Artificial Intelligence in Medicine 48 (2010) 43–50

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Coding of amino acids by texture descriptors

Loris Nanni *, Alessandra Lumini

Department of Electronic, Informatics and Systems (DEIS), Università di Bologna, Via Venezia 52, 47023 Cesena, Italy

ARTICLE INFO

Article history:
Received 12 December 2008
Received in revised form 24 September 2009
Accepted 3 October 2009

Keywords:
Protein classification
Peptide classification
Vaccine development
Locally binary patterns
Discrete cosine transform
Support vector machine

ABSTRACT

Objective: In this paper we propose a new feature extractor for peptide/protein classification based on the calculation of texture descriptors. Representing a peptide/protein using a matrix descriptor, instead of a vector, allows to deal with the peptide/protein as an image and to use texture descriptors for representation purposes.

Methods and materials: A matrix descriptor, which is a squared matrix of the dimension of the peptide/protein, is obtained considering a partial ordering of the amino acids of the peptide/protein according to their value of a given physicochemical property. Each matrix descriptor is considered as a texture image and several texture descriptors are considered to obtain a compact representation which is scale invariant (i.e. independent on the length of the peptide\protein). The texture descriptors tested in this work are: local binary patterns (LBP), discrete cosine transform (DCT) and Daubechies wavelets.

Results and conclusion: The experimental section reports several tests, aimed at supporting our ideas, performed on the following datasets: vaccine dataset for the predictions of peptides that bind human leukocyte antigens; human immunodeficiency virus (HIV-1) protease cleavage site prediction dataset and membrane proteins type dataset.

The experimental results confirm the usefulness of the novel descriptors: the performance obtained by our system on the three difficult datasets is quite high, indicating that the proposed method is a feasible system for extracting information from peptides and proteins. The performance obtained by each of the three texture descriptors calculated from the matrix-based representation, and coupled to a support vector machine classifier, is lower than the performance obtained by other vector-based descriptors based on physicochemical properties proposed in the literature. Anyway the new descriptors bring different information and our tests show that the texture descriptors and the vector-based descriptors can be combined to improve the overall performance of the system.

In particular the proposed approach improves the state-of-the-art results in two out of three tested problems (HIV-1 protease cleavage site prediction dataset and membrane proteins type dataset).

1. Introduction

Several applications need to extract features from peptides/proteins for solving a given classification problem [1], some examples are: sub-cellular localization [2], protein–protein interactions [3], HIV-1 protease cleavage site prediction [4,5].

Probably, the most used feature extractor for peptides and proteins is the Chou's pseudo amino acid (PseAA) composition [6]. In the literature, several variants of these descriptors have been proposed: hydropathy scales [7,8], physicochemical distance [9], digital code [10], complexity factor [11–12], digital signal [13], Fourier low-frequency spectrum [14], cellular automata [15], "artificial" features created by genetic programming combining one or more "original" Chou's pseudo amino acid features [16]. The interested reader can refer to [17] and [18] for a survey of the different methods for extracting features from peptides and proteins.

Most of the feature extractors proposed in the literature are based on a vectorial representation of the peptide/protein. For example, in [19] a physicochemical encoding is proposed: each amino acid is represented by a 20-dimesional vector with all values set to zero except for the one corresponding to the considered amino acid, which takes the value of the measured physicochemical property. The descriptor associated to a peptide/protein is obtained by concatenating all the 20-dimesional vectors corresponding to its amino acid sequence.

Other interesting encoding methods, here reported for completeness, are based on kernels. One of the first approaches is the Fisher kernel [20] proposed for remote homology detection. A different kernel, the mismatch string kernel, is proposed in [21], which measures similarity among two sequences of amino acids

* Corresponding author. Tel.: +39 0547 339121; fax: +39 0547 338890.
E-mail addresses: loris.nanni@unibo.it (L. Nanni), alessandra.lumini@unibo.it (A. Lumini).

0933-3657/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.artmed.2009.10.001

based on shared occurrences of subsequences. In [21] it is shown that string kernels have performance similar to Fisher kernel with a lower computational cost. A class of new kernels is developed in [22] which obtain good performance for predicting protein subcellular localization: a set of kernel functions derived from k-peptide vectors mapped by a matrix of high-scored pairs, measured by BLOSUM62 scores, of k-peptides, are used for training a support vector machine. Another interesting approach is the biobasis function neural network [23], in this method the sequences are not encoded in a feature space; instead, the distances obtained by sequence alignment are used to train the neural network.

