Amino Acids (2008) 34:653–660
DOI 10.1007/s00726-007-0018-1

ORIGINAL ARTICLE

Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization

Loris Nanni · Alessandra Lumini

Received: 26 November 2007 / Accepted: 11 December 2007 / Published online: 4 January 2008

Abstract Given a protein that is localized in the mitochondria it is very important to know the submitochondria localization of that protein to understand its function. In this work, we propose a submitochondria localizer whose feature extraction method is based on the Chou's pseudo-amino acid composition. The pseudo-amino acid based features are obtained by combining pseudo-amino acid compositions with hundreds of amino-acid indices and amino-acid substitution matrices, then from this huge set of features a small set of 15 "artificial" features is created. The feature creation is performed by genetic programming combining one or more "original" features by means of some mathematical operators. Finally, the set of combined features are used to train a radial basis function support vector machine. This method is named GP-Loc. Moreover, we also propose a very few parameterized method, named ALL-Loc, where all the "original" features are used to train a linear support vector machine. The overall prediction accuracy obtained by GP-Loc is 89% when the jackknife cross-validation is used, this result outperforms the performance obtained in the literature (85.2%) using the same dataset. While the overall prediction accuracy obtained by ALL-Loc is 83.9%.

Keywords Submitochondria localization · Chou's pseudo amino acid · Genetic programming

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

Introduction

Protein localization is a very deep studied problem in Bioinformatics, since it is important to know the exact sub-cellular localization of a protein for understanding its function (Chou and Shen 2007d). Unfortunately, it is costly and time consuming to experimentally identify the protein sub-cellular location, for this reason several systems have been developed with the aim of automatically solving this problem. These methods are mainly based on the analysis of the amino-acid sequence (Cai et al. 2000, 2002; Cedano et al. 1997; Chou 2001; Chou and Cai 2002, 2003, 2004a, b; Chou and Elrod 1998, 1999; Chou and Shen 2006; Nakai and Horton 1999; Nakai and Kanehisa 1992; Yuan 1999) as well as the relevant references cited in a recent review article (Chou and Shen 2007d).

Moreover, some proteins have multiple locations, and these proteins have some very special biological functions. The literature about this problem is quite recent: the first few automatic systems (Chou and Cai 2005; Lee et al. 2006) developed for multiple sub-cellular localization of a protein are based on the budding yeast proteins only; more recently, two web servers, i.e. Euk-mPLoc at http://chou.med.harvard.edu/bioinf/euk-multi/ (Chou and Shen 2007a) and Hum-mPLoc at http://chou.med.harvard.edu/bioinf/hum-multi/ (Shen and Chou 2007b), have been established for predicting the multiple sub-cellular localization of eukaryotic proteins and human proteins, respectively.

Another problem which has received little attention in the literature is the submitochondria location. Mitochondria are membrane enclosed organelles found in most eukaryotic cells. Mitochondria are surrounded by two layers of membrane, the inner membrane and the outer membrane, the proteins belonging to the inner membrane, outer membrane and matrix contributed in a different way to different procedures in energy metabolism. The use of a reliable automatic submitochondria localizer could speed up the drugs design for over 100 kinds of complex diseases related with mitochondria like programmed cell death (Gottlieb 2000) and ionic homeostasis (Jassem et al. 2000). The importance of mitochondrial proteins is reflected by the fact that children keep dying from mysterious illness that have been traced to tiny structures called mitochondria.

To the best of our knowledge, however, only one computational system (Du and Li 2006) for predicting protein submitochondria location is proposed in the literature. SUBMITO (Du and Li 2006) is an automatic system to predict the submitochondria location (mitochondria inner membrane, mitochondria outer membrane and mitochondria matrix) for a protein, where the features are extracted from the amino-acid sequence. The feature vector is composed by: occurrence frequencies of different residues; dipeptide composition; Chou's pseudo-amino acid composition (Chou 2005) where nine physicochemical properties are used.

Moreover, the protein is segmented into fixed length segments and the features are extracted from each segment.

Although significant progresses have been made during the last 15 years in predicting protein sub-cellular localization as summarized in a recent review article (Chou and Shen 2007d), in contrast to that much fewer methods (particularly with web-server) have been reported for predicting protein submitochondrial localization. The present study was initiated in an attempt to enrich the latter by proposing a submitochondria localizer whose feature extraction method is based on the Chou's pseudo amino acid (PseAA) composition.

