Expert Systems with Applications 38 (2011) 3185–3191

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

A new encoding technique for peptide classification

Loris Nanni *, Alessandra Lumini

DEIS – Università di Bologna, viale Risorgimento 2, 40136 Bologna, Italy

ARTICLE INFO

Keywords:
Peptide classification
Amino acid encoding
Physicochemical properties
Machine learning
HIV-protease
Human immune system

ABSTRACT

Research on peptide classification problems has focused mainly on the study of different encodings and the application of several classification algorithms to achieve improved prediction accuracies. The main drawback of the literature is the lack of an extensive comparison among the available encoding methods on a wide range of classification problems. This paper addresses the fundamental issue of which peptide encoding promises the best results for machine learning classifiers. Two novel encoding methods based on physicochemical properties of the amino acids are proposed and an extensive comparison with several standard encoding methods is performed on three different classification problems (HIV-protease, recognition of T-cell epitopes and prediction of peptides that bind human leukocyte antigens). The experimental results demonstrate the effectiveness of the new encodings and show that the frequently used orthonormal encoding is inferior compared to other methods.

1. Introduction

During the last decade there has been a tremendous growth in the amount of available biomedical data, therefore there has been an increased need of systems able to manage and process a very large amount of data. In particular, the study of interactions between proteins within the cell is a nowadays major topic in bioinformatics; such interactions commonly depend on the successful automatic recognition of suitable functional sites which support them. Due to the intrinsic complexity of the interpretation of information contained in the 3D conformations and primary structures of the proteins involved, the motivation for applying machine learning techniques in bioinformatics has shifted from being an illustrative example to becoming a real prerequisite for solving complex problems (i.e. many kinds of functional sites have been already examined, including those for protease cleavage, glycosylation, phosphorylation, acetylation and enzymatic catalysis). Systems aimed at the identification of functional sites typically uses sets of fixed length short peptides (short amino acid chains from a protein). These peptides are represented as multidimensional vectors, by means of an ad hoc encoding technique and classified, according to a given problem, by intelligent systems that learn models from data.

Machine learning methods have gained a lot of success in several bioinformatics fields due to the fact that they are able to extract hidden relationships and correlations among the data, even if the amount of knowledge available is too large for being encoded by human experts. Moreover, in cases when the performance of stand-alone methods is not good enough, ensembles of learning algorithms can be used where averaging the different responses of more classifiers could make the combined system able to produce a better solution. The ensemble of classifiers have been theoretical and empirical (Altıncay & Demirekler, 2000; Kittler, 1998; Opitz & Maclin, 1999) demonstrated to be able of improving 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) (Melville & Mooney, 2003; Whitaker & Kuncheva, 2003; Zenobi & Cunningham, 2001). Three main categories have been proposed for the analysis of existing typologies of ensembles: perturbation of the patterns, where each classifier is trained using a different training set or different weights for the patterns, perturbation of the features, where each classifier is trained using a different feature set and perturbation of the classifiers, where each classifier has different values for its parameters or different classifiers are combined.

The high performance of most ensembles arises mainly from the cooperation between informative features and efficient classifier design (Nanni, 2006). In this work a new encoding method for the representation of peptides is proposed and validated throughout an exhaustive experimentation on several classification problems. The new encoding has been tested both as stand-alone approach and in combination with other well-known encodings from the literature for the design of an ensemble based on the perturbation of the features.

The rest of the paper is organized as follows. Section 2 briefly introduce some well-known problem based on peptide classification and summarizes the previous literature on these problems; Section 3 contains the analysis of many existing encoding tech-

* Corresponding author.
E-mail address: lnanni@deis.unibo.it (L. Nanni).

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.09.005

niques for peptides and presents the new encoding proposed in this work; Section 4 reports experimental results obtained on three different classification problems and discuss the validity of the new approach through an exhaustive comparison with the other state-of-the-art methods; finally, Section 5 draws some conclusions.

2. Peptide classification problems

In the literature there are several problems based on the classification of peptides; in particular high relevance has been given in developing systems based on peptide classification which can be useful in helping the medical experts in diagnosing and vaccine discovery. In this work we test our system in three well-known case studies: HIV-protease, recognition of T-cell epitopes and prediction of peptides that bind human leukocyte antigens (two datasets).

The first case study is related to the problem of discovering efficient HIV-1 protease inhibitors. AIDS is a grave disease of the immune system transmitted through HIV; HIV depends on protease in its reproductive cycle, thus, it is important to discover efficient HIV-1 protease inhibitors as antiviral means. In terms of peptides' classification this problem can be stated as to find out which peptides are cleaved by the HIV-1 protease and which are not. The protease–peptide interactions are modeled by the "lock" and "key" model, where the amino acids in the peptide **P** (the "key") are tested to find out if they fit the positions in active site pockets **S** of the protease (the "lock").

Several works that try to solve the HIV-1 protease specificity problem by applying techniques from machine learning have been published (see Rögnvaldsson, You, and Garwicz (2007) for a good review): the first approaches were based on a standard feed-forward multilayer perceptron (Cai & Chou, 1998), next it has been shown (Rögnvaldsson & You, 2003) that HIV-1 protease cleavage is a linear problem and that the best classifier for this problem is the linear support vector machine (SVM).

The other case studies are related to human immune system. "The immune system is a complex of cells, molecules and organs with the primary role of limiting damage to the host organism by pathogens, which elicit an immune response and thus are called antigens" (De Castro & Timmis, 2002). Adaptive immune response is the main defense mechanism of the body and it is mainly constituted by lymphocytes (mainly T and B cells). The illustration of the working procedure of the immune system is shown in Fig. 1.

