Finding Transport Proteins in a General Protein Database

Sanmay Das, Milton H. Saier, Jr., and Charles Elkan

University of California, San Diego, La Jolla, CA 92093, USA

Abstract. The number of specialized databases in molecular biology is growing fast, as is the availability of molecular data. These trends necessitate the development of automatic methods for finding relevant information to include in specialized databases. We show how to use a comprehensive database (SwissProt) as a source of new entries for a specialized database (TCDB, the Transport Classification Database). Even carefully constructed keyword-based queries perform poorly in determining which SwissProt records are relevant to TCDB; we show that a machine learning approach performs well. We describe a maximum-entropy classifier, trained on SwissProt records, that achieves high precision and recall in cross-validation experiments. This classifier has been deployed as part of a pipeline for updating TCDB that allows a human expert to examine only about 2% of SwissProt records for potential inclusion in TCDB. The methods we describe are flexible and general, so they can be applied easily to other specialized databases.

1 Introduction

The number of specialized databases in molecular biology is growing fast. The 2006 Database Issue of the journal *Nucleic Acids Research* (NAR) describes 94 new databases and updates of 68 existing databases [2]. The NAR Molecular Biology Collection Database is a compilation of 968 databases as of the 2007 update, 110 more than in the previous update [6]. The vast number of these databases and the number of high-throughput projects producing molecular biological information make it difficult for human curators to keep their databases up-to-date [12]. There has recently been much research on identifying documents containing information that should be included in specialized databases [8,5,13].

The traditional approach has been to apply text classification algorithms to the primary literature to determine whether or not a paper is relevant to the database. We propose an alternative approach: to leverage an existing, general protein database (namely SwissProt, `http://www.expasy.org/sprot/`) by directly screening its records for potential inclusion.

Previously, we investigated an approach in which an expert constructed detailed queries based on keywords and gene ontology terms to identify appropriate SwissProt records. The results of this approach were encouraging, but we hypothesized that a classifier trained on the content of SwissProt records could have higher precision and recall. This paper confirms this hypothesis, in the

J.N. Kok et al. (Eds.): PKDD 2007, LNAI 4702, pp. 54–66, 2007.

context of the Transport Classification Database (TCDB), a specialized protein database created and maintained at UCSD.

Our approach to automatically updating TCDB has three steps: (1) triage to filter out SwissProt records that are not relevant to the TCDB domain; (2) deciding which of the remaining proteins are novel enough to be included in TCDB (since TCDB is intended to be representative, not comprehensive, most transport proteins will not be included); and (3) actually incorporating data from SwissProt and other sources into TCDB.

This paper focuses on step (1) and briefly describes our approach to (2) and (3). We use a maximum-entropy classifier to select relevant proteins from SwissProt. We demonstrate experimentally that this classifier discriminates effectively between transport-related proteins and others on the basis of text in SwissProt records. We show how to interleave training the classifier with creating a clean training set, which we have made publicly available. These results show that maintainers of specialized molecular biology databases can find records of interest with minimal preprocessing after the one-time effort required to create a clean training set.

Having an effective method for automatically selecting data from SwissProt is critical given the size of SwissProt and its growth rate (for example, 2015 new proteins were added to SwissProt between the releases of January 9 and January 23, 2007). Once the number of proteins to consider has been reduced, more complex analyses can then be applied in step (2). We discuss this process in the context of TCDB in Section 3. After the proteins to be included have been determined, the fact that we are working directly with SwissProt records is advantageous for step (3) because information can be transferred directly from one structured format to another.

2 A Pipeline to Identify Relevant and Novel Proteins

A transporter protein is one that imports or exports molecules through the membrane of a cell or organelle. The Transport Classification Database (TCDB) is a web-based resource (`http://www.tcdb.org`) that provides free access to data on proteins related to transmembrane transport. TCDB contains information compiled from over 3000 published papers and is intended to provide information on all recognized, functionally characterized, transmembrane molecular transport systems [11]. TCDB implements and defines the Transport Classification (TC) system, which was officially adopted by the International Union of Biochemistry and Molecular Biology in 2002 as the international standard for categorizing proteins involved in transmembrane transport.

As mentioned above, there are three well-defined steps in our pipeline for automatically updating TCDB. This section introduces terminology we use in the rest of the paper and explains our approach to each of the three steps.

