An Aberrant Splicing Database for Finding Rules of Splice-Site Selection

Takashi Yamanaka [1] **Tetsushi Yada** [2] **Kenta Nakai** [3]

yamanaka@ims.u-tokyo.ac.jp yada@ims.u-tokyo.ac.jp knakai@ims.u-tokyo.ac.jp

[1] Institute for Molecular and Cellular Biology, Osaka University, c/o Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[2] RIKEN Genomic Sciences Center, c/o Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

[3] Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

The gene finding problem, that is, finding coding sequences from a genomic sequence, is one of the most fundamental problems in bioinformatics. The problem becomes especially difficult in eukaryotic genomes where coding sequences are divided into pieces by introns. Therefore, it is very important to know how splice sites are selected in cells. It has been known that there exist consensus sequences around both 5' splice sites and 3' splice sites. However, there are great many similar sequences that are not selected as true splice sites (false positives). Moreover, some of these false positive sites are activated as so-called cryptic sites when authentic splice sites are destroyed by mutations. Therefore, it is evident that the selection of splice sites is not entirely determined from local information such as consensus sequences. To further understand more global rules of splice-site selection, Nakai and Sakamoto [5] constructed a database on aberrant splicing where mutations that caused abnormal type of RNA splicing were detected from genetic diseases. Although some interesting observations were made from the database, the total number of accumulated data was not enough to make reliable statistical analyses. Furthermore, the mutation data were not fully examined within their sequence contexts because they were not linked to sequence databases such as GenBank [1]. In this study, we updated our database which now contains more than 1000 mutations. The data are linked to sequence data if possible. We also report our sequence analyses using them.

2 Overview of Aberrant Splicing Database

Since the first version of our database has been released, there have been released some databases on mutations. The Human Gene Mutation Database (HGMD) is one of the most useful examples among them [4]. Although HGMD contains extensive information on mutated sites for aberrant splicing, it does not provide the information on caused splicing patterns. Therefore, we constructed our database by reviewing original references cited in HGMD. Our database now contains 245 genes and 1046 mutations, most of which are human origin. The gene names were taken from the list by the Human Gene Nomenclature Committee (`http://www.gene.ucl.ac.uk/nomenclature`). Like its original version, aberrant splicing patterns are classified into four types: exon skipping, activation of cryptic sites, intron retention, and creation of new splice sites. Minor variants of aberrant splicing are also contained when available. Each data is linked to corresponding entries in GenBank and PubMed. Our database will be released through WWW in the near future.

3 Observations and Discussion

From our database, we observed the following points which seem to be useful to get some insights on the molecular mechanism of RNA splicing:

1. About 80% of the mutations occurred within the so-called consensus regions around the splice sites and their frequency roughly corresponds to the conservation degree for each position in authentic splice sites. For future works, it will be interesting to examine the nature of the remaining 20% of mutations which may be related to some *cis* elements known as splicing enhancers or silencers.

2. In the four types of aberrant splicing patterns, exon skipping was the most frequently observed (about 60%). We also observed that the length of skipped exons are longer than the exons where cryptic sites are used on average. They appear to be consistent with the idea of exon recognition proposed by S. M. Berget [2]. In addition, 5' site mutations are more common than 3' site mutations in any patterns.

3. When a cryptic 3' site relatively near an original site is activated, it is selected from the downstream region. We also observed that newly-created (5' and 3') splice sites are observed in the upstream region from their authentic sites in most cases. Both of these observations are consistent with a hypothesis that 3' splice sites (and possibly 5' splice sites) are more or less scanned from upstream to downstream.

4. The type of caused aberrant splicing patterns is uniquely determined by the information of destroyed splice sites and is not dependent on the details of mutations (such as the information on whether a mutation is occurred at +1 or -1 position). This observation allows us to formulate a new prediction problem: *how a splicing pattern will be changed when a given splice site is destroyed?* As a first step, we used the data of 50 genes (220 splice sites) the sequence of which have been determined through their (almost) entire transcripts. A famous gene-finding program, GenScan [3], could not faithfully predict the outcome of these mutations (only about 30% were correct). We are developing an HMM model for the better prediction of such cases.

Acknowledgments

This work is supported by a Grant-in-Aid “Genome Science” for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Benson, D.A., Boguski, M.S., Lipman, D.J., Ostell, J., Ouellette, F., Rapp, B.A., and Wheeler, D.L., GenBank, *Nucl. Acids Res.*, 27:12–17, 1999

[2] Berget, S.M., Exon recognition in vertebrate splicing, *J. Biol. Chem.*, 270:2411–2414, 1995

[3] Burge, C. and Karlin, S., Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, 268:78–94, 1997

[4] Krawczak, M. and Cooper, D.N., The Human Gene Mutation Database, *Trends Genet.*, 13:121–122, 1997

[5] Nakai, K. and Sakamoto, H., Construction of a novel database containing aberrant splicing mutations of mammalian genes, *Gene*, 141:171–177, 1994