



Inferring sub-cellular localization through automated lexical analysis

Rajesh Nair^{1, 2} and Burkhard Rost^{1, 3,*}

¹CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 650 West 168th Street BB217, New York, NY 10032, USA, ²Department of Physics, Columbia University, 538 West 120th Street, New York, NY 10027, USA and ³Columbia University Center for Computational Biology and Bioinformatics (C2B2), Russ Berrie Pavilion, 1150 St. Nicholas Avenue, New York, NY 10032, USA

Received on January 24, 2002; revised and accepted on March 29, 2002

ABSTRACT

Motivation: The SWISS-PROT sequence database contains keywords of functional annotations for many proteins. In contrast, information about the sub-cellular localization is available for only a few proteins. Experts can often infer localization from keywords describing protein function. We developed LOCKey, a fully automated method for lexical analysis of SWISS-PROT keywords that assigns sub-cellular localization. With the rapid growth in sequence data, the biochemical characterisation of sequences has been falling behind. Our method may be a useful tool for supplementing functional information already automatically available.

Results: The method reached a level of more than 82% accuracy in a full cross-validation test. Due to a lack of functional annotations, we could infer localization for fewer than half of all proteins in SWISS-PROT. We applied LOCKey to annotate five entirely sequenced proteomes, namely *Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fly), *Arabidopsis thaliana* (plant) and a subset of all human proteins. LOCKey found about 8000 new annotations of sub-cellular localization for these eukaryotes.

Availability: Annotations of localization for eukaryotes at: <http://cubic.bioc.columbia.edu/services/LOCKey>.

Contact: nair@cubic.bioc.columbia.edu;
rost@columbia.edu

Keywords: genome sequence analysis; predicting sub-cellular localization; protein function; lexical analysis.

INTRODUCTION

Protein sequence-function gap

The number of completely sequenced genomes has been rapidly increasing. Currently, we know the full genomes for over 60 organisms (Adams *et al.*, 2000; Arabidopsis

Genome Initiative, 2000; Fleischmann *et al.*, 1995; Frishman, 2000; Goffeau *et al.*, 1996; Liu and Rost, 2000; The *C. elegans* Sequencing Consortium, 1998). This sequence explosion has widened the gap between the number of sequences deposited in public databases and the experimental characterisation of the corresponding proteins (Koonin, 2000). To bridge this gap, faster and more effective means of creating annotation are required (Baker and Brass, 1998; Eisenberg *et al.*, 2000; Fleischmann *et al.*, 1999; Lewis *et al.*, 2000). One promising approach is use of automatic annotations (Apweiler *et al.*, 1997; Gaasterland, 1996; Kretschmann *et al.*, 2001). One step towards understanding protein function is elucidating its sub-cellular localization (Eisenhaber and Bork, 1998).

Database annotations of function often very detailed

Protein function may be described best in the context of molecular interactions. The SWISS-PROT database (Bairoch and Apweiler, 2000) contains functional annotations predominantly at a very detailed level of biochemical function, e.g. a protein may be annotated as a cdc2 kinase, but not as being involved in intra-cellular communication (Apweiler, 2001; Tamames *et al.*, 1998). We would like to complement these detailed annotations with descriptions in context of higher-order processes such as the regulation of gene expression, pathways, or signalling cascades (Bork *et al.*, 1998; Riley, 1993; Riley and Labedan, 1997; Tamames *et al.*, 1998). Descriptions at this level are available for only a few proteins. To remedy this situation, a number of automatic and semi-automatic tools have been developed for functional annotation of proteins.

Classification of proteins into families of homologues

Many automatic annotations are derived from sequence similarity to proteins of known function (Andrade *et al.*, 1999; Bork and Gibson, 1996; Casari *et al.*, 1995; Devos

*To whom correspondence should be addressed.

and Valencia, 2001; Fleischmann *et al.*, 1995; Koonin, 2000; Remm *et al.*, 2001; Tatusov *et al.*, 2000). A typical similarity search starts by aligning the unknown protein *U* against databases with functional annotations through search tools such as BLAST (Altschul *et al.*, 1990), FASTA (Pearson and Lipman, 1988), or PSI-BLAST (Altschul *et al.*, 1997). If a homologue *H* is found that has an annotation and significant sequence similarity to *U*, the annotation of *H* is transferred to *U*. Such inference of function is reliable only if the sequences of *U* and *H* are very similar (Devos and Valencia, 2001; Rost, 2001). Several pitfalls of such transfers of function have been reported, e.g. inadequate knowledge of thresholds for ‘significant sequence similarity’, or using only the best database hit or ignoring the domain organisation of proteins (Bork and Koonin, 1998; Devos and Valencia, 2001; Doerks *et al.*, 1998; Galperin and Koonin, 2000).

