Journal of Molecular Graphics and Modelling 63 (2016) 22–28

Contents lists available at ScienceDirect

Journal of Molecular Graphics and Modelling

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

Highly predictive support vector machine (SVM) models for anthrax toxin lethal factor (LF) inhibitors

Xia Zhang [a], Elizabeth Ambrose Amin [a,b,*]

[a] Department of Medicinal Chemistry, College of Pharmacy, University of Minnesota, 717 Delaware St. SE, Minneapolis, MN 55414-2959, United States
[b] Minnesota Supercomputing Institute for Advanced Computational Research, 117 Pleasant St SE, Minneapolis, MN, United States

ARTICLE INFO

Article history:
Received 28 September 2015
Received in revised form 7 October 2015
Accepted 6 November 2015
Available online 17 November 2015

Keywords:
Anthrax
Anthrax toxin lethal factor
Support vector machine
SVM

ABSTRACT

Anthrax is a highly lethal, acute infectious disease caused by the rod-shaped, Gram-positive bacterium *Bacillus anthracis*. The anthrax toxin lethal factor (LF), a zinc metalloprotease secreted by the bacilli, plays a key role in anthrax pathogenesis and is chiefly responsible for anthrax-related toxemia and host death, partly via inactivation of mitogen-activated protein kinase kinase (MAPKK) enzymes and consequent disruption of key cellular signaling pathways. Antibiotics such as fluoroquinolones are capable of clearing the bacilli but have no effect on LF-mediated toxemia; LF itself therefore remains the preferred target for toxin inactivation. However, currently no LF inhibitor is available on the market as a therapeutic, partly due to the insufficiency of existing LF inhibitor scaffolds in terms of efficacy, selectivity, and toxicity. In the current work, we present novel support vector machine (SVM) models with high prediction accuracy that are designed to rapidly identify potential novel, structurally diverse LF inhibitor chemical matter from compound libraries. These SVM models were trained and validated using 508 compounds with published LF biological activity data and 847 inactive compounds deposited in the Pub Chem BioAssay database. One model, **M1,** demonstrated particularly favorable selectivity toward highly active compounds by correctly predicting 39 (95.12%) out of 41 nanomolar-level LF inhibitors, 46 (93.88%) out of 49 inactives, and 844 (99.65%) out of 847 Pub Chem inactives in external, unbiased test sets. These models are expected to facilitate the prediction of LF inhibitory activity for existing molecules, as well as identification of novel potential LF inhibitors from large datasets.

1. Introduction

Anthrax is an acute, often fatal infectious disease caused by the rod-shaped, spore-forming bacterium *Bacillus anthracis*. Primarily a zoonotic disease affecting livestock and wild animals, anthrax has more recently emerged as a lethal bioterror agent, with the inhalational form posing a particular threat to society. Anthrax-related toxicity has been attributed primarily to its plasmid-encoded, secreted exotoxin comprising the lethal factor (LF), the edema factor (EF, a calmodulin-activated adenylate cyclase), and the protective antigen (PA) [1]. LF, a zinc-dependent hydrolase, joins with PA to form the anthrax lethal toxin, which is chiefly responsible for cytotoxicity and eventual host death associated with anthrax pathogenesis [2]. The protective antigen delivers LF into the cytoplasm of host cells, where LF cleaves and inactivates mitogen-activated protein kinase kinases (MAPKKs), thereby interfering with signaling processes that are essential for cell function and survival, most notably involving the immune response [3–5]. Antibiotics such as fluoroquinolones are capable of eradicating the bacilli, however, host death from residual toxemia can occur even after *B. anthracis* is cleared from the system, and there is currently no extant therapeutic modality to directly combat LF-mediated cytotoxicity [6,7].

As *B. anthracis* continues to pose a significant threat as a biological weapon, various experimental and computational efforts have been focused on identifying small-molecule LF inhibitors as potential drugs as adjunct therapeutics with antibiotics [4,8–33]. Previous computational modeling efforts have been primarily directed toward structure-based virtual screening, pharmacophore mapping, and 3D-QSAR model development [28–33]. While these studies have been useful for the prediction of LF inhibitory activity and the identification of common molecular features in LF inhibitors, compounds addressed in these studies have chiefly been limited to one or two structural classes. Studies have demonstrated

* Corresponding author at: Department of Medicinal Chemistry, College of Pharmacy, University of Minnesota, 717 Delaware St SE, Minneapolis, MN 55416, United States. Fax: +612 626 6346.
E-mail address: eamin@umn.edu (E.A. Amin).

http://dx.doi.org/10.1016/j.jmgm.2015.11.008
1093-3263/

that models built on a structurally similar set of compounds occupying closely adjacent areas of chemical space are likely to have limited applicability in terms of identifying novel inhibitor classes, and thus may result in unreliable predictions when used in virtual screening of structurally diverse chemical databases [34,35].

With the goal of overcoming this roadblock, in the current work we have assembled a diverse set of active and inactive LF inhibitors collected from the literature, to develop novel support vector machine (SVM) models that can be used to accurately identify new compounds (or compounds based on novel scaffolds) that may exhibit favorable LF inhibitory activity. The SVM method has consistently demonstrated robust predictivity in lead identification and optimization, and has also proven useful in the prediction of drug metabolism, blood–brain barrier penetration, *p*-glycoprotein substrates, oral absorption, and the efficacy of various enzyme inhibitor therapeutics [36]. The SVM models we report here have been rigorously validated using 10-fold cross-validation, and they have demonstrated quite favorable accuracy in predicting biological activities of external, unbiased test set compounds. Specifically, as discussed below, a particularly efficacious model using MOE (Chemical Computing Group, Inc.) descriptors successfully identified 39 (95.12%) of 41 nanomolar-level LF inhibitors, while rejecting 46 (93.88%) of 49 inactives and 844 (99.65%) of 847 inactives in a series of compound set evaluations. We found that these validation and testing results support the application of our SVM models as screening tools for identifying potentially potent LF inhibitors.