The aim of this paper is to propose a novel descriptor obtained from a matrix representation of the peptides/proteins. Analogously to many of the above cited methods the physicochemical properties are considered to discriminate among the amino acids: each descriptor, which is a squared matrix of the dimension of the peptide/protein, is obtained considering a partial ordering of the amino acids of the peptide/protein according to a given physicochemical property. A more compact representation of this matrix descriptor is obtained by considering such matrix as an image and using a texture descriptor to obtain a scale-invariant representation, independent on the length of the peptide\protein. Several well-known texture descriptors are tested in this paper: local binary pattern (LBP), which extracts a histogram that describes the difference between each matrix point and its neighborhood; discrete cosine transform (DCT); Daubechies wavelet, which performs a multi-resolution analysis of the image.

The experimental section reports several tests on the following datasets: vaccine dataset for the predictions of peptides that bind human leukocyte antigens; HIV-1 protease cleavage site prediction dataset and membrane proteins type dataset. Our results show that the proposed descriptors obtain valuable classification accuracy and can be considered for a fusion with other standard descriptors to further improve the classification performance.

The remaining of the paper is organized as follows. Section 2 briefly reviews the related works on the three applications tested in this paper; Section 3 introduces the feature extraction method proposed in this work; Section 4 reports experimental results obtained on three different classification problems; finally, Section 5 draws some conclusions.

2. Related works

2.1. HIV-1 protease cleavage site prediction

For the replication of the AIDS virus, the HIV-1 protease [24–26] is essential. The inhibitors of the protease bind the active site in HIV-1 protease and do not permit the normal functioning of the protease. In the literature, several methods for HIV-1 protease cleavage sites in proteins prediction are published, most based on machine learning systems: in [27–29] a standard feed-forward multilayer perceptron (MLP) is proposed to outperform the decision tree classifier; in [24,30] a support vector machine (SVM) classifier is tested, in particular, in [24] it is shown that HIV-1 protease cleavage is a linear problem and that the best results are obtained by linear SVM. Recently, a web-server was established for predicting HIV-1 protease cleavage sites in proteins [31].

2.2. Vaccine (predictions of peptides that bind human leukocyte antigens)

In order to design useful vaccines for a large population, it is very important to predict the peptides that bind multiple human leukocyte antigen (HLA) molecules [32]. The developing of automatic systems for predicting if a peptide binds multiple HLA molecules is very useful for making the design of vaccines more time-effective. Examples of automatic systems yet proposed in the literature (see [33] for a survey) are: a system based on SVM proposed in [34]; systems based on artificial neural networks and hidden Markov model proposed in [33] where and ensembles of classifiers tested in [18,35].

2.3. Membrane proteins type

The membrane proteins type determines the function of that protein [36,37], for this reason several methods for automated classification of membrane protein types have been proposed in the literature [7,14,38,39]. Until 2007, only small datasets without being rigorously screened by a data-culling operation to avoid redundancy were available in the literature, a bigger and more reliable dataset was collected in [2].

In [2] a system based on an ensemble of optimized evidence-theoretic *k*-nearest neighbor classifiers is proposed, where the features (the pseudo amino acid composition) are extracted considering the position-specific scoring matrix.

In [16] an ensemble of SVMs, where each classifier is trained considering a different physicochemical property and a feature extraction method based on the residue couple model, is proposed. This method partially fills the performance gap between the feature extraction methods based on the amino acid sequence and the feature extraction methods based on the position-specific scoring matrix.

3. System description

The system proposed in this work is based on a matrix representation of the peptide/protein, which is treated as an image and characterized using a texture descriptor. A SVM classifier is trained using the extracted texture features to perform the classification task. A graphical schema of the proposed system is reported in Fig. 1. The following subsections describe the main steps of the approach.

3.1. Matrix representation

The aim of this step is the characterization of a peptide/protein sequence by means of a matrix able of representing the information related to both the positions of the amino acids in the sequence and also their physicochemical properties. Then, a texture descriptor is extracted from the matrix, as described in Section 3.2, in order to obtain a descriptor that can be considered as invariant for the protein sequences.

The matrix representation of the peptide/protein is constructed by considering a selected physicochemical property of amino acids, which can be obtained by the amino acid index database[1] [40]. First, the 20 amino acids are sorted according to the value of the selected property and then a "ranking value" is assigned to each of them, which weighs the position of the amino acid in the sequence [41]. The ranking rule is the following: the first amino acid has value 1, the last has value 1/20, if there are not two amino acids with the same values, otherwise the sequence is mediate on the number of different values present in the sequence. For example, if the 20 bases are sorted in the following way, according to a given physicochemical property **P**: N < K < R < Y < F = Q < S < H < M < W < G = L < V < E < I < A < D < T < P < C, the corresponding weights are: $rank_P(N) = 1/18$, $rank_P(K) = 2/18$, $rank_P(R) = 3/18$, ..., and $rank_P(C) = 18/18 = 1$.

The ranking relationships between all the pairs of amino acids that compose the sequence of the peptide/protein are collected into a square matrix, named OM(**P**) having dimensions $l \times l$, where

[1] Available at www.genome.jp/dbget/aaindex.html (accessed 15 July 2009).