To successfully use the Chou's PseAA composition for improving the prediction quality of various protein attributes, the key is how to optimally extract the features for the PseAA components. Many different approaches have been proposed, such as hydrophobicity (Wang et al. 2004), hydrophilicity (Chou 2005), physicochemical distance (Chou 2000), digital code (Gao et al. 2005), complexity factor (Xiao et al. 2005, 2006b), digital signal (Xiao and Chou 2007), Fourier low-frequency spectrum (Liu et al. 2005), cellular automata (Xiao et al. 2006a), as well as a variety of many others (see, e.g., Chen et al. 2006a, b; Chen and Li 2007b; Diao et al. 2007a; Du and Li 2006; Fang et al. 2007; Kurgan et al. 2007; Li and Li 2007; Lin and Li 2007a, b; Mondal et al. 2006; Mundra et al. 2007; Pu et al. 2007; Shi et al. 2007; Zhang et al. 2006; Zhang and Ding 2007; Zhou et al. 2007).

In this paper, we deal with the submitochondria localization problem by generating a set of "artificial" features (Lumini and Nanni 2007) (to be used instead of the "original" ones). The "original" feature set is composed by the Chou's pseudo-amino acid composition features and the set of the physicochemical properties obtained by the amino acid index database (Kawashima and Kanehisa 2000).

From the set of 11,540 "original" features, a small set of 15 "artificial" features is created by genetic programming (GP) (Yu and Bhanu 2006; Paul and Iba 2007), randomly applying mathematical operations to randomly extracted "original" features. This set of "artificial" features is used to train a radial basis function support vector machine (RSVM).

The proposed method (GP-Loc) obtains an interesting result of 89% accuracy on the same dataset used in (Du and Li 2006), when the jackknife cross-validation testing protocol is used. Moreover, our experiments demonstrate that a method (ALL-Loc) where all the 11,540 "original" features are used to train a linear support vector machine (LSVM) obtains performance (83.9% accuracy) similar to (Du and Li 2006) (85.2% accuracy) without any kind of parameter optimization (the parameter *C* of LSVM is set to the default value 1).

Notice that both GP-Loc and SUBMITO are based on a set of selected physicochemical properties, where the selection is necessarily dataset dependent; therefore some problems could happen if the dataset is not enough representative for the submithocondria localization problem. For example, in the 2007 it has been discovered that the most used (in the last 10 years) dataset for HIV-protease (Rögnvaldsson and You 2003; Kontijevskis et al. 2007) was not representative and that several biological and bioinformatics discoveries were not completely true (i.e. the best physicochemical properties) (Kontijevskis et al. 2007). On the contrary, the main advantage of ALL-Loc is that it is a very few parameterized method: no feature selection is performed and the parameter *C* of LSVM is set to the default value 1.

Methods

In this paper, we propose two systems for dealing with the submitochondria localization problem (see Fig. 1): the first and simpler one, ALL-LOC, is based on a set of 11,540 "original" features directly used to train a LSVM classifier, the second, GP-Loc, uses GP to generate from the "original" features a set of "artificial" features (Lumini and Nanni 2007) to be used for training a RSVM classifier. The "original" feature set is composed by the Chou's pseudo-amino acid composition features and the set of the physicochemical properties obtained by the Amino Acid index database (Kawashima and Kanehisa 2000).

Then GP is used to generate "artificial" features named evolved Chou's PseAA features. Our system is built by running different executions of GP to synthesize *G* evolved Chou's PseAA features; finally, these *G* ($G = 15$ in this

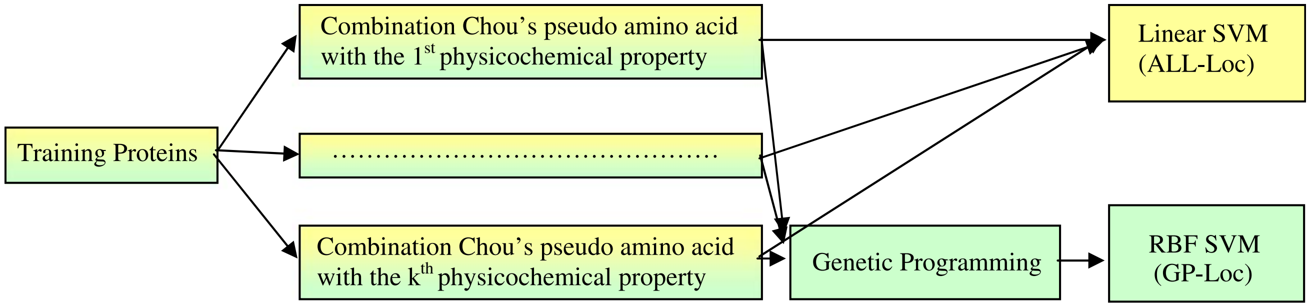
Combination Chou's pseudo amino acid with the 1st physicochemical property
Linear SVM (ALL-Loc)
Training Proteins
....
Combination Chou's pseudo amino acid with the kth physicochemical property
Genetic Programming
RBF SVM (GP-Loc)

