Published online 28 September 2021

https://doi.org/10.1093/nar/gkab829

BioSeq-BLM: a platform for analyzing DNA, RNA and protein sequences based on biological language models

Hong-Liang Li[1], Yi-He Pang[1] and Bin Liu[1,2,*]

[1]School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China and [2]Advanced Research Institute of Multidisciplinary Science, Beijing Institute of Technology, Beijing, China

Received July 18, 2021; Revised August 24, 2021; Editorial Decision September 07, 2021; Accepted September 09, 2021

ABSTRACT

In order to uncover the meanings of 'book of life', 155 different biological language models (BLMs) for DNA, RNA and protein sequence analysis are discussed in this study, which are able to extract the linguistic properties of 'book of life'. We also extend the BLMs into a system called BioSeq-BLM for automatically representing and analyzing the sequence data. Experimental results show that the predictors generated by BioSeq-BLM achieve comparable or even obviously better performance than the exiting state-of-the-art predictors published in literatures, indicating that BioSeq-BLM will provide new approaches for biological sequence analysis based on natural language processing technologies, and contribute to the development of this very important field. In order to help the readers to use BioSeq-BLM for their own experiments, the corresponding web server and stand-alone package are established and released, which can be freely accessed at http://bliulab.net/BioSeq-BLM/.

INTRODUCTION

The genome is the 'book of life', whose languages are the biological sequences (1). Natural languages and biological sequences are similar. For examples, the peptide bonds connect the amino acid residues to form a protein with certain structure and function. Similarly, words are combined by grammar and linguistic rules into a sentence with certain meanings. In this regard, the techniques grounded in linguistics are used to uncover the meanings of the 'book of life', and have greatly contributed to the development of biological sequence analysis. Protein domains can be considered as the words of proteins, and the rules for domain associations are the grammar of proteins (2). Inspired by these similarities between proteins and languages, the linguistic technique *n*-gram was employed to probe the proteome grammar, showing that a 'quasi-universal grammar' underlies the evolution of domain architectures (3). Biological sequences store all the information determining their structures and functions, and the sentences contain all the information defining their syntactic and semantic (4). Because the relationships among biological sequence, structure and function are similar as the relationships among sentence, syntactic and semantic in linguistics (see Figure 1), techniques for semantic analysis derived from natural language processing have been applied to predict the structures and functions of proteins (5), providing new ideas and approaches for solving these tasks.

All these approaches based on natural language processing are playing important roles in uncovering the meanings of the 'book of life'. Unfortunately, we still know only a little about its semantic. The existing studies focus on exploring the lexical, syntactic, or semantic of biological sequences. The biological sequences with various structures and functions share some common features with natural languages, but they also have their own linguistic properties. For examples, there are >500 physiochemical properties for amino acids (6), and >180 physiochemical properties for nucleotides (7). Even the most complicated polysemous word in a language will never have so many properties. As a result, the rule-based approaches show limited performance for some difficult tasks, such as protein disordered region prediction, enhancer identification, etc. Furthermore, these methods highly depend on the experience-based linguistic features. Therefore, models which are able to automatically and systematically capture the linguistic features are highly desired. They are critical for promoting the development of biological sequence analysis based on natural language processing. Language models can systematically and comprehensively represent and analyze the sentences, independent from the rule-based features, signif-

[*]To whom correspondence should be addressed. Tel: +86 10 68911310; Email: bliu@bliulab.net
Present address: Bin Liu, Beijing Institute of Technology, No. 5, South Zhongguancun Street, Haidian District, Beijing 100081, China.

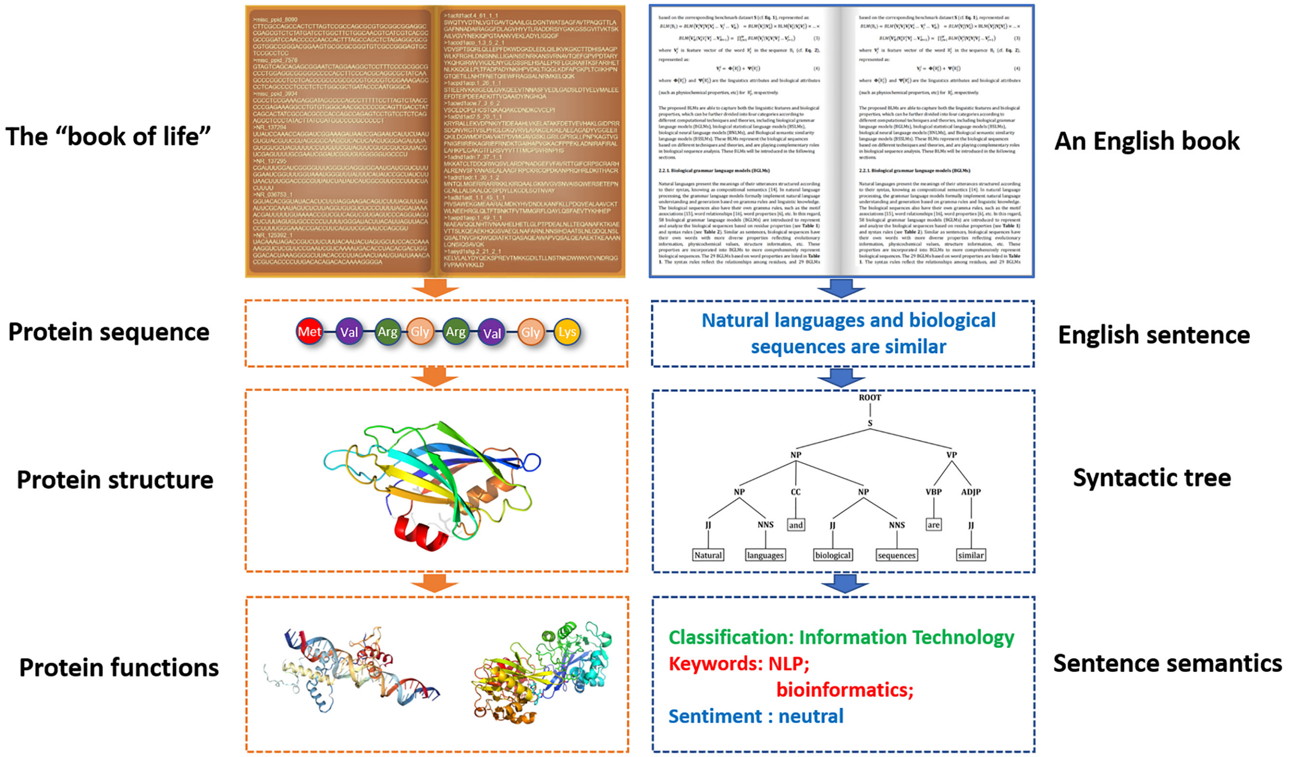


Figure 1. The similarities between protein sequence and natural language sentence.

icantly contributing to the development of the natural language processing (8). Inspired by their successes, we are to propose the biological language models (BLMs) for DNA, RNA and protein sequences. Because the deep learning techniques have been demonstrated to be key methods in bioinformatics, such as protein structure prediction (9), and function analysis (10), BLMs mainly focuses on the biological neural language models to represent and analyze biological sequences based on deep learning techniques. We extend the BLMs to an automatic system called BioSeq-BLM (http://bliulab.net/BioSeq-BLM). Given the sequence data for a specific sequence analysis task, BioSeq-BLM will automatically construct the BLM, select the predictor, evaluate the performance, and analyze the results. BioSeq-BLM is particularly useful for solving the problems of extracting the linguistic features and designing the techniques derived from natural language processing, providing a new view to explore the meanings of 'book of life'. It is anticipated that BioSeq-BLM will be a useful tool for biological sequence analysis, computational proteomics and genomics. As discussed in previous studies (11–13), a system which is able to automatically analyze the biological sequence data is highly desired, and several software tools have been established, such as BioSeq-Analysis (11), BioSeq-Analysis2.0 (12), protr (14), Rcpi (15), Kipoi (16), Janggu (17), Selene (18) and Pydna (19). However, among these existing tools, only BioSeq-BLM and BioSeq-Analysis2.0 are able to automatically construct the predictors with the benchmark datasets as the inputs. The other five tools focus on the individual steps, such as feature extraction, machine learning algorithm selection, or performance evaluation. The fundamental difference between BioSeq-BLM and other similar tools such as BioSeq-Analysis (12) is that BioSeq-BLM is the first study to define the BLMs and introduce 155 different BLMs for biological sequence analysis. Although some features or machine learning techniques also exist in BioSeq-Analysis2.0, BioSeq-BLM is the only existing tool for biological sequence analysis based on biological language models, providing new concepts and techniques for this very important field. These contributions make BioSeq-BLM unique, and more powerful than BioSeq-Analysis2.0. The comparisons between BioSeq-BLM and BioSeq-Analysis2.0 were listed in Table 1, from which we can see that BioSeq-BLM is beyond the reach of BioSeq-Analysis2.0 and any other similar tools.

Our main contributions are as follows:

(1) Based on the similarities between natural languages and biological sequences, we introduce the biological language models (BLMs) motivated by language models (LMs) in the field of natural language processing.
(2) We extend the BLMs into a platform called BioSeq-BLM, only requiring the benchmark datasets as inputs. The predictor will be automatically constructed and evaluated with the help of BioSeq-BLM. BioSeq-BLM is freely available at http://bliulab.net/BioSeq-BLM/.
(3) Experimental results showed that the predictors constructed by BioSeq-BLM are able to improve the predictive performance for some biological sequence analysis tasks.

Table 1. The differences between BioSeq-BLM and BioSeq-Analysis2.0

Modules	Descriptions	BioSeq-BLM	BioSeq-Analysis2.0
BLMs	Number of BGLMs	58	51
	Number of BSLMs	48	0
	Number of BNLMs	41	0
	Number of BSSLMs	8	0
Predictor construction	Number of machine learning algorithms	9	3
	Number of deep learning algorithms	6	0
Performance evaluation	Number of evaluation metrics	10	6
Result analysis	Number of methods for normalization	4	0
	Number of methods for clustering	5	0
	Number of algorithms for feature selection	5	0
	Number of models for dimension reduction	3	0
Other	Support GPU-accelerate or not	Yes	No

MATERIALS AND METHODS

Biological sequence analysis tasks

The aim of biological sequence analysis is to computationally analyze the sequences of DNA, RNA and proteins so as to identify their structures, functions and their associations with diseases. Given a benchmark dataset **S** for a specific biological sequence analysis task with N sequences:

$$\mathbf{S} = \{\mathrm{B}_1,\ \mathrm{B}_2,\ \mathrm{B}_3,\ \mathrm{B}_4,\ \ldots,\ \mathrm{B}_i,\ \ldots,\ \mathrm{B}_N\} \tag{1}$$

where B_i is the i-th biological sequence in **S** represented as:

$$\mathrm{B}_i = \mathrm{R}_1^i\mathrm{R}_2^i\mathrm{R}_3^i\mathrm{R}_4^i \ldots\ \mathrm{R}_j^i \ldots\ \mathrm{R}_M^i \tag{2}$$

where R_j^i represents the j-th word (the words of biological sequences will be introduced in the following sections) in B_i.

Biological sequence analysis tasks can be mainly divided into residue-level analysis and sequence-level analysis (12), aiming to identify the properties of each residue (R_j^i) and the whole sequence (B_i), respectively. Their main difference is that the residue-level analysis treats each residue as a sample, while sequence-level analysis treats each biological sequence as a sample. For more information, please refer to (12). The BLMs will play key roles in these tasks, which will be introduced in the following sections.

Biological language models

The language model (LM) creates a statistical model for English sentences based on the Markov processes (20), which is a milestone in the field of natural language processing. LM determines the joint probability of a word sequence (21) so as to accurately represent and analyze the sentences. In this study, we are to propose the biological language models (BLMs) to represent and analyze the biological sequences following the ideas of LM. A BLM can be constructed for a specific task based on the corresponding benchmark dataset **S** (cf. Equation 1), represented as:

$$\begin{aligned} BLM\ (\mathrm{B}_i) &= BLM\ \left(\mathbf{V}_1^i\mathbf{V}_2^i\mathbf{V}_3^i\mathbf{V}_4^i \ldots \mathbf{V}_j^i \ldots \mathbf{V}_M^i\right) \\ &= BLM\left(\mathbf{V}_1^i|\mathbf{V}_0^i\right) \times \mathrm{BLM}\ \left(\mathbf{V}_2^i|\mathbf{V}_0^i\mathbf{V}_1^i\right) \\ &\quad \times \ldots \times BLM\ \left(\mathbf{V}_M^i|\mathbf{V}_0^i V_1^i\mathbf{V}_2^i \ldots \mathbf{V}_{M-1}^i\right) \\ &= \prod_{j=1}^{M} BLM\left(\mathbf{V}_j^i|\mathbf{V}_0^i\mathbf{V}_1^i\mathbf{V}_2^i \ldots \mathbf{V}_{j-1}^i\right) \end{aligned} \tag{3}$$

where $\mathbf{V}_j^i$ is feature vector of the word R_j^i in the sequence B_i (cf. Equation 2), represented as:

$$\mathbf{V}_j^i = \Phi\left(\mathrm{R}_j^i\right) + \Psi\left(\mathrm{R}_j^i\right) \tag{4}$$

where $\Phi(\mathrm{R}_j^i)$ and $\Psi(\mathrm{R}_j^i)$ are the linguistics attributes and biological attributes (such as physiochemical properties, etc) for R_j^i, respectively.