The first step is a triage stage that selects only a subset of *relevant* proteins from SwissProt. We define any protein that is involved in any transmembrane transport process as relevant. Specialists in transport proteins estimate that

approximately 10% of all proteins will meet this description of relevance. It does not automatically follow that 10% of all proteins in SwissProt will be relevant, since the distribution of proteins in SwissProt may be different from the distribution of proteins overall.

It might seem that a search for keywords like "transport" should be sufficient for this task, but even keyword searches constructed by experts are problematic in this domain. This paper describes how we use machine learning techniques to build a classifier that evaluates the relevance of protein records in SwissProt. It is not clear *a priori* that learning-based methods will have high precision and recall, because it is hard for an untrained human being who is not an expert in molecular biology to become skilled at finding relevant records without substantial expert coaching. However, experiments show that we can achieve 95% precision and 95% recall.

One of the major problems we encounter in the process of building a classifier is the absence of a reliable "negative" training set of protein records that are definitely *not* relevant. We know the SwissProt accession numbers of all proteins already in TCDB, but for any protein that is in SwissProt but not in TCDB, we do not know if it is irrelevant or whether it is relevant but not included in TCDB for any of many reasons. Therefore, the process of building a classifier has to involve the assembly of a reliable training set. Details on this process, including the construction of the training set, can be found in Section 3.

After the triage step, the second step is to decide which proteins should actually be included in TCDB. This is hard because TCDB is intended to be a representative database, not a comprehensive one. For example, a protein that is homologous to a protein already in TCDB should be included only if it performs a different function (for example, it acts upon a different substrate) or is in some other way sufficiently distinct (for example, it is from an organism in a different domain and has sufficient sequence dissimilarity). We will refer to proteins that should be included as *novel*. A major advantage of classifying SwissProt records, as opposed to published papers, is that once a record is found to be relevant, we can directly retrieve multiple kinds of information to perform further analysis. For example, we can analyze the sequences of proteins classified as relevant. At this stage we use a rule-based system to decide on the novelty of relevant proteins, largely because of its transparency and comprehensibility to biologists.

In the third stage of the pipeline, if a protein is relevant and novel, a human expert assigns its TC number and the protein information is entered into TCDB. We intend to automate this process as well in the future, but do not focus on the issues involved in this paper.

Since we make extensive use of expert judgments in what follows, it is important to characterize expertise in the TC domain. We define a "Level 1" expert to be a person who is an expert in molecular biology and the TC system. Such a person can definitively decide whether or not to enter a protein into TCDB, and assign TC numbers, using his/her knowledge of the TC system. A "Level 2" expert is someone who has substantial knowledge of transport proteins, but who cannot always decide whether a protein is novel, or what TC number is most

appropriate. Note that even a Level 1 expert may not be able to make a final decision on relevance as defined above, because there may not be enough evidence about a protein. In these cases, the protein will not be included in TCDB.

3 Learning a Classifier to Determine Relevance

The first step of our pipeline consists of a classifier that uses text from certain fields of a SwissProt record to decide whether or not the record describes a protein involved in transmembrane transport. We do not consider TrEMBL records because SwissProt is carefully manually curated and we are largely interested in proteins that are sufficiently well-known and characterized for inclusion in SwissProt.

Evaluation Measures. Precision and recall are the primary measures of success for our classifier. Estimating precision and recall in the absence of known labels for all test examples is in itself a tricky problem [1,4]. We propose that the best way to estimate precision and recall in such circumstances is to perform two separate experiments. Precision is measured using an experiment in which randomly selected unseen test data are labeled as relevant or irrelevant by the classifier, and then all the examples labeled as relevant are manually checked. The proportion of examples labeled as relevant by the classifier that are also manually judged relevant gives a statistically valid estimate of precision. Unfortunately, performing a similar experiment for recall is impractical because it would require labeling the entire test set, not just the examples labeled as relevant by the classifier. Therefore, we use a ten-fold cross-validation experiment on the training data in order to measure recall, since the training set is fully labeled.

Task-specific utility functions are also sometimes used as measures of success, for example, in the document classification task of the 2005 TReC Genomics Track [7]. In our experimental results we present complete confusion matrices from which any function can be computed in addition to precision and recall numbers.

