Amino Acids (2008) 35: 65–73
DOI 10.1007/s00726-007-0628-7
Printed in The Netherlands

Amino Acids

Discrimination of outer membrane proteins using a *K*-nearest neighbor method

C. Yan, J. Hu, and **Y. Wang**

Department of Computer Science, Utah State University, Logan, UT, USA

Received September 17, 2007
Accepted October 28, 2007
Published online January 25, 2008; © Springer-Verlag 2008

Summary. Identification of outer membrane proteins (OMPs) from genome is an important task. This paper presents a *k*-nearest neighbor (*K-NN*) method for discriminating outer membrane proteins (OMPs). The method makes predictions based on a weighted Euclidean distance that is computed from residue composition. The method achieves 89.1% accuracy with 0.668 MCC (Matthews correlation coefficient) in discriminating OMPs and non-OMPs. The performance of the method is improved by including homologous information into the calculation of residue composition. The final method achieves an accuracy of 96.1%, with 0.873 MCC, 87.5% sensitivity, and 98.2% specificity. Comparisons with multiple recently published methods show that the method proposed in this study outperforms the others.

Keywords: Prediction – Transmembrane proteins – Machine learning – Gram-negative bacteria

1. Introduction

Membrane proteins are important targets of protein science and cell biology research (Chou and Shen, 2007c; Douglas et al., 2007). Two of the hot topics related to membrane proteins are to identify the type of membrane proteins (Chou and Elrod, 1999; Wang et al., 2004, 2005b, 2006; Chou and Cai, 2005a, b; Liu et al., 2005a, b; Shen and Chou, 2005b, 2007e; Shen et al., 2006; Chou and Shen, 2007c; Pu et al., 2007) and to identify transmembrane regions (e.g. Diao et al., 2007). Outer membrane proteins (OMPs) perform diverse functional roles, including bacterial adhesion, structural integrity of the cell wall, and material transport (Koebnik et al., 2000; Schulz, 2000; Wimley, 2003). These proteins consist of β-barrel transmembrane regions and are found in the outer membranes of gram-negative bacteria and the outer membranes of mitochondria and chloroplasts (Schulz, 2000; Waldispuhl et al., 2006; Wimley, 2003). Unlike α-helical membrane proteins, which can be easily identified based on long stretches of hydrophobic residues, OMPs are more difficult to predict, mainly due to shorter membrane-spanning regions with higher variations in properties (Koebnik et al., 2000). A few methods have been developed to identify OMPs. Gnanasekaran et al. (2000) used profiles developed from structure-based alignments of porins to identify OMPs. Wimley et al. (2002) analyzed the structures of 15 non-redundant OMPs and developed a method to identify OMPs based on residue composition and structural features. Martelli et al. (2002) and Bagos et al. (2004a, b) used hidden Markov models (HMMs) to discriminate OMPs from globular proteins. Liu et al. (2003) developed a method that combines the residue composition of membrane spanning regions and predicted secondary structure to identify OMPs. Garrow et al. (2005a,b) developed a method for discrimination of OMPs in genomes. Berven et al. (2004) developed the BOMP method that predicts OMPs by combining pattern search, β-barrel score, and a filter that explores the abundance of asparagine and isoleucine in the protein. Gromiha and Suwa (2005) developed a simple method to identify OMPs using a "deviation" distance based on amino acid composition. Later, they evaluated 11 machine-learning methods for the discrimination of OMPs using residue composition as input (Gromiha and Suwa, 2006). In another study, researchers from the same group used a backward-and-forward approach to select discriminative features from residue composition and dipeptide composition and used a SVM method to identify OMPs (Park et al., 2005).

K-nearest neighbor (*K-NN*) method has been successfully adopted to predict various protein attributes (Chou, 2002; Shen et al., 2007b), such as protein subcellular localization (see, e.g., (Chou and Shen, 2006a, b, c; Chou and Shen, 2007a, b; Shen and Chou, 2007b, c, f; Shen et al., 2007a)), subnuclear protein localization (Shen and Chou, 2005a), protein structural classification (Shen et al., 2005), protein fold pattern (Shen and Chou, 2006), membrane protein type (Shen and Chou, 2005b, 2007e; Shen et al., 2006; Chou and Shen, 2007c), enzyme main and sub functional classification (Shen and Chou, 2007a), and signal peptide (Chou and Shen, 2007e; Shen and Chou, 2007d). In this study, we propose a *K-NN* method for the discrimination of OMPs from non-OMPs. The method achieves 96.1% accuracy, with 0.873 MCC, 87.5% sensitivity, and 98.2% specificity. Comparisons with multiple recently published methods show that the proposed method outperforms the others.

2. Materials and methods

2.1 Datasets

Three datasets were obtained from a previous study by Park et al. (2005): *SetA* contains 208 well-annotated outer membrane proteins (OMPs); *SetB* has 673 globular proteins (which includes 155 all-α, 156 all-β, 184 $\alpha + \beta$, and 178 α/β proteins) from various structure families; and *SetC* has 206 α-helical membrane proteins (AMPs). In these datasets, the sequence identity between any two proteins is less than 40%. We first used these datasets to evaluate the proposed method. Then, we filtered these datasets so that the similarity between any two proteins is less than 25%. After the filtering, 112 OMPs, 673 globular proteins, and 178 AMPs are left. We also used the filtered datasets to evaluate the proposed method.

