Amino Acids (2008) 35:573–580
DOI 10.1007/s00726-008-0083-0

ORIGINAL ARTICLE

An ensemble of support vector machines for predicting the membrane protein type directly from the amino acid sequence

Loris Nanni · Alessandra Lumini

Received: 21 January 2008 / Accepted: 26 February 2008 / Published online: 22 April 2008

Abstract Given a particular membrane protein, it is very important to know which membrane type it belongs to because this kind of information can provide clues for better understanding its function. In this work, we propose a system for predicting the membrane protein type directly from the amino acid sequence. The feature extraction step is based on an encoding technique that combines the physicochemical amino acid properties with the residue couple model. The residue couple model is a method inspired by Chou's quasi-sequence-order model that extracts the features by utilizing the sequence order effect indirectly. A set of support vector machines, each trained using a different physicochemical amino acid property combined with the residue couple model, are combined by vote rule. The success rate obtained by our system on a difficult dataset, where the sequences in a given membrane type have a low sequence identity to any other proteins of the same membrane type, are quite high, indicating that the proposed method, where the features are extracted directly from the amino acid sequence, is a feasible system for predicting the membrane protein type.

Keywords Membrane type prediction · Residue couple model · Chou's pseudo–amino acid composition · Ensemble of classifiers · Physicochemical properties · Support vector machine

L. Nanni (✉) · A. Lumini
DEIS, IEIIT—CNR, Università di Bologna,
Viale Risorgimento 2, 40136 Bologna, Italy
e-mail: loris.nanni@unibo.it

Introduction

The function of a membrane protein is closely related to the type it belongs to (Chou and Elrod 1999a; Lodish et al. 1995). It is very interesting to note that 20–35% of genes encode membrane proteins, but only 1% of proteins whose three-dimensional structure is known are membrane proteins (Douglass et al. 2007). Unfortunately, it is expensive to experimentally identify the membrane type of a membrane protein; for this reason it is very important to develop a fast and reliable method for predicting the membrane protein type. Many membrane proteins are important targets for drug discovery (see e.g, Doyle et al. 1998; Chou 2004; Schnell and Chou 2008). In view of this, a fast and reliable method for predicting the membrane protein type is highly desired. In particular, the membrane proteins can move around the cell membrane, and hence knowing the type of a membrane protein can provide insight into this kind of motion (Chou and Shen 2007c).

In the last few years some membrane type prediction methods have been proposed in the literature (Liu et al. 2005; Pu et al. 2007; Shen et al. 2007a; Wang et al. 2004). The main drawback of these methods is that small datasets (five classes) without rigorous screened by a data-culling operation to avoid redundancy (Chou and Shen 2007c) have been used to validate the proposed methods.

In Chou and Shen (2007c), an ensemble of optimized evidence-theoretic k-nearest neighbor classifiers was proposed to predict the membrane proteins and their types. To extract the features from a given protein, a method that considers the evolution information of the protein is proposed. The pseudo–amino acid composition (Chou 2005) was adopted to extract the features from the position-specific scoring matrix (Mundra et al. 2007). The ensemble of classifiers is built, perturbing both the feature set and the parameter of the classifier. Notice that the method proposed in Chou and Shen (2007c), named MemType-2L, is a two-layer predictor: the first layer recognizes a protein as membrane or non-membrane; the second layer identifies the membrane type of the membrane proteins. Another interesting contribution of Chou and Shen (2007c) is that the authors collect a benchmark dataset of eight categories, and a redundancy cut-off was used to avoid redundancy among the sequences that belong to the dataset.

The eight categories of the proposed dataset are (1) type I, (2) type II, (3) type III, and (4) type IV transmembranes, (5) multipass transmembrane, (6) lipid-chain-anchored membrane, (7) GPI-anchored membrane, and (8) peripheral membrane. Please refer to Cedano et al. (1997) and Chou and Shen (2007c) for a detailed description of these categories.

