Chemometrics and Intelligent Laboratory Systems 122 (2013) 1–11

Contents lists available at SciVerse ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Gravitational search algorithm: A new feature selection method for QSAR study of anticancer potency of imidazo[4,5-b]pyridine derivatives

Behnam Mohseni Bababdani, Mehdi Mousavi *

Department of Chemistry, Faculty of Sciences, Shahid Bahonar University of Kerman, P.O. Box 76175-133, Kerman, Iran

ARTICLE INFO

Article history:
Received 15 August 2012
Received in revised form 10 December 2012
Accepted 11 December 2012
Available online 22 December 2012

Keywords:
Gravitational search algorithm
Feature selection
QSAR study
Imidazo[4,5-b]pyridine
Anti-cancer potency
Aurora A kinase

ABSTRACT

Choosing the most suitable subset of descriptors among a large number of structural parameters is one of the most important and challenging steps in quantitative structure–activity relationship (QSAR) studies. So far, many feature selection algorithms have been applied in these studies, but none of them behave generally. In this study, a binary version of gravitational search algorithm (GSA) as a novel feature selection method is developed and coded for QSAR studies. The GSA is applied as a descriptor selection tool for anticancer potency modeling of a set of imidazo[4,5-b]pyridine derivatives consisting of 65 compounds. The GSA selected descriptors were subjected to Bayesian regularized artificial neural networks to model the anticancer potency. The generated model satisfactorily describes the experimental variation in the biological activity of the data set compounds. The results of external validation ($R_v^2 = 0.98$) and internal cross-validation tests ($Q_{LOO}^2 = 0.94$, $R_{L4O}^2 = 0.93$, $R_{L8O}^2 = 0.92$) in conjunction with Y-randomization confirm the predictive ability, robustness and effectiveness of the generated model. Also, comparison between GSA and genetic algorithm (GA) indicates that GSA has certain advantages over the GA.

1. Introduction

Cellular division is one of the distinctive features of living organisms tightly regulated by a vast number of proteins [1]. Cancer is a generic term for uncontrolled division, growth and spread of cells, which can occur in all parts of the body. One of the defining features of cancer is the rapid creation of abnormal cells that grow beyond their usual boundaries. These cells can then attack adjoining parts of the body and spread to other organs [2]. Cancer is a major public health problem and leading cause of death in many parts of the world [3]. Deaths from cancer worldwide are projected to continue rising, with an estimated 13.1 million deaths in 2030 [2].

A new cancer treatment method is using anticancer drugs that act against proteins involved in cancer cell proliferation [4]. Among the network of regulatory proteins in cellular division, Aurora A kinases are of particular relevance as they play a crucial role, by controlling chromatid segregation [1]. The Aurora kinase family, identified in 1990 [5], is a group of cell cycle-regulated serine/threonine kinases that are important for mitosis [4]. Aurora A, which possesses catalytic effect during mitosis, is one of the isoforms of Aurora kinase enzymes [4,6]. Aurora A is considered as an interesting target for new anticancer drugs [7,8] and its inhibition mechanism can be found elsewhere [4,9].

Compounds with different inhibition activities have shown promising activity against cancer and are introduced into clinical trials for treatment of various cancers [9–12]. Nevertheless, the search of potent anticancer compounds is still on the desktop of molecular modeling and drug design specialists [13]. Many successful applications of quantitative structure–activity relationship (QSAR) studies prove the usefulness of this approach in drug design and development [14–17]. Practical use of QSAR based models, as virtual screening tools to discover biologically active molecules, is growing increasingly [18–20].

Choosing the most suitable subset of descriptors among a large number of structural parameters generated by existing software such as Molconn-Z [21], CODESSA [22] and DRAGON [23] is one of the most critical steps in QSAR modelings. So far, many feature selection algorithms such as GA, ant colony optimization (ACO) and particle swarm optimization (PSO) have been used in QSAR studies [24–30]. In these algorithms, different criteria are implemented for selecting appropriate sets of descriptors. The GA is a heuristic search algorithm that mimics the process of natural evolution [31]. The initial population has a predefined number of candidate solutions called chromosomes. Chromosomes evolve using genetic operators, i.e. selection, reproduction and mutation. The ACO algorithm is based on the behavior of ants seeking a path between their colony and a source of food [32]. The paradigm is based on the medium used by ants to communicate information regarding shortest paths to food by means of pheromone trail. While an isolated ant moves randomly, an ant encountering a previously laid trail can detect its path and consequently reinforces the trail with its own pheromone. Therefore, the probability of choosing a path increases with the number of ants that previously chose the same path. The PSO is motivated from the

* Corresponding author. Tel.: +98 341 3222033; fax: +98 341 3202113.
E-mail address: mmousavi@mail.uk.ac.ir (M. Mousavi).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2012.12.002

simulation of the social behavior of flock of birds. This algorithm is initialized with a swarm of random candidate solutions, called particles. This optimization approach updates the position and velocity of particles in the search space by applying an operator according to the fitness information obtained from the environment. Thus, the individuals of the population can be expected to move toward a better solution [33]. Comprehensive reviews of various feature selection algorithms and their comparison are presented by Walters and Goldman [34] and Goodarzi et al. [35]. Accordingly, it is revealed that no unique feature selection method behaves generally, i.e. there are no absolute rules for selecting appropriate sets of descriptors. Therefore, development and application of new feature selection methods are of prime importance and necessary in QSAR studies.

