Chemoinformatics approach to identify new compounds which inhibit biofilms formed by either Salmonella or Pseudomonas

Minta Thomas *[1], Johan AK Suykens[1]and Bart De Moor [1]

[1]ESAT-SCD / iMinds Future Health Department Katholieke Universiteit Leuven,3001, Leuven,Belgium.

Email: minta.thomas@esat.kuleuven.be;

*Corresponding author

Abstract

Background: A major difficulty in the control of *Salmonella* and *Pseudomonas* infections is the fact that these bacteria, like many other bacterial species, can form biofilm, in which they are protected against the activity of antibiotics, disinfectants or the immune system. Hence to prevent these infections, it is necessary to find compounds that are active against these biofilms. The most common representation of compounds is a connection-table, i.e., a table enumerating the atoms and a table enumerating the bonds. In this study, we intend to derive a new chemical descriptor from this table to distinguish chemical compounds which are active against biofilms.

Method: We propose a new machine learning approach for the identification of biofilm inhibitors of *Salmonella* Typhimurium and *Pseudomonas aeruginosa*. 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 is used in which a table enumerating the atoms is weighted against a table enumerating the bonds. We apply this framework to a given experimental data set on activity of 242 compounds against *Salmonella* and *Pseudomonas* biofilms. Prediction performances are evaluated and compared with other well-known chemical descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints and MACCSKeys.

Result: With the new chemical descriptor from the connection-table, the highest test areas under the receiver operating characteristic curves (AUC) were obtained, ranging from 0.8721 to 0.9890, 0.9198 to 0.9999 for *Salmonella* and *Pseudomonas* respectively.

Conclusion: The results suggest that the newly proposed chemical descriptor 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.

Introduction

It is generally accepted that biofilms are the predominant mode of bacterial growth, reflected in the observation that approximately 80% of all bacterial infections are related to biofilms (National Institutes of Health (USA)) [1, 2]. Biofilms are defined as structured communities of bacterial cells enclosed in a self-produced polymeric matrix adherent to inert or living surfaces [3–6]. Bacteria in biofilms are generally well protected against environmental stresses, antibiotics [7], disinfectants and the host immune system [8] and as a consequence are extremely difficult to eradicate [9]. Biofilm formation by *Pseudomonas aeruginosa* in the lungs of patients suffering from cystic fibrosis is a classic example of biofilm involvement in chronic infections. Because of the bacterial assembly in biofilms, this chronic infection is often noncurable and eventually results in the death of CF patients [10, 11]. Given the extent of problems caused by biofilms, there has been a significant effort to develop new anti-biofilm strategies [12, 13]. Traditionally biologists identify biofilm inhibitors by wet-laboratory experiments on 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. In the area of bacterial biofilm inhibition, chemoinformatics approaches have not been fully exploited yet. Molecular docking, a well established computational technique, was applied for the identification of potential histidine kinase inhibitors to combat *Staphylococcus epidermidis* [**?**] and for the selection of potential biofilm inhibitors of *Pseudomonas aeruginosa* [**?**].

Computational methods involving machine learning techniques are among the most popular tools used in cheminformatics tasks [14–16]. 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 [17]. Path-Length Fingerprints were generated with respect to path lengths of chemical compounds [18]. MACCS keys [19] are the sets of descriptors based on structural fragments, that have been identified a priori by domain expert [20].

The major aim of this paper is to introduce a new chemical descriptor based on the connection-table of compounds. Initially, molecular structures are converted into structural descriptors in terms of two vector spaces: atoms and bonds by principal component analysis (PCA). Then a weighted LS-SVM classifier is applied on these extracted features to establish a really useful statistical predictor for a biological system. Finally we compare the prediction

performance of the newly proposed descriptor with other well known descriptors such as Extended-Connectivity Fingerprints, Path-Length Fingerprints and MACCSKeys.

Data sets

In our analysis, we have a connection-table of 242 compounds that have the capability to inhibit biofilm formation of *Salmonella* or *Pseudomonas*. In addition, we have IC50 values which determine the effectiveness of these compounds against biofilms. A lower IC50s means more effective a compound against the biofilm.

Methods

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

Principal component analysis

Principal component analysis (PCA) [21] is mathematically defined as an orthogonal transformation that transforms the data to 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 is defined as follows:

$$A = Score * Coef^T$$

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

Least squares support vector machine classifiers

The constrained optimization problem for the least squares support vector machine (LS-SVM) [22, 23] 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 [22] 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.

