Bioinformatics, 2017, 1–7
doi: 10.1093/bioinformatics/btx401
Advance Access Publication Date: 19 June 2017
Original Paper

Genome analysis

BPP: a sequence-based algorithm for branch point prediction

Qing Zhang[1,†], Xiaodan Fan[2,†], Yejun Wang[3], Ming-an Sun[1], Jianlin Shao[4] and Dianjing Guo[1,*]

[1]School of Life Sciences and the State Key Laboratory of Agrobiotechnology, [2]Department of Statistics, The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China, [3]Department of Cell Biology and Genetics, Shenzhen University Health Science Center, Shenzhen 518060, China and [4]First Affiliated Hospital, School of Medicine, Zhejiang University, Hangzhou, China

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

Associate Editor: John Hancock

Received on January 11, 2017; revised on May 22, 2017; editorial decision on June 14, 2017; accepted on June 16, 2017

Abstract

Motivation: Although high-throughput sequencing methods have been proposed to identify splicing branch points in the human genome, these methods can only detect a small fraction of the branch points subject to the sequencing depth, experimental cost and the expression level of the mRNA. An accurate computational model for branch point prediction is therefore an ongoing objective in human genome research.
Results: We here propose a novel branch point prediction algorithm that utilizes information on the branch point sequence and the polypyrimidine tract. Using experimentally validated data, we demonstrate that our proposed method outperforms existing methods.
Availability and implementation: https://github.com/zhqingit/BPP.
Contact: djguo@cuhk.edu.hk
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Eukaryotic pre-mRNA splicing involves a set of reactions catalyzed by the spliceosome, a protein complex consisting of five small nuclear ribonucleoproteins (the U1, U2, U4, U5 and U6 snRNPs) and hundreds of other proteins (Burge *et al.*, 1999; Jurica and Moore, 2002). Through alternative splicing, the transcription of a gene can generate multiple isoforms by selectively expressing different exon sequences, which thereby contributes significantly to proteome complexity in metazoan (Graveley, 2001; Maniatis and Tasic, 2002). Recent studies suggest that ‘spliceosomal mutations’ can result in cancer-specific mis-splicing, which can be therapeutically exploited using compounds that influence the splicing process (Dvinge *et al.*, 2016). The importance of splicing is also illustrated by the fact that more than 200 human diseases arise from the disruption of splicing by mutations either in the splicing sites or in the *cis*-acting splicing regulatory sites (Cieply and Carstens, 2015; Chabot and Shkreta, 2016).

A key step in determining the intron and exon to be spliced out or retained is the recognition of the branch point sequence (BPS) by the ribonucleoprotein U2 snRNP. The conserved GTAGTA hexanucleotide in U2 snRNA can base-pair with the BPS, forcing the branch point adenosine to flip out and form a bulge between the fifth and the sixth base. This initiates a nucleophilic attack at the intronic 5 splice site, thus starting the first of the two transesterification reactions that mediate splicing. Recently, researchers have attempted to identify branch points (BPs) by high-throughput sequencing approach (Bitton *et al.*, 2014; Mercer *et al.*, 2015), in which the intronic lariats were first enriched, and then sequenced to identify the position of the human BPs. However, due to the quick degradation of the lariat structure in the nucleus (Moore, 2002), these sequencing methods can only detect a small fraction of the lariats given the moderate sequencing depth. In theory, these sequencing methods also cannot identify the BPs for tissue-specific genes or genes with low expression level. For example, Mercer *et al.* (2015) only identified 59 359 high-confidence BPs out of a total of ∼300 000 human introns.

Efficient computational tools for human BP prediction would greatly facilitate human genome research. However, *in silico*, the identification of human BPs is rather challenging, mainly because human BPSs are highly variable and extremely degenerate. BPs have been successfully predicted in fungal species based on the Hamming distance of the BPS to the U2-complementary sequence (Kupfer *et al.*, 2004), whereas this approach has proven insufficient for human genomes (Corvelo *et al.*, 2010). Corvelo *et al.* (2010) introduced SVM-BPfinder, a method utilizing Support Vector Machines to predict human BPs using a set of high-confidence putative BPSs based on conservation and positional bias across seven mammalian species. SVM-BPfinder performs very well for introns with BPS containing a 'TNA' structure but it is unable to predict non-canonical BPS without the 'TNA' structure. HSF is an online bioinformatics tool to predict the effects of mutations on splicing signals (Desmet *et al.*, 2009). It can identify BPS based on a position weight matrix (PWM) from the consensus sequence YNYCRAY. However, the lack of a standalone version limits its wider application.

In this article, we present a novel branch point prediction (BPP) method, integrating the characteristics of BPS and polypyrimidine tract. Our method is innovative in two aspects: (i) by pre-processing the raw data and selecting a suitable start point, degenerate BPS motif in human can be inferred using the mixture model (MM), a popular motif inference method; and (ii) using relative frequencies for different nucleotides instead of the most commonly used arbitrary scores to represent the effects of polypyrimidine tract (PPT). Our model not only estimates the affinity between the protein and the PPT, but also considers the co-evolution between the protein and the binding sequences. Using a set of experimentally validated data, we demonstrate that BPP outperforms previously published methods and thus provides a promising alternative method for BPP for human genome study.