Fig. 1 Global schema of our two systems: ALL-loc and GP-loc

work) features are used instead of the original ones to train the classifier.

Support vector machine is a machine learning algorithm based on statistical learning theory which was introduced by Vapnik (Cristianini and Shawe-Taylor 2000). It searches for an optimal separating hyper plane which maximizes the margin in feature space.

Extraction of the "original" features

The set of "original" features used to describe the patterns is composed by 11,540 features obtained combining the pseudo-amino acid composition with the set of the physicochemical properties obtained by the amino acid index database (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 **PC** of amino acids (Nanni and Lumini 2006a, b). This database currently contains 494 such indices and 83 substitution matrix.

In several works, that are summarized in a recent review article (Chou and Shen 2007d), Chou proposes to extract from a given protein **A** a set of PseAA based features: (20 + λ) features (where λ is a parameter denoting the maximum distance between two considered amino acids) $\mathbf{P} = (P_1,\ldots,P_{20}, P_{20+1},\ldots,P_{20+\lambda})$. The first 20 features are the amino acid composition, the features from P_{20+1} to $P_{20+\lambda}$ reflect the effect of sequence order.

In this paper, only the features from P_{20+1} to $P_{20+\lambda}$ are considered (λ is set to 20 in this work) and evaluated for all the available physicochemical properties $p \in \mathbf{PC}$, hence we have a set of $\mathbf{P}(p) = (P^p_{20+1},\ldots, P^p_{20+\lambda})$. The extraction of a generic feature P^p_{20+i}, for a physicochemical property p, from a given protein **A** is obtained as:

$$P^p_{20+i} = \left[\sum\nolimits_{j=1:L-i} \mathrm{val}(p, \mathbf{A}(j)) \times \mathrm{val}(p, \mathbf{A}(j+i))\right]/(L-i)$$
$$\mathrm{val}(p,d) = \left(\mathrm{index}(p,d) - N_p\right)/D_p$$

where $\mathbf{A}(j)$ is the jth amino acid of **A**, L is the length of the protein sequence in **A**, index(p,d) is the function returning the value of the physicochemical property p of the amino-acid d, $N_p = \mathrm{avg}_{d=1:20}(\mathrm{index}(p,d))$, is a normalization factor obtained as the average value of a given physicochemical property p for the whole set of amino-acids; $D_p = ((\Sigma_d(\mathrm{index}(p,d) - N_p)^2)/20)^{0.5}$ is a normalization factor.

Since λ is set to 20 and we use 494 indices and 83 substitution matrix we obtain a the final vector set of 11,540 "original" features ([494 + 83] × 20).

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

Genetic programming[1] for evolved Chou's pseudo amino acid features

GP may be considered an extension of the genetic algorithms. In GP the individuals can be complicated structures such as trees and graphs. An individual's program is a tree-like structure and as such there are two types of genes: functions and terminals[2]. Terminals, in tree terminology, are leaves while functions are nodes with children.

In this work the primitive features are the Chou's PseAA based features, the representation structures are binary trees and the primitive operators are the pool of unary and binary operators detailed in Table 1.

The selection for the reproduction is obtained using the well-know method named "roulette". Moreover, the best individual from both parents and children is kept for the new population (after the reproduction), independently of being a parent or a child. The remaining places in the new population are occupied by children only.

[1] This algorithm has been implemented in the Matlab toolbox named "GPLAB: A Genetic Programming Toolbox for MATLAB", http://gplab.sourceforge.net/.

[2] http://www.massey.ac.nz/~mgwalker/gp/index.html.

