Abstract

A Study on the Effect of Feature Selection against Categorical and Numerical Features in Fixed-length DNA Sequence Classification

Graduate School of
Natural Science & Technology
Kanazawa University

Division of Electrical Engineering
and Computer Science

Student ID No.: 1424042019
Name: Phan Dau
Chief advisor: Professor Kenji Satou
Date of Submission: June 30th, 2017

Abstract

In active genomics, classifying unknown DNA sequences into labeled classes is a considering problem. There have been many researches that proposed effective algorithms to tackle this kind of problems. Several models employed numerical features of DNA sequences, other models used categorical features of DNA sequences. However, until present there have not been studies that made use of the combination of numerical features and categorical features. K-mer frequency is widely used to convert sequences with various lengths into the size of feature vectors. K-mer provides quantitative information but positional information is lost. However, for fixed-length DNA sequence, each subsequence at a specific position can be used as a categorical value. It provides positional information but quantitative information is not available. Therefore, utilizing both positional and quantitative information could help improve the performance of DNA sequence classification.

In this study, we proposed an effective and efficient framework for enhancing the performance of fixed-length DNA sequence classification by using the union of numerical features (i. e. k-mer frequency) and categorical features (i. e. subsequence beginning at a specific position of DNA sequence). By performing evaluation on six benchmark datasets, the results showed that our model obtained comparable or higher classification performance than advanced models. Moreover, during conducting a two-step feature selection approach, we could also discover which group of features played a vital role in improving prediction accuracy in each dataset.

Keywords: Sequence classification, Numerical and categorical features, Feature selection

Chapter 1 : Introduction

1.1. Research context

In recent decades, biological data have been produced at a tremendous rate. Not only the number of DNA sequences contained in GenBank repository but the number of protein sequences in UniProt have been increasing dramatically. Analysis and interpretation of these data are two of the most crucial tasks in bioinformatics, and classification and prediction methods are key techniques to address such tasks.

There were active researches using numerical features like k-mer to tackle of problem DN sequence classification. There were also several studies employing categorical features. The union of numerical and categorical features has not been utilized. Therefore, we hope that the combination of them will contribute to the classification performance in a complementary manner.

1.2. Objectives

The major target of our thesis is that we develop an effective model to address the problem for classifying fixed-length DNA sequences. The specific objectives are as follows.

Applying a proposed model to classify promoter sequences

A promoter is the part of DNA sequence which are sited directly upstream of the start site of transcription. The most important step in the process of transcription is to determine where is a gene or where is the transcription start site. Promoter identification can help locate the position of gene and then analyze the process of gene expression. Therefore, promoter prediction and promoter classification are two considering problems in the field of bioinformatics and classifying promoter sequences is the first objective of our research.

Applying a proposed model to classify splice sequences

In eukaryotes, the first important works for predicting gene is to identify splice junctions. Therefore, developing an effective model for accurately predicting splice junctions is an attractive work, and it is also the second goal of our research.

Applying a proposed model to classify nucleosomal sequences

In eukaryotes, one of the fundamental parts forming chromatin includes nucleosome. Every nucleosome is composed of a segment of roughly 147 base pairs (bp) which is called a nucleosome core particle being covered stiffly around a histone

octamer [1]. Several researches indicated that nucleosome core particle played crucial roles in biological processes like DNA replication and DNA repair [2], [3]. Therefore, predicting nucleosome positioning sequence (or nucleosomal sequences) is fundamentally important in bioinformatics. This problem is also thirdly addressed in our research.

1.3. Contributions

The key contributions of present thesis can be summarized as follows:

- Developing an effective framework for improving fixed-length DNA sequence classification.
- Applying the framework to classify DNA sequences in various biological datasets.
- Discovering which type of features are more effective in each dataset

Chapter 2 : Related Works

2.1. Splice site prediction review

There have been a number studies on the prediction of splice sites. The study of using support vector machine for accurately predicting splice sites was introduced by Sonnenburg *et al.* [4] in 2007. In 2008, Baten *et al.* [5] introduced the research on identification of splice site by exploiting informative features and employing attribute selection. By using short sequence motifs, Meher *et al*. [6] in 2014 released the statistical method for predicting donor splice sites. Two years later, Meher *et al*. [7] also proposed another model based on not only neighboring but also non-neighboring dinucleotide dependencies.

