A New Model for Motif Representation Based on Suffix and Prefix Code

Mehri Bakhshian[a], Fatemeh Zare-Mirakabad[a*]

[a] Department of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

Abstract. Motif identification in DNA sequences is one of the challenging problems in computational biology and computer science. Researchers are interested in finding an efficient model to represent a DNA motif. There are two common models called position weight matrix and consensus sequence. Although these models are very simple to implement, both of these models consider all motif instances of a motif with the same length. Due to different lengths of motif instances in the experimental data, Markov chain models are extremely impressive to represent motif instances. Unfortunately, these models need a lot of data to be constructed. In this paper, we design a new motif representation model called SUF_PRE based on suffix and prefix of motif instances. Unlike Markov models which are complicated, SUF_PRE model is simple and supports motif instances with the different lengths. SUF_PRE and position weight matrix model are compared in motif search problem on the JASPAR, TRANSFAC and SCPD databases. The results show that the SUF_PRE model gives more accurate prediction than position weight matrix model.

Keywords: Information theory; Motif; Motif search; SUF_PRE model representation.

1. Introduction

A private transcription factor (TF) binds to some specific sites in promoter regions, transcription factor binding sites, to regulate the gene expression. Transcription factor binding site (TFBS) prediction is known as a challenging problem because most sites of a TF vary greatly in their promoters. In addition, mostly the length of each motif instance is short and expects to random occurrence within a few hundred base pairs. One of the most challenging problems in TFBS prediction is how to create a common pattern for a set of TFBSs.

A set of TFBSs detected by a TF is called a motif and each TFBS is named a motif instance. There are two common and simple models to represent a motif called consensus sequence and position weight matrix (PWM). The first model is built with choosing the predominant nucleotide from each position [1]. The second one displays a probability matrix $4 \times l$ where four and l show, respectively the number of nucleotides and the length of each motif instance [2]. Although these models are easy to implement, they only consider a set of motif instances with the same length. This assumption is not always true in the experimental samples, since there are some motif instances of a motif with unequal length`s. For example, Table 1 shows some of TFBSs of Gal4 in yeast. As you can see, the lengths of TFBSs are not equal.

Table1. TFBSs of TF GAL4 extracted from SCDP database [14].

>ATACTTCGGAGCACTGTTGAGCG
>AGCGCTCGGACAACTGTTGACC
>ATTGTTCGGAGCAGTGCGGCGCG
>CGGAGGAGAGTCTTCCG
>TCGGAGGGCTGTCGCCCG
>CGGCGGCTTCTAATCCG
>CGGATTAGAAGCCGCCG
>CGGGCGACAGCCCTCCGA
>CGGAAGACTCTCCTCCG
>CGCGCCGCACTGCTCCGAACAAT
>CACCGGCGGTCTTTCGTCCGTGC
>TATCGGGGCGGATCACTCCGAAC
>CGGCGCACTCTCGCCCG
>TCGGGGCAGACTATTCCGG

Different types of Makov models such as permuted Markov models [3], Markov chain optimization [4], hidden Markov models [5,6], and Bayesian hidden Markov models [7] are chosen to model these types of motifs. Unfortunately, these methods require more complicated mathematical tools, with more parameters to estimate, and require more experimental

* Corresponding author.
E-mail address: mehr_b@aut.ac.ir

data than typically available ones [3, 4, 8, 9].

In this paper, we report a new model for motif representation based on suffix and prefix of codes called SUF_PRE. Not only this model is as simple as the common models, it takes the advantages of Markov models to represent a set of motif instances of a TF with different lengths. We apply the proposed model and PWM in a motif search algorithm on JASPAR, TRANSFAC and SCPD databases. The comparison between our model and PWM in the motif search algorithms shows that our model has a better result than PWM to detect new motif instances.

2. Methods

In this section, we describe position weight matrix model as a common model to represent a motif [11]. Then, the proposed model designed based on suffix and prefix codes, SUF_PRE, is introduced.

2.1. Definition

A DNA sequence S is defined as:

$$S=s_1 \dots s_t, \quad s_i \in \{A, C, G, T\}, \quad |S|=t, \quad (1)$$

where A,C,G and T show four nucleotides, respectively adenine, cytosine, guanine, and thymine found in DNA. A subsequence S_j^l is extracted from sequence S where $S_j^l = s_j \dots s_{j+l-1}$. We display motif $M = \{M_1, \dots, M_n\}$ as a set of motif instances where $M_i = m_{i1} \dots m_{ili}$, $m_{ij} \in \{A, C, G, T\}$ and $|M_i| = l_i$.

