Amino Acids (2009) 36:409–416
DOI 10.1007/s00726-008-0076-z

ORIGINAL ARTICLE

Using ensemble of classifiers for predicting HIV protease cleavage sites in proteins

Loris Nanni · Alessandra Lumini

Received: 19 February 2008 / Accepted: 27 March 2008 / Published online: 10 April 2008

Abstract The focus of this work is the use of ensembles of classifiers for predicting HIV protease cleavage sites in proteins. Due to the complex relationships in the biological data, several recent works show that often ensembles of learning algorithms outperform stand-alone methods. We show that the fusion of approaches based on different encoding models can be useful for improving the performance of this classification problem. In particular, in this work four different feature encodings for peptides are described and tested. An extensive evaluation on a large dataset according to a blind testing protocol is reported which demonstrates how different feature extraction methods and classifiers can be combined for obtaining a robust and reliable system. The comparison with other stand-alone approaches allows quantifying the performance improvement obtained by the ensembles proposed in this work.

Keywords Machine learning · Ensembles of classifiers · HIV-1 protease prediction

Introduction

During the last decade or so, the following two strategies have been often adopted to find drugs against AIDS (acquired immunodeficiency syndrome). One is to target the HIV (human immunodeficiency virus) reverse transcriptase (Althaus et al. 1993a, b, c, 1994a, b, 1996; Chou et al. 1994); the other is to design HIV protease inhibitors (Chou 1993b, c, 1996; Poorman et al. 1991).

HIV-1 protease (Rögnvaldsson and You 2003; Chou 1993a, c, d) is one of the enzyme in the AIDS virus that is essential to its replication. HIV-1 protease inhibitors are small molecules that bind to the active site in HIV-1 protease and stay there (Chou 1996), so that the normal functioning of the enzyme is prevented. Understanding and predicting HIV-1 protease cleavage sites in proteins, i.e., knowing which amino acid sequences are cleaved by the protease and which residues play important roles for the cleavage, is therefore a major concern in medicine, since the design of an efficient inhibitor requires good understanding of the HIV-1 protease cleavage site specificity. Unfortunately, no perfect rule is yet known that determines if and where a peptide will be cleaved by the HIV-1 protease and the experimental investigation of cleavability of the patterns in the laboratory is a very expensive task, since the number of candidate sequences is very high.

In order to reduce the number of pattern to be tested the use of artificial intelligence to aid and speed up the specificity investigation is thus essential: what is need is a classification model that, given a sequence of eight amino acids (an octamer), can tell whether it will be cleaved by the HIV-1 protease or not. In the past there have been several attempts to develop various prediction methods for the HIV protease cleavage sites in proteins based on techniques from machine learning. In (Cai and Chou 1998; Thompson et al. 1995) the authors trained a standard feed-forward multilayer perceptron (MLP) to solve this problem and this choice has also been validated in (Narayanan et al. 2002) by showing that a decision tree was not able to predict the cleavage as well as MLP. Recently Support Vector Machine (SVM) has been adopted (Cai et al. 2002) to predict the cleavage, without significantly improving the

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

classification performance obtained by the neural network presented previously. In (Rögnvaldsson and You 2003) the authors showed that HIV-1 protease cleavage is a linear problem and that the best classifier for this problem is the Linear SVM. In various machine learning methods have been extensively tested, and it is found that the combination of neural networks and decompositional approach can generate a set of effective rules. Recently, a web-server was established for predicting HIV protease cleavage sites in proteins (Shen and Chou 2008).