2.2. Promoter prediction review

Czibula *et al.* in 2012 [8] proposed the method for predicting promoter using relational association rules named as "PCRAR". The combination of expectation maximization clustering and support vector machine (EMSVM) for solving the above issue in bacterial DNA sequences was presented by Maleki *el at.* in 2015 [9]. Lin *et al*. [10] proposed another model, named as "iPro54-PseKNC", in 2014.

2.3. Nucleosome positioning prediction review

In 2010, Yi *et al*. [11] introduced a model for solving the problem of predicting nucleosome positioning by using transcription factor binding sites (TFBSs) and the nearest neighbor algorithm. In the research of Guo *et al*. [1] in 2014, they proposed a predictor named as 'iNuc-PseKNC'. In 2016, Tahir and Hayat [12] introduced a

predictor (called iNuc-STNC). Here, nucleosome sequences were encoded into three different groups of features like 2-mer, 3-mer and split 3-mer composition.

2.4. Learning Machine Algorithms

There are many well-know learning algorithms used in bioinformatics. However, in this thesis, we just focus on several algorithms that are commonly used in splice site, promoter and nucleosome positioning prediction.

2.5. Feature Selection

Feature extraction can be defined as the process of projecting the input data with a higher dimensionality into a new space with lower dimensionality.

Feature selection, however, can be defined as the process of detecting relevant features and eliminating irrelevant, redundant features.

2.6. Classification Evaluation

During classification training, to obtain the optimal classifier, evaluation metrics are so important, and a choice of right evaluation metrics for evaluating a classifier plays a crucial role as well [13]. Each evaluation metrics measures each characteristics of classification performance. In this section, we describe some popular metrics used to evaluate a classifier.

Chapter 3 : Materials and Methods

3.1. Datasets

To demonstrate the validity of our method in dealing with DNA sequence classification problem, we evaluated our approach on six datasets (see Table 1).

Table 1. Description of datasets

No	Dataset	Description	Number of Classes	Number of Sample	Sequence length (base)
1	Splice	Primate splice-junction sequences.	3	3175 (762+765+1648)	60
2	Promoter	*E. coli* promoter sequences	2	106 (53 + 53)	57
3	Human	*H. sapiens* nucleosomal and linker sequences	2	4573 (2273 + 2300)	147
4	Worm	*C. elegans* nucleosomal and linker sequences	2	5175 (2567 + 2608)	147
5	Fly	*D. melanogaster* nucleosomal and linker sequences	2	5750 (2900 + 2850)	147
6	Yeast	*S.cerevisiae* nucleosomal and linker sequences	2	3620 (1880 + 1740)	150

3.2. Features

Our model used the combination of the five different vectors named as 1-categorical vector (1CAT), 2-categorical vector (2CAT), 2-mer vector (2MER), 3-mer vector (3MER), and 4-mer vector (4MER).

1CAT = (A_1, A_2, …, A_n), where A_i is a nucleotide at i^{th} position, i = 1, 2, …, n.

2CAT = (B_1, B_2, …, B_{n-1}), where B_i is two consecutive nucleotides at i^{th} and $(i+1)^{th}$ positions, i = 1, 2, …, n-1.

2MER, 3MER and (4MER)

The k-mer vector denoted as kMER is defined by kMER = ($c[s_1]$, $c[s_2]$, …, $c[s_{4^k}]$), where $c[s_i]$ is a number of times that s_i occurs in s, i = 1, 2, …, 4^k.

3.3. Algorithm

Our algorithm includes four steps (shown in Figure 1).