The immune response is activated only when the T-cell recognizes the antigen, then the T-cell clone will be activated, and the cellular immune will happen. However, not all the MHC-peptide complexes can be recognized by T-cell receptors: the portions of short binding peptides, which can be recognized, are called T-cell epitopes. Deciphering the patterns of peptides that elicit a MHC restricted T-cell response (Huang & Dai, 2005) is critical for vaccine development, therefore many automated systems have been proposed to study the interaction between peptide and MHC; the most

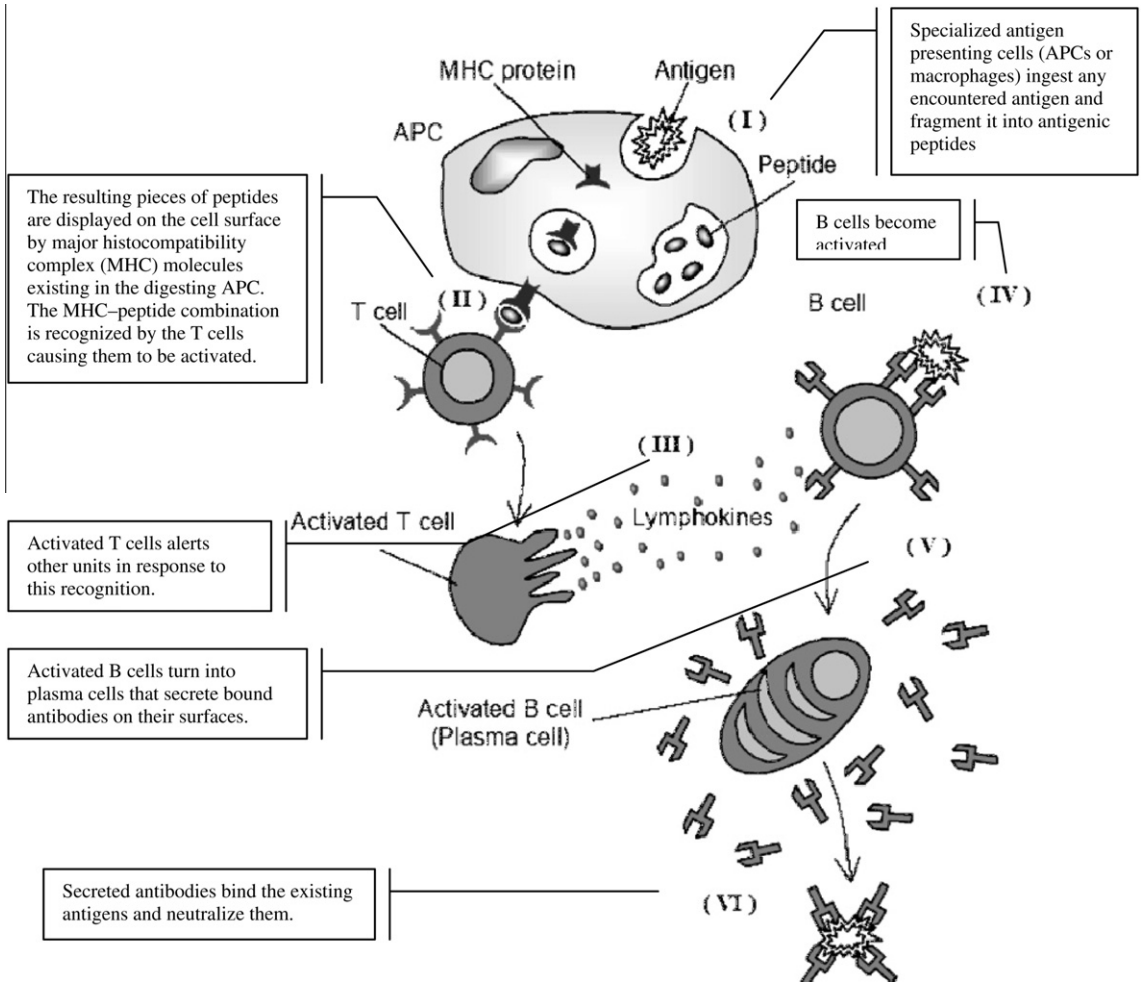

Fig. 1. Human immune system. From De Castro and Timmis (2002).

relevant are based on: structural information (Madden, 1995); mathematical approaches including binding motifs (Hammer, 1995); quantitative matrices (Sturniolo et al., 1999); artificial neural networks (Honeyman, Brusic, Stone, & Harrison, 1998; Milik et al., 1998); SVMs (Huang & Dai, 2005; Zhao, Pinilla, Valmori, Roland Martin, & Simon, 2003).

The human form of MHC is denoted as Human Leukocyte Antigen (HLA): the prediction of peptides that bind multiple HLA molecules is crucial in the designing of vaccines that are useful to a broader population. Several works have been developed for identification of HLA binding peptides, by means of SVMs (Bozic, Zhang, & Brusic, 2005), artificial neural networks (Brusic et al., 2002) and hidden Markov models (Zhang et al., 2005).

All the tests reported in this work have been conducted on four datasets (described in Section 5): a HIV dataset (**HIV**), a peptide dataset for the recognition of T-cell epitopes **(PEP)** and two vaccine datasets (**VAC1** and **VAC2**).

3. Encoding techniques for peptides

Several encoding techniques have been proposed for representing sequence of amino acids (e.g. Atchley, Zhao, Fernandes, & Drüke, 2005; Chou, 2001) in multidimensional metric spaces. Unfortunately most of the work is related to the encoding of proteins, while only few methods have been specifically developed for peptides (e.g. Atchley et al., 2005; Nanni & Lumini, 2006). In the last years, peptides have gained much interest in many medical applications (i.e. successful malaria (Lopez, Weilenman, Audran, et al., 2001) and antitumour (Knutson, Schiffman, & Disis, 2001) vaccines based on peptides). The main difference between protein and peptides is that the proteins have variable lengths while in a given problem the peptides have a fixed length. For this reason in the proteins it is very important to develop a system that considers also the sequence (as the pseudo amino acids encoding of Shen and Chou (2007) while in the peptides the sequence can be intrinsically obtained by simply concatenating the features that describes each amino acid.

