EXAMPLE-BASED CLASSIFICATION OF PROTEIN SUBCELLULAR LOCATIONS USING PENTA-GRAM FEATURES

Jinsuk Kim[1], Ho-Eun Park[2], Mi-Nyeong Hwang[1], Hyeon S. Son[2,3]*

[1]*Information Technology Department, Korea Institute of Science and Technology Information, P.O. Box 122, Yuseong, Daejeon 305-600, Republic of Korea*

Email: jinsuk@kisti.re.kr and mnhwang@kisti.re.kr

[2]*Dept. of Bioinformatics, Soongsil University, 1-1 Sangdo-dong, Dongjak-gu, Seoul, 156-743, Republic of Korea*

Email: parkon@gmail.com

*[3]*Bioinformatics Department, School of Public Health, Seoul National University, 28 Yongon-dong, Chongno-gu, Seoul 110-799, Republic of Korea*

Email: hss@kisti.re.kr

ABSTRACT

The function of a protein is closely correlated with its subcellular location(s). Given a protein sequence, how to determine its subcellular location is a vitally important problem. We have developed a new prediction system for protein subcellular location(s), called ProSLP. *The* ProSLP *is based on n-gram feature extraction method and k-nearest neighbor* (*k*NN) *classification algorithm. It classifies a protein sequence to one or more subcellular compartments based on the locations of top* k *sequences that show the highest weights against the input sequence. The weight is a kind of similarity measure which is determined by comparing n-gram features between two sequences. Currently the* ProSLP *extracts penta-grams as features of protein sequences, computes scores of the potential localization site(s) using k-nearest neighbor* (*k*NN) *algorithm, and finally presents the locations and their associated scores. We constructed a large-scale data set of protein sequences with known subcellular locations from the SWISS-PROT database. This data set contains 51,885 entries with one or more known subcellular locations. The* ProSLP *showed very high prediction precision of about 93% for this data set, and compared with other method, it also showed comparable prediction improvement for a test collection used in a previous work. The {\tt ProSLP} is available through the World-Wide Web at http://proslp.kisti.re.kr.*

Keywords: Protein sequence classification, Sub-cellular Localization, *k*NN classifier, Large-scale data set, Penta-gram, ProSLP

1 INTRODUCTION

As a result of the Human Genome Projects and other large-scale sequencing projects, the molecular sequence data of unknown functions are increasing tremendously. Accordingly, the collection size of protein databanks has also been rapidly increased these days. It has once again brought to the forefront problem of protein function prediction. Subcellular localization is a key functional characteristic of proteins. Given a protein sequence, how to determine its subcellular location as an important clue to its functions is a problem vitally important to biologists and bioinformatists (Chou & Elrod, 1999). For instance, if we want to find out virulence factors from a pathogenic bacterial genome sequence, gene products predicted or identified to be extracellular can be chosen to be primary candidates and others may be ignored in the afterward search. This will greatly reduce the further steps.

For a set of proteins to cooperate for a common physiological or structural function, proteins should be co-localized in the same cellular compartment (Eisenhaber & Bork, 1998). This means that the function of a protein is closely correlated with its subcellular location, and once again it is important to verify the protein's location(s) in the cell to understand its physiological or structural functions such as metabolic pathways, signal transduction cascades, and structural associations. Prediction of a protein's subcellular location can provide valuable clues to the protein function and allow us to screen candidates for drug discovery, to annotate gene products in an automated way, and to select proteins for further study in depth (Gardy *et al.*, 2003).

The prediction of a protein's subcellular location is primarily depending on the similarity search against the protein sequence databank. Although the subcellular location of a protein can be determined by conducting various experiments, it is time consuming and costly to acquire this kind of knowledge solely by experiments. Because the number of sequences entering into databanks has been rapidly increasing, the challenge in expediting the determination procedure for protein subcellular location has become critical and urgent.

Several automated subcellular localization predictors have been developed and made available online: TargetP (Emanuelsson *et al.*, 2000), PSORT (Nakai & Horton, 1999), NNPSL (Reinhardt & Hubbard, 1998), MitoProt and Predotar (Feng, 2002), and PLOC (Park & Kanehisa, 2003). Usually these systems adopted two kinds of feature extraction methods for predicting subcellular locations of proteins. The first method would be to predict the location only based on the amino acid compositions of protein sequences. This approach is applied on systems such as PLOC and NNPSL. The other method is based not only on amino acid composition but also strongly on the existence of signal peptides as adopted in TargetP.

In many cases, however, signal peptides cannot be found or partly assigned, thereby leading to some problems in systems depending on it. It is also known that when protein sequences are decomposed into the amino acid composition, they lose much information for prediction. Hence, it is expected that a higher accuracy should be gained when predicting the subcellular locations directly from sequences (Yuan, 1999).

