Construction and Application of Bio-Information Database

Sun Guofeng¹, Pan Wei²

¹Institute of Agricultural Information, Chinese Academy of Agricultural Sciences, Beijing, China, 100081
²Library, China National Institute of Standardization, Beijing, 100088

Abstract:

The paper describes generally both theoretical and practical issues in construction and use of bio-information database for Chinese agricultural scientific researchers. It stressed the data morns and standards used in knowledge description of the database. It also introduces methods and mode of coordination and organization of the database construction including various bio information data format defining, data collecting, processing, classification, encoding, indexing and updating etc. The construction and application of the databases in wheat, maize, rice and legume research and development shows great effectiveness. These are unique databases which are valuable to Chinese users.

Keywords: bio-information database, database construction, research and development, Database application

1. Introduction

Since 60s in 20th century, following the development of technologies in protein series test, nucleic acid series test, gene clone and PCR(polymerase chain reaction), the world research institutions may access growat amount of bio-molecule data. To collect, store, analyse and explain those data, and get useful bio-information and to help users in bio-informatics research has been critical and meaningful. The in-depth use of bio-informatics in genome analysis of agricultural crops has helped the formation of agricultural bio-informatics which is a basis of gene set research of agricultural crops. It is necessary technology in analysis of functions of gene, gene structure and gene products. At present, the main achievements in agricultural bio-informatics are in rice gene set analysis(Zheng Guoqing & Gao Liangzhi. 2002.) Bio-informatic database will be an important base for agricultural research and crop genome analysis.

2. Development of foreign bio-informatics database

2.1 UK Crop Plant Bioinformatics Network

The comparative analysis of plant genomes has become a key area of research within the last few years. The last few years have also seen an explosion in the amount of sequence information (both nucleotide and protein) available in the public databases EMBL. This growth in 'genomics' is having a profound effect on the way that scientists do science and the next few years will almost certainly see similarly huge increases in the information produced from transcriptomics and proteomics facilities. There is, therefore, a clear need for easy to use tools to manage this wealth of information. To apply a bioinformatics approach to
the study of genome evolution in crop plants with specific aims of developing software and databases that will facilitate the querying of genomic information from different crop species. Particular emphasis has been placed on developing software tools for comparative mapping. To manage and make sense of the available information, we have used the ACEDB (A C elegans Database) database system to create separate databases for each of the UK CropNet projects.

2.2 The EMBL Nucleotide Sequence Database

The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications. The database is produced in an international collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ). Each of the three groups collects a portion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis (EMBL, 2008).

2.3 The UK Crop Plant Bioinformatics Network

The UK Crop Plant Bioinformatics Network (UK CropNet) established in 1996 as part of the The Biotechnology and Biological Sciences Research Council (BBSRC)'s Plant and Animal Genome Analysis special initiative, the focus is the development, management, and distribution of information relating to comparative mapping and genome research in crop plants. BBSRC has highlighted nearly 50 top bioscience researchers and research groups for their contribution to the UK’s economic and social wellbeing. The groups and individual researchers represent a cross-section of bioscientists, from universities and institutes across the UK and from all parts of the BBSRC science base. Individual databases for Arabidopsis, Barley, Brassica, Forage grasses, Millet and comparative analysis have been developed by each of the partners and are under constant development, reflecting the ever increasing amount of genomic information made available by the scientific community. Furthermore, new databases are currently under development and will also be made available via this site. In addition new tools and displays to highlight the agronomic significance of comparative analysis are being developed (The UK Crop Plant Bioinformatics Network, 2008). The site hosts a wide range of databases and software developed by UK CropNet, as well as hosting many other plant databases developed in the USA. You can perform a keyword text search across all of these databases or use UK CropNet BLAST server to search against all of the sequences in these databases.

2.4 The Legume Information System (LIS), USDA

Legumes (soybeans, dry beans, peas, etc.) are excellent vegetable sources for proteins and oils. The Legume Information System (LIS) developed by the National Center for Genome Resources in cooperation with the USDA Agricultural Research Service (ARS), is a comparative legume resource that integrates genetic and molecular data from multiple legume species enabling cross-species genomic and transcript comparisons. The LIS virtual plant interface allows simplified and intuitive navigation of transcript data from Medicago truncatula, Lotus japonicus, Glycine max and Arabidopsis thaliana. Transcript libraries are represented as images of plant organs in different developmental stages, which are selected to
query the analyzed and annotated data. Complex queries can be accomplished by adding modifiers, keywords and sequence names. The LIS also contains annotated genomic data featuring transcript alignments to validate gene predictions as well as motif and similarity analyses. The genomic browser supports comparative analysis via novel dynamic functional annotation comparisons. CMap, developed as part of the GMOD project (http://www.gmod.org/cmap/index.shtml), has been incorporated to support comparative analyses of community linkage and physical map data. LIS is being expanded to incorporate gene expression and biochemical pathways which will be seamlessly integrated forming a knowledge discovery framework (Michael D. Gonzales, Eric Archuleta, Andrew Farmer, etc.. 2004).

Moreover, SGMDB(Soybean Genomics and Microarray Database) was established in 1999 by ARS, USDA and it contains 50 million microarray and other data.

3. Development of Chinese bio-information database

In 2000, China launched “China hybrid rice gene set research and development plan” and finished in 2003. In 2001, under cooperation of the Descendibility Research Institute of Chinese Academy of Sciences(CAS) and the national hybrid rice research center, the genome bio-informatic center of CAS completed working framework map and database of Chinese rice genome which possessed international advantages. In 2002, the refined map of Chinese rice genome was finished. The Chinese crop germplasm resources database included over 111 kinds of crops, more than 270,000 germplasm information and nearly 13 million data fields, and the data took 600M totally(Yan Jin & Xiao Langtao.2006). Moreover, the Chinese animal species cataloging database supported by CAS, the national Science and Technology Committee and Bio diversity Committee of CAS, had collected over 10000 animal basic information. In Shanghai, there has been a database which possessed Shanghai agricultural bio-gene resources and world excellent resources implemented by Shanghai Agriculture and Bio-gene Center (Long Ping, Cao Kaixiong.2004).

Principles for database construction are generally comprehensive, accurate, updated, user needed and resource sharable. To meet such objects, and make the database to be more effective and efficient, we must use international and national data description standards, rules or regulations in cataloging, classification, encoding and indexing. The mode for data collecting and updating sometimes should be decentralized and a coordinating group are needed if the database construction is implemented at the national level. Training are especially important to database construction staffs at various levels for a high quality database and better data sharing.

4. Use of bio-information database

Following the fast development of information technology, tools and softwares in information searching, series cooperation, analysis and means of bio-informatics, the using scope of agricultural bio-informatics database has been getting wider and wider. For an example, it has been used successfully in animal and plant variety improvement, in protection of rural crop genetic resources and in control of diseases and pests. The databases provided comprehensive data for users in assisting them in their research and benefit to the high yield and quality agricultural products. It also will satisfy the requirements of the green environmental market.
References:

The UK Crop Plant Bioinformatics Network, http://ukcrop.net/about.html. 2008-6-2