In particular in this work we are interested in the encoding methods well suited to be coupled with a machine learning classifier. In the following we review several existing encoding methods that can be used to extract a set of features from a peptide.

3.1. Encodings based on the amino acid sequence

The encodings described in this sub-section are based only on the amino acid sequence:

- *Amino acid composition* (**AC**): this simple encoding calculates the frequency of each amino acid inside the sequence, therefore the feature vector is composed by 20 features.
- *Orthonormal encoding* (**OE**): the most frequently used encoding is the orthonormal one, also known as distributed encoding or sparse encoding (Qian & Sejnowskij, 1988). Each amino acid of the sequence is mapped in a sparse orthonormal vector space by a 20-bit vector with 19 bits set to zero and one bit set to one. Therefore a peptide, which is a sequence of M consecutive amino acid letters can be represented by the concatenation of $M \times 20$ features.
- *2-Grams* (**2G**): this representation has been proposed for proteins (Wu, Whitson, McLarty, Ermongkonchai, & Change, 1992), but can be used also for peptides; a peptide is represented by a set of 20^2 pairs of values (v_i, c_i), where v_i is a couple of amino acids and c_i is the counts of that couple in a peptide sequence (scaled using the length of the sequence). A variant of this method is the residue couple (originally developed for proteins (Guo, Lin, & Sun, 2005). A residue couple of rank k represents the frequency with which a couple of amino acids at distance k are observed in a protein.

3.2. Encodings based on physicochemical properties

The amino acid index database (Kawashima & Kanehisa, 2000) contains a set of the physicochemical properties measured for each amino acids: currently 544 indices and 94 substitution matrices are maintained. The encodings described in this sub-section are based on physicochemical properties of the amino acids:

- *Chou's pseudo amino acid composition*: the pseudo amino acid composition is a $(20 + K)$ length feature vector where the first 20 features are the amino acid composition and the other K reflect the effect of sequence order by the correlation factors of the different ranks introduced by Chou (2001). The last K features are obtained considering a given physicochemical property p according to the following formula:

$$P^p_{20+k} = \frac{1}{M-k} \sum_{a=1,\ldots,M-k} val(p, A_a) \times val(p, A_{a+k})$$

  where $val(p, Aa)$ is a normalized value of the physicochemical property p for the amino acid in the ath position of the sequence, M is the length of the sequence and $A_a \in \{A, C, \ldots, Y\}$.
- *Physicochemical encoding* (**PE**): this encoding is particularly suited for peptides since it exploits the fixed length of the sequence. In (Zhao et al. (2003) each amino acid of the peptide is encoded by $F = 10$ factors concatenated (i.e. each amino acid is described by a vector $\mathbf{a} \in \Re^F$, therefore the feature vector is composed by $F \times M$ features. These orthogonal factors (**sa**) $\mathbf{a}_i$, $i = 1, \ldots, 20$ were obtained from 188 physical properties of 20 amino acids via multivariate statistical analyses (Kidera, Konishi, Oka, Ooi, & Scheraga, 1985). Several other encodings have been proposed based on a similar representation, but starting from a different set of factors (derived from the physicochemical properties). In Maetschke, Towsey, and Bodén (2005) the set of factors was obtained by reducing the BLOSUM62 matrix by Sammon projection (Henikoff & Henikoff, 1992) to a $20 \times F$ ($F = 5$) matrix (**bs**). In Venkatarajan and Braun (2001) a multidimensional scaling of 237 physicochemical properties is performed to derive $F = 5$ descriptors for all 20 amino acids (**ms**). In a recent work (Tong, Liu, Zhou, Wu, & Li, 2008) a novel descriptor of $F = 9$ principal component scores is derived from the principal component analysis of a matrix of 99 weighted holistic invariant molecular indices of amino acids (**sw**).
- *Weighted physicochemical encoding* (**WP**): this encoding is a variant of the above one, originally developed for proteins, where instead of considering the amino acids in their natural order, they are concatenated alphabetically and weighed according to their frequency in the sequence. Therefore the feature vector is composed by $20 \times F$ features (**fa**). The original work Mundra, Kumar, Kumar, Jayaraman, and Kulkarni (2007) is based on the $F = 5$ factor solution scores for amino acid obtained by the factor analysis in Atchley et al. (2005). The factor scores of each amino acid are multiplied for its fraction in the given peptide and concatenated together.
- *Average physicochemical encoding* (**AP**): this encoding, as the above one, is invariant to the length of the sequence, thus mainly suited for proteins. Each feature is represented by the average value of a physicochemical property with respect to the amino acid in the sequence, therefore the feature vector is composed by F features. In Huang, Tung, Huang, Hwang, and Ho (2007)) the whole set of physicochemical properties in Kawashima and Kanehisa (2000) are used ($F = 544$).

