SCIENTIFIC REPORTS

OPEN

iSS-PC: Identifying Splicing Sites via Physical-Chemical Properties Using Deep Sparse Auto-Encoder

Received: 15 May 2017
Accepted: 10 July 2017
Published online: 15 August 2017

Zhao-Chun Xu[1], Peng Wang[1], Wang-Ren Qiu[1,2] & Xuan Xiao[1,3]

Gene splicing is one of the most significant biological processes in eukaryotic gene expression, such as RNA splicing, which can cause a pre-mRNA to produce one or more mature messenger RNAs containing the coded information with multiple biological functions. Thus, identifying splicing sites in DNA/RNA sequences is significant for both the bio-medical research and the discovery of new drugs. However, it is expensive and time consuming based only on experimental technique, so new computational methods are needed. To identify the splice donor sites and splice acceptor sites accurately and quickly, a deep sparse auto-encoder model with two hidden layers, called iSS-PC, was constructed based on minimum error law, in which we incorporated twelve physical-chemical properties of the dinucleotides within DNA into PseDNC to formulate given sequence samples via a battery of cross-covariance and auto-covariance transformations. In this paper, five-fold cross-validation test results based on the same benchmark data-sets indicated that the new predictor remarkably outperformed the existing prediction methods in this field. Furthermore, it is expected that many other related problems can be also studied by this approach. To implement classification accurately and quickly, an easy-to-use web-server for identifying slicing sites has been established for free access at: http://www.jci-bioinfo.cn/iSS-PC.

Generally, the pre-mRNA, including exons and one or more introns, is transcribed from a eukaryotic gene's DNA template. In the pre-mRNA, exon-intron boundaries i.e. the 5′ ends of the introns are called splice donor sites or 5′ splice sites, and intron-exon boundaries i.e. the 3′ ends of the introns are called splice acceptor sites or 3′ splice sites, as shown in Fig. 1. There are two forms of splice sites. Before the pre-mRNA becomes a mature messenger RNA (mRNA), it must go through several biological processes (Fig. 1). The final mRNA containing only remaining exons can be directly involved in the synthesis of protein. Thus, the biological process of removing introns from its 5′ splice site to its 3′ splice site in pre-mRNA and connecting exons to form mRNA plays an important role in gene regulation and expression. In this case, accurate identification of splice sites becomes increasingly important.

Although the technology of PCR has become one of the most important identification methods to accurately identify splice sites with the development of identification technology the functional sites of genes, it is very expensive and time consuming based only on experimental technique. Hence, development of an effective computational method, so as to help researchers effectively and in a timely fashion, identifying splice sites, has become the urgent need to solve a big problem. In this situation, the computational splice-site analysis tools based on the WEB took up, such as NetGene[1, 2], SplicePredictor[3], GeneSplicer[4] and SplicePort[5]. Recently, Wei Chen *et al.*[6] built a prediction model "iSS-PseDNC" which incorporated six DNA local structural properties into pseudo dinucleotide composition to identify splice donor and acceptor sites. In 2016, M Iqbal *et al.*[7] used PseTNC and PseTetraNC methods to propose a hybrid prediction model, called iSS-Hyb-mRMR, for identifying splice sites, and Prabina Kumar Meher[8] used a hybrid feature extraction approach, which contains positional, dependency and compositional features, to develop a predictor called HSplice for predicting the donor splice sites in eukaryotic genes. These were, on balance, successful.

Based on the above information, although the remarkable progress in identification of splice sites has been made, further study about splice-site predictors can be improved and perfected, whether it is with regard to in

[1]Computer Department, Jing-De-Zhen Ceramic Institute, Jing-De-Zhen, 333403, China. [2]Department of Computer Science and Bond Life Science Center, University of Missouri, Columbia, MO, USA. [3]Gordon Life Science Institute, Boston, Massachusetts, 02478, United States of America. Correspondence and requests for materials should be addressed to Z.-C.X. (email: jdzxuzhaochun@163.com) or W.-R.Q. (email: qiuone@163.com) or X.X. (email: jdzxiaoxuan@163.com)

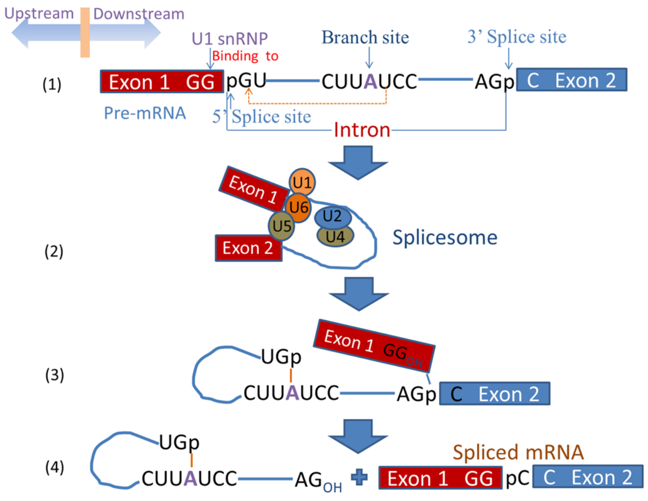


Figure 1. Sketch map showing the steps about the pre-mRNA how to become a mature messenger RNA.

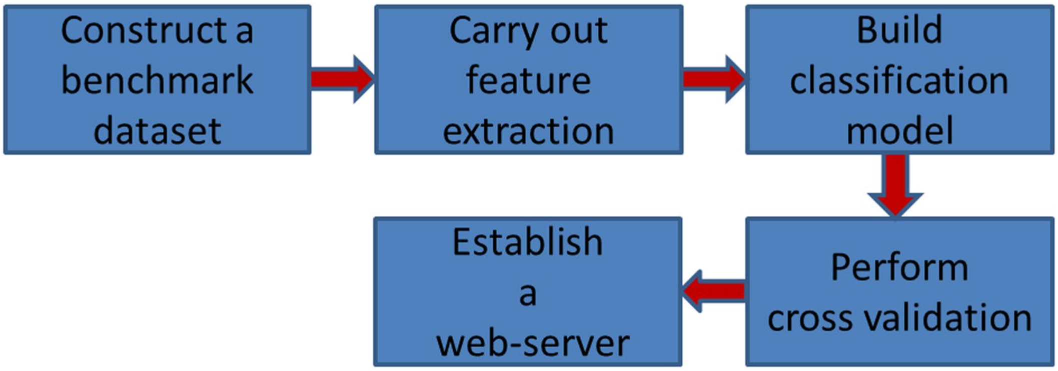

Construct a benchmark dataset
Carry out feature extraction
Build classification model
Perform cross validation
Establish a web-server

Figure 2. Sketch map showing the steps how to establish a predictor for biological system.

feature extraction, or to machine learning classification algorithms. In response to these the issue of two aspects, we have presented a solution to improve the performance of the predictive model in this paper.