2.1.1. Residue composition

The residue composition of a protein was calculated as $x_i = n_i / \Sigma_j n_j$, where n_i and n_j are the numbers of residues of types i and j. The average residue composition of OMPs was given by $\bar{x}_{i\_\text{omp}} = n_{i\_\text{omp}} / \Sigma_j n_{j\_\text{omp}}$, where n_{i_omp} and n_{j_omp} are the total numbers of residues of types i and j in OMPs. The average residue compositions of globular proteins ($\bar{x}_{i\_\text{glo}}$) and AMPs ($\bar{x}_{i\_\text{amp}}$) were also calculated in a similar way.

2.2 Weighted Euclidean distance

For a protein (*test protein*) in the test set, its distance to an OMP protein in the training set (*OMP train protein*) was calculated using $D_\text{omp} = \sqrt{\Sigma_i \frac{(x_{i\_\text{test}} - x_{i\_\text{omp\_train}})^2}{\bar{x}_{i\_\text{omp}}}}$, where x_{i_test} is the composition of residue type i in the test protein, $x_{i\_\text{omp\_train}}$ is the composition of residue type i in the *OMP train protein*, and $\bar{x}_{i\_\text{omp}}$ is the average composition of residue type i for all OMPs in the training set. Notice that $\sqrt{\Sigma_i (x_{i\_\text{test}} - x_{i\_\text{omp\_train}})^2}$ gives the Euclidean distance between the test protein and the *OMP train protein*. Here, in the calculation of D_omp, each item within the summation is weighted by a factor of $1/\bar{x}_{i\_\text{omp}}$. Therefore, D_omp is referred to as *weighted Euclidean distance* in this study. The composition of all 20 amino acid residues was used to calculate the distances for all experiments in this study.

Similarly, the weighted Euclidean distance between the test protein and a globular protein in the training set (*globular train protein*) was calculated using $D_\text{glo} = \sqrt{\Sigma_i \frac{(x_{i\_\text{test}} - x_{i\_\text{glo\_train}})^2}{\bar{x}_{i\_\text{glo}}}}$, where x_{i_test} is the composition of residue type i in the test protein, x_{i_glo} is the composition of residue type i in the *globular train protein*, and $\bar{x}_{i\_\text{glo}}$ is the average composition of residue type i for all globular proteins in the training set. The weighted Euclidean distance between the test protein and an AMP protein in the training set (*AMP train protein*) was calculated using $D_\text{amp} = \sqrt{\Sigma_i \frac{(x_{i\_\text{test}} - x_{i\_\text{amp\_train}})^2}{\bar{x}_{i\_\text{amp}}}}$, where x_{i_test} is the composition of residue type i in the test protein, $x_{i\_\text{amp\_train}}$ is the composition of residue type i in the *AMP train protein*, and $\bar{x}_{i\_\text{amp}}$ is the average composition of residue type i for all *AMPs* in the training set.

2.3 K-Nearest neighbor (K-NN) algorithm

For a test protein, its weighted Euclidean distances to every OMP in the training set were calculated separately. K smallest distances were chosen. Let them be $D_{\text{omp}-1}, D_{\text{omp}-2}, \ldots, D_{\text{omp}-k}$. The distance between the test protein and the OMP class was given by $\overline{D}_\text{omp} = \frac{1}{k}(D_{\text{omp}-1} + D_{\text{omp}-2} + \cdots + D_{\text{omp}-k})$. The distance between the test protein and the globular protein class ($\overline{D}_\text{glo}$) and the distance between the test protein and the AMP class ($\overline{D}_\text{amp}$) were computed in a similar way. In this study, the value of k was determined empirically. Various values were tried, and the best performance was achieved when $k = 4$. The test protein was classified into a class to which it has a shorter distance.

2.4 Five-fold cross-validations

Five-fold cross-validation was used to evaluate the proposed method. The overall dataset was divided into five subsets, such that the identity between any two proteins from different datasets was less than 25%. This threshold has been used in many studies for removing redundancy (Rost and Sander, 1993; Ahmad and Sarai, 2004; Deng et al., 2004; Prasad Bahadur et al., 2004; Wang and Brown, 2006). OMPs, globular proteins and AMPs were distributed into the subsets evenly. In each round of experiment, four subsets were used as the training set and the remaining subset was used as the test set. This procedure was repeated five times with each subset being used as test set once. The average performance was reported. It is instructive to point out that independent dataset test, sub-sampling (e.g., five or ten-fold cross-validation) test, and jackknife test are often used for examining the accuracy of a statistical prediction method. Among them, the jackknife test is deemed the most objective and being able to yield a unique result (Chou and Zhang, 1995), as demonstrated by an incisive analysis in a recent comprehensive review (Chou and Shen, 2007d) as well as has been increasingly and widely adopted by investigators to test the power of various prediction methods (Zhou, 1998; Zhou and Doctor, 2003; Huang and Li, 2004; Gao et al., 2005a, b; Wang et al., 2005a; Xiao et al., 2005, 2006; Cao et al., 2006; Chen et al., 2006a, b, 2007; Du and Li, 2006; Gao and Wang, 2006; Guo et al., 2006a, b; Kedarisetti et al., 2006; Mondal et al., 2006; Niu et al., 2006; Sun and Huang, 2006; Wen et al., 2006; Yan et al., 2006; Zhang et al., 2006; Chen and Li, 2007; Diao et al., 2007; Ding et al., 2007; Fang et al., 2007; Jahandideh et al., 2007; Lin and Li, 2007a, b; Liu et al., 2007; Shen and Chou, 2007e; Shen et al., 2007a; Shi et al., 2007; Tan et al., 2007; Xiao and Chou, 2007; Zhang and Ding, 2007; Zhou et al., 2007). However, in the current study, we choose to use five-fold cross-validation because it also has been widely used in previous studies and, more importantly, it is less time-consuming than jackknife test.

