A novel chemoinformatics method based on PCA conversion of the connection-table and weighted LS-SVM applied for identification of biofilm inhibitors

Minta Thomas[1*], Hans P Steenackers[2], Marc De Maeyer[4], Johan AK Suykens[1], Inge Thijs[1], Xiaoyu Qing[4], Tran Thi Thu Tran[3], Erik Van der Eycken[3], Jos Vanderleyden[2] and Bart De Moor[1]

[*]Correspondence: minta82@gmail.com
[1]KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics / iMinds Medical IT, Kasteelpark Arenberg 10, 3001 Leuven, Belgium
Full list of author information is available at the end of the article

Abstract

Background The most common representation of compounds in chemoinformatics is a connection-table, i.e., a table enumerating the atoms and another enumerating the bonds. In this study, we intend to derive a new chemical descriptor from this table allowing a better distinction between biologically active and inactive compounds. Our method was applied to the identification of inhibitors of *Salmonella*, *Pseudomonas* biofilm formation and inhibitors of thrombin, trypsin, and factor Xa. Development of this type of anti-microbial is urgently needed as biofilms, surface-associated bacterial communities embedded in a self-produced polymeric matrix, provide strong protection against the activity of antibiotics, disinfectants and the immune system. Method We propose a new machine learning approach for the identification of biologically active compounds. Principal Component Analysis (PCA) converts the connection-table of each compound into a structural descriptor of two vectors: one corresponding to atoms and the other to bonds. As a supervised classification algorithm, a weighted least squares support vector machine (LS-SVM) is used in which the table enumerating the atoms is weighted against the table enumerating the bonds. We apply this framework to, a given experimental data set on preventive activity of 2-aminoimidazole-based compounds against *Salmonella* and *Pseudomonas* biofilms and its 10 newly synthesized validation compounds, and 72 inhibitors of the benzamidine type with respect to their binding affinities toward thrombin, trypsin, and factor Xa. Prediction performances are evaluated and compared with the LS-SVM classifier on other well-known chemical descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints, MACCS Keys and Burden Eigenvalues descriptors. Result The proposed machine learning approach obtained the best test accuracy with F-score compared to the LS-SVM classifier on the other discussed chemical descriptors. In addition it is experimentally observed that the proposed approach is able to discriminate the few compounds with the highest activity from the 10 validation compounds of *Salmonella* and *Pseudomonas*. Conclusion The results suggest that the newly proposed approach, the weighted chemical descriptors of molecular structure increases, the predictive performance of chemical compounds. The proposed machine learning approach could be applicable to

any problem in which the property or activity of interest is dependent upon the molecular structure.

connection-tablechemical descriptorprincipal component analysisLS-SVM classifier

Introduction

Computational methods involving machine learning techniques are among the most popular tools used in chemoinformatics tasks [1–3]. Most of these techniques in chemoinformatics are used for the classification of chemical compounds based on a descriptor representation of the compounds. Historically, the most widely used descriptors have been based on fingerprints, such as the Extended-Connectivity Descriptor in which an iterative process assigns numeric identifiers to each atom, independent of the original numbering of the atoms [4]. Path-Length Fingerprints are generated with respect to path lengths of chemical compounds [5]. MACCS Keys [6] are the sets of descriptors based on structural fragments, that have been identified a priori by domain experts [7]. Burden eigenvalue descriptors (BCUT) rely on an eigenvalue factorization of molecular connection table [8].

Biofilms have been found to be involved in a wide variety of microbial infections in the body and it was observed that 80% of all human bacterial infections are biofilm-associated [9, 10]. Biofilms of *Pseudomonas aeruginosa* can cause chronic infections, which are a serious problem for medical care in industrialized societies, especially for immunocompromised patients and the elderly [11]. They often cannot be treated effectively with traditional antibiotic therapy. Biofilm formation is also an important survival strategy of *Salmonella* Typhimurium, both within and outside the host [12]. Given the extent of problems caused by biofilms, there has been a significant effort to develop new anti-biofilm strategies [13, 14]. Traditionally biologists identify biofilm inhibitors by wet-laboratory screening of thousands of compounds. This strategy is time consuming and expensive. Chemoinformatics approaches employ a range of methods which are primarily used for the rapid evaluation and prioritization of compounds prior to wet-laboratory testing. Although molecular docking, a well established computational technique, was applied for the identification of potential histidine kinase inhibitors to combat *Staphylococcus epidermidis* [15] and for the selection of potential biofilm inhibitors of *Pseudomonas aeruginosa* [16], chemoinformatics approaches have not been fully exploited yet in the area of bacterial biofilm inhibition.

The main purpose of this work is to introduce a machine learning approach, a weighted chemical descriptor based on the connection-table of compounds, for designing a prediction model for identification of the activity of chemical compounds in a specific biological condition. This method was applied for the identification of inhibitors of *Salmonella* and *Pseudomonas* biofilm formation and inhibitors of Thrombin, Trypsin and Factor Xa. Initially, molecular structures were converted into structural descriptors in terms of two vector spaces: atoms and bonds by Principal Component Analysis (PCA), which transforms the two dimensional connection-table into a single dimensional space. Next, a weighted least squares support vector machine (weighted LS-SVM) classifier [17] was applied on these two vector spaces to integrate them into a single vector, named as a weighted chemical descriptor (which

is completely representing the connection-table) and further predict the biological activity of the chemical compounds. Finally we compared the prediction performance of the weighted LS-SVM classifier on the newly proposed descriptor with the LS-SVM classifier on other well known descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints, MACCS Keys and BCUT descriptors.

