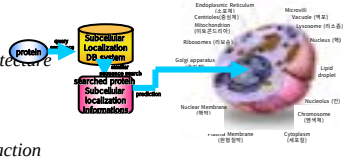
A Novel Predictor of Protein Subcellular Localization based on N-gram Features

Ho-Eun Park and Jinsuk Kim
{hepark, jinsuk}@kisti.re.kr
Center for Computational Biology and Bioinformatics,
Korea Institute of Science and Technology Information

w prediction system for protein subcellular localization site(s), called as been developed. ProSLP classifies a protein sequence to one or more r compartments based on the location of the top *k* sequences which highest weights against the input sequence. Currently ProSLP extracts ns as features of the sequence, computes scores of the potential n site(s) using *k*-nearest neighbor (*k*NN) algorithm and finally presents and their associated scores. ProSLP is available through the World- at http://proslp.kisti.re.kr.

uction

result of large-scale genome sequencing projects, the number of gene r protein sequences of unknown functions are increasing tremendously. e-consuming and costly job to identify the inherent functions of protein . Given a protein sequence, how to determine its subcellular n as an important clue to its functions is a problem vitally important to and bioinformatists (Chou and Elrod, 1999). To cooperate for a physiological function, proteins must be localized in the same cellular ent. So subcellular localization is a key functional characteristic of Eisenhaber and Bork, 1998). Several automated subcellular localization have been developed and made available online: TargetP sson *et al.*, 2000), PSORT (Nakai and Horton, 1999), NNPSL t and Hubbard, 1998), SubLoc (Hua and Sun, 2001), MitoProt and Feng, 2002). In most of theses approaches, features for a sequence are in the form of amino acid composition and/or sorting signals located at l amino acid sequence. And these features are further processed by nds of machine learning approaches such as neural networks, support achines, and k-nearest neighbor (kNN) classifiers, finally giving localization site(s). Using *k*NN classification algorithm and n-gram xtraction, a subcellular localization prediction service has been . This prediction system is called ProSLP (**Pro**tein **S**ubcellular on **P**rediction) and now available via WWW at **slp.kisti.re.kr**.



ds
rchitecture

extraction

SLP adopts a simple feature extraction scheme, n-gram tokens, which is ng intervals of fixed length n. For example, if the length is 5, protein "ACEDFIMMPAA" is segmented into "ACEDF", "CEDFI", "EDFIM", PAA". Current version of ProSLP system uses the interval length of 5 m) as its feature extraction scheme. After such tokens are extracted and or all the sequences in the database, sequences can be searched with formation retrieval algorithms.

Similarity measure

The similarity between the query and target sequence is measured by vector space model regarding n-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:

$$Sim(q,s) = \frac{1}{W_s} \cdot \sum_{t \in q \wedge s} w_{s,t} w_{q,t}$$

with:

$$W_s = \log(1 + \sum_{t \in s} f_{s,t})$$

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

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

where $f_{s,t}$ is the frequency of n-gram token t in sequence s; N is the total number of sequences in the data collection; f_t is the number of sequence where token t occurs more than or equal to once; $w_{s,t}$ is the weight of term t in sequence s; W_s is the normalization factor for the length of target sequence s.

Performance measures

To evaluate the prediction performance, we used the standard definition of precision (p) and recall (r), F, and break-even point (BeP).

$$p = \frac{locations\ relevant\ and\ retrieved}{locations\ retrieved} = \frac{tp}{tp + fp}$$

$$r = \frac{locations\ relevant\ and\ retrieved}{locations\ relevant} = \frac{tp}{tp + fn}$$

$F_1 = \frac{2pr}{p+r}$ (A special point of *F1* measure where p = r is called BeP)

Results

Prediction effectiveness

For k-nearest neighbor (kNN) classifiers, one of the most crucial steps to improve categorization performance is to determine suitable *k* for given data collection (Yang, 1999). Using two data collections, k value for the *k*NN classifier is experimentally optimized by obtaining the *k* value which shows the best prediction effectiveness. Figure 1 sh… break-eve… value i…

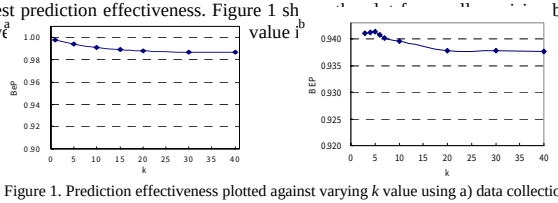


Figure 1. Prediction effectiveness plotted against varying *k* value using a) data collection 1 and b) data collection 2.

Data collection 1 consists of training set (86,720 sequence entries) a… (10,840 sequences) and data collection 2 consists of training set (43,240 entries) and test set (8,645 sequences)

Prediction accuracy of ProSLP is within the range of 93% to 98% and recall for all *k* points.

Prediction times

Since ProSLP data collection is one or two-magnitudes larger t… used in other prediction systems, one of our major focuses on this wor… prediction time, which are plotted against query sequence lengths in … These experiments were carried out on a Pentium-III dual processor… running the Linux operating system. 48 query sequences were randoml… and classification time for each query was measured. Average length of … set is 262 amino acids and average search time for each query is shown … seconds.

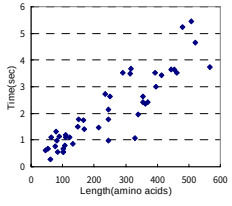


Figure 2. Prediction times plotted against the amino acid lengths of query protein sequences. We have used 48 query sequences, of which average length is 262 amino acids ranging 53 to 600.

Discussion

We showed that large-scale prediction system for subcellular loca… possible with a *k*NN classifier and n-gram features of the protein s… ProSLP, a novel subcellular location prediction system, gives incred… accuracy. In addition, its prediction time is practical even on a … computer architecture. We hope that ProSLP can help biolo… bioinformatics study various biological problems related to proteins.

References

Chou,K.C. and Elrod,D.W. (1999) Protein subcellular location prediction. Protein E… 12:107-108.

Emanuelsson,O., Nielsen,H., Brunak,S., and Heijne,G.V. (2000) Predicting … Localization of Proteins Based on their N-terminal Amino Acid sequence. … 300:1005-1016.

Feng,Z.P. (2002) An overview on predicting the subcellular location of protein. In … 2:291-303.

Gardy,J.L., Spencer,C., Wang,K., Ester,M., Tusnady,G.E., Simon,I., Hua,S., … Lambert,C., Nakai,K., and Brinkman,F.S.L. (2003) PSORT-B: improving protein… localization prediction for Gram-negative bactera. Nucleic Acids Res., 31:3613-3617

Hua,S. and Sun,Z. (2001) Support vector machine approach for protein subcellular … prediction. Bioinformatics, 17:721-728.

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting the sorting signals … and predicting their subcellular localization. Trends Biochem. Sci., 24:34-35.

van Rijsbergen,C. (1979) Information Retrieval Butterworths, London, 1979.

Yang,Y. (1999) An Evaluation of Statistical Approaches to Text Categorization. … Information Retrieval, 1:67-88.