Automatic annotation through functional descriptors

Annotation systems have been based on SWISS-PROT keywords (Bairoch and Apweiler, 2000; Eisenhaber and Bork, 1998; Fleischmann *et al.*, 1999; Hofmann *et al.*, 1999; Kretschmann *et al.*, 2001; Tamames *et al.*, 1998, 1996). The common approach to classifying function is to first extract characteristic keywords for each of the functional classes from a set of proteins classified by experts. Using these keywords, a library of rules is created that associates a certain pattern of keywords to a functional class. Creating the ‘rules library’ is a difficult task for which a variety of solutions have been proposed. EUCLID (Tamames *et al.*, 1998) uses SWISS-PROT keywords to classify proteins into 14 classes of cellular function (according to the scheme proposed by Monika Riley (Karp *et al.*, 1999; Krawiec and Riley, 1990; Riley, 1993; Riley and Labedan, 1997)). Using a simple voting scheme, the system assigns the unknown sequence to the functional class to which the majority of its keywords belong (Tamames *et al.*, 1996). A disadvantage of using such dictionaries is that they can only ‘discover’ simple correlations among the known functional keywords. The method of the Apweiler group (Fleischmann *et al.*, 1999) annotates function for the TrEMBL (Bairoch and Apweiler, 2000) database based on SWISS-PROT keywords and PROSITE motifs (Hofmann *et al.*, 1999). The system generates a ‘RuleBase’ by extracting functional annotations from all SWISS-PROT proteins that contain the same PROSITE motif. If a PROSITE motif is discovered in an un-annotated TrEMBL sequence, the functional annotation is transferred from the ‘RuleBase’. Recently (Kretschmann *et al.*, 2001), the group has implemented the C4.5 data-mining algorithm to automatically generate rules for keyword annotations found in SWISS-PROT. The rules are based on taxonomy, PROSITE motifs,

and PFAM patterns in SWISS-PROT proteins belonging to different InterPro (Apweiler *et al.*, 2000) families. The Meta_A annotator (Eisenhaber and Bork, 1998) is a partly automatic annotation evaluation system based on a combination of lexical analysis and libraries of expert rules. The rule libraries are derived from scanning the protein names, taxonomy information, commentaries and feature tables in SWISS-PROT. The system assigns one of twelve final sub-cellular localizations to each protein. Meta_A combines primary attributes with AND, OR and NOT logical operators to create a library of ‘biological rules’. The rules relating lexical patterns with functional attributes are created by expert intervention, i.e. resemble a dictionary (Eisenhaber and Bork, 1998). Thus the creation of the rule library is time-consuming and has to be repeated for each new application.

Algorithms for text categorisation

The problem of automatically extracting rules from SWISS-PROT keywords has parallels to the problem of ‘Text Categorisation’. Text categorisation (TC) is the problem of assigning predefined categories to text documents such as journal articles or abstracts. Many statistical learning methods have been applied to this problem. These include nearest neighbour classifiers (Yang and Pederson, 1997), multivariate regression models (Schutze *et al.*, 1995; Yang and Chute, 1992), probabilistic Bayesian models (Lewis and Ringuette, 1994) and symbolic rule learning (Apte, 1994). M-ary (multiple category) classifiers like the k-Nearest Neighbour (Dasarathy, 1991) and the Linear Least-squares Fit (LLSF) (Yang and Liu, 1999). Here we describe LOCKey, a novel method automatically assigning proteins to classes of sub-cellular localization based on a lexical analysis of SWISS-PROT keywords. LOCKey is based on M-ary classifiers that solve the classification problem accurately when the number of data points (proteins) and dimensionality of the feature space (number of keywords) are not too large. In contrast to dictionary-based approaches, LOCKey is fully automated and the rule libraries are generated dynamically. Our method may be applied to any database with keywords of functional information and to any task involving higher-level classifications.

SYSTEM AND METHODS

Implementation of algorithm

Instead of creating an *a priori* ‘rule library’, we generated all possible sets of rule libraries for a protein of unknown ‘class’ from a set of SWISS-PROT keywords. The protein was assigned based on the ‘rule library’ that solved the classification problem best. We applied the algorithm to infer one of ten classes of sub-cellular localization (Table 1). The algorithm involved two separate steps:

(1) build a data set of trusted vectors from proteins of known localization, and (2) classify unknown proteins (Figure 1).

Step 1: Building data set of trusted vectors

First, we compiled a data set with proteins of experimentally known localization. Then we extracted a list of keywords from SWISS-PROT for this set from the ‘keyindex’ file. Since only a partial functional annotation was available for a large number of proteins, we merged keywords from homologous sequences, to provide as complete an annotation as possible. In particular, we identified all SWISS-PROT sequences with HSP distances (equation 1) >15 to sequences in the sequence-unique subset (below) using BLAST (Altschul *et al.*, 1997; Altschul and Gish, 1996) and extracted their keywords. Finally, we built a data set of binary vectors (Salton, 1989) for these keywords that represented the presence of a certain keyword by 1 and the absence by 0. To reduce the dimensionality of feature space, we retained only keywords with ‘above random’ classifying ability based on an entropy (equation 2) and normalized entropy (equation 3) cut-off (Schutze *et al.*, 1995). The accuracy vs. coverage plots were not very sensitive to the particular cut-off chosen (plots given at cubic.columbia.edu/services/LOCKkey). We merged these keywords with the keywords found for the corresponding protein from the sequence-unique set. Most proteins had 2–5 keywords (Figure 3a).