The prediction of the membrane protein type is a problem of subcellular localization. In the literature, several subcellular localizers, mainly based on analysis of the amino acid sequence, are proposed (Cai et al. 2000, 2002; Cedano et al. 1997; Chou 2000, 2001; Chou and Cai 2002, 2003, 2004a, b, 2005; Chou and Elrod 1998, 1999a, b; Chou and Shen 2006; Nakai and Horton 1999; Nakai and Kanehisa 1992; Yuan 1999; Nanni and Lumini 2008) as well relevant references are cited in a recent review article (Chou and Shen 2007d). Moreover, recently some web-servers for predicting protein subcellular localization with both single and multiple sites have been established (Chou and Shen 2007a, 2008; Shen and Chou 2007a).

In the literature on the membrane-protein-type prediction systems, it has been shown that the methods in which the features are directly extracted from the amino acid sequence do not perform as well as the feature extraction methods based on the position-specific scoring matrix (Chou and Shen 2007c; Pu et al. 2007). In this paper, we deal with the membrane-protein-type prediction problem using an ensemble of support vector machines trained using features extracted directly from the amino acid sequence. We show that the ensemble of classifiers, where each classifier is trained considering a different physicochemical property, outperforms the standard method based on the residue couple model and that our idea partially fills the performance gap between the feature extraction methods based on the amino acid sequence and the feature extraction methods based on the position-specific scoring matrix. Since there are hundreds of physicochemical properties, a physicochemical property selection is performed by sequential forward floating selection (Nanni and Lumini 2006a, b), where the objective function is the minimization of the error rate in the training set. Notice that all the parameters of the proposed method are calculated on the training set, then the performance is calculated on an independent set.

Table 1 Number of membrane proteins in each of the eight types

Type	Training set	Independent set
Single-pass type I	610	444
Single-pass type II	312	78
Single-pass type III	24	6
Single-pass type IV	44	12
Multipass	1,316	3,265
Lipid-chain anchor	151	38
GPI anchor	182	46
Peripheral	610	444
Overall	3,249	4,333

Materials and methods

In this paper the same dataset used in Chou and Shen (2007c) is used to assess the performance of the proposed method. The protein sequences were collected from the Swiss-Prot database (http://www.ebi.ac.uk/swissprot/; version 51.0 released on 6 October 2006). The proteins that contain fewer than 50 amino acid residues were excluded; moreover, in contrast to the previous membrane type datasets, several categories (eight) are considered. To avoid redundancy among the data, the proteins were screened strictly by a cut-off procedure based on the sequence identity among proteins of the same type. In this paper, we use the training set to find the parameters of the proposed method and for training the classifiers, then we use an independent set to validate our results. The number of membrane proteins in each of the eight types is reported in Table 1.

In this work, we extend the quasi-residue couple model proposed in Nanni (2006), and we study the performance of the combination of the residue couple model with each of the physicochemical properties obtained in the amino acid index database[1] (Kawashima and Kanehisa 2000) (available at http://www.genome.jp/dbget/aaindex.html).

An amino acid index is a set of 20 numerical values representing any of the different physicochemical properties of the amino acids (Nanni and Lumini 2006a, b).

The extraction of the quasi-residue couple features[2] for a physicochemical property d from a given protein is obtained as:

$$\mathbf{P1}_{i,j}^{m} = (1/(L-m)) \times \left[\sum\nolimits_{n=1:L-m} \mathrm{H1}_{i,j}(n, n+m, d)\right] \quad (1)$$

$$\mathbf{P2}_{i,j}^{m} = (1/(L-m)) \times \left[\sum\nolimits_{n=1:L-m} \mathrm{H2}_{i,j}(n, n+m, d)\right] \quad (2)$$

where the values of i and j range from 1 to 20 representing the 20 different amino acids; $\mathrm{H1}_{i,j}(n,n+m,d) = index(i,d)$ if

[1] This database currently contains 544 such indices and 94 substitution matrices.

[2] To avoid any problem in the re-implementation of the method, the Matlab code is available in the Appendix of this paper.

