Computational Biology and Chemistry xxx (2008) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Brief communication

Genomic splice site prediction algorithm based on nucleotide sequence pattern for RNA viruses

Kun-Nan Tsai [a], Shu-Hung Lin [a], Shin-Ru Shih [b], Jhih-Siang Lai [a], Chung-Ming Chen [a,*]

[a] Institute of Biomedical Engineering, National Taiwan University, 1, Section 1, Jen-Ai Road, Taipei 100, Taiwan, ROC
[b] Department of Medical Biotechnology and Laboratory Science, Chang Gung University, 259 Wen-Hua 1st Road, Kweishan, Taoyuan 333 Taiwan, ROC

ARTICLE INFO

Article history:
Received 27 February 2008
Received in revised form 20 July 2008
Accepted 12 August 2008

Keywords:
Splice site prediction
RNA virus
Eigen-pattern
Cross-species strategy
Orthomyxovirus

ABSTRACT

Splice site prediction on an RNA virus has two potential difficulties seriously degrading the performance of most conventional splice site predictors. One is a limited number of strains available for a virus species and the other is the diversified sequence patterns around the splice sites caused by the high mutation frequency. To overcome these two difficulties, a new algorithm called Genomic Splice Site Prediction (GSSP) algorithm, was proposed for splice site prediction of RNA viruses. The key idea of the GSSP algorithm was to characterize the interdependency among the nucleotides and base positions based on the eigen-patterns. Identified by a sequence pattern mining technique, each eigen-pattern specified a unique composition of the base positions and the nucleotides occurring at the positions. To remedy the problem of insufficient training data due to the limited number of strains for an RNA virus, a cross-species strategy was employed in this study. The GSSP algorithm was shown to be effective and superior to two conventional methods in predicting the splice sites of five RNA species in the Orthomyxoviruses family. The sensitivity and specificity achieved by the GSSP algorithm was higher than 99 and 94%, respectively, for the donor sites, and was higher than 96 and 92%, respectively, for the acceptor sites. Supplementary data associated with this work are freely available for academic use at http://homepage.ntu.edu.tw/~d91548013/.

1. Introduction

RNA splicing is a post-transcriptional process that often takes place in the pre-mRNA of eukaryotes and some viruses, such as the influenza virus, simian virus 40, Adenovirus, and so on (Mount, 1982), which removes the introns and ligates the exons to form the protein coding region of a gene. Splice site prediction aims to identify the potential splice sites, i.e., the junctions of exons and introns, as the basis for further construction of the protein coding regions computationally or experimentally. Most conventional splice site prediction algorithms may be classified into two categories. One is the probabilistic approach (Burge and Karlin, 1997; Brendel and Kleffe, 1998; Pertea et al., 2001; Chen et al., 2005). The other is the neural network (NN) (Reese et al., 1997) and support vector machine (SVM) approaches (Degroeve et al., 2002, 2005; Baten et al., 2006). The probabilistic approaches computed the likelihood of the di-nucleotides (GT/AG) being a splice site by modeling the compositional characteristics of the surrounding regions with the position-specific or the region-wise nucleotide distributions. The NN and SVM approaches classified the di-nucleotides into splice or non-splice sites by encoding the nucleotide sequence information of the surrounding regions into high-dimensional features.

Although many conventional approaches have been shown to be effective in predicting the splice site of several eukaryotes, splice site prediction for an RNA virus remains a challenging task. The difficulty mainly arises from the high mutation frequency of an RNA virus. Most conventional approaches tended to capture the nucleotide sequence patterns with high nucleotide occurrence frequencies. However, the RNA mutation frequencies, though varying with individual genomic positions, are generally in the range of 10^{-5} and 10^{-3} s/nt (substitution per nucleotide) (Domingo, 1997). Most individual genomes would thus differ in one or more nucleotides from the consensus sequence of an RNA virus population. It suggests that not only might a splice site in an RNA viral sequence have one or multiple low-frequency nucleotides in the surrounding region, but also a new splice site might emerge due to mutation. If the low-frequency nucleotides of an RNA viral sequence fortuitously occur at the critical base positions, which usually contain high-frequency nucleotides in the training sequences, the splice sites are very likely missed by the conventional approaches.