Table 1
Nineteen PCA factors for amino acids.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
A	57.4	−409.7	−50.1	565.8	−243.7	305.9	−59.6	48.0	−100.2	−13.7	6.8	9.9	104.7	−26.1	6.9	96.6	−59.9	96.6	29.7
R	237.4	213.6	61.8	705.1	−73.6	253.8	81.3	162.3	−83.8	−2.6	45.3	39.7	121.0	−25.4	14.9	101.0	−62.0	96.8	33.5
N	251.7	15.7	−9.0	536.4	−46.2	132.2	−29.0	55.5	−39.6	−10.7	−7.0	49.5	118.4	−67.8	32.8	83.3	−59.4	102.5	10.1
D	93.5	−15.1	304.7	467.6	−96.1	250.3	−122.8	118.9	−75.1	−11.2	61.5	27.3	127.3	−28.5	57.8	95.2	−61.5	90.3	32.0
C	−519.8	−131.9	280.7	337.8	−70.9	254.9	73.9	42.5	−42.2	−46.2	23.9	38.8	108.6	−42.0	21.1	96.6	−62.0	97.0	36.2
Q	247.9	159.3	58.9	533.6	−46.4	155.9	−2.9	52.1	−64.6	−37.3	−3.7	−10.1	62.9	−14.4	49.4	108.0	−55.0	104.2	37.5
E	224.8	52.2	220.1	571.9	−81.2	216.0	−68.7	91.1	−74.7	−20.6	35.1	21.2	79.2	−87.5	8.3	101.5	−77.6	106.9	47.7
G	209.1	−395.4	−20.7	506.2	59.0	335.2	−21.0	114.0	13.9	17.8	45.6	18.9	87.0	−38.9	30.3	108.7	−68.3	102.1	31.6
H	−28.1	208.4	−158.7	458.8	−249.6	311.2	−0.2	58.2	−40.6	6.0	48.9	54.3	94.7	−57.7	54.5	108.4	−64.6	103.7	40.5
I	−297.2	−143.2	−4.0	641.3	−69.7	141.9	−7.3	44.0	−64.5	−8.8	33.3	7.1	132.9	−42.7	42.2	125.0	−99.0	99.0	33.6
L	−248.5	−271.8	19.4	743.8	−32.3	230.9	−15.5	59.1	−74.6	−49.0	63.3	29.6	104.8	−67.4	44.2	114.7	−31.9	98.9	35.6
K	311.1	25.6	284.1	731.6	−78.8	310.0	1.0	−56.8	−36.7	26.6	42.9	32.6	107.2	−37.0	33.6	96.3	−63.8	98.1	35.1
M	−346.0	78.0	0.5	506.0	−72.6	86.9	−57.6	49.8	−36.5	51.8	65.9	19.8	107.4	−31.5	9.0	100.2	−45.6	100.1	41.0
F	−453.6	−30.4	40.3	642.8	−32.0	240.4	−23.6	74.1	−85.7	20.2	23.0	45.0	59.4	−46.4	37.9	87.8	−73.1	70.4	30.4
P	203.3	7.3	−107.7	420.6	99.8	273.3	−31.3	6.1	−136.0	−15.2	62.2	50.6	108.0	−35.7	26.6	99.9	−68.0	101.9	37.1
S	227.3	−228.5	33.5	543.0	−31.6	203.1	−12.1	71.3	−63.4	18.9	−33.2	48.9	128.4	−48.5	41.0	102.0	−55.1	86.7	59.2
T	15.1	−162.6	224.6	536.7	−87.7	198.3	−33.7	71.1	−88.7	24.4	12.3	81.5	83.2	−33.2	28.4	142.0	−59.3	106.9	26.6
W	−554.1	240.4	76.8	596.0	12.1	377.3	−61.6	75.6	−77.7	21.1	−20.9	12.3	117.4	−48.1	31.6	107.1	−56.7	113.3	32.4
Y	−167.2	109.6	−7.6	673.3	−45.4	245.9	−103.1	55.1	−8.9	−65.2	17.3	75.0	104.4	−17.1	17.8	103.8	−69.1	97.5	41.4
V	−256.1	−302.5	91.7	651.4	−74.2	194.0	−12.5	84.0	−79.7	7.6	32.4	55.4	92.4	−28.2	46.9	71.9	−70.5	128.0	42.0

3.3. A new encoding

In this work we propose to use the physicochemical encoding with several different sets of factors obtained by two feature transform methods. In particular, starting from the whole set of physicochemical properties in Kawashima and Kanehisa (2000) (all the indices and the diagonals of the substitution matrices) the principal component analysis (PCA) and the non-linear Fisher transform (NLF)[1] are applied and tested for different values of the number of retained features *F*.

3.3.1. Principal component analysis

Given a set of 20 amino acids, each represented by n physicochemical properties $\mathbf{A} = \{\mathbf{x}_i \in \Re^n | i = 1, \ldots, 20\}$ the k dimensional eigenspace is obtained by selecting the first k eigenvectors from the covariance matrix of the original space.

In Table 1 the resulting matrix of PCA factors (**pc**) is reported which explained the whole variance.

3.3.2. Non-linear Fisher transform

Since the non-linear Fisher is a supervised feature transform we assign to each amino acid a different label. The objective function of the non-linear Fisher transform is to better discriminate patterns of different classes. One of the drawbacks of the basic Fisher transform is that it usually gives a bad approximation of the patterns in presence of multi-class problems due to the class conjunctions. In fact Fisher mapping tends to emphasizes large class distances, by which preserving the distances of already well separated classes may result into an occlusion of neighboring classes. The NLF has been proposed (Duin, Loog, & Haeb-Umbach, 2000) to deal with this problem.

Before the application of the NLF the data are processed by PCA (as in Franco, Lumini, Maio, and Nanni (2006)) to de-correlate the data. In Table 2 the resulting matrix of factors obtained by NLF (**nl**) is reported.

