NEXTDB: The Expression Pattern Map Database for *C.elegans*

Tadasu Shin-i [1] **Yuji Kohara** [2]
tshini@genes.nig.ac.jp ykohara@lab.nig.ac.jp
[1] CREST, JST
[2] Gene Network Laboratory, National Institute of Genetics, 1111 Yata Mishima 411, Japan

Abstract

We have developed a WWW-based database, named NEXTDB, to integrate all the information of ESTs (tag-sequences of cDNA clones) and gene expression patterns of C.elegans which are being produced and analyzed in this laboratory. NEXTDB incorporates and processes raw data of tag sequencing and classifies them into unique cDNA groups by comparing the 3'-tags. The database contains the information on map position of the cDNA groups, correspondence to predicted CDSs and homologies to other organisms' genes. NEXTDB incorporates image data of in situ hybridization which show the expression patterns of individual cDNA groups and provides us a platform for annotation of the images. The database also contains the cosmid contig maps obtained from AceDB . All of the information are linked each other in NEXTDB, which can be accessed through the internet.

1 Introduction

The nematode *Caenorhabiditis elegans* (*C.elegans*) is a good model system to study functional genomics with respect to animal development, nervous system and behavior at the level of single cells. *C.elegans* has only about 1,000 somatic cells, however, it has the basic structure of animals such as epidermis, muscle, nerve system, digestive organs, reproductive system and so on. This simplicity has led to the description of entire cell lineage from embryo to adult, which has allowed us to study gene functions in individual cells [1]. The genome consists of six chromosomes whose total size is about 100 Mbp and total number of genes is estimated to be about 15,000. More than 70Mb of the genome has been already sequenced by the consortium of the Sanger Centre and Washington University, and the sequencing will be completed by the end of 1998 [3].

In this laboratory, we are pursuing systematic analysis of cDNA clones of this organism with respect to tag-sequences, map positions, pattern of expression during development and gene functions. Expression pattern information has been obtained with nearly 1000 genes (Kohara et al.; in preparation).

In this report we describe an overview of the current version of the database NEXTDB and future plans to develop new functions of the database.

2 Materials and methods

2.1 ESTs

One-pass sequencing from both ends on random cDNA clones has been performed on three ABI377 sequencers in this laboratory. NEXTDB incorporates the raw sequence data and processes them to extract clean tag sequences under a criterion with respect to the rate of ambiguous base "N". The processed "clean" 3'-tag sequences were used to classify them into unique cDNA groups applying a cumulative method using FASTA, since the 3'-tags should be unique among the cDNAs derived from the same gene while the 5'-tags may be various depending on the extent of reverse transcription in the step of cDNA synthesis. Both 5'- and 3'-tag sequences were compared by BLASTN with the cosmid sequences and their predicted CDS which had been obtained from the Sanger Centre.

We also performed BLASTX search for a non-redundant database of amino acid sequences, since all the tags did not hit predicted CDS.

2.2 Expression patterns

Briefly, whole mount in situ hybridizations [2] are being performed using a set of representative clones of the unique cDNA groups. Images of the hybridization results are taken at 3 different focal planes by CCD cameras equipped on Zeiss Axioplan2 microscopes having Nomarski optics and processed manually and/or automatically.

The image data are transferred to NEXTDB and operators annotate individual images with respect to the information of developmental stages (currently, operators select from the following 10 landmark stages; 2 cell, 4 cell, 6 to 18 cell, early gastrulation, mid gastrulation, late gastrulation, comma, 1.5 fold, 2 fold, and 3 fold stages) and expressing cells/tissues. Once annotated, the images are arranged properly along development on the screen.

2.3 Links to the genome map

In order to link NEXTDB with the genome map, we applied a hierarchical model to arrange all the clones and clusters;1) chromosome, 2) cosmid clone, 3) CDS, 4) cDNA group, and 5) cDNA clone. The cosmid map data which connects 1) and 2) were obtained from the C.elegans genome database AceDB which describe the relationships among cosmid clones, predicted genes or CDS and genetically defined genes. The information about cosmids and their CDS was retrieved from the annotations of the Sanger Centre sequence data. Sequences and homologies of the probe cDNA and all the in-situ images were arranged by making links to corresponding cDNA clones. All the data are integrated based on WWW, and the information of maps and their relations are depicted visually by use of JAVA applets.

3 Result and discussion

Thus far, tag sequences of about 40,000 cDNA clones have been incorporated into NEXTDB. The database classifies the clones into about 7,500 unique cDNA groups, which correspond to a half of the total gene number. About 3,500 groups hit to the predicted CDSs in the genomic sequences, and about 1,500 hit to the non CDS region. Comparing some ESTs and the corresponding genomic sequences has revealed the presence of alternative splicings and differential termination, and sometimes bridging cosmids at gaps of cosmid contigs. CDS prediction has not been done with 30% of the sequenced cosmids. Therefore, close comparison of ESTs and the genomic sequences is very important to identify genes precisely. This task is being done by both this laboratory and the Sanger Centre. NEXTDB incorporates in situ images during embryogenesis of about 500 cDNA groups mostly from chromosome 3. The latest version is available through the following URL:

http://watson.genes.nig.ac.jp:8080/db/index.html

It can also be accessed through the home page of DDBJ (http://www.ddbj.nig.ac.jp). You can see the visual description of relationships among a cosmid, predicted genes, and clusters of cDNA clones, and you can retrieve expression pattern images which are arranged along the genome map.

Finally, the database is also regard as a prototype of laboratory automation system. For example, It can arrange a large amount of tag sequences and relevant information efficiently and flexibly, in which tag clustering modules and product prediction module work automatically. Since all the modules are integrated based on WWW, the database allows annotators to edit and manage the database from remote terminals. Once new data are installed or existing data are modified, the lab members can share the latest data immediately.

Acknowledgments

This work has been supported by CREST of JST (Japan Science and Technology Corporation) and by a Grant-in-aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports and Culture of Japan.

References

[1] Sulston, J. E., Schierenberg, E., White, E. J., and Thomson, J. N., *Dev. Biol.*, 100:64-119, 1983.

[2] Tabara, H., Motohashi, T., and Kohara, Y., *Nucl. Acids Res.*, 24:2119-2124, 1996.

[3] Waterston, R. and Sulston, J., *Proc. Natl. Acad. Sci. USA*, 92:10836-10840, 1995.