The last trend in machine learning is adopting a multi-classifier to better solve a complex classification problem; both theoretical and empirical (Opitz and Maclin 1999; Kittler 1998; Altıncay and Demirekler 2000; Breiman 1996, 2001) studies have demonstrated that a good ensemble of classifiers can improve the performance of a stand-alone classifier, in particular if the individual classifiers in the ensemble are both accurate and independent (i.e., they make errors on different regions of the feature space) (Whitaker and Kuncheva 2003; Zenobi and Cunningham 2001; Melville and Mooney 2003). In Nanni 2006 and in Nanni and Lumini 2006c several classifiers trained with different amino-acid encoding models are combined for HIV-1 Protease Cleavage Site: all the ensembles gained an error reduction with respect to the performance of the state-of-the-art stand-alone approaches. In (Nanni and Lumini 2008a) the classification task is performed by a multi-classifier system where each classifier is a trained using feature extracted by different reduced alphabets. Each alphabet is constructed by a Genetic Algorithm. Ensemble of classifiers (Chou and Shen 2007e, 2008; Shen and Chou 2007c; Shen et al. 2007) are also successfully used in improving the protein fold pattern prediction (Shen and Chou 2006), protein subcellular localization prediction (Shen and Chou 2007a, b; Chou and Shen 2006a, b, 2007a, b), membrane protein type prediction (Chou and Shen 2007c; Shen and Chou 2007e), and signal peptide prediction (Chou and Shen 2007d).

Unfortunately, several methodologies for building an ensemble of classifiers and several approaches based on physicochemical properties selection (e.g., (Nanni and Lumini 2006b)) need a training set for the parameter setting (e.g., the physicochemical properties selection); therefore if the dataset is not enough representative for the given problem the discoveries are not completely true.

In 2007, it has been reported (Kontijevskis et al. 2007) that the most used (in the last 10 years) dataset (Rögnvaldsson et al. 2007) for HIV-1 protease was not large enough to obtain reliable conclusions. For example the best physicochemical properties extracted by Nanni and Lumini (2006a) from the old dataset are not completely representative for the new dataset, and that some findings of the biologists on the most important peptides for the protease classification were true only in the old dataset (Kontijevskis et al. 2007).

In this paper, we propose four approaches based on different features extraction for HIV-1 protease prediction. The approaches are validated on the large HIV-protease dataset proposed in (Kontijevskis et al. 2007) according to a blind testing protocol. Our results are very encouraging; since the proposed ensembles drastically outperform the state-of-the-art stand alone method, which is based on the standard orthonormal encoding. Information obtained from computational approaches can timely provide very useful insights for drug development (see, e.g., (Chou 2004; Chou and Shen 2008; Chou et al. 2003; González-Díaz et al. 2008; Lubec et al. 2005; Shen and Chou 2007c, d)).

Methods

In this work, four approaches based on different features extraction have been proposed and tested for HIV-1 protease prediction. In each feature extraction method each letter of the amino acid alphabet AA = {A,R,K,V} is replaced by a given vector. Then a combination of the best methods is evaluated in an ensemble based both on the perturbation of the features (the classifiers are trained using a different feature set) and on the perturbation of the classifiers (different classifiers are combined). In the following subsections a description of the four new ensembles proposed in this work is given, the details of the final combination approach are given in the experimental section.

Physicochemical encoding representation

In the standard orthonormal representation (**O**) (Rögnvaldsson and You 2003) each amino acid is represented by a 20 bit vector with 19 bit set to zero and one bit set to one, and each amino acid vector is orthogonal to all other amino acid vectors. In this first method, named **PE**, the feature vectors used to describe the patterns are obtained by separately considering different physicochemical properties of the amino-acids (Nanni and Lumini 2006a) obtained by the Amino Acid index database[1] (Kawashima and Kanehisa 2000), the database contains 494 indices and 83 substitution matrix. The physicochemical encoding representation (Nanni and Lumini 2006a, c) is given by a 20-dimesional real vector $\mathbf{x} \in \Re^{20}$, with 19 values set to zero and the value related to the position of the amino acid takes the value of the considered physicochemical property **pe** (the physicochemical property index or the corresponding diagonal

[1] Available at http://www.genome.jp/dbget/aaindex.html.

entry of the substitution matrix). This encoding method is described by Eq. (1):

$$\mathrm{AA}_i \rightarrow (\delta_{i1} \times \mathbf{pe}(1), \ldots, \delta_{iN} \times \mathbf{pe}(N)) \quad (1)$$

δ_{ij} is the Kronecker delta symbol, $N = 20$, AA_i is the ith amino-acid letter.