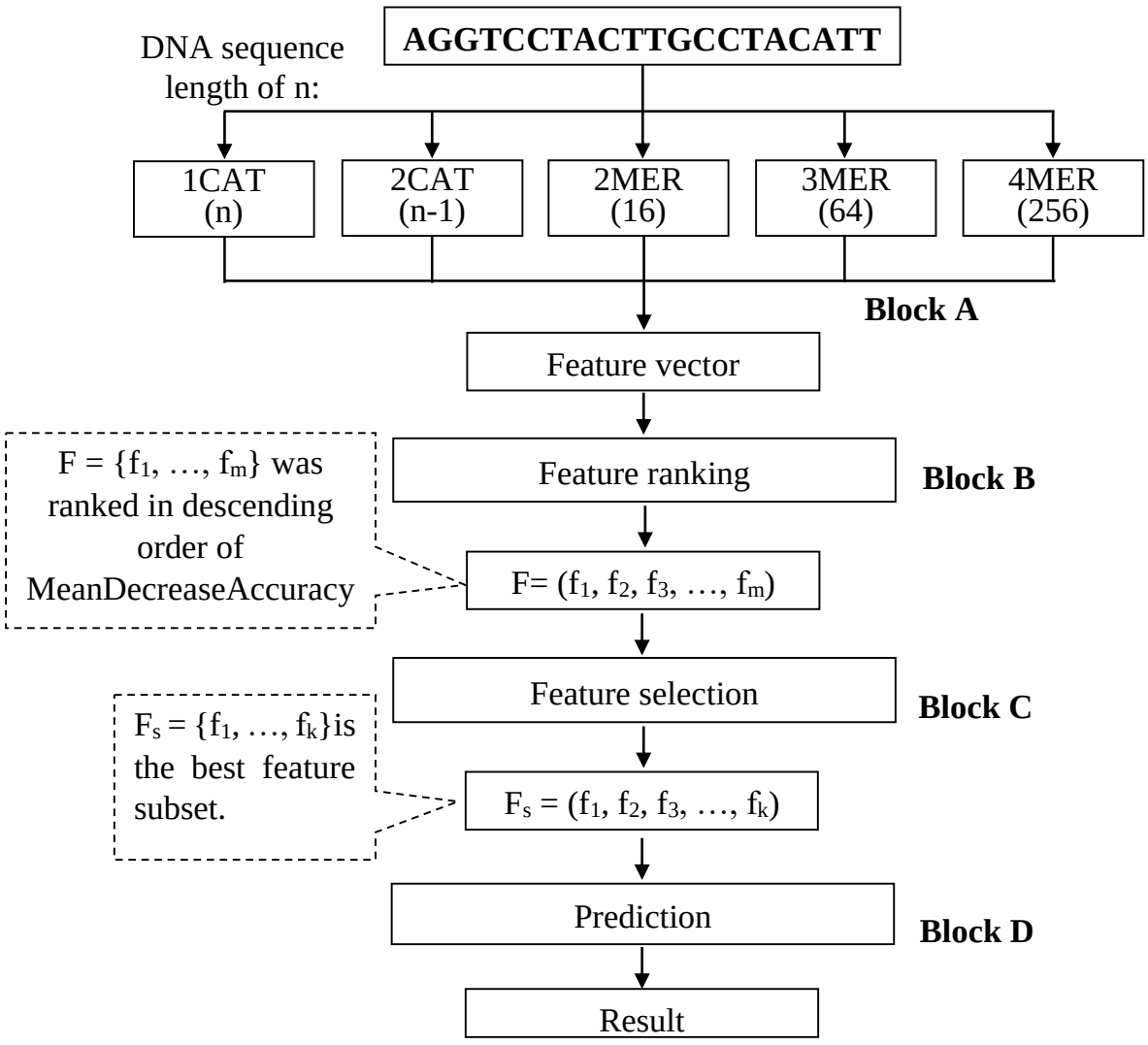

Figure 1. The flowchart of the proposed algorithm.

3.4. Feature Selection

The feature selection approach used in our research is a kind of greedy algorithm. It is a two-step feature selection approach.

Chapter 4 : Experimental Results and Discussion

4.1. Feature Ranking by Random Forest

The relationship between rank and MeanDecreaseAccuracy normalized into the range of [0, 1] in each dataset is shown in the Figure 2.

