. Nanosci.* 9, 2255 (**2012**).
18. Q. Dai and T. Wang, *BMC Bioinformatics* 9, 1 (**2008**).
19. P. Ferragina, R. Giancarlo, V. Greco, G. Manzini, and G. Valiente, *BMC Bioinformatics* 8, 1 (**2007**).
20. S. Hochreiter, M. Heusel, and K. Obermayer, *Bioinformatics* 23, 1728 (**2007**).
21. S. Vinga and J. Almeida, *Bioinformatics* 19, 513 (**2003**).
22. A. K. Brodzik, *IEEE Trans. Information Theory* 56, 756 (**2010**).
23. T. Beth, D. Jungnickel, and H. Lenz, Design Theory, Second edn., Cambridge University Press, Cambridge (**1999**), Vols. I and II.

Received: 7 October 2012. Accepted: 24 October 2012.