Independently in computer science, researches on text categorization have shown that extracting proper features representing a document is crucial for categorization effectiveness (Sebastiani, 2002). And scaling most kind of classifiers to large-scale text collection has been impractical except for example-based classification algorithms such as *k*NN classifiers (Yang, 1999). In this paper, we introduce penta-gram as an efficient feature extraction scheme that reflects sequence similarity directly in some degree. And using kNN classification algorithm as well as penta-gram method, a large-scale subcellular localization prediction service has been built. This prediction system is called *ProSLP* (*PRO*tein *S*ubcellular *L*ocalization *P*rediction) and now available online via WWW at URL:http://proslp.kisti.re.kr.

2 SYSTEMS AND METHODS

2.1 System Environment

The ProSLP runs on a Linux desktop with dual Pentium-III 1.2GHz CPUs and 2G bytes of main memory. Since many aspects of an example-based classification system is similar to information retrieval systems, we implemented our classifier on an information retrieval & management system called KRISTAL-2002[1] by incorporating k-nearest neighbor (*k*NN) classification algorithm and n-gram feature extraction method. All the n-gram features are extracted from the sequences in the collection and their inverted lists are stored in the KRISTAL-2000 system. Then, using a vector space model, *k* top-ranked sequences are retrieved against the input sequence and its subcellular locations are predicted based on the locations assigned to the *k* top-ranked sequences.

2.2 Feature Extraction

It is relatively clear to consider that texts written in natural languages are composed of understandable words. Thus, in text categorization, a document's features are generally expressed as words or terms contained in the document (Sebastiani, 2002). Protein sequences also can be regarded as texts written in language of 20 amino acid codes, but it is not clear how to extract meaningful words from the simple

[1] Visit http://giis.kisti.re.kr for more about the KRISTAL-2002 Information Retrieval & Management System.

string of amino acid codes. ProSLP adopted a simple feature extraction scheme, *n-gram*, which is overlapping intervals of fixed length *n*. For example, if *n*=5, protein sequence "ACDEFLERR" is segmented into "ACDEF", "CEDFL", "DEFLE", and "FLERR". After such tokens extracted and indexed for all the sequences in the database, sequences can be searched with traditional information retrieval algorithms.

Several n-gram methods such as *n*=3, 4, 5, 6, and 7 were tested, and since *n*=5 showed the best performance (*data not shown*), the ProSLP currently adopts the interval length of 5 (penta-gram) as its default feature extraction scheme.

2.3 Similarity Measure

As a new protein sequence is entered, our system retrieves top *k* best matching sequences from the collection and deduces appropriate localization site(s) from the locations of the top *k* sequences. The similarity between the query and target sequence is measured by a vector space model regarding penta-gram features as their representative vectors for the sequences. The similarity measure (*Sim(q,s)*) between query sequence *q* and target sequence *s* is defined as follows:

$$Sim(q,s) = \sum_{t \in q \wedge s} w_{s,t} \bullet w_{q,t} \tag{1}$$

where

$$w_{q,t} = \log(f_{q,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

$$w_{s,t} = \log(f_{s,t} + 1) \cdot \log(\frac{N}{f_t} + 1)$$

Here $f_{s,t}$ is the frequency of penta-gram token *t* in sequence *s*; *N* is the total number of sequences in the data collection; f_t is the number of sequences where token *t* occurs once or more; $w_{q,t}$ is the weight of term *t* in query sequence *q*; and $w_{s,t}$ is the weight of term *t* in sequence *s*.

2.4 Category Relevance Measure

If a set of *k* top-ranked sequences, $K = \{s_1, s_2, s_3, ..., s_k\}$, is determined by the similarity measure shown in Eq. (1), the category relevance measure ($w(q, c_i)$), which means the weight of a candidate location or category (c_i) for the query sequence *q*, is defined as

$$w(q,c_i) = \sum_{n=1}^{k} Sim(q,s_n) \bullet (s_n \wedge C_i) \quad (2)$$

where C_i is the set of sequences pre-labeled with category c_i in the data set. After the weights for all candidate categories are computed and normalized to values in the range of (0,1], categories with greater than or equal to a given threshold are chosen to be the final categories.

2.5 Data Sets

Table 1. 12 subcellular locations and their associated keywords. Locations were determined by keywords in the "CC -!- SUBCELLULAR LOCATION:" field of the SWISS-PROT database (Park & Kanehisa, 2003).

Subcellular location	Keywords
Chloroplast	chloroplast
Cytoplasmic	cytoplasmic
Cytoskeleton	cytoskeleton; filament; microtubule
Endoplasmic reticulum	endoplasmic reticulum
Extracellular	extracellular, secreted
Golgi apparatus	Golgi
Lysosomal	lysosomal
Mitochondrial	mitochondrial
Nuclear	nuclear
Peroxisomal	peroxisomal; microsomes; gloxysomal; glycosomal
Plasma membrane	integral membrane
Vacuolar	vacuolar; vacuole

In this work, two data collections of different sizes were used for measuring prediction effectiveness. The smaller data set consists of 7,580 sequence entries which has been used in a previous work (Park & Kanehisa, 2003), while the larger one built in this work consists of 51,885 sequence entries, about 7 times the former data set. Using two data sets of different sizes, we expect that the effect of data set size on prediction effectiveness can be understood. And the smaller data set is also used for comparison purpose between this work and a previous work by Park & Kanehisa (2003).