Removing ‘trivial’ keywords

To evaluate the ability of the system to discover non-trivial correlations between variables, we excluded all keywords from the vector set that were biologically correlated to localization (e.g. ‘DNA-binding’). We also excluded keywords that were observed to occur more than 90% of the time in proteins within a single sub-cellular localization. Thus, we excluded 81 keywords from our trusted vector set. Removing these keywords resulted in being unable to identify any keyword for 176 test proteins. To minimize the effects of annotation errors, we retained only keywords that occurred in at least 10 protein families.

Step 2: Classifying proteins of unknown localization

To infer the localization of a protein U of unknown localization, we first retrieved all keywords for U from SWISS-PROT that matched in our ‘trusted vector set’ of informative keywords. Thus, we retrieved a vector $V(U)$ that had the same dimension as the vectors in the ‘trusted set’. Next, we generated all possible alternatives to $V(U)$ for which one or many 1’s were flipped to 0’s. For example, for a protein with 3 keywords, we generated $2^3 - 1 = 7$ sub-vectors $V'(U)$: 111, 110, 101, 011, 100, 010 and 001. These sub-vectors constituted all possible keyword combinations for protein U . The final

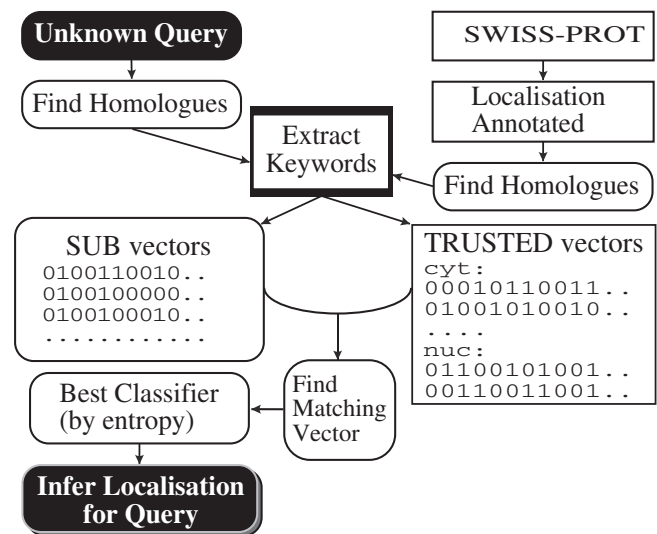


Fig. 1. LOCKkey algorithm. First, we compiled a sequence-unique data set of proteins of experimentally known sub-cellular localization. For these proteins, we then extracted keywords from SWISS-PROT. Next, we merged keywords found in homologues. We represented the keywords found in the proteins of known localization as binary vectors in the ‘Trusted Vector Set’. For proteins of unknown localization U , the goal now became to compare these to the ‘Trusted Vector Set’. Toward this end, we first identified all keywords in U and in homologues of U . Then we constructed all possible keyword combinations (SUB vectors), and compared these to the ‘Trusted’ vectors. We found the best matching vector based on entropy criteria (Methods). Finally, we used this ‘best matching vector’ to infer localisation for the query.

task was to find the keyword combination that yielded the best classification of U into one of ten classes of sub-cellular localizations, i.e. was most similar to one of the ‘trusted vectors’. To achieve this, we retrieved all exact matches of any of the sub-vectors $V'(U)$ to any of the proteins in the trusted vectors, i.e. found all proteins in the trusted set containing one of the keywords found in U . By construction of the sub-vectors, the proteins retrieved in this way may also contain keywords not found in U . Next, we simply counted how often the proteins retrieved belonged to a particular class $C(i)$, $i = 1 \dots 10$. We repeated this for each of the sub-vectors $V'(U)$, and selected the localization that was finally assigned by minimizing an entropy-based objective function (‘prediction mode’).

Data sets

We selected all proteins with unambiguously annotated sub-cellular localization in SWISS-PROT release 40 (Bairoch and Apweiler, 2000). We excluded sequences annotated as ‘POSSIBLE’, ‘PROBABLE’, ‘SPECIFIC

Table 1. Number of proteins in ‘trusted’ data sets

Sub-cellular localization	SWISS-PROT ^a	Sequence-unique ^b
Nucleus	3478	922
Extra-cellular space	2900	724
Cytoplasm	2642	544
Mitochondria	1743	467
Chloroplast	1648	197
Endoplasmic reticulum	568	108
Peroxisome	177	55
Golgi apparatus	167	52
Lysosome	163	49
Vacuolar	103	28
SUM (all 10)	13589	3146

^aNumber of proteins with known localization found in SWISS-PROT;

^bNumber of sequence-unique proteins, i.e., representative subset of all SWISS-PROT proteins found (Methods). Note: we used the sequence-unique set as test set.

