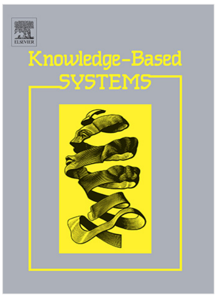
Accepted Manuscript



Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble

Wenjie You, Zijiang Yang, Guangbao Guo, Xiu-Feng Wan, Guoli Ji

PII: S0950-7051(18)30473-8
DOI: https://doi.org/10.1016/j.knosys.2018.09.023
Reference: KNOSYS 4507

To appear in: *Knowledge-Based Systems*

Received date : 7 May 2018
Revised date : 12 August 2018
Accepted date : 13 September 2018

Please cite this article as: W. You, et al., Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble, *Knowledge-Based Systems* (2018), https://doi.org/10.1016/j.knosys.2018.09.023

Prediction of DNA-binding proteins by interaction fusion feature representation and selective ensemble

Wenjie You[1,4], Zijiang Yang[2*], Guangbao Guo[3], Xiu-Feng Wan[4] and Guoli Ji[5*]

1. Fujian Provincial Key Laboratory of Nondestructive Testing, School of Electronic and Information Engineering, Fujian Normal University, Fuqing 350300, China
wenjie.you@hotmail.com

2. School of Information Technology, York University, Toronto M3J 1P3, Canada
zyang@mathstat.yorku.ca

3. Department of Statistics, Shandong University of Technology, Zibo 255000, China
ggb11111111@163.com

4. Department of Basic Sciences, College of Veterinary Medicine, Mississippi State University, Mississippi 39762, USA
wan@cvm.msstate.edu

5. Department of Automation, Xiamen University, Xiamen 361005, China
glji@xmu.edu.cn

*To whom correspondence should be addressed

Abstract

DNA-binding proteins play important roles in various cellular processes, and the identification of DNA-binding proteins is important for understanding and interpreting protein function. This manuscript presents algorithms for feature representation based on primary protein sequences and selective ensemble classification. We first propose a multi-source interaction fusion feature representation model that simultaneously considers interactions among physicochemical properties, evolutionary information, and gap distances between residues. We also provide a selective ensemble algorithm based on gap distances that yields differential base classifiers by selecting the feature subspaces. The selective ensemble algorithm improves the generalization ability of the integrated classifiers. We then compare the proposed algorithms with some state-of-the-art methods using multiple datasets. The experimental results show that the proposed algorithms are competitive and effectively identify DNA-binding proteins. The major contributions of the present study are the establishment of a model and algorithm for feature representation that involves interaction efforts and the development of a selective ensemble classification algorithm based on parameter perturbation. The proposed algorithms can also be applied to other biological questions related to amino acid sequences.

Keywords:

Combined fusion, Interaction fusion, Feature representation, Selective ensemble.

1 Introduction

In living cells, DNA-related activities occur with the aid of specific proteins and are regulated by protein-DNA interactions [1], and this type of regulation is achieved by specific or nonspecific binding between proteins and DNA. Proteins that bind to DNA and subsequently regulate DNA-related activities are called DNA-binding proteins, and these functional proteins in biological cells play a vital role in a variety of important life activities [2]. In addition, protein-DNA interactions play a key role in the genetic and evolutionary mechanisms of organisms. The investigation of protein-DNA interactions also forms the basis for human exploration and understanding of the mechanisms of activities such as growth, development, evolution and disease, and understanding protein-DNA interactions is essential for the functional interpretation of the proteome and the discovery of potential treatments for genetic diseases.

The application of traditional biological experimental techniques, such as filter-binding assays [3], chromatin immunoprecipitation with DNA microarrays (ChIP-chip) [4] and X-ray crystallography [5], allows the accurate identification of DNA-binding proteins. However, the determination of protein structure and function using biological methods requires extensive material and financial resources and is time-consuming and laborious. With the rapid development of protein sequencing technology, the amount of protein sequence data is increasing. Thus, the field of proteomics requires the use of more effective and reliable computational methods for the analysis of biological sequences, and these methods comprise one of the most important topics in the field of proteomics research.

There are two general categories of DNA-binding protein prediction methods that are based on machine learning: structure-based prediction methods [6,7] and sequence-based prediction methods [8-12]. The structure-based prediction of DNA-binding proteins can achieve higher recognition rates, but these methods cannot be widely used for the interpretation of high-throughput sequences due to the lack of sufficient information on protein structure. Most current methods predict protein function based on the amino acid sequence because many experiments have shown that if the primary structures (the sequence of amino acid residues) of polypeptides or proteins are similar, the spatial conformations of the polypeptides after folding and their functions are also very similar [13]. The sequence-based protein function prediction method consists of two main processes: (1) extraction of biological information contained in the protein sequence and transformation of the protein sequence into a corresponding numerical feature vector and (2) usage of a machine learning algorithm to train a model with the resulting numerical feature vectors and prediction of the query sequences.

Feature vector representation, which is also referred to as feature representation, involves the generation of numerical feature vectors based on protein sequences, i.e., it involves the conversion of original sequence data into numerical feature vectors for classification. During the past few decades, effective sequence-based feature representation methods have been developed, and these include (1) amino acid composition (AAC)-based methods [14,15], which consider the information encoded by adjacent and continuous amino acid residues, (2) pseudo-amino acid composition (PseAAC)-based methods [16,17], which consider the information encoded by non-adjacent (discontinuous) amino acid residues, and (3) sequence profile-based methods [18], which consider evolutionary information of proteins. AAC-based methods, such as the commonly used k-mer frequency approach, use statistical information on the sequence [19]. These methods are simple, but their generated feature dimensionality is high (20^k, k is the sliding window length), leading to over-fitting. The PseAAC-based methods, which were proposed by Kuo-chen Chou [16], consider both the local order and the global order of a sequence to better represent the order and position information within the sequence. These methods can map the position information of a sequence into the generated feature vector. Sequence-profile-based methods use a position-specific scoring matrix (PSSM) with evolutionary information representing homology information related to aligned sequences. Many applications based on PSSMs demonstrate that PSSMs with evolutionary information contain more important and relevant information than protein sequences alone [10,20-22]. The sequence profile-based methods usually have better predictive ability than other methods and are widely used for protein prediction [22].

Previous studies have shown that evolutionary information, physicochemical properties and sequence structural and locational information all play roles in the identification of DNA-binding proteins [14,15,18]. When a single method, such as an AAC information- or a sequence profile-based method, is used, the resulting numerical features are overly monotonous. Currently, the mainstream approach used in the literature is to consider different properties (such as the different physicochemical properties of proteins) and information (such as evolutionary and structural information). The feature vectors generated by these methods [23,24] are then combined, and the resulting high-dimensional feature vectors are fed into a classifier. We refer to this type of explicit feature representation as combined fusion feature representation (CFFR), which integrates the physicochemical properties of amino acids, the evolutionary information extracted from sequence profiles, and the information inherent to the sequence (information on adjacent and non-adjacent residues) to achieve better prediction performance [23,25].

Different machine learning algorithms are widely used in feature spaces generated by different feature representation methods to further improve the ability to predict DNA-binding proteins, such as support vector machine (SVM), neural network, K-nearest neighbors and random forest. In recent years, ensemble learning technology has also received extensive attention in the field of pattern recognition and bioinformatics. Ensemble learning [26] refers to learning from training samples to build a number of differential learning models (called the base classifier) and then employing a specific strategy to combine these base classifiers to solve a single learning task. In selective ensemble learning [27], an additional stage of pruning or selection of base classifiers is included between the first stage (base classifier construction) and the second stage (classifier combination) of ensemble learning, and this additional stage aims to select a subset of base classifiers that show large differences and exert a good effect. At present, an intuitive method for generating a selective ensemble involves sorting the base classifiers to achieve the purpose of pruning and integration [28-31].

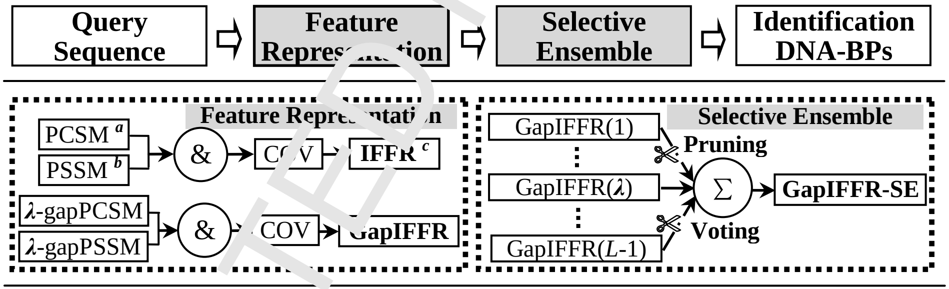
For the identification of DNA-binding proteins, the current mainstream machine learning methods are usually combined with feature representation and classification algorithms. Liu *et al*. [32] used a reduced-alphabet method to reduce the dimension of the PseAAC vector (named iDNA-Prot|dis), and it can accelerate the computational time of the Cai's algorithm. Later, they combined PseAAC with physicochemical distance conversion (named PseDNA-Pro) [33]. The results indicated that the proposed method can further improve the predictive ability from PseAAC vector. Lin *et al*. [12] proposed iDNA-Prot predictor by incorporating the features into the general form of PseAAC that were extracted from protein sequences via the grey model and by adopting the random forest operation engine. Kumar *et al*. [10] incorporated evolutionary information into sequence-based methods. They combined the evolutionary and sequential features into a SVM predictor called DNAbinder. The evolutionary features significantly improved the predictive accuracy [10], suggesting that the evolutionary information is important for distinguishing DNA-binding proteins from non-DNA-binding proteins. Similar results were reported by Ho *et al*. [20]. Later, Kumar *et al*. [34] employed random forest method, named DNA-Prot, to identify DNA binding proteins from protein sequence. They compared DNA-Prot method with DNAbinder method on three benchmark datasets. The results have shown that DNA-Prot achieves better performance. Liu *et al*. [22] proposed a new method for DNA-binding protein prediction called iDNAPro-PseAAC, which integrates the profile-based representation of the evolutionary information retrieved by PSI-BLAST into the classical PseAAC, and they found that negative samples in the training model improved the predictive performance. Dong *et al*. [19] combined SVM and the auto-cross covariance transformation. The protein sequence represented in the form of amino acids or physicochemical properties of amino acids are converted into a series of fixed-length vectors by Kmer composition and the auto-cross covariance transformation. Wei *et al*. [25] established a novel predictor named Local-DPP, which combines the local Pseudo PSSM (Pse-PSSM) features with random forest classifier. The generated features can efficiently capture the local conservation information, together with the sequence-order information. Experiments have shown that Local-DPP significantly improved the accuracy of the existing predictors.

