Clustering techniques for classification of splice sites of human exons

Muselli M (1), Romeo F (2), Pfeffer U (2)

(1) Institute of Electronics, Computer and Telecommunication Engineering,
Italian National Research Council, Genova
(2) Functional Genomics, National Cancer Research Institute, Genova

Motivation

The usage of genetic information by the cellular machinery has been greatly facilitated by the evolution of splicing, a process that permits to join fragments of coding sequences (exons) from larger genomic regions containing prevalently non-coding sequences (introns). The process relies on the precision of exon recognition since a single nucleotide shift leads to an alteration of the reading frame and hence to altered information. The analysis of the splice sites has led to the identification of the consensus sequences GURAGU at the exon-intron border and AXnYnAG (branch point, polypyrimidine tract, AG) for the intron-exon border. However, these consensi are weak: many real sequences considerably divert from them, and a great number of sequences matching the consensus patterns are not used as splice sites. We follow the hypothesis that additional sequence features that contribute to the splice site definition can be identified if specific classes of such sites are considered. Yet there is no objective criterion to classify exons. We therefore set out to classify splice sites using machine learning approaches. The classification of splice sites can be used for analyses of correlation with the biological behavior of the relative exons such as alternative splicing. This approach is of foremost importance given the introduction of whole genome exon analyses by microarray hybridization. Class specific additional sequence features may yield new information on functional single nucleotide polymorphisms and somatic mutations that, without this information, would be considered as silent.

Methods

The classification of splice sites is performed by analyzing sequences of n bases around the transition points between exons and introns. In particular, the target of the analysis is to retrieve a specific characterization that distinguishes sequences including a splice site (denoted as positive sequences) from others that do not contain it (negative sequences). Any machine learning technique for classification can be adopted to deal with this problem; however, rule generation methods are to be preferred, since they are able to produce sequences in IUB code that detect the presence of the splicing site. Shadow Clustering (SC) [1,2] is a rule generation method, based on monotone Boolean function reconstruction, which is able to achieve performances comparable to those of best machine learning techniques. SC proceeds by grouping together binary strings that belong to the same class and are close to each other according to a proper definition of distance. Since SC operates on binary strings, every sequence of bases must be previously converted in Boolean form, before the generation of the set of rules starts. To this aim, the standard basis conversion: 'A' = '0111', 'C' = '1011', 'G' = '1101', 'T' = '1110' is employed. This gives rise to a training set for SC containing binary strings with length 4n. If a huge collection of negative sequences is included in the training set, the execution of SC generates a high number of rules, many of which are obtained through specializations of a general consensus pattern. To determine these relationships a proper hierarchical clustering technique is adopted; it can be viewed as a modification of the single linkage algorithm, which takes into account the presence of an ordering among the elements to be clustered, given by the relevance associated with each rule.

Results

DNA sequences extracted from the human genome have been considered for the classification of splice sites. Two different training sets have been taken into account: the first one concerns the detection of exon-intron (EI) transition points, whereas the second one is related to intron-exon (IE) sites. In both cases sequences containing n= 120 bases around the splice sites have been selected to be included in the training set as positive examples. In particular, by using the dataset of Clark and Tanaraj [3] a collection of 14,026 and 14,309 positive sequences have been generated for the EI and

the IE problem, respectively. The negative examples for the training set in both situations have been retrieved by analyzing a 300kB fragment of human genomic DNA extracted from the region on chromosome 6 containing the estrogen receptor ?, a gene that shows extensive alternative splicing [4,5] . In this way, the whole training set for the two problems contains more than 310,000 examples. The execution of SC has produced 2,618 and 3,064 rules, written as IUB code sequences, in the EI and in the IE problem, respectively. Most of them presents consensus patterns different from the standard GURAGU. The application of the hierarchical clustering approach, developed for this particular analysis, has recognized about 25 clusters for each problem, which can be associated with specific non standard consensus patterns that can be further examined to improve the understanding of the phenomena involved in the splicing mechanism.

Contact email: marco.muselli@ieiit.cnr.it

References

1. M. MUSELLI, A. QUARATI Reconstructing positive Boolean functions with Shadow Clustering. In Proceedings of the 17th European Conference on Circuit Theory and Design (ECCTD 2005), (Cork, Ireland, August 2005).
2. M. MUSELLI Switching neural networks: A new connectionist model for classification. In WIRN/NAIS 2005, vol. 3931 of Lecture Notes in Computer Science (2006) Eds. B. Apolloni, M. Marinaro, G. Nicosia, R. Tagliaferri, Berlin: Springer-Verlag, 23-30.
3. F. CLARK, T. A. THANARAJ Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Hum Mol Genet, 11 (2002) 451-464.
4. U. Pfeffer, E. Fecarotta and G. Vidali. Coexpression of multiple estrogen receptor variant messenger RNAs in normal and neoplastic breast tissues and in MCF-7 cells. Cancer Res., 55, 2158-2165, 1995.
5. Ferro, P., Forlani, A., Muselli, M. and Pfeffer, U. Alternative splicing of the human estrogen receptor \alpha primary transcript: mechanisms of exon skipping. Int. J. Mol. Med., 12, 355-363, 2003.

Supplementary informations

Acknowledgment This work was supported by the "Ministero dell'Istruzione, dell'Università e della Ricerca" MIUR projects "Laboratory of Interdisciplinary Technologies in Bioinformatics (LITBIO)" and "Hormone Responsive Breast Cancer"