2 Materials and methods

2.1 Training dataset for BPS motif inference and octanucleotide enrichment

Firstly, all of the human intron coordinates from the UCSC Genome Browser table were obtained and the intron sequences were extracted. Then the introns used in the testing dataset (introns verified experimentally or by high-throughput sequencing) were removed. Finally, the remaining introns longer than 300 bp were kept to be the training dataset (223 606 introns).

2.2 Inferring the energy motif of a BPS

According to Stormo and Fields (1998) and Granek and Clarke (2005), an element in PWM can be interpreted as the contribution of the corresponding base to the relative free energy of the motif. For each position j, we botain a set of equations:

$$\begin{cases} \Delta G = RTln\left(f_{jb}/p_b\right), \\ \sum_b f_{jb} = 1 \end{cases} \tag{1}$$

where $b \in \{\text{A,C,G,T}\}$, $j = 1, \ldots, L$ and L is the length of the motif, R is the gas constant, T is the temperature, f_{jb} is the observed frequency of base b at position j and p_b is the background probability of base b. Here $T = 300$ K is used, and p_b is the statistical frequency based on a 50 nt segment located 300 nt upstream of all introns, chosen because no general splicing element is reported in this segment.

The program RNAcofold from the Vienna RNA package (Hofacker *et al.*, 1994) was used to calculate the binding free energy between a heptanucleotide and the U2 snRNA. Specifically, heptanucleotides with a fixed base b at position j were forced to undergo a complete pairing with nucleotides (GTAGTA) from the U2 snRNA, with the exception of the BP adenosine, which was forced not to pair with any nucleotide. The average free energy of all of these heptanucleotides was represented as the relative free energy contribution from b at position j. Because $\Delta G, R, T$ and p_b are known, f_{jb} was calculated using Equation (1) and normalized by forcing the sum to be unity.

2.3 Updating the motif of BPS using MM

2.3.1 Mixture model and expectation maximization algorithm

The expectation maximization (EM) algorithm searches for maximum likelihood estimates of the parameters of a finite MM, generating a given dataset of sequences (Bailey *et al.*, 1994). We assume that a set of heptanucleotides $S = \{s_1, s_2, \ldots, s_n\}$, where n is the number of heptanucleotides, arises from two mixture components: background and BPS. For convenience, we use 0 and 1 to represent the two components respectively. The probability of discovering s_i can be written as follows:

$$p(s_i|MM) = \lambda_0 * p\left(s_i|\theta_{background}\right) + \lambda_1 * p(s_i|\theta_{BPS}), \tag{2}$$

where λ_0 and λ_1 are the probabilities that the background and BPS components are respectively responsible for generating S, and $\lambda_0 + \lambda_1 = 1$. The parameter $\theta_{\text{background}}$ is a vector consisting of the frequencies of four nucleotides (F_b) in the background region; and θ_{BPS} represents the PWM of the BPS motif. The PWM expresses the frequency with which nucleotide b appears at position j: f_{jb}, where $j = 1, \ldots, L$, and L is the length of the motif. F_b expresses the frequency of nucleotide b: f_{0b}.

The EM algorithm for finite MMs is used to find the parameters that maximize the likelihood of the data. For details, please refer to the following articles: (Aitkin and Rubin, 1985; Bailey and Elkan, 1995; Bailey *et al.*, 1994).

2.3.2 Initial value and training dataset

To construct the training dataset, we collected a set of human intron sequences longer than 300nt, and extracted three subsequences from upstream of the 3'SS in each sequence: 21–34 nt, 187–200 nt and 3–16 nt. The three sets of subsequences respectively correspond to the BPS region, background region and PPT region.

The so-called BPS and PPT regions only represent the sequence blocks where BPS and PPT often locate and do not always contain BPS or PPT. The reasons of selecting 14 bp include: (a) many branch point sites locate at the 22nd nucleotide of upstream of the 3SS, so the PPT region can only be defined behind the 21st nucleotide; (b) a space of four nucleotides is given to avoid the repulsion between U2 snRNP and U2AF6B; (c) the last two nucleotides of 3SS locate at the splice donor site. So we have the PPT region from 16 to 3 bp upstream of the 3SS of the introns. To keep consistent, we also pick 14 bp for BPS and background regions.

To enrich the possible BPS, we decomposed the subsequences in each set into heptanucleotides, and compared the frequency of each heptanucleotide across the three sets. Specifically, we performed the

Fisher's exact test for each heptanucleotide between the BPS set and the background set, and between the BPS set and the PPT set. Only the heptanucleotides significantly enriched in both tests ($P \leq 0.05$) were selected to form the training dataset.

An initial value for the EM is essential to discover the desired motif (Bailey *et al.*, 1994) due to the highly irregular likelihood landscape in the high dimensional parameter space, especially for degenerative motifs, such as that of human BPS. Here, the energy motif of the BPS and the frequencies of four nucleotides (A,C,G,T) in the background region were used as the initial guess for the EM search.

2.4 Calculating the BPS score of a heptanucleotide based on the PWM of the BPS motif

For any heptanucleotide, the BPS score is:

$$s = \prod_{i=1}^{7} f_{ib_i}, \tag{3}$$

where b_i is the base in the i-th position of the heptanucleotide, and f_{ib_i} is the observed frequency of base b_i in the i-th position of the BPS motif.