For the identification of DNA-binding proteins, the development of an efficient feature representation method that can generate features with discriminant information from a sequence and then accurately identify and classify DNA-binding proteins has important significance for informatics and biology. In this paper, a multi-source fusion feature representation method that takes into account physicochemical properties, evolutionary information and relative position information between residues and considers their interaction effects is proposed. The proposed algorithm can generate features with strong discriminative ability and improve the prediction of DNA-binding proteins. The features generated by the algorithm help us understand the functions and roles of DNA-binding proteins from the perspective of interactions. Subsequently, we perturb the parameters of the proposed feature representation algorithm to generate multiple base classifiers and obtain differential classifiers via selection (pruning) to further improve the overall recognition performance of the ensemble classifier. Experimentally, our interaction fusion feature representation (IFFR) yields improved recognition compared with traditional CFFR. Moreover, the use of selective ensembles based on parameter perturbation significantly improves the identification of DNA-binding proteins compared with other state-of-the-art prediction methods. Furthermore, from the perspective of protein interactions, the proposed feature representation helps us understand the functions and roles of DNA-binding proteins in cellular processes.

The remainder of the paper is organized as follows. Section 2 discusses the DNA-binding protein prediction method, including IFFR and its selective ensemble algorithm. In Section 3, DNA-binding proteins in multiple protein sequence datasets are identified, and comparisons with multiple classical prediction methods are provided. We then present conclusions and discuss future work in Section 4.

2 Methodology

In the practical application of machine learning, it is generally believed that data and features determine the upper limit of learning performance and that models and algorithms can approximate this upper limit [35]. Therefore, we simultaneously pursued two goals: (1) the generation of features with strong discriminative ability through effective integration of a variety of types of information and 2) the generation of a classification model with strong generalization ability using selective ensembles of multiple classifiers. Figure 1 shows the framework of our prediction model, which consists of two key components: (1) *Feature representation process*: For an amino acid sequence of any length, the two scoring matrices, namely, PCSM and PSSM, which express physicochemical properties and evolutionary information, are given (Definition 1), and the two matrices are then combined by columns. A covariance operation is subsequently performed on the merged matrix (Definition 3) to obtain a feature vector with dual source interaction fusion (IFFR). Similarly, based on the fusion of interactions with a dual source, information on the gap distances between residues is introduced (Definition 2) to realize triple-source fusion feature representation (GapIFFR). (2) *Selective ensemble process*: The parameter λ of the feature representation algorithm GapIFFR is perturbed to generate different feature subspaces, and the different input subspaces are then subjected to selection (or pruning) to obtain a subset of differential base classifiers. The optimal selection of base classifiers is then identified to achieve selective ensemble based on gap distance (GapIFFR-SE).



[a] PCSM = physicochemical scoring matrix,
[b] PSSM = position-specific scoring matrix,
[c] IFFR = interaction fusion feature representation.

Fig. 1. DNA-binding protein prediction model framework: interaction fusion feature representation (left dashed box) and selective ensemble for classification (right dashed box)

2.1 Hypothesis

Considering and using appropriate physicochemical properties and evolutionary information is key to identifying DNA-binding proteins based on their protein sequence. CFFR, which is commonly used, considers physicochemical properties to some extent as well as evolutionary information, local position information and other features of a protein, and its use can therefore enhance the ability to identify DNA-binding proteins. However, the CFFR-based method treats physicochemical properties and evolutionary information independently and thus ignores the existence of interaction effects between various properties and evolutionary information. As a result, the features generated by CFFR carry only the explicit features of each information source itself and ignore the implicit features generated by the interaction between different information sources. In fact, based on the biochemical reactions that occur in cells, living cells have many interactions, such as interactions between proteins and interactions between amino acid residues. This paper focuses on multi-source fusion feature representation with interaction effects and examines whether interactions between different properties (physicochemical properties, etc.) and information (protein evolution information, etc.) occur and, if so, whether these interactions can improve the recognition of DNA-binding proteins. We propose the following hypotheses:

Hypothesis 1: Interaction effects exist between physicochemical properties and evolutionary information.

In this paper, we consider feature representation with interactions between physicochemical properties and evolutionary information, termed IFFR (Interaction Fusion Feature Representation). Within families of protein sequences, amino acid substitution patterns are highly specific, and there are also interactions between amino acid residues at different positions in the same protein sequence. We propose an IFFR based on different gap distances; that is, based on the fusion of interactions from two sources (physicochemical properties and evolutionary information), a gap operation with different distances (λ-gap) is introduced to achieve a triple-source fusion feature representation algorithm GapIFFR.

Multi-source fusion is an effective information processing technology. From an information theory point of view, it can (at least ideally) improve the specificity and comprehensiveness of our understanding of an entity [36]. For example, for multi-source data, by aggregating the results obtained from each single source of data, the literature [37] establishes an evaluation function for inducing three-way decision making and performing three-way concept learning, and numerical experiments have shown the effectiveness of the proposed method. Therefore, drawing on concepts from the relevant literature, we propose the following hypothesis:

Hypothesis 2: Within the IFFR framework, triple-source fusion GapIFFR is better than dual-source fusion GapPSSM.

The essence of the triple-source IFFR, GapIFFR, is the feature interaction fusion of physicochemical properties and evolutionary information with the addition of gap information at different distances. This algorithm simultaneously considers physicochemical properties, evolutionary information, local sequence location and other information from protein sequences.

2.2 Models

The feature representation process digitizes a sequence composed of characters into a fixed-dimensional feature vector based on mathematical relationships within the sequence, biochemical properties and other indicators. The generated feature vector can include both explicit and implicit features. For the feature representation of protein sequences, this section provides a new IFFR model that can consider both the internal correlations of various information (explicit features) and the interaction effects between different types of information (implicit features). A related conceptual description is given first, and from this description, a multi-source fusion feature representation model with interaction effects is derived.

Definition 1 (Scoring Matrix: SM)

Given any (protein) sequence $S = R_1, R_2, \cdots, R_L$, the scoring matrix is defined as

$$P = (p_{ij})_{L \times M} \tag{1}$$

where $p_{ij}\ (i = 1, 2, \cdots L)$ is the score of the *i*-th amino acid residue R_i on the *j*-th index, *L* is the length of sequence *S*, and *M* is the pre-determined number of indicators.

Protein sequence analysis often uses an SM, such as a PSSM, which is a matrix with *L* rows (*L* is the sequence length) and 20 columns (20 standard amino acids). The protein dataset search program PSI-BLAST can find an optimal result through multiple iterations and is useful for identifying new members of a protein family or detecting similar proteins in distantly related species [38,39]. The use of PSI-BLAST can generate a PSSM:

$$P=\begin{bmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,20} \\ \vdots & \vdots & & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,20} \end{bmatrix}_{L\times 20} \qquad (2)$$

In (2), the element p_{ij} represents the probability (log-likelihood score) of the amino acid residue R_i at the *i*-th position ($1 \le i \le L$) of a sequence mutating to the *j*-th class ($1 \le j \le 20$) amino acid during the process of protein evolution, and greater values indicate a greater likelihood of substitution. A PSSM expresses the evolutionary information of the sequence, and the detailed procedure for calculating a PSSM is given in Appendix 1.

We also use a SM for amino acid physicochemical properties (Physicochemical Scoring Matrix: PCSM). In the process of identifying DNA-binding proteins, we assume that the different physiochemical properties of amino acids will make different contributions to the predicted results. Therefore, we should consider appropriate amino acid physiochemical properties in the feature representation process. An amino acid index (AAindex) is a set of 20 numerical values representing any of the different physicochemical and biological properties of amino acids. Specifically, the AAindex is a database of numerical indices representing various physicochemical and biochemical properties of amino acids, and the AAindex1 section is a collection of published indices that currently contains 566 indices. For the *j*-th physicochemical property Q_j, any protein sequence S can be expressed as $q_{1,j}, q_{2,j}, \cdots, q_{L,j}$, where L is the sequence length, and q_{ij} $(1 \le i \le L)$ is the *j*-th physicochemical property index of the *i*-th amino acid residue R_i in the sequence. Assuming that M types of physicochemical properties exist, the PCSM for the protein sequence is as follows:

$$Q=\begin{bmatrix} q_{1,1} & q_{1,2} & \cdots & q_{1,M} \\ q_{2,1} & q_{2,2} & \cdots & q_{2,M} \\ \vdots & \vdots & & \vdots \\ q_{L,1} & q_{L,2} & \cdots & q_{L,\mathrm{M}} \end{bmatrix}_{L\times M} \qquad (3)$$

In the experimental section, to ensure the fairness of the comparison results, we use only the six physicochemical properties listed in the literature [40]: (1) hydrophobicity, (2) hydrophilicity, (3) mass, (4) pK1 (α-COOH), (5) pK2 (NH3), and

(6) pI (25°C). The detailed physicochemical indices of the amino acids used in this paper are given in Appendix 2.

Two amino acids located far from each other in the amino acid sequence might be spatially close to each other and even in contact after the protein polypeptide chain is folded. In three-dimensional space, each residue has its own specific space coordinates, and a result, the Euclidean distance between two residues, also known as spatial distance, can be obtained. If the Euclidean distance between two residues (between C$_\beta$ atoms) is less than 8 Å, the residues are biologically considered to be in *contact*. This interaction between residues (i.e., contact) has a huge impact on protein structure and function. Therefore, considering the interactions between amino acid residues located at different distances in the protein sequence based on an analysis of pseudo-amino acid composition and the drawing on the idea of pseudo-amino acid composition analysis [16], the definition of the λ-gap SM (λ-gapSM) is given.