Choice of Features. We use a text representation of selected fields from SwissProt records. The maximum-entropy classifier performs significantly better when using the chosen fields rather than the whole SwissProt record, according to preliminary experiments. This finding is consistent with the result of [3] that text classification algorithms (and, in particular, maximum-entropy methods) perform better when using only selected paragraphs from papers rather than the entire paper. Table 1 shows the fields we use.

The last feature mentioned in Table 1 is the reported number of transmembrane segments in the protein. This feature is derived from the "Features" field of the original SwissProt record, which contains position-wise annotations of the protein sequence. The number is alphabetized and concatenated with the letters "TM," so for example the word "TMFOUR" is added to the representation of the record for a protein annotated as having four transmembrane segments. Note that the number of transmembrane segments is not used until after the

relabeling process described below, and we present final results with and without this feature. Also, the tokenization performed by the software package that we use [9] removes numeric strings and special characters, and only preserves the alphabetic parts of alphanumeric strings.

Table 1. Description of SwissProt fields used in the text classification process

SwissProt Field Title	Text Code	Description
Accession Number	AC	Accession number (unique)
Protein Name	DE	Full name of the protein
References	RT	Titles of papers referenced
Comments	CC	Human annotations
Keywords	KW	Assigned by curators
Ontologies (GO)	DR GO	Gene ontology (GO) terms
Features (Transmembrane Segments)	FT TRANSMEM	# transmembrane segments

One simple way to make the coding of SwissProt records more sophisticated, and possibly more useful to the learning algorithm, would be to add a tag to each word specifying which section of the SwissProt record it is found in. This tagging would, for example, treat the word "transmembrane" in the title of a paper differently from the same word in a functional annotation. While tagging words could be useful, based on experience with the feature encoding the number of transmembrane segments (see Table 2), a dramatic improvement in precision or recall is unlikely.

Selection of Training Data. The training set is created from the version of SwissProt dated September 23, 2006, which contains 232,345 records. The training set contains 2453 SwissProt records corresponding to proteins in TCDB. The features described above are extracted for each of these records, and these 2453 are labeled as positive examples. We select twice this number, that is, 4906 random records from SwissProt excluding the 2453 records known to be in TCDB. We expect the universe of SwissProt records to be significantly unbalanced with respect to our notion of relevance, and we want to reflect this fact in the training set. However, using a very unbalanced training set is problematic because it may lead to a loss in discriminative power in the learned classifier. The classification threshold can be adjusted after training if the classifier identifies too many records as positive, which is presumably the greatest risk of using a training set that is less unbalanced than the test data.

The 4906 records are initially assumed to be negative examples, but some of them are actually positive, since they are randomly selected from the entire universe of SwissProt proteins, which includes a significant number of proteins that are involved in transmembrane transport but are not in TCDB. These may be proteins that should be in TCDB (these are the ones that it is our ultimate goal to identify), or proteins that do not meet the criteria for inclusion in TCDB for a variety of reasons. For training a classifier to determine relevance as opposed to novelty, both these types of records should be labeled positive.

Below, we describe an iterative process for relabeling the negative examples as appropriate.

For final training and experiments after the relabeling process, we used the January 9, 2007 version of SwissProt for each of the 7359 records used, since modifications might have been made to the annotations or sequences.

Choosing a Classifier. We consider two classification algorithms, the naive Bayes classifier and the maximum-entropy classifier. We use the implementations provided in the Mallet toolkit [9]. Both algorithms use Bayesian probabilistic models. Space limitations preclude a detailed description of their properties in this paper; see [10] for a detailed comparison. The naive Bayes classifier builds a generative model using the strong assumption of conditional independence of the features given the class label, while the maximum-entropy model is discriminative, in that it directly maximizes the conditional log-likelihood of the data. The motivation behind maximum-entropy methods is to prefer the most uniform model that satisfies the constraints determined by the data [10].