On the one hand, improvement of feature extraction method is of critical importance to improve the classification performance. Since S Wold[9] proposed the concept of auto-covariance function(ACF) and cross-covariance function(CCF) to analyze the relations between biopolymer sequences and chemical processes in 1993, this method had been applied to identify nuclear receptors and their subfamilies[10] and N^6-methyladenosine sites[11] via incorporating physical-chemical properties into pseudo amino acid composition(PseAAC) or pseudo dinucleotide composition(PseDNC), respectively. Encouraged by the above successes of introducing this feature extraction approach into computational proteomics, we use twelve physical-chemical properties of the dinucleotides within DNA via a battery of cross-covariance and auto-covariance transformations to obtain a mode of PseDNC to formulate given sequence samples.

On the other hand, the improved machine learning classification algorithms that can provide a better result for classification, is one of the important factors impacting on the performance of classifiers. And in general, different classification algorithms will have different performances. Conventional classification algorithms, such as Support Vector Machine(SVM)[12–15], random forest[16], hidden Markov model[17], Bayes[18], covariance discriminant (CD)[19], Minimax Probability Machine (MPM)[20] and so on, have limitations in processing the original data. Recently, a novel classification algorithm, deep learning, has been proposed based on big data, and it has overcome the former limitations. Deep learning algorithm mainly includes convolutional neural network(CNN)[21], deep belief network(DBN)[22] and stacked auto-encoder(SAE)[23, 24]. Some remarkable progress has been made in diverse fields such as speech recognition and image recognition. In 2014, L James *et al.*[25] firstly used SAE to predict θ and Tangles used to represent local backbone structure of proteins. In the same year, SP Nguyen *et al.*[26] built a model "DL-Pro" that learned a SAE network as a classifier for protein structures. In 2016, J Xu *et al.*[27] used SAE algorithm to detect on breast cancer histopathology images. W Xu *et al.*[28] constructed a model for human promoter recognition with SAE. Inspired by these achievements, the predictor called iSS-PC is constructed by using deep sparse auto-encoder in this paper and its predication performance has been greatly improved.

Basing on a series of recent studies[29–31], we can draw a conclusion that we should follow the five steps[32] shown in Fig. 2 to establish a real and effective biological predictor based on sequence. Below, we are going to discuss

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
$\tau = 2$	88.88	77.77	88.34	89.43
$\tau = 3$	80.58	61.15	81.01	80.14
$\tau = 4$	90.56	81.13	90.09	91.04
$\tau = 5$	90.74	81.49	90.77	90.71

Table 1. The test results of splice donor site sequences based on different characteristic parameter τ values.

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
$\tau = 2$	89.01	78.09	87.40	96.08
$\tau = 3$	90.02	80.04	89.69	90.36
$\tau = 4$	91.11	82.24	90.14	92.11
$\tau = 5$	90.95	81.95	99.44	92.50

Table 2. The test results of splice acceptor site sequences based on different characteristic parameter τ values.

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
iSS-PseDNC[a]	87.71	75.46	89.56	85.86
iSS-PC[b]	90.56	81.13	90.09	91.04

Table 3. The comparison of the 5-fold cross-validation test results on benchmark data-set only containing splice donor site sequences. [a]The prediction method developed by Wei Chen (2014). [b]The prediction method proposed in this paper.

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
iSS-PseDNC[a]	88.73	77.89	94.24	83.07
iSS-PC[b]	91.11	82.24	90.14	92.11

Table 4. The comparison of the 5-fold cross-validation test results on benchmark data-set only containing splice acceptor site sequences. [a]The prediction method developed by Wei Chen (2014). [b]The prediction method proposed in this paper.

how to deal with these steps one by one. Of course, the order of these steps may be appropriately adjusted to be in a format that is suitable for the journal "Scientific Reports".

Results and Discussion

Selection of the characteristic parameter. As described in Section Methods later in the article, we can obtain a feature vector containing $144 \times \tau$ components to represent the given sample sequence D. Here τ is named characteristic parameter, and its value as an integer. Obviously, the dimension I of the feature vector is increased with the increment of the characteristic parameter τ, as shown below.

$$I = \begin{cases} 288 & \tau = 2 \\ 432 & \tau = 3 \\ 576 & \tau = 4 \\ 720 & \tau = 5 \\ \vdots & \vdots \end{cases} \tag{1}$$

However, we should notice that oversized τ value will lead to the problem of the curse of dimensionality. Thus, the value of τ is set at 2, 3, 4 and 5 to carry out experiments, respectively. And the experimental results are listed in Table 1 and Table 2. As can be seen from Table 1, $\tau = 5$ gives the best results, but there is little difference between the results given by $\tau = 4$ and $\tau = 5$. Then, in order to reduce computation time, we fix the τ value into 4. As can be seen form Table 2, $\tau = 4$ gives the best results. Then we can generate a feature vector containing $144 \times 4 = 576$ components as the input of the deep sparse auto-encoder for identifying splicing donor site and splicing acceptor site.

Comparison with the existing methods. The four metrics i.e. accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew correlation coefficient (Mcc) can reflect the performance of predictors clearly. Based on the benchmark dataset composed solely of splice donor site sequences, their scores obtained by the new predictor "iSS-PC" via the five-fold cross-validation test are listed in Table 3. And the results for splice acceptor site sequences, listed in Table 4. For ease of comparison between the other methods, the results obtained by the

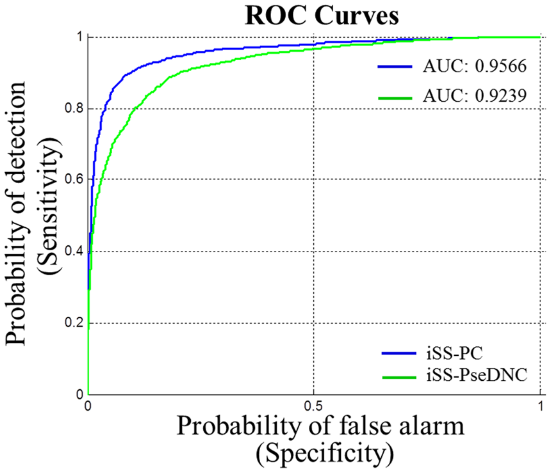


Figure 3. ROC curves of the two different predictors for the splice donor site sequences.

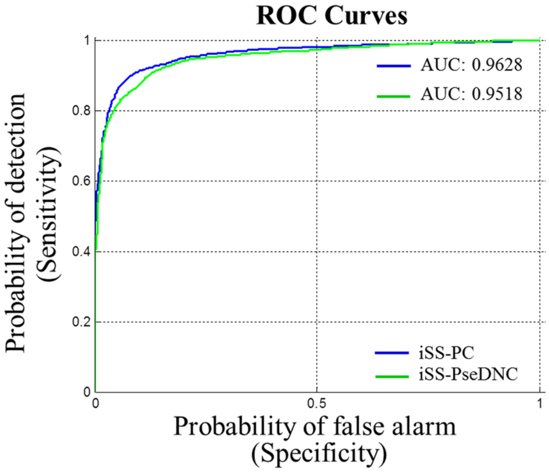


Figure 4. ROC curves of the two different predictors for the splice acceptor site sequences.

iSS-PseDNC predictor constructed by Wei Chen[6] based on the corresponding benchmark dataset are listed in these tables, respectively.

