Construction and Application of Mouse Full-length cDNA Database

Hideaki Konno [12]
hkonno@rtc.riken.go.jp

Yoshifumi Fukunishi [12]
fukunisi@rtc.riken.go.jp

Toshinori Endo [1]
tendo@rtc.riken.go.jp

Yoshihide Hayashizaki [1]
yosihide@rtc.riken.go.jp

[1] Laboratory for Exploration Research Project, Genomic Sciences Center (GSC) and Genome Science Lab, RIKEN Life Science Tsukuba Center, the Institute of Physical and Chemical Research (RIKEN), 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan
[2] CREST, Japan Science and Technology Corporation (JST)

1 Introduction

Our group has been making effort to collect mouse full-length cDNA clones as large scale Mouse cDNA project. It is aiming at collecting all kind of expressed full-length Mouse cDNA clones. The first phase of the project is to make cataloged non-redundant cDNA clone bank. It is necessary for analysis of sequence data and further applications of cDNA clones.

We sequence 3'-end part of all cDNA clones and we use them as identification sequence tags. And then we classify them in terms of cluster analysis. Each cDNA cluster consists of the identical sequences. The clustering by sequence identity is achieved by an all-against-all homology search. The parameters for the homology search were optimized to give the suitable sensitivity comparing to our sequence accuracy based on the actual sequence data. If the sequence tag was new for us, it will be searched through known sequence database and EST database.

We had developed automated system of series of the analysis. We improved this clustering system in accuracy and processing speed. And results of these analyses are stored into the database which is managed by RDBMS.

We have gotten seventy several thousands of identical cDNA clusters of over 460,000 clones. These cDNA clones and their information are provided for further analysis. Information retrieval system for the database is available through web browser and/or UNIX command line. For example, we have applied our cDNA clones to cDNA microarray. It is very useful for further analysis to link from each spot of cDNA microarray data to each clone data.

2 System and Method

The clustering is done by several steps. All these steps are automated.

After 5'-end and 3'-end parts are sequenced, they are stored into database entry with its cDNA library information. We are using SYBASE Adaptive Server 11.9.2 for managing sequence and library data.

At the first step, vector and primer sequence are removed from each sequenced data. Needleman-Wunsch algorithm [1] based method was applied to find the edge of primer sequences.

And then, picked out 100 nt sequences from each end of sequences as sequence tag. Non complex repeated sequences like poly-A are excluded before making the tags. Then, the tags are made into a homology search with in-house 3'-end or 5'-end databases using BLAST [2, 3] and FASTA [4]. We have examined this homology search conditions by computer simulation and actual data analysis.

If it is not found any equivalent sequence in existing sequence database, then new group is created and added to in-house database. Otherwise the tag is added as new member of the existing group. And then, created new group tag is made into homology search with known sequence database and EST database. The homolologous entry's information, i.e. GID, ACCESSION, DEFINITION, etc. and homology scores are stored as the homology search results. The data sets of known sequence data and EST data were taken from NCBI ftp server [5] We use nt – All Non-redundant GenBank+EMBL+DDBJ+PDB sequence (but no EST, STS, GSS, or HTGS sequences) – as known sequence data, and est_human and est_mouse as EST database.

These analysis are done automatically and each results are stored into the database.

At the beginning, we had used BLAST 1.4.10MP to compare sequences, but there were miss-clustering problems caused by gapping sequencing errors. We changed to use BLAST 2.0 – gapped BLAST – to avoid this problem. But it requires about three to four times longer processing time to rebuild BLAST2 database than previous one. We process over 5,000 sequences per day, if we update BLAST2 database sequence by sequence, it takes too much time to process. So we developed two steps clustering system. At the first step we cluster incoming batch of sequences before comparing existing sequence database, and then only clustered sequences will be compared with existing database. It is no use to rebuild BLAST2 database while clustering because incoming sequences has been non-redundant with pre-clustering step. We improved accuracy of clustering and processing speed with this clustering method.

And we made easy to retrieve the cDNA information from the database, using web browsers (i.e. Netscape Navigator). We are using apache httpd with CGI program and Java applet for retrieving cDNA information. CGI programs are mainly written in perl and C language. Some part of our sequence data has been opened to public (`http://genome.rtc.riken.go.jp/`).

Acknowledgements

This study has been supported by Special Coordination Funds and a Research Grant for the Genome Exploration Research Project from the Science and Technology Agency of the Japanese Government, CREST (Core Research for Evolutional Science and Technology) of Japan Science and Technology Corporation (JST), and a Grant-in-Aid for Scientific Research on Priority Areas and Human Genome Program from the Ministry of Education and Culture, Japan to Y.H.

References

[1] Needleman, S.B. and Wunsch, C.D., A general method applicable to the search for similarities in the amino acid sequence of two proteins, *J. Mol. Biol.*, 48(3):443–453, 1970.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–410, 1990.

[3] Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, 25:3389–3402, 1997.

[4] Pearson, W. R., and Lipman, D. J., Improved tools for biological sequence analysis, *Pro. Natl. Acad. Sci.*, 85: 2444–2448, 1988.

[5] `ftp://ncbi.nlm.nih.gov/blast/db/` .