LocDB: experimental annotations of localization for *Homo sapiens* and *Arabidopsis thaliana*  

Shruti Rastogi¹,* and Burkhard Rost¹,²,³  

¹Department of Biochemistry and Molecular Biophysics, Columbia University, 701 West, 168th Street, New York, NY 10032, USA, ²Technical University Munich, Bioinformatics, Department of Computer Science and Institute of Advanced Studies (IAS) and ³New York Consortium on Membrane Protein Structure (NYCOMPS), TUM Bioinformatics, Boltzmannstr. 3, 85748 Garching, Germany  

Received August 15, 2010; Revised September 24, 2010; Accepted September 27, 2010  

ABSTRACT  
LocDB is a manually curated database with experimental annotations for the subcellular localizations of proteins in *Homo sapiens* (HS, human) and *Arabidopsis thaliana* (AT, thale cress). Currently, it contains entries for 19 604 UniProt proteins (HS: 13 342; AT: 6262). Each database entry contains the experimentally derived localization in Gene Ontology (GO) terminology, the experimental annotation of localization, localization predictions by state-of-the-art methods and, where available, the type of experimental information. LocDB is searchable by keyword, protein name and subcellular compartment, as well as by identifiers from UniProt, Ensembl and TAIR resources. In comparison to other public databases, LocDB as a resource adds about 10 000 experimental localization annotations for HS proteins and ~900 for AS proteins. Over 40% of the proteins in LocDB have multiple localization annotations providing a better platform for development of new multiple localization prediction methods with higher coverage and accuracy. Links to all referenced databases are provided. LocDB will be updated regularly by our group (available at: http://www.rostlab.org/services/locDB).  

INTRODUCTION  
Proteins are the fundamental functional components of the machinery of life. The particular cellular compartment, in which they reside, i.e. their native subcellular localization, is a key feature that characterizes their physiological functions. Many careful, hypothesis-driven experimental studies have been contributing to our large body of annotations of cellular compartments (1–5). Recently, high-throughput experiments have stepped up to the challenge to increase the amount of annotations (6–15). These data sets capture aspects of protein function and, more generally, of global cellular processes.

UniProt (release 2010_07) (16) constitutes the most comprehensive and, arguably, the most accurate resource with experimental annotations of subcellular localization. However, even this excellent resource remains incomplete for the proteomes from *Homo sapiens* (HS) and *Arabidopsis thaliana* (AT) of the 20 282 human proteins in Swiss-Prot (17), 14402 have annotations of localization (72%), but for only 3720 (18%) these annotations are experimental. Similarly, of the 9099 AT proteins only 1495 (17%) have experimental annotations of localization. While LocDB stands on and roots UniProtKB, it encompasses this giant and adds specific value by collecting information about subcellular localization from the primary literature and from other databases. These data are enriched by annotations, links and predictions.  

DATA SET  
Curated entries with experimental data  
LocDB contains experimental annotations for subcellular localization of 19 604 UniProt proteins; 13 342 of these are from *Homo sapiens* [10 102 Swiss-Prot and 3240 TrEMBL (17)] and 6262 from AT (3466 Swiss-Prot, 2796 TrEMBL). This raises the experimental annotations for human from 3720 (18%) to 13 342 (66%), and for thale cress from 1495 (16% of the UniProt subset of AT; note that this subset may constitute as little as 30% of all AT proteins) to 6262 (69% of the UniProt subset of AT). We classify all proteins according to the Gene Ontology (18) (GO) hierarchy into 12 primary classes of subcellular localization, i.e. use the following classes: cytoplasm, endoplasmic reticulum, endosome, extracellular, Golgi apparatus, mitochondrion, nucleus, peroxisome, plasma membrane, plastid, vacuole and vesicles (Table 1).
The proteins are further classified in subclasses of above primary classes denoted as secondary protein localizations, for example, protein RL21_HUMAN is experimentally annotated to be localized in primary: Nucleus and Secondary: Nucleolus.

Statistics

Each entry in LocDB has some experimental localization data. However, we have explicit annotations of a particular experiment type for only 25% of the entries. This is a work in progress as, curation is tedious and manual, and we are planning to update details regarding experiments with every new release of LocDB. Most annotations in LocDB are for the nucleus (20%), cytoplasm (20%) and the plasma membrane (20%).

Almost two in three of all HS proteins are annotated in one of the largest three compartments (23% nucleus, 25% cytoplasm, 20% plasma membrane). Similarly, two in three of the AT proteins fall into one of the compartments (28% plastid (incl. chloroplast), 21% plasma membrane, 13% nucleus). The distribution of proteins within each region is accessible from the LocDB statistics page http://www.rostlab.org/services/locDB/statistics.php.

Multiple localizations

Many proteins travel, i.e. they stay in more than one subcellular localization at one point of their ‘life’. Most proteins annotated by traditional detailed biochemical experiments, point to one single compartment as the major native environment of each protein (19). By contrast, most high-throughput experiments identify most proteins in more than one compartment. Clearly, high-throughput experiments are noisy. Nevertheless, are noisy large-scale experiments closer to the truth than small-scale approaches? The answer remains unclear. About 40% of the LocDB entries have experimental evidence for more than one localization. This may imply that 60% of all proteins are primarily native to a single compartment. In fact, previous analyses suggest a similar value (19). However, this does not imply that only 40% of the proteins ever ‘travel through’ more than one compartment: many traveling proteins are likely not captured in the experimental data due to limited coverage and limitations in the experimental resolution (false negatives). On the other hand, some fraction of this 40% of proteins evidenced in several localizations may also indicate experimental errors (false positives). It remains unclear how to weigh those effects.