Data Sets

In our analysis, initially we used connection-table of 308 2-aminoimidazole-based compounds [18–22] that have the capability to prevent biofilm formation of *S.* Typhimurium or *P. aeruginosa*, for model development and an independent set of 10 novel compounds for validation. The analytical data of the validation compounds are given in Supplementary File I. In addition, we have IC50 values of these 308 compounds. All compounds were synthesized according to our previously reported protocols [18–22]. The IC50 value is defined as the concentration at which biofilm formation is inhibited by 50% and thus predicts the effectiveness of these compounds against biofilm formation. The lower the IC50, the more effective a compound is against the biofilm formation. Later we used Bohm Serin Protease Inhibitor Data Set to identify the inhibitors of Thrombin, Trypsin and Factor Xa [23].

Methods

The methods used for this study can be subdivided into two categories: feature selection from the connection-table (table enumerating atoms and table enumerating bonds) using PCA and prediction with either weighted LS-SVM classifier or LS-SVM classifier.

Principal Component Analysis (PCA)

PCA was applied to convert the connection-table of each compound in the data sets, i.e. inhibitors of *S.* Typhimurium and *P. aeruginosa* biofilm formation, into a novel chemical descriptor of two vectors: one corresponding to atoms and the other to bonds.

PCA [24] is mathematically defined as an orthogonal transformation that converts the data into a new coordinate system, such that the largest variance by any projection of the data lies on the first coordinate (first principal component), the second largest variance on the second coordinate, and so on. PCA of a $m \times n$ matrix A and is defined as follows:

$$A = Score * Coef^T$$

where each column of the $n \times n$ matrix $Coef$ are the coefficients of corresponding principal components and $Score$, $m \times n$ matrix, is the representation of A in the principal component space.

Weighted Least Squares-Support Vector Machine (Weighted LS-SVM) Classifier

Initially PCA converts the connection-table of each compounds into two descriptor vectors, one in atom space and the other in bond space. These two vectors should be considered together for the complete representation of a chemical compound.

These descriptors are effectively integrated back into a single space by using the weighted LS-SVM classifier proposed in [17] which also performs further prediction.

Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$. Consider the feature map $\varphi^{(1)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_1}$ and $\varphi^{(2)}(.) : \mathbb{R}^p \to \mathbb{R}^{n_2}$ to high dimensional feature space, which is possibly infinite dimensional. The centered feature matrices $\Phi_c^{(1)} \in \mathbb{R}^{n_1 \times N}$, $\Phi_c^{(2)} \in \mathbb{R}^{n_2 \times N}$ become

$$\Phi_c^{(1)} = [\varphi^{(1)}(x_1^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T; \ldots; \varphi^{(1)}(x_N^{(1)})^T - \hat{\mu}_{(\varphi_1)}^T]^T$$

$$\Phi_c^{(2)} = [\varphi^{(2)}(x_1^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T; \ldots; \varphi^{(2)}(x_N^{(2)})^T - \hat{\mu}_{(\varphi_2)}^T]^T,$$

where $\hat{\mu}_{\varphi l} = \frac{1}{N}\Sigma_{i=1}^N \varphi^{(l)}(x_i^{(l)})$, $l = 1, 2$.

The weighted LS-SVM classifier proposed in [17] optimizes the following problem

$$\boxed{\text{P}}: \quad \min_{v,e,b} \; J(v,e) \; = \; \gamma \tfrac{1}{2} e^T (\Phi_c^{(2)^T} \Phi_c^{(2)})^{-1} e + \tfrac{1}{2} v^T v$$

$$\textit{such that} \quad y = \Phi_c^{(1)^T} v + b 1_N + e,$$

with $e = [e_1, \ldots, e_N]^T$ a vector of variables to tolerate misclassifications, weight vector v in primal weight space, bias term b and regularization term $\gamma \in \mathbb{R}^+$. Solution to the above problem in dual space,

$$\begin{bmatrix} 0 & 1_N^T \\ 1_N & \Omega_c^{(1)} + \gamma^{-1}\Omega_c^{(2)} \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{1}$$

with $y = [y_1, ..., y_N]^T$, $1_N = [1, ..., 1]^T$, $\alpha = [\alpha_1, ..., \alpha_N]^T$, $\Omega_c^{(1)} = \Phi_c^{(1)^T} \Phi_c^{(1)}$ and $\Omega_c^{(2)} = \Phi_c^{(2)^T} \Phi_c^{(2)}$.

The resulting classifier in the dual space is given by

$$y(x) = \text{sign}[\sum_{i=1}^N \alpha_i ([K^{(1)}(x, x_i) + \frac{1}{\gamma} K^{(2)}(x, x_i)] + b)]. \tag{2}$$

with α_i the Lagrange multipliers, $K^{(1)}(x,z) = \varphi^{(1)}(x)^T \varphi^{(1)}(z)$, $K^{(2)}(x,z) = \varphi^{(2)}(x)^T \varphi^{(2)}(z)$ and $y(x)$ is the output corresponding to validation point x. Throughout the paper we use radial basis function (RBF) kernel: $K(x, x_i) = \exp(-\frac{||x-x_i||^2}{2\sigma^2})$ (RBF kernel with bandwidth σ).

Here the advantages of a weighted LS-SVM classifier are thus explored by designing two kernel matrices: a measure of similarity between two compounds based on their atoms and bonds description. In the primal space of the weighted LS-SVM

framework, each atom description is weighted with the bond description and the solution in dual space, provides a way to integrate these two kernel matrices efficiently (see Equation 2). The resultant chemical descriptors are completely representing the connection-table of chemical compounds.

The parameter selection for the weighted LS-SVM classifier is illustrated in the Algorithm box. For several possible values of the kernel parameters σ_1 and σ_2, the leave-one-out cross validation (LOO-CV) performance is computed for each possible combination of γ . The optimal parameters are the combinations (σ_1, σ_2, γ) with best LOO-CV performance. Note the complexity of this optimization procedure because both the kernel parameters (σ_1 and σ_2) and γ need to be optimized in the sense of the LOO-CV performance.