2.5 Including homologous sequences into the calculation of residue composition

For each protein, the BLAST program (Altschul et al., 1997) was used to search for homologous sequences from the National Center for Biotechnology Information (NCBI) non-redundant database using an *E*-value of 0.0001. 50 best hits (not including the query sequence itself) were chosen from the returned result. If less than 50 hits were returned, then all of the hits were chosen. These homologous proteins plus the query protein were used to compute the residue composition for the query protein.

2.6 Performance measures

Three types of two-class classifications were performed in this study: OMPs vs. globular, OMPs vs. AMPs, and OMPs vs. non-OMPs (i.e., globular + AMPs). In each of these experiments, OMP class was defined as the positive class, and the other was the negative class. Let TP be the number of true positives (i.e., the number of OMPs predicted as OMPs); TN be the number of true negatives (i.e., the number of proteins from the negative class that are predicted to belong to the negative class); FN be the number of false negatives (i.e., the number of OMPs incorrectly predicted as negative) and FP be the number of false positives (i.e., the number of negative proteins incorrectly predicted as OMPs). Several measures were used to evaluate the method:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$$

$$\text{MCC} = \frac{\text{TP}\times\text{TN} - \text{FP}\times\text{FN}}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}$$

3. Results

3.1 The proposed method can distinguish between OMPs and globular proteins

First, we evaluated the proposed method's ability to discriminate OMPs from globular proteins. Five-fold cross-validations were performed as described in Materials and Methods. The results (Table 1, column 2) show that the proposed method achieves 88.8% overall accuracy with 0.708 MCC. 84.1% (sensitivity) of the OMPs and 90.2% (specificity) of the globular proteins are correctly identified.

3.2 Using homologous sequences to calculate residue composition can improve the performance

For each protein, we included homologous sequences into the calculation of residue composition as described in materials and methods. The comparison of residue compositions calculated using single sequence and homologous sequences are available at http://www.cs.usu.edu/∼cyan/OMP_KNN/supplement.htm. We repeated the five-fold cross-validations using the same dataset partition used in Section 3.1. Comparisons (Table 1, columns 2 and 3) show that using homologous information can improve the performance remarkably: the accuracy is increased to 96.0%; MCC is increased to 0.888; 87.5% of the OMPs (sensitivity) and 98.7% of the globular proteins (specificity) are correctly identified.

3.3 The proposed method can distinguish OMPs from α-helical membrane proteins (AMPs)

We then evaluated the proposed method's ability to discriminate OMPs from AMPs. Five-fold cross-validations were performed such that the identity between any protein from the training set and any protein from the test set is less than 25%. The results (Table 1, columns 4 and 5) show that when using single protein sequence as input, the proposed method can discriminate between OMPs and AMPs with 90.1% overall accuracy and 0.802 MCC. 88.9% (sensitivity) of OMPs and 91.3% (specificity) of AMPs are correctly predicted. When homologous sequences are used, the performance is improved remarkably, reaching 94.7% accuracy with 0.894 MCC, 95.7% sensitivity, and 93.7% specificity.

3.4 Discrimination between OMPs and non-OMPs

Based on the results obtained in the previous sections, we designed a method for discriminating OMPs from non-OMPs (AMPs + globular proteins). In the method, the composition of a test protein was calculated. The distances from the protein to OMPs, AMPs and globular proteins were calculated based on the *K-NN* method. If the distance to OMPs was the smallest among them, then the protein was

Table 1. Performance of the K-NN method evaluated using five-fold cross-validations

Classification	OMPs vs Globular		OMPs vs AMPs		OMPs vs Non-OMPs	
Mode	Single[a]	Homologous[b]	Single	Homologous	Single	Homologous
Accuracy	88.8%	96.0%	90.1%	94.7%	89.1%	96.1%
MCC	0.708	0.888	0.802	0.894	0.668	0.873
Sensitivity	84.1%	87.5 %	88.9%	95.7%	78.8%	87.5 %
Specificity	90.2%	98.7%	91.3%	93.7%	91.5%	98.2%

[a] For each protein, only the protein itself was used to calculate residue composition

[b] For each protein, 50 homologous proteins were included in the calculation of residue composition

predicted to be OMP. Otherwise, it was predicted to be non-OMP. The method was evaluated using five-fold cross-validations. The results (Table 1, columns 6 and 7) show that the proposed method achieves 96.1% accuracy with 0.873 MCC when homologous sequences are used. The distances to OMPs, AMPs and globular proteins for each protein are available at http://www.cs.usu.edu/∼cyan/OMP_KNN/supplement.htm.

3.5 Evaluating the proposed method using datasets with lower sequence similarity

In the datasets used so far, sequence similarity between two proteins can be as high as 40%. It will be important to know how the proposed method performs if the mutual similarities between sequences in datasets are lower. To investigate this, we filtered the datasets so that the sequence similarity between any two sequences was less than 25%. After the filtering, 112 OMPs, 673 globular proteins, and 178 AMPs were left. We then used these filtered datasets to evaluate the proposed method using five-fold cross-validation. When homologous sequences were used to calculate residue composition, we also required that the homologous sequences of any two sequences did not overlap. Thus, if one protein from the NCBI non-redundant database was homologous to both proteins A and B in the datasets, it was only used to calculate the residue composition of protein A (or B). We applied this requirement to prevent the case that one test protein was correctly classified only because its homologous sequences overlaped with the homologous sequences of some proteins in the training set. The results (Table 2) show that reducing the sequence similarity in the dataset only reduces the performance slightly. The proposed method can still discriminate OMPs and non-OMPs with very high performance: 95.3% accuracy with 0.760 MCC.