2.5 Generating the weighted octanucleotides

As shown by Stormo and Fields (1998) and Granek and Clarke (2005), the relative frequency of the nucleotide at each position in one motif is related to the binding free energy. If we consider the octanucleotide as an unit, its relative frequency should correlate with the binding affinity of U2AF65. To avoid the situation that the count of one ocatanucleotide is very small in the PPT region, but the relative frequency is very high, we also consider the frequency of each ocatanucleotide in the PPT region. We decompose the sequences in the PPT region and background region into octanucleotides, and weight each octanucleotide by

$$w_{octa} = \frac{C_{ppt} * C_{ppt}}{C_{background} * \sum C_{ppt}}, \tag{4}$$

$$w^*_{octa} = w_{octa} / \sum w_{ocat} \tag{5}$$

where C_{ppt} is the octanucleotide count in the PPT set, $C_{\text{background}}$ is the octanucleotide count in the background set, and w^*_{ocat} is the normalized weight.

2.6 Finding the AGEZ

Following Corvelo *et al.* (2010) and Gooding *et al.* (2006), the region between the 3'SS and the first 'AG' dinucleotide whose distance to the 3'SS is longer than 12 nt is defined as the AGEZ in our method.

2.7 The BPP method

The BPP slides along the AGEZ. At each site, the first seven nucleotides are considered as a candidate BPS, which produces a score based on the motif of the BPS:

$$s_{bps} = \prod_{i=1}^{7} f_{ib_i}, \tag{6}$$

where b_i is the base in the i-th position of the heptanucleotide, and f_{ib_i} is the observed frequency of base b_i in the i-th position of the BPS motif. BPP estimates the contribution of U2AF6B by summarizing the weighted octanucleotides in the sequence of 20 bp immediately downstream from the candidate BPS:

$$s_{ppt} = \frac{\sum_{i=1}^{L-8+1} w^*_{octa}}{L}. \tag{7}$$

We respectively normalize the two scores to range from 0 to 1:

$$s^*_{bps} = s_{bps} / s_{(TACTAAC)}, \tag{8}$$

$$s^*_{ppt} = s_{ppt} / s_{20*T}. \tag{9}$$

Finally, the BPP score can be calculated:

$$s = s^*_{bps} * s^*_{ppt} \tag{10}$$

2.8 Z-score calculation

We normalize the BPP score for each position by calculating the Z-score:

$$z_i = \frac{s_i - \mu}{\sigma}, \tag{11}$$

where s_i is the BPP score of position i; μ and σ are respectively the mean and the standard deviation of the BPP scores of all positions in the AGEZ.

3 Results

3.1 BPP provides a sequence-based method for BPP

BPP is a method for BP prediction in pre-mRNA splicing. The prediction is based on the features of a BPS motif inferred from a MM, and a set of weighted octanucleotides representing the binding affinity between the U2AF65 and the PPT (see Fig. 1a). Using the U2 snRNP sequence, we first infer an intermediate motif, termed the 'energy motif' to represent the free energy (see Section 2). Then, the BPS motif is derived by using an MM, which initializes from the energy motif and is trained on a set of BPS-enriched heptanucleotides. The influence of the PPT is quantified through a set of octanucleotides, weighted by the relative frequencies with which they occur in the PPT and the background regions. When using the trained model for prediction, a sliding window moves along an intron sequence to calculate the score of the BPS and the affinity between the

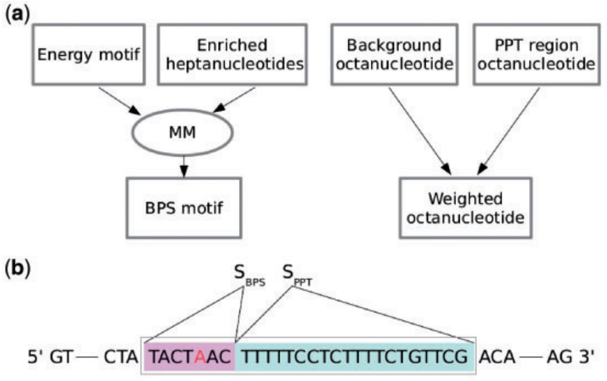
(a)
Energy motif
Enriched heptanucleotides
Background octanucleotide
PPT region octanucleotide
MM
BPS motif
Weighted octanucleotide
(b)
S_BPS
S_PPT
5' GT— CTA TACTAAC TTTTCCTCTTTTCTGTTCG ACA — AG 3'

Fig. 1. Flowchart of BPP. (**a**) Left: Based on the energy motif and the enriched heptanucleotides, the mixture model is trained to infer the motif of the BPS; right: By comparing the octanucleotide frequencies in the background and the PPT regions, each octanucleotide was weighted to represent the affinity of U2AF6B to the PPT. (**b**) A window (black rectangle) slides along the intron sequence; S_{BPS} is calculated based on the candidate BPS and S_{PPT} is calculated based on the sequence of 20 bp immediately downstream from the candidate BPS

downstream sequence and the U2AF65 at each site. Finally, the site with the highest score was selected as the predicted BP site (Fig. 1b, for details see Section 2).

