In Search of Predictive Models for Inhibitors of 5-alpha Reductase 2 Based on the Integration of Bioactivity and Molecular Descriptors Data

Joana Sousa[1,2], Rui M. M. Brito[1,3], Jorge A. R. Salvador[1,2], and Cândida G. Silva[1,3⋆]

[1] Centre for Neuroscience and Cell Biology, University of Coimbra, 3004-517 Coimbra, Portugal
[2] Laboratory of Pharmaceutical Chemistry, Faculty of Pharmacy, University of Coimbra, Pólo das Ciências da Saúde, 3000-508 Coimbra, Portugal
[3] Chemistry Department, University of Coimbra, 3004-535 Coimbra, Portugal
candidasilva@qui.uc.pt

Abstract. 5-alpha reductase (5α-reductase) is a microsomal protein that converts testosterone into dihydrotestosterone (DHT). When changes occur in the function of this enzyme, disorders such as pseudohermaphroditism, baldness, benign prostatic hyperplasia and prostate cancer may arise. Currently, there are only two marketed drugs, finasteride and dutasteride, for the therapy of benign prostatic hyperplasia, which have long term side effects, stressing the need for the development of better inhibitors. In the present study, we used a dataset of compounds with known inhibitory activity against 5α-reductase (isozyme 2; 5α-R2) obtained from the ChEMBL database, and employed machine learning methods (random forests and support vector machines) to build classifiers for high-throughput virtual screening campaigns to help prioritise molecules for further analysis. The performance of the classification models was evaluated based on sensitivity, specificity, precision, F-score and accuracy. Our results show that, overall the classification models produced by the two algorithms present similar performance. Furthermore, the classifiers show high performance on the identification and discrimination between potent and weak inhibitors.

Keywords: Virtual screening, Machine learning, 5α-reductase, Classification, SVM, Random Forests

1 Introduction

The enzyme 5-alpha reductase (5α-reductase) is a microsomal protein that plays a central role in human sexual differentiation. This enzyme reduces the Δ^4-double bond of testosterone by using nicotinamide adenine dinucleotide phosphate (NADPH) as cofactor, affording the corresponding dihydrotestosterone

⋆ Corresponding author

(DHT), which is a more potent androgen [1]. There are two isozymes of 5α-reductase: the type 1 isoform (5α-R1) and the type 2 isoform (5α-R2). 5α-R1 is widely distributed, but it is highly expressed in subcutaneous glands of the skin and the liver. By contrast, 5α-R2 is prevalent in the prostate, genital skin, seminal vesicles, liver and epididymis [2]. Recently, a third isoform was identified and designated as type 3 isoform (5α-R3). This isozyme was originally identified in tissue of prostate cancer but was also found in other tissues, such as pancreas, brain, skin and adipose tissues [3].

Increased activity of these enzymes may cause diseases such as benign prostatic hyperplasia (BPH), prostate cancer, male-pattern baldness, acne and hirsutism [1, 4, 5]. The central role of α-reductase and their product DHT in these disorders has triggered the development of inhibitors of this enzyme, such as finasteride and dutasteride [4–6]. Finasteride (Fig. 1, right) is a 4-azasteroid which selectively inhibits 5α-R2, by blocking the conversion of testosterone to DHT to reduce stimulation of the prostate. On the other hand, dutasteride (Fig. 1, left) inhibits both 5α-R1 and 5α-R2 leading to a 95% decrease in DHT concentration and showing improved clinical results for patients with BPH. However, the use of these compounds in the therapy of prostate cancer remains controversial. Although the incidence of cancer was reduced by treatment with these inhibitors of 5α-reductase, in some patients more aggressive forms of cancer were detected when compared to patients treated with placebo [6]. For this reason, further studies for the development of new and more potent inhibitors of 5α-reductase are required.

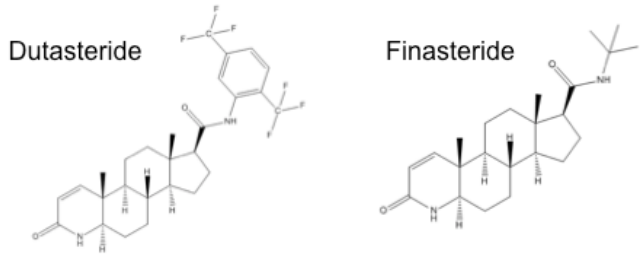

Fig. 1. Chemical structures of two steroidal inhibitors of 5α-reductase.