Table 2. Discrimination of outer membrane proteins (OMPs) from non-OMPs on datasets with sequence similarity less than 25%

	Single sequence[a]	Homologous sequences[b]
Accuracy	91.8% (*89.1%*)[c]	95.3% (*96.1%*)
MCC	0.606 (*0.668*)	0.760 (*0.873*)
Sensitivity	66.1% (*78.8%*)	72.3% (*87.5 %*)
Specificity	95.2% (*91.5%*)	98.4% (*98.2%*)

[a] For each protein, only the protein itself was used to calculate residue composition

[b] For each protein, 50 homologous proteins were included in the calculation of residue composition

[c] Values in parenthesis are the performance on the datasets with mutual sequence similarity less than 40%

Table 3. Comparison of the proposed method (K-NN) with BLAST search in the discrimination of OMPs and non-OMPs

	K-NN		BLAST search
	Single sequence[a]	Homologous sequences[b]	
Accuracy	91.8%	95.3%	77.6%
MCC	0.606	0.760	0.446
Sensitivity	66.1%	72.3%	88.4%
Specificity	95.2%	98.4%	76.1%

[a] For each protein, only the protein itself was used to calculate residue composition

[b] For each protein, 50 homologous proteins were included in the calculation of residue composition

The improvement of using homologous sequences is obvious: accuracy is improved from 91.8% to 95.3% and MCC is improved from 0.606 to 0.760.

3.6 Comparison with predictions solely based on similarity search

The results have shown that combining homologous information with the *K-NN* method can improve the performance dramatically. Then, how well it performs if we make the predictions solely based on homologous search? To explore this, for each test protein, we performed a homologous search on the training set using the BLAST program. The test protein was classified into the class of the protein from the training set that shares the highest similarly with the test protein. We evaluated this approach using the same five-fold cross-validations that we used to evaluate our *K-NN* method. The results (Table 3) show that the BLAST search approach only achieves 77.6% accuracy with 0.446 MCC. In comparison, the *K-NN* method achieves as high as 95.3% accuracy and 0.760 MCC on the same datasets.

3.7 Comparisons with other methods

We also compare the proposed method with multiple previously published methods. As discussed in Baldi et al. (2000), in a two-class classification, if the numbers of examples of the two classes are not equal, MCC is a better measure for evaluating the classification performance. In many studies, MCC has been used as the standard for comparing different predicting methods (Bao and Cui, 2005; Dobson et al., 2006; Ye et al., 2007). In the discrimination of OMPs and non-OMPs, the numbers of examples in the two classes are not equal. Therefore, we will

Table 4. Comparisons of different methods in the discrimination of OMPs and non-OMPs

	Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
K-NN method[a]	**95.7 ± 0.4**	**0.858 ± 0.011**	86.3 ± 1.0	97.9 ± 0.3
Neural Network[b,c] (Gromiha and Suwa, 2006)	91.0	0.716	79.3	93.8
Support Vector Machine[b] (Park et al., 2005)	93.9	0.816	90.9	94.7

[a] The method proposed in this study. The five-fold cross-validations performed in this study require that the sequence similarity between any protein from the training set and any protein from the test sets is less than 25%. The five-fold cross-validations are repeated five times. The average and standard deviation are reported

[b] The statistics are obtained from the original publications (Gromiha and Suwa, 2006; Park et al., 2005). Note that the original studies used the same datasets and the same type of cross-validation (five-fold cross-validations) as the current study. But the sequence similarity between training and test sets can be as high as 40%. In the original publication, only Accuracy, Sensitivity and Specificity were reported. Here, we calculate the MCC based on their published statistics

[c] In their study, Gromiha and Suwa (2006) evaluated 11 different methods. Neural network was reported to be the best

use MCC as the primary measure in the comparison of different methods. At the same time, we also report accuracy, specificity and sensitivity.

Gromiha and Suwa (2006) tried a set of 11 machine learning methods for the discrimination of OMPs using residue composition as input. One of the 11 methods was a *k*-nearest neighbor method based on Euclidean distance. Neural network method was reported to achieve the best performance in their study. In another study, researcher from the same group (Park et al., 2005) developed a support vector machine (SVM) method to discriminate OMPs. Both studies used the same datasets that we use in this study to evaluate their methods based on five-fold cross-validations. So, we compared the results we obtained in the current study with what their reported in their publications. To assess the statistical significance of the comparison, we evaluated our *K-NN* method by repeating the five-fold cross-validations five times using different data splits. The average and the standard deviation are shown in Table 4 (row 2). The results (Table 4) show that our *K-NN* outperforms all of the 11 methods used in Gromiha and Suwa's study (2006) and the SVM method developed by Park et al. (2005). Note that not all of the 11 methods from Gromiha and Suwa's study (2006) are shown in Table 4. Since neural network was reported to achieve the best performance among the 11 methods, we only show the results of the neural network method in the table. Table 4 shows that the MCC of *K-NN* is remarkably higher than those of neural network and SVM. The differences are larger than 3 times of the deviation in both cases. This confirms the statistical significance of the improvement. It is also worth to point out that in the five-fold cross-validations performed in the current study, we made sure that the sequence similarity between any protein from the training set and any protein from the test sets is less than 25%. Meanwhile, in the studies of Gromiha and Suwa (2006) and Park et al. (2005), the sequence similarity between training and test sets can be as high as 40%. Although we use a stricter criterion to evaluate our *K-NN*, the performance of our method is still better than what were reported for the other two methods using a looser criterion.