Box: Algorithm: weighted LS-SVM classifier for prediction of chemical compounds

1 Given a training data set of N points $\mathcal{D} = \{x_i^{(1)}, x_i^{(2)}, y_i\}_{i=1}^N$ with output data $y_i \in \mathbb{R}$ and input data sets $x_i^{(1)} \in \mathbb{R}^m$, $x_i^{(2)} \in \mathbb{R}^p$.

2 Find optimal γ and σ_1, σ_2 bandwidths of kernel functions $K^{(1)}$, $K^{(2)}$ by solving linear system Equation (1) using leave-one-out cross validation (LOO-CV).

3 Predict the performance of the model by plugging the optimal parameters from previous steps into Equation (2).

Least Squares-Support Vector Machine (LS-SVM) Classifier

In the previous section, we have used the weighted LS-SVM classifier for designing a weighted chemical descriptor of compounds and further perform predictions. For each data set, i.e. *S.* Typhimurium and *P. aeruginosa*, to compare prediction performances of the proposed descriptor with other existing descriptors, LS-SVM classifier is applied.

The constrained optimization problem for the least squares support vector machine (LS-SVM) [25, 26] for classification, is defined as follows. Given a training set $\{x_i, y_i\}_{i=1}^N$ with input data $x_i \in \mathbb{R}^d$ and corresponding binary class labels $y_i \in \{-1, +1\}$. Vapnik's SVM classifier formulation was modified in [25] into the following LS-SVM formulation:

$$\min_{w,b,e} \frac{1}{2} w^T w + \gamma \frac{1}{2} \Sigma_{i=1}^N e_i^2$$

subject to:

$$y_i[w^T \varphi(x_i) + b] = 1 - e_i, \qquad i = 1, \ldots, N$$

where $\varphi(.)$: $\mathbb{R}^d \to \mathbb{R}^{d_h}$ is the feature map which maps the d-dimensional input vector x from the input space to the d_h-dimensional feature space. The solution in dual space is

$$\begin{bmatrix} 0 & y^T \\ y & \Omega + \frac{I}{\gamma} \end{bmatrix} \begin{bmatrix} b \\ \beta \end{bmatrix} = \begin{bmatrix} 0 \\ 1_N \end{bmatrix}$$

with $y = [y_1, \ldots, y_N]^T$, $1_N = [1, \ldots, 1]^T$, $e = [e_1, \ldots, e_N]^T$, $\beta = [\beta_1, \ldots, \beta_N]^T$ and $\Omega_{i,j} = y_i y_j K(x_i, x_j)$ where $K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j)$.

The classifier in the dual space takes the form

$$y(x) = \text{sign}[\sum_{i=1}^{N} \beta_i y_i K(x, x_i) + b]$$

where β_i are Lagrange multipliers.

Static Peg Assay for Prevention of Biofilm Formation [18]

The device used for biofilm formation is a platform carrying 96 polystyrene pegs (Nunc no. 445497) that fits as a microtiter plate lid with a peg hanging into each microtiter plate well (Nunc no. 269789) [27]. Two-fold serial dilutions of the compounds in 100 μL of Tryptic Soy Broth diluted 1/20 (TSB 1/20; BD Biosciences) per well were prepared in the microtiter plate in duplicate Subsequently, overnight cultures (grown in LuriaBertani medium) [28] of *S.* Typhimurium ATCC14028 or *P. aeruginosa* PA14 were diluted 1:100 into the respective liquid broth and 100 μL (106 cells) was added to each well of the microtiter plate, resulting in a total amount of 200 μL of medium per well. The pegged lid was placed on the microtiter plate, and the plate was incubated for 24 h at 25C without shaking. For quantification of biofilm formation, the pegs were washed once in 200 μL of phosphate buffered saline (PBS). The remaining attached bacteria were stained for 30 min with 200 μL of 0.1% (w/v) crystal violet in an 2-propanol/methanol/PBS solution (v/v 1:1:18). Excess stain was rinsed off by placing the pegs in a 96-well plate filled with 200 μL of distilled water per well. After the pegs were air-dried (30 min), the dye bound to the adherent cells was extracted with 30% glacial acetic acid (200 μL). The OD570 of each well was measured using the Multimode Synergy MX, Biotek . The BIC50 value (i.e. IC50 for biofilm inhibition) and 95% confidence interval for each compound was determined from the concentration gradient by using the GraphPad software of Prism.

Result

The proposed method defines a new chemical descriptor of compounds from their connection-table, a representation of chemical molecules. Below is an example of the connection-table, a portion of a structure-data file (SDF file):

```
MOLCONV
3 2  0  0  1  0              1 V2000
5.9800   -0.0000   -0.0000 Br  0  0  0  0  0  0
4.4000   -0.6600    0.8300 C   0  0  0  0  0  0
3.5400   -1.3500   -0.1900 C   0  0  0  0  0  0
1  2  1  0
2  3  1  0
```

The first row of the table is the header line. The remaining rows are actual part of the connection-table in which the second row indicates: 3 atoms, 2 bonds, ..., V2000 standard. Rows 3-5 represent the atom block (1 line for each atom) and rows 6-7

represent the bond block (1 line for each bond). The columns of the atoms block specify the X,Y,Z-coordinates, atom symbol, isotope, charge, stereo code and the columns of the bond block specify the row numbers for atoms, and codes for bond type, bond stereochemistry etc. The table contains non-numerical data, i.e. atom symbol, which is converted to a numerical representation as follows while the other columns remain the same:

```
C-0  N-1  H-2  F-3  Br-4  O-5  S-6  CL-7  Si-8  K-9 ...
```

The molecule in the given example, has three atoms (3 rows with 10 columns) and two bonds (2 rows with 4 columns). We form two tables (referred to as Table 1 and Table 2) corresponding to the atoms and the bonds description of the chemical compounds, respectively.