Therefore each amino acid has a number P of physicochemical encoding representations equal to number of different physicochemical properties obtained by the Amino Acid index database. In this work $P = 577$, since 494 indices and 83 substitution matrices have been encoded. In order to reduce the number of features a Sequential Forward Floating Selection (SFFS) (Pudil et al. 1994) feature selection approach has been adopted (as in (Nanni and Lumini 2006a)), where the objective function is the maximization of the area under the ROC curve; in this way the number of the features used for classification is reduced to K (different values of K ranging from 1 to 10 have been tested in the experiments). Finally the classification step is performed by a pool of Linear SVM[2] classifiers (Cristianini and Shawe-Taylor 2000), each trained on a different physicochemical property, which are finally combined by the sum rule.

SVMs are widely considered as the state-of-the-art among the machine learning classifiers. The goal of SVMs is to establish the equation of a hyperplane that divides the feature space, leaving all the points of the same class on the same side, while maximizing the distance between the two classes and the hyperplane.

Genetic programming (GP)[3] for designing encoding techniques

In the second method we use the Genetic Programming (GP) (Bhanu and Lin 2004) to synthesize new encoding amino acid models. Different running of GP are executed in order to obtain evolved amino acid encoding models, then a multi-classifier systems is built by combining Linear SVM classifiers trained by the K resulting evolved amino acid encoding models; finally, the K classifiers are combined by the "sum" rule.

In this work the primitive features for GP are encoding models randomly drawn, where each value is a random number between −50 and 50, 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. Children occupy the remaining places in the new population only. We run each GP for 10 times using 200 individuals. We selected as evolved amino acid encoding only the best individual of the last run of a given GP.

This genetic approach is used to obtain evolved amino acid encoding in two different ways:

- **GP1**, each executions of GP is independent from the others executions. The fitness function of GP is given by the area under the ROC curve obtained by the given individual.
- **GP2**, in order to reduce the "correlation" among the evolved amino acid encoding models, the results of all the previous executions of GP are used to drive the actual one. Therefore, the fitness function of the ith GP is given by the area under ROC curve obtained by the fusion with the "sum" rule of a given individual and the evolved amino acid encodings created by the previous 1,…,$i-1$ GPs.

Let us name as **G** the encoding amino acid models obtained by GP, the amino-acid are codified as described by Eq. (2):

$$\mathrm{AA}_i \rightarrow (\delta_{i1} \times \mathbf{G}(1), \ldots, \delta_{iN} \times \mathbf{G}(N)) \quad (2)$$

Quasi-residue couple method

The third method tested in this work, named **QR**, is based on the Quasi-residue couple encoding method (proposed in (Nanni 2006). This encoding technique combines the amino-acid index together with the sequence order of the amino-acids composition. This is achieved by replacing each non-zero entry in a Residue couple model (order 3) (Guo et al. 2005) by the corresponding value appeared in

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	

[2] Implemented as in OSU toolbox. http://www.ece.osu.edu/~maj/osu_svm/.

[3] Implemented as in GAOT (Genetic Algorithms for Optimization Toolbox) http://www.ie.ncsu.edu/mirage/GAToolBox/gaot/.

the amino-acid index. A residue-couple of rank k represents the frequency with which a couple of amino acids at distance k are observed in a protein. The total number of feature is $P = 577$, then the features are reduced to K by SFFS using the same objective function of **PE**. This encoding technique is coupled with Oja's Subspace[4] classifier where for each feature a subspace is computed for each class exploiting at least a fraction 0.95 of the class variance. Finally, the K classifiers are combined by the "sum" rule.

The Oja's Subspace classifiers are based on the Karhunen-Loeve (KL) feature transform (Franco et al. 2006). For each class one KL subspace is created, with the aim of capturing the intra-class variance. A map between the original space and the reduced eigenspace is performed by means of the operator of projection (Duda et al. 2000), the norm of the projection of a pattern on each subspace is used as similarity measure between the input vector and the class related to the subspace. The input vector is then classified according to the maximal similarity value.

