Finding Characteristic Sequence Patterns for Alternative Splicing in Human Genomic DNA

Hiroki Sakai[1] hiroki@ims.u-tokyo.ac.jp
Osamu Maruyama[2] om@math.kyushu-u.ac.jp
Hiroo Murakami[3] hiroo@ims.u-tokyo.ac.jp
Nobuyoshi Sugaya[3] sugaya@ims.u-tokyo.ac.jp
Akira Imaizumi[3] akima@ims.u-tokyo.ac.jp
Makihiko Sato[4] maki@maebashi-it.ac.jp
Sachiyo Aburatani[3] sachiyo@ims.u-tokyo.ac.jp
Katsuhisa Horimoto[3] khorimot@ims.u-tokyo.ac.jp
Minoru Kanehisa[1] kanehisa@kuicr.kyoto-u.ac.jp

[1] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto, 611-0011, Japan
[2] Faculty of Mathematics, Kyushu University, 6-10-1 Hakozaki, Higashi-ku, Fukuoka, 812-8581, Japan
[3] Laboratory of Biostatistics, Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokane-dai, Minato-ku, Tokyo 108-8639, Japan
[4] Department of Systems and Information Engineering, Graduate School, Maebashi Institute of Technology, 460-1 Kamisatori-Cho, Maebashi-City, 371-0816, Japan

Keywords: alternative splicing, pattern design

1 Introduction

Alternative pre-mRNA splicing events can be classi ed into various types, including cassette, mutually exclusive, alternative 3' splice site, alternative 5' splice site, retained intron [1]. The detection of regulatory sequence elements closely related to such a particular type of alternative splicing events is an important and challenging problem in understanding the mechanism of alternative splicing. However, it seems that it has not been given enough extensive computational analysis of examining whether there are candidate regulatory sequence elements characterizing types of alternative splicing events. In this work, we consider the problem of nding signi ca nt patterns speci c to the types of alternative 5' splice site, alternative 3' splice site, and cassette, respectively, from their alternative exons and ankin g introns.

2 Materials and Methods

Lee *et al.* [2] have compiled information related to alternative splicing, and the results are available as an online database ASAP (Alternative Splicing Annotation Project). The text les of this database can be downloaded at the site, `http://www.bioinformatics.ucla.edu/HASDB/`. An entry of the database has a column indicating how much evidence we have for the alternative splicing event. The value "multiple" means that both splices have at least two ESTs or at least one mRNA observation. All other alternative splices are indicated by "single". The entries we use here are restricted to the ones where their evidences are labeled by "multiple".

Through our computational experiments, all the alternative exons involved in the alternative splicing type in question, for example, the type of alternative 5' splice site, are used as positive examples. On the other hand, all the constitutive exons are considered to be negative examples. The sequences related to those exons are called negative and positive sequences, respectively. For two alternative exons e_1 and e_2 of alternative 5' splice sites, four kinds of search regions are de ned in Fig.1. The length l of the upstream and downstream regions is set at 100 nt, which is the same as the previous work[3]. In the same way, the four kinds of search regions for

alternative exons of alternative 3' splice sites are de ned (see Fig.2). For cassette and constitutive exons e, the upstream, exonic and downstream regions are de ned (see Fig.3).

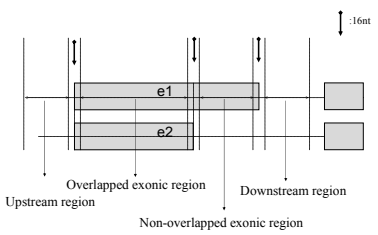

Figure 1: alternative 5'

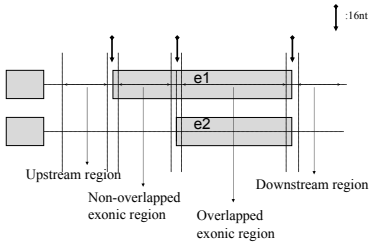

Figure 2: alternative 3'

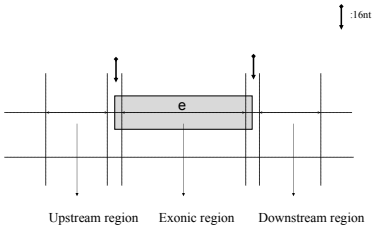

Figure 3: cassette

For pattern models are considered in this work :mismatch patterns, degenerate patterns, alphabet-indexed substring pattern, numerical indexing patterns. The returned values of the patterns are set to be binary. Thus, the conjunction (i.e., logical product) and disjunction (logical sum) of any types of patterns are de ned and can be calculated. We here describe a score function F of patterns, whose value is called a contrast score. By $p(T)$ we denote the number of the strings t in T such that there is at least one occurrence of p in t. The contrast score function F returns the a score $p(P)/|P| - p(N)/|N|$, given a positive sequence set P and a negative sequence set N.

3 Results and Discussion

Through our experiment, we have succeeded in nding discriminative features with practically high accuracies. An interesting point is that the two search regions of the conjunctions and disjunctions of two types of patterns with high contrast scores are the pair of upstream region and downstream region. As for the alphabet-indexed substring pattern, high score patterns share the same alphabet indexing which separates A and T from C and G. Furthermore, this fact is not dependent on the types of alternative splicing. Other results will be reported in the meeting.

Table 1: Example of result(alternative 3').

Region	*Patternmodel*	*Pattern*	F	$p(P)/\|P\|$	$p(Q)/\|Q\|$
Downstream	Degenerate	[CG][AG]GGG	20.34	61.04	40.69
		CCC[AC][CG]	19.93	56.62	36.69
		[CG][AC]GGG	19.72	58.63	38.90

References

[1] T.A. Thanaraj and S. Stamm. *Prediction and Statistical Analysis of Alternatively Spliced Exons*, pages 1–31. Progress in Molecular and Subcellular Biology **31**. Springer-Verlag, 2003.

[2] C. Lee, L. Atanelov, B. Modrek, and Y. Xing. ASAP: the alternative splicing annotation project. *Nucleic Acids Research*, 31:101–105, 2003.

[3] M. Brudno, M.S. Gelfand, S. Spengler, M. Zorn, I. Dubchak, and J. G. Conboy. Computational analysis of candidate intron regulatory elements for tissue-speci c alternative pre-mRNA splicing. *Nucl. Acids. Res.*, 29:2338–2348, 2001.