Berven et al. (2004) developed a method, BOMP, for the prediction of OMPs. It is one of the best scoring methods in identifying OMPs from genome. The BOMP method combines a pattern search, a β-barrel score based on amino acid distribution, and a filter that explores the abundance of Asparagine and Isoleucine in the protein. Here,

Table 5. Comparison of the proposed method (*K-NN*) and the BOMP method (Berven et al., 2004)

		Accuracy (%)	MCC	Sensitivity (%)	Specificity (%)
Datasets used in this study[a]	*K-NN*	**95.6 ± 0.2**	**0.774 ± 0.011**	74.3 ± 1.5%	98.4 ± 0.2%
	BOMP	93.1	0.623	52.7	98.5
Datasets used in Berven et al. (2004)	*K-NN*	98.8	0.870	83.1	99.6
	BOMP	98.3	0.831	88.1	98.8

[a] The dataset used in this study was submitted to the BOMP server. The dataset submitted to the BOMP server is likely overlap with the dataset that BOMP was trained on. On the contrary, in the evaluation of our K-NN method, we make sure that the sequence similarity between training and test sets is less than 25%

we also compared our *K-NN* method with BOMP. First, we submitted the filtered datasets (in which similarity between any two proteins is less than 25%) used in this study to the BOMP server. The results (Table 5) show that our *K-NN* method outperforms BOMP. The MCC of *K-NN* outperforms that of BOMP by more than 10 times of the standard deviation. This confirms the statistical significance of the improvement. It is worth to point out that in this comparison, the dataset submitted to the BOMP server is likely overlap with the dataset that BOMP was trained on. On the contrary, in the evaluation of our *K-NN* method, we make sure that the sequence similarity between training and test sets is less than 25%. We also evaluated our *K-NN* method using the same datasets that Berven et al. (2004) used to evaluate their BOMP method. The results (Table 5) show that our *K-NN* method still outperforms BOMP on their datasets. When BOMP datasets were used, leave-one-out cross validations were used to evaluate both methods as described in Berven et al. (2004). We notice that when compared using the BOMP dataset, the improvement of *K-NN* method over BOMP is not so big as when our dataset is used. The possible reason is that the BOMP dataset contains only a small number of positive examples (59 in total). Since leave-one-out cross-validations were performed, we could not calculate the standard deviation as we did in five-fold cross-validation (because there is only one possible way to split data for leave-one-out cross validation). However, the improvement of the *K-NN* method in MCC is still clear.

3.8 Receiver operating characteristic (ROC) curve

In the *K-NN* method, a protein is classified as OMP or non-OMP based on the comparison of $\overline{D}_{\mathrm{omp}}$ (its distance to the OMP group), $\overline{D}_{\mathrm{glo}}$ (its distance to the globular protein group), and $\overline{D}_{\mathrm{imp}}$ (its distance to the AMP group). A protein is predicted to be OMP if $\overline{D}_{\mathrm{omp}} < \mathrm{Min}\{\overline{D}_{\mathrm{imp}}, \overline{D}_{\mathrm{glo}}\}$, where $\mathrm{Min}\{\cdot\}$ returns the minimal value of a set. This criteria is equal to $\overline{D}_{\mathrm{omp}} - \mathrm{Min}\{\overline{D}_{\mathrm{imp}}, \overline{D}_{\mathrm{glo}}\} < 0$. In general, we can introduce a threshold parameter α, such that a protein is predict to be OMP if $\overline{D}_{\mathrm{omp}} - \mathrm{Min}\{\overline{D}_{\mathrm{imp}}, \overline{D}_{\mathrm{glo}}\} < \alpha$. Figure 1 shows the ROC curve of the *K-NN* method obtained by varying α. The ROC curve shows how the *K-NN* method can tradeoff between specificity and sensitivity by changing values of the parameter. When applying a prediction method to identify OMPs, some researchers may prefer to identify more potential OMPs (high sensitivity) at the cost of relatively low specificity, others may want to identify OMPs with very high specificity at the cost of relatively low sensitivity. The advantage of introducing this parameter α to the *K-NN* method is that users can chose a threshold based on their need. When α is set to a lower value, the *K-NN* can achieve higher specificity. On other hand, when a high value of α is chosen, the *K-NN* can achieve higher sensitivity.