A chemical compound can be denoted with variety of different connection tables describing one and the same compound but with different numbering of atoms. Canonicalization is the task which taking one of the numberings as the standard one and derive a unique code from it. It is accomplished by numbering the atoms of a molecule so that it is represented by only one connection table. Morgan algorithm is used for deriving a canonical code for the molecule [29, 30].

Initially, PCA is applied to mean centered atom and bond tables individually. The first component which captures the majority of the variance of the data (for each compound, the first component explaining minimally 70% variance) is selected for each table. Thus, we obtain two vectors, one corresponding to a table enumerating the atoms and another to a table enumerating bonds for each compound. Based on this, we construct two tables: compounds vs. atoms and compounds vs. bonds. The number of atoms and bonds are different for each compound. The entries in the descriptor vectors are filled with zeros, if the compound does not have an atom or bond corresponding to an entry. An overview of chemical descriptor formation from the connection-table of compounds is given in Figure 1.

In the weighted LS-SVM classifier, the RBF kernel function provides a measure of similarity between two compounds based on their atoms and bonds description, resulting in two kernel matrices. In the primal space of the weighted LS-SVM framework, each atom description is weighted with the bond description and the solution in dual space, provides a way to integrate these two kernel matrices efficiently (see Equation 2). The weighting scheme in the weighted LS-SVM classifier with RBF kernel function provides a way to represent chemical descriptors efficiently and perform further classification. In order to obtain the optimized values of σ_1 and σ_2, LOO-CV performance is computed with each possible combination of γ. The combinations of the parameters which offer the highest LOO-CV performance are selected as the optimal parameters.

To evaluate the performance of the proposed machine learning approach, initially we randomly split the given compounds into two groups: training (2/3rd of compounds) and test data sets. This split was performed stratifying to ensure that the relative proportion of active and inactive compounds in both training and test sets was the same. The classification model was built on the training data and the prediction performance of the method was evaluated on the test data. In the case of *Pseudomonas*, we have eliminated 78 compounds for which the corresponding IC50 values were not available. Compounds were categorized into two sets: active or less

(not) active based on the IC50 values. Selecting a cut-off value of IC50 to classify a compound as active or inactive is arbitrary. In the trained model of *Salmonella* and *Pseudomonas*, the averaged performance measurements on 30 iterations of the test data were calculated for a range of IC50 values (e.g. IC50 = 10 μM,..., IC50 = 100 μM). The prediction performance of the model was represented in terms of test areas under the receiver operating characteristic curves (AUC), accuracy and F-score. The F-score indicates the balance between precision and recall. The optimal cut-off value, below which a compound was considered active, was calculated based on the best prediction performances. Models with the optimal cut-off value could offer best performance for the future predictions. We have determined the optimal cut-off value of IC50 as 10 μM for both the *Salmonella* and *Pseudomonas* and 5 μM for Thrombin, Trypsin and Factor Xa case studies. At optimal cut-off value, the *Salmonella* data set contains 47 active and 159 inactive compounds as training set, and 23 active and 79 inactive compounds as test set. While the *Pseudomonas* data set contains 32 active and 122 inactive compounds as training set and 15 active and 61 inactive compounds as test set.

Table 3 and Figures 2, 3 compare the prediction performances (with optimal cut-off) of the weighted LS-SVM on the proposed descriptor with the LS-SVM on other well known chemical descriptors such as Extended-Connectivity Fingerprints, MACCS keys, BCUT and Path-Length Fingerprints, on *Salmonella* and *Pseudomonas* data set. Table 4 illustrates the averaged prediction performances of different descriptors on Thrombin, Trypsin and Factor Xa data sets. Each traditional chemical descriptor is completely representing a compound, thus for final prediction, just LS-SVM classifier is applied. While the PCA based descriptor spaces should be considered together to obtain a complete representation of chemical compound, which necessitates the use of the weighted LS-SVM classifier to integrate these two data sets and further perform predictions. Statistical significance tests were performed in order to allow a correct interpretation of the results. A non-parametric paired test, the signed rank test has been used in order to make general conclusions. Two results are significantly different if the value of the signed rank test (p-value) applied to both of them is lower than 0.05. In both case studies, the proposed novel machine learning approach performed well with test accuracy and F-score, while the best test AUC returned by MACCS keys for *Salmonella* and Path keys for *Pseudomonas*. In Tables 3 and 4, BCUT descriptor performed better than proposed method in terms of F-score and test accuracy, respectively.

Finally, we used the designed model to predict the activity of 10 novel synthesized compounds (see Table 5) against biofilms. These compounds were synthesized and *in vitro* tested for preventive activity against *Salmonella* and *Pseudomonas* biofilm formation according to previously published procedures [18]. First, we trained the model based on 308 compounds for different cut-off values of the IC50 (e.g. IC50 = 10 μM, IC50 = 20μM, IC50 = 30 μM, IC50 = 50 μM and IC50 = 100 μM) to define activity and non-activity of the compounds. Then we used these trained models to predict the activity of the novel compounds based on different cut-off values. Table 6 lists the compounds identified by each descriptor for different IC50 cut-off values. Finally, with the *in vitro* measured IC50 values of the 10 experimentally synthesized compounds, the prediction performances of all the discussed descriptors

were obtained and represented in Tables 7 and 8 in terms of test AUC, accuracy and F-score. The results illustrated the ability of the proposed descriptor to identify compounds with very high activity against the biofilms.

Discussion

In recent years, the development of computational techniques that build models to accurately classify chemical compounds, has been an active area of research. Most of the best performing techniques in this area use chemical descriptors of compounds. Recently, connection-tables have become the most complete structural representation of chemical compounds [31]. In this paper we proposed a new chemical descriptor derived from the connection-table of compounds by PCA analysis, which transforms the two dimensional connection-table into a single dimensional subspace in terms of two vectors: one corresponding to atoms and the other to bonds for each compound. The RBF kernel function provides a measure of similarity between two compounds based on their atoms and bonds description, resulting into two kernel matrices. In the primal space of the weighted LS-SVM framework, each atom description is weighted with the bond description and the solution in dual space provides a way to integrate these two kernel matrices efficiently.