It is well recognised that the discovery of novel drugs in the pharmaceutical industry is becoming increasingly difficult, costly and time-consuming [7–9]. In the last decade, many approaches have been suggested to decrease the cost and time spent in the drug discovery process, such as the use of virtual, or *in silico*, screening methods to complement chemical and biological approaches. Virtual screening involves the computational filtering of a large number of molecules to identify those having a high probability of being active in the biological system of interest [7–9]. Thus, a virtual screening method takes as input all those molecules that might be acquired and tested, and then outputs those few that should be tested. When the three-dimensional structure of the biological tar-

get (determined either experimentally through X-ray crystallography or NMR or computationally through homology modelling) is available, it can be employed to design new ligands or to find ligands able to satisfy the structural requirements to form a protein-ligand complex [10]. However, this is not always the case, and in many situations the only information available is about known inhibitors for specific targets. In this scenario, approaches such as similarity and substructure searching, quantitative structure-activity relationships (QSAR), and pharmacophore and three dimensional shape matching can be applied [11, 12]. These methods work under the assumption that structurally related molecules are susceptible to present similar properties, in particular, display similar activity. In fact, given that the crystal structure of 5α-reductase isozymes remains unknown, mainly due to its instability during purification, the design of 5α-reductase inhibitors has been mostly based on the knowledge of the structure of known inhibitors, the enzyme mechanism and structure-activity relationships (SAR) information ([5] and references herein). For example, Kumar and co-workers have recently reported results on ligand-based 3D-QSAR studies using self-organising molecular field analysis on several steroidal 5α-reductase inhibitors to rationalise the molecular properties and their human 5α-reductase inhibitory activities [13–16].

The recent availability of public repositories, such as ChEMBL [17] and PubChem [18], containing both chemical structure and bioactivity information, opened unprecedented opportunities for the application of a myriad of machine learning methods to search for correlations between physico-chemical properties of bioactive molecules and their activity on specific target proteins [19, 20]. Recently, machine learning predictive models have been reported for different target proteins using methods such as random forests, support vector machines, Naïve Bayes and graph analysis, among others [21–24]. Here, we evaluate the application of two different machine learning methods – random forests and support vector machines – to build classifiers for high-throughput virtual screening campaigns to help prioritise molecules capable of inhibiting 5α-R2 for further analysis.

2 Materials and Methods

2.1 Data Set

ChEMBL is a database that congregates bioactivity values (IC_{50}, K_i, etc.) to millions of compounds on thousands of different therapeutic targets [17]. In ChEMBL, for 5α-R2, there are 793 values of bioactivity reported for 642 different compounds (Table 1). For the majority of the compounds, IC_{50} values were reported. All IC_{50} values were converted to nM.

Based on the information provided by ChEMBL, all bioactivity values for 5α-R2 were obtained from scientific literature. All scientific papers were retrieved based on the PubMed ID and DOI supplied by ChEMBL, and read to check for inconsistencies in the data. Two major inconsistencies were found. Compounds with IC_{50} values wrongly assigned and duplicated values. These values were

Protein	Bioactivity			# compounds	
	Total	IC_{50}	K_i	Total	Studied
5α-R2	793	466	102	642	354

Table 1. Description of the data set found for 5α-reductase (isozyme 2) in ChEMBL (accessed in December 2012) in terms of the number of bioactivity values (total, IC_{50}, and K_i) and number of compounds (total, and studied).

removed from the data set. Additionally, many compounds had multiple IC_{50} values reported, in a number between 2 and 22. In such cases, after analysis of the distribution of the values using box plots, the median value was assumed for each of these compounds. After these pre-processing steps, the data set studied for 5α-R2 was composed of 354 distinct compounds and corresponding IC_{50} values (Table 1).

Because we were interested in exploring the application of classification methods to prioritise compounds that inhibit 5α-R2, the IC_{50} values were converted to IC_{50} classes as presented in Table 2.

Class	IC_{50} interval	# compounds
Very Good	0 – 1	107
Good	1 – 10	48
Medium	10–100	48
Bad	> 100	151

Table 2. Definition of IC_{50} classes assignment.

2.2 Molecular Descriptors Generation

The chemical structures of the 354 compounds were downloaded from ChEMBL, and molecular descriptors were calculated using ChemAxon's [25] module `cxcalc`. These molecular descriptors are organised in several categories: elemental analysis, charge, conformation, geometry, isoforms, Markush enumerations, name, partitioning, predictor, protonation and others. For each compound in the data set, we calculated a total of 40 quantitative molecular descriptors selected from the different categories.