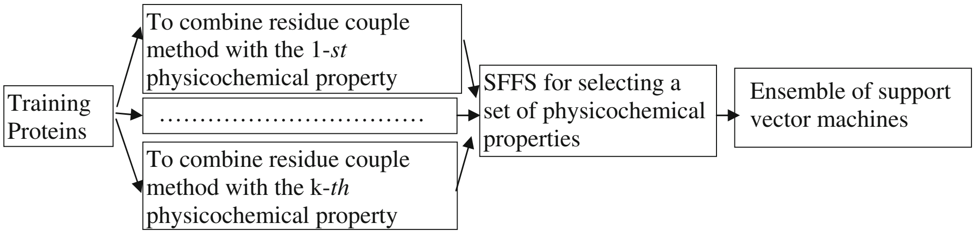

Training Proteins
To combine residue couple method with the 1-*st* physicochemical property
……………………………
To combine residue couple method with the k-*th* physicochemical property
SFFS for selecting a set of physicochemical properties
Ensemble of support vector machines

Fig. 1 System proposed

the amino acid in location n is i and the one in location $n+m$ is j, otherwise $H1_{i,j}(n,n+m,d) = 0$; $H2_{i,j}(n,n+m,d) = index(j,d)$ if the amino acid in location n is i and the one in location $n+m$ is j, otherwise $H2_{i,j}(n,n+m,d) = 0$; L is the length of the protein sequence**;** $index(p,d)$ is the function returning the value of the physicochemical property p of the amino-acid d; the parameter m is called the rank of the residue couple model. The vector that describes a given protein is given by the concatenation of **P1** and **P2**. In this paper, we extract the features using the first three ranks (i.e., m range from 1 to 3); in this way, the final vector is 2,400-dimensional.

For each physicochemical property, a different support vector machine is trained. Among these hundreds of physicochemical properties, a small set is selected by running the sequential forward floating selection[3] (SFFS) [as in Nanni and Lumini (2006a, b)], where the objective function is the minimization of the error rate in the training set. The support vector machine is a machine learning algorithm based on statistical learning theory that was introduced by Vapnik (Cristianini and Shawe-Taylor 2000). It searches for an optimal separating hyperplane that maximizes the margin in feature space.

In Fig. 1 our system is detailed.

In Fig. 2 we show an example of feature extraction where the Alpha-CH chemical-shifts property is used.

Experimental results

In this work we have used the dataset described in Chou and Shen (2007c), and the error rate in percentage is used as a parameter to evaluate the proposed system.

In Fig. 3, we report the error rate, obtained by a five-fold cross-validation on the training set, varying the number K of physicochemical properties selected by SFFS. In statistical prediction, the subsampling (such as fivefold or tenfold cross-validation) test and jackknife test are two cross-validation methods often used in the literature for examining the accuracy of a predictor (Chou and Zhang 1995). Because there are too many ways to sub-sample a given dataset (Chou and Shen 2007b), only the jackknife test can yield an objective unique result for a given benchmark dataset (Chou and Shen 2008). Accordingly, the jackknife test has been increasingly and widely adopted by investigators to test the power of various predictors (Cao et al. 2006; Chen et al. 2006a, b, 2007; Chen and Li 2007; Diao et al. 2007a, b; Ding et al. 2007; Du and Li 2006; Fang et al. 2007; Gao and Wang 2006; Gao et al. 2005a, b; Guo et al. 2006a, b; Huang and Li 2004; Jahandideh et al. 2007; Kedarisetti et al. 2006; Li and Li 2007; Lin and Li 2007a, b; Liu et al. 2007; Mondal et al. 2006; Mundra et al. 2007; Nanni and Lumini 2008a, b; Niu et al. 2006; Pugalenthi et al. 2007; Shen and Chou 2007a, b; Shen et al. 2007; Shi et al. 2007a, b; Sun and Huang 2006; Tan et al. 2007; Wang et al. 2005; Wen et al. 2007; Xiao and Chou 2007; Xiao et al. 2005, 2006; Zhang et al. 2006a, 2007; Zhang and Ding 2007; Zhang et al. 2006b; Zhou 1998; Zhou and Assa-Munt 2001; Zhou and Doctor 2003; Zhou et al. 2007a, b). However, since it would take too much computational time to perform the jackknife test (particularly by SVM), as a compromise here we used the fivefold cross-validation to test our method.