Inspired by the success of language models (LMs) in the field of natural language processing, we introduce the biological language models (BLMs) for biological sequence analysis. The main differences between LM and BLM are as follows:

(1) Different inputs and methods. The inputs of LMs are sentences, while the inputs of BLMs are biological sequences. Furthermore, the words and word segmentation methods of BLMs are more diverse than those of LMs.

(2) More information in BLMs. BLMs not only consider the linguistic attributes of biological sequences, but also include the biological attributes, such as physical and chemical properties, evolutionary information, motifs, etc.

The proposed BLMs are able to capture both the linguistic features and biological properties, which can be further divided into four categories according to different computational techniques and theories, including biological grammar language models (BGLMs), biological statistical language models (BSLMs), biological neural language models (BNLMs), and biological semantic similarity language models (BSSLMs). These BLMs represent the biological sequences based on different techniques and theories, and are playing complementary roles in biological sequence analysis. These BLMs will be introduced in the following sections.

Biological grammar language models (BGLMs). Natural languages present the meanings of their utterances structured according to their syntax, knowing as compositional semantics (22). In natural language processing, the grammar language models formally implement natural language understanding and generation based on grammar rules and linguistic knowledge. The biological sequences also have their own grammar rules, such as the motif associations (23), word relationships (24), word properties (6), etc. These grammar rules of biological sequences are important for insightfully representing the sequence characteristics. In this regard, 58 biological grammar language models (BGLMs)

Table 2. 29 BGLMs based on syntax rules

Category	Model	Description
DNA	DAC	Dinucleotide-based auto covariance (24)
	DCC	Dinucleotide-based cross covariance (24)
	DACC	Dinucleotide-based auto-cross covariance (24)
	TAC	Trinucleotide-based auto covariance (24)
	TCC	Trinucleotide-based cross covariance (24)
	TACC	Trinucleotide-based auto-cross covariance (24)
	MAC	Moran autocorrelation (104,105)
	GAC	Geary autocorrelation (104,106)
	NMBAC	Normalized Moreau-Broto autocorrelation (104,107)
	ZCPseKNC	Z curve pseudo k tuple nucleotide composition (108)
	ND	Nucleotide Density (26)
RNA	DAC	Dinucleotide-based auto covariance (24,47)
	DCC	Dinucleotide-based cross covariance (24,47)
	DACC	Dinucleotide-based auto-cross covariance (24,47)
	MAC	Moran autocorrelation (104,105)
	GAC	Geary autocorrelation (104,106)
	NMBAC	Normalized Moreau-Broto autocorrelation (104,107)
	ND	Nucleotide Density (26)
Protein	AC	Auto covariance (24,47)
	CC	Cross covariance (24,47)
	ACC	Auto-cross covariance (24,47)
	PDT	Physicochemical distance transformation (27)
	PDT-Profile	Profile-based physicochemical distance transformation (27)
	AC-PSSM	Profile-based Auto covariance (24)
	CC-PSSM	Profile-based Cross covariance (24)
	ACC-PSSM	Profile-based Auto-cross covariance (24)
	PSSM-DT	PSSM distance transformation (27)
	PSSM-RT	PSSM relation transformation (109)
	Motif-PSSM	Use PSSM as input and extract features by motifs-based CNN (23)

are used to represent and analyze the biological sequences. Among these 58 BGLMs, there are 29 models based on the syntax rules (see Table 2). Because the syntax rules reflect the relationships among residues along the biological sequences, these models are particularly useful for analyzing the structures and functions of biological sequences, such as protein disordered region prediction (25), splice site prediction (26), etc. Similar as sentences, biological sequences have their own words with more diverse properties reflecting evolutionary information, physicochemical values, structure information, etc. In order to incorporate the word properties into BGLMs, the other 29 BGLMs are based on word properties (12) (see Supplementary Table S9). Because these BGLMs based on word properties are able to capture the physicochemical properties of residues and the evolutionary information of biological sequences, they are suitable for analyzing the residue properties, sequence properties, and the evolutionary relationships, such as protein remote homology detection (27), N6-Methyladenosine Sites (28), etc.

Biological statistical language models (BSLMs). In linguistics, the statistical language models (SLMs) reflect statistical rules of languages by using the distribution functions based on the statistical principles (29). As a result, the underlying intentions and topics of languages can be discovered. Inspired by SLMs, the biological statistical language models (BSLMs) are introduced to recognize the statistical rules of biological sequences based on bag-of-words (BOW) (see Table 3), term frequency–inverse document frequency (TF-IDF) (30) (see Table 4), TextRank (31) (see Table 5), and topic models (32) (see Table 6). In this study, Kmer (33), RevKmer (33–35), Mismatch (36–38) and Subsequence (36,38,39) are treated as the words of DNA. Particularly, RevKmer is able to capture the characteristics of two strands of the double helix of DNA sequences. Kmer (40), Mismatch (36–38) and Subsequence (36,38,39) are considered as the words of RNA; Kmer (41), Mismatch (36–38), Top-n-gram (42), Distance Residue (DR) (43) and Distance Top-n-gram (DT) (43) are considered as the words of proteins. Please note that Top-n-gram and DT are the words with the evolutionary information. BOW model represents sentences as the ‘bag’ of words by word occurrence frequencies, ignoring grammar and word orders (44). Therefore, these BSLMs based on BOW are suitable for analyzing simple functions of biological sequences, such as human nucleosome occupancy prediction (34), gene regulatory sequence prediction (35), etc. TF-IDF model (45) reflects the importance of words to the biological sequences. TextRank (31), a graph-based ranking model, recognizes key sentences by ranking the criticality of sentences in the text, and assigns higher weights indicating the influence of a word. Because both the TF-IDF model and TextRank model are able to detect the key features of the biological sequences and reduce the dimensions of the feature vectors, they are suitable for constructing efficient predictors for sequence-level analysis tasks, such as RNA-binding protein prediction (46), protein–protein interaction prediction (47), etc. These three models are performed on the words of biological sequences, and generate 12 BSLMs based on BOW (see Table 3), 12 BSLMs based on TF-IDF (see Table 4) and 12 BSLMs based on TextRank (see Table 5). Furthermore, the topic model discovers the abstract ‘topics’ and the latent semantic structures of a ‘sequence document’ by using Latent Semantic Analysis (LSA) (48), Probabilistic Latent Semantic Analysis (PLSA) (32), Latent Dirichlet Allocation (LDA) (49) and Labeled-Latent Dirichlet Allocation (Labeled-LDA) (50), leading to 12 BSLMs based on topic models (see Table 6).

Biological neural language models (BNLMs). In linguistics, the neural language models (NLMs) (51) employ deep neural networks to generate the distributed representations of words. Compared with other language models, the NLMs have the following advantages: (i) deep neural networks capture the local and global distance dependencies in a language; (ii) the distributed representation of words effectively avoids the problems of data sparse and dimensional disasters; (iii) the distributed representation of words captures the dependencies in a high-dimensional continuous space, leading to a better generalization ability. In order to incorporate these advantages into the biological language models, we introduce the biological neural language models (BNLMs) based on word embedding (Table 7) and

Table 3. Twelve BSLMs based on BOW

Category	Model	Description
DNA	Kmer-BOW	Kmer-based BOW (33)
	RevKmer-BOW	Reverse-complementary-Kmer-based BOW (33–35)
	Mismatch-BOW	Mismatch-based BOW (36–38)
	Subsequence-BOW	Subsequence-based BOW (36,38,39)
RNA	Kmer-BOW	Kmer-based BOW (40)
	Mismatch-BOW	Mismatch-based BOW (36–38)
	Subsequence-BOW	Subsequence-based BOW (36,38,39)
Protein	Kmer-BOW	Kmer-based BOW (41)
	Mismatch-BOW	Mismatch-based BOW (37)
	DR-BOW	Distance-Residue-based BOW (43)
	Top-n-gram-BOW	Top-n-gram-based BOW (42)
	DT-BOW	Distance-Top-n-gram-based BOW (43)

Table 4. Twelve BSLMs based on TF-IDF

Category	Model	Description
DNA	Kmer-TF-IDF	Kmer-based TF-IDF (30,33)
	RevKmer-TF-IDF	Reverse-complementary-Kmer-based TF-IDF (30,33–35)
	Mismatch-TF-IDF	Mismatch-based TF-IDF (30,36–38)
	Subsequence-TF-IDF	Subsequence-based TF-IDF (30,36,38,39)
RNA	Kmer- TF-IDF	Kmer-based TF-IDF (30,40)
	Mismatch-TF-IDF	Mismatch-based TF-IDF (30,36–38)
	Subsequence-TF-IDF	Subsequence-based TF-IDF (30,36,38,39)
Protein	Kmer-TF-IDF	Kmer-based TF-IDF (30,41)
	Mismatch-TF-IDF	Mismatch-based TF-IDF (30,37)
	DR-TF-IDF	Distance-Residue-based TF-IDF (30,43)
	Top-n-gram-TF-IDF	Top-n-gram-based TF-IDF(30,42)
	DT-TF-IDF	Distance-Top-n-gram-based TF-IDF (30,43)

Table 5. Twelve BSLMs based on TextRank

Category	Model	Description
DNA	Kmer-TextRank	Kmer-based TextRank (31,33)
	RevKmer-TextRank	Reverse-complementary-Kmer-based TextRank (31,33–35)
	Mismatch-TextRank	Mismatch-based TextRank (31,36–38)
	Subsequence-TextRank	Subsequence-based TextRank (31,36,38,39)
RNA	Kmer-TextRank	Kmer-based TextRank (31,40)
	Mismatch-TextRank	Mismatch-based TextRank (31,36–38)
	Subsequence-TextRank	Subsequence-based TextRank (31,36,38,39)
Protein	Kmer-TextRank	Kmer-based TextRank (31,41)
	Mismatch-TextRank	Mismatch-based TextRank (31,37)
	DR-TextRank	Distance-Residue-based TextRank (31,43)
	Top-n-gram-TextRank	Top-n-gram-based TextRank (31,42)
	DT-TextRank	Distance-Top-n-gram-based TextRank (31,43)

Table 6. Twelve BSLMs based on topic models

Algorithm	Model	Description
LSA	BOW-LSA	Latent Semantic Analysis (48)
	TF-IDF-LSA	
	TextRank-LSA	
LDA	BOW-LDA	Latent Dirichlet Allocation (49)
	TF-IDF-LDA	
	TextRank-LDA	
Labeled-LDA	BOW-Labeled-LDA	Labeled Latent Dirichlet Allocation Model (50)
	TF-IDF-Labeled-LDA	
	TextRank-Labeled-LDA	
PLSA	BOW-PLSA	Probabilistic Latent Semantic Analysis (32)
	TF-IDF-PLSA	
	TextRank-PLSA	

automatic features (Table 8). Because linguistic objects with similar distributions have similar meanings (52), word embedding embeds each word into a continuous real-valued vector to represent the words. In this study, word2vec (53), GloVe (54) and fastText (55) are combined with the aforementioned words of biological sequences, and the corresponding 36 BNLMs based on word embedding are listed in Table 7. Deep learning techniques are able to automatically extract the linguistic features independent from grammar rules and other experience knowledge. Because deep learning techniques require sufficient samples to train the predictive models with high performance, BNLMs are suitable for analyzing both the residue-level and sequence-level tasks with enough training samples, such as protein structure prediction (9), protein fold recognition (56), disordered region prediction (57), etc. In this study, autoencoder (58), CNN-BiLSTM (56) and DCNN-BiLSTM (56) are used to model the dependencies among residues/words in biological sequences. MotifCNN (59) and MotifDCNN (59) are used to capture the motif-based features. Finally, five BNLMs based on automatic features are shown in Table 8.