Another practical difficulty in constructing a splice site predictor for an RNA virus is due to the limited number of RNA virus strains available for learning the complex dependency among the nucleotides and base positions. Conventional approaches usually

* Corresponding author. Tel.: +886 2 33665273; fax: +886 2 33665268.
E-mail address: ming@lotus.mc.ntu.edu.tw (C.-M. Chen).

1476-9271/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compbiolchem.2008.08.002

required a sufficient number of samples from the same population to extract the nucleotide sequence patterns that are statistically meaningful. Because the sequenced RNA virus strains are often quite limited, to predict the splice sites of an emerging or a re-emerging RNA virus, it is generally impractical to use the strains of the same RNA virus as the training data. As a consequence, the conventional approaches may not be directly applied to the RNA viruses, or at least may not be as effective because of these two difficulties.

To overcome the potential problems caused by the high mutation frequency, a new splice prediction algorithm, called Genomic Splice Site Prediction (GSSP) Algorithm, was proposed for RNA viruses in this paper. The GSSP algorithm aimed to discover the eigen-patterns, each of which was a nucleotide sequence pattern that specified a unique composition of the base positions and the nucleotides occurring at the positions. It was assumed that the nucleotides in an eigen-pattern exert the required binding forces for the splicing process. To identify all eigen-patterns, the basic idea of the proposed GSSP algorithm was to explore the inter-dependency among the nucleotides and base positions by using the sequence pattern mining techniques. Not only could the GSSP algorithm capture the nucleotide sequence patterns with high nucleotide occurrence frequencies, but also it could extract those low-frequency patterns with a minimum support.

2. Methods and Materials

2.1. RNA Viruses and Cross-species Training Data

The RNA viruses tested in this study were the five species in the Orthomyxoviridae family, i.e., Influenza A virus, Influenza B virus, Influenza C virus, Infectious salmon anemia virus, and Thogoto virus. Acquired from NCBI database, the virus sequences tested were composed of segments 7 and 8 of influenza A virus, segment 8 of the influenza B virus, segment 7 of the influenza C virus, segment 7 of the infectious salmon anemia virus, and segment 6 of the Thogoto virus. For conciseness, these six segments were denoted as IFA7, IFA8, IFB8, IFC7, ISAV7 and Tho6, respectively. The number of splice sites for each species was summarized in Table 1. Because the numbers of strains available for the species in the Orthomyxoviridae family were quite limited, we proposed a cross-species strategy and suggested using the nucleotide sequences around the splice sites of Drosophila as the training data (http://www.fruitfly.org/).

2.2. Proposed Algorithm

The key idea of the GSSP algorithm was to find out all eigen-patterns from the training data based on sequence pattern mining. An eigen-pattern specified a unique set of base positions in the vicinity of a splice site and the nucleotide occurring at each position. Each eigen-pattern was assumed to exert sufficient binding forces for the splicing process. Pyrimidine-rich phenomenon is often observed in the near-upstream region of an acceptor in vertebrates, invertebrates, plants, and viruses (Senapathy et al., 1990). Moreover, most base positions in this region have no exclusive preference for either nucleotide C or T. We hypothesized what governs the formation of binding force for splicing process of an acceptor is the class of purine or pyrimidine, not the type of nucleotides at each base position of an eigen-pattern. Therefore, sequence binarization was performed before eigen-patterns were mined by the GSSP algorithm for the prediction of acceptors.

Table 1
The numbers of real and false splice sites in the data sets for Drosophila, Influenza virus, Infectious salmon anemia virus (ISAV) and Thogoto virus

Species	Donor site		Acceptor site	
	Real	False	Real	False
Drosophila	757	2,993	757	2,466
IFA7	3286	152,493	3286	252,847
IFA8	3501	102,938	3501	249,091
IFB8	176	8,239	176	15,073
IFC7	100	2,982	100	5,718
ISAV7	5	342	5	338
Tho6	4	191	4	296

2.2.1. Consensus Sequence

The consensus sequence in this study refers to a window of nucleotide sequence with the most frequently found nucleotide at each base position around a splice site. As the basis of the GSSP algorithm, the consensus sequence was constructed from the training data, i.e., the nucleotide sequences of Drosophila. The windows of the consensus sequence for the donor site and acceptor site were denoted by (D1, D2) and (A1, A2), respectively. D1 and A1 were negative numbers, the absolute values of which were equal to the numbers of nucleotides before GT and AG, respectively. D2 and A2 were positive numbers standing for the numbers of nucleotides from GT and AG, respectively, to the right end of the windows, including GT and AG.