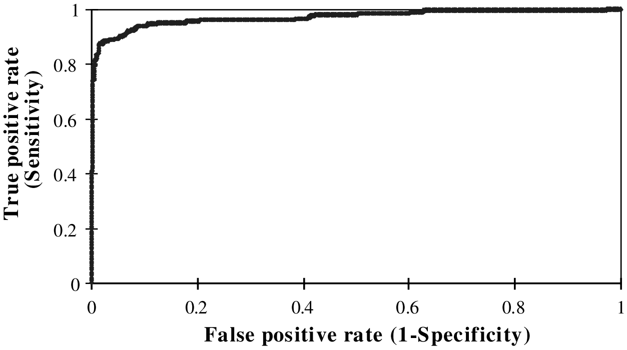

Fig. 1. ROC curve of the K-NN method

3.9 Identification of OMPs in the proteome of E.coli

We applied the *K-NN* method to search for OMPs in the proteome of *E. coli* using $\alpha = -0.11$, which corresponds to 99% specificity in the ROC curve. The E. Coli proteome consists of 4319 proteins. 123 of them were predicted to be OMP proteins. That accounts for 2.8% of the whole proteome. This ratio is consistent with the previous estimation that 2–3% of the genes in Gram-negative bacteria encodes OMPs (Wimley, 2003). Among these 123 hits, 61 proteins are annotated as OMP proteins in Swiss-Prot (Bairoch et al., 2004) or ePSORTdb (Rey et al., 2005), a database of protein subcellular locations that have been determined through laboratory experiments. In addition, 20 proteins are annotated with "Membrane", "Cell membrane" and "Multi-pass membrane protein" in Swiss-Prot. We submitted these proteins to the TMHMM (Krogh et al., 2001), a server for predicting the topology of trasmembrane α-helical proteins, and PSORTb (Gardy et al., 2005), a server for predicting subcellular locations. Only 1 of them was predicted to be trasmembrane α-helical proteins or inner membrane proteins by both methods. Thus, most of these 20 proteins are very likely OMP proteins. The remaining 42 proteins may suggest new OMP proteins that have not been previously discovered. Ultimate verifications of these predictions will require laboratory experiments to determine the subcellular locations of these proteins and are beyond the scope of this paper. The IDs of these 20 and 42 proteins are available at http://www.cs.usu.edu/∼cyan/OMP_KNN/supplement.htm.

4. Discussion

In summary, we have presented a *K-NN* method that can discriminate outer membrane proteins (OMPs) and non-OMPs with high performance: 96.1% accuracy, with 0.873 MCC, 87.5% sensitivity, and 98.2% specificity. Comparisons with other top-scoring methods show that this method achieves better performance than the others do.

4.1 Simple methods versus complicated methods

It was estimated that 2–3% of the genes in Gram-negative bacteria encodes OMPs (Wimley, 2003). Identifying all OMPs (“OMPome”) from bacterial genome is an urgent and challenging task. Due to the large size of genomes, computational methods that can accomplish this task with a fast speed are demanded. Compared with other published methods, the proposed method not only achieves better performance but also has the advantage of simplicity and fast speed. The training process of the method is simple and straightforward. The calculation of residue composition and weighted Euclidean distance can be done with a fast speed. The method proposed here will be very helpful to the discovery of “OMPome” in a genome scale.

4.2 Euclidean distance versus weighted Euclidean distance

The proposed *K-NN* method makes prediction based on the weighted Euclidean distance (i.e., $\sqrt{\Sigma_i \frac{(\bar{x}_i - x_i)^2}{\bar{x}_i}}$). Comparisons show that this method achieves better results than the *K-NN* method based on the standard Euclidian distance ($\sqrt{\Sigma_i (\bar{x}_i - x_i)^2}$). Compared with the standard Euclidian distance, the weighted Euclidian distance is a better measure to evaluate the relationship between a protein and a group. For example, for the same amount of difference between x_i and $\bar{x}_i$ (without loss of generality, let $\bar{x}_j - x_j = 0.01$), where x_i is the residue composition in the test protein, and $\bar{x}_i$ is the average residue composition of OMP proteins, if $x_i = 0.89$ and $\bar{x}_i = 0.90$, then this difference ($\bar{x}_i - x_i = 0.01$) does not present a significant distance between the test protein and the TMB group. But, if $x_i = 0.001$ and $\bar{x}_i = 0.011$, then $\bar{x}_j - x_j = 0.01$ will become a significant distance between the test protein and the OMB group. After assigning a weight of $1/\bar{x}_i$ to the term, significant differences between x_i and $\bar{x}_i$ will be given larger values in the weighted Euclidian distance.

Authors’ contributions

CY conceived of and designed the study, performed the analysis and drafted the manuscript. YW contributed to most of the computation. JH carried out the comparison between *K-NN* and BOMP using the BOMP datasets. All authors read and approved the final manuscript.

Acknowledgements

This work was supported by funding from the Department of Computer Science, Utah State University to CY.

References

Ahmad S, Sarai A (2004) Moment-based prediction of DNA-binding proteins. J Mol Biol 341: 65–71

Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402

Bagos P, Liakopoulos T, Spyropoulos I, Hamodrakas S (2004a) A hidden Markov model method, capable of predicting and discriminating beta-barrel outer membrane proteins. BMC Bioinformatics 5: 29

Bagos PG, Liakopoulos TD, Spyropoulos IC, Hamodrakas SJ (2004b) PRED-TMBB: a web server for predicting the topology of beta-barrel outer membrane proteins, Nucleic Acids Res 32: W400–W404

Bairoch A, Boeckmann B, Ferro S, Gasteiger E (2004) Swiss-Prot: Juggling between evolution and stability. Brief Bioinform 5: 39–55

Baldi P, Brunak S, Chauvin Y, Andersen CAF (2000) Assessing the accuracy of prediction algorithms for classification: an overview. Bioinformatics 16: 412–424

Bao L, Cui Y (2005) Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21: 2185–2190

Berven FS, Flikka K, Jensen HB, Eidhammer I (2004) BOMP: a program to predict integral beta-barrel outer membrane proteins encoded within genomes of Gram-negative bacteria. Nucleic Acids Res 32: W394–W399

Cao Y, Liu S, Zhang L, Qin J, Wang J, Tang K (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7: 20

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243: 444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357: 116–121

Chen YL, Li QZ (2007) Prediction of apoptosis protein subcellular location using improved hybrid approach and pseudo amino acid composition. J Theor Biol 248: 377–381

Chen J, Liu H, Yang J, Chou KC (2007) Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33: 423–428