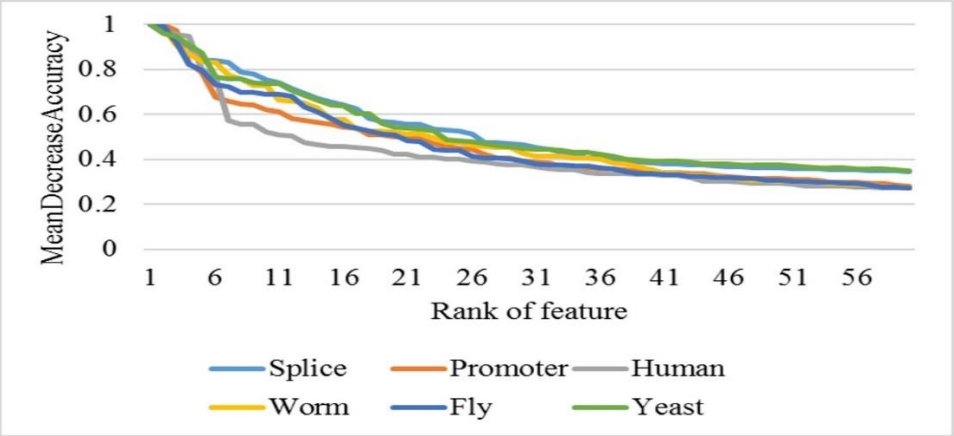

Figure 2. MeanDecreaseAccuracy along feature ranking from top 1~ 60.

Features with high importance in validation datasets are listed in Table 2.

Table 2. List of important features.

No	Dataset	List of top 10 features with high importance sorted by descending order of rank
1	Splice	B_{30}, B_{29}, B_{31}, B_{28}, A_{29}, A_{30}, B_{32}, A_{32}, B_{34}, A_{31}
2	Promoter	B_{17}, B_{16}, B_{15}, B_{14}, A_{15}, B_{39}, A_{17}, B_{18}, A_{16}, B_{38}
3	Human	TTTT, AAA, TTT, AAAA, TT, AA, AAAT, ATTT, TG, TAAA
4	Worm	B_1, AAA, AA, A_1, TTT, AAAA, AAAT, TTTT, ATTT, AATT
5	Fly	TA, GC, CG, TTT, TT, TTTT, ATA, CA, AAAA, TAT
6	yeast	AAAA, TTTT, TA, AAA, TTT, TAT, ATA, CGCG, CA, TT

4.2. Prediction Accuracy of Feature Subsets along Ranking

Prediction accuracies obtained by using either the whole set of features and the best feature subset in step 1 is presented in Table 3.

Table 3. Prediction accuracies when using the whole set of features and the best feature subset in step 1.

No	Dataset	The whole set of features		The best feature subset in step 1		Improvement (%)
		# Feature	Acc (%)	# Feature	Acc (%)	
1	Splice	455	94.55	40	96.77	2.22
2	Promoter	449	94.34	90	100	5.66
3	Human	629	85.94	420	86.35	0.41
4	Worm	629	89.06	180	89.28	0.22
5	Fly	629	80.16	140	81.79	1.63
6	Yeast	635	100	30	100	0.00

4.3. Prediction accuracy of neighbors around the best feature subset

The results are shown in Table 4.

Table 4. Prediction accuracies in step 2 compared with those in step 1.

No	Dataset	The best feature subset in step 1		The best feature subset in step 2		Improvement (%)
		# Feature	Acc(%)	# Feature	Accuracy (%)	
1	Splice	40	96.77	48	96.93	0.16
2	Promoter	90	100	90	100	0
3	Human	420	86.35	428	86.49	0.14
4	Worm	180	89.28	177	89.53	0.25
5	Fly	140	81.79	148	81.93	0.14
6	Yeast	30	100	22	100	0.00

4.4. Evaluation

To evaluate the quality of our method, four following metrics were used: accuracy, sensitivity, specificity and Matthews correlation coefficient.

We applied our model to classify the DNA sequences in the validation datasets and compared its performance with the previous researches. For evaluation, we mainly carried out 10-fold cross-validation, and then computed average prediction results. With Promoter data, however, we employed leave-one-out cross-validation due to the fact that the number of its samples is small, 106 samples.

4.5. Comparison with other methods

For splice and promoter datasets, we compared our model to model in [14] (see Table 5).

