Splice site prediction research based on location information

Bin Wei*, *Mingshu* Zhang, *Yaqiong* Niu and *Yandong* Sun

Engineering University of PAP, Xi'an, China

Keywords: Splice site, Single nucleotide polymorphisms, Hidden markov model.

Abstract. Reveal the mysteries of birth, death and so life has become one of the main purpose of bioinformatics, splice site prediction is one of the most important part, however, not been able to get this problem solved. Firstly, the third generation of genetic markers of single nucleotide polymorphisms had been used in that research to explore the influence of the SNP in splicing; Secondly, a modified hidden Markov model has been introduced; finally, experiments show that the SNP for the performance has a certain influence. In addition, location information based hidden Markov model designed also has positive effects. This method increases the effects dramatically than currently used methods.

1 Introduction

In February 2001, the genome project was completed, which the post-genome era was coming. The focus of research has shifted from sequencing to analysis of genome expression, prediction of protein structure and function, so as to reveal life phenomena such as birth, aging, disease, and death [1]. The identification of splice sites is not only the most important and critical link, but also the basis for our understanding of complex life processes, so the study is of great significance [2].

In eukaryotes, the genome is composed of intron and exon, GT/AG is the intron start and end site. In order to correctly recognition the real splice sites by computational methods, a large number of algorithms (such as variable markov model, support vector machine model, neural network model, bayesian model and so on) and software have been produced in the past years[3]. However, due to the limitations of these models and the understanding of the complexity of the splicing process, those researches has not yet reached a satisfactory results.

In the post-genome era, single nucleotide polymorphism (SNP) is considered to be one of the most important discoveries[4]. SNP is genetically stable and is the most common type of human heritable variation, accounting for 90% of all known variation[5]. One of the biggest features of SNP is that it is genetically stable.

* Corresponding author: weibin82@126.com

Therefore, involving SNP information into splice site identification may provide a new way to explore the splicing mechanism and discover splicing signals. Based on that, this paper conducts experiments by improving the hidden Markov model (HMM). The experimental results showed that the method can improve the classification accuracy remarkably. That is, the SNP information could improve the performance of the algorithm, which confirms that SNP has an influence on the splicing mechanism.

2 Location based HMM prediction method

2.1 Hidden markov model

HMM is an effective training algorithm with statistical basis and is widely used in mathematical modeling and analysis of biological sequences[6]. It is different from the Markov chain in that the events observed from the outside world do not correspond one-to-one with states, but are related through a set of probability distributions. HMM contains a double stochastic process, one is a basic stochastic process that describes state transitions, and the other describes the statistical correspondence between states and observations.

An HMM can be defined as the following:

$$\lambda = (A, B, \Pi)$$

Where, Π is the initial state probability vector, A is the state transition probability matrix, B is the observation probability matrix.

2.2 Location-HMM based splice site prediction model

The splice site prediction problem can be described as follows:

(1) Given state set S={A, T, G, C};

(2) Define the three hidden Markov models $\{\xi_n^{(1)} : 1 \le n \le L\}$, $\{\xi_n^{(2)} : 1 \le n \le L\}$ and $\{\xi_n^{(3)} : 1 \le n \le L\}$ on S corresponding to the donor site, the acceptor site and the adjacent sequences of non-splicing site, respectively;

(3) For an observation sequence $\{O_n : 1 \le n \le L\}$ on S, determine which of the above three hidden Markov models it belongs to.

According to biological knowledge, the transfer relationship between adjacent bases is related to their distance from the splice site. In order to reflect the relationship between the state transition probability and the position of the state in the sequence, the expression of the state transition probability aij of HMM is modified. The location information of the state transition is involved, and the probability of the same state transition relationship can be different when the locations are different.

Then, the state transition probability of each model can be expressed as a_{ijn}, $1 \leqslant i,j \leqslant 4$,$1 \leqslant n < L$, that is, the probability of transitioning from state i to state j at position n.

Finally, a minimum error rate Bayesian decision is used to classify the given observations, and for a given $O = \{O_n : 1 \le n \le L\}$, its most likely model is given by:

$$p(\xi^* \mid O) = \max_i P(\xi^{(i)} \mid O) = \max_i \frac{P(\xi^{(i)}, O)}{P(O)} (1 \le i \le 3) \tag{1}$$

$$P(\xi,O)=P(\xi_1)\prod_{i=2}^{L}P(\xi_i\mid\xi_{i-1})\prod_{i=1}^{L}P(O_i\mid\xi) \tag{2}$$