4. Experimental results

The tests have been conducted on the following datasets of peptides:

HIV dataset (HIV) – The dataset contains octamer protein sequences, each of which needs to be classified as an HIV-protease cleavable site or uncleavable site. An octamer protein sequence is a peptide (small protein) denoted by **P** = *P*4*P*3*P*2*P*1*P*1′*P*2′*P*3′*P*4′. The scissile bond is located between positions *P*1 and *P*1′. The dataset **HIV** (Kontijevskis, Wikberg, & Komorowski, 2007) contains 1625 octamer protein sequences (374 cleavable and 1251 uncleavable). The ten fold cross-validation testing protocol is used.

Peptide dataset (PEP) – This dataset for the recognition of T-cell epitopes contains 203 synthetic peptides and it is the same used in Zhao et al. (2003). Peptides were synthesized by the simultaneous-multiplepeptide-synthesis methods and characterized using HPLC and mass spectrometry. The ten fold cross-validation testing protocol is used.

Vaccine datasets (VAC1 and VAC2) – The two datasets (Bozic et al., 2005) contain peptides of length 9 and are built considering five HLA-A2 supertypes variants (**VAC1**) and seven HLA-A3 supertypes variants (**VAC2**). **VAC1** contains 3041 samples (664 belong to the class Binders), while **VAC2** contains 2216 samples (680 belong to the class Binders). The testing protocol proposed by Bozic et al. (2005) is used: all peptides (binders and non-binders) related to a given variant are used as test set and the remaining as training data (see Table 3).

As performance indicator the error under the ROC curve (EUC)[2] (Fawcett (2004)) is used. 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 negative examples. The EUC[3] is a scalar measure to evaluate performance which can be interpreted as the probability that the classifier will assign a lower score to a randomly picked genuine sample than to a randomly picked impostor sample.

The first test is aimed at comparing several encoding models proposed in the literature. Therefore an exhaustive evaluation of the encodings described in Section 2 is performed on the four datasets described above. For a fair comparison the same classifier is been used in all the tests: the support vector machine (SVM) which is widely considered the state-of-the-art classifier. The linear SVM is used for the orthonormal encoding, according to the suggestion given by Rögnvaldsson and You (2003) for **HIV** and confirmed by internal tests also in the other problems, while the radial basis

[1] Both implemented as in PRTools 3.1.7 MATLAB Toolbox.

[2] Implemented as in DDtool 0.95 Matlab Toolbox.

[3] EUC = 1 − AUC, AUC is the area under the ROC curve.

Table 2
Eighteen NLF factors for amino acids.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
A	0.42	−2.07	−0.67	0.01	−1.10	−0.32	−0.20	0.09	−0.20	0.09	−0.11	0.15	0.01	0.06	0.02	0.16	0.07	−0.03
R	1.65	1.40	−0.01	−0.88	−0.08	−0.07	0.60	−0.53	−0.10	0.01	0.09	−0.07	0.09	0.08	0.03	0.09	0.03	−0.02
N	1.68	0.30	−0.49	0.15	0.09	0.59	−0.06	0.02	0.14	0.00	−0.14	−0.09	0.08	−0.14	−0.11	−0.01	0.01	0.01
D	0.81	0.13	1.36	0.63	−0.15	−0.10	−0.45	−0.31	−0.10	0.03	0.15	0.02	0.16	0.12	−0.07	−0.11	−0.01	−0.05
C	−2.70	−0.32	1.19	1.37	0.04	−0.18	0.64	0.21	0.26	0.35	−0.02	−0.11	0.05	−0.04	−0.03	0.10	0.04	−0.04
Q	1.71	1.11	−0.08	0.15	0.11	0.45	0.11	0.08	0.02	0.25	−0.12	0.25	−0.20	0.16	−0.01	−0.07	0.02	0.03
E	1.56	0.48	0.87	−0.02	−0.07	0.13	−0.22	−0.15	−0.09	0.10	0.04	0.05	−0.12	−0.28	0.03	0.09	−0.06	0.02
G	1.32	−2.05	−0.60	0.31	0.61	−0.58	0.00	−0.30	0.44	−0.14	0.18	0.12	−0.12	0.01	0.06	0.00	−0.04	0.01
H	0.13	1.50	−1.22	0.52	−1.14	−0.45	0.13	0.04	0.10	−0.07	0.11	−0.13	−0.06	−0.07	−0.01	−0.16	−0.06	0.03
I	−1.52	−0.45	−0.39	−0.36	−0.01	0.55	0.06	0.10	−0.02	0.08	0.10	0.12	0.18	0.01	0.12	−0.02	−0.20	−0.01
L	−1.29	−1.21	−0.25	−0.96	0.18	0.06	−0.04	0.00	−0.09	0.26	0.18	−0.05	0.00	−0.11	0.01	−0.15	0.14	0.02
K	2.03	0.26	1.22	−0.98	−0.05	−0.32	0.10	0.73	0.11	−0.19	0.14	0.02	0.03	0.02	−0.05	0.01	−0.01	−0.02
M	−1.72	0.85	−0.34	0.44	−0.01	0.80	−0.16	0.05	0.05	−0.30	0.29	0.06	−0.02	0.03	0.03	0.09	0.14	0.00
F	−2.37	0.23	−0.09	−0.37	0.19	−0.04	0.03	−0.06	−0.14	−0.14	−0.10	0.03	−0.21	−0.04	−0.09	−0.03	−0.06	−0.18
P	1.41	0.27	−1.09	0.77	0.87	−0.33	−0.04	0.27	−0.43	0.06	0.10	−0.14	0.03	0.03	−0.01	0.04	−0.03	0.01
S	1.47	−1.11	−0.27	0.13	0.15	0.22	0.09	−0.05	0.05	−0.14	−0.30	0.01	0.16	−0.03	−0.01	−0.06	0.05	−0.06
T	0.30	−0.68	0.88	0.23	−0.10	0.23	0.03	−0.01	−0.14	−0.16	−0.15	−0.20	−0.12	0.05	0.21	−0.07	0.00	0.04
W	−2.83	1.79	0.16	−0.14	0.42	−0.84	−0.13	−0.06	−0.04	−0.18	−0.32	0.26	0.15	−0.08	0.04	−0.01	0.06	0.12
Y	−0.70	0.95	−0.36	−0.60	0.09	−0.06	−0.55	0.01	0.28	0.17	−0.12	−0.23	−0.02	0.13	0.05	0.08	−0.02	−0.02
V	−1.33	−1.39	0.15	−0.40	−0.04	0.27	0.07	−0.12	−0.10	−0.06	−0.01	−0.09	−0.07	0.10	−0.20	0.02	−0.08	0.13

Table 3
Number of binders (B) and non-binders (NB) in training and test data for the different variants of VAC1 and **VAC2**.