In order to select the best prediction model, we considered three different performance measures: accuracy, F-score and test AUC. Generally an AUC score between 0.6 and 0.7 indicates that the model has sufficient ability to discriminate positive and negative classes [32]. All discussed descriptors obtained AUC scores ranging between 0.6 and 0.7, whereas MACCS keys outperformed in the *Salmonella* and Path Keys in the *Pseudomonas* case study. While considering the prediction performance in terms of accuracy, the proposed approach significantly outperformed all other descriptors. But for highly class imbalanced data sets with many negative examples, a model can predict the value of the majority class for all predictions and achieve a high classification accuracy. In this context, additional measures such as F-score are required to evaluate a classifier. The F-score is often used as an aggregated performance score for the evaluation of algorithms, which considers both precision and recall. The F-score will be equal to 1 if the algorithm has perfect precision and recall. F-scores with values between 0 and 1 usually give a reasonable rank ordering of different algorithms/classifiers. Compared to other descriptors, the proposed weighted chemical descriptor significantly outperformed in terms of F-score in both case studies.

The results of the validation set illustrated the ability of the proposed descriptor to identify compounds with very high activity against the biofilms. It is noted that the other descriptors have missed the ability to identify active compounds at the lower IC50 cut-off values. However, with increasing cut-off values, MACCS on *Pseudomonas* outperformed all the other descriptors. The proposed model with validation data obtained the best performance at IC50 = 10 μM, similar to the test data. It is noted that for the lower IC50 cut-off values, the number of active compounds in the trained model were very small. Even with this small number of active compounds in the trained model, the weighted chemical descriptor captured enough information to distinguish active and inactive compounds. This indicates that an increase in the ratio of positive to negative training instances greatly influences the performance of the classifiers.

To the best of our knowledge, this is the first time in chemoinformatics that a prediction model is developed on a connection-table based chemical descriptors and a weighted LS-SVM classifier. We admit that the time needed for building the predictive model for the weighted LS-SVM classifier was expensive in terms of parameters optimization. In general, the use of connection-table based descriptor in combination with the weighted LS-SVM approach could be applied successfully in the area of of chemoinformatics to identify very active compounds in a given biological condition.

Conclusion

It is a challenge for the researcher in chemoinformatics to extract knowledge from the chemical structure. In our analysis, PCA was used to decompose the connection-table of each compound effectively to a low dimensional representation, as such defining a new structural descriptor of chemical compounds. Then a weighted LS-SVM approach was used to design a weighted chemical descriptor and to predict the biological activity of chemical compounds. The results illustrate that the obtained descriptor offers an improved model to identify very active compounds in a specific biological condition. The proposed machine learning technique could be applicable to any classification/prediction problem which is based on the molecular structure of compounds.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

MT has designed the weighted chemical descriptor, performed modeling, done validation of compounds on the trained model and drafted the paper. HP and JV have done anti-biofilm testing, manuscript co-writing and correction. EVDE and TTTT have synthesized the chemical compounds. MDM and XQ are the creator of database and the input of the new model. JAKS has participated in the design and implementation of framework. HP, MDM, JAKS and BDM helped draft the manuscript.

Acknowledgements

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Johan AK Suykens is professor at the Katholieke Universiteit Leuven, Belgium. MDM is is full professor at the Katholieke Universiteit Leuven, Belgium. HS is postdoctoral fellow of the FWO-Vlaanderen at KU Leuven (12F8515N). XQ doctoral student at the Katholieke Universiteit Leuven, Belgium. Research supported by: Research Council KU Leuven: GOA/10/09 MaNet, KUL PFV/10/016 SymBioSys, START 1, OT 09/052 Biomarker, several PhD/postdoc fellow grants; Industrial Research fund (IOF): IOF/HB/13/027 Logic Insulin, IOF: HB/12/022 Endometriosis; Flemish Government: FWO: PhD/postdoc grants, projects: G.0871.12N (Neural circuits), research community MLDM; IWT: PhD Grants; TBM-Logic Insulin, TBM Haplotyping, TBM Rectal Cancer, TBM IETA; Hercules Stichting: Hercules III PacBio RS;iMinds: SBO 2013; ArtD Instance;IMEC: phd grant;VLK van der Schueren: rectal cancer;VSC Tier 1: exome sequencing; Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); COST: Action BM1104: Mass Spectrometry Imaging, Action BM1006: NGS Data analysis network; CoE EF/05/006, IUAP DYSCO, FWO G.0377.12, ERC AdG A-DATADRIVE-B. MDM, HS and XQ are gratefull for support by the IWT Flanders under the grant agreement SBO NEMOA (IWT-SBO 120050), the European Commission's Seventh Framework Programme (FP7/2007–2013) under the grant agreement COATIM (Project no. 278425) and the KU Leuven (IDO/11/008; GOA 10/014). The scientific responsibility is assumed by its authors.

Author details

[1]KU Leuven, Department of Electrical Engineering (ESAT), STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics / iMinds Medical IT, Kasteelpark Arenberg 10, 3001 Leuven, Belgium. [2]Centre of Microbial and Plant Genetics (CMPG), Department of Microbial and Molecular Systems, KU Leuven, Kasteelpark Arenberg 20, box 2460, B-3001 Leuven, Belgium. [3]Laboratory for Organic & Microwave-Assisted Chemistry

(LOMAC), Department of Chemistry, KU Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium. [4]Laboratory for Biomolecular Modelling, Division of Biochemistry, Molecular and Structural Biology, Department of Chemistry, KU Leuven, Celestijnenlaan 200G, B-3001 Leuven, Belgium.