Biological semantic similarity language models (BSSLMs). Calculation of the sequence similarities of biological sequences is one of the keys in biological sequence analysis, which can be considered as the semantic similarities among sentences. The biological semantic similarity language models (BSSLMs) are able to represent the biological sequences based on the semantic similarities. The semantic similarities can be calculated by the feature vectors generated by the aforementioned three kinds of BLMs via Euclidean Distance (60–62), Manhattan Distance (63), Chebyshev Distance (64), Hamming Distance (65), Cosine Similarity (60–62), Pearson Correlation Coefficient (60–62), KL Divergence (Relative Entropy) (60–62), or Jaccard Similarity Coefficient (60–62). The resulting 8 BSSLMs are listed in Table 9. Because the BSSLMs are able to accurately calculate the similarities among biological sequences, they are suitable for analyzing the relationships among biological sequences and the associations between diseases and biological sequences, such as homology detection (60), non-coding RNA-disease association identification (66), etc.

Table 7. Thirty-six BNLMs based on word embedding

Category	Algorithm	Model	Description
DNA	word2vec	Kmer2vec	Learn word representations via word2vec model (53)
		RevKmer2vec	
		Mismatch2vec	
		Subsequence2vec	
	GloVe	Kmer-GloVe	Learn word representations via Glove model (54)
		RevKmer-GloVe	
		Mismatch-GloVe	
		Subsequence-GloVe	
	fastText	Kmer-fastText	Learn word representations via fastText model (55)
		RevKmer-fastText	
		Mismatch-fastText	
		Subsequence-fastText	
RNA	word2vec	Kmer2vec	Learn word representations via word2vec model (53)
		Mismatch2vec	
		Subsequence2vec	
	GloVe	Kmer-GloVe	Learn word representations via Glove model (54)
		Mismatch-GloVe	
		Subsequence-GloVe	
	fastText	Kmer-fastText	Learn word representations via fastText model (55)
		Mismatch-fastText	
		Subsequence-fastText	
Protein	word2vec	Kmer2vec	Learn word representations via word2vec model (53)
		Mismatch2vec	
		DR2vec	
		Top-n-gram2vec	
		DT2vec	
	GloVe	Kmer-Glove	Learn word representations via glove model (54)
		Mismatch-Glove	
		DR-Glove	
		Top-n-gram-Glove	
		DT-Glove	
	fastText	Kmer-fastText	Learn word representations via fastText model (55)
		Mismatch-fastText	
		DR-fastText	
		Top-n-gram-fastText	
		DT-fastText	

Table 8. Five BNLMs based on automatic features

Model	Description
MotifCNN	CNN construction with motifs initializing convolution kernel (59)
MotifDCNN	DCNN construction with motifs initializing convolution kernel (59)
CNN-BiLSTM	Combine CNN and BiLSTM (56)
DCNN-BiLSTM	Combine DCNN and BiLSTM (56)
Autoencoder	Learning Sequence Representations based on Autoencoders (58)

Table 9. Eight BSSLMs

Model	Description
ED	Euclidean Distance (60–62)
MD	Manhattan Distance (63)
CD	Chebyshev Distance (64)
HD	Hamming Distance (65)
CS	Cosine Similarity (60–62)
PCC	Pearson Correlation Coefficient (60–62)
KLD	KL Divergence (Relative Entropy) (60–62)
JSC	Jaccard Similarity Coefficient (60–62)

Extension of BLMs to BioSeq-BLM system

As introduced above, the BLMs represent the biological sequences in different aspects. We extend the BLMs to BioSeq-BLM system, making BLMs not only represent the biological sequences but also analyze the biological sequences, which is even out of the reach of any existing language model in linguistics. To achieve this goal, three other functions are added into BLMs, including predictor construction, performance evaluation, and result analysis, which will be introduced in the following sections. The overall flowchart of BioSeq-BLM is shown in Figure 2.

Predictor construction. We extend the BLMs to analyze the biological sequences by combining machine learning classifiers, which can be divided into three categories: classification algorithms, sequence labelling algorithm and deep learning algorithms.

For classification algorithms, the Support Vector Machine (SVM) (67) and Random Forest (RF) (68) are employed. They are widely used in classification tasks and regression tasks because of their good generalization ability (69). For the sequence labelling algorithm, the Conditional Random Field (CRF) (70) is used for the residue-level analysis tasks. Compared with the classification algorithms, CRF is able to model the biological sequences in a global fashion by considering the dependency information of all the residues along the sequences. For deep learning algorithms, the convolutional neural network (CNN) (57) captures the localized semantic association features. Long short-term memory (LSTM) (71) and Gated recurrent units (GRU) (72) capture the long-term dependence features of

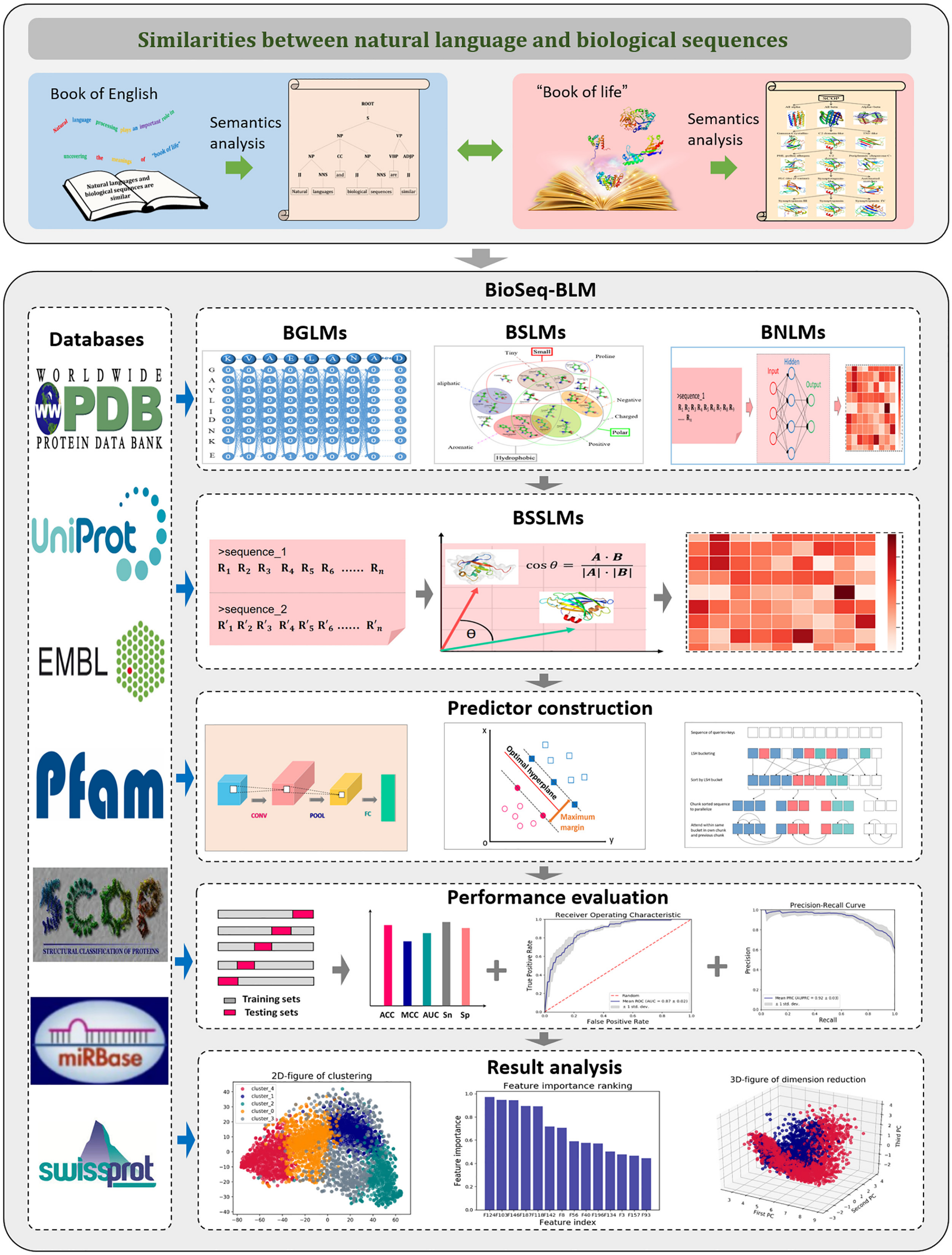


Figure 2. The main components and their relationships of BioSeq-BLM. Inspired by the similarities between the natural languages and biological sequences, the BioSeq-BLM is constructed. There are four main components in BioSeq-BLM, including Biological Language Models (BLMs), predictor construction, performance evaluation and result analysis.

sequences. Transformer (73), weighted transformer (74) and reformer (75) capture the dependencies at any distances in sequences. Compared with the classification algorithms and sequence labelling algorithms, deep learning algorithms can learn the deeper representation of the sequences and model more complex interactions, leading to better performance for the sequence analysis task.

Performance evaluation. Here two methods are employed to evaluate the performance of BioSeq-BLM, including *N*-fold cross-validation and independent test. 9 metrics are used to measure the performance of BioSeq-BLM for binary classification tasks, calculated by:

$$\begin{cases} \text{Acc} = \frac{TP + TN}{TP + FN + TN + FP} & 0 \le \text{Acc} \le 1 \\ \text{MCC} = \frac{TP * TN - FP * FN}{\sqrt{(TP+FN)(TN+FN)(TP+FP)(TN+FP)}} & -1 \le \text{MCC} \le 1 \\ \text{AUC : Area Under ROC Curve} & 0 \le \text{AUC} \le 1 \\ \text{Sn} = \frac{TP}{TP + FN} & 0 \le \text{Sn} \le 1 \\ \text{Sp} = \frac{TN}{TN + FP} & 0 \le \text{Sp} \le 1 \\ \text{Balanced Accuracy} = (\text{Sn} + \text{Sp})\ /2 & 0 \le \text{Balanced Accuracy} \le 1 \\ \text{Precision} = \frac{TP}{TP+FP} & 0 \le \text{Precision} \le 1 \\ \text{AUPR : Area Under PR Curve} & 0 \le \text{AUPR} \le 1 \\ F1 = \frac{2*Precision*Recall}{Precision+Recall} & 0 \le \text{F1} \le 1 \end{cases} \quad (5)$$

where TP represents the number of true positive samples; TN represents the number of true negative samples; FP represents the number of false positive samples; FN represents the number of false negative samples. For multiclass classification tasks, multi-classification accuracy (12) is used, calculated by:

$$\text{Acc}\,(i) = 1 - \frac{N_{-}^{+}(i) + N_{+}^{-}(i)}{N^{+}(i) + N^{-}(i)} \quad 0 \le \text{Acc}\,(i) \le 1 \quad (6)$$

where $N^{+}(i)$ represents the total number of the samples in the *i*-th class, $N_{-}^{+}(i)$ is the number of the samples in the *i*-th class wrongly predicted as the other classes, $N^{-}(i)$ represents the total number of the samples not in the *i*-th class and $N_{+}^{-}(i)$ is the number of the samples not in the *i*-th class wrongly predicted to be the *i*-th class.

The selection of performance measures is generally based on the characteristics of datasets. For most biological sequence analysis tasks, Acc, MCC, AUC and Balanced Accuracy are the most commonly used metrics for performance evaluation. For a balanced dataset with approximately equal number of samples for each label, Acc metric can accurately evaluate the performance of a predictor. For an unbalanced dataset, MCC, AUC and Balanced Accuracy can better evaluate the performance of the predictors. For example, MCC and AUC are used to evaluate the performance of different predictors for identification of intrinsically disordered regions in proteins (76) because of the imbalance of samples.

BioSeq-BLM trained with imbalanced benchmark datasets will bias the class with fewer samples. In this regard, the sampling techniques are provided to solve this problem, including over-sampling method Synthetic Minority Oversampling Technique (SMOTE) (77), under-sampling method Tomek links (78) and the combination of over-sampling and under-sampling (79).

Result analysis. We provide a result analysis framework to interpret the predictive results with four modules: normalization, clustering, feature selection and dimension reduction. L1 regularization (80), L2 regularization (81), MinMaxScaler (82) and StandardScaler (82) in the normalization module can be used to normalize the features. Clustering module provides 5 cluster algorithms to visualize and validate if the corresponding BLM is able to accurately represent the data, including K-means (83), affinity propagation algorithms (84), Density-Based Spatial Clustering of Applications with Noise algorithm (DBSCAN) (85), Gaussian mixture model (86) and Agglomerative Nesting (87). Feature selection module provides 5 methods to analyze the importance of the features generated by BLMs, including chi-square (88,89), *F*-value (88,89), mutual information (88,89), recursive feature elimination (90) and tree mode (91). Three dimension reduction methods are incorporated into the dimension reduction module to remove the noise and reduce the dimensions of the feature vectors, including principal component analysis (PCA) (92), kernel principal component analysis (93) and truncated singular value decomposition (TSVD) (94).