As can be seen from Table 3, although the Sn rate of the new predictor "iSS-PC" is a little bit higher than that of the iSS-PseDNC predictor, the score of the other three metrics has been greatly improved. For example, the ACC rate of our predictor "iSS-PC" has increased by nearly three percent, the MCC rate, nearly six percent and the Sp rate, also nearly six percent. It means that better experimental effect has been acquired, and indicates that our predictor is superior to the iSS-PseDNC predictor at identifying the splice donor site sequences.

On the other hand, as can be seen from Table 4, although the Sn rate of the new iSS-PC predictor is 4% lower than that of the iSS-PseDNC predictor, the Sp rate of our predictor has increased by over 9 percent. And most importantly, the most important indicators for ranking different algorithms have different increases, ACC, nearly 2.5 percent and MCC, nearly 4.5 percent. It indicates that our predictor is also superior to the iSS-PseDNC predictor at identifying the splice acceptor site sequences.

Then through the above analyses, we can draw the conclusion that the methods of feature extraction and classification designed in this paper are very effective based on the splice site sequences. It means that the iSS-PC predictor has higher prediction precision and consumes less time than the existing predictors.

Receiver operating characteristic (ROC) curves. Receiver operating characteristic(ROC) curve[33] is the another important gauge of performance of a predictor. It can visually present readers' eyes in graphical form. The area under the ROC curve(AUC) represents a popular evaluation index of the performance of a binary classifier. Studies[34,35] indicated that the larger the AUC meant better predictor's performance.

In the Figs 3 and 4, the blue curve is generated by new predictor "iSS-PC", and the green curve is formed by the predictor "iSS-PseDNC" constructed by Wei Chen *et al*. The corresponding values of AUC computed over five-fold cross-validation are shown in Figs 3 and 4. From Fig. 3 it can be seen that the values of AUC are 0.9566 and 0.9239 for splice donor site sequences, respectively. On the other hand, for the splice acceptor site sequences

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
iSS-PC[a]	90.56	81.13	90.09	91.04
iSS-SVM[b]	77.59	55.25	75.68	79.50
iSS-RF[c]	83.13	66.38	80.11	86.14
iSS-libD3C[d]	83.38	67.09	78.43	88.32

Table 5. The 5-fold cross-validation test results obtained from different classification algorithms with the same feature extraction method on benchmark data-set only containing splice donor site sequences. [a]The predictor with SAE proposed in this paper. [b]The predictor with SVM created in WEKA with the default parameters. [c]The predictor with Random Forest (RF) created in WEKA. [d]The predictor with an ensemble classifier libD3C.

Predictor	ACC(%)	MCC(%)	Sn(%)	Sp(%)
iSS-PC[a]	91.11	82.24	90.14	92.11
iSS-SVM[b]	73.10	46.23	71.94	74.29
iSS-RF[c]	85.80	71.60	84.70	86.90
iSS-libD3C[d]	83.15	66.55	79.38	87.04

Table 6. The 5-fold cross-validation test results obtained from different classification algorithms with the same feature extraction method on benchmark data-set only containing splice acceptor site sequences. [a]The predictor with SAE proposed in this paper. [b]The predictor with SVM created in WEKA with the default parameters. [c]The predictor with Random Forest (RF) created in WEKA. [d]The predictor with an ensemble classifier libD3C.

the value of AUC generated by predictor "iSS-PC" is found to be 0.9628, whereas the value of AUC generated by predictor "iSS-PseDNC" is found to be 0.9518, as shown in Fig. 4. Obviously, it can be seen that the AUC value of the predictor "iSS-PC" is higher than that of the predictor "iSS-PseDNC" for both the splice donor and acceptor site sequences. Therefore, we can draw the conclusion that our predictor "iSS-PC" is superior to the predictor "iSS-PseDNC", and from the experimental results, it can be proved that the predictor "iSS-PC" is accurate and stable.

Comparison with traditional high-effectiveness machine learning algorithms.

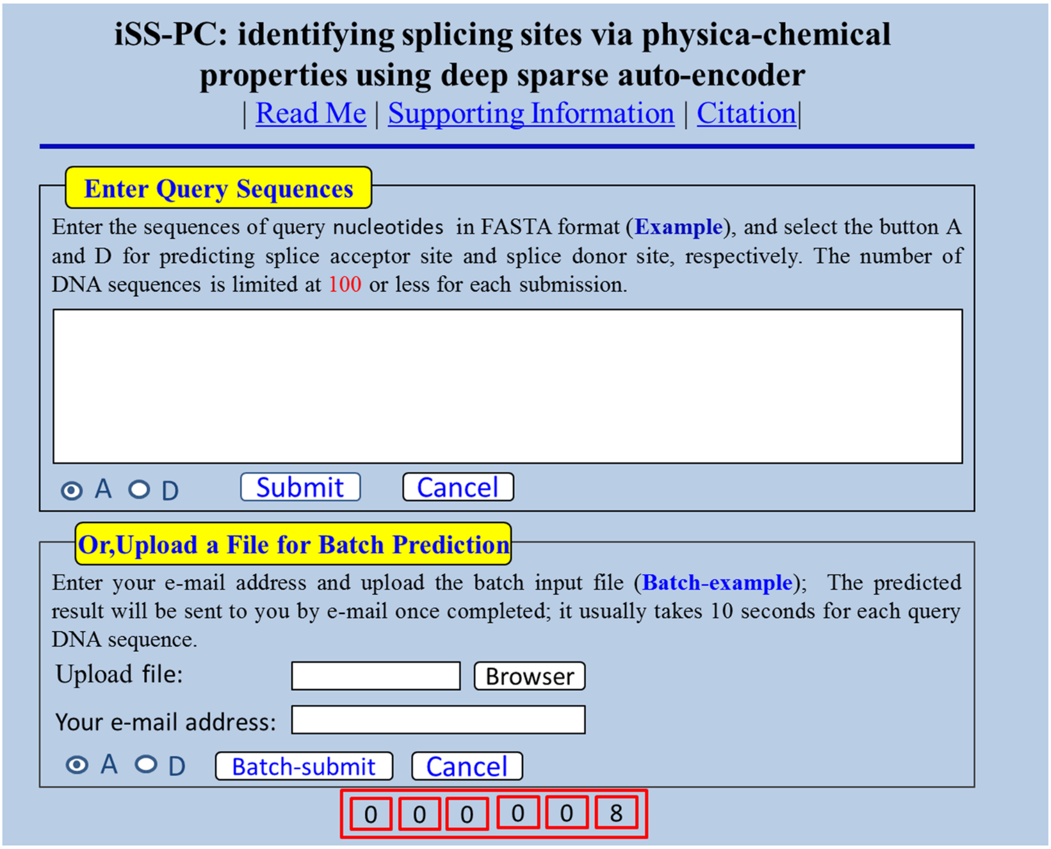
SVM and random forest (RF) are the traditional but efficient classification algorithms. In addition, Dynamic selection and Circulating Combination-based ensemble Clustering i.e. libD3C[36, 37] is a popular tool for binary classification task, too. In order to quickly and easily perform classification prediction for users, libD3C package can be downloaded from the website: http://datamining.xmu.edu.cn/~gjs/LibD3C_1.1/index.html. Meanwhile, WEKA, a free and open source software program, should be downloaded and installed. Then, the ensemble classification model constructed by libD3C can be created in WEKA. In this paper, we compare the SAE model with these traditional machine learning algorithms to examine the performance of the new predictor. And the results are listed in Tables 5 and 6.