Definition 2 (λ-gapSM).

Given an SM $P=(p_{ij})_{L\times M}$ and parameter λ, the λ-gap scoring matrix is a $(L-\lambda)\times M$ matrix G_λ, which is defined as follows:

$$G_\lambda = A_\lambda P = A_\lambda \begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_L \end{pmatrix} = \begin{pmatrix} p_1 + p_{\lambda+1} \\ p_2 + p_{\lambda+2} \\ \vdots \\ p_{L-\lambda} + p_L \end{pmatrix} \tag{4}$$

where $A_\lambda = (a_{ij})_{(L-\lambda)\times L}$ is the (0-1) matrix, $a_{ij} \in \{0,1\}$, i.e.,

$$A_\lambda = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_{L-\lambda} \end{pmatrix} = \begin{pmatrix} \underleftrightarrow{1,0,\cdots,1},0 & \cdots & 0,0,\cdots,0,0 \\ 0,\underleftrightarrow{1,0,\cdots,1}, & \cdots & 0,0,\cdots,0,0 \\ \vdots & \cdots & \vdots \\ 0,0,\cdots,0,0 & \cdots & 0,\underleftrightarrow{1,\cdots,0,1} \end{pmatrix} \tag{5}$$

λ $(0 \le \lambda \le L\text{-}1)$ represents the distance between two nonzero elements a_{ki} and a_{kj} in any row vector a_k of matrix A_λ, i.e., $\lambda = |j-i|$. In particular, if $\lambda = 0$, A_0 degenerates into the identity matrix I_L, that is, a 0-gapSM,

$$G_0 = A_0 P = I_L P = P \tag{6}$$

The λ-gapSM indirectly represents information on the relative positions of residues in a sequence, and λ represents the gap distance between any two residues, i.e., relative to the spatial distance, which is referred to herein as the sequence-based gap distance. In particular, if λ is equal to 0, the λ-gapSM does not consider information between residues; if λ is equal to 1, information between adjacent residues is considered; and if λ is greater than 1, information between non-adjacent residues is considered.

Definition 3 (Covariance SM: CovSM)

Given a λ-gapSM $G_\lambda = (g_{ij})_{(L-\lambda)\times M}$, a covariance matrix of the λ-gapSM is defined as follows:

$$\Sigma = Cov(G_\lambda) = G_\lambda^T G_\lambda = (\sigma_{ij})_{M\times M} \tag{7}$$

It follows that Σ is a symmetric matrix.

Suppose that U is an upper triangular matrix corresponding to the symmetric matrix $\Sigma = (\sigma_{ij})_{M\times M}$, i.e.,

$$U = \begin{pmatrix} \sigma_{1,1} & \sigma_{1,2} & \cdots & \sigma_{1,M} \\ & \sigma_{2,2} & \cdots & \sigma_{2,M} \\ & & & \vdots \\ & & & \sigma_{M,M} \end{pmatrix} \tag{8}$$

The matrix “vec” operator is applied to U by a column vector, and the elements σ_{ij} that satisfy $i \le j$ are retained; as a result, the “vec” operator transforms a matrix into a column vector by stacking the columns of the matrix. Thus, we derive the following remark.

Remark 1

Given any protein sequence $S = R_1, R_2, \cdots, R_L$ and gap distance λ, it is easy to derive a feature vector,

$$v = vec(U) = (\sigma_{1,1}, \sigma_{1,2}, \sigma_{2,2}, \cdots, \sigma_{1,M}, \sigma_{2,M}, \cdots, \sigma_{M,M})^T \tag{9}$$

Obviously, the dimension of this vector is ($M(M+1)/2$), and this dimension is only related to M and is independent of L (sequence length) and λ (gap distance).

The significance of Remark 1 lies in the generated vector, which contains not only the relevant information (such as length) of the original sequence but also the spacing information for adjacent and non-adjacent residues. However, the dimension of the resulting feature vector does not depend on the length of the sequence or gap distances.

A mathematical model of our proposed feature representation is presented below.

For the scoring matrices PSSM and PCSM, the corresponding scoring matrices λ-gapPSSM and λ-gapPCSM can be obtained by Definition 2. Given any protein sequence of length L, there are the PSSM matrix P and the PCSM matrix Q, and by horizontally concatenating P and Q, the matrix $W=(P,Q)=(w_{ij})_{L\times(M+20)}$ can be obtained. Based on Definition 2, the λ-gapSM can be obtained as follows:

$$G_\lambda = A_\lambda W = A_\lambda(P,Q) = \left(A_\lambda P, A_\lambda Q\right) \tag{10}$$

According to Definition 3 and block matrix operations, it is easy to obtain

$$\begin{aligned}\Sigma &= Cov(G_\lambda) = (A_\lambda P, A_\lambda Q)^T (A_\lambda P, A_\lambda Q) \\ &= \begin{pmatrix} P^T A_\lambda^T \\ Q^T A_\lambda^T \end{pmatrix} (A_\lambda P, A_\lambda Q) = \begin{pmatrix} P^T A_\lambda^T A_\lambda P & P^T A_\lambda^T A_\lambda Q \\ Q^T A_\lambda^T A_\lambda P & Q^T A_\lambda^T A_\lambda Q \end{pmatrix}_{(M+20)\times(M+20)}\end{aligned} \tag{11}$$

From Remark 1, the dimension of the feature vector corresponding to Equation (11) is related only to M and is independent of the sequence length L and the parameter λ.

In the abovementioned feature representation model, we use the correlation information $Q^T A_\lambda^T A_\lambda Q$ and $P^T A_\lambda^T A_\lambda P$ contained in the physicochemical property matrix Q and the evolutionary information matrix P, respectively, to generate explicit features. To generate an implicit feature, we also consider the interaction effect term $Q^T A_\lambda^T A_\lambda P$ (or $P^T A_\lambda^T A_\lambda Q$) between the physicochemical property matrix and the evolutionary information matrix. Here, $A_\lambda^T A_\lambda$ depicts the location information of adjacent and non-adjacent residues (distance equal to λ). In particular, if $\lambda = 0$ ($A_0 = I$), Equation (11) can be simplified to

$$\Sigma = (P,Q)^T (P,Q) = \begin{pmatrix} P^T \\ Q^T \end{pmatrix} \begin{pmatrix} P & Q \end{pmatrix} = \begin{pmatrix} P^T P & P^T Q \\ Q^T P & Q^T Q \end{pmatrix}_{(M+20)\times(M+20)} \tag{12}$$

Table 1
Feature representation methods developed within the proposed model framework.

Scoring Matrix	Matrix Dimension	CovSM (**Def. 3**)	Vector Dimension **(Remark 1)**	Feature Representation
SM (**Def. 1**)	L×M	M×M	M(1+M)/2	
PCSM	L×6	6×6	21	CovPCSM
PSSM	L×20	20×20	210	CovPSSM
			231 [b]	CFFR
	L×26 [a]	26×26	351	IFFR
λ-gapSM (**Def. 2**)	(L-λ)×M	M×M	M(1+M)/2	
λ-gapPCSM	(L-λ)×6	6×6	21	GapPCSM
λ-gapPSSM	(L-λ)×20	20×20	210	GapPSSM
			231 [b]	GapCFFR
	(L-λ)×26 [a]	26×26	351	GapIFFR

[a] Horizontal concatenation of the above two matrices;

[b] Tandem combination of the above two vectors.

Table 1 summarizes the relevant information on the different feature representation methods in the framework of our model, including the dimension of the SM, the dimension of the generated feature vector, and the abbreviation of the feature representation methods. Among these methods, CovPCSM considers six different physicochemical properties and their own internal correlations, and the dimension of the generated feature vector is 21. CovPSSM considers the evolutionary information of a sequence and their correlations for all 20 amino acids, and the feature dimension is 210. The CFFR method conducts a simple tandem combination of the previous two methods; the resulting feature dimension is the sum of the feature dimensions of the individual methods and is equal to 231. IFFR considers not only the correlations within the six physicochemical properties studied and within evolutionary information but also the interactions between physicochemical properties and evolutionary information; thus, the generated feature dimension is equal to 351. In this paper, we also consider the relative location information for residues and propose a multi-source fusion feature representation method with the gap distance *λ*. These feature representations are named GapPSSM, GapCFFR and GapIFFR. The proposed mathematical models and algorithms are universal, and the analytical methods discussed in this paper can be applied to other biological questions related to amino acid sequences.

2.3 Algorithms

For the DNA-binding protein prediction problem, this section provides the feature representation GapIFFR algorithm (Algorithm 1) and the selective ensemble classification algorithm (Algorithm 2).