The first collection, *PLoc* data set, which was built in a previous work (Park & Kanehisa, 2003), is

distributed via WWW at <URL:http://web.kuicr.kyoto-u.ac.jp/~park/Seqdata>. Park & Kanehisa (2003) collected eukaryotic protein sequences from the SWISS-PROT database (Bairoch & Apweiler, 2000) release 39.0. In order to collect proteins in 12 locations, they used the keywords in Table 1 to search against the categorization of subcellular location fields (CC -!- SUBCELLULAR LOCATION:). Prokaryotic protein sequences were removed from the data set by checking the OC field. And sequences with two or more locations, sequences with high similarity and sequences with B, Z, or X amino acid codes were further removed. The final PLoc data set contains 7,580 sequence entries[2]. The number of sequence entries for 12 subcellular locations is summarized in Table 2.

The second collection, *SLP* data set, is also built in a manner similar to the PLoc data set. All protein sequences were collected from the SWISS-PROT database (Bairoch & Apweiler, 2000) release 40.0. However, the limitations used in the PLoc data set were not applied. The final SLP data set consists of 51,885 sequence entries with 54,542 subcellular locations as summarized in Table 2.

Table 2. The number of sequence entries for each subcellular location in the two data sets, *PLoc* and *SLP*.

Subcellular locations	PLoc	SLP
Chloroplast	671	2,717
Cytoplasmic	1,241	19,449
Cytoskeleton	40	126
Endoplasmic reticulum	114	791
Extracellular	861	7,059
Golgi apparatus	47	613
Lysosomal	93	237
Mitochondrial	727	2,977
Nuclear	1,932	8,268
Peroxisomal	125	387
Plasma membrane	1,675	11,721
Vacuolar	54	197
Total locations	7,580	54,542[a]
Total sequences	7,580	51,885 [a]

[a] There are sequences with two or more locations asigned.

[2] Park & Kanehisa (2003) reported that their data set consists of 7,589 entries but their data distributed on the web contains 7,580 sequences only.

Each data set is subdivided into training set and test set. This division has two purposes. First, the classification effectiveness is measured by comparing pre-defined categories of test set with those predicted by the ProSLP against the training set. Second, using these two data sets, *k* value for the *k*NN classifier is experimentally optimized by obtaining the *k* value which shows the best prediction effectiveness.

The PLoc data set is distributed in five subdivisions of test and training set, where test set contains a fifth of total entries and training set consists of the other four fifths. The five subdivisions were named as PLoc1, PLoc2, PLoc3, PLoc4, and PLoc5, in this work (Table 3). The SLP data set was divided into training set of 43,240 sequence entries and test set of 8,645 entries by choosing every sixth entry as test set and the others as training set (Table 3).

Table 3. Test set and training set divisions of the PLoc and SLP data sets.

Data Set	Test Set	Training Set	Total
PLoc1	1,522	6,058	7,580
PLoc2	1,514	6,066	7,580
PLoc3	1,521	6,059	7,580
PLoc4	1,508	6,072	7,580
PLoc5	1,515	6,065	7,580
SLP	8,645	43,240	51,885

2.6 Performance Measures

2.6.1 Microaveraged measures

To evaluate the prediction performance, we used the standard definition of precision and recall as basic performance measures. And microaveraging method (Sebastiani, 2002) was applied to average the precision and recall across subcellular locations. Microaveraged precision (P_{mi}) and recall (R_{mi}) are defined as

$$P_{mi} = \frac{\text{locations relevant and retrieved}}{\text{locations retrieved}} = \frac{TP}{TP + FP}$$

$$R_{mi} = \frac{\text{locations relevant and retrieved}}{\text{locations relevant}} = \frac{TP}{TP + FN}$$

where *TP* = the total number of true positives; *FP* = false positives; and *FN* = false negatives (Sebastiani, 2002; Kim & Kim 2004).

Along with precision and recall, many other researches in text categorization have used microaveraged F_1 measure as the performance measure. The microaveraged F_1 measure (van Rijsbergen, 1979) is the harmonic average of micro-averaged precision and recall, which is defined as

$$F_1 = \frac{2P_{mi} \bullet R_{mi}}{P_{mi} + R_{mi}}$$

A special point of F_1 measure where precision is equal to recall is called break-even point (BeP). Since, theoretically, the BeP is always less than or equal to F_1 measure in any point, it is frequently used to compare effectiveness among different kinds of classifiers (Yang, 1999; Sebastiani, 2002). We will present micro-averaged precision, recall, and the BeP as overall prediction performance measures, and if it is unable to get the BeP, microaveraged F_1 measure will be presented instead.