2.2.2. Sequence Pattern Mining

Sequence pattern mining aimed to discover the co-occurring nucleotide patterns with reasonable supports embedded in a set of sequence data. A sequence s of length l was expressed as $s = \langle s_1, s_2, \ldots, s_l \rangle$. For donor sites, $s_j \in \{A,T,C,G\}$ and for acceptor sites, $s_j \in \{Y,R\}$. A sequence pattern P of length l was expressed as $P = \langle p_1, p_2, \ldots, p_l \rangle$. For donor sites, $p_j \in \{A, T, C, G, \bar{A}, \bar{T}, \bar{C}, \bar{G}, d\}$ and for acceptor sites, $p_j \in \{Y, R, \bar{Y}, \bar{R}, d\}$, where $\bar{X}$ represented "not-X" and d "don't care", i.e., any nucleotide. If every s_j in s could be represented by the corresponding p_j of P or equal to p_j, we might say either sequence s matched with sequence pattern P, or sequence pattern P represented sequence s.

An eigen-pattern was a minimally supported sequence pattern (MSSP), which was defined based on the training sequence database, Ω, containing the real splice sites. For each sequence pattern P, support$(P|\Omega)$ was defined as the number of sequences that matched with sequence pattern P in Ω. If support$(P|\Omega) \geq$ min_sup, we called sequence pattern P a MSSP, where min_sup was the given smallest number of sequences.

The first step of the GSSP algorithm was to generate the candidate sequence patterns. Given a set of training sequence data and a pre-specified min_sup, the mining process started from the construction of a binary tree, each node of which defined a sequence pattern, using the consensus sequence. Suppose the consensus sequence was denoted by $s = \langle s_1, s_2, \ldots, s_l \rangle$. The root node of the binary tree, which was at level-0 of the tree, defined the initial sequence pattern, $\langle d, d, \ldots, d \rangle$, where "d" denoted "don't care". Recursively, we supposed that the sequence pattern of a level-i node was $\langle s_1, \ldots, \bar{s}_i, d, \ldots, d \rangle$ and it was a MSSP. Then, two child nodes, which were at level-$(i+1)$, would be generated from this level-i node with the sequence patterns of $\langle s_1, \ldots, \bar{s}_i, s_{i+1}, d, \ldots, d \rangle$ and $\langle s_1, \ldots, \bar{s}_i, \bar{s}_{i+1}, d, \ldots, d \rangle$, respectively. Any of these two level-$(i+1)$ nodes could further spawn two child nodes only if its sequence pattern was a MSSP. If the sequence pattern of a level-i node was not a MSSP, this level-i node would be removed from the binary tree. This recursive process would continue until all level-$(l-1)$ nodes with MSSP's had generated their two child nodes at level-l according to the last nucleotide of the consensus sequence and these level-2^l nodes had been determined if they should remain in the tree

Please cite this article in press as: Tsai, K.-N., et al., Genomic splice site prediction algorithm based on nucleotide sequence pattern for RNA viruses. Comput. Biol. Chem. (2008), doi:10.1016/j.compbiolchem.2008.08.002

depending on if they were MSSP's. When the recursive process terminated, all remaining sequence patterns of the leaf nodes were considered as the candidate sequence patterns.