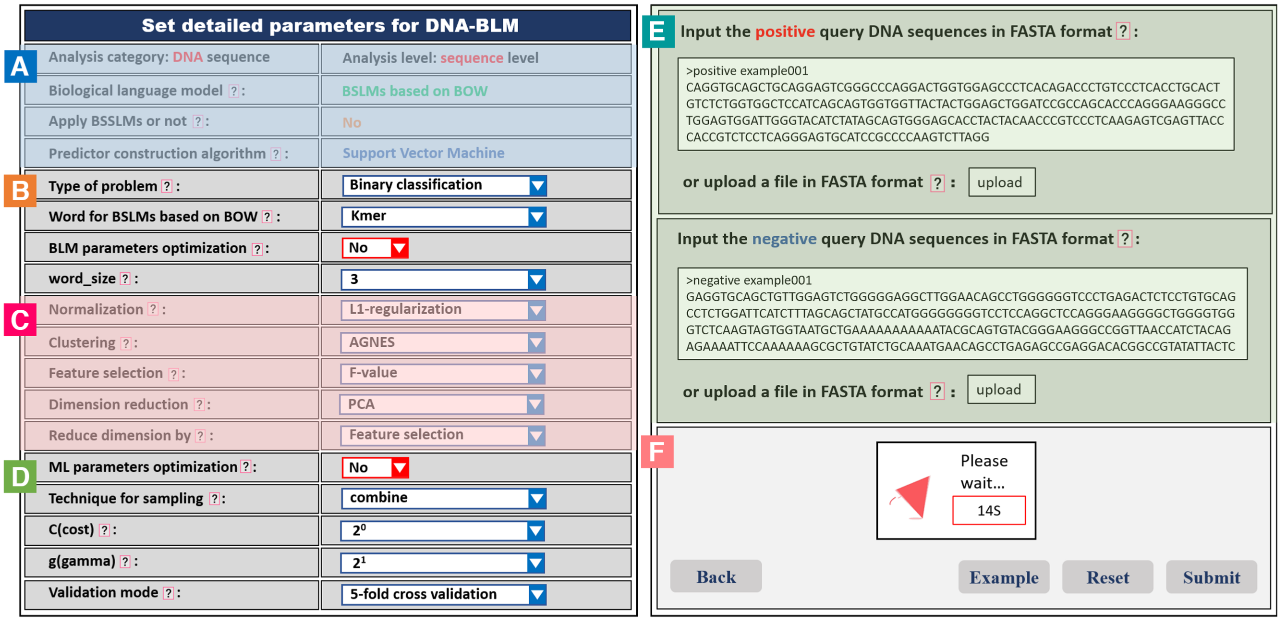
BioSeq-BLM web server and stand-alone package

In order to help the researchers to use the biological language models for biological sequence analysis, we establish the web server and stand-alone tool of BioSeq-BLM, which can be freely accessed from http://bliulab.net/BioSeq-BLM/.

Web server. After clicking the 'Server' tab, three kinds of BLMs (DNA-BLM, RNA-BLM and protein-BLM for DNA, RNA and protein sequence analysis, respectively) will be shown on the screen, and then the level of analysis (residue-level analysis and sequence-level analysis) and BLM should be selected. Next, choose to calculate semantic similarity based on BSSLMs or not for sequence-level analysis. After selecting the machine learning algorithm, the submit page will be shown on the screen (see Figure 3A–D), where the users should set the parameters of the predictors, type the datasets in FASTA format into the input box or upload FASTA files (see Figure 3E and F). Finally click the 'Submit' for calculation. The results will be shortly shown on the screen (see Figure 4).

In this example, the BLM is set as Kmer-BOW (see Table 3) combined with the SVM for DNA sequence analysis at sequence level. The results page contains six parts as shown in Figure 4, from which the users can easily see the performance of the BLM and the importance of different features. This information is a key for selecting the BLM for biological sequence analysis. Please note that for BLMs based on deep learning techniques with high computational costs, the command lines of the stand-alone package will be given, based on which the users can easily obtain the corresponding results with the help of the stand-alone package installed in their own computers.

Stand-alone package. The web server of BioSeq-BLM is easy to use. However, for high-throughput analysis, its computational cost is high, especially for the BLMs based on deep learning techniques. In this regard, the stand-alone package of BioSeq-BLM is provided, which can be downloaded from http://bliulab.net/BioSeq-BLM/



Set detailed parameters for DNA-BLM

Analysis category: DNA sequence	Analysis level: sequence level
Biological language model ? :	BSLMs based on BOW
Apply BSSLMs or not ? :	No
Predictor construction algorithm ? :	Support Vector Machine
Type of problem ? :	Binary classification
Word for BSLMs based on BOW ? :	Kmer
BLM parameters optimization ? :	No
word_size ? :	3
Normalization ? :	L1-regularization
Clustering ? :	AGNES
Feature selection ? :	F-value
Dimension reduction ? :	PCA
Reduce dimension by ? :	Feature selection
ML parameters optimization ? :	No
Technique for sampling ? :	combine
C(cost) ? :	2^0
g(gamma) ? :	2^1
Validation mode ? :	5-fold cross validation

Input the positive query DNA sequences in FASTA format ? :

```
>positive example001
CAGGTGCAGCTGCAGGAGTCGGGCCCAGGACTGGTGGAGCCCTCACAGACCCTGTCCCTCACCTGCACT
GTCTCTGGTGGCTCCATCAGCAGTGGTGGTTACTACTGGAGCTGGATCCGCCAGCACCCAGGGAAGGGCC
TGGAGTGGATTGGGTACATCTATAGCAGTGGGAGCACCTACTACAACCCGTCCCTCAAGAGTCGAGTTACC
CACCGTCTCCTCAGGGAGTGCATCCGCCCCAAGTCTTAGG
```

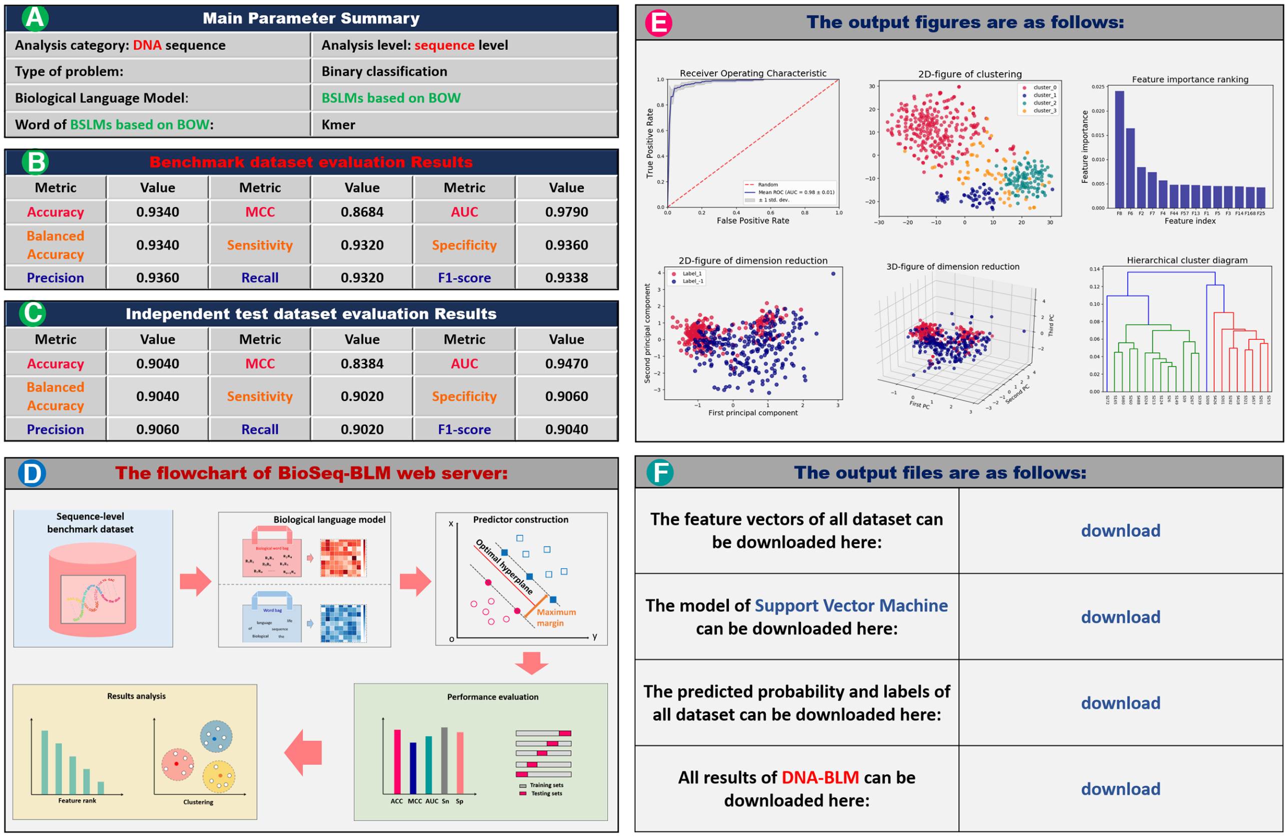
or upload a file in FASTA format ? : upload

Input the negative query DNA sequences in FASTA format ? :

```
>negative example001
GAGGTGCAGCTGTTGGAGTCTGGGGGAGGCTTGGAACAGCCTGGGGGGTCCCTGAGACTCTCCTGTGCAG
CCTCTGGATTCATCTTTAGCAGCTATGCCATGGGGGGGGTCCTCCAGGCTCCAGGGAAGGGGCTGGGGTGG
GTCTCAAGTAGTGGTAATGCTGAAAAAAAAAAAATACGCAGTGTACGGGAAGGGCCGGTTAACCATCTACAG
AGAAAATTCCAAAAAAGCGCTGTATCTGCAAATGAACAGCCTGAGAGCCGAGGACACGGCCGTATATTACTC
```

or upload a file in FASTA format ? : upload

Figure 3. Screenshot of the input page of BioSeq-BLM web server. (**A**) A summary of the main parameters; (**B**) the parameters for BLM; (**C**) the parameters of result analysis; (**D**) the parameters for predictor construction and performance evaluation; (**E**) the input box of the datasets; (**F**) the functional buttons.



Main Parameter Summary

Analysis category: DNA sequence	Analysis level: sequence level
Type of problem:	Binary classification
Biological Language Model:	BSLMs based on BOW
Word of BSLMs based on BOW:	Kmer

Benchmark dataset evaluation Results

Metric	Value	Metric	Value	Metric	Value
Accuracy	0.9340	MCC	0.8684	AUC	0.9790
Balanced Accuracy	0.9340	Sensitivity	0.9320	Specificity	0.9360
Precision	0.9360	Recall	0.9320	F1-score	0.9338

Independent test dataset evaluation Results

Metric	Value	Metric	Value	Metric	Value
Accuracy	0.9040	MCC	0.8384	AUC	0.9470
Balanced Accuracy	0.9040	Sensitivity	0.9020	Specificity	0.9060
Precision	0.9060	Recall	0.9020	F1-score	0.9040

The output figures are as follows:

The output files are as follows:

The feature vectors of all dataset can be downloaded here:	download
The model of Support Vector Machine can be downloaded here:	download
The predicted probability and labels of all dataset can be downloaded here:	download
All results of DNA-BLM can be downloaded here:	download

Figure 4. Screenshot of the result page of BioSeq-BLM web server. It contains six sections: (**A**) the summary of main parameters; (**B**) and (**C**) the evaluation results; (**D**) the flowchart; (**E**) the output figures; (**F**) the output files.

