Poster B-25

eSLDB: eukaryotic Subcellular Localization DataBase.

Authors:
Andrea Pierleoni (*Biocomputing Group, Department of Biology, University of Bologna*)
Pier Luigi Martelli (*Biocomputing Group, Department of Biology, University of Bologna*)
Piero Fariselli (*Biocomputing Group, Department of Biology, University of Bologna*)
Rita Casadio (*Biocomputing Group, Department of Biology, University of Bologna*)

Short Abstract: eESLB is the first database of eukaryotic proteomes fully annotated for subcellular localization, containing experimental annotations derived from primary protein databases, homology based annotations and computational predictions. In its first release, the database contains subcellular annotations of proteins from five entire eukaryotic proteomes. eSLDB is publicly available at http://gpcr.biocomp.unibo.it/esldb/.

Long Abstract:

MOTIVATIONS: Subcellular localization is a key feature for functional characterization of proteins. Experimental determination is an expensive and time consuming procedure, that up to now, has been achieved only for a small subset of the known proteins. Large-scale experiments have been carried out for determining the subcellular location of all the proteins of an organism.

Limitations in the experimental procedure allow, to date, to analyse only simple, unicellular species, such as Saccharomyces cerevisiae (Huh et al., 2003), and cannot be easily scaled up to more complex ones.

In this poster we present eSLDB, a database of protein subcellular localization, which aims to compensate this gap providing an annotation for the entire proteomes of eukaryotic organisms. The database contains the experimental localizations, when available, the homology-based annotations, when feasible, and predictions performed with machine learning based methods. Up to date we processed five proteomes: Homo sapiens, Mus musculus, Caenorhabditis elegans, Saccharomyces cerevisiae and Arabidospis thaliana.

RESULTS: The SwissProt annotations for subcellular localization of eukaryotes can be grouped into 16 major classes: Nucleus, Cytoplasm, Mitochondrion, Plastid, Golgi, Endoplasmic reticulum, Lysosome, Endosome, Vescicles, Peroxisome, Vacuole, Cell wall, Secretory pathway, Extracellular, Cytoskeleton and Membrane. Only 9% of all the SwissProt entries for the five species taken into account contains a record reporting the experimental subcellular localization. The rate of experimental annotation ranges from 27% of S. cerevisiae proteome to less than 3% for A. thaliana and C. elegans. In these cases eSLDB contains the annotation extracted from the SwissProt database using an automated tool that parses the SUBCELLULAR LOCALIZATION section of the COMMENT field. The words directly and/or implicitly referring to one of 16 classes are taken into account. Entries annotated as “probable”, “possible” or “by similarity” were not considered as experimental annotations.

All the proteins are then assigned using both sequence similarity and state-of-art predictors.

Based on the fact that proteins sharing high sequence identity usually have the same

subcellular localization (Pierleoni et al, 2006), we aligned each sequence with the experimentally annotated proteins belonging to the same kingdom (metazoa, fungi or viridiplantae). The annotation of the best scoring match having an E-value lower than 10-4, if existing, is then transferred to the query sequence. This procedure assigns localization to 45% of the proteins in the database. This rate ranges from 20% of C. elegans up to 62% for H. sapiens.
In order to annotate the rest of the proteins, that comprise the largest fraction of the proteome, a pipeline of predictors based on machine learning is needed. We used Spep (Fariselli et al, 2003) and ENSEMBLE (Martelli et al, 2003) for discriminating membrane proteins, then BaCelLo (Pierleoni et al, 2006) for assigning localization for soluble proteins. These are among the best available methods. To achieve a good reliability, the 16 original classes are reduced to 6 macro-classes: Membrane, Nucleus, Cytoplasm, Secretory pathway, Mitochondrion and Plastid. The decision tree structure of the prediction is also reported, deriving from the original output of the used methods.
All the data are available at our website and fully searchable either by sequence or protein name (ENSEMBL or TAIR) submission.

CONCLUSIONS: eSLDB is the only available database containing annotations for subcellular localizations of whole eukaryotic proteomes, that includes experimental data, homology-based annotations and predictions. Other available eukaryotic proteomes are currently under process and will be added to the database.
eSLDB is publicly available at http://gpcr.biocomp.unibo.it/esldb/.

REFS:

Fariselli, P., Finocchiaro, G. and Casadio, R. (2003) SPEPlip: the detection of signal peptide and lipoprotein cleavage sites. Bioinformatics, 19, 2498-2499.

Huh,W.K., Falvo,J.V., Gerke,L.C., Carroll,A.S., Howson,R.W., Weissman,J.S. and O'Shea,E.K. (2003) Global analysis of protein localization in budding yeast. Nature, 425, 686-91.

Martelli, P.L., Fariselli, P. and Casadio, R. (2003) An ENSEMBLE machine learning approach for the prediction of all-alpha membrane proteins. Bioinformatics, 19, i205-i211.

Pierleoni, A., Martelli, P.L., Fariselli, P. and Casadio, R. (2006) BaCelLo: a Balanced subCellular Localization predictor. Bioinformatics, in press. Accepted at ISMB 2006.