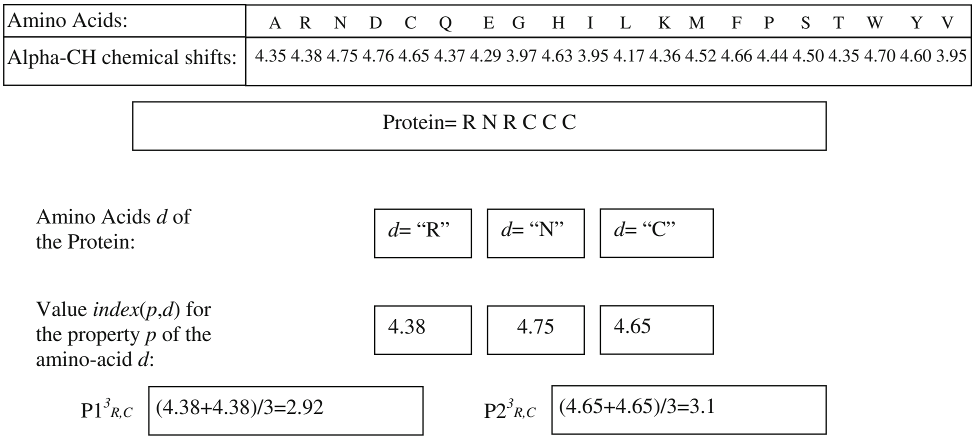
We test two different classifiers: linear support vector machine (LSVM) and radial basis function support vector machine (RSVM)[4]. Before the classification, the features are linearly normalized between 0 and 1. The best performance was obtained by RSVM with $K = 13$.

In Table 2, we report performance with the independent dataset using the 13 physicochemical properties selected by SFFS in the training set. With VOTE, we included the fusion-by-vote rule[5] among the classifiers trained using the 13 physicochemical properties. With NO, we named the RSVM trained by our feature extraction where the value of the physicochemical property for each amino-acid was not considered, as in the standard residue couple model (Pu et al. 2007) (i.e., $H1_{i,j}(n,n+m) = 1$ if the amino acid in location n is i and the one in location $n+m$ is j, otherwise

[3] Implemented as in the PRTools 3.1.7 Matlab toolbox.

[4] The support vector machine is implemented as in the OSU svm Matlab toolbox; the parameters of RSVM are $C = 0.1$ and gamma = 100.

[5] In the vote rule, all votes from the classifiers are tallied, and the class with the most votes represents the final prediction (Nanni and Lumini 2006c).



Amino Acids:	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Alpha-CH chemical shifts:	4.35	4.38	4.75	4.76	4.65	4.37	4.29	3.97	4.63	3.95	4.17	4.36	4.52	4.66	4.44	4.50	4.35	4.70	4.60	3.95

Protein= R N R C C C

Amino Acids *d* of the Protein:	*d*= “R”	*d*= “N”	*d*= “C”
Value *index*(*p*,*d*) for the property *p* of the amino-acid *d*:	4.38	4.75	4.65

$P1^{3}_{R,C}$ (4.38+4.38)/3=2.92

$P2^{3}_{R,C}$ (4.65+4.65)/3=3.1

Fig. 2 Example of feature extraction (with $m = 3$)

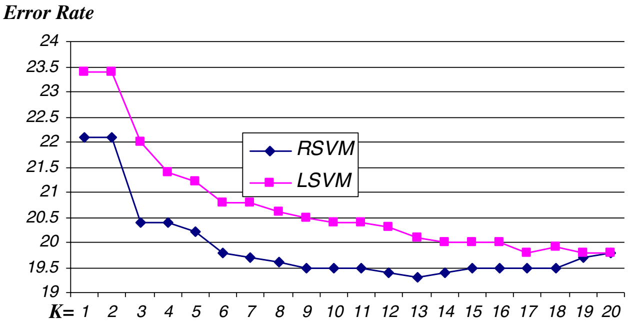


Fig. 3 Error rate in percentage obtained by varying the number K of physicochemical properties selected by SFFS