The use of Oja's Subspaces coupled with the Quasi-residue encoding is motivated by some preliminary experiments reported in (Nanni and Lumini 2008b) where Oja's Subspaces have proven to outperform SVM for this encoding method.

Genetic approach for building different alphabets

The forth method is based on representing proteins by their N-peptide compositions. In the N-peptide composition for each value of N the corresponding feature vector contains the fraction of each possible N-length substring in the sequence (i.e., it corresponds to amino acid composition for $N = 1$ and dipeptide composition for $N = 2$). In order to limit the high number of dimensions (20^N) required to the formation of feature vectors, especially for large values of N, the size of amino acid alphabet can be reduced from 20 to S using statistical techniques based on the information of certain BLOSUM matrices and justified by well-known biochemical amino acid groups (Murphy et al. 2000).

In (Nanni and Lumini 2008a), an alternative way for the construction of reduced alphabets is studied, based on a Genetic Algorithm for grouping amino-acids, whose objective function is the maximization of the performance of a given classification problem. K different alphabets are created for different couples of the size S of the reduced alphabets and the length N of the substrings. For details on this last method, named **AL**, please read (Nanni and Lumini 2008a). The classification step is performed by a pool of Linear SVM classifiers combined by mean rule.

Results and discussion

The tests have been conducted on the UPPSALA dataset, collected by the authors of (Kontijevskis et al. 2007), which is the biggest dataset ever tested for the HIV protease problem. This dataset contains 1,625 octamer protein sequences $\mathbf{P} = P4P3P2P1P1'P2'P3'P4'$ that are classified as HIV protease cleavable site (374) or uncleavable site (1,251).

The performance is evaluated using the area under the Receiver Operating Characteristic (ROC) curve; the ROC curve is a two-dimensional measure of classification performance that plots the probability of classifying correctly the positive examples against the rate of incorrectly classifying true negative examples. The Area Under the ROC curve (AUC[5]) (Fawcett 2004) is a scalar measure to evaluate performance, which can be interpreted as the probability that the classifier will assign a higher score to a randomly picked genuine sample than to a randomly picked impostor sample. We prefer AUC to accuracy (error rate) as performance indicator since it has been shown (Qin 2006; Huang and Ling 2005) that AUC is empirically and theoretically better than accuracy, due to the fact that accuracy does not considered the scores of the classifiers, it gives mere positive/negative classification results.

The results reported in Table 2 for all the methods described in "Methods" (**O** stands for orthonormal encoding, **PE** is physicochemical encoding representation, **GP1** and **GP2** are encoding techniques obtained by Genetic Programming, **QR** is the Quasi-residue couple method, **AL** is an encoding approach based on alphabets built by a Genetic approach) have been obtained using the following double cross-validation testing protocol. First, the dataset has been randomly divided into ten equally sized subsets D_i, then, we generated ten new datasets (N_i) removing once one of the D_i subsets from the original set. In each of the N_i datasets the ten-fold cross validation is used for finding the parameters of our method, the subset D_i is classified using N_i. Notice that for each of the ten tests, the encoding amino-acid models are built using only N_i.

Due to computation issue, for **QR** method only a subset of N_i (a random subset of 20% of the patterns) is used for finding the parameters (e.g., the encoding amino-acid models).

The results in Table 2 show that:

- all the proposed ensembles work well for this problem, in fact, for all the methods the performance increases with the number of fused classifiers K;
- all the proposed approaches with the exception of **AL** outperform the orthonormal encoding **O** (which is widely considered (Rögnvaldsson et al. 2007) the best feature extractor for this problem);

[4] Implemented as in PRtools 3.1.7 toolbox http://130.161.42.18/prtools/.

[5] AUC is implemented as in dd_tools 0.95 davidt@ph.tn.tudelft.nl.