2.3 Classification Models

Generation of classification models and further analyses were performed using the KNIME suite of programs [26]. KNIME provides a graphical interface to

work the whole workflow of data analysis and integrates various components of machine learning, such as input pre-processing, cross validation, training and testing, and performance evaluation. Overall, there is a wide variety of machine learning methods to perform classification tasks [27], many of which implemented within KNIME. Here, we report the results obtained with two state-of-the-art classifiers namely support vector machines and random forest which were trained to build predictive models for 5α-R2 inhibitors. A general overview of these two algorithms is provided below. The statistical method of 5-fold cross-validation, using random sampling, was applied to allow the comparison of performance between the two methods.

Support Vector Machines (SVM). SVM is a classification method which is based on the construction of an hyperplane in a multidimensional space [28], allowing objects in different classes to be differentiated. The hyperplane is positioned using the set of training examples which are known as support vectors. The confidence level is given by the distance to the hyperplane: the greater the distance, the greater the confidence in the prediction. In recent years, SVM has been widely applied to build predictive models from libraries of known active and inactive compounds [29]. Then, each new compound can be mapped to the same spatial characteristics and its activity predicted according to which side of the hyperplane it will be found. Although, generally applied for binary classification problems, SVM has been generalised for multi-class problems [30]. Fig. 2 shows the KNIME workflow for the generation and analyses of the multi-class SVM classification models. First, the input data stored in an Excel file is read, after which each input feature vector is z-normalised. Cross-validation is performed using KNIME components X-Partitioner and X-Aggregator. The SVM model training is performed using the Weka LibSVM component [31], and testing is carried out using the Weka Predictor node. In the final step of the process, the component Scorer is employed to generate multiple performance metrics.

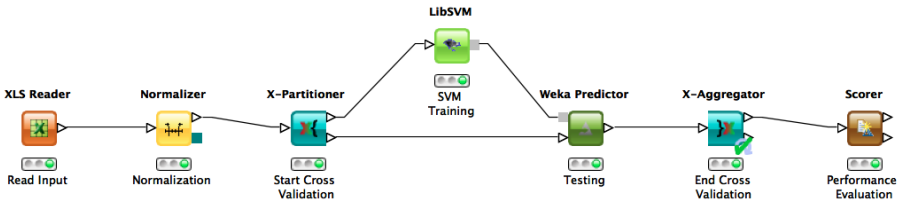

Fig. 2. KNIME workflow for the generation and analyses of the multiclass SVM classification models.

Random Forests. A decision tree defines a model for decisions and their possible consequences, including probabilities of outcomes, in a tree-like graph. From

the concept of decision tree, Breiman [32] formalised the concept of random forest. A random forest is a combination of decision trees, where each tree generated is used to classify a new object, and the final decision about the class to which the new object belongs is taken based on a majority vote. In decision trees, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees can be built by splitting the samples into subsets of samples of samples based on a certain variable. This process is then repeated on each derived subset of samples in a recursive manner [32]. The main advantages of this method are that it is fast to compute and the results are easy to interpret. Fig. 3 depicts the KNIME workflow for the generation and analyses of the random forest classification models. The major difference from the SVM workflow is that random forests, unlike SVM, are not dependent on data range or scale, and thus data normalization is not required.

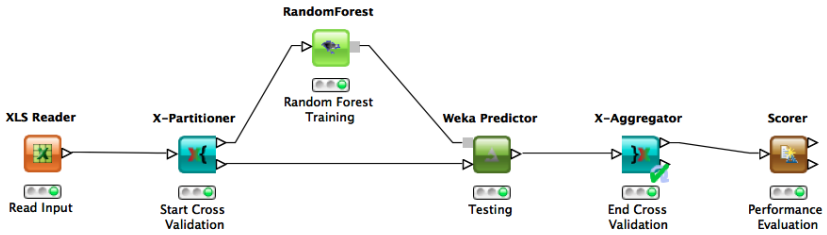

Fig. 3. KNIME workflow for the generation and analyses of the multiclass random forest classification models.

Performance Evaluation. Several performance measures were used to evaluate the classification models generated by the two algorithms. Sensitivity measures the level of positive correct prediction (Eq. 1), while specificity measures the proportion of negatives that are predicted correctly (Eq. 2). Additionally, accuracy (Eq. 5) accounts for the proximity of measurement results to the true value, whereas precision accounts for the reproducibility of the measurement (Eq. 3). The F score combines the precision measurement and sensitivity (Eq. 4).

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$F = \frac{2 \times precision \times sensitivity}{precision + sensitivity} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

where TP represents the number true positives, TN is the number true negatives, FP is the number of false positives, and FN the number of false negatives.

