Neurocomputing 199 (2016) 154–162

Contents lists available at ScienceDirect

Neurocomputing

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

Identification of DNA binding proteins using evolutionary profiles position specific scoring matrix

Muhammad Waris, Khurshid Ahmad, Muhammad Kabir, Maqsood Hayat*

Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan

ARTICLE INFO

Article history:
Received 24 August 2015
Received in revised form
12 February 2016
Accepted 7 March 2016
Communicated by L. Kurgan
Available online 6 April 2016

Keywords:
PSSM
SAAC
DPC
SVM
KNN
DNA-binding

ABSTRACT

DNA-binding plays a crucial role in different genomics processes including identification of specific nucleotides, regulation of transcription and regulation of gene expression. Various conventional methods have been used for identification of DNA-binding proteins. However, due to large explosion of protein sequences in databases, it is intricate or sometimes impossible to identify DNA-binding proteins. Therefore, it is intensively desired to establish an automated model for identification of DNA binding proteins. In this model, numerical attributes are extracted through Dipeptide composition, Split Amino Acid Composition, and position specific scoring matrix (PSSM). In order to overcome the issue of biasness and reduce true error, oversampling technique SMOTE was applied to balance the datasets. Several classification learners including *K*-nearest neighbor, Probability Neural Network, Support vector machine (SVM) and Random forest are utilized. Two benchmark datasets and jackknife test are applied to assess the performance of classification algorithms. Among various classification algorithms, SVM achieved the highest success rates in conjunction with PSSM feature space, which are 92.3% accuracy on dataset1 and 88.5% on dataset2. The empirical results revealed that our proposed model obtained the highest results so far in the literatures. It is anticipated that our proposed model might be useful and provides a substance for research and academia community.

1. Introduction

DNA binding plays a major role in formation of both the eukaryotic and prokaryotic proteins. In addition, it organizes and manages various biological processes including DNA packing, reflection, transcription regulation [1]. The researchers analyzed that 2–3% of prokaryotic and 6–7% of eukaryotic proteins can bind to DNA [2,3]. Nevertheless, it also performs an important role in finding out the potential therapeutics for genetic disorders and protein functions annotation. Therefore, the acknowledgment of DNA binding proteins has considered one of the most challenging tasks in annotation of protein functions. In this regards, most of the researchers have targeted discrimination of DNA binding and non-binding proteins. Initially, various conventional techniques, such as filter binding assays, genetics analysis, chromatin immunoprecipitation on microarrays and x-ray crystallography [4,5] have been carried out to handle this problem. These approaches have shown some reasonable performance but it is time consuming and laborious. Later, some computational methods were applied for identification of DNA binding and non-binding proteins. They have identified DNA binding and non-binding proteins on the combination of structure and sequence information [6]. However, the availability of recognized structures was limited. Owing to lack of recognized structures, few investigators have concentrated only on sequence information [7]. In computational model, the most important task is to effectively formulate the biological sequences. In this regards, various discrete methods, evolutionary profiles based methods and physicochemical properties based methods were used to extract numerical descriptors from biological sequences. Initially, researchers have widely applied amino acid composition (AAC) for protein sequences representation [8]. But AAC suffers the defect of losing sequence order information completely. In order to incorporate sequence order information Chou has introduced the concept of pseudo amino acid composition (PseAAC) [9,10]. PseAAC is the combination of AAC and correlation factors, which was extensively used by many investigators for their problems [11]. Lin et al. proposed iDNA-Prot by using random forest and grey model [5]. Xu et al. have developed a computational model for identification of DNA-binding proteins [12]. Evolutionary information and PseAAC were used as feature extraction schemes and SVM was utilized as classification algorithm. The performance of Xu et al. developed model was good but still there needed further improvement [12].

* Corresponding author. Tel./fax: +92 937 542194.
E-mail addresses: Maqsood.hayat@gmail.com, m.hayat@awkum.edu.pk (M. Hayat).

http://dx.doi.org/10.1016/j.neucom.2016.03.025
0925-2312/

Recently Liu et al. proposed iDNA-Prot|dis predictor for identification of DNA binding proteins by using amino acid distance pair and reduced alphabet composition methods [13]. PseDNA-Pro was developed by Liu et al. In their method they utilized physiochemical distance transformation and Chou's PseAAC for identification of DNA-binding proteins [14]. A new approach called iDNAPro-PseAAC was proposed by Liu et al. by incorporating PseAAC and PSSM approach for identification of DNA-binding proteins [15].