Table 5. Accuracies of our model and model in [14].

No	Dataset	Acc (%) in [14]			Acc (%) by our method			Improvement in average (%)
		Min	Max	Average	Min	Max	Average	
1	Splice	95.87	96.73	96.18	96.65	96.93	96.81	0.63
2	Promoter	99.06	99.06	99.06	100	100	100	0.94

For human, worm, fly and yeast datasets, we compared our models to methods in [1], [12], [15], [16], [17] on four metrics (Table 6).

Table 6. Performance comparison of our model and previous models.

Dataset	Method	Acc (%)	Sen (%)	Sp(%)	MCC
Human	Our method	86.33	89.77	82.93	0.73
	iNuc-PseKNC [1]	86.27	87.86	84.70	0.73
	iNuc-PseSTNC [12]	87.60	89.31	85.91	0.75
	3LS [15]	**90.01**	**91.69**	**88.35**	**0.80**
	TNS [15]	81.67	-	-	-
Worm	Our method	**89.35**	**92.45**	86.30	**0.79**
	iNuc-PseKNC [1]	86.90	90.30	83.55	0.74
	iNuc-PseSTNC [12]	88.62	91.62	86.66	0.77
	3LS [15]	87.86	86.54	**89.21**	0.76
	TNS [15]	83.94	-	-	-
Fly	Our method	81.75	79.14	**84.40**	0.64
	iNuc-PseKNC [1]	79.97	78.31	81.65	0.60
	iNuc-PseSTNC [12]	81.67	79.76	83.61	0.63
	3LS [15]	**83.41**	**84.07**	82.74	**0.67**
	TNS [15]	70.82	-	-	-
Yeast	Our method	**100**	**100**	**100**	**1.00**
	TNS [15]	**100**	-	-	-
	Chen et al. [16]	98.10	98.20	98.00	0.96
	Yi et al. [17]	99.06	-	-	-

4.6. Discussion and Conclusion

The combination vector can reflect not only the positional information (categorical features) of DNA sequence, but also the quantitative information (k-mer features) of sequence. It can characterize a genetic sequence. We proposed a simple but powerful model for solving DNA sequence classification problems. The model was tested on six different datasets. In terms of accuracy, sensitivity and MCC, our method achieved better performance than any other competing methods.

Therefore, it can be concluded that our model is effective for DNA sequence classification.

Chapter 5 : Summary and Future Research

5.1. Dissertation summary

The target of our research is to develop an effective framework for classification of fixed-length DNA sequences by using five feature vectors (1CAT, 2CAT, 2MER, 3MER and 4MER). So as to achieve better the performance, the two-step feature selection algorithm was also utilized. The proposed model in present thesis was evaluated on six benchmark datasets.

Four evaluation metrics (accuracy, sensitivity, specificity and Mathews correlation coefficient), 10-fold cross-validation were used to weigh our model. Through the performance evaluation on six benchmark datasets of fixed-length DNA sequences, our algorithm achieved comparable or higher performance than other advanced algorithms. The most thing to note is that our model reaches the accuracy of 100 % on two datasets, promoter and yeast.

5.2. Future Research

Application of the proposed model to protein prediction.

For predicting beta-turns and beta-turn types, the combination of categorical features with the below numerical features will be considered. These features are Position Specific Scoring Matrices (PSSMs), predicted shape strings, and predicted protein blocks. For phosphorylation site prediction, we will combine categorical features with other numerical features used in the research of Ismail *et al.* [18].

Improving the performance of DNA sequence classification.

We will incorporate other numerical features used by previous studies into our model. These features consist of “Pseudo k-tuple nucleotide composition” [1] and “General series correlation pseudo trinucleotide composition”) [19].

References

[1] S. H. Guo, E. Z. Deng, L. Q. Xu, H. Ding, H. Lin, W. Chen, and K. C. Chou, "INuc-PseKNC: A sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, p. 1522–1529, 2014.

[2] N. M. Berbenetz, C. Nislow, and G. W. Brown, "Diversity of eukaryotic DNA replication origins revealed by genome-wide analysis of chromatin structure," *PLoS Genetics*, vol. 6, 2010.