The results show in the Tables 5 and 6: the rates of the two most important indicators, ACC and MCC obtained from our predictor "iSS-PC" are significantly higher than those of others, respectively. It indicates the SAE classification algorithm is more effective to identify the splice sites and the new predictor "iSS-PC" would be a very useful tool in this regard.

Web server and its user guide.

In this paper, a simple and practical network predictor shown in Fig. 5, called iSS-PC, has been developed, in order to help the researchers identify splicing sites in real-time and easily. And we provide service consumers with a Web site link http://www.jci-bioinfo.cn/iSS-PC. Below, this article provides details on how to use the network predictor "iSS-PC".

(a) If you want to get the information about the network predictor, please click the Read Me button. Then you can obtain a brief introduction of our predictor and the caveats for using it.
(b) If you want to obtain the benchmark data-set for the iSS-PC predictor training and testing in this paper, please click the Supporting Information button. Here are a few data-sets for download, such as S_1 only containing splice donor site sequences, S_2 only containing splice acceptor site sequences.
(c) If you want to get some important references and resources in establishing the iSS-PC predictor, please click the Citation button.
(d) Before entering query sequences or uploading a file for batch prediction, you should choose types of splice sites: splice donor site or splice acceptor site.
(e) The network predictor "iSS-PC" accepts single or multiple sequence queries. But the input sequences must be in FASTA format, or the network predictor may report errors and will request you to re-input your query sequence. Click the Example button on top of the first input box to see the input format.
(f) If you want to obtain the prediction results, please click the Submit button. After entering query sequences in the first input box in the Example window, you will see how much you've been doing with the job on your screen. When the job is over, the results will be displayed in the page as "The number of DNA sequences investigated: X", and "The DNA #xx is splice donor/acceptor site sequences" or " The DNA #xx is non-splice donor/acceptor site sequences".



iSS-PC: identifying splicing sites via physica-chemical properties using deep sparse auto-encoder

| Read Me | Supporting Information | Citation|

Enter Query Sequences

Enter the sequences of query nucleotides in FASTA format (Example), and select the button A and D for predicting splice acceptor site and splice donor site, respectively. The number of DNA sequences is limited at 100 or less for each submission.

● A ○ D

Or,Upload a File for Batch Prediction

Enter your e-mail address and upload the batch input file (Batch-example); The predicted result will be sent to you by e-mail once completed; it usually takes 10 seconds for each query DNA sequence.

Upload file:

Your e-mail address:

● A ○ D

Figure 5. A semi-screenshot of the homepage for the web-server "iSS-PC".

(g) The lower panel of Fig. 5 offers the option for batch prediction. If you want to submit your batch of multiple sequences in FASTA format for prediction in order to avoid constantly online awaiting, please click the Browse button. The prediction results of each batch job will be sent to your e-mail address. Clicking the Batch-example button, you will see the examples of batch file in FASTA format.

(h) Running times of the network predictor "iSS-PC" are shown underneath the above graph in mathematical terms. And the corresponding number stands for popularity of our predictor to a certain extent.

Conclusions

Feature extraction is the key problem in the research on bioinformatics. In this article, we incorporated twelve physical-chemical properties of the dinucleotides within DNA into PseDNC to formulate the given sequence samples via a battery of cross-covariance and auto-covariance transformations, and achieved good results. However, with the further research of feature extraction methods and the development of computer technology, more and more web servers have been emerged, such as Pse-in-One[38], repRNA[39], and repDNA[40]. Then, many features such as pseudo amino acid composition (PseAAC), pseudo dinucleotide composition (PseDNC), pseudo trinucleotide composition (PseTNC), dinucleotide-based auto covariance (DAC) and dinucleotide-based cross covariance (DCC) can be generated by using these web servers. Therefore, for the future, we can try to study more other similar genomic problems by using the feature extraction methods based on these web servers.

Classification algorithm design is another important step that can affect the performance of a predictor. In this paper, we used deep sparse auto-encoder to construct the iSS-PC predictor. By using the same feature extraction method on benchmark data-sets, we compared the SAE model with those traditional machine learning algorithms, and found that the SAE classification algorithm was stable and reliable. Therefore, the new approach could be used to solve many important tasks in bioinformatics, such as iRSpot-EL[41], iDHS-EL[42], iEnhancer-2L[43]. And these are the work which should be completed in the next phase. In fact, we had constructed a predictor called "iDHSs-PseTNC"[44] to identify DNase I hypersensitive sites with pseudo trinucleotide component by deep sparse auto-encoder, and the results of the predictor iDHSs-PseTNC was superior to that of iDHS-EL.

In conclusion, the timely identification of the splicing sites in DNA sequence is significant for the intensive study on DNA function and the development of new drugs. The experimental results by five-fold cross-validation on the same benchmark datasets indicated that the iSS-PC predictor was superior to other predictors in this area. And the results were promising enough for our predictor to be used as an analytic solution to more genomic problems, such as DNA-binding protein prediction[45], detection of tubule boundary[46], methylation site prediction[47], phosphorylation site prediction[48], and protein-protein interaction prediction[49].

Methods

Benchmark dataset. In this paper, the benchmark dataset is composed of two parts: splice donor site sequences and splice acceptor site sequences. The former can be denoted by S_1, the latter can be formulated by S_2, as shown below.

$$S_1 = S_1^+ \bigcup S_1^-; S_2 = S_2^+ \bigcup S_2^- \tag{2}$$

where S_1^+ represents the positive dataset containing 2796 true splice donor site sequences, while S_1^- represents the negative dataset consisting of 2800 false splice donor site sequences. S_2^+, the positive dataset composed of 2880 true splice acceptor site sequences, while S_2^-, the negative dataset composed of 2800 false splice acceptor site sequences. The symbol $\bigcup$ denotes "union" in the Cantor set theory. Datasets $S_1$and S_2 provided by Wei Chen[6] can be downloaded from the website: http://dx.doi.org/10.1155/2014/623149, or these datasets can be obtained from Supplementary Information.

Feature extraction. Generally, input of nearly all the machine learning based classifiers must be numerical features but not sequences[50], therefore, splice site sequences should be transformed into numerical feature vectors. Below, let's describe how to formulate a sample sequence into a discrete vector model.

A sequence sample in the current benchmark dataset can be generally expressed as

$$\mathrm{D} = N_1N_2N_3N_4N_5N_6N_7\cdots N_L \tag{3}$$

where N_i ($i = 1, 2, \ldots, L$) represents the ith nucleotide of the sequence sample. It can be any one of the four nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T), respectively. While L represents the length of the given sequence sample.