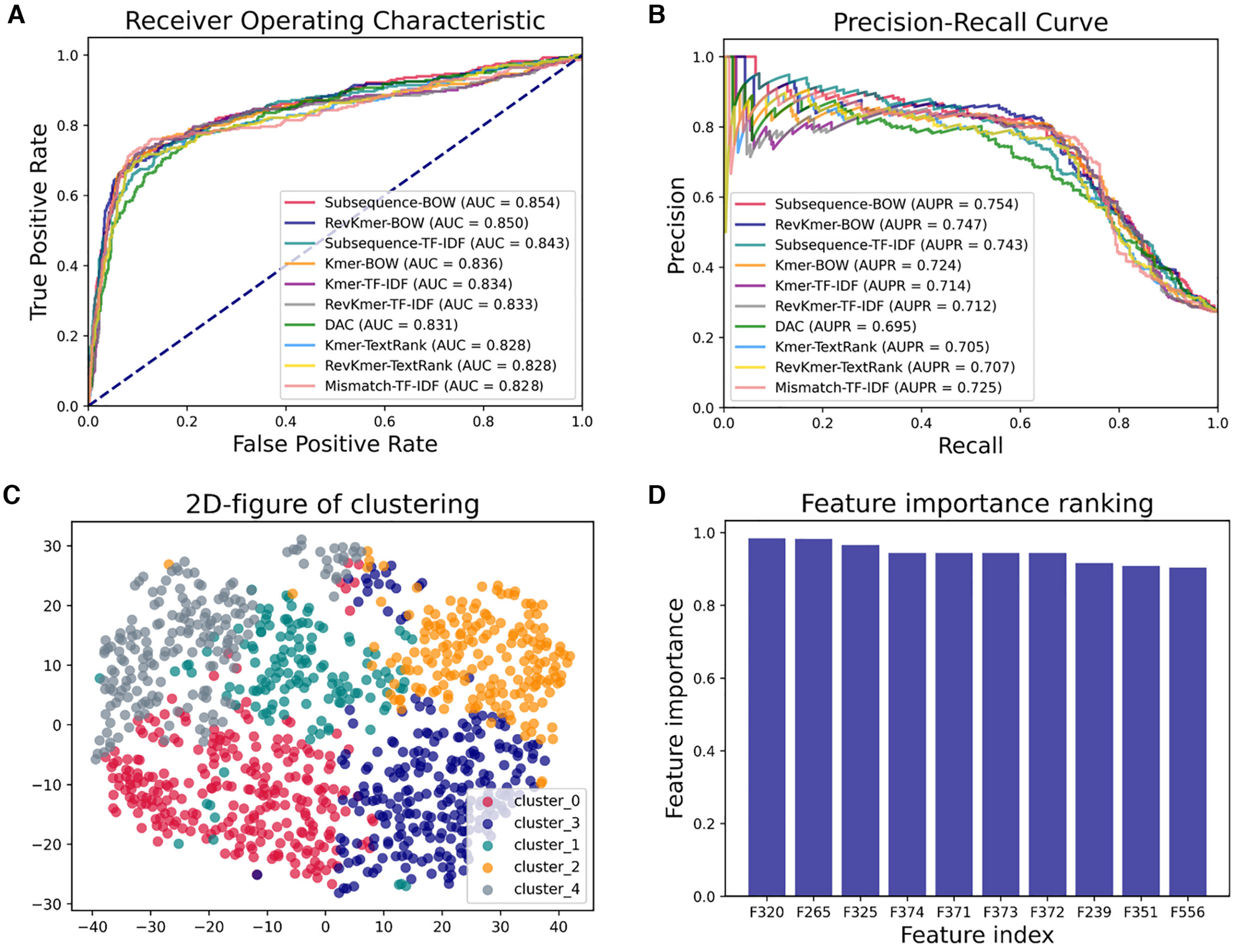


Figure 5. (**A**) The Receiver Operating Characteristic (ROC) curves of top 10 predictors constructed by SVMs and different BLMs for identification DNase I hypersensitive sites; (**B**) the corresponding Precision-Recall (PR) curves of these 10 predictors; (**C**) the clustering results by *K*-means algorithm; (**D**) the 10 most important features and their corresponding feature importance values evaluated in terms of *F*-value.

download/. Different from web server, the stand-alone package based on multithreading and GPU acceleration can make full use of local computing resources to implement computing. The trained models generated by BioSeq-BLM can be loaded to predict the unknown samples by using the stand-alone package of BioSeq-BLM. Furthermore, the bash scripts for automatically selecting the best models for specific biological sequence analysis tasks are incorporated into the stand-alone package, which will be particularly useful for biologists to choose the suitable models. For more information, please refer to the manual, which can be accessed at http://bliulab.net/BioSeq-BLM/static/download/BioSeq-BLM_manual.pdf.

How to choose web server or stand-alone package. The Stand-alone package and web server are complementary. The choice of web server and stand-alone package is related to the number of input sequences and the selected model. For a small number of sequences, web server is generally recommended. When there are many input sequences, it is recommended to use stand-alone package for calculation. If we need to use the model based on deep learning for biological sequence analysis, we suggest the users to use the web server to generate the command lines, and then use the corresponding command lines to run the stand-alone package. If users want to batch select the best BLM and machine learning algorithm for a specific task, the stand-alone package provides relevant scripts to facilitate the relevant functions.

RESULTS AND DISCUSSION

The BioSeq-BLM incorporates 155 different BLMs for biological sequence analysis. In this section, we will show how to use BioSeq-BLM to solve some specific biological sequence analysis tasks, which will be particularly helpful for researchers to select the BLMs.

Identification DNase I hypersensitive sites

Identification of DNase I hypersensitive sites (DHSs) is important for understanding the functions of noncoding genomic regions (95), which is a DNA sequence analysis task at sequence level. Here, we will show how to construct

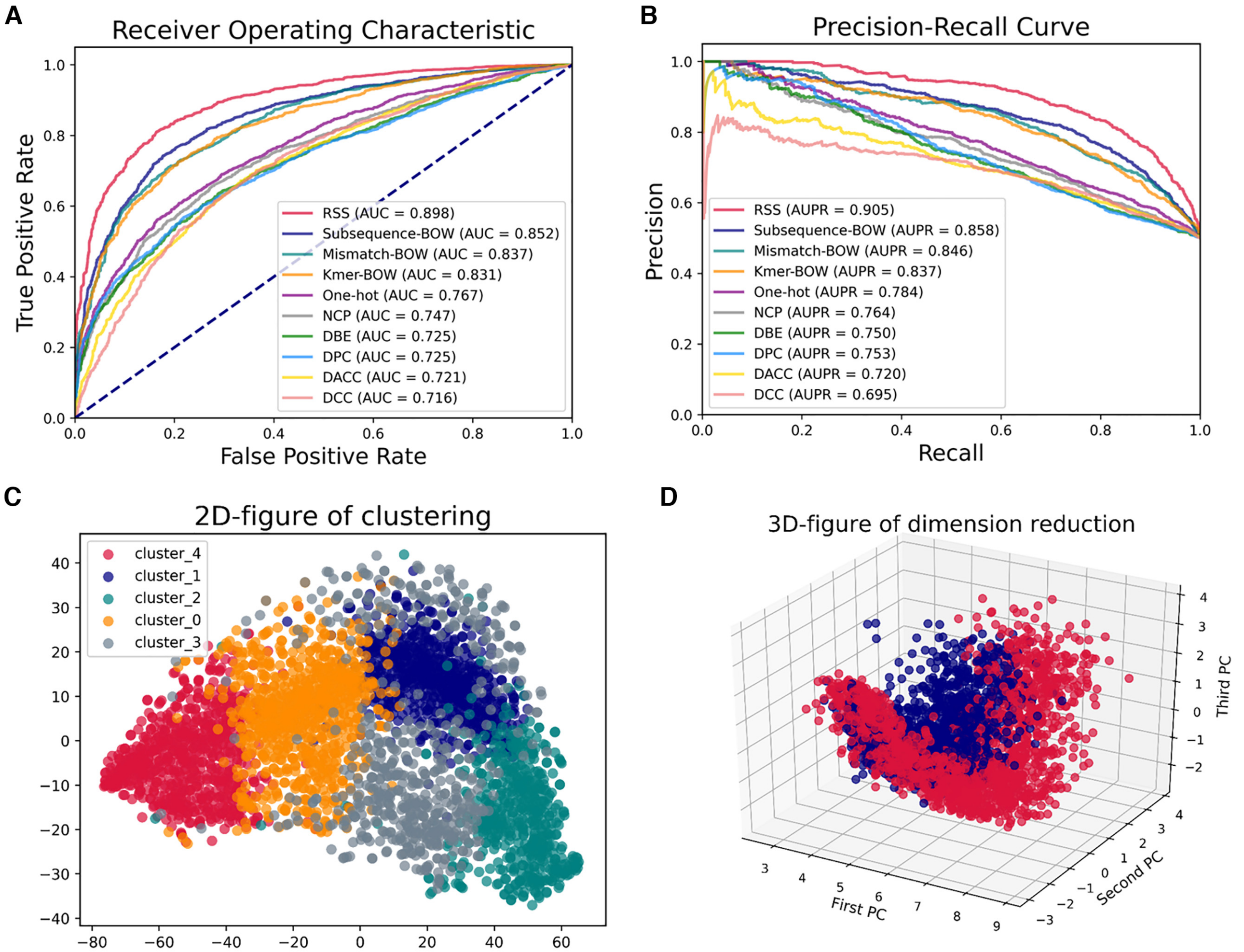


Figure 6. (**A**) The ROC curves of top 10 predictors constructed by SVMs and different BLMs for identification of real microRNA precursors; (**B**) the corresponding PR curves of these 10 predictors; (**C**) the clustering result for *K*-means algorithm; (**D**) the 3D-figure for dimension reduction when applying TSVD method.

computational predictors for this task based on different BLMs and SVMs with the help of the stand-alone package of BioSeq-BLM. The benchmark dataset (95) downloaded from http://bliulab.net/iDHS-EL/data is used as the inputs of BioSeq-BLM. For example, a predictor Subsequence-BOW combines the BSLM of Subsequence-BOW (see Table 3) and SVM can be easily constructed with the help of the stand-alone package with the following command line:

```
python BioSeq-BLM_Seq.py -category DNA -mode BOW -words
Subsequence -word_size 3 -cl Kmeans -nc 5 -fs F-value -nf 128 -rdb fs -ml
SVM -sp combine -seq_file pos_file neg _file -label + 1 -1
```

The performance of the top 10 best predictors generated by BioSeq-BLM is shown in Figure 5 and Supplementary Table S1, from which we can see that Subsequence-BOW predictor is highly comparable with the state-of-the-art predictor iDHS-EL reported in (95). The BOW model based on Subsequence words (36,38,39) is able to extract the discriminative features leading to better performance. All the complicated processes for constructing a computational predictor can be easily implemented by BioSeq-BLM with only one command line. Furthermore, the different BLMs incorporated in BioSeq-BLM would be potential candidates for constructing more efficient predictors for this task.

Identification of real microRNA precursors

As miRNAs are deeply implicated with many cancers and other diseases, it is important for both basic research and miRNA-based therapy to discriminate the real pre-miRNAs from the false ones (96), which is a RNA sequence analysis task at sequence level. Given the benchmark dataset (96), the RSS predictor for miRNA prediction based on the BGLM of RSS (RNA Secondary Structure) (97) and SVM can be easily constructed with the help of BioSeq-BLM by using the following command line:

```
python BioSeq-BLM_Seq.py -category RNA -mode OHE -method RSS
-cl Kmeans -nc 5 -dr TSVD -np 128 -rdb dr -ml SVM -seq_file pos_file
neg _file -label + 1 -1
```

The performance of the top 10 best predictors generated by BioSeq-BLM is shown in Figure 6 and Supplementary Table S2. Because the BGLM of RSS can capture the 'hairpin' characteristics of miRNAs in their secondary