[3] W. Chen, L. Luo, and L. Zhang, "The organization of nucleosomes around splice sites," *Nucleic Acids Research*, vol. 38, p. 2788–2798, 2010.

[4] S. Sonnenburg, G. Schweikert, P. Philips, J. Behr, and G. Rätsch, "Accurate splice site prediction using support vector machines," *BMC Bioinformatics*, vol. 8, no. Suppl 10, p. S7, 2007.

[5] a K. M. a Baten, S. K. Halgamuge, and B. C. H. Chang, "Fast splice site detection using information content and feature reduction," *BMC bioinformatics*, vol. 9, no. Suppl 12, p. S8, 2008.

[6] P. K. Meher, T. K. Sahu, A. R. Rao, and S. D. Wahi, "A statistical approach for 5’ splice site prediction using short sequence motifs and without encoding sequence data," *BMC bioinformatics*, vol. 15, no. 1, p. 362, 2014.

[7] P. K. Meher, T. K. Sahu, A. R. Rao, and S. D. Wahi, "A computational approach for prediction of donor splice sites with improved accuracy," *Journal of Theoretical Biology*, vol. 404, p. 285–294, 2016.

[8] G. Czibula, M. I. Bocicor, and I. G. Czibula, "Promoter sequences prediction Using Relational Association Rule Mining," *Evolutionary Bioinformatics*, vol. 8, pp. 81-196, 2012.

[9] A. Maleki, V. Vaezinia, and A. Fekri, "Promoter Prediction in Bacterial DNA Sequences Using Expectation Maximization and Support Vector Machine Learning Approach," *Data Mining in Genomics & Proteomics*, vol. 6, no. 2, 2015.

[10] H. Lin, E. Z. Deng, H. Ding, W. Chen, and K. C. Chou, "IPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research,* vol. 42, no. 21, p. 12961–12972, 2014.

[11] X. Yi, Y.-D. Cai, Z. He, W. Cui, and X. Kong, "Prediction of nucleosome positioning based on transcription factor binding sites," *PLOS one,* vol. 5, no. 9, p. 1–7, 2010.

[12] M. Tahir and M. Hayat, "iNuc-STNC: a sequence-based predictor for identification of nucleosome positioning in genomes by extending the concept of SAAC and Chou's PseAAC," *Molecular BioSystems*, vol. 12, p. 2587–2593, 2016.

[13] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International Journal of Data Mining & Knowledge Management Process,* vol. 5, no. 2, 2015.

[14] N. G. Nguyen, V. A. Tran, D. L. Ngo, D. Phan, F. R. Lumbanraja, M. R. Faisal, B. Abapihi, M. Kubo, and K. Satou, "DNA Sequence Classification by Convolutional Neural Network," *J. Biomedical Science and Engineering,* vol. 9, no. 9, p. 280–286, 2016.

[15] A. Awazu, "Prediction of nucleosome positioning by the incorporation of frequencies and distributions of three different nucleotide segment lengths into a general pseudo k-tuple nucleotide composition," *Bioinformatics,* vol. 33, no. 1, p. 42–48, 2017.

[16] W. Chen, P. Feng, H. Ding, H. Lin, and K. C. Chou, "Using deformation energy to analyze nucleosome positioning in genomes," *Genomics,* vol. 107, no. 2–3, p. 69–75, 2016.

[17] X. F. Yi, Z. S. He, K. C. Chou, and X. Y. Kong, "Nucleosome positioning based on the sequence word composition," *Protein and Peptide Letters,* vol. 19, pp. 79-90, 2012.

[18] H. D. Ismail, A. Jones, J. H. Kim, R. H. Newman, and D. B. KC, "RF-Phos: A Novel General Phosphorylation Site Prediction Tool Based on Random Forest," *BioMed Research International,* vol. 2016, 2016.

[19] W. Chen, T. Y. Lei, D. C. Jin, H. Lin, and K. C. Chou, "PseKNC: a flexible web server for generating pseudo K-tuple nucleotide composition," *Analytical biochemistry*, vol. 456, pp. 53-60, 2014.