Japanese Journal of Inherited Metabolic Diseases 2003;19:249

The 3rd Asian Symposium on Inherited Metabolic Disease

AS-1

Development of the Chinese Gene Variation Database (CGVdb)

○Kwang-Jen Hsiao[1,2,3,4], Chien-Han Lin[3], Jung-Wei Fan[3], Chia-Hsiu Tu[2,4], Szu-Hui Chiang[4], Tze-Tze Liu[2]

[1]Institute of Genetics, [2]Genome Research Center, and [3]Bioinformatics Program, National Yang-Ming University; [4]Dept. of Medical Research and Education, Taipei Veterans General Hospital; Taipei, Taiwan 112, R.O.C

In order to collect genetic mutations and other variations information in the Chinese population for clinical application and medical genetic research, a Chinese Gene Variation Database (CGVdb) <http://www.CGVdb.org.tw> was established. A semi-automatically electronic data collection method was developed to help us to collect the genetic mutation and polymorphism study reports from PubMed, BIOSIS Previews, and EMBASE databases. These mutation and variation reports related to genetic diseases in Chinese populations, but not including somatic mutations, are defined as Chinese gene Variation Reports (CVRs). Customized searching string composed of index terms (e.g. MeSH) and search field tags was used to retrieve CVRs from the databases.

In order to evaluate the effectiveness of our electronic data collection method, we have established a standard data set which contains those papers expected to be true CVRs by manual reviewing all the papers published in 18 selected journals related to human genetics (9 in U.S. and Europe, 5 in Mainland China and 4 in Taiwan) published in the period of 1997 and 1998. PubMed has collected 6,942 papers published by these journals during this period of time. Manual review has found 116 CVRs. However, our electronic data collection method detected 263 papers. By comparison of the standard data set and the result of our electronic data collection method, recall rate of 0.974 (113/116) and precision of 0.43 (113/263) were found. Using this electronic data collection method, we have found 766 reports from searching all the papers contained in PubMed in the period of 1997 and 1998. Among them, 266 were found to be CVRs (precision: 0.347). There were 33 known false negatives (recall ≤ 0.89) in this data set, but 32 of them can be detected by searching BIOSIS and EMBASE. Based on this data set, search strings were developed to search BIOSIS and EMBASE. The recall and precision were found to be ≤ 0.956, 0.441 and ≤ 0.958, 0.417 for BIOSIS and EMBASE, respectively. A 59% (156 extra) increase of CVRs was found by searching the BIOSIS and EMBASE in addition to PubMed (1997-1998). This result indicated that the total CVRs in the PubMed (1966-2002), BIOSIS (1985-2002), and EMBASE (1982-2002) could be estimated to be around 2,620 from the 7,518 non-redundant reports retrieved by our electronic method on June 4th, 2003. A dictionary-based gene name identification program, namely Idgene, with a recall around 0.8-0.9 for CVRs was developed in aid of CVRs text annotation. This program may be useful to annotated gene name for other documents and is assessable through Internet <http://www.cgvdb.org.tw/idgene/>.

220