The second step was to generate the unraveled sequence patterns. In this step, the candidate sequence pattern in every leaf node was unraveled according to all $\bar{X}$'s in the pattern. Suppose a candidate sequence pattern was $\langle s_1, \bar{C}, \ldots, \bar{T}, \ldots, s_l\rangle$. According to $\bar{C}$, three sequence sub-patterns would be generated, i.e., $\langle s_1, A, \ldots, \bar{T}, \ldots, s_l\rangle$, $\langle s_1, G, \ldots, \bar{T}, \ldots, s_l\rangle$ and $\langle s_1, T, \ldots, \bar{T}, \ldots, s_l\rangle$. Then, the ratio of the support of each sup-pattern and the support of $\langle s_1, \bar{C}, \ldots, \bar{T}, \ldots, s_l\rangle$ was computed. If the largest ratio was greater than a pre-specified ratio, min_ratio, the sequence sub-pattern with the largest ratio was retained, where $0 \leq \text{min\_ratio} \leq 1$. Similarly, three sequence sub-patterns could be generated according to $\bar{T}$. The sequence sub-pattern with the largest ratio, which was greater than min_ratio, was retained. Suppose the retained sequence sub-patterns were $\langle s_1, A, \ldots, \bar{T}, \ldots, s_l\rangle$ and $\langle s_1, \bar{C}, \ldots, C, \ldots, s_l\rangle$. The unraveled sequence pattern was defined to be $\langle s_1, A, \ldots, C, \ldots, s_l\rangle$.

The third step of the GSSP algorithm was to determine the eigen-patterns from the candidate sequence patterns and the unraveled sequence patterns. For each pair of candidate and unraveled sequence patterns, denoted by P_c and P_u, respectively, in theory, the unraveled sequence pattern had a higher specificity, whereas the candidate sequence pattern tended to have a better sensitivity. Let Ω and Φ denote the training sequence sets of the real splice sites and the false splice sites, respectively. To determine which of the P_c and P_u was better, we computed $r_c = \text{support}(P_c|\Phi)/\text{support}(P_c|\Omega)$ and $r_u = \text{support}(P_u|\Phi)/\text{support}(P_u|\Omega)$. If $r_c < r_u$ and $r_c < T_{fr}$, P_c was defined as an eigen-pattern, where T_{fr} was a pre-specified threshold. If $r_u < r_c$ and $T_{fr} < r_c$, P_u was defined as an eigen-pattern.

With the eigen-patterns, we could easily determine if a test splice site was a real splice site or a false one by checking if the sequence around the test splice site matched with any eigen-pattern. However, since the eigen-patterns were mined from *Drosophila* rather than from the RNA viruses, to account for the genetic difference among species, a tolerance was allowed in determining if a test sequence matched with an eigen-pattern. We allowed up to N_m mismatching base positions when we concluded that a test sequence matched with an eigen-pattern. If a test sequence matched with more than one eigen-pattern, the eigen-pattern with the smallest number of mismatching base positions was selected. Once the donors and acceptors had been predicted separately, we further performed a minimal donor–acceptor pair-check, i.e., we checked if each acceptor site had at least one donor site in its upstream. If not, the predicted acceptor site was removed. This was based on the biological phenomenon that the donor and acceptor sites should appear in pairs.

Since there was no way to know the optimal N_m value, the strategy that we had taken to determine N_m was to increase the sensitivity as much as possible while controlling the specificity in an acceptable range. More specifically, the value of N_m was decided as follows. Before the minimal donor–acceptor pair-check, starting from 0, we increased N_m gradually. For each N_m value, we computed the positive ratio, which was defined as the ratio of the positive number and the total number, where the positive number was the number of predicted splice sites and the total number was the total number of test splice sites. It should be noted that the positive number and the total number for the donor and the acceptor sites were calculated separately. Supposed that the positive ratio was greater than 15% when $N_m = N_e$. We would carry out the minimal donor–acceptor pair-check for the splice sites predicted by letting $N_m = N_e - 1, N_e - 2, \ldots, 1$, in the descending order. The iterative process terminated when the positive ratio first became less than 10% after the minimal donor–acceptor pair-check, which defined the N_m employed by the GSSP algorithm. Therefore, the specificity could be controlled in the range of 90–100%. Of course, the threshold, 10%, could be changed as desired.

2.3. Performance Analysis

The performance of the proposed GSSP algorithm was evaluated using the five species of RNA viruses in the Orthomyxoviridae family. To compare with the conventional approaches, we had chosen SplicePredictor (Brendel and Kleffe, 1998) and NNsplice (Reese et al., 1997) for comparison. The performance figures computed for each algorithm were sensitivity and specificity, which were defined as TP/(TP + FN) and TN/(TN + FP), respectively, where TP, TN, FP and FN stood for true positive, true negative, false positive and false negative. True positive means that a real splice site was predicted as a splice site.