1) Gap-based IFFR algorithm

Based on the proposed feature representation model, a new multi-source feature representation algorithm, GapIFFR, is proposed. This algorithm considers the interaction effects among specific physicochemical properties, evolutionary information, and location information between (adjacent and non-adjacent) amino acid residues. The detailed algorithm is as follows:

Algorithm 1 Gap-based Interaction Fusion Feature Representation (**GapIFFR**)

```
Input:    seq_FASTA  // Query protein sequence
          $\lambda$          // Distance of gaps
Output:   v          // Numeric vector
1:  Initialization: L = length of sequence seq_FASTA,  $\lambda \le L$-1
2:  Obtain PSSM matrix $P$ by calling PSI-BLAST (Set evalue=0.001
     num_iterations=3):  $P = (p_{ij})_{L\times 20}$
3:  Obtain PCSM matrix $Q$ from AAindex dataset:  $Q = (q_{ij})_{L\times M}$
4:  Horizontally concatenate $P$ and $Q$:  $W = [P, Q] = (w_{ij})_{L\times(20+M)}$
5:  Compute matrix $G_\lambda$ in term of Definition2:
            $G_\lambda = A_\lambda W = (g_{ij})_{(L-\lambda)\times(20+M)}$
6:  Compute matrix $\sum$ in term of Definition3:
            $\sum = \mathrm{cov}(G_\lambda) = G_\lambda^T G_\lambda = (\sigma_{ij})_{(20+M)\times(20+M)}$
7:  Return a column vector v in term of Remark1:
            $v = (\sigma_{1,1}, \sigma_{1,2}, \sigma_{2,2}, \cdots, \sigma_{1,20+M}, \sigma_{2,20+M}, \cdots, \sigma_{20+M,20+M})$
```

In particular, if *W* = *P* (ignoring *Q*) in Step 4 of Algorithm 1, the algorithm GapPSSM is produced; similarly, if *W* = *Q* (ignoring *P*) in Step 4 of Algorithm 1, the algorithm GapPCSM is produced. The feature vector returned by the GapCFFR algorithm is the combination of the feature vectors generated by these two algorithms (details given in Table 1). The input parameter *λ* of Algorithm 1 is the gap distance between residues. If *λ* = 0, the above feature representation algorithm considers only physicochemical properties and evolutionary information, and Algorithm 1 degenerates into dual-source CFFR. Similarly, if *λ* = 0 and *W* = *P* in Step 4 of Algorithm 1, the algorithm degenerates into a feature representation based only on evolutionary information, denoted CovPSSM; if *λ* = 0 and *W* = *Q* in Step 4 of Algorithm 1, the algorithm degenerates into a feature representation based only on physicochemical properties, denoted CovPCSM. The CFFR algorithm is a tandem combination of the two feature vectors generated by CovPSSM and CovPCSM (details are given in Table 1). If number of sequences in the NR database is represented by N_library_seq and the average length of the sequence is *L*, the complexity of Algorithm 1 is $O(\text{N\_library\_seq}\cdot L^2)$. The main advantages of Algorithm 1 are that it can be applied to amino acid sequences of any length and that it considers the interactions among multiple sources of information. As a result, it can mine potentially useful biological information hidden in protein sequences and generate features with strong discriminating abilities.

2) GapIFFR-based selective ensemble algorithm

Given a set of protein sequences, a training set S_{trn}, a validation set S_{val} and a test set S_{tst} are randomly divided. If $D_{trn}^{(\lambda)} = \{(\mathrm{x}_i^{(\lambda)}, y_i)\}$ is the training set corresponding to S_{trn}, an input variable $\mathrm{x}_i^{(\lambda)} = (x_{i1}^{(\lambda)}, x_{i2}^{(\lambda)}, \cdots, x_{ip}^{(\lambda)}) \in \Re^p$ within the training sample $(\mathrm{x}_i^{(\lambda)}, y_i)$ is the p-dimensional feature vector obtained by Algorithm 1 with gap distance λ, and the output variable is $y_i \in Y = \{+1, -1\}$. Similarly, the validation set $D_{val}^{(\lambda)}$ and the test set $D_{tst}^{(\lambda)}$ can be obtained. The base classifier C_λ can be trained on $D_{trn}^{(\lambda)}$ $(1 \le \lambda \le L-1)$ to obtain a set of $T = \{C_1, C_2, \cdots, C_{L-1}\}$, where $\bar{T}$ is any subset of *T*. The validation error $\varepsilon(\bar{T})$ of the ensemble classifiers corresponding to subset $\bar{T}$ in the validation set $D_{val}^{(\lambda)}$ can be calculated, and the subset $T^* = \arg\min_{\bar{T} \subset T} \varepsilon(\bar{T})$ with the smallest validation error is selected.

The optimal base classifier subset T^* can be obtained by an exhaustive search. However, if *L* is large, the associated computation would be excessive. A simple and intuitive selection strategy is to sort the base classifier C_i by the performance index *M*, select a subset T^* consisting of the top k (odd) base classifiers, as is performed when pruning ensemble classifiers, and then vote on the subset $T^* \subset T$ using the Max-Wins Voting (MWV) strategy [41]. The following depicts the selective ensemble algorithm based on GapIFFR:

```
Algorithm 2 GapIFFR-based Selective Ensemble (GapIFFR-SE)
Input:   $S_{trn}$, $S_{val}$, $S_{tst}$, $C$, $M$, $k$      // $C$ is a base classifier algorithm,
         // $M$ is the evaluation criteria (such as Accuracy, MCC, etc.)
Output:  $Y$   // class label of the test dataset $S_{tst}$.
(1) Initialization process:
    —Set $T=\Phi$, $L$=minimum sequence length of $S_{trn}$, $S_{val}$ and $S_{tst}$, calculate
      $D_{trn}(\lambda)$, $D_{val}(\lambda)$ and $D_{tst}(\lambda)$ by calling GapIFFR with $\lambda=\{1,2,\ldots,L-1\}$
(2) Training base classifiers process:
    —For $i$=1 to $L$-1 do
    ——Update $T=T\cup C_i$, where the base classifier $C_i$ is trained on the training
        dataset $D_{trn}(i)$ using the given classifier $C$.
    —EndFor
(3) Selection (Pruning) process:
    —For $j$=1 to $L$-1 do
    ——Calculate $M_j$ for each base classifier $C_j\in T$ on the validation dataset
        $D_{val}(j)$ using the evaluation criteria $M$.
    —EndFor
    —Sort $M_j$ in descending order, and select $T^*=\{C_{\lambda 1},C_{\lambda 2},\ldots,C_{\lambda k}\}\subset T$,
        where $C_{\lambda 1},C_{\lambda 2},\ldots,C_{\lambda k}$ correspond to the top $k$ of the $M_j$ values.
(4) Ensemble (Voting) process:
    —Predict the class label of the test dataset $S_{tst}$,
            $Y = sign\{\sum_{t=1}^{k} C_{\lambda_t}(\mathrm{x})\}$
      where $C_{\lambda t}$ is the $\lambda_t$-th base predictor on the dataset $\mathrm{x}\in D_{tst}(\lambda_t)$,
      $\{\lambda_1,\lambda_2,\ldots,\lambda_k\}\subset\{1,2,\ldots,L-1\}$.
Return $Y$
```

Algorithm 2 is a GapIFFR-based selective ensemble, GapIFFR-SE, that essentially perturbs the parameter λ to generate different input feature subspaces and then uses the strategy of selection (pruning) to obtain a subset of differential base classifiers and thereby improve the performance of the integrated classifier. The time complexity of Algorithm 2 is $O(\text{n\_seq}\cdot\text{N\_library\_seq}\cdot L^3)$, where *N_library_seq* is the number of sequences in the NR database and L is the average length of the sequence.

3 Experiments

3.1 Experimental datasets and evaluation indicators

To verify the effectiveness of the proposed method, six sequence datasets of DNA-binding proteins (including one group of independent testing sets) are selected for analysis. Their sample sizes are relatively large (≥300), and these datasets have sequence homologies less than 40%, guaranteeing the relative credibility of the experimental results. Table 2 provides a summary of the datasets used and lists their sources[1].

Table 2
Summary of datasets

Dataset	Number of Proteins			Min Length	Similarity
	DNA-BP	non-DNA-BP	Total		
Alternate Dataset [10]	1153	1153	2306	51	≤25%
PDB1075 Dataset [32]	525	550	1075	50	≤25%
Independent 1 Dataset [34]	823	823	1646	35	≤40%
Independent 2 Dataset [34]	88	233	321	30	≤40%
Training Dataset [10]	146	250	396	26	≤25%
Testing Dataset [42]	92	100	192	45	≤25%

To objectively and systematically evaluate the predictive performance of the proposed method, the Jackknife validation method, k-fold cross validation (k-foldCV) and the HoldOut method are used to compare and evaluate the algorithms proposed in this paper. k-foldCV can effectively reduce the over-learning and under-learning states caused by insufficient data. In practice, 10-fold CV is considered a standard method. The Jackknife validation method is considered a more objective statistical test; it can avoid randomness due to random division of the training and test data, thereby ensuring the reproducibility of the experimental results. The HoldOut method can determine the predictive ability of the algorithm for fresh samples (independent test sets).

The evaluation indices used for algorithm performance are accuracy (ACC), sensitivity (SE), specificity (SP) and the Matthews Correlation Coefficient (MCC), which are defined below

$$ACC = \frac{TP + TN}{TP + TN + FN + FP} \times 100\%$$

$$SE = \frac{TP}{TP + FN} \times 100\%$$

[1] http://www.imtech.res.in/raghava/dnabinder/download.html
http://server.malab.cn/Local-DPP/Datasets.html
http://www3.ntu.edu.sg/home/EPNSugan/index_files/dnaprot.htm

$$SP = \frac{TN}{TN + FP} \times 100\%$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(FP + TP)(TP + FN)(TN + FP)(TN + FN)}}$$

Here, TP (true positive) indicates the number of DNA-binding proteins that are correctly predicted as DNA-binding proteins, TN (true negative) indicates the number of non-DNA-binding proteins that are correctly predicted as non-DNA-binding proteins, FP (false positive) is the number of non-DNA-binding proteins that are incorrectly predicted as DNA-binding proteins, and FN (false negative) indicates the number of DNA-binding proteins that are incorrectly predicted as non-DNA-binding proteins.

ACC represents the percentage of the sum of correctly classified samples (TP and TN) among the total number of classified samples. SE represents the percentage of TP among all predicted positives, and SP represents the percentage of TN among all predicted negatives. In a perfect prediction system, these three indicators would achieve scores of 100%. However, in unbalanced datasets, increases in SE lead to decreases in SP, and vice versa. Thus, these indicators do not evaluate prediction results well. In comparison, MCC is a more balanced evaluation criterion with the range [-1, +1]: a value of 1 indicates that the prediction result correlates perfectly with the true categories, a value of 0 indicates a completely random prediction, and a value of -1 indicates total disagreement between the prediction result and the true categories. The area under the curve (AUC) of the receiver operating characteristic (ROC) curve can be used as a more objective classification performance evaluation criterion. The ROC curve is a unit square with values along two axes (false positive rate and true positive rate) ranging from 0 to 1; the AUC maximum of 1 corresponds to a perfect classifier.