2. Methodology

2.1. Data sets

Compound structures and biological activities for 546 LF inhibitors of varying potency (database **DB**) were collected from the literature as described in our previously published work [37]. A total of 102 compounds with LF IC_{50} or *Ki* values less than 1 μM were considered to be active LF inhibitors. These displayed high structural diversity and included sulfonamide hydroxamates, rhodanine-based derivatives, guanidinylated 2,5-dideoxystreptamine derivatives, guanidinylated derivatives of neamine, aniline, and γ-ether, an *N*-sulfonylated phenylfuran derivative, and an *N*-hydroxyhexanamide analog, among other scaffold types. 122 compounds with specified IC_{50} or *Ki* values larger than 100 μM, or nonspecified IC_{50} or *Ki* values larger than 40 μM, were considered to be inactive. Taken together, these 224 compounds (subset database **DBA**) were used for SVM model development and validation. From among the remaining 320 compounds in **DB**, 284 compounds (subset database **DBB**) with IC_{50} or *Ki* values ranging from 1 μM to 40 μM were treated as weakly active compounds and were set aside for model validation. In addition to **DB**, 847 inactive compounds from two recently reported high-throughput screening experiments deposited on Pubchem BioAssay (AID: 602142 and 602326) were used as an external validation set and were termed database **DBC**. Although 13 compounds in **DBC** were reported to be active, they lacked specific IC_{50} values and were therefore not included in the validation set.

2.2. Computational methods

2.2.1. 3D Structure generation.

Three-dimensional conformations of all dataset structures were generated via geometry optimization by energy minimization in Pipeline Pilot, and were further geometry optimized in MOE 2011.10 (Chemical Computing Group, Inc.) using the MMFF94s force field with a convergence criterion of 0.01 kcal/mol Å [38].

2.2.2. Molecular descriptor calculation.

2.2.2.1. MOE Descriptors. Molecular descriptors were used in this study to quantitatively represent structural and physicochemical properties of compounds. A total of 334 2D and 3D molecular descriptors were calculated using MOE 2011.10 [39]. These included subdivided surface areas, atom counts and bond counts, Kier & Hall connectivity and Kappa Shape indices, and physical property-related, adjacency and distance matrix, pharmacophore feature, partial charge, potential energy, MOPAC, surface area, volume and shape, and conformation-dependent charge descriptors. Any descriptors with missing values were eliminated, resulting in a final set of 313 descriptors.

2.2.2.2. Schrödinger descriptors. We incorporated a total of 292 topological, MOPAC, and ADME-tox related descriptors (relevant to potential therapeutic design and optimization) from Schrödinger, Inc., using Maestro 9.3 [40].

2.2.2.3. ISIDA Fragment descriptors. The Online Chemical Modeling Environment was used to calculate a series of ISIDA 2D fragment descriptors [41]. Descriptors with low variance (less than 0.01) or with fewer than two unique values were removed. Also, if the correlation coefficient between two descriptors was larger than 0.95, one descriptor was eliminated. A total of 748 ISIDA fragment descriptors were utilized in this work.

2.2.3. SVM Modeling approaches

2.2.3.1. Data set division for model development and validation. Database **DBA** was randomly split into a training set (Train 1) of 134 compounds (61 actives and 73 inactives, 60% of **DBA**) and an external test set (Test 1) of 90 compounds (41 actives and 49 inactives, 40% of **DBA**). In addition, in order to assess the ability of the resulting SVM models to classify compounds that are structurally dissimilar to the training set, active and inactive **DBA** compounds were clustered based on ECFP_4 descriptors in Pipeline Pilot 8.0 (Accelrys, Inc.). One cluster containing 44 actives and one cluster containing 51 inactives were extracted from **DBA** as an external test set (Test 2). The remaining structures in DBA were retained as a training set (Train 2), in order to ensure that Test 2 compounds would be structurally dissimilar to those in Train 2.

2.2.3.2. Support vector machine (SVM). SVM is a popular and effective classification algorithm in which data points (in this case, inhibitor compounds) are mapped onto descriptor-based feature space, and a decision boundary (expressed as $\omega^T x + b = 0$) is identified using support vectors to separate compounds into two categories (actives and inactives) by the widest gap (margin) via a hyperplane. Support vectors often constitute a small portion of examples in the training set, allowing an SVM model to be less prone to overfitting while maintaining generalizability [42].

Specifically, for an input set of pairs $(x^{(i)}, y^{(i)})$, $i = 1, ..., m$, $x^{(i)} \in R^P$ (P is defined as the dimension of the input space), $y^{(i)} \in \{-1, 1\}$, presenting the classes of an sample $x^{(i)}$, the following optimization can be formulated:

$$\min_{\gamma,\ \omega,\ b} \frac{1}{2}\|\omega\|^2$$

$$\text{s.t.} y^{(i)}(\omega^T x^{(i)} + b) \geq 1, \quad i = 1, ..., m$$

However, sometimes the data may not be easily separable. Also, where outliers exist, finding a separating hyperplane may not offer the best solution to a problem. In order for the algorithm to function for non-separable data and exhibit less sensitivity to outliers, the