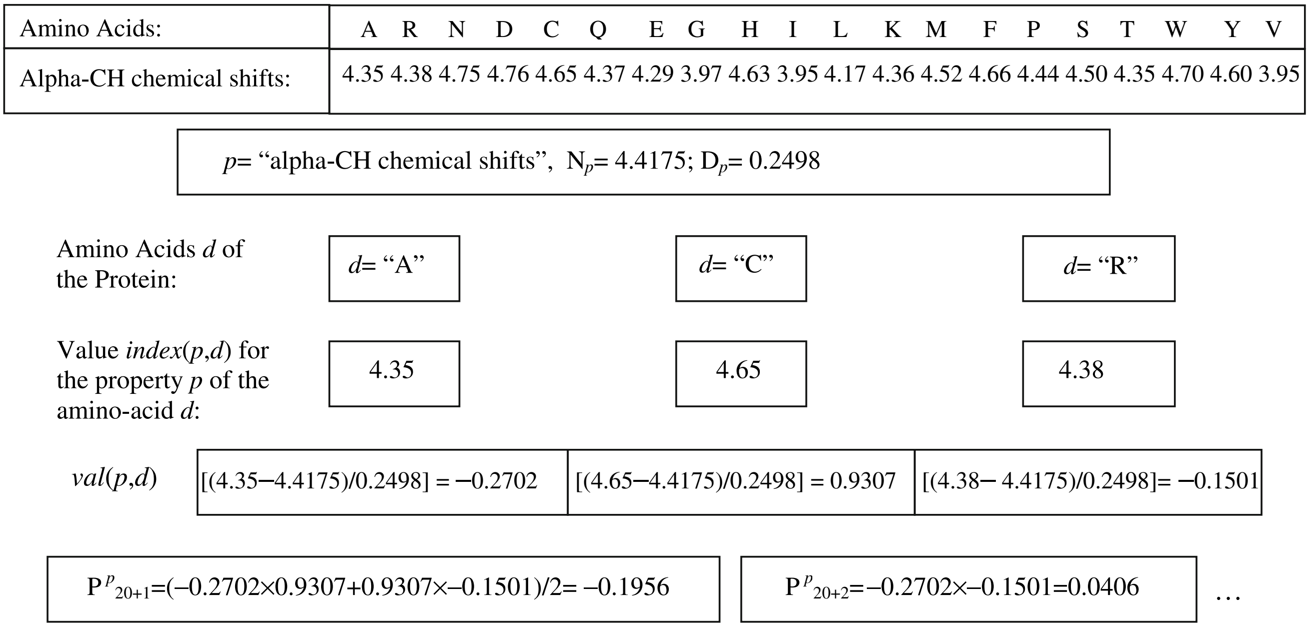
Amino Acids:
A R N D C Q E G H I L K M F P S T W Y V
Alpha-CH chemical shifts:
4.35 4.38 4.75 4.76 4.65 4.37 4.29 3.97 4.63 3.95 4.17 4.36 4.52 4.66 4.44 4.50 4.35 4.70 4.60 3.95
p= "alpha-CH chemical shifts", Np= 4.4175; Dp= 0.2498
Amino Acids d of the Protein:
d= "A"
d= "C"
d= "R"
Value index(p,d) for the property p of the amino-acid d:
4.35
4.65
4.38
val(p,d)
[(4.35–4.4175)/0.2498] = –0.2702
[(4.65–4.4175)/0.2498] = 0.9307
[(4.38– 4.4175)/0.2498]= –0.1501
P p 20+1=(–0.2702×0.9307+0.9307×–0.1501)/2= –0.1956
P p 20+2=–0.2702×–0.1501=0.0406
…

Fig. 2 Example of feature extraction

Table 1 Primitive operators

Unary operators	Binary operators
SQ: square	SUM: sum
SQRT: square root	SUB: subtraction
SIN: sin	PROD: product
COS: cosine	DIV: division
ASIN: arc sin	
REC: reciprocal	
ACOS: arc cosine	
LOG: logarithm	
TAN: tangent	
ABS: absolute value	
TANH: hyperbolic tangent	
NEG: negative value	
NO: nothing	

Our system is built running different executions of GP to synthesize G evolved Chou's PseAA features. The results of all the previous executions of GP are used to drive the actual one. The fitness function of the ith GP is given by the Accuracy Rate obtained using a fivefold cross-validation, where the support vector machine is trained considering also the evolved Chou's PseAA features created by the previous $1,\ldots,i-1$ GP executions.