Most proteins unique

LocDB also clusters proteins into families or groups of related proteins (Figure 1). For instance, 1160 (8%) of all HS proteins and 74 (1%) of all AS proteins have

![Figure 1. Clustering of LocDB. We clustered the LocDB entries by BLASTclust (26) to explore whether or not some families are highly over-represented in LocDB, and found that they are not. For instance, 46% of HS and 43% of AT proteins in LocDB have levels of PIDE <25%, i.e. differ substantially in sequence. On the other end of the spectrum, only 8% of HS and 1% of AT proteins are very similar to each other (PIDE >98%). Note that levels of PIDE>70% usually suffice to infer similarity in localization at levels of about 75% (31), i.e. for over 80% of the LocDB entries no other entry could be used to predict localization by homology.]
more than 98 percentage pair wise sequence identity (PIDE) to another protein in the data set. Clustering at PIDE < 25% yields 5587 proteins in HS (42%) and 2744 proteins in AS (47%). This implies that conversely about 7755 proteins annotated in HS and 3518 in AT are sequence-unique at the 25% PIDE threshold. The percentage of proteins with multiple localizations is higher when considering sequence-unique subsets, e.g. while 40% of all proteins are annotated with multiple localizations, 4.6% of those clustered at 98% PIDE and 45% of those clustered at 25% PIDE.

**Experimental and predicted localization**

Each LocDB entry corresponds to one protein, and contains protein identifiers, experimental annotations of protein localization, types of experiments performed and the respective publication PubMed (20) identifiers, as well as predicted localization annotations from LOCtree (19), WOLFPSORT (21), MultiLoc (22), TargetP (23), PredictNLS (24) and NuPred (25). Prediction results are given in both basic and detailed formats along with the respective reliability and probability scores (Figure 2).

**Data mining from primary literature**

Data for LocDB are collected from reports of many low- and high-throughput experiments. Citations to the appropriate experiments are displayed on the LocDB protein entry pages. Protein sequences and identifiers from the experimental papers are extracted and BLASTed (26) against UniProt. The sequences with ≥ 98% PIDE over the entire sequence are assigned UniProt and Ensembl (27) identifiers for HS and TAIR (28) identifiers for AT.
Data mining from external databases

Data are also mined from external databases, e.g. LOCATE (1), SUBA (4) and many other resources. LocDB reports all the references with the entries in the database which link directly to their PubMed (20) abstracts.

Comparison with other resources

Many excellent subcellular localization resources are available with experimental annotations of proteins for HT and AT such as LOCATE (1) for HT and SUBA (4) for AT. The comparison and overlap between these resources together with UniProt release (2010_07) are shown in Figure 3a and b. In addition, the comparison in number of proteins annotated in various compartments in these resources is shown in Table 1. These comparisons show that we have added ∼10000 human protein localization annotations and ∼900 Arabidopsis protein localization annotations over LOCATE, SUBA and UniProt.

As mentioned above, UniProt database contains both experimental and general annotations such as ‘Probable’, ‘By similarity’ and ‘Potential’ for protein subcellular locations. A very high level of discrepancies is found in the annotations for locations involved in secretory pathway such as Golgi apparatus, endoplasmic reticulum etc., especially in human proteins (shown in Figure 1a and b in Supplementary Data). In Arabidopsis, there is high discrepancy in all the compartments except nucleus and plastid. Comparison with databases DBSubLoc (29) and eSLDB (30) is also done; however, they are not shown as the annotations in these resources are mostly derived from Swiss-Prot database.

LocDB will be updated once every 3 months. There is also a provision for users to contribute to the resource by adding information on the contribution page of website as well as by sending an email to locdb@rostlab.org, if they come across any inaccuracies. We will use the database as a portal to access state-of-the-art prediction methods, which will enable users and developers to test prediction methods. We will also add predictions for proteins without experimental annotations that will be clearly marked as predictions. More eukaryotic and prokaryotic proteomes will be available in future through the database such as Escherichia coli and yeast. Moreover, we plan to add curated protein expression data and protein–protein interaction data in the following versions of locDB.

Availability

LocDB data can be retrieved as individual entries or downloaded as HTML and text files from http://www.rostlab.org/services/locDB. The database is a MySQL database and can be obtained upon request (locdb@rostlab.org) as an SQL file.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We are pleased to thank Amos Bairoch (SIB, Geneva), Rolf Apweiler (EBI, Hinxton), Phil Bourne (San Diego University) and their crew for maintaining excellent databases. Furthermore, thanks to all experimentalists who enabled this analysis by making their data publicly available.

FUNDING

Funding for open access charge: The National Institute of General Medical Sciences (NIGMS; grant R01-GM079767) at the National Institutes of Health (NIH).

Conflict of interest statement. None declared.

REFERENCES
