Comparative analysis of information contents relevant to intron recognition in many species

Hiroaki Iwata
iwata@genome.ist.i.kyoto-u.ac.jp

Osamu Gotoh
o.gotoh@i.kyoto-u.ac.jp

1 Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Keywords: RNA splicing, intron recognition, information content, clustering

1 Introduction

The removal of intron is the process that is important for RNA synthesis in most eukaryotes and some prokaryotes. The process is called RNA splicing. In the usual circumstances, splicing occurs only between the 5’ and 3’ sites of the same intron. The main components necessary for splicing are the short consensus sequences at the 5’ and 3’ splice sites and the branch site. However, the molecular basis of proper recognition of correct splice site pairs remains unclear.

Here, two questions arise. First, how the pair of right splice sites is recognized properly from a lot of candidate sites? Second, whether the mechanism is conserved in the evolutionary process, i.e. whether similar mechanisms are used in related species? We have investigated these problems by the method of Lim and Burge [2] with an increased number of species. We identified 5’ splice site, 3’ splice site, and branch site for 58 species, and analyzed the properties relevant to splicing from the viewpoint of evolution and information theory.

2 Method

2.1 Data set

We perform the following procedures for all the 58 species. Spaln [1] is used to align Unigene and, if available, additional EST sequences to the genomic sequence of the same species. Spaln can map exon-intron structures quite accurately. Using the alignment results, intron and flanking exon sequences are extracted from the genomic sequence. To obtain only reliable introns, some criteria are applied on the quality of the alignment. Moreover, only representative introns are retained when several exon-intron boundaries share highly similar sequences (> 75% identities).

Our algorithm assigns information (log-odds) score to each of five features (5’ss, 3’ss, branch point, intron length and nucleotide compositions within the intron). We enumerate all possible 5’ and 3’ splice signal pairs that are separated by typical lengths of short introns, and assign them the “splicing score” defined as the sum of information scores of the five properties. We select the pair with the highest splicing score as the predicted intron boundaries. The intron detection accuracy is defined as \( Ac = ( \text{the number of correctly predicted introns}) / (\text{the total number of true introns}) \).

2.2 Evaluation of contribution of each transcript feature

We measure the contribution of each transcript feature to intron detection as follows. First, we examine accuracy of intron recognition with all the five features, and then with all but one. Next, the accuracy is converted into an information value defined as –log(1.0-Ac). After Lim and Burge [2], we regard that intron recognition is precise (100%) if Ac is greater than 0.98. The 58 species examined are classified into four groups according to the relative contributions of the five features plus the fraction of “information lack”. For the classification, we first obtain the first and second components of principal component analysis (PCA) from the six contributions and then apply the k-means method to the same data with the initial conditions inferred from the PCA analysis.
Figure 1: Relative contributions of five transcript features to intron detection.

3 Results

Fig.1 shows the results of the clustering analysis applied to 58 species. Cluster (a) consists of various species. The sum of the contributions of “5’ss” and “3’ss” is approximately 50%. The rest is shared by “branch point”, “length”, and “composition”. There is little “information lack”, suggesting that the five features possess enough information for intron recognition. Cluster (b) is occupied mostly by animals other than vertebrates. There is “information lack” of around 10%. The contributions of “5’ss” and “3’ss” are large, accounting for 60% of the total contribution. Cluster (c) mostly consists of mammals and another vertebrate. The fraction of “information lack” is around 20% being the largest among the four clusters. Like the tendency observed in cluster (b), “5’ss” and “3’ss” highly contribute, whereas the contribution of length is minimal. Most of fungi belong to the last cluster (d). Only one animal is seen in this cluster. The contributions of “5’ss” and “3’ss” are the lowest among the four clusters. On the other hand, the contribution of “branch point” is generally large, maximally reaching about 30%. The contribution of “length” is also large, and the sum of these two contributions often exceeds 50% of the total.

4 Discussion

This study presents a large-scale computational analysis of pre-mRNA splicing, which has become feasible by the availability of sequence data from 58 complete genomes and their transcripts. Some correlation was observed between the relative contributions of the splicing features and the genealogical relationships among the species. However, some species exhibit very different patterns of intron recognition compared with other species in the same taxon. As the five features are not enough for precise intron recognition in many species, there must be other features to be explored.

References