2.6.2 Macroaveraged measures

With microaveraging method, the measures are overwhelmed by categories with a large number of positive test instances, especially if the different locations are unevenly populated. Therefore it is possible that microaveraged performance is somewhat misleading because more frequent locations overwhelm the microaveraged precision and recall (Wiener *et al.*, 1995). To supplement this problem and compare our system with previous works, we present additional performance measures, macro-averaged precision (P_{ma}) and macro-averaged recall (R_{ma}) defined as

$$P_{ma} = \frac{\sum_{i=1}^{m} p_i}{m} \qquad \text{where} \qquad p_i = \frac{TP_i}{TP_i + FP_i}$$

$$R_{ma} = \frac{\sum_{i=1}^{m} r_i}{m} \qquad \text{where} \qquad r_i = \frac{TP_i}{TP_i + FN_i}$$

Here, p_i is the local precision of location i, r_i is the local recall, m is the number of subcellular locations (m=12), TP_i is the correctly predicted number of location i (true positives), FP_i is the incorrectly assigned number of location i (false positives), and FN_i is the unassigned number of location i while it is pre-labeled with location i (false negatives).

In macroaveraging method, *local* precision and recall for each category is computed at first and then *global* precision and recall is averaged over all local precisions and recalls, respectively. Macro-averaged F_1 measure and its special point, break-even point, also can be defined similarly (Sebastiani, 2002) as defined in microaveraging method.

2.6.3 About nomenclature for evaluation measures

In a previous work by Park and Kanehisa (2003), the terms *accuracy* or *total accuracy* and *location accuracy*, were used as evaluation measures. The actual definition of total accuracy in this literature is the same as micro-averaged recall in our nomenclature. And the location accuracy is the same as macro-averaged recall. However, in the field of text categorization, the term *accuracy* is used in different meaning and defined as $Accuracy=(TP+TN)/(TP+TN+FP+FN)$ (Yang, 1999; Sebastiani, 2002). Therefore, we will avoid using the term *accuracy* here after in this work. Instead we will use the terms microaveraged recall and macroaveraged recall, especially which will be measured at the point where the microaveraged precision and microaveraged recall matches exactly.

3 RESULTS

3.1 Selection of Optimal *k* Values

For *k*-nearest neighbor (*k*NN) classifiers, one of the most crucial steps to improve categorization performance is to determine suitable *k* value for a given data collection (Yang, 1994; Yang, 1999; Sebastiani, 2002). Figure 1 shows the plot for microaveraged precision and recall break-even point (BeP) against varying *k* values for the PLoc2 and SLP data sets. For the PLoc2 data set, there can be seen a trend that as *k* increases, BeP increases slightly from k=1 to k=20 then reaches the maximal value at and after k=20. On the other hand, for the SLP data set, there is no significant difference in classification effectiveness among *k* values tested.

For reference purpose, micro-averaged precision, recall and BeP for the SLP data set are shown in the order of their associated *k* values in Table 4. The ProSLP classifier shows the best BeP (in fact this is an F_1 measure) at k=1 in Table 4, but it is marginally higher than those of other *k* values. The BeP values for the SLP data set ranges from 0.927 to 0.929 (Figure 1; Table 4) while 0.805 to 0.832 for the PLoc2 subdivision (Figure 1).

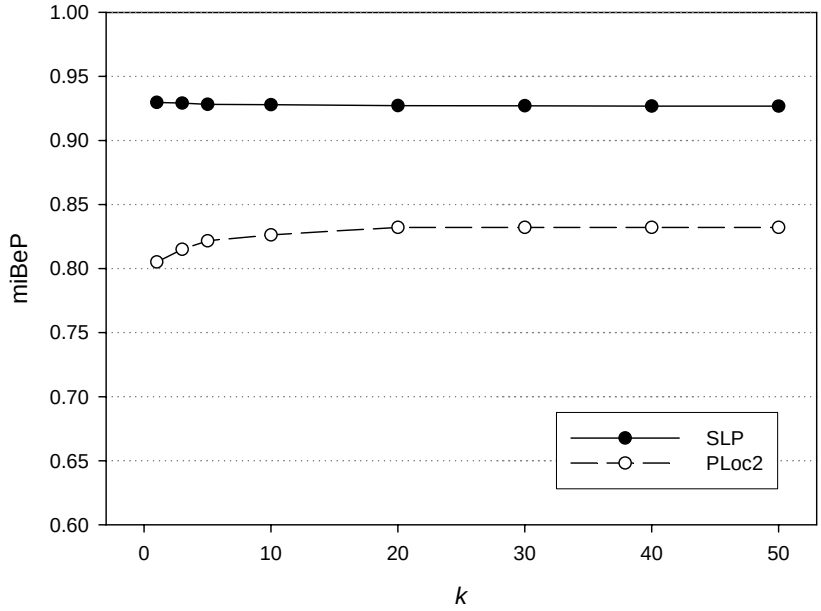

Figure 1. Microaveraged precision and recall break-even point against *k* for the PLoc2 and SLP data sets. "miBeP" means the microaveraged break-even point and "*k*" is the number of top-ranked similar sequences used in categorization of subcellular locations.