Table 2 Performance with the independent dataset

Physicochemical property	Ensemble (%)
Average relative fractional occurrence in AL(*i*-1)	11.5
Transfer energy, organic solvent/water	10
Average non-bonded energy per atom	13.1
Short- and medium-range non-bonded energy per atom	13.1
Average relative fractional occurrence in AL(*i*)	11.4
A parameter of charge transfer donor capability	11.4
Net charge	21.7
Weights for alpha-helix at the window position of –2	10.8
Long-range non-bonded energy per atom	13.1
Average non-bonded energy per residue	12.4
Number of full nonbonding orbitals	12.9
Partition coefficient	13.1
The number of atoms in the side chain labelled 1 + 1	11.1
VOTE	10.2
NO	12

$H1 = 0$; $H2_{i,j}(n,n+m) = 1$ if the amino acid in location n is i and the one in location $n+m$ is j, otherwise $H2 = 0$).

From Table 2, it is clear that the ensemble reduces the error rate by 20% with respect to the error rate obtained by the standard stand-alone method (i.e., the method named NO in that table). Notice that the best property obtains an error rate of 10% (the ensemble obtains an error rate of 10.2%), unfortunately it is not possible to understand the good behavior of that property using only the training data (the best property of the training data obtains an error rate of 11.5% on the independent dataset).

The main drawback of the proposed system is the high dimensionality of the feature vector (2,400-dimensional). To reduce the dimensions of the feature vector, we have modified the feature extraction method in the following way:

$$\mathbf{P1}^{m}_{i,j} = (1(L-m)) \times \left[\sum\nolimits_{n=1:L-m} \left(\mathrm{H1}_{i,j}(n,n+m,d) + \mathrm{H2}_{i,j}(n,n+m,d)\right)\right] \quad (3)$$

Each protein is represented only using **P1**. In this way we have a feature vector of 1,200 elements. We named this feature extraction method the short quasi-residue couple. We also ran the SFFS selection using the short quasi-residue couple method.

In Table 3, the performance obtained in each class is reported. In this table we compare the performance of the proposed systems with the state-of-art [i.e., the MemType-2L system proposed in Chou and Shen (2007c)]. ENS1 is the ensemble based on the quasi-residue couple method, and ENS2 is the ensemble based on the short quasi-residue couple method. Moreover, for ENS2 we report ENS2-BORDA[6] as well as the performance obtained by

[6] The properties selected for ENS2-BORDA are normalized hydrophobicity scales for alpha/beta-proteins, a parameter of charge transfer capability, short- and medium-range non-bonded energy per residue, hydrophobicity factor, net charge, free energy of solution in water (kcal/mole), long-range non-bonded energy per atom, partition coefficient, hydropathy index, transfer free energy, average non-bonded energy per atom, weights for beta-sheet at the window position of −1, positive charge, transfer energy, organic solvent/water, weights for alpha-helix at the window position of −2, direction of hydrophobic moment.

Table 3 Error rate obtained in each class

Type	ENS1 (%)	ENS2 (%)	ENS2-BORDA (%)	MemType-2L (%)
Single-pass type I	8.3	9.2	10.3	13.1
Single-pass type II	37.2	29.4	25.6	29.5
Single-pass type III	100	83.3	83.3	66.7
Single-pass type IV	66.7	58.3	66.6	32.3
Multipass	7.2	7.3	6.2	5.0
Lipid-chain anchor	68.4	65.8	60.5	57.9
GPI anchor	37	39.1	32.6	23.9
Peripheral	17.1	17.1	16.2	17.8
Overall	10.2	10	9	8.4

combining the classifiers using the Borda Count[7] (Nanni and Lumini 2006c).