Weighted LS-SVM classifier

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 $\varphi^{(2)}(.) : \mathbb{R}^m \to \mathbb{R}^{n_2}$ 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 [24] 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

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

with α_i are 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. An overview of the weighted LS-SVM classifier for predicting chemical compounds is given in Algorithm [24]:

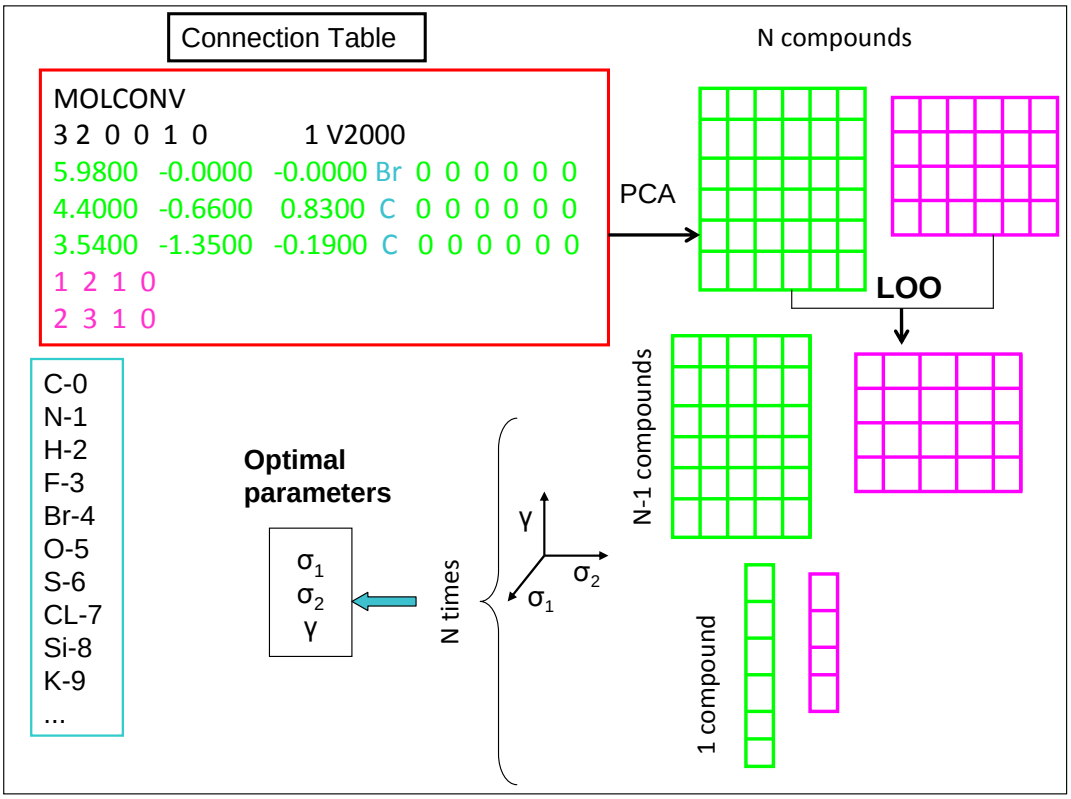

Figure 1: Overview of the weighted LS-SVM classifier model building strategy. PCA applied to the connection-table of each compounds to define a new structural descriptor in terms of two vectors . This resulted into two matrices: atoms vs. compounds and bonds vs. compounds. LOO-CV is applied to select the optimal parameters.

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).

The proposed weighted LS-SVM classifier is a machine learning approach in chemoinformatics for predicting the activity of a compound in a specific biological condition. In this study, the advantages of a weighted LS-SVM classifier were explored, by designing a prediction method for chemical compounds. The newly proposed chemical descriptors are represented in terms of two vectors: atoms and bonds. For each data sets, i.e., *Salmonella* Typhimurium and *Pseudomonas aeruginosa*, we compared the prediction performance of the new descriptors with other existing descriptors. The parameter selections for the weighted LS-SVM classifier are illustrated on Figure 1 and Algorithm. For several possible values of the kernel parameters σ_1 and σ_2, the 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.