In motif search problem, DNA sequence S and set M are given as inputs and the goal is to investigate a new motif instance in the sequence S.

2.2. Position weight matrix model

In PWM model, matrix $B_{\{A, C, G, T\}\times l}$ is defined as follows to depict motif $M = \{M_1, \dots, M_n\}$ where $|M_1| = \dots = |M_n| = l$:

$$B[i,j] = \log_2\left(\frac{p(i,j)}{p(i)}\right),$$

where $p(i,j)$ shows the probability of occurrence nucleotide i in position j of the motif instances in set M. The probability of occurrence nucleotide i in seuquence S is displayed by $p(i)$. A new motif instance S_k^l is found from the sequence S as follows:

$$k = \arg max_{i=1\dots|S|-l+1} Sc(i),$$

$$Sc(i) = \sum_{j=0}^{l-1} B[s_{i+j}, j+1].$$

2.3. SUF_PRE model

The proposed model, SUF_PRE, is defined based on a type of suffix-prefix codes to represent a motif. This model is built in four following steps:

1. All suffix subsequences of each motif instance from set M are extracted and saved in a set called suffix.
2. All prefix subsequences of each sequence from set suffix are added to a set named C as suffix-prefix codes.
3. $F(c_i)$ function computes the occurrence of each code $c_i \in C$, $|c_i| = r$, in all motif instances of set M.
4. $P(c_i)$ Function calculates the probability of code c_i as follows:
   a. Subset $Cr = \{c \in C | \, |c| = r\} \subseteq C$ is generated.
   b. $P(c_i) = \frac{F(c_i)}{\sum_{c \in C_r} F(c)}$.
5. The background probability of code c_i is computed by $B(c_i) = \prod_{j=1}^{r} p(c_{ij})$ function, where $p(c_{ij})$ shows the probability of nucleotide c_{ij} in sequence S while $c_i = c_{i1} \dots c_{ir}$ and $c_{ij} \in \{A, C, G, T\}$.
6. The value of each code $c_i \in C$ is calculated relative to the background as $V(c_i) = log_2(P(c_i) / B(c_i))$.

Unlike PWM representation which the length of all motif instances is equal, this model can be constructed for motif instances with various lengths. For this reason, to extract a subsequence from position i of sequence S, l_i is defined as the length of subsequence where $\forall\ j=1, \dots, l_i$, $S_i^j \in C$, $S_i^{li+1} \in C$.

A new motif instance S_k^{lk} is found from sequence S as follows:

$$k = argmax_{i=1\dots|S|} Sc(i), \qquad Sc(i) = \sum_{j=1}^{l_i} V(S_i^j).$$

3. Results and discussion

In this section, some TFs are extracted from three public databases JASPAR [12], TRANSFAC [13] and SCPD [14] to analyze the proposed model, SUF_PRE, for motif representation. We select 107 TFs from the JASPAR database and implant TFBSs of these TFs in some random sequences which are generated similar to [15]. For extracting motifs from the TRANSFAC database, we used the generated benchmark by Sandve et al [16]. Sandve generated three data set versions from TRANSFAC based on the collections of binding site fragments that are ranked according to the optimal level of discrimination. These data sets are called 'algorithm_Markov', 'algorithm_real' and 'model_real'. In addition, we used the real biological sequences from the SCPD database [14] constructed based on Yeast promoters.

3.1. Methods for comparison

We compare SUF_PRE and PWM models on the above datasets based on introduced criteria explained by Tompa et al. [17]. They introduced different measurements such as Sensitivity (*nSn*), Positive predicted value (*nPPV*), Nucleotide Specificity (*nSp*), Nucleotide Performance Coefficient (*nPC*) [18] and Nucleotide Correlation Coefficient (*nCC*) [19]. The definitions of these measurements are given in Table 2.

In Table 2, the variable *nTP* is the number of nucleotide positions in both known sites and the predicted sites, *nFP* is the number of nucleotide positions not in the known sites but in the predicted sites, *nFN* is the number of nucleotide

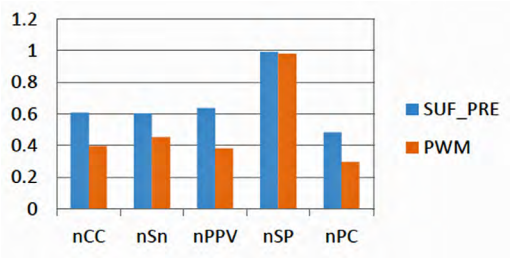

Fig. 1. The result obtained from SUF_PRE and PWM model for motif search on the JASPAR database. These results include nCC, nSn, nPPV, nSP and nPC.