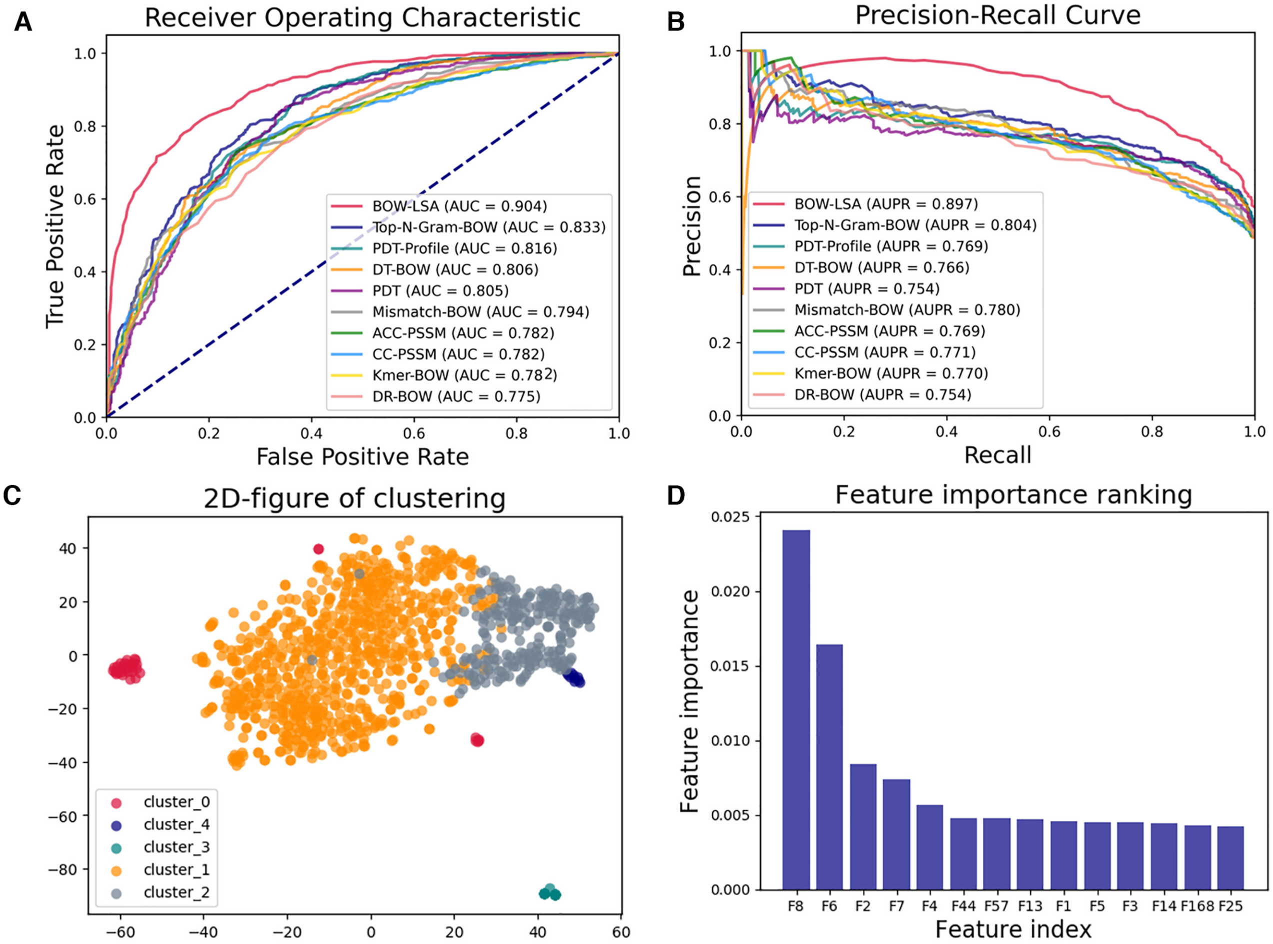


Figure 7. (**A**) The ROC curves of top 10 predictors constructed by RFs and different BLMs for identification of DNA binding proteins; (**B**) the corresponding PR curves of these 10 predictors; (**C**) the clustering result for *K*-means algorithm; (**D**) the 14 most important features and their corresponding feature importance evaluated by tree-based feature selection.

structures, the computational predictor RSS generated by BioSeq-BLM achieves comparable performance with the iMcRNA predictor reported in (96), further confirming the usefulness of BioSeq-BLM for RNA sequence analysis.

Identification of DNA-binding proteins and RNA-binding proteins

Identification of DNA-binding proteins (DBPs) and RNA-binding proteins (RBPs) are important protein sequence analysis tasks at the sequence level. Identification of DBPs and RBPs play important roles in biological processes, such as replication, translation and transcription of genetic material.

A predictor BOW-LSA for DBP prediction can be generated by BioSeq-BLM by combining the BSLM of BOW-LSA (see Table 6) and RF with the following command line:

```
python BioSeq-BLM_Seq.py -category Protein -mode TM -method LSA
-in_tm BOW -words Top-N-Gram -top_n 2 -com_prop 0.7 -cl Kmeans -nc
5 -fs Tree -nf 128 -ml RF -seq_file pos_file neg _file -label + 1 -1
```

A predictor PDT-Profile for RBP prediction can be generated by BioSeq-BLM by combining the BGLM of PDT-Profile (see Table 2) and SVM with the following command line:

```
python BioSeq-BLM_Seq.py -category Protein -mode SR -method
PDT-Profile -ml SVM -seq_file pos_file neg _file -label + 1 -1
```

Evaluated on the benchmark dataset (98) for DBP identification, the performance of the top 10 best predictors generated by BioSeq-BLM is shown in Figure 7 and Supplementary Table S3. Because the BSLM of BOW-LSA is able to capture the global information, BOW-LSA shows the best performance and achieves an ACC of 81.58%, outperforming the predictor PseDNA-Pro reported in (98). Evaluated on the Salmonella benchmark dataset (46) for RBP identification, the performance of the top 10 best predictors generated by BioSeq-BLM is shown in Figure 8 and Supplementary Table S4. Because the BGLM of PDT-Profile (27) is able to efficiently extract the evolutionary information from the profiles, the PDT-Profile predictor combining PDT-Profile and SVM achieves the best performance with an AUC of 0.915, outperforming other three existing state-of-the-art predictors (TriPepSVM (46), RNAPred (99) and RBPPred (100)) by 6–13% in terms of AUC (see Supplementary Table S5). These results indicate that the

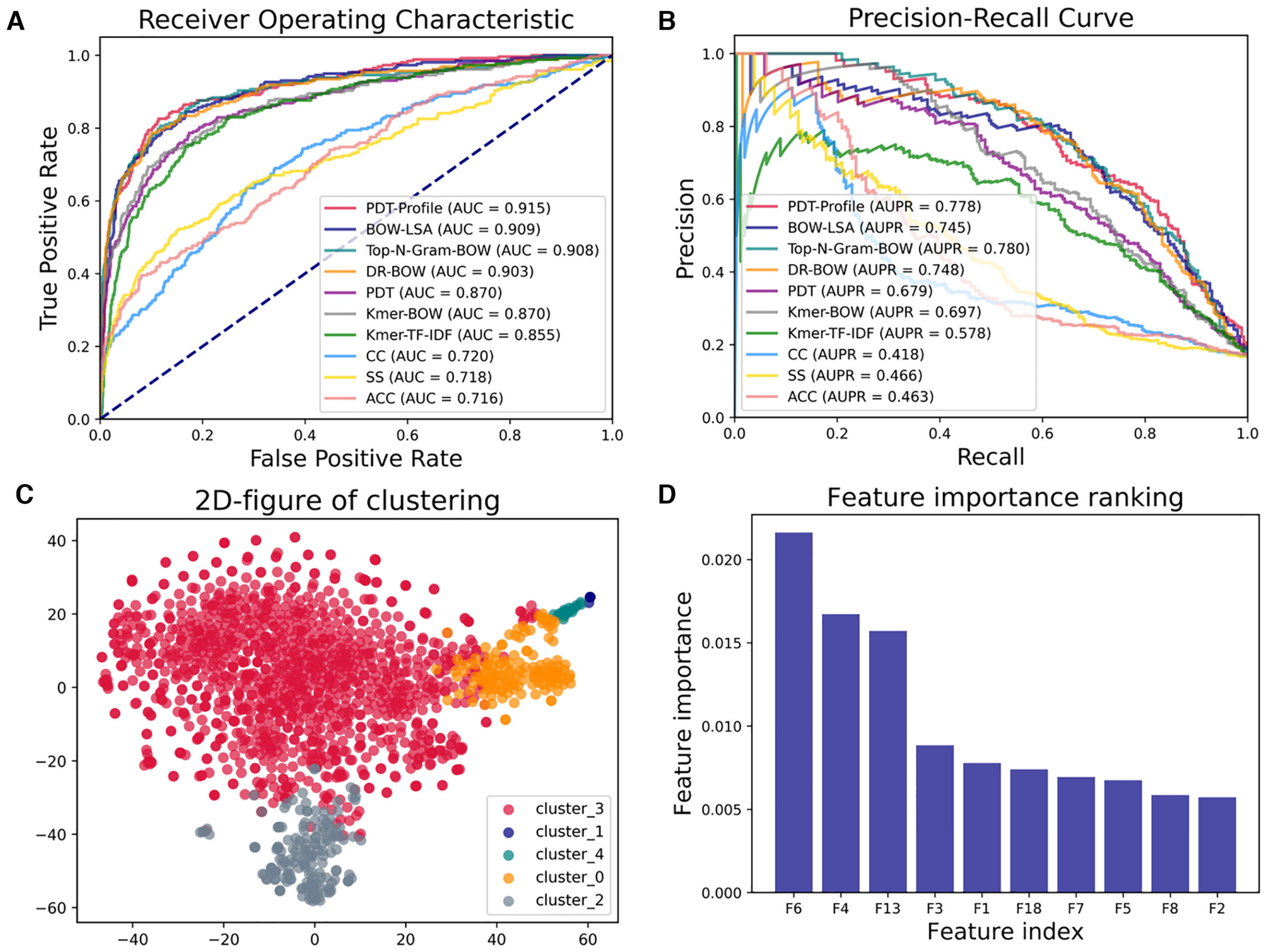


Figure 8. (**A**) The ROC curves of top 10 predictors constructed by SVMs and different BLMs for identification of RNA binding proteins; (**B**) the corresponding PR curves of these 10 predictors; (**C**) the clustering result for *K*-means algorithm; (**D**) the 10 most important features and their corresponding feature importance evaluated by tree-based feature selection.

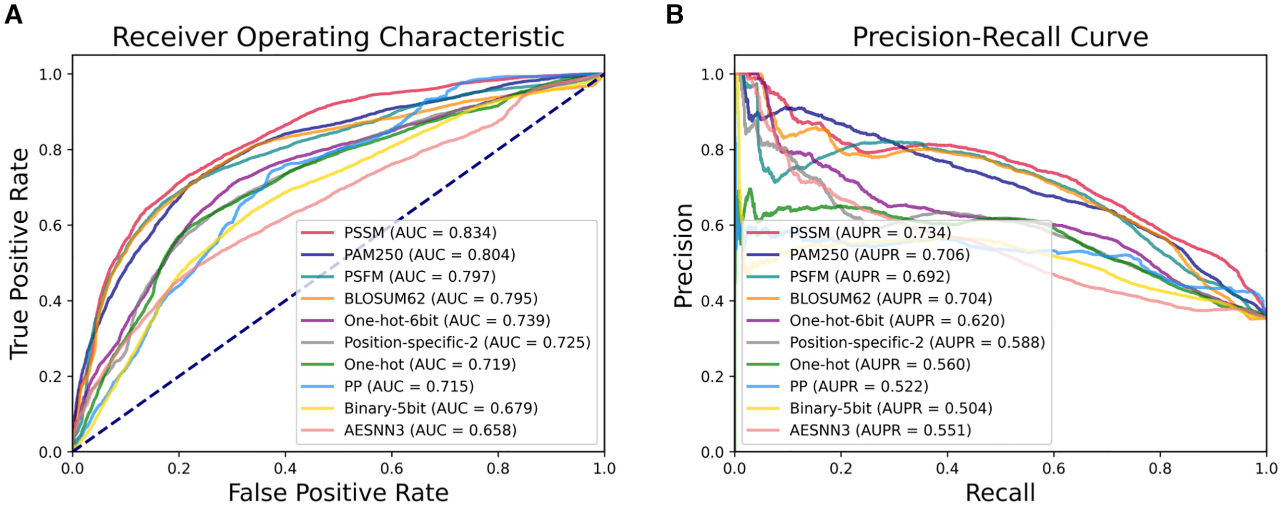


Figure 9. (**A**) The ROC curves of top 10 predictors constructed by LSTMs and different BLMs for identification of intrinsically disordered regions in proteins; (**B**) the corresponding PR curves of these 10 predictors.

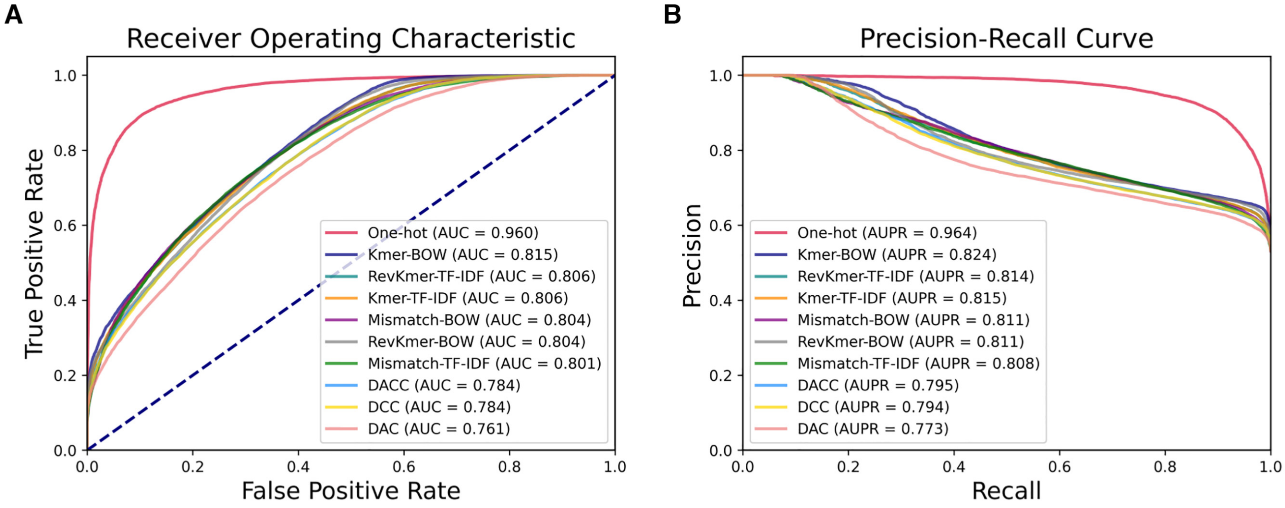


Figure 10. (**A**) The ROC curves of top 10 predictors constructed by RFs and different BLMs for RNA secondary structure prediction; (**B**) the corresponding PR curves of these 10 predictors.

predictors automatically generated by BioSeq-BLM can even achieve obviously better results than other existing approaches, which is another big step for the applications of the artificial intelligence to protein sequence analysis following the contributions of Alphfold2 (101) to the protein structure prediction.

