BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 18 2005, pages 3694–3696
doi:10.1093/bioinformatics/bti580

Databases and ontologies

DEPD: a novel database for differentially expressed proteins

Quan-Yuan He[1], Jing Cao[1], Xiang-Hua Liu[2], Miao-Xin Li[3], Yi-Song Liu[1], Jin-Yun Xie[1] and Song-Ping Liang[1,*]

[1]College of Life Science, Hunan Normal University, Changsha 410081, People's Republic of China, [2]State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Handan Road 220, Shanghai 200433, People's Republic of China and [3]Shanghai Center for Bioinformation Technology (SCBIT), 100 Qinzhou Road, 12th Floor Shanghai 200235, People's Republic of China

Received on March 13, 2005; revised on June 13, 2005; accepted on July 11, 2005
Advance Access publication July 14, 2005

ABSTRACT

Summary: The Differentially Expressed Protein Database was designed to store the output of comparative proteomics studies and provides a publicly available query and analysis platform for data mining. The database contains information about more than 3000 differentially expressed proteins (DEPs) manually extracted from the published literature, including relevant biological, experimental and methodological elements. Tools for visualization and functional analysis of DEPs are provided via a user-friendly web interface.

Availability: http://protchem.hunnu.edu.cn/depd/

Contact: liangsp@hunnu.edu.cn

INTRODUCTION

The goal of comparative proteomics is to systematically compare global protein expression profiles, focusing on quantitative changes that occur as a function of disease, treatment and environment (Somiari *et al*., 2003). Such an approach may provide comprehensive insight into the dynamics of the proteome and reveal protein markers of disease. The increasing volume of data from high-throughput proteomics urgently needs to be made accessible in databases with standards for data storage, exchange and analysis. However, as yet, no database of differentially expressed proteins (DEPs) has been publicly available. We have addressed this by developing the Differentially Expressed Protein Database (DEPD), a relational database that provides a platform for query and analysis. Currently, the DEPD contains information about more than 3000 DEPs, manually extracted from published literature, largely from studies of serious human diseases including lung cancer, breast cancer and liver cancer.

The DEPD integrates functional information from the Swiss-Prot/TrEMBL (Bairoch and Apweiler, 2000), GO (The Gene Ontology Consortium, 2001), KEGG (Kanehisa *et al*., 2004) and Pfam (Bateman *et al*., 2004) databases with published studies of DEPs. We developed an XML schema named CPXS 0.1 (Comparative Proteomics XML Schema) as a data exchange standard for comparative proteomics. We also set up a user-friendly web interface with tools for querying, visualization and analysis of the results of published comparative proteomics studies. All of the DEPD data can be downloaded freely from the download page of website.

*To whom correspondence should be addressed.

DATA COLLECTION AND XML SCHEMA

The data in DEPD were curated from published literature by trained biologists, ensuring consistently high quality. The core data were obtained through keyword searches of PubMed (Wheeler *et al*., 2000) for comparative proteomics articles. The PubMed search hits were then filtered manually by expert biologists. After critical reading of the articles, the curators recorded names, accession numbers, types of expression change, folds of expression change and relevant information about contexts, according to our CPXS 0.1 schema (see below). Protein sequence, taxonomy information, domain architecture, functional annotation and pathway category were automatically integrated into DEPD from the Swiss-Prot/TrEMBL, Pfam, GO and KEGG databases using our DEPDManager tool.

CPXS 0.1 is a minimum specification of the data and metadata that should be captured from comparative proteomics experiments. Its purpose is to highlight the biological significance and context of the DEP rather than to provide sufficiently detailed information to reproduce the experiments. A key feature of CPXS is that it takes a comprehensive description of context as a crucial component of data entry and it does so in an efficient and flexible way. For example, a basic data entry describing a comparative proteomics experiment comprises not only a list of DEPs but also several context elements including SamplePair, EnvironmentalFactorList and MethodologicalContext. This is essential because data generated by proteomics are meaningful only in the appropriate biological, environmental and methodological context. Further details of the schema and these elements can be found at the DEPD website.

SYSTEM ARCHITECTURE AND IMPLEMENTATION

The DEPD system is composed of three tiers: the relational database back-end implemented in MySQL4.1, the Apache/Tomcat web- and application-server system, and the web clients. A series of servlets extend the Java HttpServlet class and take POST actions. These servlets are driven by Apache Tomcat 4.1, which generates the dynamic web pages. The free third-party packages iText (http://www.lowagie.com/iText/) and JfreeChart (http://www.jfree.org/jfreechart/) are used to create PDF files and charts in