3.2 The inferred energy motif of BPS is consistent with the experimental results

In this paper, the BPS region is defined as the 21st to the 34th nucleotides (nt) upstream of the 3′ splice site (3′SS), where most branch points are located (Mackereth *et al.*, 2011). The PPT region is defined as the 3rd to the 16th nt upstream of the 3′SS. The background region is defined as the 187th to the 200th nt upstream of the 3′SS, because no general splicing element is reported in this region. Through their contrast with the BPS and PPT regions, the background regions provide statistical clues about the features of the true signal. Importantly, these defined regions are only used for model training, not for prediction.

MM has been successfully used for motif discover in a set of DNA/RNA sequences (Bailey *et al.*, 1994). However, the human BPS motif is highly degenerate (Gao *et al.*, 2008) and the fitting of the MM is very sensitive to the initial value. MM is therefore unable to predict the true BPS motif reliably if it is trained using only upstream sequences. Additionally, the co-occurrence of the PPT sequence and the BPS may interfere with the inference process (Corvelo *et al.*, 2010). To overcome these problems, we use a set of enriched heptanucleotides in the BPS region to train the MM, which starts from an inferred energy motif of the BPS.

The RNA-RNA base pairing between the GUAGUA motif in the U2 snRNP and the BPS is important in human pre-mRNA splicing (Wu and Manley, 1989; Zhuang and Weiner, 1989). To supply the MM with an initial guess as close as possible to the real motif, we first infer an energy motif of the BPS based on the binding energy between ‘GTATGA’ in the U2 snRNP and all of the heptanucleotides. We name it energy motif to distinguish it from the motif derived from the MM. As the BP is not complementary to any nucleotide in the U2 snRNP, and experimental evidence has shown that the nucleotide Adenine (A) is strongly preferred even though Cytosine (C) and Thymine (T) can also function as the branch nucleotide (Hartmuth and Barta, 1988), we set the relative frequencies of the four nucleotides as: A:0.97, C:0.01, G:0.01, T:0.01 (see Fig. 2a).

This energy motif possesses some experimentally validated characteristics. For example, the fourth and seventh positions are mostly pyrimidines (Gao *et al.*, 2008), and the second and fifth positions are mostly purine (Pastuszak *et al.*, 2011). The heptanucleotide with the highest score (TACTA*C) based on the energy motif is completely complementary to the conserved motif (GTAGTA) in the U2 snRNP, which complies with the finding that TACTAAC is the most efficient BPS for mammalian mRNA splicing (Zhuang *et al.*, 1989). The fact that the derived scores based on the inferred energy motif are highly correlated with the binding energy highlights the rationality of this inference (Supplementary Fig. S1).

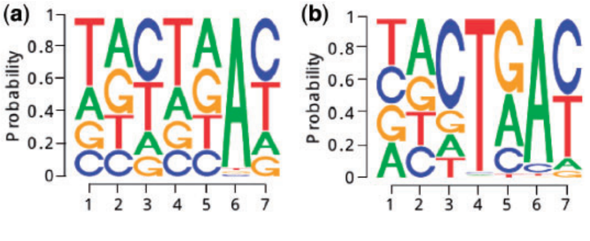

Fig. 2. BPS motifs. (**a**) Energy motif; (**b**) BPS motif inferred by the mixture model

3.3 Customized MM successfully infers the degenerate BPS motif

The BPS motif inferred by MM is shown in Figure 2b. Compared with the above energy motif, this inferred BPS motif captures most characteristics of the human BPS: (i) the frequency of T at the fourth position reaches ~98%, and pyrimidines frequently appear at the third and the seventh positions. This supports the experimental finding that the human branch point consensus sequence is nyUnAyn (Gao *et al.*, 2008); (ii) the frequency of T at the 1st position declines, and the 1st and the 2nd positions carry little information; and (iii) in accordance with results from previous studies (Corvelo *et al.*, 2010; Harris and Senapathy, 1990), enriched purines and depleted T are preferentiably found at the fifth position. In contrast, the inferred motif was not successfully located when the MM was trained using an un-enriched dataset and a random start. The relative frequencies of the four nucleotides at the BP site did not affect the final inferred motif (see Supplementary Fig. S2).

3.4 The weighted octanucleotides show binding affinity to the U2AF65

Because the PPT-U2AF65 binding is essential for efficient BP utilization and 3’SS recognition in metazoans (Frendewey and Keller, 1985; García-Blanco *et al.*, 1989; Reed, 1989; Reed and Maniatis, 1985; Roscigno *et al.*, 1993; Ruskin and Green, 1985; Wieringa *et al.*, 1984), the PPT also incorporates the binding affinity between the U2AF65 and the sequences downstream of the BPS.

In previous studies, the PPT score was calculated as the contribution of U2AF65 based on its pyrimidine content (Clark and Thanaraj, 2002; Corvelo *et al.*, 2010) or by statistical tests (Schwartz *et al.*, 2008), and these scoring systems are universal across different species. Considering, however, that the PPT may have co-evolved with the U2AF65 RNA recognition motifs (Schwartz *et al.*, 2008), we deem it more reasonable to construct a species-specific PPT scoring system.