Identification of intrinsically disordered regions in proteins

Intrinsically disordered regions (IDRs) in proteins are important for protein structure and function analysis, which is a protein sequence analysis task at residue level. BioSeq-Analysis2.0 (12) is another software tool based on machine learning techniques to automatically analyze biological sequences. In this regard, we compare the predictors constructed by BioSeq-BLM for IDR prediction with the predictors generated by BioSeq-Analysis2.0 on the benchmark dataset (76). A predictor PSSM for IDR prediction can be generated by BioSeq-BLM via combining the BLM of PSSM (102) and LSTM with the following command line:

```
python BioSeq-BLM_Res.py -category Protein -method PSSM -ml
LSTM -epoch 10 -batch_size 20 -n_layer 2 -hidden_dim 64 -seq_file
protein_seq_file -label_file protein_label_file
```

The performance of the top 10 predictors built by BioSeq-BLM is shown in Figure 9 and Supplementary Table S6. Because the LSTM captures the deep-level dependencies of residues in biological sequences in a global fashion, the PSSM predictor based on LSTM achieves the best performance among the 10 predictors. It also outperforms all the five top performing predictors generated by BioSeq-Analysis2.0 (12) by 8.7–12.6% in terms of AUC (see Supplementary Table S7). These results are not surprising because the BioSeq-BLM is based on the biological language models and deep learning methods, which are able to more accurately represent and analyze the biological sequences, and therefore, it achieves better performance than BioSeq-Analysis2.0.

RNA secondary structure prediction

Identification of RNA secondary structure is an important step to understand RNA functions, which is a residue level analysis task. For example, we can use BioSeq-BLM to generate the One-hot predictor based on BGLM of One-hot and Random Forest by using the following command line:

```
python BioSeq-BLM_Seq.py -category RNA -mode OHE -method
One-hot -ml RF -seq_file pos_file neg _file -label + 1 -1 -fixed_len 37
```

Given the benchmark dataset (PARS-Yeast dataset (103)), the performance of the top 10 predictors built by BioSeq-BLM is shown in Figure 10 and Supplementary Table S8. Because the combination of biological word properties and machine learning algorithms can improve the generalization ability of a predictor, the One-hot predictor achieves the best performance among the 10 predictors with an AUC of 0.960, which is highly comparable with the state-of-the-art predictor GRASP achieving an AUC of 0.967 (103).

CONCLUSION

As discussed above, the techniques derived from natural language processing (NLP) are the keys to uncover the meanings of the 'book of life'. As a result, with the rapid growth of the biological sequence data, the NLP techniques are playing more and more important roles in prediction of the structures and functions of these sequence data. Unfortunately, it is never an easy task to find the suitable NLP techniques to solve a specific task. In order to solve this challenging problem, in this study, we introduce 155 different BLMs for DNA, RNA and protein sequence analysis, and extend these BLMs into a system called BioSeq-BLM, which is able to automatically represent and analyze the sequence data only requiring the sequence data in FASTA format as inputs. With its help, the predictors can be easily constructed. Experimental results show that the predictor even outperforms the existing state-of-the-art approaches for specific tasks. BioSeq-BLM provides new approaches for biological sequence analysis based on the techniques from NLP, which is particularly useful for constructing the computational predictors, or at the very least, it will play a commentary role with the existing methods to contribute to the development of this very important field.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are also very much indebted to the three anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper.

FUNDING

National Key R&D Program of China [2018AAA0100100]; National Natural Science Foundation of China [61822306, 61861146002, 61732012]; Beijing Natural Science Foundation [JQ19019]. Funding for open access charge: National Key R&D Program of China [2018AAA0100100].
Conflict of Interest statement. None declared.

REFERENCES

1. Searls,D.B. (2002) The language of genes. *Nature*, **420**, 211–217.
2. Scaiewicz,A. and Levitt,M. (2015) The language of the protein universe. *Curr. Opin. Genet. Dev.*, **35**, 50–56.
3. Yu,L.J., Tanwar,D.K., Penha,E.D.S., Wolf,Y.I., Koonin,E.V. and Basu,M.K. (2019) Grammar of protein domain architectures. In: *Proc. Natl. Acad. Sci. U.S.A*. Vol. **116**, pp. 3636–3645.
4. Searls,D.B. (2001) Reading the book of life. *Bioinformatics*, **17**, 579–580.
5. Gimona,M. (2006) Protein linguistics - a grammar for modular protein assembly? *Nat. Rev. Mol. Cell Biol.*, **7**, 68–73.
6. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.*, **36**, D202–D205.
7. Friedel,M., Nikolajewa,S., Suhnel,J. and Wilhelm,T. (2009) DiProDB: a database for dinucleotide properties. *Nucleic Acids Res.*, **37**, D37–D40.
8. Chen,Z., Eavani,H., Chen,W., Liu,Y. and Wang,W.Y. (2020) Few-Shot NLG with Pre-Trained Language Model. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. pp. 183–190.
9. Senior,A.W., Evans,R., Jumper,J., Kirkpatrick,J., Sifre,L., Green,T., Qin,C., Zidek,A., Nelson,A.W.R., Bridgland,A. *et al.* (2020) Improved protein structure prediction using potentials from deep learning. *Nature*, **577**, 706–710.
10. Alipanahi,B., Delong,A., Weirauch,M.T. and Frey,B.J. (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.
11. Liu,B. (2019) BioSeq-Analysis: a platform for DNA, RNA and protein sequence analysis based on machine learning approaches. *Brief. Bioinform.*, **20**, 1280–1294.
12. Liu,B., Gao,X. and Zhang,H.Y. (2019) BioSeq-Analysis2.0: an updated platform for analyzing DNA, RNA and protein sequences at sequence level and residue level based on machine learning approaches. *Nucleic Acids Res.*, **47**, e127.
13. Chen,Z., Zhao,P., Li,F., Marquez-Lago,T.T., Leier,A., Revote,J., Zhu,Y., Powell,D.R., Akutsu,T., Webb,G.I. *et al.* (2019) iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data. *Brief. Bioinform.*, **21**, 1047–1057.
14. Xiao,N., Cao,D.S., Zhu,M.F. and Xu,Q.S. (2015) protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. *Bioinformatics*, **31**, 1857–1859.
15. Cao,D.S., Xiao,N., Xu,Q.S. and Chen,A.F. (2015) Rcpi: R/Bioconductor package to generate various descriptors of proteins, compounds and their interactions. *Bioinformatics*, **31**, 279–281.
16. Avsec,Z., Kreuzhuber,R., Israeli,J., Xu,N., Cheng,J., Shrikumar,A., Banerjee,A., Kim,D.S., Beier,T., Urban,L. *et al.* (2019) The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. *Nat. Biotechnol.*, **37**, 592–600.
17. Kopp,W., Monti,R., Tamburrini,A., Ohler,U. and Akalin,A. (2020) Deep learning for genomics using Janggu. *Nat. Commun.*, **11**, 3488.
18. Chen,K.M., Cofer,E.M., Zhou,J. and Troyanskaya,O.G. (2019) Selene: a PyTorch-based deep learning library for sequence data. *Nat. Methods*, **16**, 315–318.
19. Pereira,F., Azevedo,F., Carvalho,Â., Ribeiro,G.F., Budde,M.W. and Johansson,B. (2015) Pydna: a simulation and documentation tool for DNA assembly strategies using python. *BMC Bioinformatics*, **16**, 142.
20. Shannon,C.E. (1948) A mathematical theory of communication. *Bell Syst. Tech. J.*, **27**, 379–423.
21. Goodman,J. (2001) A bit of progress in language modeling. *Comput. Speech Lang.*, **15**, 403–434.
22. Chomsky,N. (1956) Three models for the description of language. *IRE Trans. Inf. Theory*, **2**, 113–124.
23. Zhang,J., Chen,Q.C. and Liu,B. (2020) iDRBP_MMC: identifying DNA-binding proteins and RNA-binding proteins based on multi-label learning model and motif-based convolutional neural network. *J. Mol. Biol.*, **432**, 5860–5875.
24. Dong,Q.W., Zhou,S.G. and Guan,J.H. (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. *Bioinformatics*, **25**, 2655–2662.
25. Hanson,J., Paliwal,K.K., Litfin,T. and Zhou,Y. (2019) SPOT-Disorder2: improved protein intrinsic disorder prediction by ensembled deep learning. *Genomics Proteomics Bioinformatics*, **17**, 645–656.
26. Bari,A.T., Golam,M., Reaz,M.R., Choi,H.-J. and Jeong,B.-S. (2013) DNA Encoding for Splice Site Prediction in Large DNA Sequence. In: *Proceedings of the 18th International Conference on Database Systems for Advanced Applications*. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 46–58.
27. Liu,B., Wang,X.L., Chen,Q.C., Dong,Q.W. and Lan,X. (2012) Using amino acid physicochemical distance transformation for fast protein remote homology detection. *PLoS One*, **7**, e46633.
28. Qiang,X., Chen,H., Ye,X., Su,R. and Wei,L. (2018) M6AMRFS: robust prediction of N6-methyladenosine sites with sequence-based features in multiple species. *Front. Genet.*, **9**, 495.
29. Bahl,L.R., Brown,P.F., Souza,P.V. and Mercer,R.L. (1989) A tree-based statistical language model for natural language speech recognition. *IEEE Trans. Acoust. Speech Signal Process.*, **37**, 1001–1008.
30. Zhang,W., Yoshida,T. and Tang,X. (2011) A comparative study of TF*IDF, LSI and multi-words for text classification. *Expert Syst. Appl.*, **38**, 2758–2765.
31. Mihalcea,R. and Tarau,P. (2004) Textrank: Bringing order into text. In: *Proceedings of the 2004 conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Barcelona, Spain, pp. 404–411.
32. Blei,D.M. (2012) Probabilistic topic models. *Commun. ACM*, **55**, 77–84.
33. Chen,W., Lei,T.Y., Jin,D.C., Lin,H. and Chou,K.C. (2014) PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Anal. Biochem.*, **456**, 53–60.
34. Gupta,S., Dennis,J., Thurman,R.E., Kingston,R., Stamatoyannopoulos,J.A. and Noble,W.S.J.P.C.B. (2008) Predicting human nucleosome occupancy from primary sequence. *PLoS Comput. Biol.*, **4**, e1000134.
35. Noble,W.S., Kuehn,S., Thurman,R., Yu,M. and Stamatoyannopoulos,J. (2005) Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics*, **21**, I338–I343.
36. El-Manzalawy,Y., Dobbs,D. and Honavar,V. (2008) Predicting flexible length linear B-cell epitopes. *Comput. Syst. Bioinformatics Conf.*, **7**, 121–132.
37. Leslie,C.S., Eskin,E., Cohen,A., Weston,J. and Noble,W.S. (2004) Mismatch string kernels for discriminative protein classification. *Bioinformatics*, **20**, 467–476.

