Database and Analysis System of cDNA Sequences Obtained from Full-Length Enriched cDNA Library

Tetsuo Nishikawa [1]
nisikawa@hri.co.jp

Toshio Ota [1]
ota@hri.co.jp

Sumio Sugano [2]
ssugano@ims.u-tokyo.ac.jp

Asaf Salamov [1]
asaf@hri.co.jp

Yoshitaka Nakamura [1]

Tateo Nagai [1]

Atsushi Oogane [1]

Yutaka Suzuki [2]

Yuri Kawai [1]

Shizuko Ishii [1]

Kaoru Saito [1]

Jun-ichi Yamamoto [1]

Takao Isogai [1]
isogai@hri.co.jp

[1] Helix Research Institute
1532-3 Yana, Kisarazu-shi, Chiba 292, Japan

[2] Institute of Medical Science, University of Tokyo
4-6-1 Shirokanedai, Minato-ku, Tokyo 108, Japan

Abstract

We are developing an efficient sequence analysis system and a database system for clones from full-length enriched cDNA libraries using oligo-capping method [1]. Sequences of 1) 5'-end and 3'-end and of 2) full-length, obtained from full-length enriched cDNA clones, respectively, are dealt in this system. We have developed a semi-automatic analysis system for 5'-end and 3'end sequences. This consists of pre-processing of the raw-sequences, grouping of sequences, similarity searching against public databases, and analysis of ORFs in the sequences. In each step, newly developed or improved tools are used. All analyzed data are registered in our database and can be retrieved with the analysis results as queries. This is the first cDNA database system containing sequences from full-length enriched cDNA library in large scale.

1 Introduction

Several EST projects have already produced cDNA fragments of more than half of human genes. However, ESTs are insufficient in 5'-region of full length cDNA sequence, because clones obtained by conventional methods lack their 5'-regions of mRNA specifically. They often lack the initiation codons for their translation. To analyze function of genes, it is important to obtain the clones including the intact coding sequences. Maruyama & Sugano developed an excellent method [1], "oligo-capping method", to obtain the intact 5'-ends of mRNA efficiently. We, therefore, have started to collect a bulk of full length human cDNA clones [2] by using the oligo-capping method to develop a high throughput functional gene-analysis system. We aimed at developing a semi-automatic sequence analysis system and database system which is optimized for dealing with these full length cDNA clones.

2 Analysis system of cDNA data

Analysis system consists of pre-processing of the raw-sequences, clustering of sequences, similarity searching against public databases, and analysis of ORFs in the sequences. The pre-processing system have the steps of automatical recognition of vector-sequences and low-accuracy regions with an user-interface for editing. The vector sequences were recognized by using dynamic programming method

and the low-accuracy regions were identified based on the local N-letter rates. The user interface was designed to show and edit easily the recognition-information of the vector and low-accuracy region. The clustering of sequences was performed using all-by-all blast2 comparisons of the 5'-sequences. The clustering parameters were optimized using the parameter-dependency of the number of groups.

The similarity searching was performed against GenBank and Swiss-Prot. We classified the sequences based on the searching results, and set similarity-flags for each sequence. We also set a flag for each sequence which evaluates the sequence-fullness based on the comparison with EST sequences. To judge the correctness of ORFs, we developed a prediction tool for initiation ATG. For predicted ORFs, analyses such as signal peptide prediction and motif search are performed. Signal peptide prediction could be possible with high accuracy by the combination of full-length cDNA library and the prediction tools for the sequence-fullness and for the initiation ATG. We are now developing a system which assembles our sequences with dbEST sequences.

3 Database system of full-length cDNA

All analyzed data were registered in the database. We have developed a new retrieval system for the analyzed data. This system enables searching the data with the combination of parameters concerning to analysis results as a query; these parameters are, for example, alignment information with the searched sequences, GenBank contents of the searched sequences, the similarity and fullness flags, psort analysis results, and so on. This system can also show several kinds of information at the same time on the divided windows. Interesting clones can be easily retrieved by using this searching system.

4 Evaluation of full-length library

We have entried totally 8,816 5'- and 4,239 3'-sequences from four full-length enriched libraries and 2235 5'-sequences from single 5'-tag library (not containing 3' ends of mRNAs) into the database. We evaluated the rate of full clones in these libraries by comparing 5' sequences with known complete mRNA sequences. Average rates of fullness were 66% in full-length enriched cDNA libraries and 95% in 5'-tag library, respectively. The comparison with EST sequences showed that our 5'sequences often have longer 5' end than those of 5'-EST sequences.

5 Acknowledgments

We would like to thank Dr. Nakai at Osaka University for letting us use the Psort program.

References

[1] Maruyama, K., Sugano, S. "Oligo-capping:a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides" *Gene*, 138:171–174, 1994.

[2] Ota, T., et al. "Full-Length cDNA Project toward a High Throughput Functional Analysis" *9th International Genome Sequencing and Analysis Conference*, 204, 1997.