Artificial Intelligence in Medicine 75 (2017) 16–23

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A high-order representation and classification method for transcription factor binding sites recognition in *Escherichia coli*☆

Shiquan Sun[a,b], Xiongpan Zhang[a], Qinke Peng[a,*]

[a] *Systems Engineering Institute, Xi'an Jiaotong University, 28 Xianning West Road, Xi'an, Shaanxi 710049, China*
[b] *Department of Biostatistics, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109, USA*

ARTICLE INFO

Article history:
Received 7 June 2016
Accepted 23 November 2016

Keywords:
Tensor
Partial least squares
Transcription factor binding sites
Machine learning
Classification
Computational biology

ABSTRACT

Background: Identifying transcription factors binding sites (TFBSs) plays an important role in understanding gene regulatory processes. The underlying mechanism of the specific binding for transcription factors (TFs) is still poorly understood. Previous machine learning-based approaches to identifying TFBSs commonly map a known TFBS to a one-dimensional vector using its physicochemical properties. However, when the dimension-sample rate is large (i.e., number of dimensions/number of samples), concatenating different physicochemical properties to a one-dimensional vector not only is likely to lose some structural information, but also poses significant challenges to recognition methods.
Materials and method: In this paper, we introduce a purely geometric representation method, tensor (also called multidimensional array), to represent TFs using their physicochemical properties. Accompanying the multidimensional array representation, we also develop a tensor-based recognition method, tensor partial least squares classifier (abbreviated as TPLSC). Intuitively, multidimensional arrays enable borrowing more information than one-dimensional arrays. The performance of each method is evaluated by average *F*-measure on 51 *Escherichia coli* TFs from RegulonDB database.
Results: In our first experiment, the results show that multiple nucleotide properties can obtain more power than dinucleotide properties. In the second experiment, the results demonstrate that our method can gain increased prediction power, roughly 33% improvements more than the best result from existing methods.
Conclusion: The representation method for TFs is an important step in TFBSs recognition. We illustrate the benefits of this representation on real data application via a series of experiments. This method can gain further insights into the mechanism of TF binding and be of great use for metabolic engineering applications.

1. Introduction

Transcription factors (TFs) are one of groups of proteins that bind to specific regions on the DNA sequence, thereby activating or repressing the rate of gene transcription [1,2]. In practical bioengineering applications, an effective method for identifying new TFBSs plays an important role in providing insights into cellular behavior, and helps us further understand the complex gene regulatory networks in cells [3,4].

Generally, the method for identifying TFBSs can be roughly divided into two categories: the experimental and computational approach. However, both categories are not mutually exclusive. Experimental methods can identify binding sites in some cases, such as DNase footprinting [5,6] and electrophoretic mobility shift assays [7,8]. However, due to the relatively short length and high degrees of degeneracy of such TFBSs, showing how the specificity of protein-DNA interactions is challenging. More specifically, with the advances in high-throughput sequencing technologies, the resolution is limited in hundreds of base-pairs (bps), and the procedure to identify TFBSs is still laborious and difficult in *in vivo* protein binding across the whole genome [9].

As supplement to the experimental method, the computational method not only identifies the real TFBSs in practice, but also provides useful instructions about the distribution of probes and potential binding sites. For example, in previous studies, consensus sequence and position-specific weight matrix (PWM) have been commonly used to model the sequence motifs [10–13]. In principle, these two methods can predict the binding sites via comparing

☆ Supported by China Natural Science Fund.
* Corresponding author.
E-mail addresses: shiquan_sun@126.com (S. Sun), panpan18022@163.com (X. Zhang), qkpeng@xjtu.edu.cn (Q. Peng).

http://dx.doi.org/10.1016/j.artmed.2016.11.004
0933-3657/

test sequences and consensus sequences. However, both methods result in a low identification rate because they both assume that the relationship between the nucleotide positions is independent. To address this issue, physicochemical properties (e.g., shape) are frequently introduced to help gain more information about the original DNA sequence [14–18]. To increase the prediction power, extensive studies leverage machine learning methods to train a prediction model, providing a promising way to identify TFBSs, such as support vector machine (SVM) [19,20,14], random forest (RF) [21,22], and deep learning [23].

Therefore, we can conclude that a well-performing method for identifying TFBSs mainly depends not only on a powerful prediction model but also a good representation method, which contains as much information about sequences as possible. However, there are several potential drawbacks when a DNA sequence is represented as a one-dimensional numeric vector. Theoretically, randomly permuting (or re-ranking) features do not affect the accuracy of the prediction model. In other words, the one-dimensional numeric vector and its corresponding DNA sequence do not necessarily have one-to-one correspondence, and the different binding sites might have the same distribution pattern after we re-rank the features, which contradicts with our original intention. On the other hand, the letter features (Section 2) will become useless if the identifying procedure incorporates a feature selection step. Because four features together represent one type of nucleobases, separating the four features becomes meaningless in practice. A promising way to deal with this issue is to use multidimensional array-based representation [24,25]. This type of representation has been successfully applied to EEG signals classification in biomedical engineering [26–28], image processing in computer vision or pattern recognition [29–31], and other fields [32–34].

