Chapter 5

Molecular Similarity: Combining Neural Networks and KNN Methods

Abdelmalek Amine
Tahar Moulay University & Djillali Liabes University, Algeria

Zakaria Elberrichi
Djillali Liabes University, Algeria

Michel Simonet
Joseph Fourier University, France

Ali Rahmouni
Tahar Moulay University, Algeria

ABSTRACT

In order to identify new molecules susceptible to become medicines, the pharmaceutical research has more and more resort to new technologies to synthesize big number of molecules simultaneously and to test their actions on given therapeutic target. This data can be exploited to construct the models permitting to predict the properties of molecules not yet tested, even not yet synthesized. Such predictive models are very important because they make it possible to suggest the synthesis of new molecules, and to eliminate very early in the the molecule’s search process the molecules whose properties would prevent their use as medicine. The authors call it virtual sifting. It is within this framework that research by similarity is registered. It is a practical approach to identify molecules candidates (to become medicines) from the data bases or the virtual chemical libraries by comparing the compounds two by two. Many statistical models and learning tools have been developed to correlate the molecule’s structure with their chemical, physical or biological properties. The large majority of these methods start by transforming each molecule in a vector of great dimension (using molecular descriptors), then use a learning algorithm on these vectorial descriptions. The objective of this chapter is to study molecular similarity using a particular type of neural networks: the Kohonen networks (also called “SOM” Self-Organizing Maps), applying the nearest neighbor algorithm to the projection of the molecules (coordinates) in the constructed MAP.
INTRODUCTION

Research by molecular similarity is nowadays a very important tool in various processes of conception of products having particular properties such as medicines (van de Waterbeemd & Gifford, 2003). This tool is one of the virtual sifting methods that permits to reach the searched molecule in a very short time and to lower the costs of these processes.

The virtual sifting methods developed profit from the availability of both the increasing computer’s powerfulness and the sophisticated algorithms. The method compares some features called descriptors of the molecule studied with those of molecules of one or several data bases (Mozziconacci, 2003; Willett, 1998). This operation permits the elimination of an important number of molecules that doesn’t present the sought properties. The experimental tests will only be done on a reduced number of molecules that research by similarity would have given like potential candidates.

Many statistical models and learning tools can correlate the structures of the molecules with their chemical, physical or biological properties (Mahé & Vert, 2007). The vast majority of these methods represents the molecules by vectors of high dimension whose components are the molecular descriptors. They use an algorithm combined to a chosen measure to determine the molecules having descriptors close to those of the studied molecule.

Among these methods we can mention the linear models as the linear regression and the multiple linear regression, and the nonlinear models as the artificial neuron networks, the genetic algorithms and the SVM (support vector machine) (Mahé & Vert, 2007).

In this chapter we use a particular type of neural networks; the Kohonen networks known also SOM (Self Organizing Maps). This type of networks permits to reduce the dimensionality of the molecule’s representation and to classify the molecules by groups of similar molecules. The nearest neighbor algorithm is applied then on the bidimensional map generated after the first phase to search for the molecules most similar to a given molecule.

After this introduction, section 2 presents some general notions on the similarity and the molecular similarity in particular. Some methods are described there as well as similarity measures and distance. Section 3 presents the principles of similarity based on descriptors. Section 4 is dedicated to molecular descriptors. Section 5 introduces artificial neuron networks and nearest neighbor methods applied in the used approach. Section 6 presents the tests of the developed application. Some results are presented there. A conclusion will put an end to this chapter.

SIMILARITY

Functions of similarity are used in many fields, in particular in Data Analysis, Form Recognitions, Symbolic Machine Learning, and Cognitive Sciences.

In a general way, a function of similarity is defined in a universe $U$ that can be modelled using a quadruplet: $(L_d, L_s, T, F_S)$.

- $L_d$ is the language of representation used to describe the data.
- $L_s$ is the language of representation of the similarities.
- $T$ is a set of knowledge that we possess on the studied universe.
- $F_S$ is the binary function of similarity, such as: $F_S: L_d \times L_d \rightarrow L_s$

When, the function of similarity has for object to quantify the resemblances between the data, the $L_s$ language corresponds to the set of the values in the interval $[0...1]$ or in the $R^+$ set and we will speak then of similarity measurement (Bisson, 2000).
Related Content

Symbolic Equation for the Instantaneous Amount of Substance in Linear Compartmental Systems: Software Furnishing the Coefficients Involved in it
www.igi-global.com/chapter/symbolic-equation-instantaneous-amount-substance/56463?camid=4v1a

Modeling of Fluid Interaction Produced by Water Hammer
www.igi-global.com/chapter/modeling-fluid-interaction-produced-water/77067?camid=4v1a

Prediction of Compound-Protein Interactions with Machine Learning Methods
www.igi-global.com/chapter/prediction-compound-protein-interactions-machine/45477?camid=4v1a

Application of Machine Leaning in Drug Discovery and Development
www.igi-global.com/chapter/application-machine-leaning-drug-discovery/45473?camid=4v1a