The MemType-2L did outperform our system, but we want to stress that MemType-2L extracts the features considering the evolution information of the proteins. The aim of our work is to show how an ensemble of classifiers can improve the performance of a method based on the features extracted directly from the amino-acid sequence. Notice that Pu et al. (2007) show that the features extracted from the amino-acid sequence and from the evolution information of the proteins are partially complementary and that this property can be used for building a multi-classifier.

Moreover, several improvements of the system (e.g., a grid search for the parameters of the support vector machine, different ways to combine the physicochemical properties with the residue couple model, different methods to combine the classifiers of the ensemble) could improve the performance.

As further comparison we report the error rate on the independent set obtained by two well-known systems: from Chou and Shen (2007c), least Euclidean distance 38.6%; from Cedano et al. (1997), ProtLoc 62.8%. From these results, it is clear that the proposed method obtains a very low error rate with a very difficult dataset.

Conclusions

In this paper, we propose a new algorithm that uses the residue couple model in conjunction with the physicochemical properties to obtain a novel method for predicting the membrane protein type directly from the amino acid sequence. A reduced set of classifiers—radial basis function support vector machines—are selected by running the sequential forward floating selection, where the objective function is the minimization of the error rate in the training set.

The validity of the novel approach is proved by comparison with other state-of-the-art methods using the tested problem.

Appendix: Matlab code of the quasi residue couple

The following function implements the base feature extraction method as detailed in Materials and methods.

```
function X=QRcouple(seq,m,P)
alfabeto=['A' 'R' 'N' 'D' 'C' 'Q' 'E' 'G' 'H' 'I' 'L' 'K' 'M' 'F' 'P' 'S' 'T' 'W' 'Y' 'V'];
%seq is a given protein
%P is the property
N=size(seq,2);
t=1;
for ii=1:m
  for i=1:20
    for j=1:20
      X(t)=0;
      X(t+1)=0;
      for n=1:N-ii
        if seq(n)==alfabeto(i) & seq(n+ii)==alfabeto(j)
          X(t)=X(t)+P(i);
          X(t+1)=X(t+1)+P(j);
        end
      end
      if X(t)>0
        X(t:t+1)=X(t:t+1).*(1/(N-ii));
      end
      t=t+2;
    end
  end
end
```

The following function implements the short quasi residue couple method:

```
function X=QRcouple(seq,m,P)
alfabeto=['A' 'R' 'N' 'D' 'C' 'Q' 'E' 'G' 'H' 'I' 'L' 'K' 'M' 'F' 'P' 'S' 'T' 'W' 'Y' 'V'];
N=size(seq,2);
t=1;
for ii=1:m
  for i=1:20
    for j=1:20
      X(t)=0;
      for n=1:N-ii
        if seq(n)==alfabeto(i) & seq(n+ii)==alfabeto(j)
          X(t)=X(t)+P(i)+P(j);
        end
      end
      if X(t)>0
        X(t)=X(t).*(1/(N-ii));
      end
      t=t+1;
    end
  end
end
```

[7] Borda Count is defined as a mapping from a set of individual rankings to a combined ranking leading to the most relevant decision. Each class gets one point for each last place vote received, two points for each next-to-last point vote, etc., all the way up to *M* points for each first place vote (where *M* is the number of candidates/alternatives).