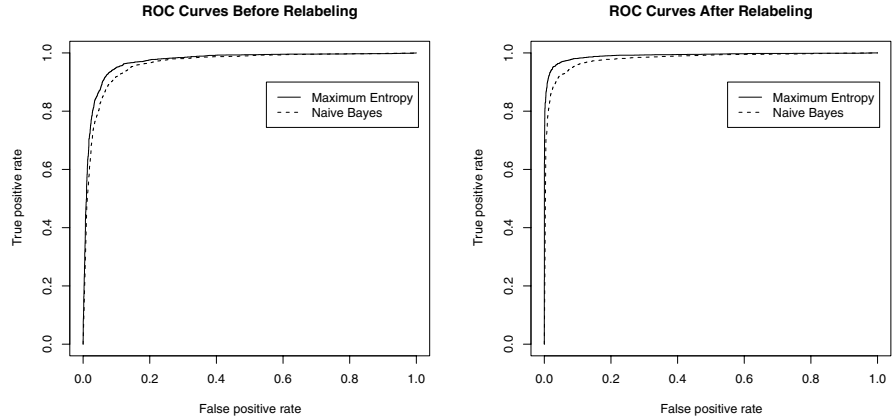

Fig. 1. ROC curves for the maximum-entropy (ME) and naive Bayes (NB) classifiers on the original (left) and relabeled (right) datasets. In the original dataset the random selections from SwissProt are assumed to all be negative examples. The areas under the curves are as follows: NB (original): 0.962, ME (original): 0.971, NB (relabeled): 0.979, ME (relabeled): 0.991. Curves are generated from ten-fold cross-validation experiments.

We compared the performance of the two classifiers on the original training data, assuming that all 4906 randomly selected proteins were, in fact, negative examples. In our application the ROC curve for the maximum-entropy method dominated that for Naive Bayes. We thus proceeded to relabel negative examples (described next) using the maximum entropy method. However, we repeated the ROC curve comparison after relabeling. Both ROC curves are shown in Figure 1. The success of the maximum-entropy method is likely related to its ability to adjust for redundant information, which is prevalent in our domain because the

Table 2. Confusion matrices (above) and precision and recall (below) after each iteration of relabeling and retraining. Performance improves with each iteration, both because incorrect labels are fixed, and because the correct labels are used in the next iteration of training.

Iteration 1

	Neg	Pos
Neg	4628	278
Pos	260	2193

Iteration 2

	Neg	Pos
Neg	4530	183
Pos	261	2385

Iteration 3

	Neg	Pos
Neg	4491	132
Pos	141	2595

Iteration 4

	Neg	Pos
Neg	4456	130
Pos	146	2627

(Rows represent ground truth and columns represent classifier predictions. The row sums change between iterations due to relabeling.)

	Iteration 1	**Iteration 2**	**Iteration 3**	**Iteration 4**
Precision	88.75%	92.87%	95.16%	95.28%
Recall	89.40%	90.14%	94.85%	94.73%
"False" positives that were actually positive	192/278	91/183	37/132	28/130

same information can appear in SwissProt records in multiple places, including in the titles of papers, functional annotations, and keywords.

Updating the Negative Set. One major problem with the training data is that we have no assurance that the 4906 randomly selected proteins are, in fact, negative examples. There are a number of different approaches to learning with partially labeled data, but we focus on creating a clean training set.

Our approach relies on explicitly examining examples classified as positive and deciding on their relevance. We use this approach for two reasons. First, we need to examine these records in detail in any case for the second part of our task, determining whether or not an entry is novel, i.e. should be added to TCDB. Verifying relevance is a necessary part of evaluating novelty. Second, going through the relabeling process allows us to determine which factors are important and develop a screening process for the second part of the pipeline, which is described in detail in the next section. Third, once we have created a reliable negative set of records, it can be re-used by us and by others. We are making this database publicly accessible for researchers.

The process of relabeling is performed in several iterations as follows. Each iteration assumes that the relabeling from the previous iteration is the "ground truth" about relevance. The first iteration assumes that all 4906 randomly selected records are not relevant. Each iteration consists of the following steps:

1. All training examples are randomly divided into ten folds (sets of equal size).
2. For each fold F_i, a classifier is trained on the other nine folds and then applied to each example in F_i.
3. A confusion matrix and precision and recall numbers are computed using the predicted labels and the assumed true labels.
4. Each example in the "false positive" category of the confusion matrix is manually examined to determine relevance (details of this process are in the next subsection).