VAC1	Training		Test		**VAC2**	Training		Test	
	B	NB	B	NB		B	NB	B	NB
0201	224	378	440	1999	0301	573	1447	107	89
0202	619	2361	45	25	0302	534	1277	146	259
0204	641	2162	23	224	1101	538	1313	142	223
0205	648	2346	16	40	1102	538	1325	142	211
0206	621	2349	43	37	3101	636	1482	44	54
					3301	645	1474	35	62
					6801	621	898	59	638

Table 4
EUC obtained by the encoding methods tested in this paper on the different datasets.

	AA	**OE**	**2G**	**WP**	**AP**	**PE**
HIV	0.122	**0.014**	0.018	0.063	0.064	0.031
PEP	0.158	0.102	0.106	0.198	0.163	**0.089**
VAC1	0.218	0.135	0.155	0.217	0.219	**0.127**
VAC2	0.317	0.136	0.204	0.279	0.214	**0.135**

SVM is used for the other tested encodings. No ad hoc optimization of the SVM parameters has been of performed ($C = 1$ and Gamma = 0.1). Moreover, before the classification the data are linearly normalized to [0–1] using the training data.

In the following Table 4 the EUC obtained by several encoding methods is reported:

- **AC**, amino acid composition;
- **OE**, the orthonormal encoding;
- **2G**, the 2-gram composition;
- **WP**, the weighted physicochemical encoding proposed by Mundra et al. (2007);
- **AP**, the average physicochemical encoding used in Huang et al. (2007);
- **PE**, the physicochemical encoding with the factors used by Mundra et al. (2007) (**fa**).

Please note that Chou's pseudo amino acid composition has a too low dimensionality using only the few physicochemical properties suggested in Chou (2001), while performing a selection among all the indices available in Kawashima and Kanehisa (2000) is beyond the aim of this paper (the interested reader can refer to Nanni and Lumini (in press) for recent results concerning feature selection).

From the results reported in Table 4 it is clear that **PE** is the best encoding among those that use physicochemical properties since it is able to directly encode the amino acid sequence. In the **HIV** dataset the best approach is **OE** (as yet stated by Rögnvaldsson et al. (2007)) while in the other problems **PE** outperforms also the basic encodings not based on physicochemical properties.

Once stated that **PE** is an encoding very suited for several peptide classification problems, the second experiment is aimed at comparing different factor matrices proposed in the literature and the new matrices proposed in this paper:

- **fa**, $F = 5$ factors used in Mundra et al. (2007);
- **bs**, $F = 5$ factors used in BLOMAP (Maetschke et al., 2005);
- **sa**, $F = 10$ orthogonal factors proposed by Kidera et al. (1985) and Zhao et al. (2003);
- **ms**, $F = 5$ factors obtained by multidimensional scaling in Venkatarajan and Braun (2001);
- **sv**, $F = 9$ descriptors proposed in Tong et al. (2008);
- **pc**, the first F eigenfeatures listed in Table 1;
- **nl**, F=19 non-linear Fisher features listed in Table 2.

Please note that the results related to the PCA features (***pc-F***) are shown as a function of F in order to determine the best number of features to be selected.

From the results reported in Table 5 the following conclusions can be drawn:

- the best performance is obtained by the methods with the largest number of factors F, in particular the best results are obtained by the proposed matrices (**pc-19** and **nl**);
- the number of features used for **pc** should be larger than 15 (15 or 19)
- **sa** is the best among the yet proposed factor matrices, even if it is the older one (it was proposed in 1985);
- **pc-19** outperforms also the orthonormal encoding (**OE**) in the **HIV** dataset; to the best of our knowledge, this is the best result obtained on **HIV** from a stand-alone method (Rögnvaldsson et al., 2007).

The third test is aimed at evaluating the performance of the new encoding **nl** in combination with other well-known encodings

Table 5
EUC obtained by the **PE** encoding using different factor matrices.

	fa	bs	sa	ms	sv	pc				nl
F	5	5	10	5	9	5	10	15	19	19
HIV	0.031	0.029	0.014	0.023	0.038	0.034	0.020	0.018	**0.012**	0.014
PEP	0.089	0.119	0.093	0.109	0.088	0.134	0.135	0.091	0.090	**0.087**
VAC1	0.127	0.121	0.108	0.129	0.108	0.128	0.097	0.097	0.101	**0.094**
VAC2	0.135	0.133	0.137	0.132	0.133	0.144	0.130	0.127	0.130	**0.124**

Table 6
EUC obtained by fusing **nl** with the method in the first row of the table.

	AA	OE	2G	WP	AP	sa	sv
HIV	0.022	**0.010**	**0.010**	0.018	0.019	0.011	0.017
PEP	0.115	0.085	0.079	0.118	0.104	**0.060**	0.063
VAC1	0.123	0.101	0.099	0.125	0.117	0.096	**0.094**
VAC2	0.172	0.127	0.140	0.164	0.143	0.128	**0.126**

Table 7
EUC obtained on HIV by fusing the couple of methods in the first row and column.