Table 4. Prediction effectiveness against *k* values on the SLP data set. The SLP data set consists of training set of 43,240 sequence entries and test set of 8,645 sequences.

k	Microaveraged Precision	Microaveraged Recall	Microaveraged BeP
1	0.9292	0.9302	0.9297[a]
3	0.9292	0.9292	0.9292
5	0.9282	0.9282	0.9282
10	0.9279	0.9279	0.9279
20	0.9272	0.9272	0.9272
30	0.9271	0.9271	0.9271
40	0.9269	0.9269	0.9269
50	0.9268	0.9268	0.9268

[a] Micro-averaged F_1 measure, not BeP, *i.e.*, precision does not equal recall.

From Figure 1, Table 4, and some empirical observations of our own, we chose *k*=20 for the PLoc data set and *k*=10 for the SLP data set. These *k* values are used for all the experiments presented afterward.

3.2 Performance for Two Data Sets

Table 5 summarizes the categorization performance for two data sets, the PLoc and SLP. As mentioned above, since the PLoc data set has five subdivisions, performance for each subdivision was measured and the average local recalls for 12 subcellular locations are shown in Table 5. For the SLP data set, the local recalls for 12 locations were measured with 8,645 test sequences against 43,240 training sequences. For the PLoc data set, local recalls for 12 locations range from 0.567 to 0.926, macroaveraged recall is 0.738 and microaveraged recall is 0.814. For the SLP data set, each location's local recall ranges from 0.650 to 0.976, and macroaveraged recall and microaveraged recall are 0.863 and 0.928 respectively. Note that, in Table 5, though labeled as "microaveraged recall (R_{mi})", the values are microaveraged break-even points since they were measured at the point where microaveraged precision equals microaveraged recall.

Table 5. Effectiveness for the PLoc and SLP data sets. For the PLoc data set, each local recall is averaged over those from five subdivisions.

Subcellular locations	PLoc	SLP
Chloroplast	0.793	0.922
Cytoplasmic	0.789	0.976
Cytoskeleton	0.858	0.650
Endoplasmic reticulum	0.755	0.906
Extracellular	0.765	0.891
Golgi apparatus	0.573	0.792
Lysosomal	0.762	0.892
Mitochondrial	0.579	0.821
Nuclear	0.926	0.916
Peroxisomal	0.567	0.857
Plasma membrane	0.875	0.925
Vacuolar	0.613	0.813
Macroaveraged recall (R_{ma})	0.738	0.863
Microaveraged recall (R_{mi})	0.814	0.928

92.8\% of microaveraged BeP for the SLP data is very high compared with that of the PLoc data set and those of other works. It is probably due to the magnitude of data collections used to train the classifiers. While other systems are trained with training set of hundreds to thousand of pre-labeled sequences, the ProSLP classifier learns from one or two magnitudes larger training set of the SLP data set. This means that one or two magnitude more chances to match similar protein sequences for query sequence is

possible in the case of the ProSLP system than other systems, resulting in much higher prediction effectiveness.

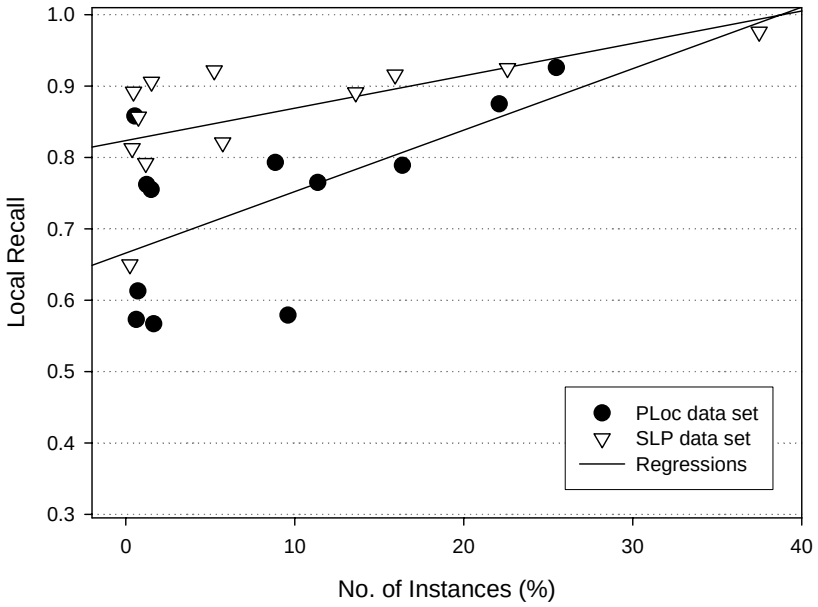

Figure 2. Local recalls for 12 subcellular locations against their number of instances (in percentile) in the PLoc and SLP test collections. Linear regressions for two data sets are also shown.

