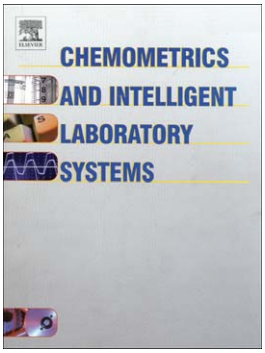

Performance of knowledge based biological models in higher dimensional chemical space

A.P. Kavitha, U.C. Abdul Jaleel, V.M. Abdul Mujeeb, K. Muraleedharan

PII: S0169-7439(16)30031-4
DOI: doi: 10.1016/j.chemolab.2016.02.009
Reference: CHEMOM 3185

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received date: 1 November 2015
Revised date: 2 January 2016
Accepted date: 22 February 2016

Please cite this article as: A.P. Kavitha, U.C. Abdul Jaleel, V.M. Abdul Mujeeb, K. Muraleedharan, Performance of knowledge based biological models in higher dimensional chemical space, *Chemometrics and Intelligent Laboratory Systems* (2016), doi: 10.1016/j.chemolab.2016.02.009

Performance of knowledge based biological models in higher dimensional chemical space

A. P. Kavitha[a], U. C. Abdul Jaleel[b], V. M. Abdul Mujeeb[a], K. Muraleedharan[a*]

[a] *Department of Chemistry, University of Calicut, Malappuram 673635, India*

[b] *Cheminformatics, Indian Institute of Science, Bangalore 560012, India*

* kmuralika@gmail.com; Tel: +91 494 2407413; Fax: +91 494 2400269

Abstract

This study evaluates the improvement of the knowledge based biological models by incorporating additional advanced molecular descriptors to the existing classical descriptors. It was found that the inclusion of constitutional, topological and hybrid descriptors in the generation of biological models trained on Mtb (Mycobacterium Tuberculosis) bioassay dataset using classifiers like Random Forest, J48, Naive Bayes and SMO (Sequential Minimal Optimization) have found to enhance the performance of these models.

Keywords: Artificial intelligence; Data mining; Machine learning; Predictive modelling; Molecular descriptors; Anti-tubercular activity.

1. Introduction

The inclusion of vivid perspectives like electronic, graph theoretical approach, etc., provides an enhanced multidimensional space analysis of the physico-chemical properties. This in turn can helps to determine the biological activity of the molecule by generating *in silico* predictive models based on molecular descriptors derived from different representations of the molecule. For this, data mining and machine learning tools and techniques can be employed [1–6]. These models can be applied to virtually screen a set of molecules for their specific chemical property or bioactivity. Such computational models also provide structural insight to design new compounds with enhanced reactivity or bioactivity by utilizing predictive data mining models on high throughput biological whole-cell screening dataset using various combinations of pharmacological, classical and graph theoretical descriptors. The major objective of this work was to reveal the relevance of such descriptors in predictive modeling and to explore the statistical significance of enhanced descriptor space upon the performance and predictive power of such models.