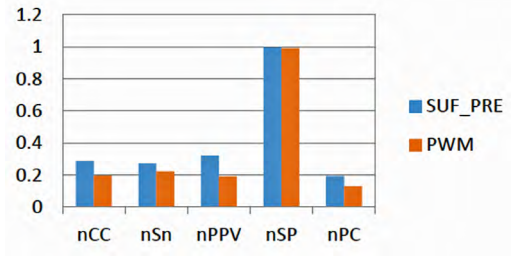

Fig. 2. The comparison between SUF_PRE and PWM models for motif search on the 'algorithm_real' sandve's benchmark. These results include nCC, nSn, nPPV, nSP and nPC.

positions in known sites but not in the predicted sites and nTN is the number of nucleotide positions in neither known sites nor the predicted sites.

Table 2. Comparison measurement formulas.

Mesurment	Formula
nSn	$nTP/(nTP+nFN)$
$nPPV$	$nTP/(nTP+nFP)$
nSp	$nTn/(nTn+nFP)$
nPC	$nTP/(nTP+nFN+nFP)$
nCC	$\frac{nTP.nTN-nFN.nFP}{\sqrt{(nTP+nFN).(nTN+nFP).(nTP+nFP).(nTN+nFN)}}$

3.2. Motif search based on SUF_PRE and PWM model

We implement a jackknife test based on SUF_PRE and PWM models to search a new motif instance in a given sequence. Assume that a motif M has n known TFBSs (motif instances) which are implanted in n sequences. For each j = 1 … n, we ignore j-th TFBS of the set M and calculate the SUF_PRE model (PWM model) for n-1 remaining TFBSs. Then, based on the mentioned search strategy (see section Methods), we search j-th sequence based on SUF_PRE (PWM) model to find a new motif instance. Fig. 1 shows the average results of motif search algorithm based on SUF_PRE and PWM model performed on the TFs from the JASPAR database. This result displays that SUF_PRE model is more successful than PWM model in the motif search problem on JASPAR database.

Fig. 2, 3 and 4 represent the average results of motif search algorithm based on SUF_PRE and PWM models performed on the TFBSs from the 'algorithm_real', 'algorithm_Markov' and 'model_real' sandve's benchmarks. The results show that

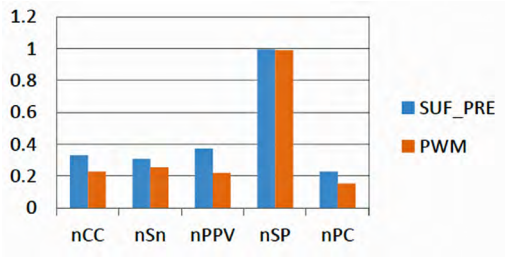

Fig. 3. The comparison between SUF_PRE and PWM models for motif search on the 'algorithm_Markov' sandve's benchmark. These results include nCC, nSn, nPPV, nSP and nPC.

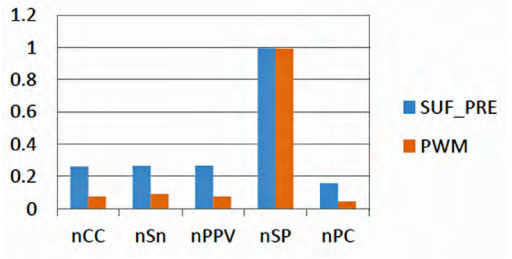

Fig. 4. The comparison between SUF_PRE and PWM models for motif search on the 'model_real' sandve's benchmark. These results include nCC, nSn, nPPV, nSP and nPC.

SUF_PRE model is more successful than PWM model in the motif search problem on these benchmarks.

We apply SCPD database to show that the proposed model can represent each motif with different length motif instances. We could not compare our model with PWM model, because PWM model is not able to represent a motif with unequal length motif instances. Fig. 5 shows the average results of motif search based on SUF_PRE model on the SCPD database.

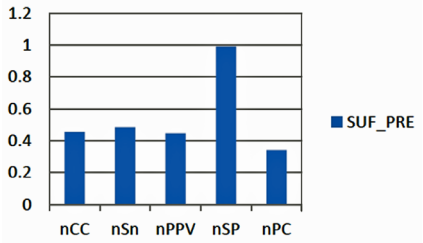

Fig. 5. Result obtained from SUF_PRE model by the jackknife test on SCPD database. This results include nCC, nSn, nPPV, nSP and nPC.