Chou KC (2002) A new branch of proteomics: prediction of protein cellular attributes. In: Weinrer PW, Lu Q (eds) Gene cloning and expression technologies, Chapter 4. Eaton Publishing, Westborough, MA, pp 57–70

Chou KC, Cai YD (2005a) Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Modeling 45: 407–413

Chou KC, Cai YD (2005b) Using GO-PseAA predictor to identify membrane proteins and their types. Biochem Biophys Res Commun 327: 845–847

Chou KC, Elrod DW (1999) Prediction of membrane protein types and subcellular locations. Proteins Struct Funct Genet 34: 137–153

Chou KC, Shen HB (2006a) Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun 347: 150–157

Chou KC, Shen HB (2006b) Large-scale predictions of Gram-negative bacterial protein subcellular locations. J Proteome Res 5: 3420–3428

Chou KC, Shen HB (2006c) Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J Proteome Res 5: 1888–1897

Chou KC, Shen HB (2007a) Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res 6: 1728–1734

Chou KC, Shen HB (2007b) Large-scale plant protein subcellular location prediction. J Cell Biochem 100: 665–678

Chou KC, Shen HB (2007c) MemType-2L: a web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Commun 360: 339–345

Chou KC, Shen HB (2007d) Review: recent progresses in protein subcellular location prediction. Anal Biochem 370: 1–16

Chou KC, Shen HB (2007e) Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Commun 357: 633–640

Chou KC, Zhang CT (1995) Review: prediction of protein structural classes. Crit Rev Biochem Mol 30: 275–349

Deng Y, Liu Q, Li Y-X (2004) Scoring hidden Markov models to discriminate beta-barrel membrane proteins. Comput Biol Chem 28: 189–194

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2007) Using pseudo amino acid composition to predict transmembrane regions in protein: cellular automata and Lempel-Ziv complexity. Amino Acids, DOI: 10.1007/s00726-007-0550-z

Ding YS, Zhang TL, Chou KC (2007) Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Peptide Lett 14: 811–815

Fang Y, Guo Y, Feng Y, Li M (2007) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids, DOI: 10.1007/s00726-007-0568-2

Dobson R, Munroe P, Caulfield M, Saqi M (2006) Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics 7: 217

Douglas SM, Chou JJ, Shih WM (2007) DNA-nanotube-induced alignment of membrane proteins for NMR structure determination. Proc Natl Acad Sci USA 104: 6644–6648

Du P, Li Y (2006) Prediction of protein submitochondria locations by hybridizing pseudo-amino acid composition with various physicochemical features of segmented sequence. BMC Bioinformatics 7: 518

Fang Y, Guo Y, Feng Y, Li M (2007) Predicting DNA-binding proteins: approached from Chou's pseudo amino acid composition and other specific sequence features. Amino Acids, DOI: 10.1007/s00726-007-0568-2

Gao QB, Wang ZZ (2006) Classification of G-protein coupled receptors at four levels. Protein Eng Des Sel 19: 511–516

Gao QB, Wang ZZ, Yan C, Du YH (2005a) Prediction of protein subcellular location using a combined feature of sequence. FEBS Lett 579: 3444–3448

Gao Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005b) Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter. Amino Acids 28: 373–376

Gardy JL, Laird MR, Chen F, Rey S, Walsh CJ, Ester M, Brinkman FSL (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis, Bioinformatics 21: 617–623

Garrow A, Agnew A, Westhead D (2005a) TMB-Hunt: an amino acid composition based method to screen proteomes for beta-barrel transmembrane proteins. BMC Bioinformatics 6: 56

Garrow AG, Agnew A, Westhead DR (2005b) TMB-Hunt: a web server to screen sequence sets for transmembrane beta-barrel proteins. Nucleic Acids Res 33: W188–W192

Gnanasekaran TV, Peri S, Arockiasamy A, Krishnaswamy S (2000) Profiles from structure based sequence alignment of porins can identify beta stranded integral membrane proteins. Bioinformatics 16: 839–842

Gromiha MM, Suwa M (2005) A simple statistical method for discriminating outer membrane proteins with better accuracy. Bioinformatics 21: 961–968

Gromiha MM, Suwa M (2006) Discrimination of outer membrane proteins using machine learning algorithms. Proteins 63: 1031–1037

Guo J, Lin Y, Liu X (2006a) GNBSL: a new integrative system to predict the subcellular location for Gram-negative bacteria proteins. Proteomics 6: 5099–5105

Guo YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006b) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30: 397–402

Huang Y, Li Y (2004) Prediction of protein subcellular locations using fuzzy k-NN method. Bioinformatics 20: 21–28

Jahandideh S, Abdolmaleki P, Jahandideh M, Asadabadi EB (2007) Novel two-stage hybrid neural discriminant model for predicting proteins structural classes. Biophys Chem 128: 87–93

Kedarisetti KD, Kurgan LA, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348: 981–988

Koebnik R, Locher KP, Van Gelder P (2000) Structure and function of bacterial outer membrane proteins: barrels in a nutshell. Mol Microbiol 37: 239–253

Krogh A, Larsson B, Heijne Gv, Sonnhammer ELL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567–580

Li FM, Li QZ (2007) Using pseudo amino acid composition to predict protein subnuclear location with improved hybrid approach. Amino Acids, DOI: 10.1007/s00726-007-0545-9

Lin H, Li QZ (2007a) Predicting conotoxin superfamily and family by using pseudo amino acid composition and modified Mahalanobis discriminant. Biochem Biophys Res Commun 354: 548–551

Lin H, Li QZ (2007b) Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components. J Comput Chem 28: 1463–1466