We also find that the test collection size is correlated with prediction performance for each location (local recall) as shown in Table 5 and Figure 2. In most cases, local recalls of the SLP data set are greater than those of the PLoc data set. Figure 2 plots each local recall against the number of location instances in the whole data set. Since the dimension of two data sets differ in magnitude, “No. of instances” were normalized to percentage against the total number of entries in the data sets. In this figure, one can see a trend that the more the number of instances is, the higher the local recall. Furthermore, it is shown that the slope of linear regression of the PLoc data set is much steeper than that of the SLP data set. This is that the PLoc data set is much highly affected by the number of test instances in each location than the SLP data set. From these observations, we expect that, as the size of data set grows, better subcellular localization prediction will be achieved in example-based classifiers.

3.3 Classification Speed

Another factor to concern is the speed. Since the SLP test collection is seven times greater than the PLoc data set, one of our major focuses on this work was the prediction speed. We randomly selected 48 protein

sequences from another protein database which are not belonged to the SLP data collection. Average length of the proteins is 262 amino acids. They were classified with the ProSLP system against the SLP data set and measured classification time for each sequence, resulting in average 2.28 seconds per sequence. Reminding that the ProSLP runs on a Pentium-III dual processor machine running the Linux operating system, the classification speed seems to be feasible to a practical service. Furthermore, the ProSLP service is two to five times faster than other online subcellular localization predictors (*data not shown*).

3.4 Conclusion

Entering into the post-genomic era, it is very important to identify new gene products in the form of exons or coding regions in the genomic sequences. If one can identify their subcellular locations, they will give some important clues to physiological or structural functions of proteins. In this paper, we showed that large-scale prediction system for subcellular localization is possible with a kNN classifier using penta-gram features.

The ProSLP, a novel subcellular location prediction system, gives very high precision for blindly assigning subcellular locations to large numbers of potential proteins. In addition, its prediction time is practical even on a out-of-date computer architecture such as Pentium-III based desktop machine. The ProSLP service and its associated data collection are available on the World Wide Web at <URL:http://proslp.kisti.re.kr>. As a meta-search engine, it provides not only the ProSLP's prediction result but also other services' predictions for user convenience. We hope that the ProSLP can help biologists and bioinformatists study various biological problems related to proteins.

4 DISCUSSION

4.1 Relationships Among Various Measures

The ProSLP classifier is an example-based classification system which classifies a protein's subcellular location based on those of k top-ranked sequences. In this process, one or more candidate locations can be suggested and candidate locations with weights above a given threshold are assigned to the input sequence. Figure 3 shows plots of various measures against this threshold used in determination of subcellular locations for each test sequence of the SLP data set. There can be seen that as the threshold increases, the microaveraged recall (R_{mi}) and macroaveraged recall (R_{ma}) decrease, but microaveraged precision (P_{mi}) and microaveraged F_1 measure increase.

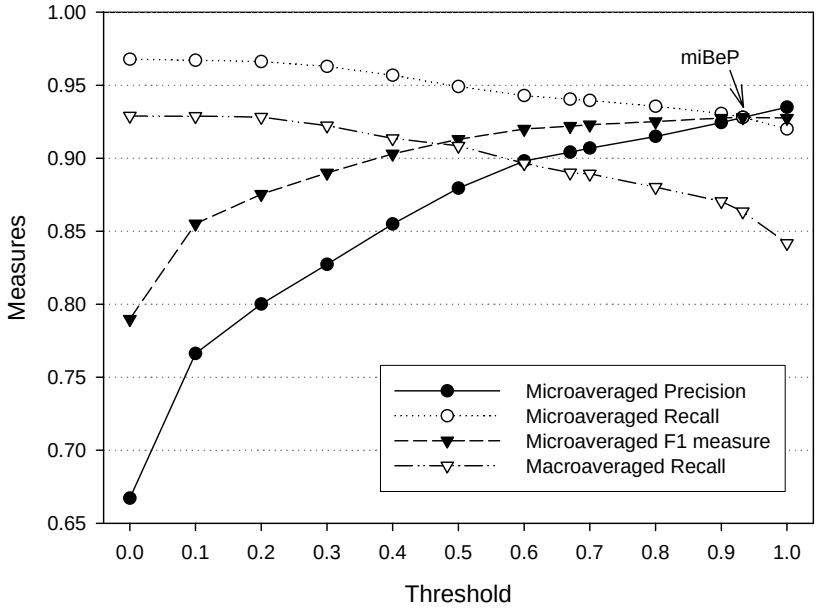

Figure 3. Plots of various evaluation measures against the categorization threshold for the SLP data set. "miBeP" is the microaveraged precision and recall break-even point which is a special point of F_1 measure where the precision equals the recall.

