Journal of Molecular Graphics and Modelling 60 (2015) 55–62

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Predicting activity approach based on new atoms similarity kernel function

Ahmed H. Abu El-Atta[a,b,*], M.I. Moussa[b], Aboul Ella Hassanien[a,c]

[a] Scientific Research Group in Egypt (SRGE)[1], Egypt
[b] Faculty of Computers and Information, Benha University, Benha, Egypt
[c] Faculty of Computers and Information, Cairo University, Egypt

ARTICLE INFO

Article history:
Received 24 December 2014
Received in revised form 28 March 2015
Accepted 26 May 2015
Available online 10 June 2015

Keywords:
Chemoinformatics
Graph kernel
Machine learning
Activity prediction
Drug discovery

ABSTRACT

Drug design is a high cost and long term process. To reduce time and costs for drugs discoveries, new techniques are needed. Chemoinformatics field implements the informational techniques and computer science like machine learning and graph theory to discover the chemical compounds properties, such as toxicity or biological activity. This is done through analyzing their molecular structure (molecular graph). To overcome this problem there is an increasing need for algorithms to analyze and classify graph data to predict the activity of molecules. Kernels methods provide a powerful framework which combines machine learning with graph theory techniques. These kernels methods have led to impressive performance results in many several chemoinformatics problems like biological activity prediction. This paper presents a new approach based on kernel functions to solve activity prediction problem for chemical compounds. First we encode all atoms depending on their neighbors then we use these codes to find a relationship between those atoms each other. Then we use relation between different atoms to find similarity between chemical compounds. The proposed approach was compared with many other classification methods and the results show competitive accuracy with these methods.

1. Introduction

Chemoinformatics (chemical informatics) is the field that seeks to use informational techniques like computer science, mathematics and information techniques to predict or analyze molecule's (chemical compounds) properties. One of the major principles in this research field is the similarity principle, which states that two structurally similar molecules should have similar activities and properties.

Graphs are flexible models that have been used to present data in many scientific, engineering, and business fields. For example, in finance data analysis, graphs are used to model dynamic stock price changes [1]. To analyze biological data, graphs have been utilized in modeling chemical structures [2], protein sequences [3], protein structures [4], and gene regulation networks [5]. The structure of a molecule is encoded by a labeled graph $G=(V, E, \mu, \pi)$, where the unlabeled graph (V, E) encodes the structure of the molecule while μ maps each vertex to an atom's label and π maps each edge to a type of bond between two atoms (single, double, triple or aromatic).

Presenting chemical compounds by graphs enables us to apply graph classification techniques to predict chemical compounds properties. In graph classification, each graph is associated with a target value and the aim is to find a good function that maps graphs to their target values. The existing algorithms of classifying graph data can be divided into three categories [6,7].

The first approach is introduced within the quantitative structure activity relationship (QSAR) field which is based on finding the correlation between molecule's descriptors and molecule's properties [8]. Vectors of molecular descriptors (MDs) may be defined from structural information [9] besides physical properties or biological activities. Molecular descriptors may be classified into three categories. The first is simple one-dimensional (1D) descriptors which represent bulk properties of compounds, the second is two-dimensional (2D) descriptors such as topological and charge indices, and the last is complex three-dimensional (3D) descriptors which often rely on 3D representation and conformational aspects of a molecule [10]. Molecular descriptors may be used within any statistical machine learning algorithm to predict molecule's

* Corresponding author at: Faculty of Computers and Information, Benha University, Benha, Egypt.

E-mail addresses: ahmed.aboalatah@fci.bu.edu.eg (A.H. Abu El-Atta), mahmoud.mossa@fci.bu.edu.eg (M.I. Moussa), abo@egyptscience.net (A.E. Hassanien).

[1] egyptscience.net.

http://dx.doi.org/10.1016/j.jmgm.2015.05.014
1093-3263/© 2015 Elsevier Inc. All rights reserved.

properties. Such a scheme allows benefiting from the large set of tools available within the statistical machine learning framework.

Another approach is to explicitly collect a set of features from the graphs. Features may be chosen from paths, cycles, trees, and subgraphs. Once a set of features is determined, a graph is described by a feature vector. With a collection of vectorized graph data, any existing data mining method that works in n-dimensional Euclidean space may be applied to do graph classification. In the context of chemoinformatics, explicit pattern features for compounds are often known as structural keys. A structural key is a bit string denoting presence of certain patterns (such as paths, cycles, trees, etc.) of interest [11].

The third approach of graph classification is to implicitly collect a set of features (possibly an infinite number of such features) and compute the similarity of two graphs via a kernel function. The term kernel function refers to an operation of computing the inner product between two points in a Hilbert space, so this may lead to avoiding the explicit computation of coordinates in that feature space. Graph kernel functions are simply kernel functions that have been defined to compute the inner product between two graphs.

Many of graph kernel functions have been developed, with promising application results as described in [12]. Among these methods, some kernel functions draw on graph features such as walks [13] or cycles [14], while others may use different approaches such as genetic algorithms [15], frequent subgraphs [16], or graph alignment [17]. In [18], a local subgraph is used to encode the stereo-isomerism property of each atom of a molecule. A kernel between bags of such subgraphs provides a similarity measure incorporating stereo-isomerism properties.