It should be noted that the experimental results used in the comparisons performed in this study are all based on the use of the base classifier, which is a linear kernel SVM, i.e., C-SVM in libsvm toolkit (default SVM type), the parameter kernel_type is set to linear kernel (no parameters required). Because we focus our attention on the feature representation of protein sequences, we did not optimize the classifier. Obviously, better predictions can be obtained by selecting the classifier and adjusting its parameters. Moreover, to better demonstrate the effectiveness of the proposed method, we also did not deliberately select the more favorable physicochemical properties. In fact, we have found that better predictions can be obtained by selecting a more efficient subset of physicochemical properties than just using the six physicochemical properties listed in the literature [40].

3.2 Evaluation of dual-source IFFR

In this experiment, we focus on the assessment of the proposed feature representation algorithms (without ensemble techniques); in other words, we discuss model selection for feature representation. For the DNA-binding protein prediction problem, we first compare and evaluate the performance of dual-source IFFR (i.e., Algorithm 1 with λ = 0) based on physicochemical properties and evolutionary information. Some state-of-the-art feature representation algorithms are also used for comparison.

First, the performances of the four algorithms, CovPCSM, CovPSSM, CFFR and IFFR, are validated and compared using the Jackknife method with the benchmark datasets. The results are shown in Table 3. The CovPCSM method considers only physicochemical properties, and its recognition ability is mediocre. The CovPSSM method considers only evolutionary information but has better recognition ability. The CFFR method is a simple tandem combination of CovPCSM and CovPSSM; the generated feature vector considers both physicochemical properties and evolutionary information, and its recognition ability is slightly better than that of CovPSSM. The IFFR approach considers not only correlations within physicochemical properties and within evolutionary information but also the interaction effects between physicochemical properties and evolutionary information and thereby achieves better recognition.

We then further examine and compare the performance of the proposed feature representation IFFR with three feature representation algorithms, pseudo-PSSM (PsePSSM) [43], PseAAC and AAC, with the four independent datasets. To produce a more objective and reliable comparison, we use 30 random results of 10-foldCV for the analysis.

Table 3
Comparison of the prediction performances of different feature representations (Jackknife validation test)

Dataset	Evaluation Indices	Feature Representation Method			
		CovPCSM	CovPSSM	CFFR	IFFR
Alternate Dataset	MCC	0.3015	0.4701	0.4735	**0.4824**
	ACC (%)	63.62	73.11	73.29	**73.76**
	SE (%)	85.08	82.22	**82.31**	**82.31**
	SP (%)	42.15	64.01	64.27	**65.22**
PDB1075 Dataset	MCC	0.3882	0.5266	0.5504	**0.5533**
	ACC (%)	68.65	76.00	77.21	**77.40**
	SE (%)	57.27	69.64	71.09	**71.82**
	SP (%)	80.57	82.67	**83.62**	83.24
Independent 1 Dataset	MCC	0.6881	0.9612	0.9612	**0.9624**
	ACC (%)	84.14	98.06	98.06	**98.12**
	SE (%)	78.01	97.57	97.45	**97.57**
	SP (%)	90.28	98.54	**98.66**	**98.66**
Independent 2 Dataset	MCC	NaN	0.6826	0.6761	**0.6937**
	ACC (%)	72.59	87.23	86.92	**87.85**
	SE (%)	0.00	**78.41**	**78.41**	77.27
	SP (%)	100.00	90.56	90.13	**91.85**
Training Dataset	MCC	0.4050	0.6922	0.7099	**0.7197**
	ACC (%)	72.22	85.61	86.36	**86.87**
	SE (%)	77.60	88.00	88.00	**88.80**
	SP (%)	63.01	81.51	**83.56**	**83.56**

Note: The values shown in bold are the best prediction results.

As shown in Figure 2, the feature representation algorithm IFFR shows excellent performance with the Alternate Dataset, the PDB1075 Dataset and the Independent 2 Dataset, and its average performance is superior to those of the other algorithms (PsePSSM, PseAAC and AAC). For all the datasets, the IFFR usually has a small standard deviation; this finding indicates that to some extent, the IFFR is not sensitive to the random composition of the training set, and thus, the proposed algorithm is more robust. With the Independent 1 Dataset, the feature representation algorithm PsePSSM also demonstrates good performance and is significantly better than PseAAC and AAC. Because both IFFR and PsePSSM use evolutionary information, the results therefore suggest that the evolutionary information in PSSM is more abundant and more important than the information contained in the sequence itself. Therefore, prediction performance can be improved by considering evolutionary information. In conclusion, our feature representation, IFFR, shows superior performance with four independent datasets than three state-of-the-art algorithms (PsePSSM, PseAAC and AAC).

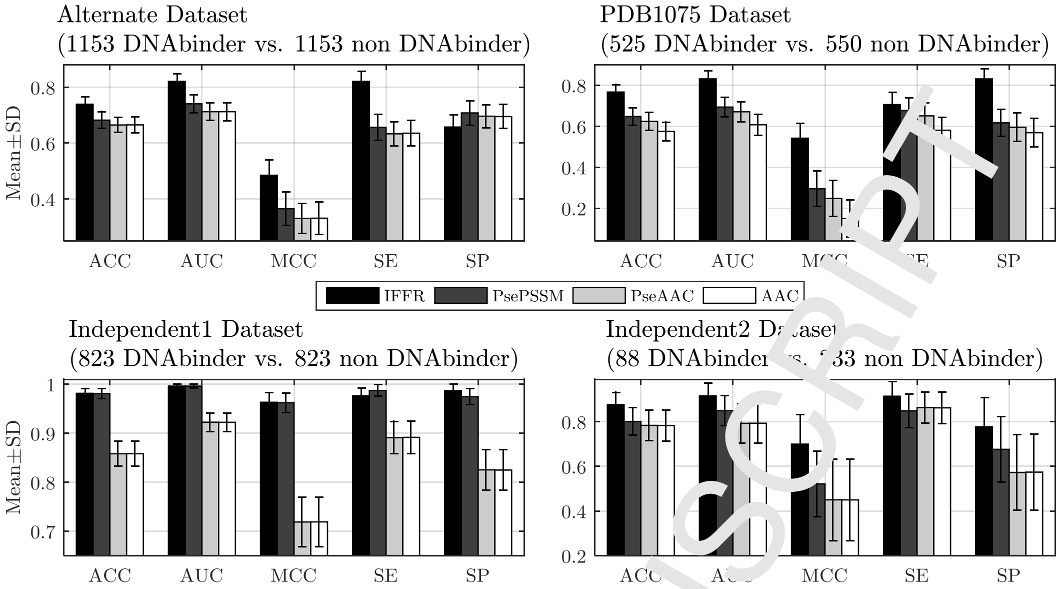


Fig. 2. Comparison of IFFR with existing state-of-the-art feature representation algorithms (AAC, PseAAC and PsePSSM) using four independent datasets. (30 random results of 10-fold CV, base classifier: linear SVM)

In summary, as shown in Table 3 and Figure 2, the recognition rate of the dual-source IFFR based on physicochemical properties and evolutionary information is higher than the recognition rates achieved by single-source feature representation algorithms such as CovPCSM, CovPSSM and AAC. IFFR also achieves a higher recognition rate than other dual-source (or combined) fusion algorithms such as CFFR, PseAAC and PsePSSM. These experimental results show the existence of interaction effects between the physicochemical properties and evolutionary information of DNA-binding proteins and demonstrate that the recognition rate can be improved by using these implicit features and considering their interaction effects. Thus, our IFFR depicts explicit and implicit features simultaneously and can more fully mine the information hidden in a protein sequence. This result validates Hypothesis 1.

3.3 Sensitivity analysis of parameter and comparison of models

In this section, we discuss the parameter selection problem in the multi-source fusion feature representation model and examine the sensitivity of the parameter *λ*, i.e., the effect of different gap distances *λ* on the results obtained within the framework of the proposed model. To ensure the reproducibility of the experimental results and their subsequent comparability, we continue to use the Jackknife validation method and linear kernel SVM classifier in our analysis.

We conduct comparisons using the same four independent datasets and the Jackknife validation method while varying the algorithm parameter λ (gap distance) continuously from 1 to L-1 (L is the length of the protein sequence) and observe the differences in the results obtained using three different algorithms (GapPSSM, GapCFFR and GapIFFR). The results in terms of MCC are shown in Figure 3.

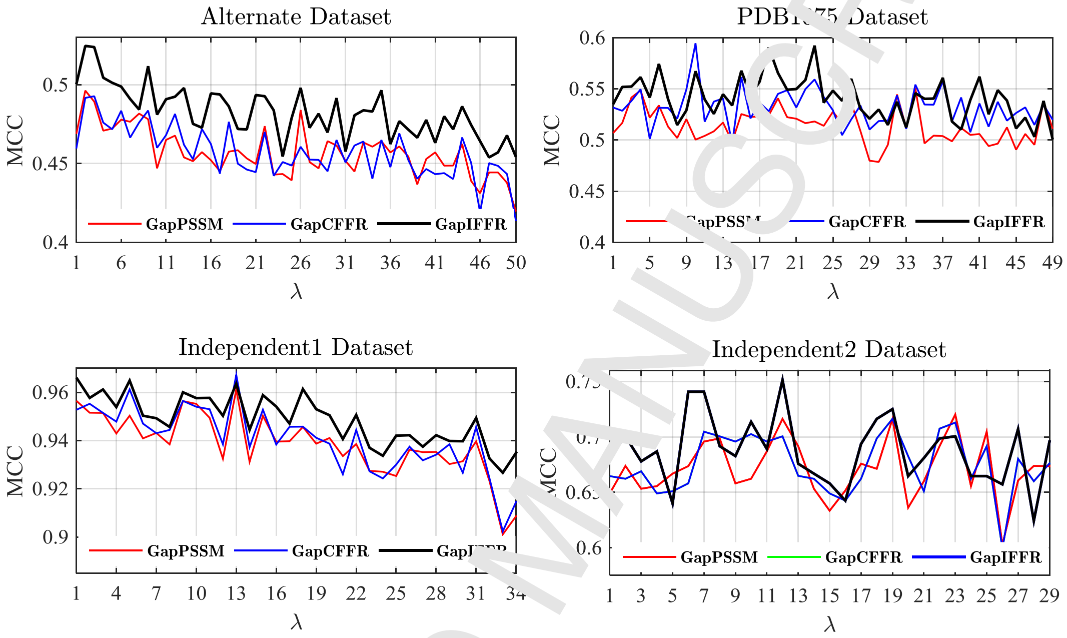


Fig. 3. Comparison of the performances of GapPSSM, GapCFFR, and GapIFFR, showing the effect of parameter λ on the results. (Jackknife, classifier: linear SVM)

