Automated cDNA Information System for Large Scale cDNA Project

Hideaki Konno [12] hkonno@rtc.riken.go.jp
Yuichi Sugahara [1] sugahara@rtc.riken.go.jp
Yoshifumi Fukunishi [12] fukunisi@rtc.riken.go.jp
Kazuhiro Shibata [1] shibata@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

We have been making effort to collect various Mouse 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 this project is to collect non-redundant cataloged cDNA clone bank. It is necessary for sequence data analysis and further applications of cDNA clones.

Comparing two full length cDNA clones, if the end part of sequences are different, these clones are obviously different clones. If the end part of sequence is equivalent, these clones may be equivalent clone practically. It is obviously far easier to compare both end tip of tagged sequences than to compare whole sequences, especially in case of the sequence is long. And it may be very rare case that different clones with same 5'-end and 3'-end sequences.

It is effective way to classify cDNA clones, sequencing 5'-end and 3'-end and comparing previously sequenced data. We have developed automated cDNA classification system to make non-redundant cDNA clone collection.

2 System and Method

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

After 5'-end and 3'-end parts are sequenced, they are stored into database entry with its cDNA library informations. We are using SYBASE System XI RDBMS 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 100b 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] and FASTA [3]. We have examined this homology search conditions by computer simulation and actual data analysis.

If it is not found any equivalent sequence, 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 [4]. 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 analyses are done automatically and each results are stored into the database.

And we made easy to retrieve the cDNA information from the database, using WWW browsers (i.e. Netscape Navigator). We are using apache httpd with CGI program and Java applet for retrieving cDNA informations. CGI programs are mainly written in perl and C language. This server has not been opened to public, because it is part of our laboratory information system. Sequenced cDNA clone information will be opened near future.

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., 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):443–53, 1970.

[2] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.*, 215:403–10, 1990.

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

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