3. Results and Discussion

The difficulties of splice site prediction on RNA viruses lay in the limited number of strains available for a species and the diversified sequence patterns around the splice sites caused by the high mutation frequency. Both difficulties made most conventional approaches ineffective in predicting the slice sites of RNA viruses. The former led to the problem of insufficient training data, whereas the latter degraded those conventional approaches counting on the high nucleotide occurrence frequency around the splice sites. To cope with these two difficulties, the proposed GSSP algorithm was parameterized to account for the cross-species variation in eigen-patterns. For example, the larger the N_m was, the larger the variation was allowed. Moreover, the idea of eigen-patterns provided a new approach to identifying the important sequence patterns around splice sites but probably with nucleotides of low-occurrence frequencies.

The parameters of the GSSP algorithm included a consensus sequence window, min_sup, T_{fr}, min_ratio, and N_m. While the optimal parameter values were different for different species, we had chosen to use the same parameter values, except N_m, for all five species in this study. The performance reported for each species may be further improved by using other parameter values. The value of N_m was dependent on the ratio of the positive number to the total number. More precisely, the parameters of donor sites were set to $(D_t1, D_t2) = (3,4)$, min_sup = 1, $T_{fr} = 2$, min_ratio = 0.8, $N_m = 0$. The parameters of acceptor sites were set to $(A_t1, A_t2) = (17,1)$, min_sup = 1, $T_{fr} = 2$, min_ratio = 0.8. The value of N_m varied with species. For IFA7, IFA8, IFB8, IFC7, ISAV7 and Tho6, N_m was set to 3, 3, 4, 3, 3 and 3, respectively.

The performances achieved by the GSSP, NNsplice and SplicePredictor algorithms were summarized in Tables 2 and 3, for donor and acceptor predictions, respectively, for all five species. In these two tables, "Sen" and "Spe" stood for sensitivity and specificity, respectively. All three algorithms had achieved similar specificities for the five species of RNA viruses in the Orthomyxoviridae family. However, the proposed GSSP algorithm had better sensitivities than the NNSplice and SplicePredictor. More specifically, for the prediction of donor sites, the proposed GSSP algorithm was able to attain sensitivity higher than 99% and specificity higher than 94% for the tested species. The GSSP algorithm was clearly superior to the NNSplice in the prediction of the donor sites of IFA8 and ISAV7. In particular, the NNSplice was not able to predict any of the donor sites of ISAV7. In comparison with the SplicePredictor, the GSSP algorithm far outperformed its counterpart in predicting the donor sites of IFA7, IFA8, IFB8, and IFC7. The SplicePredictor could barely identify donor sites of the influenza A virus. For the prediction of acceptor sites, the GSSP algorithm achieved sensitivity higher than 96% and specificity higher than 92%. While the GSSP algorithm and NNSplice

Please cite this article in press as: Tsai, K.-N., et al., Genomic splice site prediction algorithm based on nucleotide sequence pattern for RNA viruses. Comput. Biol. Chem. (2008), doi:10.1016/j.compbiolchem.2008.08.002

Table 2
The performances of the GSSP, NNSplice, and SplicePredictor algorithms on the donor sites of the five RNA viruses in the Orthomyxoviridae family

Species	GSSP		NNSplice		SplicePredictor	
	Sen (%)	Spe (%)	Sen (%)	Spe (%)	Sen (%)	Spe (%)
IFA7	99.63	94.86	99.48	95.56	0	98.84
IFA8	99.8	99.13	66.32	96.39	0.94	98.84
IFB8	99.43	99.9	98.3	97.6	4.55	92.28
IFC7	100	99.9	100	100	58	98.56
ISAV7	100	97.66	0	91.5	100	100
Tho6	100	97.38	100	94.76	100	98.43

Table 3
The performances of the GSSP, NNSplice, and SplicePredictor algorithms on the acceptor sites of the five RNA viruses in the Orthomyxoviridae family

