Protein sequence databases in the context of genome projects

Amos Bairoch
bairoch@cmu.unige.ch

Medical Biochemistry Department, University of Geneva
1211 Geneva 4, Switzerland

Recent developments concerning the SWISS-PROT and PROSITE databases are discussed in the context of genome projects and of new network access tools

SWISS-PROT[1] is a curated protein sequence database which strives to provide a high level of annotations (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc), a minimal level of redundancy and high level of integration with other databases. In the recent months we have developed the database in the following directions:

- We have selected a number of organisms that are the target of genome sequencing and/or mapping projects and for which we intend to be as complete as possible as well as to provide a high level of annotations. Entries originating from these organisms are cross-referenced to specialized database(s) that contain, among other data, some genetic information about the genes that code for these proteins. The organisms currently selected are (the associated specialized database is listed in brackets): B.subtilis (SubtiList); C.elegans (WormPep); D.discoideum (DictyDB); D.melanogaster (FlyBase); E.coli (EcoGene); H.sapiens (MIM) and S.cerevisiae (LISTA).
- We have made an important effort in the implementation, in SWISS-PROT, of data relevant to human genetic diseases and of their characterization at the molecular level. Information concerning disease causing mutations is now available in the database.
- SWISS-PROT has committed itself to work in close collaboration with a number of groups developing 2D gel databases. In particular we provide cross-references to the identificators for the spots corresponding to known or unknown microsequenced proteins. We also create new entries for micro- sequences that correspond to novel, yet unidentified, proteins.

PROSITE[2] is a compilation of sites and patterns found in protein sequences; it can be used as a method of determining the function of uncharacterized proteins translated from genomic or cDNA sequences. Recent developments include:

- The extension of the collection to include profile-based motif descriptions in addition to regular expression-like patterns. This will allow the detection of protein families and domains that cannot be detected using patterns due to their extreme sequence divergence. Typical examples of important functional domains which are weakly conserved are the Ig domains, the SH2 and SH3 domains, or the Fn-III domain.
- A significant increase in the number of patterns stored in PROSITE. In the current release there are 1029 patterns that allow the characterization of 18786 out of a total of 38303 entries in SWISS-PROT (close to 50

Both SWISS-PROT and PROSITE are available through the ExPASy World-Wide Web (WWW) server[3]. WWW is a powerful global information system merging networked information retrieval and hypertext. The ExPASy server allows access to the SWISS-PROT, PROSITE, SWISS-2DPAGE and SWISS-3DIMAGE databases and, through any SWISS-PROT protein sequence entry, to other databases such as EMBL, REBASE, FlyBase, GCRDb, MaizeDB, OMIM, PDB and Medline.

References

[1] Bairoch A., Boeckmann B. *Nucleic Acids Res.* 22: 3578-3580 (1994).

[2] Bairoch A., Bucher P. *Nucleic Acids Res.* 22: 3583-3589 (1994).

[3] Appel R.D., Bairoch A., Hochstrasser D.F. *Trends Biochem. Sci.* 19: 258-260 (1994).