Upgraded Aberrant Splicing Database Supports the Scanning Mechanism of 3' Splice-Site Selection

Takashi Yamanaka[1]
yamanaka@ims.u-tokyo.ac.jp

Tetsushi Yada[2]
yada@gsc.riken.go.jp

Toshihisa Takagi[3]
takagi@ims.u-tokyo.ac.jp

Kenta Nakai[3]
knakai@ims.u-tokyo.ac.jp

[1] Institute for Molecular and Cellular Biology, Osaka University,
1-3 Yamada-oka, Suita, Osaka 565-0879, Japan

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

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

Keywords: aberrant splicing, database, point mutation, scanning model

1 Introduction

RNA splicing is an essential step of gene expression for eukaryotes. If it occurs inaccurately in a gene, an abnormal mRNA, and then, an abnormal protein, will be produced. Usually, this will cause some deficiency in the organism. Thus, there must be certain mechanisms ensuring the correct selection of splice sites. Although it is well-known that the consensus sequences around the splice sites are necessary, they are not sufficient; *i.e.,* there are many other spurious consensus-like sequences. Moreover, when a splice site is destroyed by a mutation, another cryptic site is often selected, instead. Thus, it is evident that a splice site is not determined by its local sequence information only but there are some contextual rules. One way of finding them is to collect the examples of aberrant splicing, namely, abnormal RNA splicing caused by a mutation and revealed as a disease. By examining the position of such mutations, one will be able to find which positions are important for specifying splice sites. Moreover, such a compilation can be used for understanding subtle balances on whether a potential site will be selected or not. With this in mind, Nakai and Sakamoto constructed such a database containing 90 genes and 209 mutations [5]. Although they reported several observations that support the exon recognition hypothesis [1], the amount of data was not enough for further detailed analyses. For example, sequence analyses of aberrant splicing were not tried because there were so few completely sequenced genes at that time. In this work, we significantly increased the data size of the aberrant splicing database and performed several analyses based on it.

2 Database

Since the first version of our database was constructed, many databases on human mutations have been released through the Internet. Amongst them, the Human Gene Mutation Database (HGMD) contains 20,000 mutations on January, 2000 [3]. It also involves a comprehensive collection of mutations that affect RNA splicing. We used it for updating our database. However, we had to consult all original papers since HGMD only provides the information only on where mutations occurred but not on how the splicing patterns were changed. Now, our database contains 245 genes and 1042 mutations based on 631 references. Only mammalian mutations are collected and most of them are human origin. For

704 mutations, we could find the information on the change of RNA splicing pattern. Moreover, we examined the corresponding GenBank entries of them; 51 genes containing 224 mutations are found to be (almost) completely sequenced and thus can be used for further sequence analyses. Our database will be released through the Internet in the near future.

3 Results and Discussion

In most cases, previously-reported tendencies remain unchanged in this updating; the ratio between 5' and 3' splice-site mutations is 64:36; 85% of mutations occur on the consensus regions; aberrant splicing patterns can be classified into four types: the exon skipping, the cryptic-site activation, the new-site creation, and the intron retention. Two novel observations were made. First, cryptic 3' splice-sites are usually selected from the downstream of the authentic sites; 42 out of 55 (75%) sites were on the downstream while the figure was 55% for 5' splice-sites. More specifically, most of the 3' cryptic sites were located within the downstream 25 bp region. Second, new 3' splice-sites are created within the 25 bp region upstream from the authentic sites in 14 out of 21 cases while 4 out of 13 cases were found on the corresponding region for new 5' sites. Both of these observations strongly support the "scanning" hypothesis proposed by Mount [4]; namely, 3' splice sites are first recognized at the branch point and an appropriate 'AG' is searched unidirectionally on the downstream side from there.

We also performed some sequence analyses. The strength of authentic sites, cryptic sites, and nearby other candidates were evaluated based on the method of Shapiro and Senapathy [6]. In many cases, authentic sites showed the highest score and then the second highest site within the neighboring 50 bp regions on both directions were selected as a cryptic site. However, there remain a significant number of cases where cryptic sites were selected in spite of the presence of apparently more appropriate sites. It is likely not due to the incompleteness of the score-calculation method but there seem to lie other reasons. Further examinations of such cases may reveal some rules of splice-site selection. In addition, it will become an interesting project to predict the result of a mutation on a given splice site. Finally, we confirmed that with a famous gene-finding program, GENSCAN, the original exon/intron boundaries could be predicted rather accurately (about 80 %). However, only about 30 % of the aberrant splicing pattern could be predicted. It clearly reflects the fact that such a program heavily relies on the coding information. Our data will become a good benchmark dataset for the development of next-generation gene-finding programs.

References

[1] Berget, S.M., Exon recognition in vertebrate splicing, *J. Biol. Chem.*, 270(6):2411–2414, 1995.

[2] Burge, C. and Karlin, S., Prediction of complete gene structures in human genomic DNA, *J. Mol. Biol.*, 268(1):78–94, 1997.

[3] Krawczak, M., Ball, E.V., Fenton, I., Stenson, P.D., Abeysinghe, S., Thomas, N., and Cooper, D.N., Human gene mutation database—A biomedical information and research resource, *Hum. Mutat.*, 15(1):45–51, 2000.

[4] Mount, S.M., A catalogue of splice junction sequences, *Nucleic Acids Res.*, 10(2):459–472, 1982.

[5] Nakai, K. and Sakamoto, H., Construction of a novel database containing aberrant splicing mutations of mammalian genes, *Gene*, 141(2):171–177, 1994.

[6] Shapiro, M.B. and Senapathy, P., RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression, *Nucleic Acids Res.*, 15(17):7155–7174, 1987.