Databases and ontologies

MPSS: an integrated database system for surveying a set of proteins

Pei Hao1,2, Wei-Zhong He3, Yin Huang2, Liang-Xiao Ma3, Ying Xu2, Hong Xi3, Chuan Wang2, Bo-Shu Liu2, Jin-Miao Wang2, Yi-Xue Li2,* and Yang Zhong1,∗

1School of Life Sciences, Fudan University, Shanghai 200433, China, 2Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China and 3Shanghai Center for Bioinformation Technology, Shanghai 200235, China

Received on November 24, 2004; revised on January 16, 2005; accepted on January 30, 2005

Advance Access publication February 4, 2005

ABSTRACT

Summary: We design and implement an integrated database system called ‘multi-protein survey system’ (MPSS), which provides a platform to retrieve information about many proteins at a time. This system integrates several important and widely used databases including SwissProt, TrEMBL, PDB and InterPro, plus useful references such as GO and KEGG to other databases. Users may submit a group of protein IDs, entry names, SwissProt/TrEMBL accession numbers or GenBank GIs through MPSS’ web interface, and obtain protein annotation information from public databases and pre-computed molecular properties speedily. MPSS can also supply comprehensive information about query proteins, including 3D structures, domains, pathway, gene ontology and visual presentation of mapping to the GO tree and KEGG pathway, to provide an up-to-date view of available knowledge with regard to the structures and molecular functions of proteins under study.

Availability: MPSS is freely accessible at http://www.scbit.org/mpss/

Contact: yangzhong@fudan.edu.cn, yxli@sibs.ac.cn

INTRODUCTION

Proteomics is the large-scale study of all proteins expressed in a cell, tissue, organ or the whole organism. It gives rise to a need for some efficient systems to retrieve, correlate and analyze large datasets. However, many existing online protein information systems process only one protein at a time. Therefore, it is a laborious and time-consuming task for people working on high-throughput proteomics projects to gather all relevant information for a large set of proteins under study.

We had such a frustrating experience first-hand when carrying out the annotation work for the rat liver proteome project (Jiang et al., 2004). As a result, we created a database system called ‘multi-protein survey system’ (MPSS) to integrate several popular protein information databases including sequence databases SwissProt plus TrEMBL (Boeckmann et al., 2003), structural database PDB (Deshpande et al., 2005), protein family databases such as Pfam (Bateman et al., 2004) and InterPro (Mulder et al., 2003), and also contain many useful references to other databases such as Gene Ontology (Gene Ontology Consortium, 2004) and KEGG (Kanehisa et al., 2004). We also pre-computed some parameters for each protein, such as iso-electric point (PI), aliphatic index (Ikai, 1980) and grand average of hydrophaticity, or GRAVY value (Kyte and Doolittle, 1982).

USAGE

The MPSS is made for public access as a web application with a very simple user interface. Users may enter a group of protein IDs, entry names or accession numbers from SwissProt, TrEMBL or GenBank on the submission page. The system will return basic protein information in a tabular view, which includes SwissProt annotations such as keywords, cellular location and function, and pre-computed molecular weight, PI, aliphatic index and GRAVY value. The various IDs on the result page are also externally linked to their origin, such as SwissProt, where the user can get a detailed view. Users can also go to ‘Domain Information’, ‘Gene Ontology’, ‘Pathway’ and ‘3D Structure’ pages to obtain the respective result for the same set of protein IDs. ‘Save As...’ links on each page allow the user to save the result to a local file in tab-delimited format that may then be imported into a spreadsheet program such as Microsoft Excel for further analysis or reporting. On the Gene Ontology result page, the ‘Tree View’ link shows the functional distribution of the set of proteins at a glance. The Pathway result contains links leading to a pictorial view highlighting the location of the query protein in the pathway.