$$l^{(1)}(o)=\frac{P(\xi^{(1)},O)}{P(O)}/\frac{P(\xi^{(3)},O)}{P(O)}=P(\xi^{(1)},O)/P(\xi^{(3)},O) \tag{3}$$

$$l^{(2)}(o)=P(\xi^{(2)},O)/P(\xi^{(3)},O) \tag{4}$$

$$l^{(3)}(o)=P(\xi^{(1)},O)/P(\xi^{(2)},O) \tag{5}$$

If $l^{(1)}(o)\geq 1$, and $l^{(3)}(o)\geq 1$, then $\xi^{(1)}$ is the most probable mode of O, that is, the center of O is the donor site; similarly, if $l^{(2)}(o)\geq 1$ and $l^{(3)}(o)\leq 1$, then $\xi^{(2)}$ is the most probable mode of O, that is, the center of O is the donor site; otherwise, the center of O is an acceptor site.

3 Integration of SNP information with splice site data

3.1 DNA sequence

The splice site data used in this paper were obtained from Gene Bank Release 146. A total of 70 human DNA sequences were used in this paper, which contained a total of 346 donor sites and 346 acceptor sites.

3.2 Data integration

SNP refers to the polymorphism of the DNA sequence caused by the variation of a single base in the genome with a probability of more than 1%, that is, any base in the DNA sequence is replaced by any one of the other three nucleotides. As shown in Figure 1, the base G in the person 1 sample is replaced by A, which is usually called G-A allele, according to the frequency of the two bases. It is divided into wild type (Wild Type, corresponding to G) and mutant (Mutant, corresponding to A).

```
Person 1    AAGCTAAATTTG
Person 2    AAGCTAAGTTTG
Person 3    AAGCTAAGTTTG
Person 4    AAGCTAAATTTG
Person 5    AAGCTAAGTTTG
```

Fig. 1. Schematic diagram of SNP.

Based on the sorting and analysis of more than 25 million human genome SNPs in the dbSNP database of the National Center for Biotechnology Information, we use them to annotate and replace the DNA sequences involved in this article. The wild type and mutant type are replaced with W and M respectively. Therefore, the state set S*={A, T, G, C, W, M} involved in this paper.

3.3 Sample sequence length selection

This paper use a sliding window to obtain a subsequence of length L from the DNA sequence, as shown in Figure 2. In this experiment, the middle site of the window is

classified according to whether it belongs to a donor site, an acceptor site or a non-splicing site, and the window size can be adjusted.

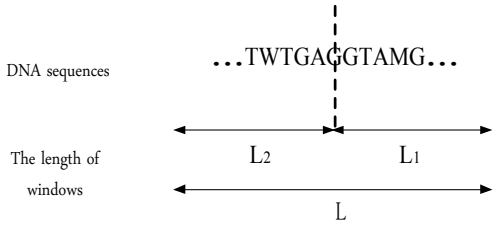


Fig. 2. Schematic diagram of the sequence around the splice site.

4 Experiment and result analysis

4.1 Evaluation indicators

In DNA sequences, the number of pseudo-splicing sites is much higher than that of splicing sites, and some prediction methods often mistake many pseudo-splicing sites as splice sites when they reach a certain accuracy. Therefore, this paper and are used to evaluate the performance of the methods.

$$P_x = \frac{TP}{TP + FN} \tag{6}$$

$$N_x = \frac{TN}{TN + FP} \tag{7}$$

Where: *TN*, *FN*, *TP* and *FP* are defined in Table 1.

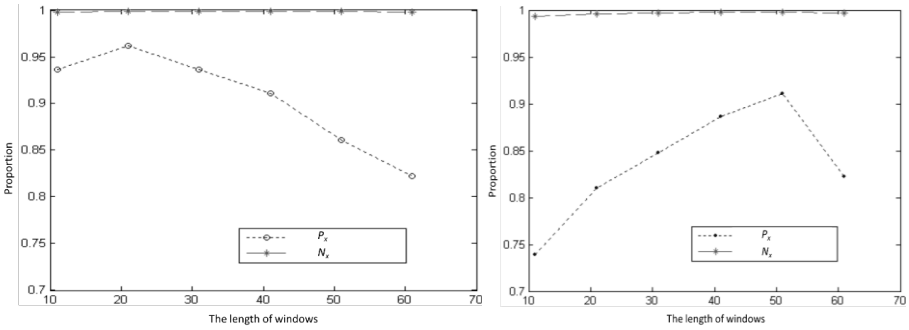
Table 1. Definition of TN, FN, TP and FP.