We use weighted octanucleotides to represent the U2AF65 binding affinity because either octanucleotides or nonanucleotides may be the basic binding elements of U2AF65 (Ito *et al.*, 1999; Mackereth *et al.*, 2011; Sickmier *et al.*, 2006). The weight score reflects the relative frequency of an octanucleotide in the PPT and in the background regions (Section 2), and therefore, a small fraction of octanucleotides (the long right tail in Supplementary Fig. S3) frequently occur at the PPT region (Supplementary Fig. S2). This fraction includes octanucleotides known to be preferred by human U2AF65, including TTTTTTTT, TTTCTTTT, TTTTCTTT and others. The finding that higher scores correlate with a stronger affinity between the octanucleotide and the U2AF65 validates the effectiveness of our method.

3.5 Performance comparison between BPP and other methods

A performance comparison was conducted for BPP, SVM-BPfinder and HSF. Of these three methods, both BPP and SVM-BPfinder predict BPs in the AGEZ region (Gooding *et al.*, 2006) of the introns. BPP starts from 4 bp upstream of 3′SS, and if the distance between the candidate BPS and the 3′SS is shorter than 9 bp, then BPP does not integrate the contribution of the PPT. In contrast, SVM-BPfinder starts from 15 bp upstream of 3′SS. Based on the output of the online system, HSF predicts BPs in the region from 18 to 100 bp upstream of the 3′SS.

To comprehensively evaluate the three methods, a set of 86 introns with experimentally verified BPs (Ajiro and Zheng, 2015;

Burrows *et al.*, 1998; Chavanas *et al.*, 1999; ; Corvelo *et al.*, 2010; Darman *et al.*, 2015; Goux-Pelletan *et al.*, 1990; ; Gooding *et al.*, 2006; Gao *et al.*, 2008; Helfman and Ricci, 1989; Janssen *et al.*, 2000; Li and Pritchard, 2000; Maslen *et al.*, 1997; Mayer *et al.*, 2000; Smith and Nadal-Ginard, 1989; Southby *et al.*, 1999; Webb *et al.*, 1996;) and a set of 47 294 introns with sequencing-verified BPs (Mercer *et al.*, 2015) were collected and used for the comparison. The comparisons were conducted in the following three ways:

1. In their default settings, the methods predict the position with the highest score to be the BP. Based on the corresponding results, we compared the correctly predicted introns of each method.
2. A receiver operating characteristic (ROC) curve was built at the intron level for the three methods based on a testing set consisting of positive introns with known BPs (86 experimentally verified and 47 294 NGS verified BPs) and a set of randomly generated negative sequences.
3. A ROC curve was built at the position level for each method, taking the BP and non-BP positions in the introns as the positives and negatives respectively.

3.6 Comparison based on the default setting outputs of each method

For performance comparison, only real intron sequences were used as inputs and the positions with the highest score for each intron were selected. For HSF, only positions in the AGEZ were selected. Here, precision was defined as the number of introns with a correctly predicted BP divided by the total number of introns. Given the fact that the U2 snRNA can hybridize to the adenosines at either the fifth or the sixth position of the BPS and bulge out of the other (Query *et al.*, 1994), and that reverse transcriptase can stop within 1 nt of the branch point, (Rodriguez *et al.*, 1984; Zeitlin and Efstratiadis, 1984). The prediction was also considered correct if the predicted BP was −1 or 1 nt away from of the real BP position. We supplied both one nucleotide error tolerant and no error tolerant versions. For introns containing more than one BP, the prediction was marked correct if any one of the BPs was detected. Because SVM-BPfinder only finds BPS containing TNA, its prediction was considered false for introns that did not include any TNA-containing BPS.

For the 86 experimentally verified introns, the precision levels of BPP, SVM-BPfinder and HSF were 68.6, 61.6 and 30.2%, respectively (Fig. 3a, Supplementary Fig. S4a). Even when only considering the introns with a TNA structure, BPP still outperformed SVM-BPfinder: in these cases, BPP correctly predicted 77.6% of the BPs, compared with 59.7% for SVM-BPfinder and 34.2% for HSF. The full list of the 86 introns with the corresponding predictions by the three methods can be accessed in Supplementary Table S1. Subsequently, we compared the three methods based on the 47 294 introns with NGS-verified BPs. Again, BPP outperformed the other methods (Supplementary Fig. S5a and c).

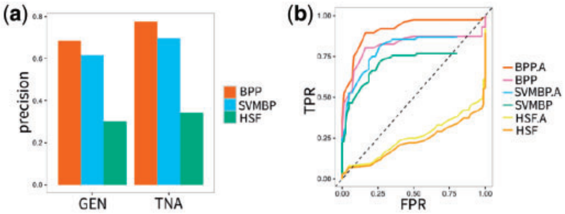

Fig. 3. Performances of BPP, SVM-BPfinder (labelled by SVMBP in the figure) and HSF based on the experimentally verified BPs (1 nucleotide error tolerant). (**a**) Performances of the three methods using their default settings: GEN corresponds to the general cases (i.e. including all instrons) and TNA corresponds to specifically those introns with a TNA structure; (**b**) ROC curves of the three methods at the intron level; the suffix 'A' corresponds to the TNA cases

3.7 Comparison using the ROC curves at the intron level

To build the ROC curves, a set of sequences with the same lengths as the introns in the positive dataset was randomly generated as the negative dataset. For the experimentally verified BPs, the testing dataset consisted of 86 positives and 86 negatives; for the NGS-verified BPs, the testing dataset consisted of 47 294 positives and 47 297 negatives. As described earlier, the prediction was considered correct if the predicted BP was located within −1 or 1 nt of the real BP position. The ROC curves under the general and TNA-only conditions are illustrated in Figure 3b (Supplementary Fig. S4b) and Supplementary Figure S5b (Supplementary Fig. S5d), and show that BPP again outperformed the other two methods.