As shown in Figure 3, the fluctuation of the performance curve is relatively large, and the gap distance (parameter λ) has a significant effect on the performance of the classifier, which indicates that the recognition rate is sensitive to parameter λ. For different datasets, the influence of parameter λ on the prediction results differs, and the values of parameter λ that yield the best performance also differ among the datasets. For example, for the Alternate Dataset, the three algorithms achieve the optimal MCC value at λ = 2, and a short-range interaction was observed between residues at the corresponding positions, i.e., contact between residues. For the Independent 2 Dataset, the maximum MCC value is obtained at λ = 12, indicating that a remote interaction between residues at the corresponding position (i.e., contact between residues). Similarly, the three performance curves show consistent fluctuations because all three algorithms consider information of gap distances, i.e., the interaction of residue pairs corresponding to different λ values; therefore, the algorithms show a consistent trend in their fluctuations. Figure 3 also shows that the performance of the triple-source interaction fusion algorithm GapIFFR is better than that of the combined fusion algorithm GapCFFR and that of the dual-source fusion algorithm GapPSSM.

To strengthen the credibility of the results, we also analyze the statistical significance of the Jackknife results in terms of four indicators (MCC, ACC, SE and SP) to determine whether there are significant differences in the results obtained using three algorithms. Specifically, we use both parametric and nonparametric statistical tests (the paired t-test and the Wilcoxon signed rank test) to determine whether there are significant differences in prediction performance among the different feature representation methods. Because the prediction performance of these methods was measured using the same training and test sets, i.e., there are no differences in the random compositions of the sample sets, any differences revealed by the paired statistical tests can be attributed to a difference in the algorithms.

Table 4
Statistical significance of the performance differences between algorithms.

Dataset	Evaluation	GapPSSM vs. GapIFFR		GapCFFR vs. GapIFFR	
	Indices	Paired T-test	Signed-rank Test	Paired T-test	Signed-rank Test
Alternate Dataset	MCC	(-) 1.[illegible]53×10$^{-2[illegible]}$	(-) 7.557×10^{-10}	(-) 1.018×10^{-19}	(-) 7.557×10^{-10}
	ACC	(-) [illegible].756×10^{-20}	(-) 7.513×10^{-10}	(-) 5.259×10^{-19}	(-) 7.977×10^{-10}
	SE	(-) 2.50[illegible]×10$^{-[illegible]}$	(-) 6.470×10^{-7}	(-) 4.528×10^{-14}	(-) 2.710×10^{-9}
	SP	(-) 1.624×10^{-15}	(-) 2.159×10^{-9}	(-) 3.492×10^{-12}	(-) 2.056×10^{-8}
PDB1075 Dataset	MCC	(-) 2.[illegible]3×10^{-13}	(-) 3.775×10^{-9}	(-) 0.0016	(-) 0.0026
	ACC	(-) 1.207×10^{-13}	(-) 4.657×10^{-9}	(-) 0.0013	(-) 0.0023
	SE	(-) 3.765×10^{-13}	(-) 5.556×10^{-9}	(-) 0.0037	(-) 0.0096
	SP	(-) 6.848×10^{-7}	(-) 2.244×10^{-6}	(-) 0.0248	(-) 0.0342
Independent1 Dataset	MCC	(-) 3.390×10^{-12}	(-) 3.653×10^{-7}	(-) 2.585×10^{-9}	(-) 7.443×10^{-7}
	ACC	(-) 2.768×10^{-12}	(-) 3.444×10^{-7}	(-) 2.325×10^{-9}	(-) 6.871×10^{-7}
	SE	(-) 5.013×10^{-11}	(-) 3.495×10^{-7}	(-) 2.159×10^{-8}	(-) 4.131×10^{-6}
	SP	(-) 5.993×10^{-5}	(-) 1.278×10^{-4}	(-) 6.478×10^{-4}	(-) 8.413×10^{-4}
Independent2 Dataset	MCC	(-) 0.0045	(-) 0.0064	(-) 0.0170	(-) 0.0264
	ACC	(-) 0.0067	(-) 0.0092	(-) 0.0202	(-) 0.0322
	SE	(=) 0.0994	(=) 0.0810	(=) 0.1160	(=) 0.1247
	SP	(-) 0.0011	(-) 0.0018	(-) 0.0202	(-) 0.0232

Here, (-) implies that the second algorithm is statistically better than the first one, (=) means that the two algorithms show no significant differences between them, and the p-values are given.

Table 4 shows the results of the comparisons. For GapPSSM (GapCFFR) and GapIFFR, with the exception of the SE index obtained with the Independent 2 Dataset, the p-values of the paired tests for the four performance indices across all datasets were less than the significance level of 0.05. This finding suggests that the prediction performances of these three algorithms are significantly different; specifically, the prediction performance of GapIFFR is significantly better than those of GapPSSM and GapCFFR, and this finding was obtained with the four datasets. In contrast, there was no statistically significant difference in the SE index obtained with the Independent 2 Dataset. Thus, the feature representation obtained with GapIFFR is significantly better than that of the other two algorithms studied. This result indicates that within the framework of IFFR models, the triple-source fusion GapIFFR method is significantly better than the dual-source fusion GapPSSM. GapIFFR is also significantly superior to the combined fusion feature representation GapCFFR. This result validates Hypothesis 2.

3.4 Evaluation of selective ensemble based on parameter perturbation

In this experiment, we discuss the selective ensemble based on different gap distances λ, that is, we perturb parameter λ to generate different input feature subspaces and then construct different base classifiers to enhance the generalization ability of the integrated learner. To ensure the comparability of the experimental results, we again use the Jackknife validation method. Assuming that the protein dataset has N sequences, each of these sequences is used as a sample to be tested, and the remaining N-1 protein sequences are divided using the K-fold CV (K = 10 in this study); of these, (K-1)-fold sequences are used as the training set ($(\mathrm{K}\text{-}1)/\mathrm{K}\times(\mathrm{N}-1)$ samples) for model training, and 1-fold sequences are used as the validation set ($1/\mathrm{K}\times(\mathrm{N}-1)$ samples) for selection (pruning) to determine the structure of the integrated learner.

To save space, we select only the MCC for the comparison because this index can better reflect the generalization ability of the learner; the other indicators can be used for similar analyses. The four datasets are used in the experiments to compare the performance of the proposed selective ensemble algorithm GapIFFR-SE (here, k is directly set to 3) with those of the IFFR (as a benchmark algorithm) and the GapIFFR. The results are shown in Figure 4 below.

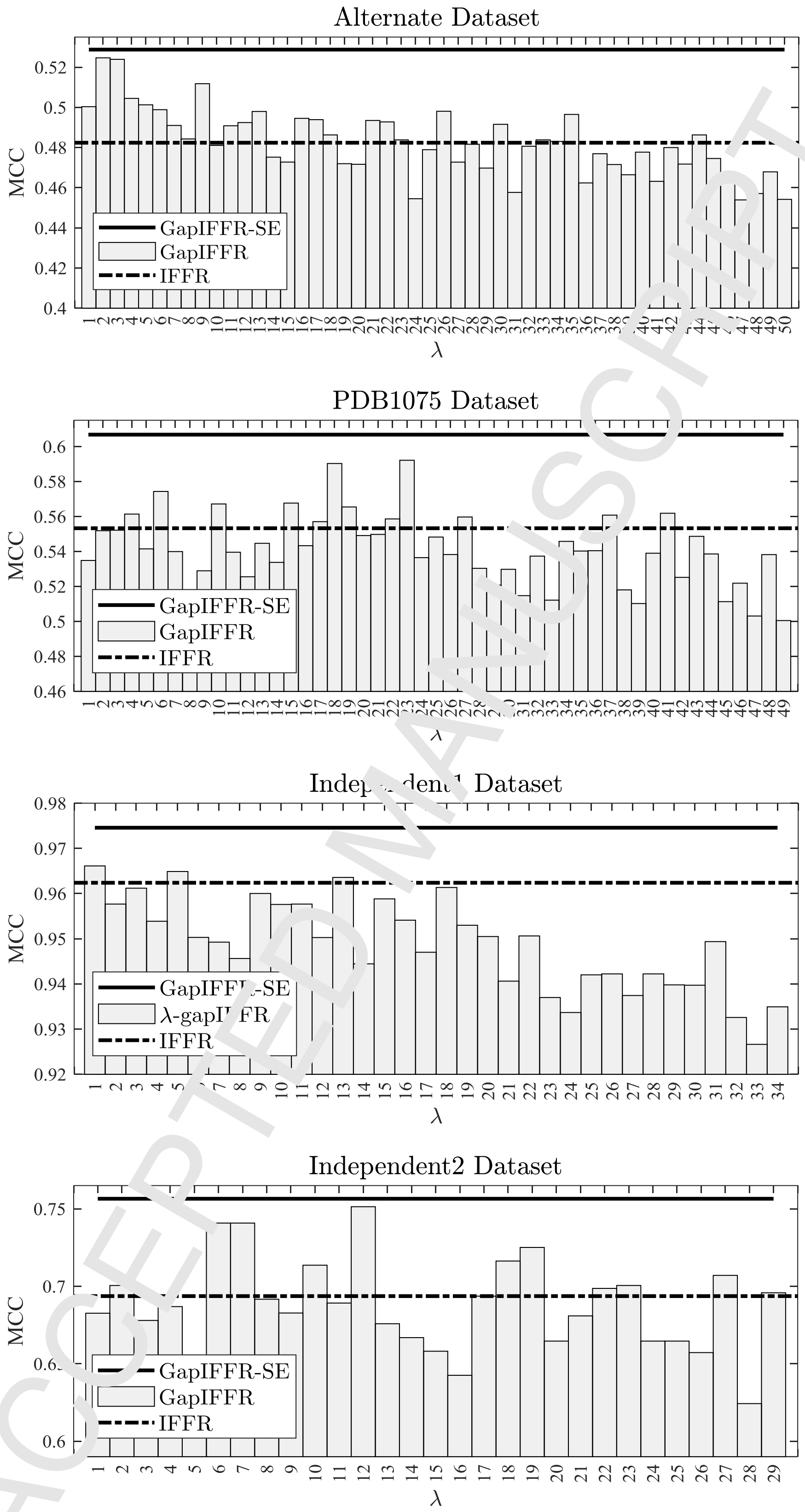


Fig. 4. Comparison of the performances of different algorithms. The effects of parameter λ on the MCC (histogram) of the GapIFFR (IFFR is used as the baseline, black dotted line) are compared to those of the selective ensemble (GapIFFR-SE, black solid line). (Jackknife test)

