



Fig. 3. Performance improves with number of keywords. (a) Keyword distribution in test set: Most test proteins had 2–5 keywords. (b) Performance as a function of keywords: The prediction accuracy and coverage were both nearly 100% for proteins with more than 30 keywords. The coverage (thin line) tends to increase with the number of keywords. The accuracy was observed to decrease first (thick line) before increasing.

Table 3. Automatically annotating sub-cellular localization for five proteomes

Organism	Nprot ^a	OneKey ^b	LOCKey ^c	Homology ^d	signalP ^e	predictNLS ^f
<i>Arabidopsis thaliana</i> (plant)	25456	6703	3598	1961	100	16
<i>Caenorhabditis elegans</i> (worm)	18898	3584	1999	1240	60	22
<i>Drosophila melanogaster</i> (fly)	14184	4010	2430	1501	66	24
<i>Homo sapiens</i> (human, partial)	31073	16522	10174	6057	100	23
<i>Saccharomyces cerevisiae</i> (yeast)	6306	3691	1747	837	3	20
SUM	95917	34510	19948	11596		

^aNprot: Number of proteins in proteome; ^bOneKey: Number of proteins with at least one keyword in SWISS-PROT that matches our trusted vectors (System); ^cLOCKey: number of proteins for which LOCKey inferred sub-cellular localization in ten classes (Table 1; note: these results were obtained using the entropy thresholds that gave 87% testing accuracy, Figure 2); ^dHomology: sub-cellular localization inferred using homology, i.e. sequence similarity to proteins of known localization taken from SWISS-PROT (at a threshold of HSP-distance > 15; at this distance the assignment through homology yielded levels around 90% accuracy, Nair and Rost, unpublished); ^esignalP: percentage of predicted extra-cellular proteins also predicted to contain a signal peptide (Nielsen *et al.*, 1997); ^fpredictNLS: percentage of predicted nuclear proteins also predicted to have a nuclear localization signal (Cokol *et al.*, 2000). Note that LOCKey enabled to annotate 8352 eukaryotic proteins of unknown localization (19 948–11 596).

the major source of error in predicting nuclear and extra-cellular proteins. One reason could be that experimental annotations are less accurate for cytoplasmic proteins. Another reason could be that proteins do in fact shuttle between the cytoplasm and other localizations and that our ‘errors’ really captured proteins that could also occur in the predicted class. This interpretation was somewhat supported by the finding that LOCKey often found the correct class in the first two hits. In other words, when replacing the binary classification accuracy (a protein can only be in one single localization) by a probabilistic measure (one protein can be in many compartments), LOCKey appeared more accurate.

We applied LOCKey to five (yeast, worm, fly, human, and arabidopsis) entirely sequenced eukaryotic proteomes. We could infer localization for over 8300 proteins for which localization could not have been detected by any other automatic system. Three types of methods can infer or predict localization in the context of entire proteomes: (1) homology to proteins of known localization, (2) detection of sequence motifs, and (3) prediction from sequence and structure. In our group, we simultaneously work on all these types of methods. LOCKey is most relevant for the coverage achieved by homology-based methods, since it allows one to automatically increase the data set of proteins of known localization for which we can apply