References

1. Geppert, H., Vogt, M., Bajorath, J.: Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. J Chem Inf Model **50**, 205–216 (2010)
2. Melville, J.L., Burke, E.K., Hirst, J.D.: Machine learning in virtual screening. Comb Chem High Throughput Screen **12**, 332–343 (2009)
3. Schwaighofer, A., Schroeter, T., Mika, S., Blanchard, G.: How wrong can we get?a review of machine learning approaches and error bars. Comb Chem High Throughput Screen **12**, 453–468 (2009)
4. Rogers, D., Brown, R., Hahn, M.: Using extended-connectivity finger prints with laplacian-modified bayesian analysis in high-throughput screening. J.Biomol.Screen **7**(10), 682–6 (2005)
5. Todeschini, R., Consonni, V.: Molecular Descriptors for Chemoinformatics. Wiley-VCH Verlag GmbH & Co. KGaA, ??? (1995)
6. Barnard, J.M., Downs, G.M.: Chemical fragment generation and clustering software. J Chem Inf Comput Sci **37**, 141–142 (1997)
7. Durant, J.L., Leland, B.A., Henry, D.R., Nourse, J.G.: Reoptimization of mdl keys for use in drug discovery. J Chem Inf Model **6**(42), 1273–1280 (2002)
8. Burden, F.R.: Molecular identification number for substructures search. J. Chem. Inf. Comput. Sci. **29**, 225–227 (1989)
9. Davies, D.: Understanding biofilm resistance to antibacterial agents. nature reviews. Drug Discovery **2**(2), 114–122 (2003)
10. Hall-Stoodley, L., Stoodley, P.: Evolving concepts in biofilm infections. Cellular Microbiology **11**(7), 1034–1043 (2009)
11. Romling, U., Balsalobre, C.: Biofilm infections, their resilience to therapy and innovative treatment strategies. Journal of internal medicine (272), 541–561 (2012)
12. Steenackers, H., Hermans, K., Vanderleyden, J., De Keersmaecker, S.: Salmonella biofilms: an overview on occurrence, regulation and eradication. Food Research International **2**(45) (2012)
13. Bjarnsholt, T., Tolker-Nielsen, T., Hoiby, N., Givskov, M.: Interference of pseudomonas aeruginosa signalling and biofilm formation for infection control. Expert Reviews in Molecular Medicine **12**, 11 (2010)
14. Landini, P., Antoniani, D., Burgess, J.G., Nijland, R.: Molecular mechanisms of compounds affecting bacterial biofilm formation and dispersal. Applied Microbiology and Biotechnology **3**(86), 813–823 (2010)
15. Qin, Z., Zhang, J., Xu, B., Chen, L., Wu, Y., Yang, X., Shen, X., Molin, S., Danchin, A., Jiang, H., Qu, D.: Structure-based discovery of inhibitors of the yycg histidine kinase: new chemical leads to combat staphylococcus epidermidis infections. BMC Microbiol **6**, 96 (2006)
16. Zeng, Z., Qian, L., Cao, L., Tan, H., Huang, Y., Xue, X., Shen, Y., Zhou, S.: Virtual screening for novel quorum sensing inhibitors to eradicate biofilm formation of pseudomonas aeruginosa. Appl Microbiol Biotechnol **79**, 119–126 (2008)
17. Thomas, M., De Brabanter, K., Suykens, J.A.K., De Moor, B.: Predicting breast cancer using an expression values weighted clinical classifier. BMC Bionformatics **15**, 1–17 (2014)
18. Steenackers, H., Ermolat'ev, D.S., Savaliya, B., de Weerdt, A., de Coster, D., van der Eycken, E., de Vos, D., Vanderleyden, J., de Keersmaecker, S.C.: Structure activity relationship of 4(5)-phenyl-2-amino-1h-imidazoles, n1-substituted 2-aminoimidazoles and imidazo[1,2-a]pyrimidinium salts as inhibitors of the biofilm formation by salmonella typhimurium and pseudomonas aeruginosa. J. Med. Chem. (54), 472–482 (2010)
19. Steenackers, H.P., Ermolat'ev, D.S., Savaliya, B., Weerdt, A.D., Coster, D.D., Shah, A., van der Eycken, E.V., de Vos, D.E., Vanderleyden, J., de Keersmaecker, S.C.: Structure-activity relationship of 2-hydroxy-2-aryl-2,3-dihydro-imidazo[1,2-a]pyrimidinium salts and 2n-substituted 4(5)-aryl-2-amino-1h-imidazoles as inhibitors of biofilm formation by salmonella typhimurium and pseudomonas aeruginosa. Bioorg. Med. Chem. (19), 3462–3473 (2011)
20. Ermolat'ev, D.S., Bariwal, J.B., Steenackers, H.P., de Keersmaecker, S.C., van der Eycken, E.V.: Concise and diversity-oriented route toward polysubstituted 2-aminoimidazole alkaloids and their analogues. Angew. Chem. Int. Ed. Engl. (49), 9465–9468 (2010)
21. Steenackers, H., Ermolat'ev, D., Trang, T., Savalia, B., de Weerdt, A., Shah, A., Vanderleyden, J., van der Eycken, E.: Microwave-assisted one-pot synthesis and anti-biofilm activity of 2-amino-1h-imidazole/triazole conjugates. Org. Biomol. Chem. (12), 3671–3678 (2014)
22. Steenackers, H., Dubey, A., Robijns, S., Ermolat'ev, D., Delattin, N., Dovgan, B., Girandon, L., Fröhlich, M., De Brucker, K., Cammue, B.P.A., Thevissen, K., Balzarini, J., Van der Eycken, E.V., J, V.: Evaluation of the toxicity of 5-aryl-2-aminoimidazole-based biofilm inhibitors against eukaryotic cell lines. Bone Cells and the Nematode Caenorhabditis elegans Molecules (19), 16707–16723 (2014)
23. Bohm, M., Sturzebecher, J., Klebe, G.: Three-dimensional quantitative structureactivity relationship analyses using comparative molecular field analysis and comparative molecular similarity indices analysis to elucidate selectivity differences of inhibitors binding to trypsin, thrombin, and factor xa. J. Med. Chem. **42**, 458–477 (1999)
24. Pearson, K.: On lines and planes of closest fit to systems of points in space. Philosophical Magazine 2 **11**, 559–572 (1901)
25. Suykens, J.A.K., Vandewalle, J.: Least squares support vector machine classifiers. Neural Processing Letters **9**, 293–300 (1999)
26. Suykens, J.A.K., Van Gestel, T., De Brabanter, J., De Moor, B., Vandewalle, J.: Least Squares Support Vector Machines. World Scientific, Singapore (2002)
27. De Keersmaecker, S.C., Varszegi, C., van Boxel, N., Habel, L.W., Metzger, K., Daniels, R., Marchal, K.,