In the example reported in Fig. 3 there are two binary operators (× and +) and three terminals (P^{p}_{20+1}, P^{p}_{20+2} and P^{p}_{20+3}) obtained for p = "alpha-CH chemical shifts": the "artificial" feature created by GP is given by: $P^{p}_{20+1} + (P^{p}_{20+2} \times P^{p}_{20+3})$.

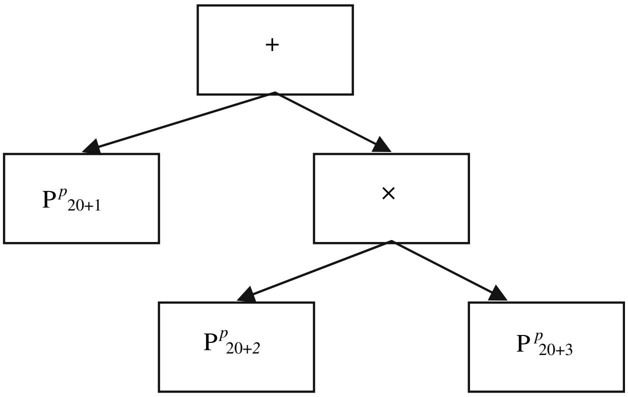

Fig. 3 Example of a GP individual

Experimental results

In this work we have used the dataset described in (Du and Li 2006), which contains 317 proteins classified into three submitochondria locations: 131 inner membrane proteins; 41 outer membrane proteins; 145 matrix proteins. The identity cut off is set to 40% (i.e. the identity between any two sequences in the processed dataset is less than 40%) in order to get a balance between the homologous bias and the size of the training set. The authors (Du and Li 2006) claim that if a cut off value at level 25% was used, it should not be possible to obtain enough sequences to build sufficient large training set.

The prediction accuracy in percentage and Matthew's correlation coefficient (MCC) for each location are used as parameters to evaluate the proposed system:

$$\text{Total accuracy} = \frac{100}{\#D}\sum_{i=1}^{c}\text{TP}(i)$$

$$\text{Accuracy}(i) = \frac{\text{TP}(i)}{\text{TP}(i)+\text{FN}(i)} \times 100$$

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

where: $\#D$ is the total number of sequences, c is the number of classes ($c = 3$ in this problem, “inner membrane proteins”, “outer membrane proteins”, “matrix proteins”); TP(i) is the number of correctly predicted sequences (true positives) of location i ($i \in \{1,\ldots,c\}$); TN(i), FP(i) and FN(i) are the numbers of true negatives, false positives and false negatives of the ith location, respectively.

Notice that the MCC considers not only the number of true positives but also the number of false positives, false negatives and true negatives, for this reason it is more reliable than the overall accuracy, especially when the studied dataset is unbalanced.

Among the independent dataset test, subsampling (e.g., five or tenfold cross-validation) test, and jackknife test, which are often used for examining the quality of a statistical prediction method, the jackknife test was deemed the most rigorous and objective (Chou and Zhang 1995). Besides, in contrast to the subsampling test which may lead to many different results even for a same benchmark dataset, the jackknife test can always yield a unique result as demonstrated by an incisive analysis in a recent comprehensive review (Chou and Shen 2007d) and Eq. 50 thereof. Accordingly, the jackknife test has been increasingly and widely adopted by investigators to test the power of various prediction methods (see, e.g., Chen et al. 2007; Chen and Li 2007a; Chou and Shen 2007b, c, e; Diao et al. 2007b; Ding et al. 2007; Fang et al. 2007; Gao et al. 2005; Guo et al. 2006; Li and Li 2007; Lin and Li 2007a; Lin and Li 2007b; Liu et al. 2007; Mondal et al. 2006; Mundra et al. 2007; Niu et al. 2006; Shen and Chou 2007a, c, d; Shen et al. 2007; Shi et al. 2007; Sun and Huang 2006; Tan et al. 2007; Wang et al. 2005; Wen et al. 2006; Zhang et al. 2006; Zhang and Ding 2007; Zhou et al. 2007). In view of this, here we use the objective and rigorous jackknife