PERIODS’ or ‘BY SIMILARITY’, and proteins with multiple annotations of localization. This left 13589 proteins in the ‘Experimental data set’ (Table 1). To reduce bias, we built a representative subset of sequence-unique proteins by using a simple greedy algorithm (Hobohm *et al.*, 1992). In particular, we accepted only pairs with an HSSP-distances below 15 (Rost, 1999; Sander and Schneider, 1994):

$$\begin{aligned} \text{HSSP DISTANCE} &= \text{PIDE} - \text{HSSP_PIDE} & (1) \\ \text{HSSP_PIDE} &= \begin{cases} 100, & \text{for } L \leq 11 \\ 480 \cdot L^{-0.32 \cdot \{1 + e^{-L/1000}\}}, & \text{for } L \leq 450 \\ 19.5, & \text{for } L > 450 \end{cases} \end{aligned}$$

where PIDE is the percentage of pairwise identical residues and L is the alignment length. The final sequence-unique subset of known localization contained 3146 proteins. For entire-proteome predictions, we obtained the sequences with their alignments to SWISS-PROT proteins from <http://cubic.bioc.columbia.edu/genomes> (Liu and Rost, 2000).

Sorting assignments by keyword entropy

For each of the remaining keywords, we calculated the Shannon Information SI (Shannon, 1951) according to:

$$SI = - \sum_{i=1}^N P_i \log P_i \quad (2)$$

where N is the number of localization classes (10) and P_i are the probabilities of finding the keyword in one of the 10 classes of localization. Since the Shannon Information does not take into account the background

distribution of proteins among the various localizations, we calculated a normalized Shannon Information $normSI$ for each keyword:

$$\begin{aligned} normSI &= - \sum_{i=1}^M N_i \log N_i \\ N_i &= X_i / \sum_{i=1}^M X_i, \end{aligned} \quad (3)$$

where X_i was the fraction of proteins belonging to a given localization identified by the keyword and M the number of localizations in which the keyword was found. Finally, we defined the percent of fractional change in SI (and $normSI$) as:

$$\begin{aligned} fracSI &= 100 \cdot \frac{maxSI - SI}{maxSI} \\ fracNormSI &= 100 \cdot \frac{maxNormSI - normSI}{maxNormSI} \end{aligned} \quad (4)$$

where $maxSI$ is the maximum possible Shannon Information and $maxNormSI$ the maximum possible normalized Shannon Information. We included only those keywords in our set of trusted vectors that satisfied the criteria: $fracSI > 25$ and $maxNormSI > 25$.

Inferring localization from keywords (prediction mode)

To infer localization for test proteins we identified the keyword combinations that maximized $fracSI$ and $maxNormSI$ (equation 4). Predictions were made only if at least one keyword combination could be found such that $fracSI > 70$ and $maxNormSI > 70$. Additionally, we required that the keyword combination was present in at least five families in the training set.

Evaluating performance accuracy

We evaluated performance by a five-fold cross-validation experiment, i.e., we partitioned the sequence-unique subset into five sets. Then we used four of the five sets to generate the data set of trusted keyword vectors (training set), and inferred (predicted) localization for the remaining fifth set (test set). We repeated this procedure five times such that each of the sets was used for testing once. The final levels of accuracy and coverage constituted averages over all five tests. The partitioning was performed using a greedy clustering algorithm starting with the largest and longest families (Hobohm *et al.*, 1992). An HSSP distance = 5 (equation 1) was chosen for the clustering. This ensured that two 100 residue long sequences chosen from the training and test set have fewer than 35% pairwise identical residues.

