Applied Soft Computing 10 (2010) 170–182

Contents lists available at ScienceDirect

Applied Soft Computing

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

Genetic programming for QSAR investigation of docking energy

Francesco Archetti [a,b], Ilaria Giordani [a,c], Leonardo Vanneschi [a,*]

[1] Dipartimento di Informatica, Sistemistica e Comunicazione (D.I.S.Co.), University of Milano-Bicocca, via Bicocca degli Arcimboldi, 8, 20126 Milan, Italy
[2] Consorzio Milano Ricerche, 20126 Milan, Italy
[3] DELOS Srl, 20091 Bresso (Milan), Italy

ARTICLE INFO

Article history:
Received 21 February 2008
Received in revised form 19 June 2009
Accepted 28 June 2009
Available online 5 July 2009

Keywords:
Genetic Programming
Machine learning
Regression
Docking energy
Computational biology
Drug design
QSAR

ABSTRACT

Statistical methods, and in particular Machine Learning, have been increasingly used in the drug development workflow to accelerate the discovery phase and to eliminate possible failures early during clinical developments. In the past, the authors of this paper have been working specifically on two problems: (i) prediction of drug induced toxicity and (ii) evaluation of the target–drug chemical interaction based on chemical descriptors. Among the numerous existing Machine Learning methods and their application to drug development (see for instance [F. Yoshida, J.G. Topliss, QSAR model for drug human oral bioavailability, Journal of Medicinal Chemistry 43 (2000) 2575–2585; Frohlich, J. Wegner, F. Sieker, A. Zell, Kernel functions for attributed molecular graphs—a new similarity based approach to ADME prediction in classification and regression, QSAR and Combinatorial Science, 38(4) (2003) 427–431; C.W. Andrews, L. Bennett, L.X. Yu, Predicting human oral bioavailability of a compound: development of a novel quantitative structure–bioavailability relationship, Pharmacological Research 17 (2000) 639–644; J Feng, L. Lurati, H. Ouyang, T. Robinson, Y. Wang, S. Yuan, S.S. Young, Predictive toxicology: benchmarking molecular descriptors and statistical methods, Journal of Chemical Information Computer Science 43 (2003) 1463–1470; T.M. Martin, D.M. Young, Prediction of the acute toxicity (96-h LC50) of organic compounds to the fat head minnow (*Pimephales promelas*) using a group contribution method, Chemical Research in Toxicology 14(10) (2001) 1378–1385; G. Colmenarejo, A. Alvarez-Pedraglio, J.L. Lavandera, Chemoinformatic models to predict binding affinities to human serum albumin, Journal of Medicinal Chemistry 44 (2001) 4370–4378; J. Zupan, P. Gasteiger, Neural Networks in Chemistry and Drug Design: An Introduction, 2nd edition, Wiley, 1999]), we have been specifically concerned with Genetic Programming. A first paper [F. Archetti, E. Messina, S. Lanzeni, L. Vanneschi, Genetic programming for computational pharmacokinetics in drug discovery and development, Genetic Programming and Evolvable Machines 8(4) (2007) 17–26] has been devoted to problem (i). The present contribution aims at developing a Genetic Programming based framework on which to build specific strategies which are then shown to be a valuable tool for problem (ii). In this paper, we use target estrogen receptor molecules and genistein based drug compounds. Being able to precisely and efficiently predict their mutual interaction energy is a very important task: for example, it may have an immediate relationship with the efficacy of genistein based drugs in menopause therapy and also as a natural prevention of some tumors. We compare the experimental results obtained by Genetic Programming with the ones of a set of "non-evolutionary" Machine Learning methods, including Support Vector Machines, Artificial Neural Networks, Linear and Least Square Regression. Experimental results confirm that Genetic Programming is a promising technique from the viewpoint of the accuracy of the proposed solutions, of the generalization ability and of the correlation between predicted data and correct ones.

1. Introduction

The goal of this paper is to investigate the usefulness of Genetic Programming (GP) [9,10] for automatically generating the underlying functional relationship between a set of molecular descriptors of drug-like compounds and their value of the interaction, or *docking*, energy with a particular estrogen receptor. Being able to develop automatic computer systems to successfully and efficiently predict the mutual interaction energy between drug-like compounds and estrogen receptors would have a great impact, given that this interaction energy has an immediate relationship with the efficacy of those drugs.

GP is an evolutionary approach which extends the genetic model of learning to the space of programs. It is a major variation of

* Corresponding author. Tel.: +39 02 64487874; fax: +39 02 64487805.
E-mail address: vanneschi@disco.unimib.it (L. Vanneschi).

1568-4946/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2009.06.013

Genetic Algorithms [11,12] in which the evolving individuals are themselves computer programs instead of fixed length strings from a limited alphabet of symbols. In the last few years, GP has become more and more popular for biomedical and pharmacokinetic applications. In particular, GP has been recently used to mine large datasets with the goal of automatically generating the underlying (hidden) functional relationship between data and correlate the behavior of latent features with some interesting pharmacokinetic parameters bound to drug activity patterns. For instance, in [13] GP has been used to classify drug-like molecules in terms of their bioavailability, in [14] it has been used with mutual information methods for analyzing complex molecular data, in [8] it has been used for quantitative prediction of drug induced toxicity and in [15] it has been applied to cancer expression profiling data to select features and build molecular classifiers by mathematical integration of genes.