5. All "false positives" that are manually determined to be relevant are relabeled as positive for the next iteration.

The above process is repeated for four iterations. The process of relabeling becomes faster with each iteration, because proteins are often brought up as false positives repeatedly, but they only need to be examined for relevance once. The results from each iteration are shown in Table 2. The relabeled dataset is available at `http://people.csail.mit.edu/sanmay/data/`.

Table 3. Final training results are similar with and without the feature that codes the number of transmembrane segments

	Without TM feature		**With TM feature**
	Neg Pos		Neg Pos
Neg	4444 114	Neg	4438 120
Pos	139 2662	Pos	135 2666
Precision	95.89%	Precision	95.69%
Recall	95.04%	Recall	95.18%

Manually Determining Relevance. The process of determining whether or not a record is relevant proceeds in a series of steps that go from minimum to maximum expert intervention in the process. First, a number of features are derived to help in making the judgment.

1. The protein sequence is retrieved from SwissProt and then a BLAST search is performed against TCDB.
2. An indicator variable called "Interesting Papers" is defined for each SwissProt record. A paper is thought to be interesting if it may have functional information about the protein. We use a heuristic to make this determination by eliminating any paper that contains two out of the three words "complete," "genome," and "sequence," or their cognates in the title.
3. Similar variables are defined for whether or not the protein may be hypothetical (if the name of the protein starts with "Hypothetical" or "Putative" or is of the form "Protein y⟨xxx⟩") and whether it is likely to be involved in bulk transport (indicated by the presence of the word "vacuole" or its cognates, or the term "endoplasmic reticulum," although transmembrane transporters can also be present in the vacuoles or the endoplasmic reticulum).

We then categorize many of the proteins as follows:

1. Those with best TC-BLAST e-value scores better than 10^{-50} (better means closer to 0, or smaller in magnitude) are automatically assumed to be relevant (TC-BLAST is NCBI blastp applied to all the proteins in TCDB).
2. Those with best TC-BLAST scores worse than 10^{-1} (further from 0, or greater in magnitude) that also have no interesting papers or are hypothetical are assumed to be irrelevant.

3. Proteins involved in bulk transport, which is indicated by the presence of the words mentioned above, but is verified by a human (who can be a non-expert) reading the functional annotation, are also assumed to be irrelevant.
4. Proteins with best TC-BLAST scores better than 10^{-10} which have functional annotations that indicate they perform the same function as their best BLAST hit in TCDB are assumed to be relevant.

The remaining proteins are analyzed by experts. Many of these can be judged by a Level 2 expert, but some need the judgment of a Level 1 expert.

Final Precision and Recall Results. The cross-validation experimental results in Table 2 show that a maximum-entropy classifier trained after the relabeling process achieves recall over 95%. This is a fair estimate of recall on unseen test data, and sufficiently high to be of excellent practical value. Figure 2 shows precision and recall when different proportions of the entire relabeled dataset are used for training and testing. The figure shows that performance continues to improve as more training data is used, perhaps indicating that further improvements may be achievable as we add more entries to TCDB that can serve as positive examples for future iterations of the classifier.

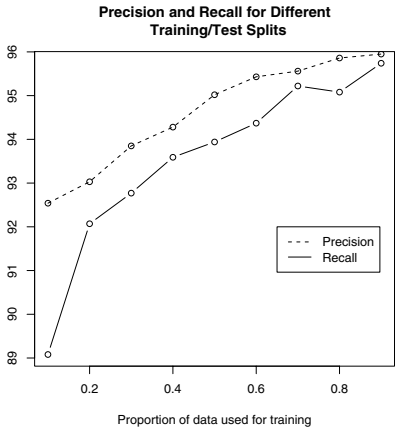

Fig. 2. Precision and recall of the maximum-entropy classifier when different fractions of the entire relabeled dataset are used for training. Results are averaged over ten different random splits for each data point.

To estimate precision on real test data, we use the same experiment that we use to judge the success of the second stage of the pipeline (described below). Out of 1000 randomly selected SwissProt records, the maximum-entropy classifier labels 99 as relevant. Of these 99, 82 are truly relevant, yielding an estimated "real-world" precision of 83%. This value is certainly high enough to be useful in practice.