Table 2 AUC obtained by different methods, varying the value *K* of combined classifiers

K	1	5	10
O	0.9859 (0.0081)	–	–
PE	0.9861 (0.0077)	0.9876 (0.0073)	0.9881 (0.0075)
GP1	0.9860 (0.0071)	0.9870 (0.0070)	0.9878 (0.0071)
GP2	**0.9863** (0.0069)	0.9873 (0.0073)	0.9877 (0.0070)
QR	0.9850 (0.0070)	**0.9880** (0.0065)	**0.9895** (0.0070)
AL	0.9578 (0.0176)	0.9728 (0.0123)	0.9720 (0.0160)

The bold numbers are the higher performance for each column

Table 3 AUC obtained by different methods using the second protocol, varying the value *K* of combined classifiers

K	1	5	10
O	0.9813 (0.0119)	–	–
PE	0.9857 (0.0101)	0.9871 (0.0101)	0.9872 (0.0102)
GP1	**0.9861** (0.0112)	0.9875 (0.0105)	0.9879 (0.0102)
GP2	0.9860 (0.0109)	**0.9879** (0.0105)	**0.9881** (0.0100)
QR	0.9696 (0.0125)	0.9856 (0.0110)	0.9870 (0.0105)
AL	0.9709 (0.0095)	0.9731 (0.0090)	0.9742 (0.0096)

The bold numbers are the higher performance for each column

- it is interesting to note that the artificial methods for building the amino-acid encoding (**GP1** and **GP2**) gain performance comparable with that obtained by **PE** where the real physicochemical properties discovered by scientists are considered;
- the encoding technique based on alphabets built by a Genetic approach (**AL**) obtains the worst performance in this problem; this is probably due to the fact that it is based on the 2-g encoding which is not well suited (Rögnvaldsson et al. 2007) for this classification problem. In fact the AUC obtained by the standard 2-g is 0.9640 in this dataset, while the ensemble of alphabets proposed in (Ogul and Mumcuoglu 2007), for a protein classification problem, obtains an AUC of 0.9600.

The previous results are obtained selecting a different set of properties for each validation set, in order to obtain a unique list of selected properties we run some tests using a different protocol: the properties are selected as those that maximize the AUC on a first set of 10 experiments, in each run 1,400 patterns have been used to build the training set, the other patterns to the test set. Then, another set of 10 experiments is performed and the average results are reported in Table 3. The results obtained with this protocol confirm the outcomes listed above. The list of physicochemical encoding representations found using this protocol for **PE** and **QR** is reported in Table 4.

It is well known in the literature that combining systems based on different classifiers and different feature extractions allows for robust and reliable systems to be obtained. In Table 5 we evaluate the performance of two multi-classifiers obtained as the fusion[6] by sum rule between **QR** and **PE** (**FUS1**) and the fusion[6] by sum rule among **QR**, **PE** and **AL** (**FUS2**); the other methods reported in Table 5 are evaluated for $K = 10$. These tests demonstrate the usefulness of combining different systems: the improving in terms of AUC obtained by the two multi-classifiers is very impressive.

In order to confirm the benefit of our method the DET curve has been also considered. The DET curve (Martin et al. 1997) is a two-dimensional measure of classification performance that plots the probability of false positive against the rate of false negative. In Fig. 1 the DET curves obtained by the five method considered above is reported. It is clear that the two multi-classifiers **FUS1** and **FUS2** obtain the best results.

As further experiment, we have run the Wilcoxon Signed-Rank test (Demsar 2006) for comparing the results (the AUC is used as performance indicator) of **FUS1** and **FUS2** with **O**. The null hypothesis is that there is no difference between the AUC of the considered classifiers (Demsar 2006). We reject the null hypothesis (level of significance 0.05) and accept that in both cases the multi-classifiers significantly improve the AUC of the stand-alone approach.