3.8 Comparison using ROC curves at the position level

For this comparison, two ROCs were created for each method based on two different values: (i) the absolute score of each position and (ii) the ranking of each position based on its score relative to all of the others. The ranking values were used because for each position, the score was inferred independently, i.e. neglecting the competition of other positions, and because the scores across the introns might not be mutually compatible. The true-positive rate was unable to reach 1 because some of the real BP positions lay outside of the regions selected for prediction by the three methods. As shown in Supplementary Figure S4c and d, for all three methods, the evaluation methods by ranking performed better than the one by score, indicating the positions of the BPs were influenced by the competitive effects of other positions. Again, BPP gave the best performance and HSF outperformed SVM-BPfinder.

In general, BPP and SVM-BPfinder are recommended if the goal is to predict the most likely BP for each intron. However, if the goal is to predict multiple BPs in one intron, BPP and HSF should be preferred methods.

3.9 Genome-wide BP prediction in human

Using BPP, all of the introns in the human genome ($N = 281\,787$) were scanned for BPs (Supplementary Table S2). As shown in Figure 4a, 94.6% ($N = 266\,707$) of the predicted BPs were located within 0–50 upstream of the 3′SS, consistent with previous reports

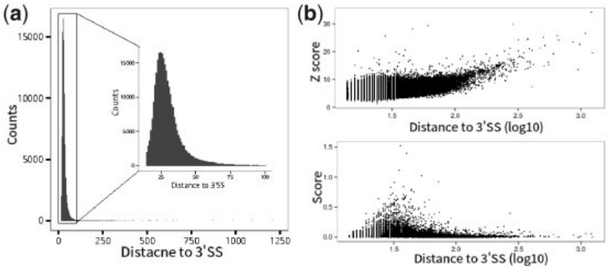

Fig. 4. Genome-wide analysis of human BPs. (**a**) Distribution of the distances between human branch points and their corresponding 3′SS; (**b**) top panel: Variation of the Z-score with the distance from the corresponding BP to its 3′SS; bottom panel: Variation of the BPP score with the distance from the corresponding BP to its 3′SS

(Corvelo *et al.*, 2010; Mercer *et al.*, 2015), that most BPs are discovered at 23–25 nt upstream of the 3′SS.

The presence of distal BPs in the predictions prompted us to speculate on how they are selected in the splicing process. The distances of the BPs to the 3′SS were plotted against their relative BPP scores, i.e. the normalized Z scores across the BPP scores of all positions in the AGEZ (Fig. 4b top). Interestingly, the Z scores of the distal BPs increased rapidly with the distance to the 3′SS at distances longer than ~100 bp. However, the absolute BPP scores were still very low for most of the distal BPs (Fig. 4b bottom). Hence this suggests that the distal BPs are selected due to the even much lower scores of the positions close to the 3′SS. Given the fact the absolute BPP scores reflect the physical affinity between the RNA sequence and the splicing proteins, it seems likely that most of the introns with distal BPs are partially spliced, which is consistent with the conclusion from paper (Bitton *et al.*, 2014). We further plotted the frequencies of the alternative ends of the introns under different BP distance to 3′SS (Supplementary Fig. S6). If the end of the intron can be found in all the transcripts, we define it as the constitute end, otherwise the alternative end. In accord with the Figure 4b, the frequencies of the alternative ends increase when the distal BPs are used. However it is surprising that we found a minimal value at 122 bp for the downstream alternative ends, which we cannot explain based on known knowledge. But it might be a good start to explore some specific biological functions in the RNA splicing process in the future.

Searching the ClinVar (Landrum *et al.*, 2014) database, we found that 42 BPs located in gene related to various human diseases, including hyperoxalurea type III, colorectal cancer, supratentorial primitive neuroectodermal tumours and Alzheimer's disease, among others (Supplementary Table S2). In future, it will be interesting to examine the possibility that the mutation of BPs might alter the mRNA splicing process and result in these diseases.

4 Discussion

With the development of high-throughput sequencing technology, it is now possible to identify thousands of human BPs in one experiment. However, due to experimental costs and low gene expression levels, only ~20% of the total human introns have been validated by the sequencing approach. Computational methods for human BP prediction are complicated by the degeneracy of the motifs and the involvement of other auxiliary elements, such as the PPT. To overcome these limits, we proposed an algorithm (BPP) to predict the branch points of human introns based only on the sequence information. By using a set of experimentally verified BPs, we showed that BPP outperformed the other currently available programs. However, BPP still has some limitations. For example, the BPS motif inference and octanucleotide weighting process are conducted separately, which may hinder the accurate weight estimation of certain heptanucleotides or octanucleotides, and result in loss of information from other elements, because the BPS and the PPT co-exist in a contiguous sequence. In general, a framework combining these two processes into a unified function may further improve the predictive accuracy of BPP. In future work, we will refine the current MM to accommodate the combination of different elements.

