Chapter 93
Use of SciDBMaker as Tool for the Design of Specialized Biological Databases

Riadh Hammami
Université Laval, Canada

Ismail Fliss
Université Laval, Canada

ABSTRACT

The exponential growth of molecular biology research in recent decades has brought concomitant growth in the number and size of genomic and proteomic databases used to interpret experimental findings. Particularly, growth of protein sequence records created the need for smaller and manually annotated databases. Since scientists are continually developing new specific databases to enhance their understanding of biological processes, the authors created SciDBMaker to provide a tool for easy building of new specialized protein knowledge bases. This chapter also suggests best practices for specialized biological databases design, and provides examples for the implementation of these practices.

INTRODUCTION

The exponential growth of research in molecular biology has brought concomitant proliferation of databases for stocking its findings. A variety of protein sequence databases exist. While all of these strive for completeness, the range of user interests is often beyond their scope. Large databases covering a broad range of domains tend to offer less detailed information than smaller, more specialized resources, often creating a need to combine data from many sources in order to obtain a complete picture. Scientific researchers are continually developing new specific databases to enhance their understanding of biological processes. In this chapter, we present the implementation of a new tool for protein data analysis. SciDBMaker is stand-alone software that allows the extraction

DOI: 10.4018/978-1-4666-3604-0.ch093

of protein data from the Swiss-Prot database, sequence analysis comprising physicochemical profile calculations, homologous sequences search and multiple sequence alignments (Riadh Hammami, Zouhir, Naghmouchi, Ben Hamida, & Fliss, 2008). Furthermore, with its easy-to-use user interface, this software provides the opportunity to build more specialized protein databases from the universal protein sequence database Swiss-Prot. It compiles information with relative ease, updates and compares various data relevant to a given protein family and could solve the problem of dispersed biological search results. Using SciDBMaker, two databases were developed, namely BACTIBASE (R Hammami, Zouhir, Ben Hamida, & Fliss, 2007; Riadh Hammami, Zouhir, Le Lay, Ben Hamida, & Fliss, 2010) and PhytAMP (Riadh Hammami, Ben Hamida, Vergoten, & Fliss, 2009) and analyzed here as 'proof of concept'. Here, we share our knowledge in protein database development and provide advices that we hope are useful to people designing their own biological database.

BACKGROUND

Bioinformatics and computational biology methods are increasingly used to study biological systems and widely applied to facilitate collecting, organizing, and analyzing of large-scale of data in molecular biology. Biological databases appeared as invaluable method for managing these data and for making them accessible to scientific community. In this mold, molecular biological databases could contain either the result of large amounts of molecular biological experiments or manual extraction of literature data. Depending on the type of biological data that they enclose, these biodatabases fulfill different functions. Most molecular data are in the form of a biosequence of a DNA, RNA, or a protein molecule.

Dr. Dayhoff and her research group were pioneers in the development of computer methods for the analysis of protein sequences evolution. This led to the establishment in 1984 of the Protein Information Resource (PIR) as a resource to assist researchers in the identification and interpretation of protein sequence information (Wu et al., 2003). This has inspired Amos Bairoch in 1986 for the creation and public release of Swiss-Prot sequence database (Bairoch, 2000). The increasing quantities of nucleic acid sequence data being generated worldwide in 1980s created the need to the construction of nucleic acid sequence databases, notably GenBank (Benson, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2009), European Molecular Biology Laboratory Nucleotide Sequence Database (EMBL) (Hingamp, van den Broek, Stoesser, & Baker, 1999) and DNA Data Bank of Japan (DDBJ) (Tateno, Fukami-Kobayashi, Miyazaki, Sugawara, & Gojobori, 1998). Together, these databases form the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org) which archives and makes publically available more than 80 million individual molecular sequences (Benson, et al., 2009). In 2002 PIR, along with its international partners, EBI (European Bioinformatics Institute) and SIB (Swiss Institute of Bioinformatics), unified the PIR, Swiss-Prot, and TrEMBL databases by creating UniProt, a single worldwide database of protein sequence and function. Today, an important collection of biological databases are available in the public domain, spanning the worlds of sequence, family and structure of DNA, RNA and proteins, organisms, genomes, signaling and metabolic pathways, microarrays, biodiversity, and so on (Ellis & Attwood, 2001). Currently, there are many different types of biodatabases, including:

- **Bibliographic databases**: are considered as important information sources for biomedical research and contain summary information taken from a variety of sources including journals, books, conference reports and patents. PubMed is one of the largest databases of life science abstracts with more than 19 million citations for bio-