Finally, we have investigated the relationship between the different approaches combined in the methods **FUS1** and **FUS2** by evaluating the error independence by *Q*-statistic (Kuncheva and Whitaker 2003). For two classifiers G_i and G_j the *Q*-statistic a posteriori measure is defined as:

$$Q_{i,k} = \frac{N^{11}N^{00} - N^{01}N^{10}}{N^{11}N^{00} + N^{01}N^{10}}$$

where N^{ab} is the number of instances in the test set, classified correctly ($a = 1$) or incorrectly ($a = 0$) by the classifier G_i, and correctly ($b = 1$) or incorrectly ($b = 0$) by the classifier G_j. Q varies between -1 and 1; $Q_{i,j} = 0$ for statistically independent classifiers. Classifiers that tend to recognize the same patterns correctly will have $Q > 0$, and those, which commit errors on different patterns, will have $Q < 0$. Table 6 reports the *Q*-static among the methods **QR**, **PE** and **AL**. These results partially motivate the good result obtained by **FUS1** and **FUS2**.

Conclusions

In this paper, we have presented methods based on ensemble of classifiers for HIV-1 protease prediction. The

[6] Before the fusion the scores of the classifiers are normalized to mean 0 and standard deviation 1.

Table 4 Physicochemical properties selected by the **PE** and **QR**

Method	**PE**	**QR**
1st	Entropy of formation	Proportion of residues 100% buried
2nd	Volume	Normalized frequency of C-terminal helix
3rd	Hydropathy scale based on self-information values in the two-state model	The 250 PAM transmembrane protein exchange matrix
4th	Hydrophobicity	Structure-based comparison table for outside beta class
5th	Context-dependent optimal substitution matrices for buried helix	Signal sequence helical potential
6th	Percentage of buried residues	Normalized frequency of alpha-helix
7th	Residue accessible surface area in folded protein	AA composition of mt-proteins from fungi and plant
8th	Accessible surface area	A parameter of charge transfer capability
9th	Absolute entropy	Residue accessible surface area in folded protein
10th	Context-dependent optimal substitution matrices for exposed beta	Average gain in surrounding hydrophobicity

Table 5 Comparison among several methods (AUC)

O	0.9859 (0.0081)
PE	0.9881 (0.0075)
QR	0.9895 (0.0070)
FUS1	0.9940 (0.0027)
FUS2	**0.9950** (0.0024)

The bold number is the higher performance of the column

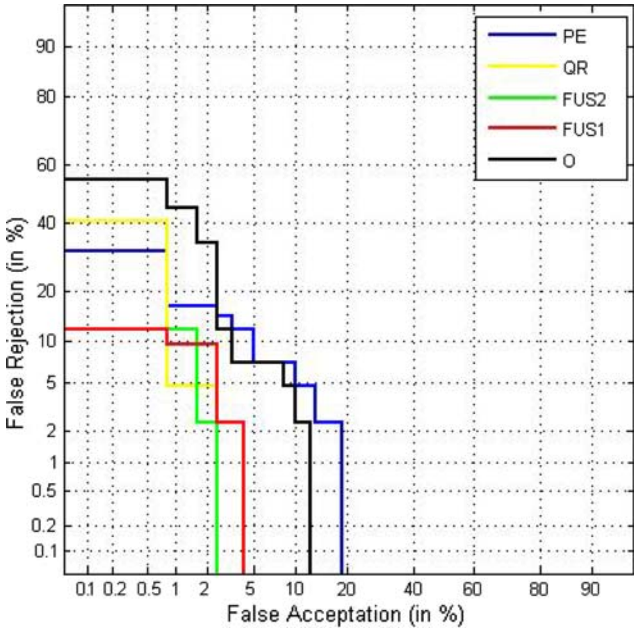

Fig. 1 The DET curves obtained by O, PE, QR, FUS1 and FUS2

Table 6 *Q*-statistic among **QR**, **PE** and **AL**

	PE	**QR**	**AL**
PE	1	0.60	0.85
QR	–	1	0.7
AL	–	–	1

four ensembles described in this work are based on different feature extractions from peptides: two based on the physicochemical properties, another based on the generation of artificial encodings by Genetic Programming, the last based on different alphabets for the *N*-peptide composition. An extensive evaluation on a large dataset according to a blind testing protocol has demonstrated the superiority of these four ensembles with respect to the stand-alone approaches. The experiments demonstrated that all the approaches are well-suited for this classification problem, and above all some of them (in particular **QR** and **PE**) are quite “independent” from each other, thus leading to a further performance improvement if combined together.