To present a classifier's performance correctly, break-even point or F_1 measure is important rather than precision only or recall only. However, in practical service, recall may be more important than precision since high precision at the expense of recall (Figure 3) may result in missing true positives if they have low category weights.

4.2 Comparison With Other Method

To compare our method with previous works, we chose the PLoc data set and its associated method, PLOC, reported in Park & Kanehisa (2003). They used support vector machines (SVMs) approach and showed that the PLOC method is superior to a previous work by Chou & Elrod (1999). The comparison between the PLOC and the ProSLP approaches is shown in Table 6. Microaveraged recall, which is denoted as *total accuracy* in Park & Kanehisa (2003), has been improved slightly from 78.2% to 81.4%, while macroaveraged recall, denoted as *location accuracy*, significantly improved from 57.9% to 73.8%.

Table 6. Comparison of the ProSLP prediction effectiveness with a previous approach, PLOC method.

Subcellular locations	PLOC method	ProSLP method
Chloroplast	72.3	79.3
Cytoplasmic	72.2	78.9
Cytoskeleton	58.5	85.8
Endoplasmic reticulum	46.5	75.5
Extracellular	78.0	76.5
Golgi apparatus	14.6	57.3
Lysosomal	61.8	76.2
Mitochondrial	57.4	57.9
Nuclear	89.6	92.6
Peroxisomal	25.2	56.7
Plasma membrane	92.2	87.5
Vacuolar	25.0	61.3
Macroaveraged recall (%)	57.9	73.8
Microaveraged recall (%)	78.2	81.4

Furthermore, the result of the PLOC method still shows over-fitting for large groups and under-fitting for small groups such as Golgi (14.6%), Peroxisomal (25.2%) and Vacuolar (25.0%). But our method is much more balanced among small and large groups as shown in Figure 4. The slope of linear regression for the ProSLP method is slower than that of the PLOC method. This means that the number of positive examples in each location has smaller effect on the ProSLP method compared with that on the PLOC method. The local recalls by the PLOC method range from 14.6% to 92.2% while those by the ProSLP range from 56.7% to 92.6%.

Classifiers based on *busy* learning algorithms such as neural networks and support vector machines can be suffered from over-fitting to large groups and under-fitting to small groups. However, this problem becomes loosened in classifiers based on *lazy* learning algorithms such as examples-based classifiers, due to lack or delay of learning steps (Sebastiani, 2002). Since the ProSLP classifier is a kind of example-based classifiers, it also seems not to be suffered from over-fitting and under-fitting problem as shown in Table 6 and Figure 4.

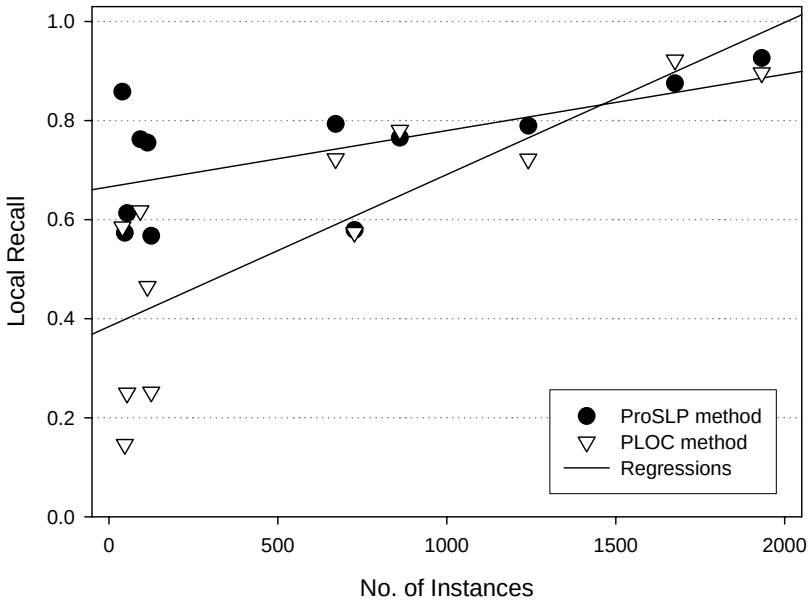

Figure 4. Comparison of the distribution of local recalls for 12 subcellular locations against their number of instances in the PLoc data set by applying the ProSLP approach and PLOC method. Linear regressions for two methods are also shown.

4.3 Suggestions For Further Improvements

The time required to classify a sequence in the ProSLP system is approximately proportional to the sequence length (*data not shown*). This is a characteristic of our feature extraction scheme, penta-gram method. Since the number of penta-grams for a sequence of length l is l-5+1, there are more features to be evaluated in retrieving k top-ranked sequences and thus more computation is required as l increases. In text categorization, it has shown that *feature selection*, also known as dimensionality reduction, further improves the classification speed as well as classification effectiveness (Yang, 1999; Sebastiani, 2002, Kim & Kim, 2004) by removing relatively unimportant features from the feature pool. If such feature selection mechanism for penta-gram features is applied, the ProSLP system is hoped to be further improved and speeded up.