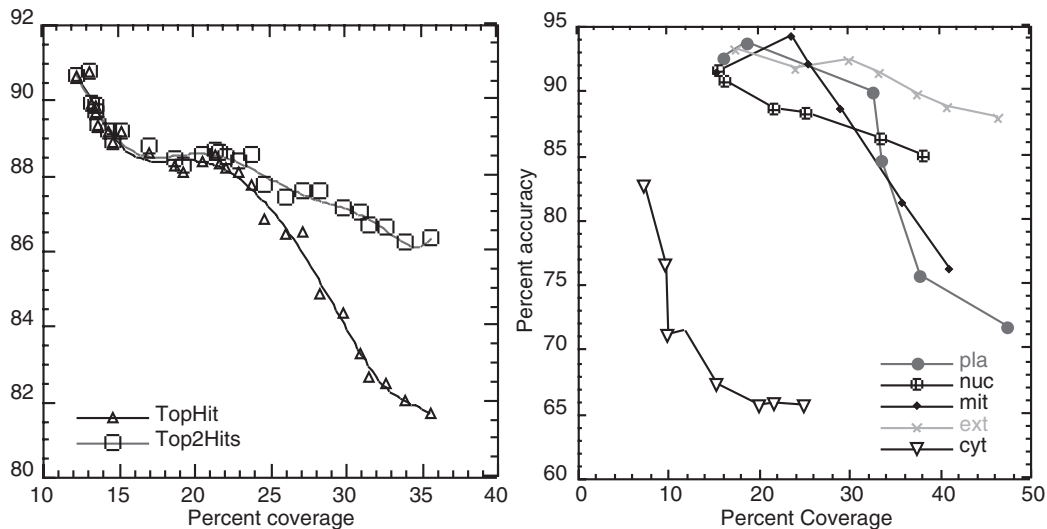


Fig. 2. Results for five-fold cross-validation. (a) Average over 10 classes. Results were obtained for a sequence-unique set. Keywords thought to be biologically correlated with sub-cellular localization and those observed to occur with high specificity in a single localization were excluded (Methods). The bold line represents accuracy versus coverage for the predicted localization. For example, at 25% coverage the system approaches a prediction accuracy of 87%. Prediction accuracy appeared higher, when we considered our findings correct if the correct localization was one of top two predicted localizations (grey line). (b) Major classes. Nuclear, extra-cellular and mitochondrial classes showed similar accuracy versus coverage statistics. Cytoplasmic and chloroplast proteins were predicted with a much lower accuracy.

RESULTS AND DISCUSSION

LOCKey yielded high accuracy but low coverage

We tested LOCKey in a five-fold cross-validation experiment (Methods). Some SWISS-PROT keywords are such that all biologists would immediately know the localization of the respective protein (e.g. DNA-binding), others need more expertise. To assess the ability of the system to discover non-trivial correlations between variables, we excluded all keywords predominantly associated with a single localization. This ‘filtering out of the most obvious’ required removing 81 keywords from the ‘trusted vectors’. For 176 of the proteins in the test set, we found no keywords. When choosing the entropy cut-off such that we could classify one fourth of all proteins (coverage in Figure 2a), our system correctly inferred one of ten classes of sub-cellular localization for 87% of all proteins (accuracy in Figure 2a). We noticed that the top two hits contained the observed localization for a similar level of accuracy (87%) at an entropy cut-off at which we assigned localization for about 35% of all proteins (dashed line in Figure 2a). Interestingly, this improved performance when considering the top two hits, was mostly due to proteins from the chloroplast that were often confused with mitochondrial proteins. The reasons for the low coverage of the system were manifold. First, some proteins had no keyword (176 of 3146). Second, we required that the keyword pattern was present in at least five proteins in the

vector set. Third, the vector set was too small (3146 proteins) to provide a good sample for all proteins. In other words, many keywords found in the testing set were not present in the ‘training’ set.

Performance varied substantially between classes

LOCKey was more successful for some classes than for others. In particular, extra-cellular, nuclear and mitochondrial proteins could be inferred more reliably than the other classes (Figure 2b). On the other hand, performance was worst for proteins from the cytoplasm and chloroplasts (Figure 2b). When we considered our findings correct if any of the top two hits was predicted in the observed localization, we noticed a considerable improvement for most of the major classes (data not shown). The only exceptions were cytoplasmic proteins for which the keywords yielded levels above 75% accuracy at entropy thresholds corresponding to levels of coverage around 10%. The detailed ‘confusion matrix’ (Table 2) revealed that cytoplasmic proteins were most often confused with nuclear proteins, and proteins from the chloroplast were most often assigned incorrectly to mitochondria. Although the minor classes (Golgi, Endoplasmic reticulum, peroxisome, vacuoles, and lysosome) contained too few proteins to allow statistically significant conclusions, we noted that proteins from vacuoles and the lysosome were mostly confused with extra-cellular proteins (Table 2).

Table 2. ‘Confusion Matrix’ for LOCKey

Prd ^a ⇒ Obs ^b ↓	nuc	ext	cyt	mit	pla	ret	oxi	gol	lys	vac	SUM ^{obs}
nuc	340	5	4	1	0	0	0	0	0	0	350
ext	5	321	11	0	0	0	0	0	0	0	337
cyt	37	25	59	14	1	2	0	0	0	0	138
mit	10	0	5	151	20	4	0	0	0	0	190
pla	5	1	5	24	56	1	0	0	0	0	92
ret	1	4	2	1	1	2	0	2	0	0	13
oxi	2	0	1	5	0	1	0	0	0	0	9
gol	0	2	2	0	0	0	0	17	0	0	21
lys	0	5	0	1	0	0	0	0	0	0	6
vac	0	2	1	1	0	0	0	0	1	0	5
SUM ^{prd}	400	365	90	198	78	10	0	19	1	0	1161

^aPrd: predicted localization; ^bObs: annotated localization; Abbreviations for localizations: nuc: nucleus; ext: extra-cellular space; cyt: cytoplasm; mit: mitochondria; pla: chloroplast; ret: Endoplasmic reticulum; oxi: peroxysome; gol: Golgi apparatus; lys: lysosome; vac: vacuoles. The numbers give the proteins used in the five-fold cross-validation experiment for which LOCKey assigned any localization (correct classifications in bold letters).