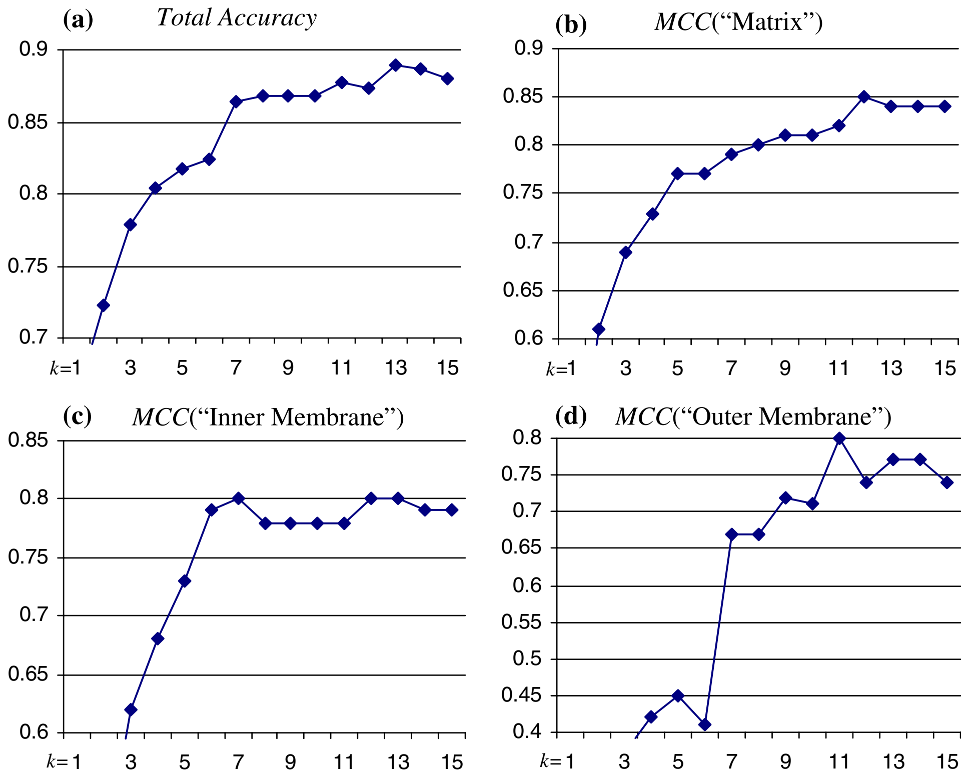

Fig. 4 Performance (total accuracy or MCC) of GP-Loc in function of the number k of features used to train SVM: **a** Total accuracy; **b** MCC(“Matrix”); **c** MCC(“Inner Membrane”); **d** MCC(“outer Membrane”)

Table 2 Comparison among several methods

Indicator		GP-LOC	SUBMITO	ALL-LOC	ALL(B)-LOC	2-GRAM
Total accuracy		89%	85.2%	83.9%	83.9%	71.3%
Accuracy	Inner membrane	83.21%	85.5%	78.6%	79.4%	71.8%
	Outer membrane	78.05%	51.2%	58.5%	56.1%	31.7%
	Matrix	97.24%	94.5%	95.9%	95.9%	82.1%
MCC	Inner membrane	0.8	0.79	0.75	0.75	0.56
	Outer membrane	0.77	0.64	0.62	0.60	0.30
	Matrix	0.85	0.77	0.75	0.75	0.497

Table 3 The five best evolved features and the original features used to their creation

Evolved features	Original features	Properties
TANH(TAN(P^{p1}_{20+4}))	P^{p1}_{20+4}	$p1$ = Proportion of residues 95% buried
P^{p2}_{20+2}	P^{p2}_{20+2}	$p2$ = Hydrophobicity
P^{p3}_{20+5}	P^{p3}_{20+5}	$p3$ = Hydropathy index
PROD(P^{p4}_{20+16}, LOG(P^{p5}_{20+16}))	P^{p4}_{20+16}	$p4$ = Relative preference value at N4
	P^{p5}_{20+16}	$p5$ = Normalized positional residue frequency at helix termini N1
P^{p6}_{20+2}	P^{p6}_{20+2}	$p6$ = Normalized frequency of zeta L

cross-validation method as testing protocol, while the feature generation by GP is performed using a fivefold cross-validation in the following way: remove one observation from the data set; perform the feature extraction using fivefold cross-validation on the remaining data; train a classifier that uses the found features on the remaining data; test this classifier on the left out observation; Repeat these steps over the entire data set.

In Fig. 4 the performance of GP-Loc in function of the number of features used to train the support vector machine are plotted.