Some literatures have shown that among the discrete vector models for a DNA sample, nucleic acid composition (NAC) is the simplest one. According to the NAC-discrete vector model, the given sequence sample D of Eq. (3) can be defined as

$$\mathrm{D} = [\, f(A) \quad f(C) \quad f(G) \quad f(T)]^T \tag{4}$$

where $f_i = f(\,\cdot\,)$, ($i = 1, 2, 3, 4$) is the normalized occurrence frequency of the corresponding descriptor in the DNA sequence. And T is the transpose operator. But in this way all the sequence order information of sequence D would be entirely lost.

As mentioned in the literature[51], in order to incorporate more short-range sequence-order or local information, the k-tuple nucleotide composition or k-mers approach can be used to formulate the given sequence D into a feature vector containing 4^k components, i.e.

$$\mathrm{D} = [\, f_1 \quad f_2 \quad f_3 \quad \cdots \quad f_{4^{k-1}} \quad f_{4^k}]^T \tag{5}$$

where f_1 is the normalized occurrence frequency of the first k-mer; f_2, that of the second k-mer, and so on. It should be noted however, that k is usually not more than 4, otherwise it may cause over-fitting problem, "high-dimension disaster"[52] and increase of computational run-time with the feature vector dimensions increasing.

To incorporate long-range or global sequence order information, the pseudo components were proposed to deal with not only peptide/protein sequences, but also RNA/DNA sequences. As mentioned in the recent paper[53], the sequence D of Eq. (2) can be formulated as below by using the pseudo nucleotide composition (PseKNC).

$$\mathrm{D} = [\xi_1 \quad \xi_2 \quad \xi_3 \quad \cdots \quad \xi_\mu \quad \cdots \quad \xi_I]^T \tag{6}$$

where subscript I, the vector dimension, is an integer. Its value as well as the components in Eq. (6) will depend on how to extract the desired information from the sequence D.

Below, the "physical-chemical property matrix" and "auto-covariance and covariance transformations" will be used to define the value of subscript I in Eq. (6).

Physical-chemical property matrix. DNA physical-chemical(PC) property is the most intuitive feature of biochemical reactions. And it has different PC properties for each of sixteen different dinucleotides or dimers that are AA, AC, AG, AT, CA, …, TT in a DNA sequence, respectively. In this paper, the following twelve PC properties were adopted: (1) HC^1: A-philicity[54]; (2) HC^2: base stacking[55]; (3) HC^3: B-DNA twist[56]; (4) HC^4: bendability[57]; (5) HC^5: DNA bending stiffness[58]; (6) HC^6: DNA denaturation[59]; (7) HC^7: duplex disrupt energy[60]; (8) HC^8: duplex free energy[61]; (9) HC^9: propeller twist[56];(10) HC^{10}: protein deformation[62]; (11) HC^{11}: protein-DNA twist[62]; (12) HC^{12}: Z-DNA[63]. The original values of the twelve descriptors for each dinucleotide are listed in Table 7. Then we can obtain a 12 × (L − 1) PC property matrix as shown below.

$$\mathrm{D} = \begin{bmatrix} PC^1(N_1N_2) & PC^1(N_2N_3) & \cdots & PC^1(N_{L-2}N_{L-1}) \\ PC^2(N_1N_2) & PC^2(N_2N_3) & \cdots & PC^2(N_{L-2}N_{L-1}) \\ \vdots & \vdots & \vdots & \vdots \\ PC^{12}(N_1N_2) & PC^{12}(N_2N_3) & & PC^{12}(N_{L-2}N_{L-1}) \end{bmatrix} \tag{7}$$

Code	HC^1	HC^2	HC^3	HC^4	HC^5	HC^6	HC^7	HC^8	HC^9	HC^{10}	HC^{11}	HC^{12}
AA	0.97	−5.37	35.5	−0.27	35	66.51	1.9	−1.2	−18.66	12.1	35.1	3.9
AC	0.13	−10.5	33.1	−0.21	60	108.8	1.3	−1.5	−13.1	9.8	31.5	4.6
AG	0.33	−6.78	30.6	−0.08	60	85.12	1.6	−1.5	−14	6.3	31.9	3.4
AT	0.58	−6.57	43.2	−0.28	20	72.29	0.9	−0.9	−15.01	2.1	29.3	5.9
CA	1.04	−6.57	37.7	−0.01	60	64.92	1.9	−1.7	−9.45	6.1	37.3	1.3
CC	0.19	−8.26	35.3	−0.03	130	99.31	3.1	−2.3	−8.11	2.9	32.9	2.4
CG	0.52	−9.69	31.3	−0.03	85	88.84	3.6	−2.8	−10.03	4.5	36.1	0.7
CT	0.33	−6.78	30.6	−0.18	60	85.12	1.6	−1.5	−14	1.6	31.9	3.4
GA	0.98	−9.81	39.6	0.03	60	80.03	1.6	−1.5	−13.48	2.3	36.3	3.4
GC	0.73	−14.6	38.4	0.02	85	135.8	3.1	−2.3	−11.08	4	33.6	4
GG	0.19	−8.26	35.3	−0.06	130	99.31	3.1	−2.3	−8.11	6.1	32.9	2.4
GT	0.13	−10.51	33.1	−0.18	60	108.8	1.3	−1.5	−13.1	2.1	31.5	4.6
TA	0.73	−3.82	31.6	0.18	20	50.11	1.5	−0.9	−11.85	2.3	37.8	2.5
TC	0.98	−9.81	39.6	−0.11	60	80.03	1.6	−1.5	−13.48	4.5	36.3	3.4
TG	1.04	−6.57	37.7	0.13	60	64.92	1.9	−1.7	−9.45	9.8	37.3	1.3
TT	0.97	−5.37	35.5	−0.28	35	66.51	1.9	−1.2	−18.66	2.8	35.1	3.9

Table 7. The original values of the twelve PC properties for each dinucleotide.

where $PC^i(N_jN_{j+1})$ represents the ith ($i = 1, 2, \ldots, 12$) PC property value for the dinucleotide N_jN_{j+1} in Eq. (3). However, the data of Table 7 should be normalized by the following equation before they were substituted into Eq. (7).

$$y_k = (x_k - mean(x))/std(x) \tag{8}$$

where x_k represents the original PC property value in Table 7 of the kth ($k = 1, 2, \ldots, 16$) dinucleotide. While mean (x) represents the average value for the sixteen dinucleotides; and std (x), the corresponding standard deviation; y_k, the corresponding converted values, will remain unchanged if they go through the same conversion procedure again.

Auto-covariance and cross covariance.

