A Refinement System for Large Amount of cDNA Data

Haretsugu Hishigaki [1 2] hisigaki@ims.u-tokyo.ac.jp
Kagehiko Kitano [1] kitano@ims.u-tokyo.ac.jp
Yusuke Nakamura [1] yusuke@ims.u-tokyo.ac.jp
Hiroumi Maekawa [2] maekawa@otsuka.genome.ad.jp
Toshihisa Takagi [1] takagi@ims.u-tokyo.ac.jp

[1] Human Genome Center,
Institute of Medical Science, The University of Tokyo,
4-6-1 Shirokanedai, Minatoku, Tokyo 108 Japan

[2] Otsuka GEN Research Institute, OTSUKA Pharmaceutical Co., Ltd.,
463-10 Kagasuno, Kawauchi-cho, Tokushima 771-01 Japan

Abstract

As a part of the Human Genome Project, large scale sequencing of cDNA clones from various tissues have been performed and many cDNA sequences have been stored in the public databases. From a lot of sequence data, to obtain more useful biological information, it is indispensable to refine and classify them [1]. We developed a prototype system for refinement and classification of many sequence data.

1 Introduction

Large scale cDNA sequencing is an efficient way not only to understand all gene structures in genome, but also to get information about the gene expressions in various tissues.

After being sequenced, those sequences are compared with published sequences in EMBL or GenBank to predict their functions. But sequences stored in each laboratory include the following problems: (1)redundancy of sequence data, (2)uncertainty of letters in sequences, (3)existence of very short sequences, (4)inclusion of junk sequence fragments such as ALU, L1, and so forth, and (5)inclusion of partial vector sequences. Especially, existence of data redundancy, junk sequences and vector fragments reduce the accuracy and sensitivity of the homology search as well as the efficiency of these data analyses. Thus it is important to solve the above problems. We developed a prototype system for refinement and classification of many cDNA sequences.

2 System and Method

We developed the system with Perl (version 4), and the system is running on a Sun workstation with SunOS 5.X. This system allows the Pearson/FastA format for nucleotide sequences.
Our system refines sequence data with the following steps:

1. conversion from unformatted sequences, that is, raw sequences (from a sequence analyzer) to sequences in the Pearson/FastA format,

2. conversion from 3' DNA sequences to 5' ones,
3. striking out simple nucleotide repeat fragments (3' polyA signal) and dinucleotide repeat sequences ('CA', 'AT', and so on) occured at the ends of sequences,
4. detection and removal of vector fragments in sequences,
5. striking out human repetitive sequence fragments identical with some entries in the repetitive sequence database (repbase), and
6. exclusion of redundant data by making contig sequences according to results of the BLAST program.

The refined data are compared with sequences in the public database (EMBL or GenBank) and are classified into three groups by OPT score and identity. Three groups are high homology scoring group (HSG), medium homology scoring group (MSG) and low homology scoring group (LSG). Generally speaking, HSG and LSG correspond to the known gene category and the unknown one respectively. The HSG threshold values are defined as OPT $\geq$ 400 and identity $\geq$ 90(%). The MSG's as OPT $\geq$ 400 and identity $<$ 90(%) or 200 $\leq$ OPT $<$ 400. The LSG's as OPT $<$ 200.

3 Result

With this system, we analyzed about 30,000 cDNA sequences from brain, aorta, placenta and heart, which were sequenced by Otsuka GEN Research Institute, OTSUKA Pharmaceutical Co., Ltd. (Tokushima, Japan) and Human Genome Center, Institute of Medical Science, The University of Tokyo (Tokyo, Japan). We excluded redundant data and junk and vector fragments. As the result, the number of cDNA sequences was reduced from 33,102 to 19,538, and the accuracy of homology search was improved. And we classified all cDNA sequence data into three categories (HSG, MSG and LSG). Table 1 shows the results of the data category.

Table 1 : The Result of Data Category

	brain	aorta	placenta	heart	total
HSG	3464	1215	1119	863	6661
MSG	1724	471	412	85	2692
LSG	5729	2469	1299	688	10185

These results with our system are available via GenomeNet WWW server (http://genotk.genome.ad.jp:8010).

Acknowledgement

We thank Dr. Atsushi Ogiwara (National Institute for Basic Biology) for technical advices. This work was supported in part by a Grant-in-Aid (08283103) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

[1] G. Grillo, M. Attimonelli, S. Liuni and G. Pesole, "CLEANUP: a fast computer program for removing redundancies from nucleotide sequence databases," *CABIOS*, Vol. 12, No. 1, pp. 1-8, 1996.