Bio-kernel Self-organizing Map for HIV Drug Resistance Classification

Zheng Rong Yang and Natasha Young

Department of Computer Science, University of Exeter,
Exeter EX4 4QF, UK
z.r.yang@ex.ac.uk
http://www.dcs.ex.ac.uk/~zryang

Abstract. Kernel self-organizing map has been recently studied by Fyfe and his colleagues [1]. This paper investigates the use of a novel bio-kernel function for the kernel self-organizing map. For verification, the application of the proposed new kernel self-organizing map to HIV drug resistance classification using mutation patterns in protease sequences is presented. The original self-organizing map together with the distributed encoding method was compared. It has been found that the use of the kernel self-organizing map with the novel bio-kernel function leads to better classification and faster convergence rate…

1 Introduction

In analysing molecular sequences, we need to select a proper feature extraction which can convert the non-numerical attributes in sequences to numerical features prior to using a machine learning algorithm. Suppose we denote by \( x \) a sequence and \( \phi(x) \) a feature extraction function, the mapping using a feature extraction function is \( F : (\mathbb{F} \rightarrow \phi) \in \mathbb{R}^d \). Finding an appropriate feature extraction approach is a non-trivial task.

It is known that each protein sequence is an ordered list of 20 amino acids while a DNA sequence is an ordered list of four nucleic acids. Both amino acids and nucleic acids are non-numerical attributes. In order to analyze molecular sequences, these non-numerical attributes must be converted to numerical attributes through a feature extraction process for using a machine learning algorithm. The distributed encoding method [2] was proposed in 1988 for extracting features for molecular sequences. The principle is to find orthogonal binary vectors to represent amino (nucleic) acids. With this method, amino acid Alanine is represented by 0000000000 0000000001 while Cystine 0000000000 0000000010, etc. With the introduction of this feature extraction method, the application of machine learning algorithms to bioinformatics has been very successful. For instance, this method has been applied to the prediction of protease cleavage sites [3], signal peptide cleavage sites [4], linkage sites in glycoproteins [5], enzyme active sites [6], phosphorylation sites [7] and water active sites [8].

However, as indicated in the earlier work [9], [10], [11] such a method has its inherent limit in two aspects. First, the dimension of an input space has been enlarged 20 times weakening the significance of a set of training data. Second, the biological
content in a molecule sequence may not be efficiently coded. This is because the
similarity between any pair of different amino (nucleic) acids varies while the dis-
tance between such encoded orthogonal vectors of two different amino (nucleic) acids
is fixed.

The second method for extracting features from protein sequences is to calculate
the frequency. It has been used for the prediction of membrane protein types [12], the
prediction of protein structural classes [13], subcellular location prediction [14] and
the prediction of secondary structures [15]. However, the method ignores the coupling
effects among the neighbouring residues in sequences leading to potential bias in
modelling. Therefore, di-peptides method was proposed where the frequency of each
pair of amino acids occurred as neighbouring residues is counted and is regarded as a
feature. Dipeptides, gapped (up to two gaps) transitions and the occurrence of some
motifs as additive numerical attributes were used for the prediction of subcellular
locations [16] and gene identification [17]. Descriptors were also used, for instance, to
predict multi-class protein folds [18], to classify proteins [19] and to recognise rRNA-
, RNA-, and DNA-binding proteins [20], [21]. Taking into account the high order
interaction among the residues, multi-peptides can also be used. It can be seen that
there are 400 di-peptides, 8,000 tri-peptides and 16,000 tetra-peptides. Such a feature
space can be therefore computational impractical for modelling.

The third class of methods is using profile measurement. A profile of a sequence
can be generated by subjecting it to a homology alignment method or Hidden Markov
Models (HMMs) [22], [23], [24], [25].

It can be seen that either finding an appropriate approach to define $\phi(x)$ is difficult
or the defined approach may lead to a very large dimension, i.e., $d \rightarrow \infty$. If an ap-
proach which can quantify the distance or similarity between two molecular se-
quenches is available, an alternative learning method can be proposed to avoid the
difficulty in searching for a proper and efficient feature extraction method. This
means that we can define a reference system to quantify the distance among the mo-
lecular sequences. With such a reference system, all the sequences are quantitatively
featured by measuring the distance or similarity with the reference sequences.

One of the important issues in using machine learning algorithms for analysing
molecular sequences is investigating sequence distribution or visualising sequence
space. Self-organizing map [26] has been one of the most important machine learning
algorithms for this purpose. For instance, SOM has been employed to identify motifs
and families in the context of unsupervised learning [27], [28], [29], [30], [31]. SOM
has also been used for partitioning gene data [32]. In these applications, feature ex-
traction methods like the distributed encoding method were used.

In order to enable SOM to deal with complicated applications where feature extrac-
tion is difficult, kernel method has been introduced recently by Fyfe and his col-
leagues [1]. Kernel methods were firstly used in cluster analysis for K-means algo-
rithms [33], where the Euclidean distance between an input vector $x$ and a mean vec-
tor $m$ is minimized in a feature space spanned by kernels. In the kernel feature space,
both $x$ and $m$ were the expansion on the training data. Fyfe and his colleagues devel-
oped so-called kernel self-organizing maps [34], [35]. This paper aims to introduce a
bio-kernel function for kernel SOM. The method is verified on HIV drug resistance
classification. A stochastic learning process is used with a regularization term.