Table 1: Table enumerating atoms: data source I

5.9800	-0.0000	-0.0000	Br	0	0	0	0	0	0
4.4000	-0.6600	0.8300	C	0	0	0	0	0	0
3.5400	-1.3500	-0.1900	C	0	0	0	0	0	0

Table 2: Table enumerating bonds: data source II

1	2	1	0
2	3	1	0

De Vos, D., Vanderleyden, J.: Chemical synthesis of (s)-4,5-dihydroxy-2,3-pentanedione, a bacterial signal molecule precursor, and validation of its activity in salmonella typhimurium. J. Biol. Chem. **20**(280), 19563–8 (2005)
28. Sambrook, J., Fritsch, E.F., Maniatis, T.: Molecular Cloning: A Laboratory Manual. 2nd ed.; Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, ??? (1989)
29. Gasteiger, J., Engel, T.: Chemoinformatics: A Textbook. WILEY-VCH GMbH and Co. KGaA, ??? (2003)
30. Morgan, H.J.: The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. J.Chem. Doc. **5**(2), 107–113 (1965)
31. Leach, A.R., Gillet, V.J.: Representation and Manipulation of 2D Molecular Structures, pp. 1–4. Springer, ??? (2007)
32. Šimundić, A.: Measures of diagnostic accuracy: basic definitions. Med Biol Sci. **4**(22), 61–5 (2008)

Figures

Figure 1: An overview of chemical descriptor formation from the connection-table of compounds. PCA is applied to the connection-table of each compounds to define a new structural descriptor in terms of two vectors. This results into two matrices: atoms vs. compounds and bonds vs. compounds. The weighted LS-SVM framework integrates these two vectors into a single vector named as weighted chemical descriptor and performs further prediction. LOO-CV is applied to select the optimal parameters.

Figure 2: *Salmonella*: Error Bar of Accuracy, test AUC and F-score on 30 iterations for different discussed descriptors, where x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score.

Figure 3: *Pseudomonas*: Error Bar of Accuracy, test AUC and F-score on 30 iterations for different discussed descriptors, where x-axis denotes the descriptors and y-axis denotes the test AUC/test accuracy/F-score

Tables

Additional Files

Supplementary File 1 — The analytical data of the 10 validation compounds

Table 3: Comparison of averaged prediction performances of different descriptors - *Salmonella* and *Pseudomonas*.

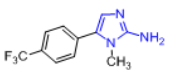
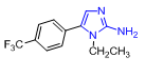
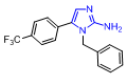
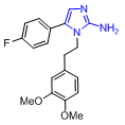
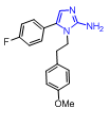
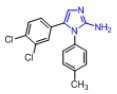
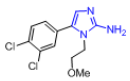
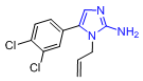
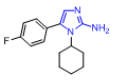
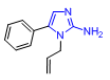
	Proposed Method	MACCS Keys	ECF	Path keys	BCUT
Salmonella					
Accuracy (std)	**0.8071(0.0581)**	0.7706(0.0500)	0.7686(0.0438)	0.7735(0.0454)	0.766(0.04)
p-value		0.0100	0.0100	0.0005	0.012
Test AUC (std)	0.6453(0.0409)	**0.6988(0.0294)**	0.6525(0.0263)	0.6425(0.0328)	0.665 (0.077)
p-value		6.73E-05	0.3528	0.2800	0.031
F-score (std)	**0.3058(0.0375)**	0.0212(0.0333)	0.0315(0.0241)	0.036(0.0142)	0.484(0.0823)
p-value		1.29E-16	5.52E-17	1.00E-17	0.064
Pseudomonas					
Accuracy (std)	**0.8179(0.0492)**	0.7908(0.0441)	0.7829(0.0360)	0.7882(0.0313)	0.65(0.055)
p-value		0.0001	0.0019	0.0004	0.012
Test AUC (std)	0.6549(0.0267)	0.7041(0.0258)	0.6559(0.0397)	**0.7118(0.0395))**	0.588(0.066)
p-value		5.06E-06	0.9104	3.30E-05	0.212
F-score (std)	**0.3277(0.0346)**	0.0211(0.0330)	0.1401(0.0854)	0.0874(0.0469)	0.743(0.059)
p-value		2.34E-17	1.38E-07	8.71E-14	0.024

Accuracy: (true positives + true negatives)/total population. **AUC**: probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.
F-score: 2* (precision*recall)/(precision+recall). **p-value**: two-sided sign test.

Table 4: Comparison of averaged prediction performances of different descriptors - *Thrombin*, *Trypsin* and *FactorXa*.