HIV	OE	AI	pc-19
OE	0.014	0.011	**0.008**
AI	–	0.012	**0.008**
pc-19	–	–	0.012

from the literature for the design of an ensemble based on the perturbation of the features. In Table 6 the EUC obtained by combining with sum rule the methods listed in the first row and **nl** is reported.

The perturbation of features for the design of multi-classifiers is very useful in the first two datasets where the performance of the best combination is lower than that of the best stand-alone approach. In particular the proposed encoding **nl** is well suited to be fused with the orthonormal encoding **OE** and with **sa**. Different conclusions could be drawn for the vaccine datasets where none of the fusions allows a performance improvement with respect the best stand-alone approach. In this problem other methods for combining classifiers and other fusion rules should be evaluated in order to find an effective multi-classifier: for example in Nanni and Lumini (in press) it has been shown that good performance can been obtained designing a multi-classifier based on the perturbation of classifiers.

Finally, a further test is carried out in the **HIV** dataset, in order to evaluate the performance obtained by combining the best approaches published in the literature for this problem. In particular, in Table 7 all the 2 by 2 combinations (by sum rule) of the orthonormal encoding, the method (named **AI**) based on the amino acid indices proposed in Nanni and Lumini (2006), and the new encoding **pc-19** are reported (the performance of the stand-alone method is in diagonal). In order to perform the optimization of parameters for **AI**, the experiments have been conducted using the following double cross-validation testing protocol: first, the dataset has been randomly divided into ten equally sized subsets D_i used as test sets, then, ten new datasets (N_i) used as training sets are generated removing once one of the D_i subsets from the original set. In each of the N_i datasets the 10-fold cross validation is used for finding the parameters of **AI** (i.e. the amino acid indices), the subset D_i is classified using the parameters optimized in N_i.

The results are very impressive: the performance of the fusions including the proposed encoding (third column) improves the state-of-the-art approaches. If we combine all the three classifiers we obtain an EUC of 0.007.

Notice that before the fusion the scores of the classifiers are normalized to mean 0 and standard deviation 1.

5. Conclusions

The motivation of this work is the lack of an extensive comparison among the available encoding methods for peptides in the literature. In this paper the fundamental issue of which peptide encoding promises the best results for machine learning classifiers in different peptide classification problems is addressed. In the context of three fundamental classification problems like HIV-protease, recognition of T-cell epitopes and prediction of peptides that bind human leukocyte antigens, we have proposed two novel encoding methods based on physicochemical properties of the amino acids that have gained very valuable performance. In particular the encoding based on non-linear Fisher transform of the physicochemical properties has outperformed several other encoding methods used for comparison in all the tested problems. The reasons of the good performance gained by the novel approach is the use of a longer codify (18 factors per amino acid) and of a supervised feature transform for calculating the best factors.

The other encoding based on principal component analysis of the physicochemical properties has obtained very good results in particular in the HIV-protease problem, where it has gained the best performance both as stand-alone approach and fused with other state-of-the-art methods.

Possible improvements and extensions of this work are related to the analysis of the many feature transforms proposed in the literature to be applied for the calculation of different factor sets used for the amino acid encoding: e.g. several kernel feature transforms that have been successfully applied in many pattern recognition problems.

Acknowledgments

The authors would like to thank: I. Bozic for sharing the vaccine dataset; L. Huang for sharing the peptide dataset; Kontijevskis for sharing the HIV dataset.

Appendix A. Matlab code

To extract (PRTools 3.1.7 Matlab Toolbox) the **pc**-*F* factor matrix:

```
w = pca(dataset(MM'), F);
%the matrix MM contains the physicochemical
  properties
MM = +(w * dataset(MM'));
```

To extract (PRTools 3.1.7 Matlab Toolbox) the **nl** factor matrix:

```
w = pca(dataset(MM'));
y = 1:20; MM = +(w * dataset(MM'));
w = nlfisherm(dataset(MM, y'));
MM = +(w * dataset(MM'));
```

References

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

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Science, 102*, 6395–6400.

Bozic, I., Zhang, G. L., & Brusic, V. (2005). Predictive vaccinology: Optimization of predictions using support vector machine classifiers. *Ideal*, 375–381.

Brusic, V., Petrovsky, N., Zhang, G., & Bajic, V. B. (2002). Prediction of promiscuous peptides that bind HLA class I molecules. *Immunology and Cell Biology, 80*, 280–285.

Cai, Y. D., & Chou, K. C. (1998). Artificial neural network model for predicting HIV protease cleavage sites in protein. *Advances in Engineering Software, 29*, 119–128.

Chou, K.-C. (2001). Prediction of protein cellular attributes using pseudo-amino acid composition. *PROTEINS: Structure, Function, and Genetics, 43*, 246–255.

De Castro, L. N., & Timmis, J. (2002). *Artificial immune systems: A new computational intelligence approach*. UK: Springer.

Duin, R. P. W., Loog, M., & Haeb-Umbach, R., (2000). Multi-class linear feature extraction by nonlinear PCA. In *ICPR2000, September 03–08, Barcelona, Spain*.

Fawcett, T. (2004). *ROC graphs: Notes and practical considerations for researchers*. Technical report. Palo Alto, USA: HP Laboratories.

Franco, A., Lumini, A., Maio, D., & Nanni, Loris (2006). An enhanced subspace method for face recognition 1 January 2006. *Pattern Recognition Letters, 27*(1), 76–84.

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).

Hammer, J. (1995). New methods to predict MHC-binding sequences within protein antigens. *Current Opinion Immunology, 7*(2), 263–269.