4. Conclusions

In the last three decades, different algorithms have been improved to predict motifs in a set of co-expressed genes [20-22]. One of the most challenging in this problem is the various lengths of motif instances. In this paper, we proposed a new model called SUF_PRE to represent different lengths of motif instances based on suffix and prefix of instances. Not only this model is very simple such as PWM but also it takes the advantage of the complicated models. In other words, we used the proposed and PWM models in the jackknife test for motif search problem. The comparison between PWM

and SUF_PRE in motif search problem on the JASPAR and TRANSFAC databases shows that our model is more powerful to represent a motif. Finally, we performed our model to search motif on the SCPD database. We could not represent the extracted motifs from this database by PWM model, because some motifs of this database contain a set of instances with different lengths. In future, we are interested in improving some of well-known de novo motif finding algorithms by this motif representation model.

References

1. Schneider, Thomas D. "Consensus sequence zen." Applied bioinformatics 1.3 (2002): 111.
2. Xia, Xuhua. "Position weight matrix, gibbs sampler, and the associated significance tests in motif characterization and prediction." Scientifica 2012 (2012).
3. Zhao, Xiaoyue, Haiyan Huang, and Terence P. Speed. "Finding short DNA motifs using permuted Markov models." Journal of Computational Biology 12.6 (2005): 894-906.
4. Ellrott, Kyle, et al. "Identifying transcription factor binding sites through Markov chain optimization." Bioinformatics 18.suppl 2 (2002): S100-S109.
5. Marinescu, Voichita D., Isaac S. Kohane, and Alberto Riva. "MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes." BMC bioinformatics 6.1 (2005): 79.
6. Maaskola, Jonas, and Nikolaus Rajewsky. "Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models." Nucleic acids research 42.21 (2014): 12995-13011.
7. Gelfond, Jonathan AL, Mayetri Gupta, and Joseph G. Ibrahim. "A Bayesian Hidden Markov Model for Motif Discovery Through Joint Modeling of Genomic Sequence and ChIP-Chip Data." Biometrics 65.4 (2009): 1087-1095.
8. Barash, Yoseph, et al. "Modeling dependencies in protein-DNA binding sites." Proceedings of the seventh annual international conference on Research in computational molecular biology. ACM, 2003.
9. King, Oliver D., and Frederick P. Roth. "A non-parametric model for transcription factor binding sites." Nucleic acids research 31.19 (2003): e116-e116.
10. Ma, Qin, et al. "An integrated toolkit for accurate prediction and analysis of cis-regulatory motifs at a genome scale." Bioinformatics 29.18 (2013): 2261-2268.
11. Zhang, Yipu, and Ping Wang. "A fast cluster motif finding algorithm for ChIP-Seq data sets." BioMed research international 2015 (2015).
12. Sandelin, Albin, et al. "JASPAR: an open-access database for eukaryotic transcription factor binding profiles." Nucleic acids research 32.suppl 1 (2004): D91-D94.
13. Wingender, Edgar, et al. "TRANSFAC: a database on transcription factors and their DNA binding sites." Nucleic acids research 24.1 (1996): 238-241.
14. Zhu, Jian, and Michael Q. Zhang. "SCPD: a promoter database of the yeast Saccharomyces cerevisiae." Bioinformatics (Oxford, England) 15.7 (1999): 607-611.
15. Tomovic, Andrija, and Edward J. Oakeley. "Position dependencies in transcription factor binding sites." Bioinformatics 23.8 (2007): 933-941.
16. Sandve, Geir Kjetil, et al. "Improved benchmarks for computational motif discovery." BMC bioinformatics 8.1 (2007): 193.
17. Tompa, Martin, et al. "Assessing computational tools for the discovery of transcription factor binding sites." Nature biotechnology 23.1 (2005): 137-144.
18. Pevzner, Pavel A., and Sing-Hoi Sze. "Combinatorial approaches to finding subtle signals in DNA sequences." ISMB. Vol. 8. 2000.
19. Burset, Moises, and Roderic Guigo. "Evaluation of gene structure prediction programs." genomics 34.3 (1996): 353-367.
20. Zhang, Yipu, Ping Wang, and Maode Yan. "An Entropy-Based Position Projection Algorithm for Motif Discovery." BioMed Research International 2016 (2016).
21. Rajasekaran, Sanguthevar, Sudha Balla, and C-H. Huang. "Exact algorithms for planted motif problems." Journal of Computational Biology 12.8 (2005): 1117-1128.
22. Hendriks, Marleen E., et al. "Hypertension in sub-Saharan Africa: cross-sectional surveys in four rural and urban communities." PloS one 7.3 (2012): e32638.