Please note that all the reported results have been obtained without any kind of parameter optimization for the SVMs; other works (Rögnvaldsson et al. 2007) have reported the possibility of obtaining performance improvement by means of a fine parameter tuning.

Reproducible research

We try to explain in the better way our methods, anyway we know that some errors in the explanations are always possible, for this reason some Matlab code used for obtaining the results reported in this paper are available at: http://bias.csr.unibo.it/nanni/softwareHIV.rar

Acknowledgments The authors would like to thank: A. Kontijevskis from Uppsala University for sharing the UPPSALA dataset.

References

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993a) Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32:6548–6554

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993b)

Steady-state kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-87201E. J Biol Chem 268:6119–6124

Althaus IW, Gonzales AJ, Chou JJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1993c) The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem 268:14875–14880

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F (1994a) Steady-state kinetic studies with the polysulfonate U-9843, an HIV reverse transcriptase inhibitor. Experientia 50:23–28

Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Chou KC, Kezdy FJ, Romero DL, Thomas RC, Aristoff PA, Tarpley WG, Reusser F (1994b) Kinetic studies with the non-nucleoside HIV-1 reverse transcriptase inhibitor U-90152E. Biochem Pharmacol 47:2017–2028

Althaus IW, Chou KC, Franks KM, Diebel MR, Kezdy FJ, Romero DL, Thomas RC, Aristoff PA, Tarpley WG, Reusser F (1996) The benzylthio-pyrididine U-31,355 is a potent inhibitor of HIV-1 reverse transcriptase. Biochem Pharmacol 51:743–750

Altıncay H, Demirekler M (2000) An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification. Speech Commun 30(4):255–272

Bhanu B, Lin Y (2004) Object detection in multi-modal images using genetic programming. Appl Soft Comput J, vol 4, pp 175–201

Breiman L (1996) Bagging predictors. Mach Learn 24:123–140

Breinman L (2001) Random forest. Mach Learn 45(1):5–32

Cai YD, Chou KC (1998) Artificial neural network model for predicting HIV protease cleavage sites in protein. Adv Eng Softw 29:119–128

Cai YD, Liu X, Xu XB, Chou KC (2002) Support vector machines for predicting HIV protease cleavage sites in protein. J Comput Chem 23:267–274

Chou JJ (1993a) Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J Protein Chem 12:291–302

Chou JJ (1993b) Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J Protein Chem 12:291–302

Chou JJ (1993c) A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. Biopolymers 33:1405–1414

Chou KC (1993d) A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 268:16938–16948

Chou KC (1996) Review: Prediction of HIV protease cleavage sites in proteins. Anal Biochem 233:1–14

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

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) 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) Large-scale plant protein subcellular location prediction. J Cell Biochem 100:665–678

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

Chou KC, Shen HB (2007e) 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, Kezdy FJ, Reusser F (1994) Review: steady-state inhibition kinetics of processive nucleic acid polymerases and nucleases. Anal Biochem 221:217–230

Chou KC, Wei DQ, Zhong WZ (2003) Binding mechanism of coronavirus main proteinase with ligands and its implication to drug design against SARS. (Erratum: ibid., 2003, vol 310, 675). Biochem Biophys Res Commun 308:148–151

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

Demsar J (2006) Statistical comparisons of classifiers over multiple data sets. J Mach Learn Res 7:1–30

Duda RO, Hart PE, Stork G (2000) Pattern classification, 2nd edn. Wiley, New York

Fawcett T (2004) ROC graphs: notes and practical considerations for researchers, technical report. HP Laboratories, Palo Alto

Franco A, Lumini A, Maio D, Nanni L (2006) An enhanced subspace method for face recognition. Pattern Recognit Lett 27:76–84