Species	GSSP		NNSplice		SplicePredictor	
	Sen (%)	Spe (%)	Sen (%)	Spe (%)	Sen (%)	Spe (%)
IFA7	96.1	96.73	99.97	95.95	45.13	99.94
IFA8	99.89	96.9	99.91	96.11	37.79	99.83
IFB8	98.86	93.01	95.45	96.76	97.16	96.11
IFC7	100	92.86	100	96.33	58	98.58
ISAV7	100	97.93	100	98.5	60	98.83
Tho6	100	95.95	100	95.55	100	99.33

had comparable sensitivities, the GSSP algorithm was obviously superior to the SplicePredictor in predicting the acceptor sites of IFA7, IFA8, IFC7, and ISAV7.

The superior performance achieved by the GSSP algorithm might be primarily attributed to the idea of using the sequence pattern mining technique to mine the eigen-patterns. An eigen-pattern was a co-occurring pattern specifying not only the critical positions, but also the position-specific nucleotides. The critical positions and the position-specific nucleotides referred to the nucleotide positions and the nucleotides at these positions in the surrounding region of the di-nucleotides GT/AG, which were assumed to provide sufficient binding force for forming a splice site. This idea had two potential advantages. The first advantage was that the complex dependency among the nucleotides within the surrounding region, which is a high-order function of positions and nucleotides, might be modeled by a set of eigen-patterns. Each eigen-pattern represented a unique data dependency, which was made up of nucleotides and their critical positions. Note that different eigen-patterns might have different critical positions. The second advantage was using the sequence pattern mining technique allowed us to find the eigen-patterns of low nucleotide occurrence frequencies but with high confidences. On the other hand, the nucleotides frequency, at each position, played an important role in both the NNSplice and the SplicePredictor. Those splice sites with low-frequency nucleotides in the vicinity tended to be easily missed even though both algorithms were much more sophisticated than simply counting the nucleotide frequencies. More detailed analyses on the performance of the GSSP algorithm may be found in the Supplemental Data at http://homepage.ntu.edu.tw/~d91548013/Supplementary_data.pdf.

To check the applicability of the GSSP algorithm to other species of RNA viruses, we had applied the GSSP algorithm to the human immunodeficiency virus type 1 (HIV-1). HIV-1 was chosen because it was a highly important RNA virus and had 8 different strains, which was a reasonable sample size statistically. Note that the numbers of different strains available are mostly smaller than 8 for the RNA viruses currently. As a result, the numbers of real and false HIV-1 donor sites used were 14 and 3351. The sensitivity and specificity attained by the GSSP algorithm for the donor site were 100 and 93.97%, respectively. The numbers of real and false HIV-1 acceptor sites used were 14 and 8004. The sensitivity and specificity attained by the GSSP algorithm for the acceptor site were 100 and 91.14%, respectively. The performances partially supported that the GSSP algorithm might be used for various RNA viruses.

Since RNA viruses are known to infect vertebrates, which means the viruses exploit the spliceosomes in their vertebrate hosts, one reasonable choice for training data would be the human nucleotide sequences. Nevertheless, our study showed that the performances achieved by using *Drosophila* as the training data was better than those by using human nucleotide sequences. While the real reason requires further investigation, one possible explanation may be that not all human splice sites are connected to RNA viruses. Using all human nucleotide sequences as the training data may bring in unnecessary noises in pattern mining process. On the other hand, *Drosophila* has recently been considered as a reasonable model for studying the interactions between RNA viruses and host cells. For examples, RNA viruses, such as Flock house virus and Influenza virus, has been shown to be suppressors of RNA silencing in *Drosophila* via mediating nucleic acid-based antiviral response (Li et al., 2002, 2004). RNA virus infection mechanism has been investigated in *Drosophila* (Adamson et al., 2005; Chotkowski et al., 2008). *Drosophila* RNAi screen model has been employed to identify host genes, which are essential for influenza virus replication (Hao et al., 2008).