Test result	Disease stutas	
	+	-
+	True positive(*TP*)	False positive(*FP*)
-	False negative(*FN*)	True negative(*TN*)

4.2 Experimental results

The size of the data window can be adjusted. In order to analyze its impact on performance, six cases of window 11, 21, 31, 41, 51, and 61 are selected for experiments. Due to the fact that the conservation of sequences near the donor site is better than that of the acceptor site, it can be seen from the Figure 3 that the prediction effect of the donor site is better than that of the acceptor site.

Secondly, we tested the performance of the method used in this paper. Table 1 shows the comparison between the results obtained by the standard HMM and the method in this paper. From this table, we can see that using the method in this paper can not only greatly improve the prediction accuracy of splice sites, but also greatly reduce the number of incorrectly predicted non-splicing sites. Therefore, location information is an important factor in splice site identification, which means that our improvement of HMM in this way is successful in this study.



(a) Donor sites (b) Acceptor sites

Fig. 3. The effect of window size on prediction performance.

Table 2. Comparison of prediction effects.

Sites	method	P_X	N_X
donor sites	HMM	0.9367	0.9960
	Our method	0.9620	0.9985
acceptor sites	HMM	0.8734	0.9949
	Our method	0.9114	0.9980

Thirdly, we tested the influence of SNP information on the performance of the algorithm, and the results are shown in Table 2. It can be seen from the table that the results obtained with SNP information are significantly better, that is, the method proposed in this paper greatly reduces the number of false splice sites falsely predicted. Therefore, the SNP information introduced in this paper is a successful attempt to identify the splicing site, that is, the variation in the neighborhood of the splicing site has an important effect on the splicing mechanism.

Table 3. Influence of SNP information on the results.

Sites	Method	P_X	N_X
Donor sites	Don't have SNPs	0.9247	0.9873
	Have SNPs	0.9620	0.9985
Acceptor sites	Don't have SNPs	0.8481	0.9857
	Have SNPs	0.9114	0.9980

Finally, in order to further illustrate the effectiveness of our method, Table 3 compares its prediction effect with several commonly used methods (GENIO[7] and FSPLICE[8]). It can be seen from the table that the method in this paper is significantly better than the other two methods. In particular, the number of misidentified non-spliced sites is greatly reduced.

Table 4. Comparison of the effects of our method and several commonly used splice site prediction algorithms.

Sites	Method	P_X	N_X
Donor sites	GENIO	0.9373	0.9870
	FSPLICE	0.9347	0.9876
	Our method	0.9620	0.9985
Acceptor sites	GENIO	0.8487	0.9866
	FSPLICE	0.8561	0.9960
	Our method	0.9114	0.9980

5 Conclusion

Splicing is one of the most complex mechanisms in biological cells. Therefore, it is important to propose new or improve existing splicing algorithms based on methods such as data mining, machine learning, and optimization to improve algorithm performance. However, it is more important to conduct an in-depth analysis of the splicing process, especially to discover new features or signals that contribute to splicing. In this paper, SNP-related information is introduced in the process of splice site prediction, and location-dependent HMM is used to identify splice sites. The focus of our study is not the algorithm itself, but the evaluation of newly added splice site features based on SNP information. Finally, the hypothesis of this paper is verified by experiments, that is, the introduction of SNP information is helpful for the identification of splice sites. That is, it has an effect on the splicing mechanism. In the near future, we will apply this method to some other larger samples to investigate more complex dataset.

References

1. Villemin, J.-P., A cell-to-patient machine learning transfer approach uncovers novel basal-like breast cancer prognostic markers amongst alternative splice variants. BMC Biology, 2021. 19(70).
2. Esaie Kuitche, S.J.a.A.O., SimSpliceEvol: alternative splicing-aware simulation of biological sequence evolution. BMC Bioinformatics 2019. 20(Suppl 20)(640).
3. Wen, J., A heuristic model for computational prediction of human branch point sequence. BMC Bioinformatics, 2017. 18(459).
4. Yao, Y., CERENKOV2: improved detection of functional noncoding SNPs using data-space geometric features. BMC Bioinformatics, 2019. 20(63).
5. Xie, J., Modeling allele-specific expression at the gene and SNP levels simultaneously by a Bayesian logistic mixed regression model. BMC Bioinformatics, 2019. 20(530).
6. Zubair, A., Bayesian model selection for the Drosophila gap gene network. BMC Bioinformatics, 2019. 20(327).
7. GENIO/splice: Splice Site and Exon Prediction in Human Genomic DNA. http://www.biogenio.com/sp-lice/splice.cgi.
8. FSPLICE: FSPLICE 1.0, Prediction of potential splice sites in Homo_sapiens genomic DNA. .http://sun1.softberry.com/berry.phtml?topic=fsplice&group=programs&%20subgroup=gfind.