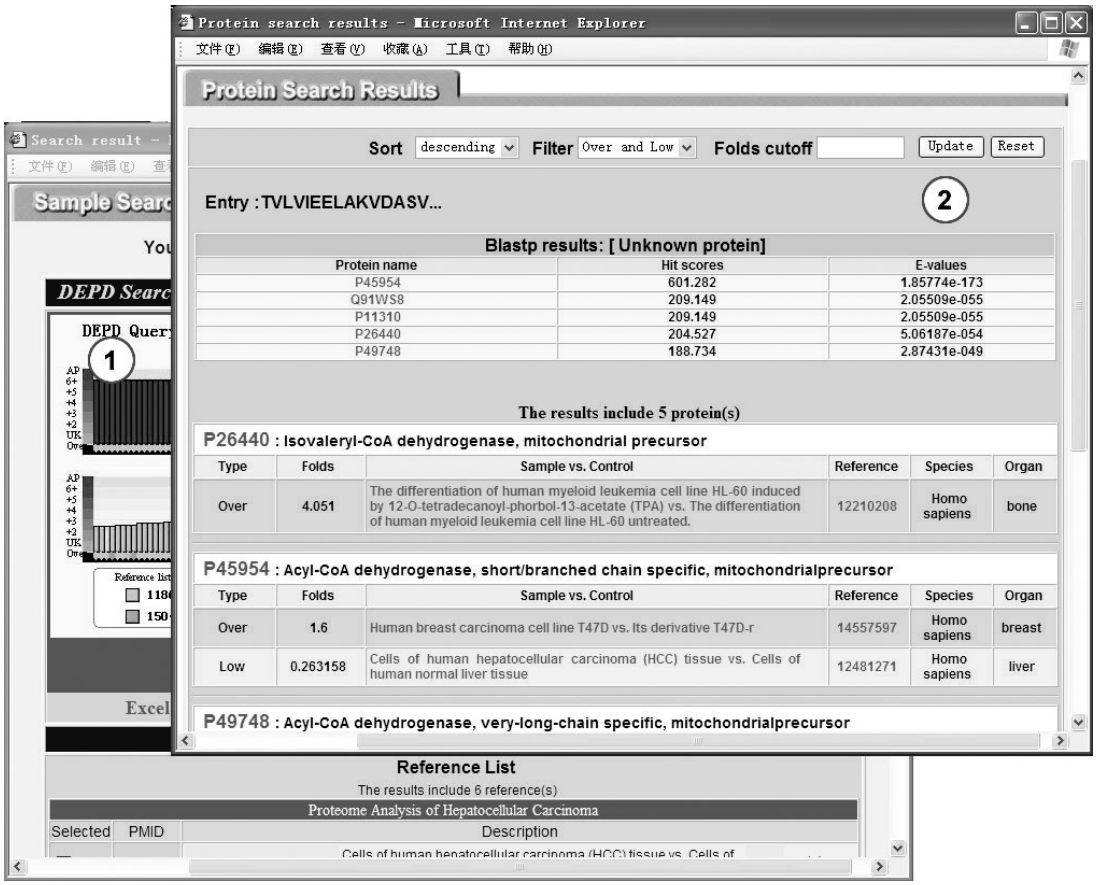

Fig. 1. DEPD's database query results. (1) Query with sample attributes retrieves a reference list and a graph view for results summary. (2) Query results of protein name/sequence search include a protein list with key information.

web pages. A web interface of the DEPD has been developed for data query and analysis.

WEB INTERFACE

DEPD provides a user-friendly web interface for data query, visualization and analysis. There are two kinds of query processes for different cases. For users who want to find their favorite proteins in the database, Protein Search, as a flexible tool, supports many different approaches to query DEPD by using protein accession numbers, protein/gene names, key words as well as protein sequences. BLAST software (Altschul *et al*., 1990) has been used to search proteins similar to users' favorite one in DEPD. Query results include a protein list with key information such as type, folds, sample versus control, species and organ. It also can be filtered by expression type and sorted by ascending or descending order of folds. For users who want to find any researches for same/similar samples or diseases, Browse and Sample Search pages were designed. These queries can be refined by combining two or more sample features. Query results comprise a reference list and a bar graph in which DEPs are represented by columns and sorted on the basis of their type and folds of expression change (Fig. 1). The height of the column is proportional to the fold change of expression. Over- and under-expressed DEPs are colored red and green, respectively. Detailed information about DEPs and literature references can also be made available by clicking the hyperlink on the graph. Users can also manually choose experiments that they are interested in for further functional analysis.

In the functional analysis process, all DEPs of selected experiments are categorized into three non-redundant DEP groups including total, over-expressed and low-expressed DEP groups. Then, users can classify DEPs in each group by their GO and KEGG annotations. These features of DEPD's web application make it easy for users to find DEPs that appear repeatedly in different experiments and obtain details about DEPs' functional classification.

DATA EXCHANGE AND FUTURE WORK

All data stored in DEPD can be downloaded freely in the XML format along with its schema or as flat files. Temporary search results can also be saved as Excel and PDF files in the Search results page. At the same time, researchers are encouraged to submit relevant data to DEPD.

We are continuing to collect and compile more entries for the database, optimize the structure of the database to accommodate the development of comparative proteomics and append some important elements such as protein modification or relevant DNA microarray data to enhance data quality. In the future we intend to integrate protein–protein interactions from BIND and DIP (Xenarios *et al*., 2002) into DEPD.

ACKNOWLEDGEMENTS

The authors thank Dr Jingchu Luo (Peking University, China), Dr David J. Studholme (Sainsbury Laboratory, John Innes Centre, UK) and anonymous reviewers for suggestions and comments on the database systems and the manuscript. This work was supported by a grant from National 973 Project of China (2001 CB510208) and a grant from National Natural Science Foundation of China (90408017).

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1990) A basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bader,G.D. and Hogue,C.W.V. (2000) BIND—a data specification for storing and describing biomolecular interactions, molecular complexes and pathways. *Bioinformatics*, **16**, 465–477.

Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **2**, 45–48.

Bateman,A. *et al*. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32** (Database issue), 138–141.

Gasteiger,E., Hoogland,C., Gattiker,A., Duvaud,S., Wilkins,M.R., Appel,R.D., Bairoch,A. and Walker,J.M. (2005) *Protein Identification and Analysis Tools on the ExPASy Server*. The Proteomics Protocols Handbook, Humana Press, pp. 571–607.

Kanehisa,M. *et al*. (2004) The KEGG resources for deciphering the genome. *Nucleic Acids Res.*, **32**, 277-280.

Somiari,R.I. *et al*. (2003) High-throughput proteomic analysis of human infiltrating ductal carcinoma of the breast. *Proteomics*, **3**, 1836–1873.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.

Wheeler,D.L. *et al*. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **28**, 10–14.

Xenarios,I. *et al*. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.*, **30**, 303–305.