Chemometrics and Intelligent Laboratory Systems 130 (2014) 151–158

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Artificial neural network modeling of antimycobacterial chemical space to introduce efficient descriptors employed for drug design

Soroush Sardari [a], Houshmand Kohanzad [c], Ghazaleh Ghavami [a,b,*]

[a] Drug Design and Bioinformatics Unit, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute, #69, Pasteur Avenue, Tehran 13164, Iran
[b] Eastern Mediterranean Health Genomics and Biotechnology Network (EMGEN), Pasteur Institute, #69, Pasteur Avenue, Tehran 13164, Iran
[c] Deputy of Research, Ministry of Health and Medical Education, Tehran, Iran

ARTICLE INFO

Article history:
Received 2 January 2013
Received in revised form 7 September 2013
Accepted 14 September 2013
Available online 16 October 2013

Keywords:
Computational biology
ANN modeling
Chemical informatics
Anti-infective
Anti-tubercular agents
Antimycobacterial
Descriptors
Drug design
Drug Modeling

ABSTRACT

Tuberculosis has become a serious condition with an estimated 2 million deaths each year in the world. According to WHO report, multi-resistant tuberculosis is responsible for approximately 460 thousand recent cases per year and for about 740 thousand patients infected by both *Mycobacterium tuberculosis* and HIV/AIDS. In the current study, several bioactive structure databases were analyzed using cheminformatics tools to correlate the chemical structures of different compounds with their pharmacological activities; in addition, these tools were tried to identify molecules that could be candidate for experimental assays. In this regard, for defining the effective chemical compounds against *Mycobacterium*, a database consisting of 400 antimycobacterial compounds has been constructed. In the next step, more than 1400 molecular descriptors were defined by DRAGON application server for each compound. Then, the resulting descriptors were clustered by kNN and k-means clustering methods to be employed for ANN modeling. Utilizing PLS and ANN modeling methods led to building a model for predicting minimum inhibitory concentration (MIC) with $R^2 = 0.98$ and MSE = 0.0002. Applying the mentioned cheminformatics tools, it would be possible to design and introduce new compounds with broad applications in antimycobacterial drug discovery and development.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

The genus *Mycobacterium* related to Mycobacteriaceae family is defined as one of the most diverse bacterial groups with 85 identified species. Various species of this bacterial genus is widely spread in nature and some of its species have been known to be pathogenic for human and animals, causing widespread diseases such as the primary pulmonary disease, tuberculosis, and two skin diseases, leprosy and buruli ulcer [1].

Tuberculosis (TB) is one of the oldest diseases in the medicinal history maintaining to be a major disease of worldwide scale, infecting at least one third (two billion) of the world's population. It is mainly caused by *Mycobacterium tuberculosis* (MTB), and to a lesser degree *Mycobacterium bovis* and *Mycobacterium africanum*. Based on World Health Organization (WHO) report in 2006, it was determined that for approximately 9 million new cases of TB, there were about 2 million cases that led to death annually [2].

Based on the mentioned aspects regarding a wide range of TB impact on global health, it is obvious that designing and discovering novel efficient TB drugs with minimum side effects could be defined as a main goal of relevant researchers around the word. There is no doubt that one of the key steps in TB drug discovery is investigating molecular and cellular mechanisms of action of MTB in host cell as well as molecular response of infected host cells to help in finding novel suitable targets for TB drug design procedures. The process of cell wall metabolism as well as DNA replication or protein synthesis could be defined as one of the hot spots for most existing TB drugs. Many compounds undergoing preclinical and clinical development target other key enzymes required for survival or the bacterial cell membrane [3]. The value of any novel target is known not only in the context of its essentiality for survival *in vitro*, but also for a range of properties related to the drug discovery procedures such as selectivity, suitability for structural studies, and ability to monitor inhibitory effects of candidate drugs against MTB in the entire cells [3].

Indeed, the appearance of multi-drug resistant MTB strains causing multidrug resistant phenomena against antimycobacterial compounds (MDR-TB), the growing rate of TB frequency, the lethal combination caused by HIV co-infection, and the lack of any novel effective antimycobacterial compounds in the last 40 years, all could be considered as the most considerable challenges regarding TB treatment and emphasize a critical need for the development of new TB therapies, especially, novel lead structures that are necessary with new modes of action [2].

With the development of relatively time-saving, shorter incubation period assay methods with lower risk of contamination, the recent decades have witnessed a surge in the amount of compounds

* Corresponding author at: Drug Design and Bioinformatics Unit, Department of Medical Biotechnology, Biotechnology Research Center, Pasteur Institute, #69, Pasteur Avenue, Tehran 13164, Iran. Tel./fax: +98 21 66954324.
E-mail address: ghghavami@gmail.com (G. Ghavami).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.09.011

that have been added into the scientific literature as potential anti-mycobacterial agents [2]. However, it obvious that drug discovery and development based on current protocols are complex and expensive defining a successful direction mostly depends on spending years of research and many resources to start from an initial disease treatment concept to a new drug application (NDA). The cost of introducing a novel drug to market is estimated to be between $800 million and $1 billion [3]. Regarding decreasing the cost, time, and failure rate in delivering novel medications to the marketplace, greater effort is being developed recently to construct better predictors of safety and pharmacokinetic properties at earlier stages of the discovery and exploratory development process [3].