In this study, a quite promising computational mode has been proposed for identification of DNA binding and non-binding proteins. In this model, protein sequences are formulated into three discrete models, such as Dipeptide composition, Split Amino Acid Composition and evolutionary profiles based method position specific scoring matrix. The number of protein sequences in the datasets was imbalanced. However, in case of imbalanced dataset, mostly classification algorithms biased towards majority classes. In order to cover the issue of uneven numbers of protein classes, Synthetic minority over-sampling technique is used to balance the data. Several classification algorithms namely: *K*-nearest neighbor, probabilistic neural network, support vector machine, and random forest are thoroughly investigated to choose the best one among these classifiers. Cross validation jackknife test is applied to evaluate the performance of classifiers.

As demonstrated by a series of recent publications [16–20] in response to the call [21], to establish a really useful sequence-based statistical predictor for a biological system, we need to consider the following procedures: (a) construct or select a valid benchmark dataset to train and test the predictor; (b) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (c) introduce or develop a powerful algorithm (or engine) to operate the prediction; (d) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (e) establish a user-friendly web-server for the predictor that is accessible to the public.

The remaining paper is organized as follow: Section 2 represents materials and methods, Section 3 describes performance measures, Section 4 evaluates results and discussion and finally conclusion has been drawn in final section.

2. Materials and methods

2.1. Benchmark dataset

In order to develop an auspicious computational model, it is required to have a valid dataset for training the model. For this purpose, two benchmark datasets are selected. The formulation of benchmark dataset is given below

$$S = S^{+} \cup S^{-} \tag{1}$$

where subset S^{+} comprises 134 DNA binding protein sequences and 247 non-binding protein sequences. The symbol U represents the union of two subsets. This dataset was initially used by Berman et al. [22], who selected DNA binding protein samples from protein data bank [22], whereas non-binding protein samples were derived from [3]. Further, only those sequences were included which had less than 25% sequence identity [23]. In addition, those proteins were excluded whose length was less than 50 residues because they might belong to a fragment rather than a complete protein.

The second dataset called independent dataset is formulated as

$$S_{ind} = S_{ind}^{+} \cup S_{ind}^{-} \tag{2}$$

In this dataset, S_{ind}^{+} contains 81 DNA binding protein sequences, while S_{ind}^{-} comprises of 99 non-binding protein sequences. This dataset was constructed by Berman et al. from Swiss-port dataset [12,22]. The interesting fact about Independent dataset is that none of the proteins had neither more than 25% pairwise sequence identity with any other in the same subset, nor had more than 40% pairwise sequence identity with those in the Benchmark dataset.

2.2. Feature representation

Protein sequence formulation is one of the preliminary steps in the field of pattern recognition. In order to extract salient features from a dataset, the protein sequences are converted to numerical values because the statistical predictors use numerical descriptors for training and testing a model. One of the most important but also most challenging problem in computational biology and biomedicine is how to formulate a biological sequence with a discrete model or a vector. This is because all the existing operation engines, such as SVM (Support vector machine) and NN (Neural Network), can only handle vector but not sequence samples [24]. Therefore, a simple and the most common strategy which has been used many researchers for formulation of protein sequences is amino acid composition (AAC). However, in AAC the correlation between two residues is lost. Recently, Chou [25] proposed a powerful method called pseudo amino acid composition (PseAAC) which can not only describe the feature of amino acid composition, but also the long distance interaction between residues. This strategy has been widely applied in almost all the areas of computational proteomics [26–34]. In this paper, the authors incorporated Dipeptide composition, SAAC and PSSM into Chou's general PseAAC to predict DNA binding proteins. Most recently a new predictor called Pse-in-One was developed by Liu et al. to generate various modes of pseudo components of DNA, RNA and protein sequences [35].

2.2.1. Dipeptide composition

Dipeptide composition (DPC) is a feature extraction technique which works by incorporating sequence neighborhood information. In DPC each two adjacent amino acid residues occurrence frequency is computed. Subsequently, 400-D feature vector is extracted against each protein sequence. DPC is considered better than amino acid composition because it retains sequence order information into account. DPC is formulated as [36]:

$$\mathrm{DPC}(i) = \frac{\text{Total number of DP } (i)}{\text{Total number of DP}} \tag{3}$$

where DPC(i) represents the occurrence frequency of ith sequence. DP(i) represents a single instance out of 400 instances and DP is the total number of dipeptide instances.

2.2.2. Split Amino Acid Composition (SAAC)

Split Amino Acid Composition (SAAC) is a protein representation method in which protein sequence is divided into more than one different components and composition of each component is computed independently [37,38]. In case of our SAAC model the protein sequence is divided into three different components; 25 residues N-terminus, 25 residues C-terminus and the region between these two termini. Unlike conventional amino acid composition, the resultant feature vector is 60D.

2.2.3. Position specific scoring matrix

Position specific scoring matrix (PSSM) is an evolutionary profiles and patterns based feature extraction scheme. It exploits multiple alignments and information pertaining protein families. Initially, it was used for identification of distinct information belonging to proteins. Further, it is used to detect hidden and