	BCUT	ECF	MACCS Keys	Path keys	Proposed Method
Thrombin					
Test AUC (std)	0.625(0.101)	0.673(0.012)	0.643(0.013)	0.5(0.0)	0.666(0.133)
Test Accuracy (std)	0.855(0.058)	0.784(0.21)	0.801(0.124)	0.610(0.070)	0.822(0.083)
F-Score(std)	NIL	NIL	NIL	NIL	0.600(0.050)
Trypsin					
Test AUC (std)	0.53(0.085)	0.625(0.091)	0.695(0.0)	0.500(0.0)	0.653(0.125)
Test Accuracy (std)	0.817(0.027)	0.764(0.042)	0.768(0.0)	0.166(0.125)	0.777(0.058)
F-Score (std)	0.2(0.03)	NIL	NIL	NIL	0.58(0.021)
FactorXa					
Test AUC (std)	0.594(0.078)	0.493(0.015)	0.571(0.022)	0.654(0.033)	0.582(0.099)
Test Accuracy(std)	0.727(0.0924)	0.766(0.023)	0.733(0.035)	0.694(0.047)	0.790(0.078)
F-Score (std)	0.835(0.070)	NIL	NIL	NIL	0.802(0.053)

Table 5: Compound Id, Structure and IC50 for biofilm prevention of 10 novel compounds

N^{o}	Structure	*Salmonella* IC50 in μM	*Pseudomonas* IC50 in μM
2- aminoimidazoles			
1		69.12(50.43 to 94.74)	211.90(142.2 to 315.9)
2		42.86(19.02 to 96.56)	60.21(27.25 to 133.0)
3		44.46(22.41 to 88.20)	greater than 400
4		115.1(65.28 to 203.0)	358.20(223.5 to 574.1)
5		48.62(36.67 to 64.45)	54.53(24.69 to 120.4)
6		53.79(21.01 to 137.7)	greater than 400
7		27.47(18.96 to 39.82)	74.74(43.41 to 128.7)
8		10.78(6.488 to 17.92)	13.92(7.851 to 24.69)
9		25.06(15.38 to 40.83)	28.70(16.04 to 51.37)
10		119.00(80.56 to 175.8)	15.92(2.630 to 96.32)

The values between brackets are the 95% confidence interval

Table 6: The compounds identified as active by each descriptor with different cut-off values for IC50 - *Salmonella* and *Pseudomonas*.

SALMONELLA

	Compounds identified	as active	(compound N^o)			
IC50 cut-off (μM)	10	20	30	40	50	100
Proposed Method	NIL	1,2,**9**,10	4,**7,9**,10	**3**,4,6,**7,9**,10	**3**,4,**6,7,9**,10	**3**,4,**5,6,7,8,9**
MACCS	NIL	NIL	NIL	**3,9**	**3,9**	**3,5,7,8,9**,10
ECF	NIL	NIL	**3,7,8**	**3,7,8**	**7,8**	**3,5,6,7,8**,10
Path	NIL	**9**	**9**	**9**	**9**	**1,3,5,7,8,9**

PSEUDOMONAS

	Compounds identified	as active	(compound N^o)			
IC50 cut-off (μM)	10	20	30	40	50	100
Proposed Method	**8,10**	6,7,**8**	3,6,7,**8**	3,6,7,**8**	3,4,6,7,**8,9**	1,**2**,3,4,**5**,6,**7,8,9**
MACCS	NIL	**8**,9	**8,10**	3,**8,9,10**	3,8,9,10	1,**2**,3,4,**5**,**7,8,9,10**
ECF	NIL	**9**	NIL	3,**10**	3,7,**10**	1,**2**,3,4,**5**,6,**7,8,9,10**
Path	NIL	**9**	**9**	**8,9**	3,**9**	1,**2**,3,4,**5**,6,**7,8,9,10**

NIL: If all 10 compounds are identified as inactive by the specific descriptor with corresponding IC50 cut-off

Table 7: Comparison of prediction performances of different descriptors - *Salmonella*.

	IC50	10	20	30	40	50	100
SALMONELLA							
Proposed Approach	AUC	0.500	0.722	0.691	0.548	0.548	0.625
	F-score	NIL	0.286	0.571	0.444	0.546	0.800
	accuracy	0.900	0.600	0.700	0.500	0.500	0.700
MACCS	AUC	0.500	0.500	0.500	0.700	0.643	0.625
	F-score	NIL	NIL	NIL	0.714	0.444	0.714
	accuracy	0.900	0.800	0.700	0.700	0.500	0.700
ECF	AUC	0.500	0.500	0.762	0.762	0.643	0.562
	F-score	NIL	NIL	0.667	0.667	0.444	0.769
	accuracy	0.900	0.800	0.800	0.800	0.500	0.600
Path	AUC	0.500	0.750	0.667	0.667	0.571	0.875
	F-score	NIL	0.200	0.500	0.500	0.250	0.857
	accuracy	0.900	0.900	0.800	0.600	0.400	0.800

NIL: If all 10 compounds are identified as inactive by the specific descriptor with corresponding IC50

Table 8: Comparison of prediction performances of different descriptors - *Pseudomonas*.

	IC50	10	20	30	40	50	100
PSEUDOMONAS							
Proposed Approach	AUC	1.000	0.625	0.4524	0.4524	0.4167	0.4167
	F-score	1.000	0.400	0.444	0.444	0.500	0.286
	accuracy	1.000	0.700	0.500	0.500	0.400	0.500
MACCS	AUC	0.500	0.687	0.833	0.929	0.792	0.625
	F-score	NIL	0.500	0.800	0.857	0.750	0.800
	accuracy	0.800	0.800	0.900	0.900	0.800	0.700
ECF	AUC	0.500	0.563	0.500	0.595	0.542	0.500
	F-score	NIL	0.363	NIL	0.400	0.545	0.750
	accuracy	0.800	0.700	0.700	0.700	0.500	0.600
Path	AUC	0.500	0.563	0.667	0.833	0.595	0.500
	F-score	NIL	0.364	0.500	0.800	0.400	0.571
	accuracy	0.800	0.700	0.800	0.900	0.600	0.600

NIL: If all 10 compounds are identified as inactive by the specific descriptor with corresponding IC50