38. Luo,L.Q., Li,D.F., Zhang,W., Tu,S.K., Zhu,X.P. and Tian,G. (2016) Accurate prediction of transposon-derived piRNAs by integrating various sequential and physicochemical features. *PLoS One*, **11**, e0153268.
39. Lodhi,H., Saunders,C., Shawe-Taylor,J., Cristianini,N. and Watkins,C. (2002) Text classification using string kernels. *J. Mach. Learn. Res.*, **2**, 419–444.
40. Lin,H., Deng,E.Z., Ding,H., Chen,W. and Chou,K.C. (2014) iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Res.*, **42**, 12961–12972.
41. Liu,B., Liu,F.L., Wang,X.L., Chen,J.J., Fang,L.Y. and Chou,K.C. (2015) Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Res.*, **43**, W65–W71.
42. Liu,B., Wang,X., Lin,L., Dong,Q. and Wang,X. (2008) A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis. *BMC Bioinformatics*, **9**, 510.
43. Liu,B., Xu,J., Zou,Q., Xu,R., Wang,X. and Chen,Q. (2014) Using distances between Top-n-gram and residue pairs for protein remote homology detection. *BMC Bioinformatics*, **15**, S3.
44. Harris,Z.S. (1954) Distributional structure. *Word*, **10**, 146–162.
45. Ramos,J. (2003) Using tf-idf to determine word relevance in document queries. In: *Proceedings of the First Instructional Conference on Machine Learning*. New Jersey, USA, Vol. **242**, pp. 133–142.
46. Bressin,A., Schulte-Sasse,R., Figini,D., Urdaneta,E.C., Beckmann,B.M. and Marsico,A. (2019) TriPepSVM: de novo prediction of RNA-binding proteins based on short amino acid motifs. *Nucleic Acids Res.*, **47**, 4406–4417.
47. Guo,Y.Z., Yu,L.Z., Wen,Z.N. and Li,M.L. (2008) Using support vector machine combined with auto covariance to predict proteinprotein interactions from protein sequences. *Nucleic Acids Res.*, **36**, 3025–3030.
48. Landauer,T.K., Foltz,P.W. and Laham,D. (1998) An introduction to latent semantic analysis. *Discourse Processes*, **25**, 259–284.
49. Blei,D.M., Ng,A.Y. and Jordan,M.I. (2003) Latent dirichlet allocation. *J. Mach. Learn. Res.*, **3**, 993–1022.
50. Ramage,D., Hall,D., Nallapati,R. and Manning,C.D. (2009) Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, pp. 248–256.
51. Bengio,Y., Ducharme,R., Vincent,P. and Jauvin,C. (2003) A neural probabilistic language model. *J. Mach. Learn. Res.*, **3**, 1137–1155.
52. HARRIS,Z. (1954) Distributional Structure. *Word*, **10**, 142–146.
53. Mikolov,T., Chen,K., Corrado,G. and Dean,J. (2013) Efficient estimation of word representations in vector space. arXiv doi: https://arxiv.org/abs//1301.3781, 07 September 2013, preprint: not peer reviewed.
54. Pennington,J., Socher,R. and Manning,C.D. (2014) Glove: Global vectors for word representation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1532–1543.
55. Joulin,A., Grave,E., Bojanowski,P. and Mikolov,T. (2017) Bag of Tricks for Efficient Text Classification. In: *Conference of the European Chapter of the Association for Computational Linguistics*. Vol. **2**, pp. 427–431.
56. Liu,B., Li,C.C. and Yan,K. (2020) DeepSVM-fold: protein fold recognition by combining support vector machines and pairwise sequence similarity scores generated by deep learning networks. *Brief. Bioinform.*, **21**, 1733–1741.
57. Hanson,J., Yang,Y.D., Paliwal,K. and Zhou,Y.Q. (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–692.
58. Lebret,R. and Collobert,R. (2015) "The Sum of Its Parts": joint learning of word and phrase representations with autoencoders. arXiv doi: https://arxiv.org/abs/1506.05703, 18 June 20155,preprint: not peer reviewed.
59. Li,C.C. and Liu,B. (2020) MotifCNN-fold: protein fold recognition based on fold-specific features extracted by motif-based convolutional neural networks. *Brief. Bioinform.*, **21**, 2133–2141.
60. Ye,X.G., Wang,G.L. and Altschul,S.F. (2011) An assessment of substitution scores for protein profile-profile comparison. *Bioinformatics*, **27**, 3356–3363.
61. Rangwala,H. and Karypis,G. (2005) Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, **21**, 4239–4247.
62. Mittelman,D., Sadreyev,R. and Grishin,N. (2003) Probabilistic scoring measures for profile-profile comparison yield more accurate short seed alignments. *Bioinformatics*, **19**, 1531–1539.
63. Strauss,T. and von Maltitz,M.J. (2017) Generalising Ward's method for use with Manhattan distances. *PLoS One*, **12**, e0168288.
64. Weinberger,K.Q. and Saul,L.K. (2009) Distance metric learning for large margin nearest neighbor classification. *J. Mach. Learn. Res.*, **10**, 207–244.
65. Laboulais,C., Ouali,M., Le Bret,M. and Gabarro-Arpa,J. (2002) Hamming distance geometry of a protein conformational space: application to the clustering of a 4-ns molecular dynamics trajectory of the HIV-1 integrase catalytic core. *Proteins-Struct. Funct. Genet.*, **47**, 169–179.
66. Wang,L., You,Z.H., Huang,Y.A., Huang,D.S. and Chan,K.C.C. (2020) An efficient approach based on multi-sources information to predict circRNA-disease associations using deep convolutional neural network. *Bioinformatics*, **36**, 4038–4046.
67. Chang,C.C. and Lin,C.J. (2011) LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.*, **2**, 27.
68. Biau,G. (2012) Analysis of a random forests model. *J. Mach. Learn. Res.*, **13**, 1063–1095.
69. Chen,Z., Zhao,P., Li,C., Li,F., Xiang,D., Chen,Y.Z., Akutsu,T., Daly,R.J., Webb,G.I., Zhao,Q. *et al.* (2021) iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization. *Nucleic. Acids. Res.*, **49**, e60.
70. Sutton,C. and McCallum,A. (2012) An introduction to conditional random fields. *Found. Trends Mach. Learn.*, **4**, 267–373.
71. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
72. Cho,K., van Merriënboer,B., Gulcehre,C., Bahdanau,D., Bougares,F., Schwenk,H. and Bengio,Y. (2014) Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pp. 1724–1734.
73. Vaswani,A., Shazeer,N., Parmar,N., Uszkoreit,J., Jones,L., Gomez,A.N., Kaiser,Ł. and Polosukhin,I. (2017) Attention is all you need. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., Long Beach, California, USA, pp. 6000–6010.
74. Ahmed,K., Keskar,N.S. and Socher,R. (2017) Weighted transformer network for machine translation. arXiv doi: https://arxiv.org/abs/1711.02132, 06 November 2017, preprint: not peer reviewed.
75. Kitaev,N., Kaiser,Ł. and Levskaya,A. (2020) Reformer: the efficient transformer. arXiv doi: https://arxiv.org/abs/2001.04451, 18 February 2020, preprint: not peer reviewed.
76. Liu,Y., Wang,X. and Liu,B. (2018) IDP–CRF: intrinsically disordered protein/region identification based on conditional random fields. *Int. J. Mol. Sci.*, **19**, 2483.
77. Chawla,N.V., Bowyer,K.W., Hall,L.O. and Kegelmeyer,W.P. (2002) SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.*, **16**, 321–357.
78. Farquad,M.A.H. and Bose,I. (2012) Preprocessing unbalanced data using support vector machine. *Decision Support Systems*, **53**, 226–233.
79. Junsomboon,N. and Phienthrakul,T. (2017) Combining Over-Sampling and Under-Sampling Techniques for Imbalance Dataset. In: *Proceedings of the 9th International Conference on Machine Learning and Computing*. Association for Computing Machinery, Singapore, Singapore, pp. 243–247.
80. Schmidt,M., Fung,G. and Rosales,R. (2007) Fast Optimization Methods for L1 Regularization: A Comparative Study and Two New Approaches. In: *Proceedings of the 18th European conference on Machine Learning*. Springer-Verlag, Warsaw, Poland, pp. 286–297.

81. Bilgic,B., Chatnuntawech,I., Fan,A.P., Setsompop,K., Cauley,S.F., Wald,L.L. and Adalsteinsson,E. (2014) Fast image reconstruction with L2-regularization. *J. Magn. Reson. Imaging*, **40**, 181–191.
82. Pedregosa,F., Varoquaux,G., Gramfort,A., Michel,V., Thirion,B., Grisel,O., Blondel,M., Prettenhofer,P., Weiss,R., Dubourg,V. *et al.* (2011) Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.
83. Jain,A.K., Murty,M.N. and Flynn,P.J. (1999) Data clustering: a review. *ACM computing surveys*, **31**, 264–323.
84. Frey,B.J. and Dueck,D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.
85. Ester,M., Kriegel,H.-P., Sander,J. and Xu,X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. AAAI Press, Portland, Oregon, pp. 226–231.
86. Kim,S.C. and Kang,T.J. (2007) Texture classification and segmentation using wavelet packet frame and Gaussian mixture model. *Pattern Recogn.*, **40**, 1207–1221.
87. Skarmeta,A.G., Bensaid,A. and Tazi,N. (2000) Data mining for text categorization with semi-supervised agglomerative hierarchical clustering. *Int. J. Intell. Syst.*, **15**, 633–646.
88. Chandrashekar,G. and Sahin,F. (2014) A survey on feature selection methods. *Comput. Electr. Eng.*, **40**, 16–28.
89. Guyon,I. and Elisseeff,A. (2003) An introduction to variable and feature selection. *J. Mach. Learn. Res.*, **3**, 1157–1182.
90. Darst,B.F., Malecki,K.C. and Engelman,C.D. (2018) Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.*, **19**, 353–363.
91. Sugumaran,V., Muralidharan,V. and Ramachandran,K. (2007) Feature selection using decision tree and classification through proximal support vector machine for fault diagnostics of roller bearing. *Mech. Syst. Signal Process.*, **21**, 930–942.
92. Yeung,K.Y. and Ruzzo,W.L. (2001) Principal component analysis for clustering gene expression data. *Bioinformatics*, **17**, 763–774.
93. Schölkopf,B., Smola,A.J. and Müller,K.-R. (1997) Kernel Principal Component Analysis. In: *Proceedings of the 7th International Conference on Artificial Neural Networks*. Springer-Verlag, pp. 583–588.
94. Wei,J.-J., Chang,C.-J., Chou,N.-K. and Jan,G.-J. (2001) ECG data compression using truncated singular value decomposition. *Trans. Info. Tech. Biomed.*, **5**, 290–299.
95. Liu,B., Long,R. and Chou,K.C. (2016) iDHS-EL: identifying DNase I hypersensitive sites by fusing three different modes of pseudo nucleotide composition into an ensemble learning framework. *Bioinformatics*, **32**, 2411–2418.
96. Liu,B., Fang,L., Liu,F., Wang,X., Chen,J. and Chou,K.-C. (2015) Identification of real microRNA precursors with a pseudo structure status composition approach. *PLoS One*, **10**, e0121501.
97. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of rna secondary structures. *Monatsh. Chem.*, **125**, 167–188.
98. Liu,B., Xu,J.H., Fan,S.X., Xu,R.F., Zhou,J.Y. and Wang,X.L. (2015) PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. *Mol. Inf.*, **34**, 8–17.
99. Kumar,M., Gromiha,M.M. and Raghava,G.P.S. (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. *J. Mol. Recognit.*, **24**, 303–313.
100. Zhang,X. and Liu,S. (2017) RBPPred: predicting RNA-binding proteins from sequence using SVM. *Bioinformatics*, **33**, 854–862.
101. Callaway,E. (2020) 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature*, **588**, 203–204.
102. Altschul,S.F. and Koonin,E.V. (1998) Iterated profile searches with PSI-BLAST - a tool for discovery in protein databases. *Trends Biochem. Sci.*, **23**, 444–447.
103. Ke,Y., Rao,J., Zhao,H., Lu,Y., Xiao,N. and Yang,Y. (2020) Accurate prediction of genome-wide RNA secondary structure profile based on extreme gradient boosting. *Bioinformatics*, **36**, 4576–4582.
104. Chen,W., Zhang,X., Brooker,J., Lin,H., Zhang,L. and Chou,K.C. (2015) PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. *Bioinformatics*, **31**, 119–120.
105. Horne,D.S. (1988) Prediction of protein helix content from an auto-correlation analysis of sequence hydrophobicities. *Biopolymers*, **27**, 451–477.
106. Sokal,R.R. and Thomson,B.A. (2006) Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population. *Am. J. Phys. Anthropol.*, **129**, 121–131.
107. Feng,Z.P. and Zhang,C.T. (2000) Prediction of membrane protein types based on the hydrophobic index of amino acids. *J. Protein Chem.*, **19**, 269–275.
108. Chen,J.H., Liu,Y.M., Liao,Q. and Liu,B. (2019) iEsGene-ZCPseKNC: identify essential genes based on Z curve pseudo k-tuple nucleotide composition. *Ieee Access*, **7**, 165241–165247.
109. Zhou,J., Lu,Q., Xu,R., He,Y. and Wang,H. (2017) EL_PSSM-RT: DNA-binding residue prediction by integrating ensemble learning with PSSM relation transformation. *BMC Bioinformatics*, **18**, 379.