González-Díaz H, González-Díaz Y, Santana L, Ubeira FM, Uriarte E (2008) Proteomics, networks, and connectivity indices. Proteomics 8:750–778

Guo J, Lin Y, Sun Z (2005) A novel method for protein subcellular localization: combining residue-couple model and SVM. In: Proceedings of third Asia-Pacific bioinformatics conference, pp 117–129

Huang J, Ling CX (2005) Using AUC and accuracy in evaluating learning algorithms. IEEE Trans Knowl Data Eng 17(3):299–310

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

Kittler J (1998) On combining classifiers, IEEE Trans. Pattern Anal Mach Intell 20(3):226–239

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

Kuncheva LI, Whitaker CJ (2003) Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Mach Learn 51:181–207

Lubec G, Afjehi-Sadat L, Yang JW, John JP (2005) Searching for hypothetical proteins: theory and practice based upon original data and literature. Prog Neurobiol 77:90–127

Martin A et al (1997) The DET curve in assessment of decision task performance. In: Proc of EuroSpeech, pp 1895–1898

Melville P, Mooney RJ (2003) Constructing diverse classifier ensembles using artificial training examples. In: Proceedings of the IJCAI, pp 505–510

Murphy LR, Wallqvist A, Levy RM (2000) Simplified amino acid alphabets for protein fold recognition and implications for folding. Protein Eng 13:149–152

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) MppS: an ensemble of support vector machine based on multiple physicochemical properties of amino-acids, NeuroComputing, vol 69, no.13, pp.1688–1690, August 2006

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

Nanni L, Lumini A (2006c) A reliable method for HIV-1 protease cleavage site prediction. Neurocomputing 69:838–841

Nanni L, Lumini A (2008a) A genetic approach for building different alphabets for peptide and protein classification. BMC Bioinformatics 9:45

Nanni L, Lumini A (2008b) Using ensemble of classifiers in Bioinformatics. In: *C*olumbus F (ed) Machine learning research progress, Hauppauge, New York, Nova (to appear)

Narayanan A, Wu X, Yang Z (2002) Mining viral protease data to extract cleavage knowledge. Bioinformatics 18:5–13

Ogul H, Mumcuoglu EU (2007) Subcellular localization prediction with new protein encoding schemes, IEEE Trans on Computational Biology and Bioinformatics

Opitz D, Maclin R (1999) Popular ensemble methods: an empirical study. J Artif Intell Res 11:169–198

Poorman RA, Tomasselli AG, Heinrikson RL, Kezdy FJ (1991) A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J Biol Chem 266:14554–14561

Pudil P, Novovicova J, Kittler J (1994) Flotating search methods in feature selection. Pattern Recognit Lett 15:1119–1125

Qin ZC (2006). ROC analysis for predictions made by probabilistic classifiers. In: Proceedings of the fourth international conference on machine learning and cybernetics, vol 5, pp 3119–3124

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

Rögnvaldsson T, You L, Garwicz D (2007) Bioinformatic approaches for modeling the substrate specificity of HIV-1 protease: an overview. Expert Rev Mol Diagn 7(4):435–451

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

Shen HB, Chou KC (2007a) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85(3):233–240

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(4):1006–1011

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) EzyPred: a top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Commun 364:53–59

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

Shen HB, Chou KC (2008) HIVcleave: a web-server for predicting HIV protease cleavage sites in proteins. Anal Biochem 375:388–390

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

Thompson TB, Chou KC, Zheng C (1995) Neural network prediction of the HIV-1 protease cleavage sites. J Theor Biol 177:369–379

Whitaker CJ, Kuncheva LI (2003) Examining the relationship between majority vote accuracy and diversity in bagging and boosting. In: Technical Report, School of Informatics, University of Wales, Bangor

Zenobi G, Cunningham P (2001) Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In: Raedt LD, Flach PA (eds) Proceedings of the 12th conference on machine learning, Lecture Notes in Computer Science, vol 2167, pp 576–587