Funding

This work was supported by a grant from the Hong Kong University Grants Committee Hong Kong UGC/AoE Plant & Agricultural Biotechnology Project grant AoE-B-07/09, a grant from the Shenzhen Science and Technology Committee [grant no. JCYJ20140425184428456] and a grant from the Hong Kong Research Grants Council [Project No. CUHK400913].

Conflict of Interest: none declared.

References

Aitkin,M. and Rubin,D.B. (1985) Estimation and hypothesis testing in finite mixture models. *J. Roy. Stat. Soc. B (Methodological)*, **47**, 67–75.

Ajiro,M. and Zheng,Z.-M. (2015) Vemurafenib-resistant braf selects alternative branch points different from its wild-type braf in intron 8 for rna splicing. *Cell Biosci.*, **5**, 1.

Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.*, **21**, 51–80.

Bailey,T.L. *et al.* (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. In: *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, California, pp. 28–36.

Bitton,D.A. *et al.* (2014) Lasso, a strategy for genome-wide mapping of intronic lariats and branch points using rna-seq. *Genome Res.*, **24**, 1169–1179.

Burge,C.B. *et al.* (1999) 20 splicing of precursors to mrnas by the spliceosomes. *Cold Spring Harbor Monogr. Arch.*, **37**, 525–560.

Burrows,N.P. *et al.* (1998) A point mutation in an intronic branch site results in aberrant splicing of col5a1 and in ehlers-danlos syndrome type ii in two british families. *Am. J. Hum. Genet.*, **63**, 390–398.

Chabot,B. and Shkreta,L. (2016) Defective control of pre–messenger rna splicing in human disease. *J. Cell Biol.*, **212**, 13–27.

Chavanas,S. *et al.* (1999) Splicing modulation of integrin (*β*4 pre-mrna carrying a branch point mutation underlies epidermolysis bullosa with pyloric atresia undergoing spontaneous amelioration with ageing. *Hum. Mol. Genet.*, **8**, 2097–2105.

Cieply,B. and Carstens,R.P. (2015) Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip. Rev. RNA*, **6**, 311–326.

Clark,F. and Thanaraj,T. (2002) Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. *Hum. Mol. Genet.*, **11**, 451–464.

Corvelo,A. *et al.* (2010) Genome-wide association between branch point properties and alternative splicing. *PLoS Comput. Biol.*, **6**, e1001016.

Darman,R.B. *et al.* (2015) Cancer-associated sf3b1 hotspot mutations induce cryptic 3 splice site selection through use of a different branch point. *Cell Rep.*, **13**, 1033–1045.

Desmet,F.-O. *et al.* (2009) Human splicing finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.*, **37**, e67.

Dvinge,H. *et al.* (2016) Rna splicing factors as oncoproteins and tumour suppressors. *Nat. Rev. Cancer*, **16**, 413–430.

Frendewey,D. and Keller,W. (1985) Stepwise assembly of a pre-mrna splicing complex requires u-snrnps and specific intron sequences. *Cell*, **42**, 355–367.

Gao,K. *et al.* (2008) Human branch point consensus sequence is yunay. *Nucleic Acids Res.*, **36**, 2257–2267.

García-Blanco,M.A. *et al.* (1989) Identification and purification of a 62,000-dalton protein that binds specifically to the polypyrimidine tract of introns. *Genes Dev.*, **3**, 1874–1886.

Gooding,C. *et al.* (2006) A class of human exons with predicted distant branch points revealed by analysis of ag dinucleotide exclusion zones. *Genome Biol.*, **7**, 1.

Goux-Pelletan,M. *et al.* (1990) In vitro splicing of mutually exclusive exons from the chicken beta-tropomyosin gene: role of the branch point location and very long pyrimidine stretch. *EMBO J.*, **9**, 241.

Granek,J.A. and Clarke,N.D. (2005) Explicit equilibrium modeling of transcription-factor binding and gene regulation. *Genome Biol.*, **6**, 1.

Graveley,B.R. (2001) Alternative splicing: increasing diversity in the proteomic world. *Trends Genet.*, **17**, 100–107.

Harris,N.L. and Senapathy,P. (1990) Distribution and consenus of branch point signals in eukaryotic genes: a computerized statistical analysis. *Nucleic Acids Res.*, **18**, 3015–3015.

Hartmuth,K. and Barta,A. (1988) Unusual branch point selection in processing of human growth hormone pre-mrna. *Mol. Cell. Biol.*, **8**, 2011–2020.

Helfman,D.M. and Ricci,W.M. (1989) Branch point selection in alternative splicing of tropomyosin pre-mrnas. *Nucleic Acids Res.*, **17**, 5633–5650.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of rna secondary structures. *Chem. Mon.*, **125**, 167–188.

