CAMUS DB for Amino Acid Sequence Data

Sadahiko Misu
smisu@genes.nig.ac.jp

Takayasu Iizuka
tiizuka@genes.nig.ac.jp

Yuichi Kawanishi
ykawanis@genes.nig.ac.jp

Kaoru Fukami-Kobayashi
kfukami@genes.nig.ac.jp

Naruya Saitou
nsaitou@genes.nig.ac.jp

DNA Data Bank of Japan [DDBJ], National Institute of Genetics,1,111 Yata,
Mishima, Shizuoka 411-8540, Japan

Keywords: CAMUS, DDBJ, DAD, homology search, compressed database, multiple alignment

1 Introduction

DDBJ/EMBL/GenBank International Nucleotide Sequence Database is still increasing, keeping doubling time only slightly longer than one year for the last these years. This situation affects the increase of DDBJ Amino acid sequence Database (DAD) [4], which is made from translation of nucleotide sequences in CDS regions annotated in DDBJ [3], and makes computation time for homology search of DAD longer as well as that of DNA database. We therefore created compressed sequence database, consisting of highly homologous sequence clusters in multiple aligned form with representative sequences, the DAD version of CAMUS (Compressed database for homology searches And MUltiple aligned Sequence database) [2].

2 Method

Data compression has done essentially in accordance with the strategy for the data compression of nucleotide sequences as follows. Each division of DAD was treated separately. Because the taxonomic divisions such as VRT (vertebrates) and PLN (plants) are made according to major classification of organisms, the sequences in each of those divisions are expected to be closely related if they are orthologous, which makes our compression process easier. We first sorted all sequences in one division by amino acid length, then sequences longer than 1,000aa were chopped into 1,000aa length fragments. Those short sequences were used as query sequence for BLAST [1] homology search. We used several cutting points for sequence identity, ranging from 30% to 99%, where the sequence length threshold was set to more than 10aa in all cases. When one sequence is found to be homologous with query sequence, it is eliminated from the target database. By this procedure, evolutionarily highly homologous sequences were grouped into one representative sequence, and the set of representative sequences are called “Compressed Database”, which can be used as target database for any homology search. We have implemented BLAST search for the secondary database as well as for the DDBJ version of CAMUS.

Sequence groups with high homology are also obtained through the construction of Compressed Database. We visualize them as a multiple alignment by unifying BLAST pairwise alignments between a representative sequence and group members. This Multiple-aligned Sequence database is also part of the DAD version of CAMUS DB.

3 Results

We used DDBJ DAD Release 16 that were made from DDBJ nucleotide sequence database Release 46. DAD consists of 18 divisions as in the case for the DDBJ nucleotide sequence database. While nucleotide EST division has vast number of sequences, few entries have CDS. We thus treated the small EST division of DAD as one group, and compressed the 18 divisions separately. Compression ratios for the taxonomic divisions were 5-76% in terms of the number of entries, when we used the cutting point of 99%. If we count the number of total letters, the compression ratios are within 16-77% range. We recognized nested structure in sequence groups using different cutting point.

Website URL of CAMUS DB is http://hypernig.nig.ac.jp/camus/.

4 Future Perspectives

We proposed a way to attack the data explosion in nucleotide and/or amino acid sequence database by compressing them in CAMUS. We are currently updating both versions of CAMUS DBs when DDBJ and DAD databases are updated. As the volume of these databases increases, the compression procedure itself will be getting burdensome. We are thus planning to speed up the procedure by using parallel computers and by taking more efficient algorithms.

References

[1] Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J., Basic local alignment search tool, *J. Mol. Biol.,* 215(3):403–410, 1990.

[2] Kikuchi, M., Misu, S., Imanishi, T., and Saitou, N., CAMUS DB: development of structural database for homology search, *Currents in Computational Molecular Biology*, Miyano S., Shamir R. and Takagi T. (eds.), Universal Academy press, Tokyo, 80–81, 2000.

[3] http://www.ddbj.nig.ac.jp/

[4] http://www.ddbj.nig.ac.jp/ddbjnew/nl17/news-j.html