In this paper, moving beyond the one-dimensional representation of TFBSs, we first represent a TFBS as a multidimensional array where the rows exhibit physicochemical properties of the DNA sequence, such as shear, stretch and shift, and the columns denote the different base pair steps (k-mers) within subsequent motifs. The elements in the multidimensional array indicate the value of physicochemical features with respect to k-mers. Accompanying the multidimensional array representation, we also develop a multidimensional array-based PLS classifier (TPLSC) to predict TFBSs. The experiments were conducted on 51 TFs in *Escherichia coli* from RegulonDB, and the results demonstrate that our method can significantly improve the recognition rate, especially for the integration host factor (IHF), which is well-known to exhibit both features specific to each base and DNA structural properties [35].

The rest of the paper is organized as follows: in Section 2, we illustrate the detailed process of multidimensional array-based representation for TFBSs. In Section 3, we discuss the standard partial least squares classifier and tensor partial least squares classifier together to demonstrate the relationship between two types of classifiers. The results are given in Section 4. Some concluding remarks are presented in Section 5.

2. Materials and TFBSs representation

In this section, we illustrate the detailed process of high-order representation for TFBSs. The real data sets confirmed by experiments can be downloaded from the RegulonDB v8.0 database (http://regulondb.ccg.unam.mx/ (accessed: 10.03.16)). This database collects the *E. coli* k-12 transcription information, and aims to build a comprehensive transcription regulation network [36]. In the current study, the real transcription factor binding sites were derived from the reference sequences (*E. coli* k-12 genome MG1655 (NCBI: NC_000913.3)), according to the starting position and the ending position which were from the RegulonDB database. To make comprehensive comparison, we randomly selected 1000 sequences from background genome sequences (non-coding sequences) as the negative samples to distinguish from the known TFBSs (positive samples).

Table 1
The combinations of different properties, and their corresponding values were collected from [19,37]. n is the number of binding sites for a specific TF, and the number n can be found in Fig. 4.

Combination	Description	Dimension
Di	All possible 2-mers properties	$n \times 111 \times 35$
DiL	All possible 2-mers properties and the letter features	$n \times 115 \times 35$
Mu	3-mers, 4-mers, and 7-mers properties	$n \times 70 \times 35$
MuL	3-mers, 4-mers and 7-mers properties, and the letter features	$n \times 74 \times 35$
DiMu	2-mers, 3-mers, 4-mers, and 7-mers properties	$n \times 181 \times 35$
DiMuL	2-mers, 3-mers, 4-mers and 7-mers properties, and the letter features	$n \times 185 \times 35$

Briefly, we summarized two ways to represent TFBSs from previous studies: base pair steps (e.g., 2-mer, 3-mer, and 7-mer), and geometrical parameters of base pairs (e.g., shear, stretch, and shift). In this paper, we focused on the physicochemical properties recorded as 2-mers to characterize the specific TFBSs, and the extended physicochemical properties recorded as 3-mers, 4-mers, and 7-mers from two recent studies [37,19]. For 2-mers, we collected all dinucleotide properties from DiProDB database (http://diprodb.fli-leibniz.de/ShowTable.php (accessed: 10.03.16)), and the total number of corresponding properties was 110. For k-mers (k = 3, 4, 7), all dinucleotide properties were collected from the Additional Materials in the paper [19] and the total number of corresponding properties for 3-mers was 62, 4-mers was 6, and 7-mers was 2. The papers have not provided the properties for 5-mers, 6-mers or other base pair steps; therefore, we left out these features in our study. Additionally, we also incorporated the letter features to provide the same information as used in PWM-based approaches. As described in previous studies [37,19], letter features were generated by designating the four kinds of nucleotides – A, C, G, and T – as mutually orthogonal 4D vectors (1,0,0,0), (0,1,0,0), (0,0,1,0), and (0,0,0,1), respectively.

We extended the length of all TFBSs with flanking nucleotides to 41 base pairs. As shown in the first step of Fig. 1, if we slide a subwindow from left to right on a 41 base pairs sequence, it will generate 35 features for 7-mers, 40 features for 2-mers, 39 features for 3-mers, and 38 features for 4-mers. To make a unified representation, we symmetrically discarded the nucleotides from both sides to ensure all k-mers with the same length (35). To clearly show the process of tensor representation, we take a binding site from AgaR TF as an example (Fig. 1), for 3-mers, we have 62 physicochemical properties and 35 features which form a 62×35 matrix, and the element $a_{1,1}$ in the matrix indicates the value of the physicochemical properties (such as 'shear') with respect to the first 3-mer feature, TTA; for 4-mers, 6 physicochemical properties and 35 features which form a 6×35 matrix; for 7-mers, 2 physicochemical properties and 35 features which form a 2×35 matrix. Then we simply concatenate the three matrices to form a tensor $\mathbf{X}^{(\mathbf{n})}$ ($1 \times 70 \times 35$). Assuming there are 11 binding sites for AgaR TF, therefore, we can obtain a third order tensor $\mathcal{X}$ in which the order is number of binding sites $\times$ number of physicochemical properties $\times$ Number of features ($11 \times 70 \times 35$). We did not illustrate the 2-mers (110×35 matrix) and the letter features (4×35 matrix) in Fig. 1. However, the process is similar to what we described above. The dimensionality of each tensor $\mathcal{X}$ is shown in Table 1.