We used the MCC value of the IFFR algorithm as the baseline (dotted black line), and as shown in Figure 4, for each dataset, some columns of the histogram exceed the dotted black lines and appear above the baseline. This finding suggests that certain parameters *λ* in GapIFFR yield MCC index values that exceed the baseline (dotted black line). In addition, based on the distribution of the histogram, the class discrimination ability of feature vectors corresponding to different λ values is quite different (performance boundary of the algorithm), and this is due to the presence of residue pairs in the amino acid sequence that interact with each other, i.e., contact between residues. Residue interactions can be divided into short-range interactions and remote interactions, and remote interactions play an important role in determining the overall structural framework. For all datasets except the Independent 1 Dataset, the maximum MCC value corresponds to a λ value greater than 1, i.e., there is contact between non-adjacent residues in these three datasets. For example, a remote interaction is found in the PDB1075 Dataset (λ = 23) and in the Independent 2 Dataset (λ = 12). In fact, it is precisely because of this spatial information (gap distance) that the feature vector generated by Algorithm 1 exhibits stronger class discriminating ability.

The solid black line in Figure 4 corresponds to the MCC value of Algorithm 2, GapIFFR-SE. As shown in the figure, the performance of the proposed selective ensemble classification algorithm is significantly improved due to the existence of polypeptide chain folds in the protein and potentially multiple pairs of residue contacts in the corresponding amino acid sequence (i.e., several different λ values). Therefore, the motivation of the selective ensemble algorithm is to find a set of potential λ values. As shown in Figure 4, there are multiple different λ values for all datasets, and these yield higher MCC values. For example, there are three different λ values (λ = 23, 18, and 6) in the PDB1075 Dataset, and Algorithm 2 can capture these λ values, perform classification prediction using these generated feature spaces, and perform a majority vote for these prediction results. As a result, GapIFFR-SE showed better overall classification performance.

3.5 Further comparison with current methods

For a further comparison of feature representation methods, we use the HoldOut method to evaluate the performance of feature representation models with a different dataset (the training and test sets are described in separate studies in the literature). We generate features using a given training set and train the classifiers with the resulting feature space. The classification model is then verified with the corresponding feature space of the given test set, and the recognition rate of the classifier is used to indirectly evaluate the performance of the different feature representations. Table 5 shows the experimental results obtained with the different feature representation methods (CovPCSM, CovPSSM, CFFR and IFFR) in the proposed model framework and with three state-of-the-art feature representation algorithms (PsePSSM, PseAAC and AAC) for the test set. Table 5 shows that CFFR achieves maximum values of the three evaluation indices ACC, MCC and AUC of 87.50%, 0.7562 and 0.9297, respectively. However, we are more interested in identifying the positive cases (DNA-binding proteins). The SE index shows that the SE values obtained for the PsePSSM and IFFR are higher than 80%; these two algorithms identify 77 and 74 positive cases, respectively, from their confusion matrices. We also note that the smaller number of support vectors (nSV) used by the classifier (LinearSVM) indicates that the classification model has better generalization capabilities. A comprehensive comparison reveals that the feature representation IFFR demonstrates better performance, with an MCC index of 0.7204, 74 recognized positive cases, and the smallest number of support vectors in the classification model. These findings also demonstrate the validity of IFFR to some extent. In contrast, the feature representation PsePSSM shows the highest SE index (SE = 83.70%) but also exhibits a low specificity index SP and a non-ideal MCC. This phenomenon is also consistent with the results shown in Figure 2.

Table 5
Comparison of the performances of various methods with a testing dataset containing 92 DNA-binding proteins and 100 non-DNA-binding proteins.

Method	Evaluation Indices						
	Confusion Matrix	SE (%)	SP (%)	ACC (%)	MCC	AUC	nSV
AAC (d = 20)	63 29 21 79	68.48	79.00	73.96	0.4781	0.7765	(87,90)
PseAAC (d = 420)	70 22 14 86	76.09	86.00	81.25	0.6252	0.8843	(99,102)
PsePSSM (Ref. λ = 0~24)	77 15 18 82	83.70	82.00	82.81	0.6564	0.8935	(89,98)
CovPCSM	57 35 2 98	61.96	98.00	80.73	0.6492	0.9104	(127,127)
CovPSSM	71 21	77.17	93.00	85.42	0.7138	0.9218	(74,78)

	7 ⁝ 93						
CFFR	73 ⁝ 19 5 ⁝ 95	79.35	95.00	87.50	0.7562	0.9297	(72,79)
IFFR	74 ⁝ 18 9 ⁝ 91	80.43	91.00	85.94	0.7204	0.9230	(69,77)
	a ⁝ *b* *c* ⁝ *d*	*a* = true positive; *b* = false negative (type II error); *c* = false positive (type I error); *d* = true negative.					

For further comparison of the prediction methods, we compare the proposed selective ensemble prediction method, GapIFFR-SE, with other prediction methods using the benchmark dataset PDB1075. Eight state-of-the-art methods are used in the comparison: iDNA-Prot|dis [39], PseDNA-Pro [33], iDNA-Prot [12], DNA-Prot [34], DNAbinder [10], iDNAPro-PseAAC [22], Kmer1+AAC [19] and Local-DPP [25]. The comparison results based on Jackknife validation are shown in Table 6. Our selective ensemble algorithm, GapIFFR-SE, exhibits the best prediction performance among the compared methods, with a maximal recognition rate of 79.91%, a maximal MCC value of 0.61, and a maximal SE value of 87.43. Thus, compared with existing methods, the proposed method demonstrates superior performance. This finding indirectly indicates that the IFFR method proposed in our study can generate features that carry strongly discriminative information and that the use of a selective ensemble can further enhance the generalization ability of ensemble learning, ultimately ensuring the accurate prediction of DNA-binding proteins.

Table 6
Comparison of the performances of various methods with the PDB1075 dataset (Jackknife test).

Methods	Evaluation Indices			
	ACC (%)	MCC	SE (%)	SP (%)
iDNA-Prot|dis [32]	77.30	0.54	79.40	75.27
PseDNA-Pro [33]	76.55	0.53	79.61	73.63
iDNA-Prot [12]	75.40	0.50	83.81	64.73
DNA-Prot [34]	72.55	0.44	82.67	59.76
DNAbinder (dimension = 400) [10]	73.58	0.47	66.47	**80.36**
DNAbinder (dimension = 21) [10]	73.95	0.48	68.57	79.09
iDNAPro-PseAAC [22]	76.56	0.53	75.62	77.45
Kmer1+AAC [19]	75.23	0.50	76.76	73.76
Local-DPP (n = 3, lambda = 1) [25]	79.10	0.59	84.80	73.60
Local-DPP (n = 2, lambda = 2) [25]	79.20	0.59	84.00	74.50
Proposed method (k = 3)	**79.91**	**0.61**	**87.43**	72.73

4 Conclusion

The prediction of protein structure and function from protein sequences (primary structures) using machine learning methods is currently a popular and important topic in research, particularly in bioinformatics research. The development of methods that can be used to adequately and effectively express feature information from sequence data is currently a focus of the field. For protein sequences, the AAC, polypeptide composition (adjacent residues), PseAAC (non-adjacent residues), physicochemical properties and evolutionary information are commonly used to generate explicit features and to combine these feature vectors. The use of this type of CFFR can achieve good results.

In this paper, we propose a feature representation algorithm with multi-source interaction fusion. The basic principle of this method is that it considers the interaction effects among different physicochemical properties, evolutionary information, and local position information between different amino acid residues in protein sequences. Experimental data demonstrate that the feature-level fusion of physicochemical properties, evolutionary information and position information between non-adjacent residues from the perspective of interactions can significantly improve the prediction of DNA-binding proteins. The fact that the generated feature vector demonstrates better performance in recognizing DNA-binding proteins indicates that our feature representation algorithm can mine the potential information hidden in a protein sequence. Furthermore, the parameters of the feature representation algorithm can be perturbed to generate different input feature subspaces. The selective ensemble algorithm improves the generalization ability of the ensemble classifier by obtaining differential classifiers via selection (pruning). The proposed model and algorithms are mathematical descriptions based on specific biological problems but are nonetheless universal, and the analytical methods described in this paper can be applied to other questions related to protein structure and function prediction. Due to its applicability to the in-depth analysis of proteins and for aiding the understanding of frontier issues, our method is of some significance to the field of bioinformatics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No.

61473329, No. 61573296, and No. 61673323), the Fujian Provincial Natural Science Foundation of China (No. 2015J01009), the National Science and Technology Major Project by the Ministry of Industry and Information Technology of China (Application of a new intelligent manufacturing mode of tractors with newly designed wheels, No. 2016-0744) and the Natural Sciences and Engineering Research Council of Canada.