The MPSS application also provides a convenient way for users to study homologues. Blast results can be pasted directly into the submission text box, and IDs after the ‘>’ in blast output are selected automatically. If a researcher is examining a new sequence without any known public identifier, they are advised to blast the new sequence against a suitable resource (e.g., GenBank ‘nr’ database, Benson et al., 2004) to find a set of homologous proteins. This can be done locally or on the web. Many public websites provide such facilities including NCBI, EBI or the SRS installation at our website http://www.scbit.org/. The resulting section of the blast output containing hit IDs can be pasted into the MPSS submission page. This approach can give users a set of comprehensive annotations relevant to the query sequence.

APPLICATION DESIGN

The architecture of MPSS is shown in Figure 1. Data content is automatically updated at a configurable frequency. Currently, it is set to update once every two weeks. (At the time of writing, the MPSS...
Integrated database system for surveying a set of proteins

Fig. 1. An overview of the MPSS architecture. The normalized protein ID(s) are mapped onto related information in MPSS. Data sources include SwissProt, TrEMBL, PDB, GO and KEGG. The main functions of MPSS are listed in the gray panels on the right.

database consists of 1,612,378 non-redundant sequences—163,496 from version 45.1 of SwissProt and 1,448,882 from version 28.1 of TrEMBL.) Together, these protein sequence entries are involved in 10,000 pathways from KEGG and are linked to over 6190 protein families, 7000 domains from Pfam and 21,000 3D structures from PDB.

MPSS collects and integrates Gene Ontology information into the local database, then maps submitted proteins onto the GO tree and creates rich links to other important databases. MPSS also provides useful molecular properties for the proteins under study including molecular weight, theoretical PI, aliphatic index and GRAVY value. This can help to confirm protein identities and facilitate further experimental design. These properties are pre-computed with the ProtParam tool from Expasy (http://www.expasy.org/tools/protpar-ref.html) to achieve high performance.

Domains, the building blocks of protein structure and function, in query proteins can also be retrieved with MPSS. This information can help to classify and predict the functions of the query protein. Domain information from a number of sources, such as Pfam and InterPro, is included.

In addition, MPSS also provides access to the pathway information in KEGG, which is useful for mapping the relationship amongst a whole system of proteins and is especially valuable for those who are interested in signal transduction networks. 3D protein structures, if available in the PDB, can be obtained easily from MPSS. Users can also view the protein classification information based on gene ontology and even map proteins onto the GO tree.

DISCUSSION AND FUTURE WORK

MPSS provides a novel information service concept that emphasizes fast and flexible access to comprehensive protein information. It integrates rich data from widely used protein-related databases including SwissProt, TrEMBL, InterPro and PDB with links to GO and KEGG. People working on proteomics and microarray studies may benefit from this flexible batch-processing approach of MPSS by allowing them to spend more time on addressing interesting biological questions. For instance, during the rat proteomics project (Jiang et al., 2004), we identified a large set of proteins and relied on MPSS to retrieve most up-to-date annotations from public domain databases.

Given the rapid advancement in biological sciences, the abilities to integrate new data types and adapt to new data formats are necessary for the success of an information service system. The simple internal database structure of MPSS makes it simple to incorporate additional protein information in the future. As a result, users will have increasing amounts of valuable information at their fingertips through MPSS.

Future work will be focused on customizing the service for different users. For example, researchers who are interested in signal transduction would like to see more information about the relationship among proteins together with some functional information, while those in protein engineering might be happy to get very detailed information of the whole structure along with the exact properties of 20 amino acids.

We believe that this new approach to obtaining molecular information will become increasingly prevalent due to the increasing volume and complexity of data involved in bench work. Thus, our ultimate goal is to turn MPSS into a fully automatic pipeline for researchers to retrieve protein information.

ACKNOWLEDGEMENTS

We would like to thank Mr. Alex Michie for improving the readability of the manuscript. This work was supported in part by grants from the EU-China Scientific Collaborative Project, Shanghai Commission for Science and Technology (03DJ14011), the ‘863’ National High-Tech Programs (2003AA231010, 2004BA711A21) and the ‘973’ National Basic Research Programs (2001CB510203, 2002CB512801, 2003CB715091).

REFERENCES