3 Results and Discussion

In this study, we used a data set of compounds with experimentally determined IC_{50} values for 5α-reductase 2 (5α-R2; CHEMBL1856) publicly available on ChEMBL database (accessed December 2012). A total of 793 different bioactivity values were available for 641 compounds, with 466 of these bioactivity values being IC_{50} values. Further analyses on the IC_{50} data revealed some inconsistencies primarily related with incorrectly assigned values between compounds reported by the authors and duplicated entries (the same value was presented in different units). Additionally, some compounds presented multiple IC_{50} values from which we selected the median value based on the analysis of values dispersion using box plots. The final data set included 354 compounds with a unique IC_{50} assigned (Table 1). We considered a multi-class classification problem by discretising the IC_{50} values into four distinct classes as described in Table 2. Our purpose was to explore different groups of compounds with very diverse IC_{50} values and characterised them with a large set of molecular descriptors, and check if this setting offered the discriminative power to correctly prioritise compounds for screening experiments.

IC_{50} Class	**SVM**				**Random Forests**			
	Sens.	**Spec.**	**Prec.**	**F**	**Sens.**	**Spec.**	**Prec.**	**F**
Very Good	86.9	86.1	93.9	86.5	85	85	93.5	85
Good	56.2	55.1	92.8	55.7	60.4	54.7	92.2	57.4
Medium	14.6	46.7	97.4	22.2	31.2	50	95.1	38.5
Bad	90.1	74.7	77.3	81.7	88.1	88.1	84.7	84.4
Accuracy	74.3				75.7			

Table 3. Evaluation of the classification models for 5α-R2 using 5-fold cross validation. The different performance measurements: sensitivity (Sens.), specificity (Spec.), precision (Prec.), F score (F) and accuracy are shown for SVM and random forests learning algorithms. All values are shown in percentage (%).

Classification models were trained and tested for two state-of-the-art machine learning algorithms – SVM and random forests –, using KNIME. All classification

models were trained based on 40 molecular descriptors and the corresponding IC_{50} class of 354 compounds. Both algorithms presented similar execution times. The performance of these algorithms in discriminating compounds belonging to different IC_{50} classes was evaluated using several metrics derived from 5-fold cross-validation (Table 3). Overall, the two algorithms present similar performance in distinguishing compounds from the different classes of IC_{50} (Accuracy $\simeq$ 75%). Furthermore, the classifiers reveal a better performance for compounds in IC_{50} classes Very Good and Bad.

Sensitivity and specificity were used to evaluate the classifiers' ability to correctly identify if a compound belongs or not to a particular class of IC_{50}. An optimal prediction is obtained when sensitivity and specificity are equal to 100%. For the IC_{50} classes of Very Good and Bad, the classifiers are very sensitive in their predictions with sensitivity between 85-90%, still the SVM classifier seems to be less specific for IC_{50} ($\simeq$ 75%). For the intermediate IC_{50} class Medium we observe the larger differences between the two algorithms, in particular in what concerns sensitivity – SVM (14.6%) and random forests (31.2%). In general, the performance of the classifiers for IC_{50} classes Good and Medium is less satisfactory as shown by the low values of sensitivity, specificity and F-score. These results may be partially justified by the lower number of compounds in these classes when compared to IC_{50} Very Good and Bad, which may have affected the training step. In fact, the total number of compounds in IC_{50} classes Good and Medium is half and one third the number of compounds from IC_{50} classes Very Good and Bad, respectively. However, it is also possible that compounds in the intermediate classes do not possess a group of characteristics (molecular descriptors) that might easily discriminate them.

4 Conclusions

The increasing number of diseases, such as prostate cancer and benign prostatic hyperplasia among others, mainly caused by disturbances of the function of isozyme 2 of 5α-reductase (5α-R2) triggered the development of inhibitors of this enzyme. However, the only two currently marketed inhibitors (finasteride and dutasteride) cause undesirable side effects, stressing the need to search for more potent and selective inhibitors.

The public access to bioactivity data of hundreds or thousands of chemical compounds offers the possibility to generate machine learning predictive models to screen molecules using *in silico* approaches. The aim is to employ the generated models to search large databases of chemical compounds and improve the identification of true hits. The methodology proposed here involves the development of machine learning workflows to generate classification models based on the integration of experimentally determined activity data and a large set of molecular descriptors for 5α-R2 inhibitors. We performed a comparison of performance of two classification algorithms – SVM and random forests – and concluded that both show an excellent performance in discriminating between compounds with very good and bad IC_{50} values.