References

[1] M. Ptashne, Regulation of transcription: from lambda to eukaryotes, Trends in biochemical sciences, 30 (2005) 275-279.

[2] K.A. Jones, J.T. Kadonaga, P.J. Rosenfeld, T.J. Kelly, R. Tjian, A cellular DNA-binding protein that activates eukaryotic transcription and DNA replication, Cell, 48 (1987) 79-89.

[3] F. Cajone, M. Salina, A. Benelli-Zazzera, 4-Hydroxynonenal induces a DNA-binding protein similar to the heat-shock factor, Biochemical Journal, 262 (1989) 977-979.

[4] M.J. Buck, J.D. Lieb, ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments, Genomics, 83 (2004) 349-360.

[5] C.-C. Chou, T.-W. Lin, C.-Y. Chen, A.H.-J. Wang, Crystal structure of the hyperthermophilic archaeal DNA-binding protein Sso10b2 at a resolution of 1.85 Angstroms, Journal of bacteriology, 185 (2003) 4066-4073.

[6] H. Zhao, Y. Yang, Y. Zhou, Structure-based prediction of RNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function, Bioinformatics, 26 (2010) 1857-1863.

[7] H. Tjong, H.-X. Zhou, DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces, Nucleic Acids Research, 35 (2007) 1465-1477.

[8] R.E. Langlois, H. Lu, Boosting the prediction and understanding of DNA-binding domains from sequence, Nucleic acids research, 38 (2010) 3149-3158.

[9] H.-L. Huang, I.-C. Lin, Y.-F. Liou, C.-T. Tsai, K.-T. Hsu, W.-L. Huang, S.-J. Ho, S.-Y. Ho, Predicting and analyzing DNA-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties, Bmc Bioinformatics, 12 (2011) S47.

[10] M. Kumar, M.M. Gromiha, G.P. Raghava, Identification of DNA-binding proteins using support vector machines and evolutionary profiles, BMC Bioinformatics, 8 (2007) 463.

[11] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, N. Deng, Predicting DNA-and RNA-binding proteins from sequences with kernel methods, Journal of Theoretical Biology, 258 (2009) 289-293.

[12] W.-Z. Lin, J.-A. Fang, X. Xiao, K.-C. Chou, iDNA-Prot: identification of DNA binding proteins using random forest with grey model, PloS one, 6 (2011) e24756.

[13] Y.-D. Cai, A.J. Doig, Prediction of Saccharomyces cerevisiae protein functional class from functional domain composition, Bioinformatics, 20 (2004) 1292-1300.

[14] A. Szilágyi, J. Skolnick, Efficient prediction of nucleic acid binding function from low-resolution protein structures, Journal of molecular biology, 358 (2006) 922-933.

[15] X. Yu, J. Cao, Y. Cai, T. Shi, Y. Li, Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines, Journal of Theoretical Biology, 240 (2006) 175-184.

[16] K.C. Chou, Prediction of protein cellular attributes using pseudo‐amino acid composition, Proteins: Structure, Function, and Bioinformatics, 43 (2001) 246-255.

[17] K.-C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics, 21 (2004) 10-19.

[18] S. Ahmad, A. Sarai, PSSM-based prediction of DNA binding sites in proteins, BMC bioinformatics, 6 (2005) 33.

[19] Q. Dong, S. Wang, K. Wang, X. Liu, B. Liu, Identification of DNA-binding proteins by auto-cross covariance transformation, Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on, IEEE2015, pp. 470-475.

[20] S.-Y. Ho, F.-C. Yu, C.-Y. Chang, H.-L. Huang, Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method, Biosystems, 90 (2007) 234-241.

[21] R. Xu, J. Zhou, B. Liu, Y. He, Q. Zou, X. Wang, K.-C. Chou, Identification of DNA-binding proteins by

incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach, Journal of Biomolecular Structure and Dynamics, 33 (2015) 1720-1730.

[22] B. Liu, S. Wang, X. Wang, DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation, Scientific reports, 5 (2015) 15479.

[23] J. Zhang, B. Gao, H. Chai, Z. Ma, G. Yang, Identification of DNA-binding proteins using multi-features fusion and binary firefly optimization algorithm, BMC bioinformatics, 17 (2016) 323.

[24] L. Li, X. Cui, S. Yu, Y. Zhang, Z. Luo, H. Yang, Y. Zhou, X. Zheng, PSSP-RFE: accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical-chemical property and functional annotations, PLoS One, 9 (2014) e92863.

[25] L. Wei, J. Tang, Q. Zou, Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information, Information Sciences, 384 (2017) 135-144.

[26] Z.-H. Zhou, Ensemble Learning, Encyclopedia of Biometrics, 270–273, Springer US2009.

[27] Z.-H. Zhou, J. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial intelligence, 137 (2002) 239-263.

[28] G. Tsoumakas, I. Partalas, I. Vlahavas, A taxonomy and short review of ensemble selection, Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications2008, pp. 1-6.

[29] G. Martınez-Munoz, D. Hernández-Lobato, A. Suárez, An analysis of ensemble pruning techniques based on ordered aggregation, IEEE Transactions on Pattern Analysis and Machine Intelligence, 31 (2009) 245-259.

[30] L. Rokach, Ensemble-based classifiers, Artificial Intelligence Review, 33 (2010) 1-39.

[31] C.-X. Zhang, J.-S. Zhang, G.-Y. Zhang, Using boosting to prune double-bagging ensembles, Computational Statistics & Data Analysis, 53 (2009) 1218-1231.

[32] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, K.-C. Chou, iDNA-Prot| dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition, PloS one, 9 (2014) e106691.

[33] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, X. Wang, PseDNA‐Pro: DNA‐binding protein identification by combining Chou’s PseAAC and physicochemical distance transformation, Molecular Informatics, 34 (2015) 8-17.

[34] K.K. Kumar, G. Pugalenthi, P. Suganthan, DNA-Prot: identification of DNA binding proteins from protein sequence information using random forest, Journal of Biomolecular Structure and Dynamics, 26 (2009) 679-686.

[35] C. Cortes, L.D. Jackel, W.-P. Chiang, Limits on learning machine accuracy imposed by data quality, Advances in Neural Information Processing Systems1995, pp. 239-246.

[36] J.R. Raol, Data fusion mathematics: theory and practice, CRC Press2015.

[37] C. Huang, J. Li, C. Mei, W.-Z. Wu, Three-way concept learning based on cognitive operators: An information fusion viewpoint, International Journal of Approximate Reasoning, 83 (2017) 218-242.

[38] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic acids research, 25 (1997) 3389-3402.

[39] A.A. Schäffer, L. Aravind, T.L. Madden, S. Shavirin, J.L. Spouge, Y.I. Wolf, E.V. Koonin, S.F. Altschul, Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements, Nucleic acids research, 29 (2001) 2994-3005.

[40] H.-B. Shen, K.-C. Chou, PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition, Analytical biochemistry, 373 (2008) 386-388.

[41] M. Moreira, E. Mayoraz, Improved pairwise coupling classification with correcting classifiers, European conference on machine learning, Springer1998, pp. 160-171.

[42] L. Wang, S.J. Brown, BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences, Nucleic acids research, 34 (2006) W243-W248.

[43] K.-C. Chou, H.-B. Shen, MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM, Biochemical and biophysical research communications, 360 (2007) 339-345.

Appendix 1

The NCBI non-redundant database (ftp://ftp.ncbi.nlm.nih.gov/blast/db/) should first be downloaded.

A query protein sequence can then be obtained by setting parameter E to 0.001 (evalue = 0.001) and the number of iterations to 3 (num_iterations = 3). The NR protein database is then searched for this sequence using PSI-BLAST.

Finally, a 20-dimensional vector representing the probabilities of conservation against mutations to 20 different amino acids, including itself, is returned. The matrix consisting of such vector representations for all residues in a given sequence is called the position-specific SM (PSSM) [18].

Appendix 2

Table 7.
Values of six physicochemical properties for each amino acid.

Amino Acid		Physicochemical Index					
		Hydrophobicity	Hydrophilicity	Mass	pK1(α-COOH)	pK2(NH3)	pI(25°C)
		$\boldsymbol{Q}^{(1)}$	$\boldsymbol{Q}^{(2)}$	$\boldsymbol{Q}^{(3)}$	$\boldsymbol{Q}^{(4)}$	$\boldsymbol{Q}^{(5)}$	$\boldsymbol{Q}^{(6)}$
A	Ala	0.62	-0.5	15	2.35	9.87	6.11
C	Cys	0.29	-1.0	47	1.71	10.78	5.02
D	Asp	-0.9	3.0	59	1.88	9.60	2.98
E	Glu	-0.74	3.0	73	2.19	9.67	3.08
F	Phe	1.19	-2.5	91	2.58	9.24	5.91
G	Gly	0.48	0.0	1	2.34	9.60	6.06
H	His	-0.40	-0.5	82	1.78	8.97	7.64
I	Ile	1.38	-1.8	57	2.32	9.76	6.04
K	Lys	-1.50	3.0	73	2.20	8.90	9.47
L	Leu	1.06	-1.8	57	2.36	9.60	6.04
M	Met	0.64	-1.3	75	2.28	9.21	5.74
N	Asn	-0.78	0.2	58	2.18	9.09	10.76
P	Pro	0.12	0.0	42	1.99	10.60	6.30
Q	Gln	-0.85	0.2	72	2.17	9.13	5.65
R	Arg	-2.53	3.0	101	2.18	9.09	10.76
S	Ser	-0.18	0.3	31	2.21	9.15	5.68
T	Thr	-0.05	-0.4	45	2.15	9.12	5.60
V	Val	1.08	-1.5	43	2.29	9.74	6.02
W	Trp	0.81	-3.4	130	2.38	9.39	5.88
Y	Tyr	0.26	-2.3	107	2.20	9.11	5.63