Recently, a new population-based heuristic algorithm, namely gravitational search algorithm (GSA), based on the metaphor of gravitational interaction between masses was introduced mainly designed for problems in electrical engineering [36]. In this algorithm, the searcher agents are a collection of masses which interact with each other based on the Newtonian gravity and the laws of motion. A comprehensive comparison between GSA and some well-known heuristic algorithms such as GA and PSO was presented by Rashedi et al. Their results indicated that in optimization field, the GSA approach has some merit over other methods [36,37].

For the first time, in this paper the gravitational search algorithm is coded, developed and applied as a new feature selection method in chemical systems, mainly QSAR approach. Here, the GSA is used for choosing the most informative descriptors as inputs of artificial neural networks (ANNs) for anticancer property modeling of a series of imidazo[4,5-b]pyridine derivatives.

2. Methods

2.1. Feature selection

In QSAR studies, feature selection refers to a procedure which selects a subset of descriptors from a pool of candidate descriptors. This procedure increases reliability and reduces cost and time of modeling process [38]. So far, various feature selection methods have been developed which are based on different strategies and logics [34,35,39]. Deterministic variable selection algorithms such as forward selection, backward elimination and stepwise multiple linear regression (MLR) and partial least square (PLS) always select a specific subset of descriptors when applied on a given problem. These algorithms do not guarantee to create the optimal results, since they do not examine all possible subsets. Sometimes the results obtained are far from the global best subset.

To overcome this problem, it is possible to use either exhaustive search method or to apply binary version of heuristic algorithms to the descriptor space. In the first method, if one is supposed to select a set of d-descriptor from the descriptor space (X) with cardinality n, one should examine all $\binom{n}{d}$ possible d-subsets of the X. Obviously; this approach guarantees to reach global best optima. But, as the number of n increases, the number of d-descriptor combinations increases exponentially and hence the computational time grows dramatically [38]. Therefore, application of the exhaustive search method in problems with moderate and large number of descriptors is too time consuming or even impractical [38].

Alternatively, it is possible to use a binary version of heuristic algorithms which look for good (near-optimal) solutions at a reasonable computational cost and time. Most of these algorithms do their search in a parallel fashion with multiple random initial points which could produce different subsets on every run. So far, various heuristic approaches have been adopted [31–33,40–42]. Implementation of several well-known and some recently proposed feature selection algorithms, verifies significant improvement in modeling results. This is due to proper selection of descriptors [43–45].

In swarm-based algorithms, such as ACO, PSO and artificial bee colony algorithm (ABC) [42] each member executes a series of particular operations and shares its information with others [36]. Although these operations are almost simple, their collective effect, known as swarm intelligence, produces a surprising result [32,36,46]. Local interactions between members provide a global result which permits the system to solve the problem without using any central controller. Furthermore, in swarm-based heuristic algorithms, a heuristic search algorithm explores the search space to find new solutions and avoid trapping in a local optimum in leading iterations (exploration). By lapse of iterations, the algorithm tunes itself in semi-optimal points (exploitation). All swarm-based heuristic algorithms employ exploration and exploitation through using different approaches and operators, i.e. all with a common framework. A vital point for achieving a high performance search is a suitable trade-off between exploration and exploitation [36].

2.2. Fundamentals of binary gravitational search algorithm (BGSA)

The gravitation is the force that causes two objects to move toward each other because of their masses. The gravity is everywhere and each particle in the universe attracts any other particles. The inevitability of gravity makes it distinguishable from all other natural forces. In the Newton law of gravity, the attraction force among particles is formulated. The nature of this gravitational force is such that the heavier and the nearer the masses, the higher the force exerted on each other. An increase in the distance between two particles results in decreasing the gravitational force between them and vice versa [36,37].

Binary gravitational search algorithm is a multi-step process which applies the above mentioned rules to select appropriate variables. The steps of this search algorithm [37] are as follows:

(a) Search space identification
(b) Randomized initialization (generation) of binary agents
(c) Fitness evaluation of agents
(d) Updating $G(t)$, $best(t)$, $worst(t)$ and $M_i(t)$ for $i = 1, 2, \ldots, N$
(e) Calculation of the total force in different directions
(f) Calculation of acceleration and velocity
(g) Updating agents' position
(h) Repeating steps c to g until the stopping criteria is reached.

A brief description of BGSA terms and steps are presented here. The BGSA begins by agent generation. A string of variables with predefined length is called an agent. In binary format, agent string is a series of 1 and 0 values which indicate the presence or absence of descriptors in the agent, respectively. In a system with N agents, the position of the ith agent, X_i, in an n-dimensional space is represented by Eq. (1):

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots, x_i^n\right); \quad i = 1, 2, \ldots, N \tag{1}$$

where x_i^d represents the position of ith agent in the dth dimension. In discrete binary environment, every dimension can only take the value of 0 or 1. The binary search space is considered as a hypercube in which an agent may move to nearer and farther corners of it by flipping between various numbers of bits. Moving through a dimension means that the corresponding variable value changes from 0 to 1 and vice versa [36,37].

In BGSA, a mass is attributed to each agent. The values of masses, M, are calculated by fitness evaluation criteria, i.e. $fit_i(t)$, $worst(t)$ and $best(t)$, according to Eqs. (2) and (3):

$$q_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \tag{2}$$