The concept of auto-covariance function and cross-covariance function was proposed in 1993, when analyzing the relations between biopolymer sequences and chemical processes. Recently, according to the description to auto-covariance and cross-covariance transformations in literatures[10, 11], these transformations could be expressed by the following mathematical expressions.

$$\mathrm{AC}(\mu, \tau) = \frac{\sum_{j=1}^{L-1-\tau}\left[PC^{\mu}\left(N_jN_{j+1}\right) - \overline{PC^{\mu}}\right]\left[PC^{\mu}\left(N_{j+\tau}N_{j+1+\tau}\right) - \overline{PC^{\mu}}\right]}{L - 1 - \tau}\ (\mu = 1, 2, \cdots, 12) \tag{9}$$

where AC represents the correlation of the same PC property between two sub-sequences separated by τ dinucleotides, $\tau = 1, 2, \ldots, L-2$. While $\overline{PC^{\mu}} = \frac{\sum_{j=1}^{L-1} PC^{\mu}\left(N_jN_{j+1}\right)}{L-1}$ is the mean of the data along the μth row in the matrix of Eq. (7).

$$\mathrm{CC}(n_1, n_2, \tau) = \frac{\sum_{j=1}^{L-1-\tau}\left[PC^{n_1}\left(N_jN_{j+1}\right) - \overline{PC^{n_1}}\right]\left[PC^{n_2}\left(N_{j+\tau}N_{j+1+\tau}\right) - \overline{PC^{n_2}}\right]}{L - 1 - \tau}\ (n_1 \neq n_2) \tag{10}$$

where CC represents the correlation between two subsequences each belonging to a different PC property.

As we can see from Eq. (9), we can generate $12 \times \tau$ components associated with the PC properties of a sample sequence D in Eq. (3) and from Eq. (10), $12 \times 11 \times \tau$ components. Then we can generate $(12 \times \tau + 12 \times 11 \times \tau) = 144 \times \tau$ components by ACF and CCF via 12 different PC properties. Therefore, the sample sequence D can be eventually formulated by

$$\mathrm{D} = [\xi_1\ \ \xi_2\ \ \xi_3\ \ \cdots\ \ \xi_{\mu}\ \ \cdots\ \ \xi_{144\times\tau}]^T \tag{11}$$

where ξ_{μ} represents the μth of the $144 \times \tau$ components generated by Eqs (9) and (10) as described above.

Deep sparse auto-encoder.

In 1986, DE Rumelhart *et al.*[64] firstly proposed the concept of an auto-encoder to process the large complex high-dimensional data. In 2006, GE Hinton *et al.*[22] improved the prototype structure of the auto-encoder, thus making deep auto-encoder (DAE) appear. Thereafter, in 2008, Y Bengio *et al.*[65] proposed the concept of sparse auto-encoder, therefore, the study of DAE went much deeper. And in 2010, P Vincent[24] developed stacked de-noising auto-encoder to yield significantly lower classification error.

Based on the research[22], we constructed a deep sparse auto-encoder model with two hidden layers in this paper, as shown in the Fig. 6. In order to implement classification accurately and quickly based on minimum error law, we can use deep learning software packages, including SAE and NN software, which can be obtained

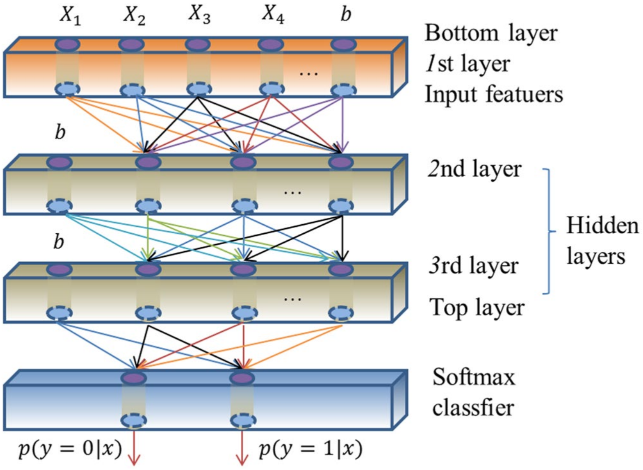


Figure 6. A sketch map of a deep sparse auto-encoder model with two hidden layers.

from the website https://github.com/rasmusbergpalm/DeepLearnToolbox. Note that, in order to optimize the effectiveness of the SAE algorithm, we should fine tune the model parameters by loop optimization. Finally, we can get the best results.

The predictor established according to the above-mentioned procedures is called 'iSS-PC', where 'i' stands for 'identifying', 'SS' for 'splicing sites' and 'PC' for 'physical-chemical property'.

There are two issues to be dealt with: one is 'what metrics should be used to examine the accuracy of the predictor?' The other is 'what validation method should be taken to calculate the metric values?'

A set of metrics for measuring prediction quality.

As mentioned in the literature, accuracy (Acc), sensitivity (Sn), specificity (Sp), and Matthew correlation coefficient (Mcc) introduced by Chou[66] are the most frequently used metrics to evaluate the performance of the predictor in bioinformatics. To make these easier to understand for the researchers, the four metrics can be formulated as below[30, 67].

$$\begin{cases} ACC = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} \\ Mcc = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} \\ Sn = 1 - \dfrac{N_-^+}{N^+} \\ Sp = 1 - \dfrac{N_+^-}{N^-} \end{cases} \tag{12}$$

where N^+ the total number of the true splice donor site sequences (true splice acceptor site sequences) detected, N_-^+ the number of the true splice donor site sequences (true splice acceptor site sequences) misidentified as the false splice donor site sequences(false splice acceptor site sequences); whereas, N^- the total number of the false splice donor site sequences (false splice acceptor site sequences) observed, N_+^- the number of the false splice donor site sequences (false splice acceptor site sequences) mis-predicted as the true splice donor site sequences (true splice acceptor site sequences).

However, it should be noted that the four metrics formulated in Eq. (12) are valid only for the single-label systems, but unsuitable for multi-label systems appearing frequently in system biology and system medicine. For the latter, an utterly different set of metrics is needed as elaborated in the literature[68].

Cross-validation.

After the four well-known metrics mentioned above have been adopted to evaluate the performance of predictors, another thing we should consider at this moment is what validation method should be used to calculate the value of the four metrics. Generally speaking, there are three popular cross-validation approaches in prediction and analysis on the statistics, i.e., independent dataset test, K-fold cross-validation and jackknife test. Although the jackknife test always yielding a unique output for a given benchmark dataset seems the least arbitrary, K-fold cross-validation has more advantages in the computational time than that of the former. Therefore, in this paper, we adopt five-fold cross-validation to score the four metrics. Below, let's introduce specific methods about five-fold cross-validation.

Firstly, for the benchmark dataset S_1 of Eq. (2) consisting of splice donor site sequences, we randomly divided the data-sets S_1^+ and S_1^- into five subsets which size was approximately equal to each other, respectively, as shown below

$$\begin{cases} S_1^+ = S_{11}^+ \bigcup S_{12}^+ \bigcup S_{13}^+ \bigcup S_{14}^+ \bigcup S_{15}^+ \\ S_1^- = S_{11}^- \bigcup S_{12}^- \bigcup S_{13}^- \bigcup S_{14}^- \bigcup S_{15}^- \end{cases} \tag{13}$$

where S_{1i}^+, the subset of S_1^+, its label for the dividing category is set to $i(i = 1, 2, \ldots, 5)$. Similarly, S_{1i}^-, the subset of S_1^-, its label for the dividing category is set to i, too. Both S_{1i}^+ and S_{1i}^- satisfied the following conditions.

$$\begin{cases} |S_{11}^+| \approx |S_{12}^+| \approx |S_{13}^+| \approx |S_{14}^+| \approx |S_{15}^+| \\ |S_{11}^-| \approx |S_{12}^-| \approx |S_{13}^-| \approx |S_{14}^-| \approx |S_{15}^-| \end{cases} \tag{14}$$

where $|S_{11}^+|$ denotes the number of elements (samples) in S_{11}^+, and so forth.