Henikoff, S., & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National academy of Sciences of the United States of America, 89*, 10915–10919.

Honeyman, M. C., Brusic, V., Stone, N. L., & Harrison, L. C. (1998). Neural network-based prediction of candidate T-cell epitopes. *Nature Biotechnology, 16*(10), 966–969.

Huang, L., & Dai, Y. (2005). A support vector machine approach for prediction of T cell epitopes. In *Proceedings of the third Asia–Pacific bioinformatics conference (APBC2005), Singapore, January 17–21, 2005* (pp. 312–328).

Huang, W.-L., Tung, C.-W., Huang, H.-L., Hwang, S.-F., & Ho, S.-Y. (2007). ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features. *Biosystems, 90*, 573–581.

Kawashima, S., & Kanehisa, M. (2000). AAindex: Amino acid index database. *Nucleic Acids Research, 28*, 374<http://www.genome.jp/dbget/aaindex.html>.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., & Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry, 4*, 23–55.

Kittler, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20*(3), 226–239.

Knutson, K. L., Schiffman, K., & Disis, M. L. (2001). Immunization with a HER-2/neu helper peptide vaccine generates HER-2/neu CD8 T-cell immunity in cancer patients. *Journal of Clinical Investigation, 107*, 477–484.

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

Lopez, J. A., Weilenman, C., Audran, R., et al. (2001). A synthetic malaria vaccine elicits a potent CD8 (+) and CD4 (+) T lymphocyte immune response in humans. Implications for vaccination strategies. *European Journal of Immunology, 31*, 1989–1998.

Madden, D. R. (1995). The three-dimensional structure of peptide–MHC complexes. *Annual Review of Immunology, 13*(5), 587–622.

Maetschke, S., Towsey, M., & Bodén, M. (2005). BLOMAP: An encoding of amino acids which improves signal peptide cleavage prediction. In M. Jayakumar & M. Ramya (Eds.), *Proceedings of the Asia–Pacific bioinformatics conference* (pp. 141–150). Singapore: Imperial College Press.

Melville, P., & Mooney, R. J. (2003). Constructing diverse classifier ensembles using artificial training examples. in: *Proceedings of the IJCAI* (pp. 505–510).

Milik, M., Sauer, D., Brunmark, A. P., Yuan, L., Vitiello, A., Jackson, M. R., et al. (1998). Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nature Biotechnology, 16*(8), 753–756.

Mundra, P., Kumar, M., Kumar, K. K., Jayaraman, V. K., & Kulkarni, B. D. (2007). Using pseudo amino acid composition to predict protein subnuclear localization: Approached with PSSM. *Pattern Recognition Letters, 28*, 1610–1615.

Nanni, L. (2006). Comparison among feature extraction methods for HIV-1 protease cleavage site prediction. *Pattern Recognition, 39*(4), 711–713.

Nanni, L., & Lumini, A. (in press). Machine learning multi-classifiers for peptide classification. *Neural Computing & Applications*, 11–12. doi:10.1007/s00521-007-0170-2.

Nanni, L., & Lumini, A. (2006). MppS: An ensemble of support vector machine based on multiple physicochemical properties of amino-acids. *NeuroComputing, 69*(13), 1688–1690.

Nanni, L., & Lumini, A. (2008). Genetic programming for creating Chou's pseudo amino acid based features for submitochondria localization. *Amino-Acids*. doi:10.1007/s00726-007-0018-1.

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research, 11*, 169–198.

Qian, N., & Sejnowskij, T. J. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology, 202*, 865–884.

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

Rögnvaldsson, T., You, L., & Garwicz, D. (2007). Bioinformatic approaches for modeling the substrate specificity of HIV-1 protease: An overview. *Expert Review of Molecular Diagnostics, 7*(4), 435–451.

Shen, H.-B., & Chou, K-C. (2007). Hum-mPLoc: An ensemble classifier for large-scale human protein subcellular location prediction by incorporating samples with multiple sites. *Biochemical and Biophysical Research Communications*(355), 1006–1011.

Sturniolo, T., Bono, E., Ding, J., Raddrizzani, L., Tuereci, O., Sahin, U., et al. (1999). Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nature Biotechnology, 17*(6), 555–561.

Tong, J., Liu, S., Zhou, P., Wu, B., & Li, Z. (2008). A novel descriptor of amino acids and its application in peptide QSAR. *Journal of Theoretical Biology*, 11–12. doi:10.1016/j.jtbi.2008.02.030.

Venkatarajan, M. S., & Braun, W. (2001). New quantitative descriptors of amino acids based on multidimensional scaling of a large number of physical–chemical properties. *Journal of Molecular Modeling, 7*, 445–453.

Whitaker, C. J., & Kuncheva, L. I. (2003). *Examining the relationship between majority vote accuracy and diversity in bagging and boosting*. Technical report. Bangor: School of Informatics, University of Wales.

Wu, C. H., Whitson, G., McLarty, J., Ermongkonchai, A., & Change, T. C. (1992). PROCANS: Protein classification artificial neural system. *Protein Science*, 667–677.

Zenobi, G., & Cunningham, P. (2001). Using diversity in preparing ensembles of classifiers based on different feature subsets to minimize generalization error. In L. D. Raedt, & P. A. Flach (Eds.), *Proceedings of the 12th conference on machine learning, Lecture notes in computer science* (Vol. 2167, pp. 576–587).

Zhang, G. L. et al. (2005). Neural models for predicting viral vaccine targets. *Journal of Bioinformatics and Computational Biology, 3*, 1207–1225.

Zhao, Y., Pinilla, C., Valmori, D., Roland Martin, R., & Simon, R. (2003). Application of support vector machines for T-cell epitopes prediction. *Bioinformatics, 19*(15), 1978–1984.