Predicting HIV Protease-Cleavable Peptides by Discrete Support Vector Machines

Carlotta Orsenigo[1] and Carlo Vercellis[2]

[1] Dip. di Scienze Economiche, Aziendali e Statistiche, Università di Milano, Italy
carlotta.orsenigo@unimi.it
[2] Dip. di Ingegneria Gestionale, Politecnico di Milano, Italy
carlo.vercellis@polimi.it

Abstract. The Human Immunodeficiency Virus (HIV) encodes an enzyme, called HIV protease, which is responsible for the generation of infectious viral particles by cleaving the virus polypeptides. Many efforts have been devoted to perform accurate predictions on the HIV-protease cleavability of peptides, in order to design efficient inhibitor drugs. Over the last decade, linear and nonlinear supervised learning methods have been extensively used to discriminate between protease-cleavable and non cleavable peptides. In this paper we consider four different proteins encoding schemes and we apply a discrete variant of linear support vector machines to predict their HIV protease-cleavable status. Empirical results indicate the effectiveness of the proposed method, that is able to classify with the highest accuracy the cleavable and non cleavable peptides contained in two publicly available benchmark datasets. Moreover, the optimal classification rules generated are characterized by a strong generalization capability, as shown by their accuracy in predicting the HIV protease cleavable status of peptides in out-of-sample datasets.

Keywords: HIV protease, cleavable peptides prediction, discrete support vector machines.

1 Introduction

The Human Immunodeficiency Virus (HIV) encodes an enzyme, called HIV protease, which is responsible for the generation of infectious viral particles. HIV protease function is to cleave virus polypeptides at defined susceptible sites. This cut gives rise to new viral proteins which are able to spread from the native cell and infect other cells. Thus, HIV protease plays a fundamental role in enabling the replication of the virus.

In molecular biology, many efforts have been devoted to investigating the HIV protease problem specificity. The interaction between the enzyme and the virus polyprotein is based upon the following paradigm, also known as "lock and key" model (Chou, 1996; Rögnvaldsson and You, 2004). The active peptide in the protein, which is generally composed by a sequence of eight amino acids around the cleavage site, must fit as a key for binding to the HIV protease

E. Marchiori, J.H. Moore, and J.C. Rajapakse (Eds.): EvoBIO 2007, LNCS 4447, pp. 197–206, 2007.

active region. This means that new infectious particles are generated if the HIV protease-cleavable site properly satisfies the substrate specificity of the active enzyme region (Chou, 1996). If the chemical combination is not verified, the bounding to the active protease site can still be accomplished but no cleavage is performed or the production of immature noninfectious viral proteins takes place.

According to this paradigm, effective HIV protease drugs can be designed with the aim of smoothing the cleavage ability of the enzyme. This task may require to identify inhibitors that, binding to the protease site, are able to prevent any further binding (competitive inhibition) or change the structure of the enzyme (non-competitive inhibition) (Narayanan et al., 2002). In both cases, the prediction of which peptides can be cleaved and which can instead act as protease inhibitors is of great importance.

The discrimination between cleavable and non cleavable sequences can be naturally formulated as a binary classification problem. For this reason, several supervised learning techniques have been applied to provide fast and accurate predictions. In order to explain the complex relationship between peptides and cleavability, many authors have resorted to nonlinear classification models; see (Rögnvaldsson and You, 2004) and the references therein. However, there are situations in which the problem of devising protease-cleavable sequences is linear in nature, and can be efficiently solved by means of linear supervised learning methods (Rögnvaldsson and You, 2004).

Recently, support vector machines (SVM) have been successfully applied to predict the HIV protease-cleavable status of two sets of peptides. The first dataset has been extensively used for the comparison of different classification approaches (Cai et al., 2002; Yang and Chou, 2004; Rögnvaldsson and You, 2004; Nanni, 2006). It contains amino acid sequences that can be cleaved or not by the protease from the Human Immunodeficiency Virus of type 1 (HIV-1 protease). The second dataset, used in (Poorman et al., 1991; Chou, 1996), contains sequences known cleavable by the enzyme from the Human Immunodeficiency Virus of type 2 (HIV-2 protease).

In this paper, we perform the classification of these two datasets by means of an alternative method based on *discrete support vector machines* (DSVM). By this term we denote SVM in which the empirical classification error is represented by a discrete function, called *misclassification rate*, counting the number of misclassified examples, in place of a proxy of the misclassification distance considered by traditional SVM approaches (Orsenigo and Vercellis, 2003, 2004). The inclusion of the discrete term leads to the formulation of a mixed-integer programming problem, whose objective function is composed by the weighted sum of three terms, expressing a trade-off between accuracy and potential of generalization.

The empirical results obtained on the benchmark datasets show the effectiveness of the proposed method that, compared to traditional support vector machines, is able to discriminate between cleavable and non cleavable peptides with the highest accuracy. Moreover, the optimal classification rules generated