Chapter 113
A Hybrid Approach Based on Self-Organizing Neural Networks and the K-Nearest Neighbors Method to Study Molecular Similarity

Abdelmalek Amine
Tahar Moulay University, Algeria

Zakaria Elberrichi
Djillali Liabes University, Algeria

Michel Simonet
Joseph Fourier University, France

Ali Rahmouni
Tahar Moulay University, Algeria

ABSTRACT

The "Molecular Similarity Principle" states that structurally similar molecules tend to have similar properties—physicochemical and biological. The question then is how to define "structural similarity" algorithmically and confirm its usefulness. Within this framework, research by similarity is registered, which is a practical approach to identify molecule candidates (to become drugs or medicines) from databases or virtual chemical libraries by comparing the compounds two by two. Many statistical models and learning tools have been developed to correlate the molecules' structure with their chemical, physical or biological properties. The role of data mining in chemistry is to evaluate "hidden" information in a set of chemical data. Each molecule is represented by a vector of great dimension (using molecular descriptors), the applying a learning algorithm on these vectors. In this paper, the authors study the molecular similarity using a hybrid approach based on Self-Organizing Neural Networks and Knn Method.

DOI: 10.4018/978-1-4666-2455-9.ch113

INTRODUCTION

Introduction of a new drug into the market is often the culmination of a long and arduous process of laboratory experimentation. This process, from hit to lead to marketable drug, is typically as long as 5-10 years. In order to identify new molecules susceptible to become medicines, the pharmaceutical research has more and more resort to technologies permitting to synthesize a very big number of molecules simultaneously and to test their actions on a given therapeutic target. These data can be exploited to construct the models permitting to predict the properties of molecules not yet tested, even not yet synthesized. Looking for molecular similarity is an intelligent way to design drug. Its use is based on the principle that structurally more similar molecules are more likely to exhibit similar properties than structurally less similar molecules (Monev, 2004; Johnson & Maggiora, 1990). 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 molecule's search process the molecules whose properties would prevent their use as medicine. We speak then of virtual sifting.

Hence, searching for functionally similar molecules, which is very important in drug design, can be accomplished by searching for structurally similar molecules (van de Waterbeemd & Gifford, 2003). But the problem is to define molecular similarity.

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: *(Ld, Ls, T, FS)*.

- *Ld* is the language of representation used to describe the data.
- *Ls* is the language of representation of the similarities.
- *T* is a set of knowledge that we possess on the studied universe.
- *FS* is the binary function of similarity, such as: $FS: Ld \ x \ Ld \rightarrow Ls$

When, the function of similarity has for object to quantify the resemblances between the data, the Ls 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).

Most works concerning the similarity measures have as base the mathematical concept of distance (the inverse notion of similarity) which was well studied in DA (Mahé & Vert, 2007; Bisson, 2000).

It is defined in the following way: let Ω the set of the individuals of the studied domain a metric D which is a function of $\Omega \ X \ \Omega$ in $R+$, $\forall \ \boldsymbol{a, b, c} \in \Omega$.

1. $D \ \boldsymbol{(a, a) = 0}$ (property of minimality)
2. $D \ \boldsymbol{(a, b) = D \ (b, a)}$ (property of symmetry)

When the function D verifies the properties 1 and 2, it is called index of dissimilarity (or more simply a dissimilarity).

The other properties are also interesting:

3. $D \ \boldsymbol{(a, b) = 0 \Rightarrow a = b}$ (property of identity)
4. $D \ \boldsymbol{(a, c)} \leq D \ \boldsymbol{(a, b)} + D \ \boldsymbol{(b, c)}$ (triangular inequality)
5. $D \ \boldsymbol{(a, c) \leq Max} \ [D \ \boldsymbol{(a, b)}, D \ \boldsymbol{(b, c)]}$

If the function D verifies the properties 1, 2 and 3 we speak of a distance index. If this index also verifies the property 4 we call it a distance and if it also verifies the property 5 it is called a ultrametric distance. In addition, when the function D verifies properties 1, 2 and 4 we speak of a variation (a gap), and when it verifies properties 1, 2 and 5 we speak of a variation (gap) ultrametric.