Liu DQ, Liu H, Shen HB, Yang J, Chou KC (2007) Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32: 493–496

Liu H, Wang M, Chou KC (2005a) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336: 737–739

Liu H, Yang J, Wang M, Xue L, Chou KC (2005b) Using Fourier spectrum analysis and pseudo amino acid composition for prediction of membrane protein types. Protein J 24: 385–389

Liu Q, Zhu Y, Wang B, Li Y (2003) Identification of beta-barrel membrane proteins based on amino acid composition properties and predicted secondary structure. Comp Biol Chem 27: 355–361

Martelli PL, Fariselli P, Krogh A, Casadio R (2002) A sequence-profile-based HMM for predicting and discriminating beta barrel membrane proteins. Bioinformatics 18: S46–S53

Mondal S, Bhavna R, Mohan Babu R, Ramakumar S (2006) Pseudo amino acid composition and multi-class support vector machines approach for conotoxin superfamily classification. J Theor Biol 243: 252–260

Niu B, Cai YD, Lu WC, Zheng GY, Chou KC (2006) Predicting protein structural class with AdaBoost learner. Protein Peptide Lett 13: 489–492

Park K-J, Gromiha MM, Horton P, Suwa M (2005) Discrimination of outer membrane proteins using support vector machines. Bioinformatics 21: 4223–4229

Prasad Bahadur R, Chakrabarti P, Rodier F, Janin J (2004) A dissection of specific and non-specific protein–protein interfaces. J Mol Biol 336: 943–955

Pu X, Guo J, Leung H, Lin Y (2007) Prediction of membrane protein types from sequences and position-specific scoring matrices. J Theor Biol 247: 259–265

Rey S, Acab M, Gardy JL, Laird MR, deFays K, Lambert C, Brinkman FSL (2005) PSORTdb: a protein subcellular localization database for bacteria. Nucleic Acids Res 33: 164–168

Rost B, Sander C (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 90: 7558–7562

Schulz GE (2000) Beta-barrel membrane proteins. Curr Opin Struct Biol 10: 443–447

Shen HB, Chou KC (2005a) Predicting protein subnuclear location with optimized evidence-theoretic K-nearest classifier and pseudo amino acid composition. Biochem Biophys Res Commun 337: 752–756

Shen HB, Chou KC (2005b) Using optimized evidence-theoretic K-nearest neighbor classifier and pseudo amino acid composition to predict membrane protein types. Biochem Biophys Res Commun 334: 288–292

Shen HB, Chou KC (2006) Ensemble classifier for protein fold pattern recognition. Bioinformatics 22: 1717–1722

Shen HB, Chou KC (2007a) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364: 53–59

Shen HB, Chou KC (2007b) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Design Selection 20: 39–46

Shen HB, Chou KC (2007c). Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355: 1006–1011

Shen HB, Chou KC (2007) Using ensemble classifier to identify membrane protein types. Amino Acids 32: 483–488

Shen HB, Yang J, Chou KC (2007) Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33: 57–67

Shi JY, Zhang SW, Pan Q, Cheng Y-M, Xie J (2007) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33: 69–74

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30: 469–475

Tan F, Feng X, Fang Z, Li M, Guo Y, Jiang L (2007) Prediction of mitochondrial proteins based on genetic algorithm – partial least squares and support vector machine. Amino Acids, DOI: 10.1007/s00726-006-0465-0

Waldispuhl J, Berger B, Clote P, Steyaert J-M (2006) TransFold: a web server for predicting the structure and residue contacts of transmembrane beta-barrels. Nucleic Acids Res 34: W189–W193

Wang M, Yang J, Chou KC (2005) Using string kernel to predict signal peptide cleavage site based on subsite coupling model. Amino Acids (Erratum, ibid. 2005, 29: 301) 28: 395–402

Wen Z, Li M, Li Y, Guo Y, Wang K (2006) Delaunay triangulation with partial least squares projection to latent structures: a model for G-protein coupled receptors classification and fast structure recognition. Amino Acids 32: 277–283

Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34: W243–W248

Wimley WC (2002) Toward genomic identification of beta-barrel membrane proteins: composition and architecture of known structures. Protein Sci 11: 301–312

Wimley WC (2003) The versatile beta-barrel membrane protein. Curr Opin Struct Biol 13: 404–411

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Peptide Lett 14: 871–875

Xiao X, Shao S, Ding Y, Huang Z, Huang Y, Chou KC (2005) Using complexity measure factor to predict protein subcellular location. Amino Acids 28: 57–61

Xiao X, Shao SH, Ding YS, Huang ZD, Chou KC (2006) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30: 49–54

Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D, Honavar V (2006) Predicting DNA-binding sites of proteins from amino acid sequence. BMC Bioinformatics 7: 262

Ye Z-Q, Zhao S-Q, Gao G, Liu X-Q, Langlois RE, Lu H, Wei L (2007) Finding new structural and sequence attributes to predict possible disease association of single amino acid polymorphism (SAP). Bioinformatics 23: 1444–1450

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006) Prediction protein homo-oligomer types by pseudo amino acid composition: approached with an improved feature extraction and naive Bayes feature fusion. Amino Acids 30: 461–468

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids, 10.1007/s00726-007-0496-1

Zhou XB, Chen C, Li ZC, Zou XY (2007) Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. J Theor Biol 248: 546–551

Authors' address: Changhui Yan, Old Main Hill 4205, Logan, UT 84322-4205, USA,
Fax: (435) 797-3265, E-mail: charles.yan@usu.edu