the following code is the main of our program:

```
RAN(4333,8)=0;
for ii=1:16
  QRCtraining(3249,1200)=0;
  QRCtesting(4333,1200)=0;

   for j=1:size(TR,2)
     QRCtraining(j,:)=single(QRcouple2(char(TR{j}),3,MM(F(ii),:)));
   end
   for j=1:size(TE,2)
     QRCtesting(j,:)=single(QRcouple2(char(TE{j}),3,MM(F(ii),:)));
   end
   %MM stores the properties, F are the selected properties

  massimo=max(QRCtraining)+0.00001;
  minimo=min(QRCtesting);
  training=[];
  testing=[];
  training=QRCtraining;
  testing=QRCtesting;
   for i=1:size(QRCtraining,2)
     training(1:size(QRCtraining,1),i)=double(QRCtraining(1:size(QRCtraining,1),
     i)-minimo(i))/(massimo(i)-minimo(i));

   end
   for i=1:size(QRCtesting,2)
     testing(1:size(QRCtesting,1),i)=double(QRCtesting(1:size(QRCtesting,1)
    ,i)-minimo(i))/(massimo(i)-minimo(i));
   end
   tra=[];
  QRCtraining=training;
  QRCtesting=testing;

  CO=[];
   for i=1:max(y)
     A=QRCtraining;A(find(y==i),:)=[];
     B=QRCtraining(find(y==i),:);
     yA=y;yA(find(y==i))=[];yA=(yA.*0)+1;
     yB=y(find(y==i));yB=(yB.*0)+2;
     [AlphaY, SVs, Bias, Parameters, nSV, nLabel] = rbfSVC(double([A; B]'),
     double([yA yB]),0.1,100);
     [ER, Decision, Ns, ConfMatrix, S]= SVMTest(double(QRCtesting'), double(yy), AlphaY,
    SVs, Bias, Parameters, nSV,nLabel);
     CO(:,i)=Decision;
   end
  [a,b]=sort(CO*-1,2);
   for j=1:8
      for jj=1:size(CO,1)
       RAN(jj,j)=RAN(jj,j)+find(b(jj,:)==j);
      end
   end
end
%the scores are stored in the matrix RAN
```

References

Cai YD, Liu XJ, Xu XB, Chou KC (2000) Support vector machines for prediction of protein subcellular location. Mol Cell Biol Res Commun 4:230–233

Cai YD, Liu XJ, Xu XB, Chou KC (2002) Support vector machines for prediction of protein subcellular location by incorporating quasi-sequence-order effect. J Cell Biochem 84:343–348

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

Cedano J, Aloy P, P'erez-Pons JA, Querol E (1997) Relation between amino acid composition and cellular location of proteins. J Mol Biol 266:594–600

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 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 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 (2000) Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. Biochem Biophys Res Commun 278:477–483

Chou KC (2001) Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet 43:246–255 (Erratum: 44:60)

Chou KC (2004) Review: structural bioinformatics and its impact to biomedical science. Curr Med Chem 11:2105–2134

Chou KC (2005) Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21:10–19

Chou KC, Cai YD (2002) Using functional domain composition and support vector machines for prediction of protein subcellular location. J Biol Chem 277:45765–45769

Chou KC, Cai YD (2003) A new hybrid approach to predict subcellular localization of proteins by incorporating gene ontology. Biochem Biophys Res Commun 311:743–747

Chou KC, Cai YD (2004a) Predicting subcellular localization of proteins by hybridizing functional domain composition and pseudo-amino acid composition. J Cell Biochem 91:1197–1203

Chou KC, Cai YD (2004b) Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem Biophys Res Commun 320:1236–1239

Chou KC, Cai YD (2005) Predicting protein localization in budding yeast. Bioinformatics 21:944–950

Chou KC, Elrod DW (1998) Using discriminant function for prediction of subcellular location of prokaryotic proteins. Biochem Biophys Res Commun 252:63–68

Chou KC, Elrod DW (1999a) Protein subcellular location prediction. Protein Eng 12:107–118

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

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

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 Comm 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 (2008) Cell-PLoc: a package of web-servers for predicting subcellular localization of proteins in various organisms. Nat Protoc 3:153–162

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

Cristianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge

Diao Y, Li M, Feng Z, Yin J, Pan Y (2007a) The community structure of human cellular signaling network. J Theor Biol 247:608–615

Diao Y, Ma D, Wen Z, Yin J, Xiang J, Li M (2007b) 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 Pept Lett 14:811–815

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

Doyle DA, Morais CJ, Pfuetzner RA, Kuo A, Gulbis JM, Cohen SL, Chait BT, MacKinnon R (1998) The structure of the potassium channel: molecular basis of K+ conduction and selectivity. Science 280:69–77

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

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

Kawashima S, Kanehisa M (2000) AAindex: amino acid index database. Nucleic Acids Res 28:374