It is fair to compare these results with the estimated precision and recall of rule sets for determining relevance that we had previously designed in consultation

with experts. One set of less complex rules achieved 67% precision and 73.5% recall while another, more complex, set of rules achieved 78% precision and 71.5% recall. Therefore, it is clear that the classifier is learning non-obvious relations among words from different parts of the text of SwissProt records. The classifier can use combinations of terms appearing in the titles of papers, human annotations of the protein record, the gene ontology record of the protein, and the SwissProt keywords. Similar performance cannot be achieved from a rule-based search, even though the rule-based method focuses on terms thought to be the most important by human experts. For example, while the top twenty words ranked by information gain on the training set include terms like "transport," "transmembrane," and "transporter," which were included among the terms in the expert rulesets, they also include words like "multi-pass" that refer to the topology of the protein, which were not included in human rule-sets. The more important gain in performance is from the classifier's ability to combine and weight different terms in a flexible manner.

Rules for Deciding Novelty. A protein should be included in TCDB only if it is sufficiently novel to add value to the database. Many proteins that are relevant are not sufficiently novel. The most common reason for a relevant protein not to be included is that it is homologous with a high degree of sequence similarity and identical function to a protein already in TCDB. Another common reason for not adding a protein to TCDB is that it does not have sufficient experimental functional characterization.

We devised rules for recognizing when proteins identified as relevant are nevertheless not novel. The proteins not eliminated by these rules are analyzed by a Level 2 expert, and then if necessary by a Level 1 expert, who makes the final decision on which proteins to include in TCDB.

Measuring Performance. In order to estimate the precision of the classifier that predicts relevance (and the success of the rules that evaluate novelty), we train a final classifier on the entire relabeled training set. We then classify 1000 fresh records selected randomly from SwissProt. All fresh records classified as positive are then examined to determine whether they are genuinely relevant, and whether they are novel.

Of the 1000 fresh records, 99 are labeled positive by the final classifier. This is reasonable if approximately 10% of proteins are related to transmembrane transport. As expected, many of the 99 records classified as positive are eliminated by the rules for evaluating novelty and relevance. The rules label 67 of the 99 as relevant but not novel, and another 11 as not relevant. The remaining 21 records were presented to the Level 1 expert. In this set, 5 have been or will be added to TCDB, while 16 will not be included in TCDB: 3 are interesting but do not yet have sufficient functional data, 6 are irrelevant, 1 is a chaperone protein that is marginally relevant, and 6 are too similar to proteins already in TCDB.

Looking just at relevance, of the 99 records labeled as positive by the classifier, 82 are genuinely relevant. Precision is lower (83%) in this real-world experiment than in the cross-validation experiments (96%). There are at least three possible

explanations for the decline in precision. One, the training set on which cross-validation experiments are performed is biased by virtue of containing many proteins from TCDB.[1] Two, the final classifier may simply overfit the training set. The third reason is drift over time in the concept of relevance: some proteins currently in TCDB are not relevant by the criteria used today, but were relevant according to criteria that were previously applied.

A Biological Success: Channel Toxins. A case study shows the benefits that automation can bring to databases like TCDB. While evaluating relevance and novelty, we came across several proteins that are toxins which specifically target transport mechanisms. Since the classifier repeatedly found these toxins, the Level 1 expert decided to introduce a new category into the TC system for them, named 8.B. (The broad Category 8 is for auxiliary proteins that are involved in regulating transport.) This experience shows that our automated methods can find a whole set of proteins that are sufficiently interesting to constitute a new category in TCDB. These proteins were not unknown prior to our experiments, but the expert never thought to include them until the classifier kept flagging them as relevant. The new 8.B category adds value to TCDB as new knowledge and also because channel toxins are important in medicine.

4 Discussion

Our work shows that it is possible to build a classifier that operates on an established general database like SwissProt to select records for potential inclusion in a more specialized database like TCDB, with high precision and recall. Similar classifiers should perform equally well for other specialized databases. Using the classifier to filter out about 90% of SwissProt makes it feasible to apply techniques like BLAST searches to the remaining records that are too expensive, or too inaccurate, to apply to all of SwissProt. The software described above is in real-world use by the biologists who maintain TCDB. For real-world use tools must be convenient, comprehensible, and transparent. Our pipeline meets these criteria.