In this paper, we present an approach based on the similarity principle to classify chemical compounds. Chemical compounds with similar properties have similar structure and compounds with similar structure contain common subgraphs. How an atom is connected to its neighbors, plays a great role to determine the properties of that atom. Based on this information we design an approach in which chemical graphs are not explicitly factored into patterns but only the count of these patterns is used.

The proposed approach starts by collecting all atoms in a given dataset then encoding them by unique codes. We use atoms codes to find similarity relationship between atoms by counting the common subgraphs between each two atoms and their neighbors. After that, the relationship between atoms is used to create a similarity vectors between compounds each other. Finally the kernel function is applied to similarity vectors to predict the activity of compounds. The rest of this paper is organized as the following in Section 2, we discuss the proposed approach and state its steps. In Section 3, we describe the atom coding system and how to employ it to find similarity between atoms. In Section 4, we describe how to compute similarity between chemical compounds and the kernel function. In Section 5, we present our results and performance evaluation. Conclusion to our work is presented in Section 6.

2. Basic steps of the prediction approach

Activity of an atom in a chemical compound depends on the bonds which connect this atom with its neighbors. We present coding system to represent each atom. This coding system preserves the atom type, types of all adjacent atoms and bonds types between that atom and its neighbors. As Fig. 1 shows, we firstly use this coding system to encode all atoms in all chemical compounds of a given dataset. This coding system enables us to sort all atoms according to that code to create an index for them.

Moreover, from this code we construct a relation between these atoms each other. Finally, the relation matrix between atoms is used to find a similarity matrix between chemical compounds each other. Rows of this similarity matrix represent the features vectors for each compound. These features vectors may be directly passed to the kernel function to create the kernel matrix for dataset or a features selection based on ranking is done before using kernel function. Steps of finding atoms similarity are described in the following section.

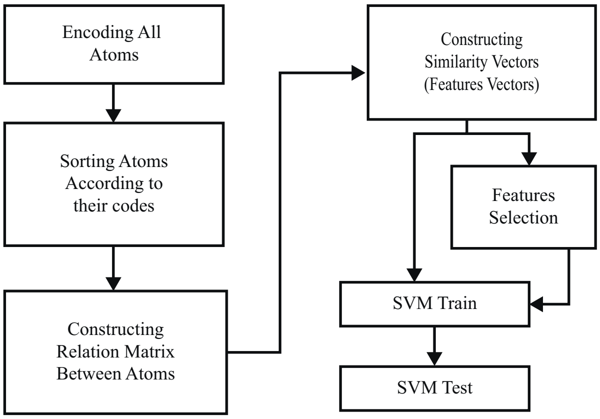
Encoding All Atoms
Sorting Atoms According to their codes
Constructing Relation Matrix Between Atoms
Constructing Similarity Vectors (Features Vectors)
Features Selection
SVM Train
SVM Test

Fig. 1. Steps of the prediction approach.

3. Atoms similarity

3.1. Atoms coding system

Any composite number is the result of multiplication of at least two prime numbers and any prime number is only divisible by one or itself. From the previous facts we build a coding system depending on prime numbers to represent each atom and its incident bonds and adjacent atoms. Prime numbers were used previously in many algorithms. An algorithm proposed by Weininger et al. [19] uses prime numbers to improve performance of Morgan algorithm [20] and to create canonical SMILES. Morgan's algorithm proceeds in two steps. In the first step, each atom is labeled by the number of neighbors of that atom (atom degree). In the second step, atoms labels are calculated in a recursive manner. At each recursive the label of any given atom is the sum of the labels of its neighbors computed in the previous step. The main criticism of Morgan's algorithm is the ambiguity of the summation when computing atom labels.

To overcome this problem, Weininger et al. [19] proposed a solution using prime numbers. In this implementation the initial labels are substituted by primes, that is, degree 1 is replaced by 2, degree 2 by 3, degree 3 by 5, and so on. Next, instead of summing the labels of the neighbors, the product of the primes is computed. According to the prime factorization theorem, the solution of Weininger et al. [19] is unambiguous. Also, prime numbers are used as labels for bonds to differ between bonds types. Prime ID number is a modification of the Randić connectivity ID number, which aimed to improving the discriminating power of Randić connectivity ID number. In practice, prime ID numbers are calculated by substituting the edge connectivity of the connectivity ID number with a different edge weight based on the first nine prime numbers [21].

The proposed coding system combines the benefits of using prime numbers as labels for atoms and bonds. Fig. 2, shows two atoms and their neighbors. The one on the left is carbon atom placed in the center and it has two single bonds with two other atoms, another carbon atom and an oxygen atom, also it has double bond with another oxygen atom. On the right a carbon atom placed in the center and it has two single bonds with two other atoms, another carbon atom and an oxygen atom, also it has an aromatic bond with another carbon atom. These atoms and their neighbors represent a