GP can be regarded as an optimization method, which makes no assumption on the objective functions and data. Furthermore, as pointed out in [8] and explained in details also further in this paper, GP often automatically performs a feature selection, maintaining into the population expressions that use subsets of data. Thus, the motivation behind our choice of investigating the usefulness of GP for assessing large biomedical datasets is twofold:

- biological/chemical data are not independent of each other. Rather, it has been verified that in most of the complex biochemical systems, small subsets of components work in cohesion [16]. These phenomena lead to high multi-dependency among the features. Hence, the underlying algorithm should make no assumption on the inter-dependencies between the different variables. Furthermore, the algorithm should be capable of extracting underlying features governing the biochemical reactions from high-dimensional correlated data.
- The dimensionality of the feature space in biomedical datasets is normally much higher than the number of observations available for training. Hence, automatic feature selection as well as other methods to handle overfitting and minimizing the generalization error should be encouraged.

Pharmacokinetics prediction tools are usually based on two approaches: *molecular modelling*, which uses intensive protein structure calculations and *data modelling*. Methods based on data modelling are widely reported in literature; they all belong to the category of Quantitative Structure Activity Relationship (QSAR) models [17] and they are adopted in the present work. To quantify the real usefulness of GP for the presented application, experimental results are compared with the ones of a set of well-known Machine Learning (ML) methods, including Support Vector Machines (SVM), Artificial Neural Networks, Linear and Least Square Regression. These ones will be referred to as "non-evolutionary" methods for simplicity.

This paper is structured as follows: Section 2 discusses previous and related work; in Section 3 we describe the method employed to build the dataset used in our experiments; Section 4 briefly describes the non-evolutionary ML methods used in this paper and discusses their experimental results on our dataset; in Section 5 we introduce the different versions of GP that we have tested in this work and we discuss their experimental results; Section 6 contains the description of a method to improve GP results for the studied problem; finally Section 8 concludes the paper and offers hints for future research.

2. Previous and related work

As outlined above, the goal of this paper is investigating the usefulness of GP for generating the hidden relationship between molecular descriptors and docking energy. Virtual molecular docking represents a basic step in rational drug design. Its objective is to predict how any macromolecules (typically a protein or nucleic acid) interact with other molecules called "ligands" (may be other proteins, peptides or small drug-like molecules) by calculating their interaction energy in some particular positions. Considerable efforts have been directed in understanding this process and optimizing it by computer simulations using many different computational methods, including Evolutionary Algorithms; see for instance [18–22]. We do not analyze in details all these contributions here, because this paper does not present a docking application, but a QSAR approach, where docking energy values are used as target.

Also, many software environments for molecular docking have been developed and commercialized. For the sake of brevity, here we only quote [23–25] and the DELOS software platform [26], which has been recently developed and which we have used to build our dataset, as described in Section 3. For a more detailed survey and discussion of the numerous existing software environments for docking optimization see for instance [19]. In the present work, we choose a particular macromolecule and ligands. As ligands, we use a set of drug-like compounds belonging to the genistein family. Genistein (*genesteina* or *genista tinctoria*) is an isoflavone $C_{15}H_{10}O_5$ found especially in soybeans which has been shown in laboratory experiments to be effective as a natural prevention of some tumors. As a macromolecule, we have used the estrogen receptor *ERα*, a member of the nuclear hormone family of intracellular receptors which is activated by the 17β-estradiol. The important effects of genistein on estrogen receptors is pointed out in many contributions; see for instance [27,28].

Many contributions have appeared to date using ML methods for training QSAR models. For instance fuzzy adaptive Least Squares are used in [1], GAs and Self Organizing Maps are used in [29], SVM are used in [2], various kinds of multivariate and Partial Least Square Regressions are used in [3], recursive partitioning and Partial Least Square regression has been tested in [4], multivariate Linear Regression and Artificial Neural Networks have been applied in [5], and a technique called Genetic Function Approximation has been proposed in [6]. Artificial Neural Networks, often used for

QSAR [7], are frequently integrated in existing commercial packages developed by software vendors involved in the field of molecular modelling. Some of these tools are analyzed in [30]. Among the leaders in this field, Accelrys Inc. [31] and Pharma Algorithms Inc. [32] essentially provide black-box mathematical models and/or Data Mining tools that can be used to build new predictors.

In the last few years, GP is becoming popular for QSAR modelling and related biomedical applications. For instance, in [13] GP is used to classify molecules in terms of their bioavailability; in [14] it has been used together with mutual information methods for analyzing QSAR data; in [8] it is used for quantitative prediction of drug induced toxicity, and in [15] it is applied to cancer expression profiling data to select feature genes and build molecular classifiers. To the best of our knowledge, the present contribution represents the first effort to develop a QSAR model for docking energy assessment using GP. This work is inspired by [27], were the goal was to discover new compounds that display the benefits of estrogens while avoiding the risk of reproductive tissue cancer. Authors applied a Virtual High-Throughput Screening, based on docking simulations, for the identification of new possible selective receptor compounds and discovered good values of the docking energy when some genistein molecules were used with *ERα* estrogen receptor.

3. Dataset

We have collected from the RCSB PDB database [33] a small set of estrogen–genistein virtual molecules. Successively we