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

Lee K, Kim DW, Na D, Lee KH, Lee D (2006) PLPD: reliable protein localization prediction from imbalanced and overlapped datasets. Nucleic Acids Res 34:4655–4666

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

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

Lodish H, Baltimore D, Berk A, Zipursky SL, Matsudaira P, Darnell J (1995) Molecular cell biology, 3rd edn. Scientific American, New York

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

Mundra P, Kumar M, Kumar KK, Jayaraman VK, Kulkarni BD (2007) Using pseudo amino acid composition to predict protein subnuclear localization: approached with PSSM. Pattern Recognition Lett 28:1610–1615

Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24:34–36

Nakai K, Kanehisa M (1992) A knowledge base for predicting protein localization sites in eukaryotic cells. Genomics 14:897–911

Nanni L (2006) Comparison among feature extraction methods for HIV-1 protease cleavage site prediction. Pattern Recognition 39:711–713

Nanni L, Lumini A (2006a) An ensemble of K-local hyperplane for predicting protein-protein interactions. BioInformatics 22(10):1207–1210

Nanni L, Lumini A (2006b) MppS: an ensemble of support vector machine based on multiple physicochemical properties of amino-acids. NeuroComputing 69:1688–1690

Nanni L, Lumini A (2006c) Detector of image orientation based on Borda-count. Pattern Recognition Lett 27:180–186

Nanni L, Lumini A (2008a) Combing ontologies and dipeptide composition for predicting DNA-binding proteins. Amino Acids. doi:10.1007/s00726-007-0018-1

Nanni L, Lumini A (2008b) Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. Amino Acids. doi:10.1007/s00726-007-0016-3

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

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

Pugalenthi G, Tang K, Suganthan PN, Archunan G, Sowdhamini R (2007) A machine learning approach for the identification of odorant binding proteins from sequence-derived properties. BMC Bioinformatics 8:351

Schnell JR, Chou JJ (2008) Structure and mechanism of the M2 proton channel of influenza A virus. Nature 451:591–595

Shen HB, Chou KC (2007a) 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 (2007b) 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 (2007a) Prediction of protein subcellular localization by support vector machines using multi-scale energy and pseudo amino acid composition. Amino Acids 33:69–74

Shi JY, Zhang SW, Pan Q, Zhou GP (2007b) Using pseudo amino acid composition to predict protein subcellular location: approached with amino acid composition distribution. Amino Acids. doi:10.1007/s00726-007-0623-z

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 33:669–675

Wang M, Yang J, Liu GP, Xu ZJ, Chou KC (2004) Weighted-support vector machines for predicting membrane protein types based on pseudo amino acid composition. Protein Eng Des Sel 17:509–516

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

Wen Z, Li M, Li Y, Guo Y, Wang K (2007) 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

Xiao X, Chou KC (2007) Digital coding of amino acids based on hydrophobic index. Protein Pept 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

Yuan Z (1999) Prediction of protein subcellular locations using Markov chain models. FEBS Letters 451:23–26

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33:623–629

Zhang SW, Pan Q, Zhang HC, Shao ZC, Shi JY (2006a) 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 ZH, Wang ZH, Zhang ZR, Wang YX (2006b) A novel method for apoptosis protein subcellular localization prediction combining encoding based on grouped weight and support vector machine. FEBS Lett 580:6169–6174

Zhang SW, Zhang YL, Yang HF, Zhao CH, Pan Q (2007) Using the concept of Chou's pseudo amino acid composition to predict protein subcellular localization: an approach by incorporating evolutionary information and von Neumann entropies. Amino Acids. doi:10.1007/s00726-007-0010-9

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17:729–738

Zhou GP, Assa-Munt N (2001) Some insights into protein structural class prediction. Proteins Struct Funct Genet 44:57–59

Zhou GP, Doctor K (2003) Subcellular location prediction of apoptosis proteins. Proteins Struct Funct Genet 50:44–48

Zhou XB, Chen C, Li ZC, Zou XY (2007a) Improved prediction of subcellular location for apoptosis proteins by the dual-layer support vector machine. Amino Acids. doi:10.1007/s00726-007-0608-y

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