As shown in Table 5 and Figure 2, the test collection size seems to be very important in precise determination of subcellular locations. These days, the protein database sizes increase rapidly, and thus subcellular localization data will also be increased accordingly resulting in more accurate predictions.

Finally we consider new n-gram methods other than penta-gram and similarity measure in retrieving k

top-ranked sequences. In this work, we did not introduced any previously known properties of protein sequences such as scoring matrices used in sequence alignment. We hope that further improvement will be possible by introducing biological knowledge in feature extraction and sequence retrieval steps.

4.4 Suggestion For Performance Measures

Many measures and synonyms - such as precision and recall, specificity and sensitivity, total accuracy and local accuracy, and success rate - have been used in previous works of protein subcellular localization prediction, even misleadingly. In this paper, we suggest that microaveraged break-even point (miBeP) is better for a classifier's effectiveness to be compared directly with other classification methods. Or microaveraged precision, recall and F_1 measure together can be used to compare indirectly the effectiveness of different classifiers, especially if the miBeP can not be obtained. Note that higher precision expenses recall, and vice versa, as shown in Figure 3. Precision only or recall only does not represent correct performance of a classifier.

To describe overfitting and underfitting, we suggest local precision and local recall together for each group should be presented. Furthermore, since microaveraged measures can be easily optimized too much for large groups at the expense of small groups, macroaveraged break-even point or F_1 measure might be a good indicator to show the balance of prediction accuracies between large and small groups.

5 ACKNOWLEDGMENTS

This work was supported in part by a grant from IBM Korea, Inc. The first author thanks the GIIS group of KISTI for helpful conversations, and Changmin Kim and Jieun Chong for supporting this work.

6 REFERENCES

Bairoch, A. & Apweiler, R. (2000) The SWISS-PROT Protein Sequence Database and its Supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45-48

Chou, K.C. & Elrod, D.W. (1999) Protein Subcellular Location Prediction. *Protein Engineering*, **12**, 107-118

Eisenhaber, F. & Bork, P. (1998) Evaluation of Human-readable Annotation in Biomolecular Sequence Database with Biological Rule Libraries. *Bioinformatics*, **15**, 528-535

Emanuelsson, O., Nielsen, H., Brunak, S., & Heijne, G.V. (2000) Predicting Subcellular Localization of Proteins Based on their N-terminal Amino Acid Sequence. *J. Mol. Biol.*, **300**, 1005-1016

Feng, Z. P. (2002) An Overview on Predicting the Subcellular Location of Protein. *In silico Biol.*, **2**, 291-303

Gardy, J.L., Spencer, C., Wang, K., Ester, M., Tusnady, G.E., Simon, I., Hua, S., deFays, K., Lambert, C., Nakai, K., & Brinkman, F.S.L. (2003) PSORT-B: Improving Protein Subcellular Localization Prediction for Gram-negative Bacteria. *Nucleic Acids Res.*, **31**, 3613-3617

Kim, J. & Kim, M.H. (2004) An Evaluation of Passage-based Text Categorization. *Journal of Intelligent Information Systems*, **23**(1), 47-65

Park, K.J, & Kanehisa, M. (2003) Prediction of Protein Subcellular Locations by Support Vector Machines Using Compositions of Amino Acids and Amino Acid Pairs. *Bioinformatics*, **19**, 1656-1663

Nakai, K. & Horton, P. (1999) PSORT: A Program for Detecting the Sorting Signals of Proteins and Predicting their Subcellular Localization. *Trends Biochem. Sci.*, **24**, 34-35

Reinhardt, A. & Hubbard, T. (1998) Using Neural Networks for Prediction of the Subcellular Location of Proteins. *Nucleic Acids Res.*, **26**, 2230-2236

Sebastiani, F. (2002) Machine Learning in Automated Text Categorization. *ACM computing Surveys*, **34**, 1-47

van Rijsbergen, C. (1979) *Informational Retrieval*. London, UK: Butterworths

Wiener, E., Pedersen, J.O., & Weigend, A.S. (1995) A Neural Network Approach to Topic Spotting. *In Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, pp 317-332.

Yang, Y. (1994) Expert Network: Effective and Efficient Learning from Human Decisions in Text Categorization and Retrieval. *Proceeding of the 17th Annual International ACM/SIGIR Conferences on Research and Development in Information Retrieval*, pp 13-22

Yang, Y. (1999) An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, **1**, 69-90

Yuan, Z. (1999) Prediction of Protein Subcellular Locations Using Markov Chain Models. *FEBS Letters*, **451**, 23-26