It may be, indeed, true that the introduction and development of combinatorial chemistry and high-throughput screening can revolute drug discovery by permitting great number of chemical agents to be synthesized and screened in less time as well as increase the rate of success in several stages of drug discovery procedures toward the drug marketplace. Advanced computational evolutionary analysis techniques combined with the rising accessibility of sequence data can facilitate the application of systematic approaches related to targets and pathways for drug discovery concepts. Cheminformatics based methodologies, particularly similarity-based virtual screening method could be applicable in drug discovery process to analyze related data from many different sources, in addition to classifying and summarizing their relationships identified. Similarity-based virtual screening is the mostly used term to describe a diversity of computational methods estimating biologically relevant properties of compounds. Most common involvement of similarity-based virtual screening is in prediction of biological activity. One of the main techniques applied in similarity-based virtual screening is Artificial Neural Networks (ANNs) with perceived ability to mimic activities of the human brain, albeit in a simplistic way [4]. In the current study, according to *in silico* methodologies, a database containing 400 antimycobacterial compounds has been made; following that, 1400 descriptors were defined and classified to construct and introduce a predictive model, which can play a key role for designing efficient compounds against *Mycobacterium* sp.

2. Materials and methods

2.1. Constructing database

A global literature search within published articles in valid scientific databases such as ScienceDirect, PubMed Central, BioMed Central, MEDLINE/PubMed, Hindawi Publishing Corporation, Springer and Wiley-Blackwell was performed to construct a data set including 400 synthetic and natural antimycobacterial compounds with known minimum inhibitory concentration (MIC) values against *Mycobacterium* sp. (especially for *M. tuberculosis*) and wherein relevant information of mentioned compounds was published during the years 1986–2010 [2,3,5–137]. In some cases, derivatives of chemical compounds with the same core and various additional groups were reported in the articles/databases. Therefore, to maintain the diversity of our database selection of compounds among each family was performed based on significant bioactivity within each group. All compounds were processed by the ChemDraw Ultra (9.0—2005) to generate their SMILES codes and minimize 3D structures.

2.2. Calculating and clustering descriptors

Recently, a wide range of physico-chemical properties, from purely empirical to quantum–mechanical could be calculated by available applications. Such diversity of computable descriptors in combination with plentiful efficient statistical and machine learning techniques could permit the design of applicable models. In the current study, there were a set of 1497 descriptors calculated using DRAGON (version 3.0—2003) based on stored SDF format of chemical molecules as the input set.

In the next step, entire computed descriptors were clustered by kNN and k-means clustering methods (MATLAB 6.5—2002). This function performs a cluster analysis using the k Nearest Neighbor (kNN) and k-means clustering algorithm and plots a dendrogram. Inputs are the data matrix (dat), and an optional matrix of sample labels (labels). The output of the function is a dendrogram showing the distances between the samples. It is notable that kNN is a method for classifying objects based on closest training examples in the feature space. kNN is defined as a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. kNN algorithm is among the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). On the subject of k-means clustering, in data mining, this method is one of the cluster analysis methods which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. This results in a partitioning of the data space into Voronoi cells. Based on mentioned summary regarding the two employed clustering methods, whole 1497 descriptors were clustered and filtered based on their difference in distances calculated by machine learning algorithms to introduce a set of 16 descriptors with minimum correlation which may be defined as the most competent atomic and molecular properties for modeling antimycobacterial compounds. In addition, mean difference among finalized descriptors was calculated by repeated measures of ANOVA (GraphPad Prism 5.01—2007) to confirm efficient diversity of clustered descriptors ($p < 0.05$ was considered significant statistically).

2.3. Modeling antimycobacterial compounds

At the core of our approach is the application of computational technologies and adoption of *in silico* methods to help similarity-based virtual screening of compounds with desired biochemical properties before these compounds are tested, acquired or even made [4]. ANN is a mathematical model or computational model on the basis of biological neural networks defined as one of the fruitful methods with broad range of applications in similarity-based virtual screening. In the current study, ANN and partial least square (PLS) methods, defined as the favorable method in realizing the existing relations and in analyzing the data, in the step-by-step manner to structure the computational model for predicting MICs, were employed. As mentioned, kNN and k-means cluster analysis were employed to cluster entire 1497 descriptors. Indeed, analyzed clustering findings were utilized to filter and select descriptors based on their differentiation in evaluated distances in the frame of dendrograms. In principle, for each compound clustered 16 descriptors with minimum correlation in calculated distances as the paragons of their distinction in physicochemical properties were utilized as the input set for PLS NIPALS algorithm (MATLAB 6.5—2002), following that the calculated output set of PLS as the matrix [400 * 16] was obtained for the final input set of ANN (MATLAB 6.5—2002).

To construct the model based on the ANN method, feed-forward network was trained to perform a nonlinear regression between clustered descriptors and MICs. After, normalizing the inputs and response, the data was divided into training, validation and test sets. It is notable that the testing set started with the second point and took every fourth point. The validation set started with the fourth point and took every fourth point. The training set took the remaining points. This method of classifying data was utilized to create a feed forward network with 5 hidden neurons, 1 output neurons, TANSIG hidden neurons and linear output neurons. Furthermore, the Levenberg–Marquardt training function TRAINLM assigned as well as the NEWFF command initialized the weights in the network.

Download English Version:

https://daneshyari.com/en/article/1180678

Download Persian Version:

https://daneshyari.com/article/1180678

Daneshyari.com