Finally, we can obtain five subsets of the benchmark dataset S_1 according to their labels for the dividing category, as shown below

$$S_1 = S_1' \bigcup S_2' \bigcup S_3' \bigcup S_4' \bigcup S_5' \tag{15}$$

where $S_1' = S_{11}^+ \cup S_{11}^-$, $S_2' = S_{12}^+ \cup S_{12}^-$, and so forth.

with

$$|S_1'| \approx |S_2'| \approx |S_3^{'}| \approx |S_4'| \approx |S_5'| \tag{16}$$

Therefore, we can single out each of the five subsets of Eq. (15) one by one to test the model that were trained with the remaining four subsets for identifying the splice donor site sequences. The cross validation is carried out five times, and the average scores among the output are regarded as the final outcome. It's remarkable that the same cross-validation process can be used for the benchmark data-set S_2 of Eq. (2) consisting of splice acceptor site sequences.

References

1. Brunak, S., Engelbrecht, J. & Knudsen, S. Prediction of human mRNA donor and acceptor sites from the DNA sequence. *Journal of Molecular Biology* **220**, 49–65 (1991).
2. Hebsgaard, S. M., Korning, P. G., Tolstrup, N., Engelbrecht, J. & Rouz, P. Splice site prediction in Arabidopsis thaliana pre-mRNA by combining local and global sequence information. *Nucleic Acids Research* **24**, 3439–3452 (1996).
3. Brendel, V. & Kleffe, J. Prediction of locally optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. *Nucleic Acids Research* **26**, 4748–4757 (1998).
4. Pertea, M., Lin, X. & Salzberg, S. L. GeneSplicer: a new computational method for splice site prediction. *Nucleic Acids Research* **29**, 1185–1190 (2001).
5. Dogan, R. I., Getoor, L., Wilbur, W. J. & Mount, S. M. SplicePort–an interactive splice-site analysis tool. *Nucleic Acids Research* **35**, W285–291 (2007).
6. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iSS-PseDNC: Identifying Splicing Sites Using Pseudo Dinucleotide Composition. *Biomed Research International* **2014**, 623149 (2014).
7. Iqbal, M. & Hayat, M. "iSS-Hyb-mRMR": Identification of splicing sites using hybrid space of pseudo trinucleotide and pseudo tetranucleotide composition. *Computer Methods & Programs in Biomedicine* **128**, 1–11 (2016).
8. Meher, P. K., Sahu, T. K., Rao, A. R. & Wahi, S. D. Identification of donor splice sites using support vector machine: a computational approach based on positional, compositional and dependency features. *Algorithms for Molecular Biology* **11**, 16 (2016).
9. Wold, S., Jonsson, J., Sjörström, M., Sandberg, M. & Rännar, S. DNA and peptide sequences and chemical processes multivariately modelled by principal component analysis and partial least-squares projections to latent structures. *Analytica Chimica Acta* **277**, 239–253 (1993).
10. Xiao, X., Wang, P. & Chou, K. C. iNR-PhysChem: a sequence-based predictor for identifying nuclear receptors and their subfamilies via physical-chemical property matrix. *Plos One.* **7**, e30869 (2012).
11. Liu, Z. *et al.* pRNAm-PC: Predicting N(6)-methyladenosine sites in RNA sequences via physical-chemical properties. *Analytical Biochemistry.* **497**, 60–67 (2015).
12. Cai, Y. D., Ricardo, P. W., Jen, C. H. & Chou, K. C. Application of SVM to predict membrane protein types. *Journal of Theoretical Biology* **226**, 373–376 (2004).
13. Gu, B. & Sheng, V. S. A Robust Regularization Path Algorithm for ν-Support Vector Classification. *IEEE Transactions on Neural Networks & Learning Systems* **99**, 1–8 (2016).
14. Gu, B. *et al.* Incremental learning for ν -Support Vector Regression. *Neural Networks the Official Journal of the International Neural Network Society* **67**, 140–150 (2015).
15. Gu, B., Sheng, V. S. & Li, S. Bi-parameter space partition for cost-sensitive SVM. *AAAI Press* **1**, 3532–3539 (2015).
16. Kandaswamy, K. K. *et al.* AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. *Journal of Theoretical Biology* **270**, 56–62 (2011).
17. Krogh, A., Larsson, B., Von, H. G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305**, 567–580 (2001).
18. Yang, Z., Wong, W. S. W. & Nielsen, R. Bayes empirical bayes inference of amino acid sites under positive selection. *Molecular Biology & Evolution* **22**, 1107–1118 (2005).
19. Chou, K. C. A Key Driving Force in Determination of Protein Structural Classes. *Biochemical & Biophysical Research Communications* **264**, 216–224 (1999).
20. Gu, B., Sun, X. & Sheng, V. S. Structural Minimax Probability Machine. *IEEE Transactions on Neural Networks & Learning Systems* **99**, 1–11 (2016).
21. Lawrence, S., Giles, C. L., Tsoi, A. C. & Back, A. D. Face recognition: a convolutional neural-network approach. *IEEE Transactions on Neural Networks* **8**, 98–113 (1997).
22. Hinton, G. E., Osindero, S. & Teh, Y. W. A fast learning algorithm for deep belief nets. *Neural Computation.* **18**, 1527–1543 (2006).
23. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature.* **381**, 607–609 (1996).
24. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P. A. Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion. *Journal of Machine Learning Research* **11**, 3371–3408 (2010).

