

Functional organization of transcript sequences as gene expression data

Hidemasa Bono¹
bono@dbcls.rois.ac.jp

Shoko Kawamoto¹
shoko@dbcls.rois.ac.jp

Kousaku Okubo²
kousaku@genomatrix.com

Toshihisa Takagi¹
takagi@dbcls.rois.ac.jp

¹ Database Center for Life Science (DBCLS), Research Organization of Information and Systems, 2-11-16 Yayoi, Bunkyo-ku, Tokyo 113-0032, Japan

² National Institute of Genetics, Research Organization of Information and Systems, Yata 1111 Mishima, Shizuoka 411-8540, Japan

Abstract

In order to make use of transcript sequences as gene expression data, we have developed repository and analysis pipeline for those produced in Japan.

Keywords: gene expression, database, EST, cDNA, pipeline

1 Introduction

Up to now, hundreds of biological databases and various computational tools to make use of these have been developed. It is now not so easy to keep track of those including biological papers even in a specific field, and the integration of those is crucial for life science. Considering these situations, the Ministry of Education, Culture, Sports, Science and Technology (MEXT) in Japan launched Integrated Database Project in 2006 [4] and Database Center for Life Science (DBCLS) was founded in 2007 to develop the integrated database for life science in Japan [3].

In Japan, sequencing projects for transcripts from various model organisms were launched, while most of the products from these, complementary DNA (cDNA) sequence data, have not been uniformly annotated and some of these are not even included in public DNA databank for redundant feature of cDNA sequence data. These can be invaluable resource for experimental design for molecular biologists especially for comparative genomic studies when sequence information for these is put together in the central archive.

2 Method and Results

Sequence information from transcripts can be treated as gene expression data [2]. It is because too many transcripts for various organisms have been sequenced. In order to make use of cDNA sequence data as gene expression data, we have developed repository and analysis pipeline for those produced in Japan (Fig.1). Sequences in the International Nucleotide Sequence Database (INSD: DDBJ/EMBL/GenBank) are currently taken for this pilot study, but the system is designed to accept sequence records not in INSD.

For genome-sequenced organisms, groups of cDNA sequences are mapped to genomes utilizing the analysis pipeline developed, and the mapping data are served in DAS server maintained at DBCLS. Genome mapping also includes synthetic probe sequences for microarray experiments and conceptual transcription factor binding sites. This is natural extension of SayaMatcher pipeline that locates short but meaningful DNA elements in a genome scale [1,5]. The data formatted in expression matrix is also available for visualization in heat map that is often used for microarray data.

In addition to genome mapping, the pipeline includes the visualization of statistical property of data. For example, the extraction of upstream sequences in the genome for corresponding transcripts enables

sequence analyses of regulatory elements for unknown transcription factor. Tentative visualization of those sequence analyses is available as one of output of the pipeline developed.

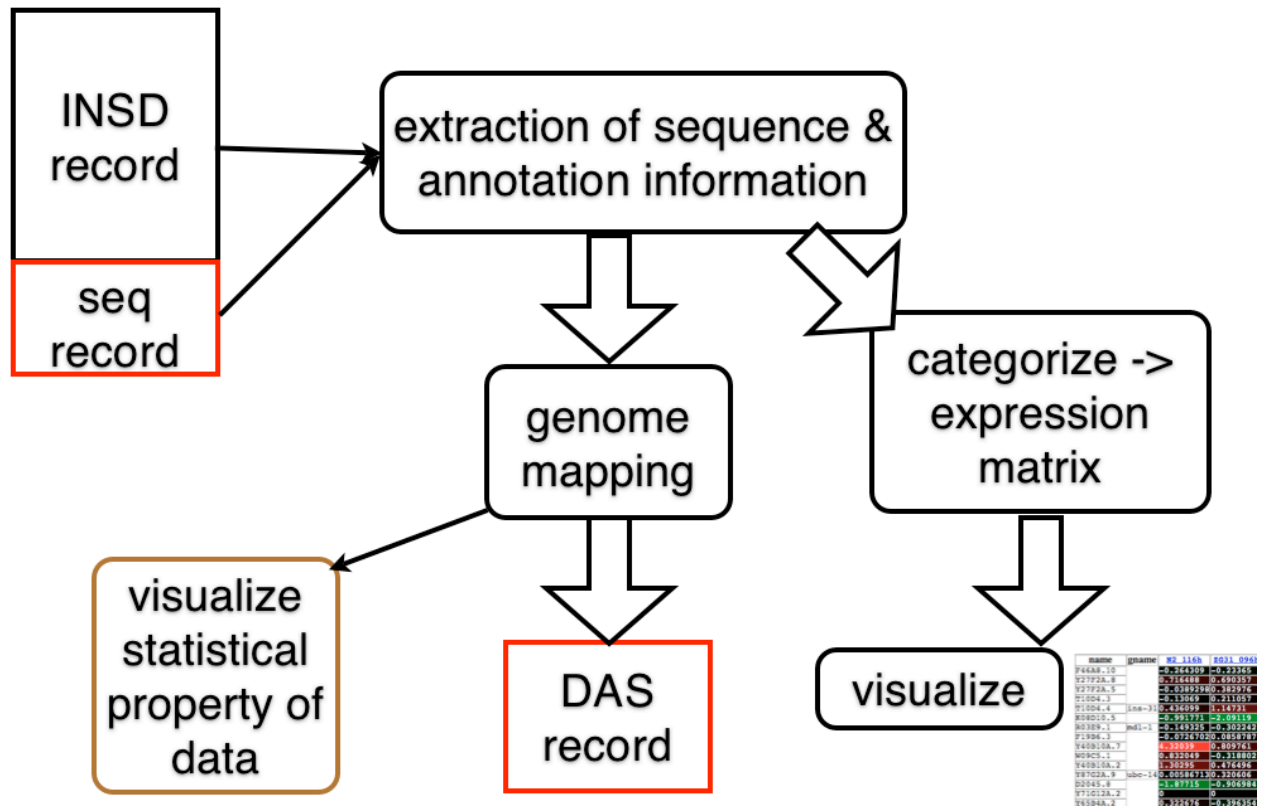


Figure 1: Pipeline for transcript sequence developed.

3 Discussions

The analysis pipeline developed can be applied for the analysis of short read sequences (RNA-Seq) from the next generation sequencers expected to be accumulated in coming few years. We believe that the analysis pipeline will accelerate biological interpretation of sequence data as gene expression data.

References

- [1] Bono, H.U., SayaMatcher: genome scale organization and systematic analysis of nuclear receptor response elements, *Gene*, 364:74-78, 2005.
- [2] Okubo, K., Hori, N., Matoba, R., Niiyama, T., Fukushima, A., Kojima, Y., Matsubara, K., Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression, *Nat Genet.* 2(3):173-179, 1992.
- [3] <http://dbcls.rois.ac.jp/en/>
- [4] <http://lifesciencedb.jp/en/>
- [5] <http://sayamatcher.sourceforge.jp/>