Performance improves with number of keywords

LOCKey inferred the correct localization for almost all proteins for which we had many keywords. In fact, both accuracy and coverage approached 100% for proteins with more than 25 keywords (Figure 3b). For proteins with few keywords, the accuracy decreased with increasing coverage (Figure 3b). The improved coverage with increasing number of keywords was the result of discovering new keyword combinations that meet the entropy criteria (Methods). Since the keywords were non-specific to any localization class, the initial drop in accuracy was due to a larger fraction of keyword combinations that closely met the entropy criteria. These combinations were more often incorrectly predicted.

Annotating entire proteomes

Using LOCKey, we could provide sub-cellular localization annotations to 38–55% more proteins than by simple annotation transfer using homology (Table 3). To provide independent confirmation for our annotations, we checked the annotations inferred for nuclear proteins for which we found nuclear localization signals (NLS; Cokol *et al.*, 2000) and extra-cellular proteins with predicted signal peptides (Nielsen *et al.*, 1997). More than 20% of all nuclear proteins identified by LOCKey (~9600) contained known NLSs (Table 3). All 3130 extra-cellular proteins identified by LOCKey in human and arabidopsis and over 60% of those in fly and worm contained signal peptides (Table 3). In contrast, only 2 of the predicted 76 extra-cellular proteins in yeast contained predicted signal peptides. Note that the numbers of proteins with signal peptides or NLSs that were not identified by LOCKey are not relevant, since neither SignalP nor our NLS database find all extra-cellular or nuclear proteins at 100%

accuracy. The relevant number in this context were the values for accuracy vs. coverage for the cross-validation experiment (Figure 2). Unfortunately, the only way to assess whether or not LOCKey correctly identified nuclear and extra-cellular proteins without motifs is to await the respective experiments. If we can generalize the levels of accuracy found in the cross-validation experiment, we conclude that LOCKey indeed identified many proteins that could not have been classified reliably by motif-based methods.

LOCKey assignments were often not trivial

Experts can often assign localization from SWISS-PROT keywords. This may be impractical in the context of assigning localization for entire proteomes. However, when analysing the assignments from LOCKey in more detail (<http://cubic.bioc.columbia.edu/services/LOCKey/>), we found that often the biologists in our lab could not clearly infer localization from the SWISS-PROT keywords. In other words, LOCKey ‘discovered’ relations that require more specific expertise than we can expect from a typical expert annotator.

CONCLUSIONS AND FUTURE WORK

LOCKey automatically provided high quality annotations of sub-cellular localization from SWISS-PROT keywords. However, for most test proteins, we could not infer localization, at all. This low coverage originated from a lack of relevant functional information for many proteins. One solution to this problem could be to extract keywords from bibliography databases such as MEDLINE (Andrade and Valencia, 1998).

Of all the major localization classes, cytoplasmic proteins were predicted worst (Table 2); these were also

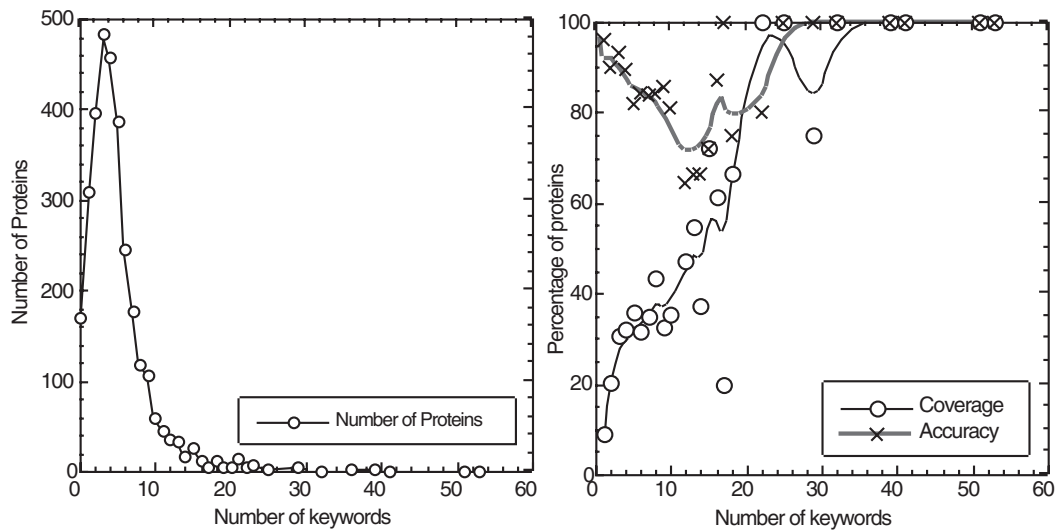


Fig. 3. Performance improves with number of keywords. (a) Keyword distribution in test set: Most test proteins had 2–5 keywords. (b) Performance as a function of keywords: The prediction accuracy and coverage were both nearly 100% for proteins with more than 30 keywords. The coverage (thin line) tends 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
<i>SUM</i>	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

homology thresholds (Nair & Rost, unpublished).