In Table 2 the following approaches are compared in terms of accuracy and MCC: GP-LOC is our best method where the "artificial" features are used to train a RSVM; SUBMITO is the system proposed in (Du and Li 2006); ALL-LOC is our simpler method, where all the "original" features are used to train a LSVM; ALL(B)-LOC is a variant of the method above where the "original" features are concatenated with the occurrence frequencies of different residues and the dipeptide composition before to be used to train a LSVM; 2-GRAM, only the dipeptide composition features are used to train a LSVM.

In Table 3 the five best evolved features obtained by running GP[3] are listed. Notice that the second and the third features correspond to a simple selection of an "original" feature, no combination is performed.

From the analysis of the experimental results, the following observations may be made: the standard dipeptide composition does not permit to obtain a reliable method; Using GP to select a small set of features among the huge set of "original" features, that can be obtained combining pseudo-amino acid composition with the set of the physicochemical properties obtained by the amino acid index database, permits to outperform SUBMITO (Du and Li 2006), the best method proposed in the literature; both GP-Loc and SUBMITO are based on the assumption that the collected database is a reliable database for the studied problem (both the methods need a parameter optimization), we also propose a non-parameterized method (ALL-Loc) that obtains good performance.

Conclusions

In this paper, we propose a new algorithm which uses GP in conjunction with Chou's PseAA to obtain a novel sub-mitochondria localizer.

All the amino-acid indices and amino-acid substitution matrices reported in amino acid index database are used to obtain the pseudo-amino acid based features, then the generation of the new "combined" features is carried out by GP, combining one or more "original" features, by means of some mathematical operators. Finally, a small set of combined features are used to train a prediction algorithm (the RSVM).

The validity of the novel approach is proved by the performance improvements obtained with respect to other state-of-the-art methods in the tested problem.

[3] The GP is performed using a fivefold cross-validation for each pattern, so for each pattern the best features could be quite different, we report the best five features on average.

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

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 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

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

Chen YL, Li QZ (2007b) Prediction of the subcellular location of apoptosis proteins. J Theor Biol 245:775–783

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 (Erratum: ibid., 2001, 44:60) 43:246–255

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 predicor. 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 (1999) Protein subcellular location prediction. Protein Eng 12:107–118

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 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 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 Peptide Lett 14:811–815

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 Y, Shao SH, Xiao X, Ding YS, Huang YS, Huang ZD, Chou KC (2005) 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 YZ, Li M, Lu M, Wen Z, Wang K, Li G, Wu J (2006) Classifying G protein-coupled receptors and nuclear receptors based on protein power spectrum from fast Fourier transform. Amino Acids 30:397–402

Gottlieb RA (2000) Programmed cell death. Drug News Perspect 13:471–476

Jassem W, Fuggle SV, Rela M, Koo DD, Heaton ND (2000) The role of mitochondria in ischemia/reperfusion injury. Transplantation 73:493–499

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

Kontijevskis A, Wikberg JES, Komorowski J (2007) Computational proteomics analysis of HIV-1 protease interactome. Proteins: Struct, Funct, Bioinformatics 68(1):305–312

Kurgan LA, Stach W, Ruan J (2007) Novel scales based on hydrophobicity indices for secondary protein structure. J Theor Biol 248:354–366

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 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 (2005) Low-frequency Fourier spectrum for predicting membrane protein types. Biochem Biophys Res Commun 336:737–739

Lumini A, Nanni L (2007) Over-complete feature generation and feature selection for biometry. Expert Syst Appl. doi: 10.1016/j.eswa.2007.08.097

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 243:252–60

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 Recognit 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, 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

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

Paul TK, Iba H (2007) Prediction of cancer class with majority voting genetic programming classifier using gene expression data. IEEE Trans Comp Biol Bioinformatics. http://doi.ieeecomputersociety.org/10.1109/TCBB.2007.70245

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

Rögnvaldsson T, You L (2003) Why neural networks should not be used for HIV-1 protease cleavage site prediction. Bioinformatics 20(11):1702–1709

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) Hum-mPLoc: an ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. Biochem Biophys Res Commun 355:1006–1111

Shen HB, Chou KC (2007c) Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Commun 363:297–303

Shen HB, Chou KC (2007d) 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

Yu, Bhanu B (2006) Evolutionary feature synthesis for facial expression recognition. Pattern Recognit Lett 27:1289–1298

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

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

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

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 (2006a) Using cellular automata images and pseudo amino acid composition to predict protein subcellular location. Amino Acids 30:49–54

Xiao X, Shao SH, Huang ZD, Chou KC (2006b) Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor. J Comput Chem 27:478–482

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

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. doi: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