Ito,T. *et al.* (1999) Solution structures of the first and second rna-binding domains of human u2 small nuclear ribonucleoprotein particle auxiliary factor (u2af65). *EMBO J.*, **18**, 4523–4534.

Janssen,R. *et al.* (2000) A branch site mutation leading to aberrant splicing of the human tyrosine hydroxylase gene in a child with a severe extrapyramidal movement disorder. *Ann. Hum. Genet.*, **64**, 375–382.

Jurica,M.S. and Moore,M.J. (2002) Capturing splicing complexes to study structure and mechanism. *Methods*, **28**, 336–345.

Kupfer,D.M. *et al.* (2004) Introns and splicing elements of five diverse fungi. *Eukaryotic Cell*, **3**, 1088–1100.

Landrum,M.J. *et al.* (2014) Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.*, **42**, D980–D985.

Li,M. and Pritchard,P.H. (2000) Characterization of the effects of mutations in the putative branchpoint sequence of intron 4 on the splicing within the human lecithin: cholesterol acyltransferase gene. *J. Biol. Chem.*, **275**, 18079–18084.

Mackereth,C.D. *et al.* (2011) Multi-domain conformational selection underlies pre-mrna splicing regulation by u2af. *Nature*, **475**, 408–411.

Maniatis,T., and Tasic,B. (2002) Alternative pre-mrna splicing and proteome expansion in metazoans. *Nature*, **418**, 236–243.

Maslen,C. *et al.* (1997) A rare branch-point mutation is associated with missplicing of fibrillin-2 in a large family with congenital contractural arachnodactyly. *Am. J. Hum. Genet.*, **60**, 1389–1398.

Mayer,K. *et al.* (2000) Three novel types of splicing aberrations in the tuberous sclerosis tsc2 gene caused by mutations apart from splice consensus sequences. *Biochim. Biophys. Acta (BBA)*, **1502**, 495–507.

Mercer,T.R. *et al.* (2015) Genome-wide discovery of human splicing branchpoints. *Genome Res.*, **25**, 290–303.

Moore,M.J. (2002) Nuclear rna turnover. *Cell*, **108**, 431–434.

Pastuszak,A.W. *et al.* (2011) An sf1 affinity model to identify branch point sequences in human introns. *Nucleic Acids Res.*, **39**, 2344–2356.

Query,C.C. *et al.* (1994) Branch nucleophile selection in pre-mrna splicing: evidence for the bulged duplex model. *Genes Dev.*, **8**, 587–597.

Reed,R. (1989) The organization of 3'splice-site sequences in mammalian introns. *Genes Dev.*, **3**, 2113–2123.

Reed,R. and Maniatis,T. (1985) Intron sequences involved in lariat formation during pre-mrna splicing. *Cell*, **41**, 95–105.

Rodriguez,J.R. *et al.* (1984) In vivo characterization of yeast mrna processing intermediates. *Cell*, **39**, 603–610.

Roscigno,R. *et al.* (1993) A mutational analysis of the polypyrimidine tract of introns. effects of sequence differences in pyrimidine tracts on splicing. *J. Biol. Chem.*, **268**, 11222–11229.

Ruskin,B. and Green,M.R. (1985) Role of the 3 splice site consensus sequence in mammalian pre-mrna splicing. *Nature*, **317**, 732–734.

Schwartz,S. *et al.* (2008) Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res.*, **18**, 88–103.

Sickmier,E.A. *et al.* (2006) Structural basis for polypyrimidine tract recognition by the essential pre-mrna splicing factor u2af65. *Mol. Cell*, **23**, 49–59.

Smith,C.W. and Nadal-Ginard,B. (1989) Mutually exclusive splicing of α-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing. *Cell*, **56**, 749–758.

Southby,J. *et al.* (1999) Polypyrimidine tract binding protein functions as a repressor to regulate alternative splicing of α-actinin mutally exclusive exons. *Mol. Cell. Biol.*, **19**, 2699–2711.

Stormo,G.D., and Fields,D.S. (1998) Specificity, free energy and information content in protein–dna interactions. *Trends Biochem. Sci.*, **23**, 109–113.

Webb,J.C. *et al.* (1996) Genetic variation at a splicing branch point in intron 9 of the low density lipoprotein (ldl)-receptor gene: a rare mutation that disrupts mrna splicing in a patient with familial hypercholesterolaemia and a common polymorphism. *Hum. Mol. Genet.*, **5**, 1325–1331.

Wieringa,B. *et al.* (1984) A minimal intron length but no specific internal sequence is required for splicing the large rabbit β-globin intron. *Cell*, **37**, 915–925.

Wu,J. and Manley,J.L. (1989) Mammalian pre-mrna branch site selection by u2 snrnp involves base pairing. *Genes Dev.*, **3**, 1553–1561.

Zeitlin,S. and Efstratiadis,A. (1984) In vivo splicing products of the rabbit β-globin pre-mrna. *Cell*, **39**, 589–602.

Zhuang,Y. and Weiner,A.M. (1989) A compensatory base change in human u2 snrna can suppress a branch site mutation. *Genes Dev.*, **3**, 1545–1552.

Zhuang,Y. *et al.* (1989) Uacuaac is the preferred branch site for mammalian mrna splicing. *Proc. Natl Acad. Sci. USA*, **86**, 2752–2756.