Results

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 (SDL) 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 table contains non-numerical data, i.e. atom name, which is converted to a numerical representation as follows:

```
 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 formed two tables (refer Table 1 and Table 2) corresponding to the atoms and the bonds description

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

of the chemical compounds, respectively. Initially, PCA is applied to each of these tables individually. The first component which captures the majority of variance of the data (for each compound, the first component explaining minimum 70% variance) is selected for each table. Thus, we obtained two vectors, one corresponding to a table enumerating the atoms and other to a table enumerating bonds for each compound. Finally, we obtain two tables: compounds vs. atoms and compounds vs. bonds. The number of atoms and bonds are different for each compounds. The entries in the descriptor vectors are filled with zeros, if the compound does not have an atom or bond corresponding to an entry.

The weighted LS-SVM classifier is applied on these newly defined chemical descriptor. The weighting scheme in 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 optimal values of σ_1 and σ_2, LOO-CV performance is computed with each possible combination of γ. The combinations of the parameters which offered the highest LOO-CV performance were selected as the optimum parameters.

To evaluate the performance of the proposed method, we split the given data sets into a training and a test data set. The classification model is built on the training data set and the prediction performance of the method is evaluated with the test data set. For the case of *Pseudomonas*, we eliminated 97 compounds for which corresponding IC50 values are not available. We randomly splitted entire 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 non-active compounds in both training and test sets. Compounds are categorized into two sets: very active or less (not) active based on the IC50 values. The cut-off value, below which a compound was considered active, was calculated based on the cross validation approach proposed in [25]. The optimal cut-off can be determined by the

Table 3: Comparison of prediction performance of different descriptors in terms of test AUC.

Descriptor	*Salmonella*	*Pseudomonas*
Extended-Connectivity Fingerprints	0.8789(0.0663)	0.8324(0.0544)
MACCSKeys	0.8314(0.0480)	0.8862(0.0649)
Path-Length Fingerprints	0.8259(0.0567)	0.8868(0.0621)
proposed descriptor	**0.9278**(0.0468)	**0.9471**(0.0305)

desire to see as few false positives as possible in future prediction. Thus we have chosen the cut-off values as 150 and 50 for *Salmonella* and *Pseudomonas* respectively. That means, if $IC50 < 50$ of a compound, then it will be active against *Pseudomonas*. All the descriptors, mentioned in Table 3, achieved the best prediction performance with the selected cut-off values.

The prediction performance of the model is represented in terms of test areas under the receiver operating characteristic curves (AUC). On this experiment, LOO-areas under the receiver operating characteristic curves (AUC) were obtained when using the weighted LS-SVM classifier ranging from 0.8721 to 0.9890 and 0.9198 to 0.9999 for *Salmonella* and *Pseudomonas* respectively. Table 3 compares the prediction performance of the proposed descriptor with other well known chemical descriptors such as Extended-Connectivity Fingerprints, MACCS keys and Path-Length Fingerprints. The results prove that the newly proposed chemical descriptor improve the prediction performance on both case studies.

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 uses chemical descriptors of compounds. Recently, connection-tables have become as the most complete structural representation of chemical compounds. In this paper we proposed a new chemical descriptor from the connection-table of compounds in terms of two vectors: one corresponding to atoms and the other to bonds of each compound. RBF kernel function provides a measures of similarity between two compounds based on their atoms and bonds description, resulting into two kernel matrices. In primal space of the weighted LS-SVM framework, each atom description is weighted with bond description and the solution in dual space, provides a way to integrate these two kernel matrices efficiently.

The results showed that weighted LS-SVM with the proposed descriptor offers very good separation between

active and non-active compounds than LS-SVM classifier on Extended-Connectivity Fingerprints, Path-Length Fingerprints and MACCSKeys. This indicates that the proposed descriptor could effectively capture the structural property of compounds. To the best of our knowledge, this is the first time in chemoinformatics the prediction model is developed on the connection-table based chemical descriptors and weighted LS-SVM classifier.

We admit that the time needed for building predictive model for the weighted LS-SVM classifier was expensive in terms of parameters optimizations. In general, the use of connection-table based descriptor in combination with the weighted LS-SVM approach could applied successfully in the area of bacterial biofilm inhibition for identifying compounds which are active against biofilm.