The PERL-code (Wall and Schwartz, 1990) of LOCKey was optimized to provide fast annotations. Annotating the entire *C. elegans* proteome took less than four hours on a PIII 900 MHz machine. The algorithm is limited to problems with few data points in the vector set ($n \ll 1\,000\,000$) and with few keywords ($n \ll 10\,000$).

Since the algorithm was not tailored to inferring sub-cellular localization, we are currently implementing the same idea to recognize distant similarities to proteins of known structure, i.e. to the problem of fold recognition. Our preliminary results are encouraging.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Columbia) for computer assistance and the genome data; to Dariusz Przybylski (Columbia) and Trevor Siggers (Columbia) for helpful discussions and to Kazimierz Wrzeszczynski (Columbia) and Henry Bigelow (Columbia) for valuable comments on the manuscript. We also thank the undisclosed reviewers for their helpful comments. The work was supported by the grants 1-P50-GM62413-01 and RO1-GM63029-01 from the National Institute of Health. Last, not least, thanks to all those who deposit their experimental data in public databases, and to those who maintain these databases.

REFERENCES

- Adams,M.D. *et al.* (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.
- Altschul,S., Madden,T., Shaffer,A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D. (1997) Gapped Blast and PSI-Blast: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Altschul,S.F. and Gish,W. (1996) Local alignment statistics. *Meth. Enzymol.*, **266**, 460–480.
- Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Andrade,M.A., Brown,N.P., Leroy,C., Hoersch,S., de Daruvar,A., Reich,C., Franchini,A., Tamames,J., Valencia,A., Ouzounis,C. and Sander,C. (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15**, 391–412.
- Andrade,M.A. and Valencia,A. (1998) Automatic extraction of keywords from scientific text: application to the knowledge domain of protein families. *Bioinformatics*, **14**, 600–607.
- Apte,C., Damerau,F. and Weiss,S. (1994) Towards language independent automated learning of text categorization models. *Proceedings of the 17th Annual ACM/SIGIR conference*.
- Apweiler,R. (2001) Functional information in SWISS-PROT: the basis for large-scale characterisation of protein sequences. *Brief Bioinform.*, **2**, 9–18.
- Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D., Durbin,R., Falquet,L., Fleischmann,W., Gouzy,J., Hermjakob,H., Hulo,N., Jonassen,I., Kahn,D., Kanapin,A., Kavavopoulou,Y., Lopez,R., Marx,B., Mulder,N.J., Oinn,T.M., Pagni,M., Servant,F., Sigrist,C.J. and Zdobnov,E.M. (2000) InterPro—an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.
- Apweiler,R., Gateau,A., Contrino,S., Martin,M.J., Junker,V., O'Donovan,C., Lang,F., Mitaritonna,N., Kappus,S. and Bairoch,A. (1997) Protein sequence annotation in the genome era: the annotation concept of SWISS-PROT+TREMBL. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 33–43.
- Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, **408**, 796–815.
- Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
- Baker,P.G. and Brass,A. (1998) Recent developments in biological sequence databases. *Curr. Opin. Biotechnol.*, **9**, 54–58.
- Bork,P., Dandekar,T., Diaz-Lazcoz,Y., Eisenhaber,F., Huynen,M. and Yuan,Y. (1998) Predicting function: from genes to genomes and back. *J. Mol. Biol.*, **283**, 707–725.
- Bork,P. and Gibson,T.J. (1996) Applying motif and profile searches. *Meth. Enzymol.*, **266**, 162–184.
- Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences—where are the bottlenecks? *Nature Genet.*, **18**, 313–318.
- Casari,G., Andrade,M.A., Bork,P., Boyle,J., Daruvar,A., Ouzounis,C., Schneider,R., Tamames,J., Valencia,A. and Sander,C. (1995) Challenging times for bioinformatics. *Nature*, **376**, 647–648.
- Cokol,M., Nair,R. and Rost,B. (2000) Finding nuclear localisation signals. *EMBO Reports*, **1**, 411–415.
- Dasarathy,B.V. (1991) *Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques*. IEEE Computer Society Press, Las Alamitos, California.
- Devos,D. and Valencia,A. (2001) Intrinsic errors in genome annotation. *Trends Genet.*, **17**, 429–431.
- Doerks,T., Bairoch,A. and Bork,P. (1998) Protein annotation: detective work for function prediction. *Trends Genet.*, **14**, 248–250.
- Eisenberg,D., Marcotte,E.M., Xenarios,I. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
- Eisenhaber,F. and Bork,P. (1998) Wanted: subcellular localization of proteins based on sequence. *Trends Cell Biol.*, **8**, 169–170.
- Eisenhaber,F. and Bork,P. (1999) Evaluation of human-readable annotation in biomolecular sequence databases with biological rule libraries. *Bioinformatics*, **15**, 528–535.
- Fleischmann,R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
- Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.
- Frishman,D. (2000) *PEDANT: Protein Extraction, Description, and Analysis Tool*. Max-Planck-Institute, Munich.
- Gaasterland,T. and Sensen,C.W. (1996) MAGPIE: automated genome interpretation. *Trends Genet.*, **12**, 76–78.
- Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics. *Nat. Biotechnol.*, **18**, 609–613.