It is important to consider two stages of evolution of the project for updating TCDB, or another specialized database. The first stage is to bring the database up-to-date, based on information already in SwissProt that was missed in the previous manual construction and updating. The second stage is to use the pipeline to screen fresh records continuously as they are entered into SwissProt.

SwissProt contained 270,778 protein entries as of June 12, 2007. Our experiments show that the maximum-entropy classifier can reduce the set of proteins

[1] For example, the words "escherichia," and "coli" have high information gain on the training set, because TCDB preferentially includes proteins from well-characterized species. A classifier trained primarily on records in TCDB might be too willing to include a protein from *E. coli* in the absence of words indicating a transport function. Similarly, transmembrane transport is overrepresented compared to bulk or intracellular transport; three of the false positives functioned in these types of transport.

we need to consider in more detail by a factor of 10, to around 27,000 proteins. The additional rules we have devised can be used by a combination of software and non-experts to eliminate perhaps 80% of these proteins, still leaving an additional 5400 for an expert to examine. The most critical direction to pursue next is to prioritize these records. The most useful criteria may be the expected quality of functional information present for a protein, which can be estimated from certain attributes of the papers cited in the SwissProt record. For example, prioritizing records that point to recent papers in particularly important journals is the approach currently preferred by the expert maintainers of TCDB.

We hope that this pipeline can continue to be used many years into the future so that experts can restrict the time they spend on updating the database manually. To achieve this, it will be necessary to screen SwissProt (new versions are released bi-weekly) for new proteins, as well as for proteins with updated functional annotations. Between the releases of January 9 and January 23, 2007, for example, 2015 records were added to SwissProt, and 102,269 entries had their annotations revised. Obviously, new records will have to be screened using the pipeline described in this paper, and this seems a feasible goal. While 102,269 is a daunting number, when screening revised SwissProt records, we will only be concerned with proteins that either are already in TCDB or have been marked as potentially interesting given more functional information. Therefore, we expect SwissProt to continue to serve effectively as a data source for updating TCDB.

Acknowledgments. This research is supported by NIH R01 grant number GM077402. The authors thank Aditya Sehgal, who was involved in designing and evaluating the keyword-based strategy for finding relevant SwissProt records.

References

1. Aslam, J.A., Pavlu, V., Yilmaz, E.: A statistical method for system evaluation using incomplete judgments. In: Proc. ACM SIGIR, pp. 541–548. ACM, New York (2006)
2. Bateman, A.: Editorial. Nucleic Acids Res. Database Issue, 34(D1) (2006)
3. Brow, T., Settles, B., Craven, M.: Classifying biomedical articles by making localized decisions. In: Proc. TReC 2005 (2005)
4. Craven, M., Kumlien, J.: Constructing biological knowledge bases by extracting information from text sources. In: Proc. 7th Intl. Conf. on Intelligent Systems for Molecular Biol. (1999)
5. Donaldson, I., et al.: PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. BMC Bioinformatics, 4(1) (2003)
6. Galperin, M.Y.: The molecular biology database collection: 2007 update. Nucleic Acids Res. Database Issue, 35 (2007)
7. William Hersh, A., Cohen, J., Yang, R.T., Roberts, B.P., Hearst, M.: Trec 2005 genomics track overview. In: Proc. TREC (2005)
8. Krallinger, M., Valencia, A.: Text-mining and information-retrieval services for molecular biology. Genome Biol. 6(7), 224–230 (2005)

9. McCallum, A.K.: Mallet: A machine learning for language toolkit (2002), http://mallet.cs.umass.edu
10. Nigam, K., Lafferty, J., McCallum, A.: Using maximum entropy for text classification. In: Proc. IJCAI-99 Workshop on Machine Learning for Inf. Filtering, pp. 61–67 (1999)
11. Saier Jr., M.H., Tran, C.V., Barabote, R.D.: TCDB: the Transporter Classification Database for membrane transport protein analyses and information. Nucleic Acids Res. 36(Database Issue) D181–D186 (2006)
12. Shatkay, H.: Hairpins in bookstacks: Information retrieval from biomedical text. Briefings in Bioinformatics 6(3), 222–238 (2005)
13. Yeh, A.S., Hirschman, L., Morgan, A.A.: Evaluation of text data mining for database curation: Lessons learned from the KDD Challenge Cup. Bioinformatics, 19(Suppl. 1) i331–i339 (2003)