Conclusion

It is a challenge for the researcher in chemoinformatics, to discover knowledge from the chemical structure information. 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 classifier for predicting the activity of chemical compounds. Results prove that a weighted LS-SVM classifier is an accurate prediction tool. The proposed machine learning approach could be applicable to any problem for which the activity of interest is prediction based on molecular structure.

Acknowledgements

BDM is full professor at the Katholieke Universiteit Leuven, Belgium. Johan AK Suykens is professor 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. The scientific responsibility is assumed by its authors.

References

1. Davies D: **Understanding biofilm resistance to antibacterial agents. Nature Reviews**. *Drug Discovery* 2003, **2**(2):114–122.
2. Hall-Stoodley L, Stoodley P: **Evolving concepts in biofilm infections**. *Cellular Microbiology* 2009, **11**(7):1034–1043.
3. Costerton JW, Stewart PS, Greenberg EP: **Bacterial biofilms: A common cause of persistent infections**. *Science* 1999, **5418**(284):1318–1322.
4. Donlan RM, Costerton JW: **Biofilms: survival mechanisms of clinically relevant microorganisms**. *Clinical Microbiology Reviews* 2002, **2**(15):167–193.
5. Hall-Stoodley L, Hu FZ, Gieseke A, Nistico L, Nguyen D, Hayes J, Forbes M, Greenberg DP, Dice B, Burrows A, Wackym PA, Stoodley P, Post JC, Ehrlich GD, Kerschner JE: **Direct detection of bacterial biofilms on the middle-ear mucosa of children with chronic otitis media**. *JAMA* 2006, **2**(296):202–211.
6. Homoe P, Bjarnsholt T, Wessman M, Sorensen HC, Johansen HK: **Morphological evidence of biofilm formation in Greenlanders with chronic suppurative otitis media**. *European Archives of Otorhinolaryngology* 2009, **10**(266):1533–1538.
7. Hoiby N, Bjarnsholt T, Givskov M, Molin S, Ciofu O: **Antibiotic resistance of bacterial biofilms**. *International Journal of Antimicrobial Agents* 2010, **4**(35):322–332.
8. Jensen PO, Givskov M, Bjarnsholt T, Moser C: **The immune system vs. Pseudomonas aeruginosa biofilms**. *FEMS Immunology and Medical Microbiology* 2010, **3**(59):292–305.
9. Burmolle M, Thomsen TR, Fazli M, Dige I, Christensen L, Homoe P, Tvede M, Nyvad B, Tolker-Nielsen T, Givskov M, Moser C, Kirketerp-Moller K, Johansen HK, Hoiby N, Jensen PO, Sorensen SJ, Bjarnsholt T: **Biofilms in chronic infections—a matter of opportunity—monospecies biofilms in multispecies infections**. *FEMS Immunology and Medical Microbiology* 2010, **3**(59):324–336.
10. Hassett DJ, Korfhagen TR, Irvin RT, Schurr MJ, Sauer K, Lau GW, Sutton MD, Yu H, Hoiby N: **Pseudomonas aeruginosa biofilm infections in cystic fibrosis: Insights into pathogenic processes and treatment strategies**. *Expert Opinion on Therapeutic Targets* 2010, **2**(14):117–130.
11. Bjarnsholt T, Jensen PO, Fiandaca MJ, Pedersen J, Hansen CR, Andersen CB, Pressler T, Givskov M, Hoiby N: **Pseudomonas aeruginosa biofilms in the respiratory tract of cystic fibrosis patients**. *Pediatric Pulmonology* 2009, **6**(44):547–558.
12. 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* 2010, **12**:e11.
13. Landini P, Antoniani D, Burgess JG, Nijland R: **Molecular mechanisms of compounds affecting bacterial biofilm formation and dispersal**. *Applied Microbiology and Biotechnology* 2010, **3**(86):813–823.
14. 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* 2010, **50**:205–216.
15. Melville JL, Burke EK, Hirst JD: **Machine learning in virtual screening**. *Comb Chem High Throughput Screen* 2009, **12**:332–343.
16. 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* 2009, **12**:453—-468.
17. Rogers D, Brown R, Hahn M: **Using extended-connectivity finger prints with laplacian-modified bayesian analysis in high-throughput screening**. *J.Biomol.Screen* 2005, **7**(10):682–6.
18. Todeschini R, Consonni V: *Molecular descriptors for chemoinformatics*. Wiley-VCH Verlag GmbH Co. KGaA 1995.
19. Barnard JM, Downs GM: **Chemical fragment generation and clustering software**. *J Chem Inf Comput Sci* 1997, **37**:141–142.
20. Durant JL, Leland BA, Henry DR, Nourse JG: **Reoptimization of mdl keys for use in drug discovery**. *J Chem Inf Model* 2002, **6**(42):1273–1280.
21. Pearson K: **On Lines and Planes of Closest Fit to Systems of Points in Space**. *Philosophical Magazine 2* 1901, **11**:559–572.

22. Suykens JAK, Vandewalle J: **Least Squares Support Vector Machine classifiers**. *Neural Processing Letters* 1999, **9**:293–300.

23. Suykens JAK, Van Gestel T, De Brabanter J, De Moor B, Vandewalle J: *Least Squares Support Vector Machines*. Singapore: World Scientific 2002.

24. Thomas M, De Brabanter K, Suykens JAK, De Moor B: **Predicting breast cancer using an expression values weighted clinical classifier**. In *Internal Report 13-39, ESAT-SISTA, K.U.Leuven (Leuven, Belgium)* 2013:1–17.

25. Langdon SR, Mulgrew J, Paolini GV, van Hoorn WP: **Predicting cytotoxicity from hetrogeneous data sources with Bayesian learning**. *Journal of Cheminformatics* 2010, (2):11.