25. James, L. *et al.* Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry* **35**, 2040–2046 (2014).
26. Nguyen, S. P., Shang, Y. & Xu, D. DL-PRO: A Novel Deep Learning Method for Protein Model Quality Assessment. *International Joint Conference on Neural Networks.* **2014**, 2071–2078 (2014).
27. Xu, J. *et al.* Stacked Sparse Autoencoder (SSAE) for Nuclei Detection on Breast Cancer Histopathology images. *IEEE Transactions on Medical Imaging* **35**, 119–130 (2016).
28. Xu, W., Zhang, L. & Lu, Y. SD-MSAEs: Promoter Recognition in Human Genome based on Deep Feature Extraction. *Journal of Biomedical Informatics* **61**, 55–62 (2016).
29. Chou, K. C. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins-structure Function & Bioinformatics* **43**, 246–255 (2001).
30. Chen, W., Feng, P. M., Lin, H. & Chou, K. C. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. *Nucleic Acids Research* **41**, e68 (2013).
31. Chen, W., Feng, P. M., Deng, E. Z., Lin, H. & Chou, K. C. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical Biochemistry.* **462**, 76–83 (2014).
32. Chou, K. C. Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of Theoretical Biology* **273**, 236–247 (2011).
33. Lu, Q., Obuchowski, N., Won, S., Zhu, X. & Elston, R. C. Using the optimal robust receiver operating characteristic (ROC) curve for predictive genetic tests. *Biometrics.* **66**, 586–593 (2010).
34. Fawcett, T. ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. *Machine Learning.* **31**, 1–38 (2004).
35. Grau, J., Grosse, I. & Keilwagen, J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics.* **31**, 2595–2616 (2015).
36. Zou, Q. *et al.* An approach for identifying cytokines based on a novel ensemble classifier. *Biomed Research International* **2013**, 1–11 (2013).
37. Lin, C. *et al.* LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing.* **123**, 424–435 (2014).
38. Liu, B. *et al.* Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. *Nucleic Acids Research* **43**, 65–71 (2015).
39. Liu, B., Liu, F., Wang, X. & Chou, K. C. repRNA: a web server for generating various feature vectors of RNA sequences. *Molecular Genetics and Genomics* **291**, 473–481 (2016).
40. Liu, B., Liu, F., Fang, L., Wang, X. & Chou, K. C. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. *Bioinformatics.* **31**, 1307–1309 (2015).
41. Liu, B., Wang, S., Long, R. & Chou, K. C. iRSpot-EL: identify recombination spots with an ensemble learning approach. *Bioinformatics.* **33**, 35–41 (2016).
42. Liu, B., Long, R. & Chou, K. C. iDHS-EL: Identifying DNase I hypersensitive-sites by fusing three different modes of pseu-do nucleotide composition into an ensemble learning framework. *Bioinformatics.* **32**, 2411–2418 (2016).
43. Liu, B., Fang, L., Ren, L., Lan, X. & Chou, K. C. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. *Bioinformatics.* **32**, 362–270 (2016).
44. Xu, Z. C., Jiang, S. Y., Qiu, W. R., Liu, Y. C. & Xiao,X. iDHSs-PseTNC: Identifying DNase I Hypersensitive Sites with Pseuo Trinucleotide Component by Deep Sparse Auto-Encoder. *Letters in Organic Chemistry*. **14**, http://www.eurekaselect.com/150033 (2017).
45. Wei, L., Tang, J. & Zou, Q. Local-DPP: An improved DNA-binding protein prediction method by exploring local evolutionary information. *Information Sciences.* **384**, 135–144 (2016).
46. Su, R. *et al.* Detection of tubule boundaries based on circular shortest path and polar-transformation of arbitrary shapes. *Journal of Microscopy* **264**, 127–142 (2016).
47. Wei, L., Xing, P., Shi, G., Ji, Z. L. & Zou, Q. Fast prediction of protein methylation sites using a sequence-based feature selection technique. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*. **99**, doi:10.1109/TCBB.2017.2670558 (2017).
48. Wei, L., Xing, P., Tang, J. & Zou, Q. PhosPred-RF: a novel sequence-based predictor for phosphorylation sites using sequential information only. *IEEE Trans Nanobioscience*. **99**, doi:10.1109/TNB.2017.2661756 (2017).
49. Wei, L. *et al.* Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier. *Artificial Intelligence in Medicine*. doi:10.1016/j.artmed.2017.03.001 (2017).
50. Chou, K. C. Impacts of bioinformatics to medicinal chemistry. *Medicinal Chemistry.* **11**, 218–234 (2014).
51. Chen, W., Lei, T. Y., Jin, D. C., Lin, H. & Chou, K. C. PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition. *Analytical Biochemistry.* **456**, 53–60 (2014).
52. Wang, T., Yang, J., Shen, H. B. & Chou, K. C. Predicting membrane protein types by the LLDA algorithm. *Protein & Peptide Letters* **15**, 915–921 (2008).
53. Wei, C., Hao, L. & Chou, K. C. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. *Molecular Biosystems.* **11**, 2620–2634 (2015).
54. Ivanov, V. I. *et al.* CRP-DNA complexes: inducing the A-like form in the binding sites with an extended central spacer. *Journal of Molecular Biology* **245**, 228–240 (1995).
55. Ornstein, R. L. & Rein, R. An optimized potential function for the calculation of nucleic acid interaction energies I. Base stacking. *Biopolymers.* **17**, 2341–2360 (1978).
56. Gorin, A. A., Zhurkin, V. B. & Olson, W. K. B-DNA twisting correlates with base-pair morphology. *Journal of Molecular Biology* **247**, 34–48 (1995).
57. Vlahoviček, K., Kaján, L. & Pongor, S. DNA analysis servers: plot.it, bend.it, model.it and IS. *Nucleic Acids Research* **31**, 3686–3687 (2003).
58. Sivolob, A. V. & Khrapunov, S. N. Translational positioning of nucleosomes on DNA: the role of sequence-dependent isotropic DNA bending stiffness. *Journal of Molecular Biology* **247**, 918–931 (1995).
59. Bram, J. Encyclopedia of molecular biology and molecular medicine. *Cell Biochemistry & Function* **95**, 73–74 (1997).
60. Breslauer, K. J., Frank, R., Blöcker, H. & Marky, L. A. Predicting DNA duplex stability from the base sequence. *Proceedings of the National Academy of Sciences* **83**, 3746–3750 (1986).
61. Sugimoto, N., Nakano, S., Yoneyama, M. & Honda, K. Improved Thermodynamic Parameters and Helix Initiation Factor to Predict Stability of DNA Duplexes. *Nucleic Acids Research* **24**, 4501–4505 (1996).
62. Olson, W. K., Gorin, A. A., Lu, X. J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 11163–11168 (1998).
63. Ho, P. S., Ellison, M. J., Quigley, G. J. & Rich, A. A computer aided thermodynamic approach for predicting the formation of Z-DNA in naturally occurring sequences. *Embo Journal.* **5**, 2737–2744 (1986).
64. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature.* **323**, 533–536 (1986).
65. Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. Advances in Neural Information Processing Systems 19. *Chinese Medical Ethics* **23**, 80–83 (2008).
66. Chou, K. C. Using subsite coupling to predict signal peptides. *Protein Engineering* **14**, 75–79 (2001).

67. Xu, Y., Ding, J., Wu, L. Y. & Chou, K. C. iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *Plos One.* **8**, e55844 (2013).
68. Chou, K. C. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. *Molecular Biosystems.* **9**, 1092–1100 (2013).

Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (No. 31560316, 61261027, 61300139), the China Scholarship Council (No. 201508360047), Natural Science Foundation of Jiangxi Province, China (No. 20142BAB207013, 20171ACB20023, 20171BAB202020), the Department of Education of Jiangxi Province (GJJ160866, GJJ160909, GJJ160910), China Postdoctoral Science Foundation Funded Project (Project No.2017M612949). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Xuan Xiao designed the study. Peng Wang collected data. Wang-Ren Qiu conceived and developed the computational model. Zhao-Chun Xu established the websever iSS-PC and wrote the article. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-08523-8

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.