

LSUHSC Proteomics Core Facility

Applications Newsletter

February 5nd, 2007

Q&A: Databases for Protein Identifications

A. What is the database search program for protein ID in Proteomics Core Facility?

LSUHSC Proteomics Core Facility uses in-house database search program, *Mascot* (Matrix Science, London, UK). It uses statistic algorithms to screen the databases to match the mass spectrometry data of protein tryptic digests with the best candidates for identification. By their statistic significance, the matches can be separated from other random events.

B. What databases does Mascot use?

This widely used program is compatible with several public databases, such as NCBI-nr, EST, TrEMBL, Swiss-Prot etc. NCBI-nr and Swiss-Prot are the most popular databases that are periodically downloaded and updated into the in-house Mascot program.

C. How often are the in-house databases updated?

At least every 1 to 2 months, due to the rapid increase of the databases.

D. Are bacteria or viruses included in the databases?

The databases contain entries for all species that are available to the public, including archaea, eukaryotes, plants, viruses, bacteria etc. More specific taxonomy chosen in the search parameters gives less redundant results.

E. What is NCBI nr database?

National Center for Biotechnology Information (NCBI) compiles the nr database, containing the non-identical sequences from GenBank, EMBL, and DNA databank of Japan (DDBJ). Up to 12/15/2006, more than ~64 million reported sequences have been included.

The major advantages of nr database over other databases:

1. With the largest and rapidly growing number of sequence entries, it yields the highest possibility to find matches.
2. It is updated frequently.

**The shortcoming of nr database is to find many redundant protein search results.

F. What is Swiss-Prot database?

It is a manually annotated protein knowledgebase established in 1986 by the Swiss Institute of Bioinformatics (SIB). Up to 1/23/07, Swiss-Prot contains 254,000 + sequence entries.

The major advantages of Swiss-Prot database over other databases:

1. Detailed Annotation of the protein- Functions, domains, sites, secondary structures, etc.
2. Minimum Redundancy-An entry contains all protein variations generated by a same gene.

?? How does the Core Facility search your data??

They are usually searched against NCBI nr database with the specific taxonomy that you provide. The search against Swiss-Prot will be performed by request. If you have questions regarding to the database search, please contact Dr. Chou (cchou@lsuhsc.edu) for information.

Information above is quoted from the following websites:

<http://www.matrixscience.com/>; <http://ca.expasy.org/>; <http://www.ncbi.nlm.nih.gov/>