4. Conclusions

A new algorithm, called GSSP algorithm, was proposed for the splice site prediction of RNA viruses in this paper. The uniqueness of the GSSP algorithm consisted in its idea of characterizing the interdependency among the nucleotides and base positions based on the eigen-patterns. Compared with the conventional approaches, the GSSP algorithm was shown to have similar specificities but much better sensitivities than its counterparts, especially in the prediction of donor sites. In addition, the GSSP algorithm had the advantage of identifying the splice sites with multiple nucleotides of low-occurrence frequency in the vicinity of the GT/AG di-nulceotides. It is a phenomenon frequently observed around the splice sites of the RNA viruses due to the high mutation frequency. Moreover, the capability of cross-species prediction rendered the GSSP algorithm a powerful tool to predict the

Please cite this article in press as: Tsai, K.-N., et al., Genomic splice site prediction algorithm based on nucleotide sequence pattern for RNA viruses. Comput. Biol. Chem. (2008), doi:10.1016/j.compbiolchem.2008.08.002

splice sites of an RNA virus with a very limited number of known strains.

References

Adamson, A.L., Wright, N., LaJeunesse, D.R., 2005. Modeling early Epstein-Barr virus infection in Drosophila melanogaster: the BZLF1 protein. Genetics 171, 1125–1135.

Baten, A.K., Chang, B.C., Halgamuge, S.K., Li, J., 2006. Splice site identification using probabilistic parameters and SVM classification. BMC Bioinformatics 7, S15.

Brendel, V., Kleffe, J., 1998. Prediction of local optimal splice sites in plant pre-mRNA with applications to gene identification in Arabidopsis thaliana genomic DNA. Nucl. Acids Res. 26, 4748–4757.

Burge, C., Karlin, S., 1997. Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268, 78–94.

Chen, T.M., Lu, C.C., Li, W.H., 2005. Prediction of splice sites with dependency graphs and their expanded Bayesian networks. Bioinformatics 21, 471–482.

Chotkowski, H.L., Ciota, A.T., Jia, Y., Puig-Basagoiti, F., Kramer, L.D., Shi, P.Y., Glaser, R.L., 2008. West Nile virus infection of Drosophila melanogaster induces a protective RNAi response. Virology 377, 197–206.

Degroeve, S., De Baets, B., Van de Peer, Y., Rouzé, P., 2002. Feature subset selection for splice site prediction. Bioinformatics 18, S75–S83.

Degroeve, S., Saeys, Y., De Baets, B., Rouzé, P., Van de Peer, Y., 2005. SpliceMachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21, 1332–1338.

Domingo, E., 1997. Rapid evolution of viral RNA genomes. J. Nutr. 127, 958S–961S.

Hao, L., Sakurai, A., Watanabe, T., Sorensen, E., Nidom, C.A., Newton, M.A., Ahlquist, P., Kawaoka, Y., 2008. Drosophila RNAi screen identifies host genes important for influenza virus replication. Nature, doi:10.1038/nature07151.

Li, H., Li, W.X., Ding, S.W., 2002. Induction and suppression of RNA silencing by an animal virus. Science 296, 1319–1321.

Li, W.X., Li, H., Lu, R., Li, F., Dus, M., Atkinson, P., Brydon, E.W., Johnson, K.L., Garcia-Sastre, A., Ball, L.A., Palese, P., Ding, S.W., 2004. Interferon antagonist proteins of influenza and vaccinia viruses are suppressors of RNA silencing. Proc. Natl. Acad. Sci. U.S.A. 101, 1350–1355.

Mount, S.M., 1982. A catalogue of splice junction sequences. Nucl. Acids Res. 10, 459–472.

Pertea, M., Lin, X., Salzberg, S.L., 2001. GeneSplicer: a new computational method for splice site prediction. Nucl. Acids Res. 29, 1185–1190.

Reese, M.G., Eeckman, F.H., Kulp, D., Haussler, D., 1997. Improved splice site detection in Genie. J. Comput. Biol. 4, 311–323.

Senapathy, P., Sharpiro, M.B., Harris, N.L., 1990. Splice junctions, branch point sites, and exons: sequence statistics, identification, and applications to genome project. Methods Enzymol. 183, 252–278.

Please cite this article in press as: Tsai, K.-N., et al., Genomic splice site prediction algorithm based on nucleotide sequence pattern for RNA viruses. Comput. Biol. Chem. (2008), doi:10.1016/j.compbiolchem.2008.08.002