- Goffeau,A., Barrell,B.G., Bussey,H., Davis,R.W., Dujon,B., Feldmann,H., Galibert,F., Hoheisel,J.D., Jacq,C., Johnston,M., Louis,E.J., Mewes,H.W., Murakami,Y., Philippsen,P., Tettelin,H. and Oliver,S.G. (1996) Life with 6000 genes. *Science*, **274**, 546–567.
- Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) Selection of representative protein data sets. *Protein Sci.*, **1**, 409–417.
- Hofmann,K., Bucher,P., Falquet,L. and Bairoch,A. (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.*, **27**, 215–219.
- Karp,P.D., Riley,M., Paley,S.M., Pellegrini-Toole,A. and Krummenacker,M. (1999) Eco Cyc: encyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.*, **27**, 55–58.
- Koonin,E.V. (2000) Bridging the gap between sequence and function. *Trends Genet.*, **16**, 16.
- Krawiec,S. and Riley,M. (1990) Organization of the bacterial chromosome. *Microbiol. Rev.*, **54**, 502–539.
- Kretschmann,E., Fleischmann,W. and Apweiler,R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17**, 920–926.
- Lewis,D.D. and Ringuette,M. (1994) Comparison of two learning algorithms for text categorization. *Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval (SDAIR'94)*.
- Lewis,S., Ashburner,M. and Reese,M.G. (2000) Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.*, **10**, 349–354.
- Liu,J. and Rost,B. (2000) *Analysing All Proteins in Entire Genomes*, CUBIC, Columbia University, Department of Biochemistry and Molecular Biophysics.
- Nielsen,H., Engelbrecht,J., Brunak,S. and von Heijne,G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int. J. Neural Syst.*, **8**, 581–599.
- Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
- Remm,M., Storm,C.E. and Sonnhammer,E.L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.*, **314**, 1041–1052.
- Riley,M. (1993) Function of the gene products in *Escherichia coli*. *Microbiol. Rev.*, **57**, 862–952.
- Riley,M. and Labedan,B. (1997) Protein evolution viewed through *Escherichia coli* protein sequences: introducing the notion of a structural segment of homology, the module. *J. Mol. Biol.*, **268**, 857–868.
- Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
- Rost,B. (2001) Enzyme function less conserved than anticipated. *J. Mol. Biol.*, submitted.
- Salton,G. (1989) *Automatic Text Processing*. Addison-Wesley, Reading, MA.
- Sander,C. and Schneider,R. (1994) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res.*, **22**, 3597–3599.
- Schutze,H., Hull,D.A. and Pederson,J.O. (1995) A comparison of classifiers and document representation for the routing problem. *18th Ann Int ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '95)*. pp. 229–237.
- Shannon,C.E. (1951) Prediction and entropy of printed English. *Bell System Tech. J.*, **30**, 50–64.
- Tamames,J., Ouzounis,C., Casari,G., Sander,C. and Valencia,A. (1998) EUCLID: automatic classification of proteins in functional classes by their database annotations. *Bioinformatics*, **14**, 542–543.
- Tamames,J., Ouzounis,C., Sander,C. and Valencia,A. (1996) Genomes with distinct function composition. *FEBS Lett.*, **389**, 96–101.
- Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.
- The *C. elegans* Sequencing Consortium, (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.
- Wall,L. and Schwartz,R.L. (1990) *Programming Perl*. O'Reilly, Sebastopol, CA.
- Yang,Y. and Chute,C.G. (1992) An application of least squares fit mapping to clinical classification. *Proceedings of the Annual Symposium on Computer Applications in Medical Care*. pp. 460–464.
- Yang,Y. and Liu,X. (1999) A re-examination of text categorisation methods. *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 42–49.
- Yang,Y. and Pederson,J.P. (1997) A comparative study on feature selection in text categorization. *The Fourteenth International Conference on Machine Learning*. pp. 412–420.