Acknowledgments. This work is funded by ERDF – European Regional Development Fund through the COMPETE Programme (Operational Programme for Competitiveness) and by National Funds through FCT – Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) and projects PTDC/QUI-QUI/122900/2010 and Pest-C/SAU/LA0001/2013-2014. Exchange of ideas with Carlos J. V. Simões is acknowledge.

References

1. Andersson, S., Russell, D. W.: Structural and biochemical properties of cloned and expressed human and rat steroid 5 alpha-reductases. Proc. Natl. Acad. Sci. U. S. A. 87(10): 3640–3644 (1990)
2. Jin Y, Penning TM.: Steroid 5α-reductases and 3α-hydroxysteroid dehydrogenases: key enzymes in androgen metabolism. Best Pract. Res. Clin. Endocrinol. Metab. 5(1): 79–94 (2001)
3. Kapp, F. G., Sommer, A., Kiefer, T., Dolken, G., Haendler, B.: 5-alpha-reductase type I (SRD5A1) is up-regulated in non-small cell lung cancer but does not impact proliferation, cell cycle distribution or apoptosis. Cancer Cell Int. 12(1): 1–10 (2012)
4. Salvador, J. A. R., Carvalho, J. F. S., Neves, M. C., Silvestre, S. M., Leito, A. J., Silva, M. M. C., Sá e Melo, M. L. : Anticancer steroids: Linking Natural and Semi Synthetic Compounds. Nat. Prod. Rep. 30(2): 324–374 (2012)
5. Salvador, J. A.R., Pinto, R. M. A., Silvestre, S. M.: Steroidal 5α-reductase and 17α-hydroxylase/17,20-lyase (CYP17) inhibitors useful in the treatment of prostatic diseases. J. Steroid Biochem. Mol. Biol. 137: 199–222 (2013)
6. Amory, J. K., Anawalt, B. D., Matsumoto, A. M., Page, S. T., Bremner, W. J., Wang, C., Swerdloff, R. S., Clark, R. V.: The effect of 5alpha-reductase inhibition with dutasteride and finasteride on bone mineral density, serum lipoproteins, hemoglobin, prostate specific antigen and sexual function in healthy young men. J. Uro. 179(6): 2333–2338 (2008)
7. Guido, R. V. C., Oliva, G., Andricopulo, A. D.: Virtual screening and its integration with modern drug design technologies. Curr Med Chem. 15(1): 37–46 (2008)
8. Kar, S., Roy, K.: How far can virtual screening take us in drug discovery?. Expert Opin. Drug Discov. 8(3), 245–261 (2013)
9. Lavecchia, A., Di Giovanni, C.: Virtual Screening Strategies in Drug Discovery: A Critical Review. Curr. Med. Chem. 20(23): 2839-2860 (2013)
10. Kitchen, D.B., Decornez, H., Furr, J.R., Bajorath, J.: Docking and scoring in virtual screening for drug discovery: methods and applications. Nat. Rev. Drug Discov. 3(11), 935–949 (2004)
11. Chen, B., Harrison, R. F., Papadatos, G., Willett, P., Wood, D. J., Lewell, X. Q., Greenidge, P., et al.: Evaluation of machine-learning methods for ligand-based virtual screening. J. Computer-Aided Mol. Design 21(1–3): 53–62 (2007)
12. 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(2), 205–216 (2010)
13. Thareja, S., Aggarwal, S., Bhardwaj, T. R., Kumar, M.: Self organizing molecular field analysis on a series of human 5alpha-reductase inhibitors: unsaturated 3-carboxysteroid. Eur. J. Med. Chem. 44(12): 4920–4925 (2009)
14. Aggarwal, S., Thareja, S., Bhardwaj, T. R., Kumar, M.: Self-organizing molecular field analysis on pregnane derivatives as human steroidal 5α-reductase inhibitors, Steroids 75: 411–418 (2010)

15. Aggarwal, S., Thareja, S., Bhardwaj, T. R., Kumar, M.: 3D-QSAR studies on unsaturated 4-azasteroids as human 5alpha-reductase inhibitors: a self organizing molecular field analysis approach. Eur. J. Med. Chem. 45(2): 476–481 (2010)
16. Aggarwal, S., Thareja, S., Verma, A., Bhardwaj, T. R., Kumar, M.: QSAR studies on human 5alpha-reductase inhibitors: unsaturated 3-carboxysteroids. Acta Pol. Pharm. 68(3): 447-452 (2011)
17. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J. P.: ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. (Database issue): D1100–D1107 (2012)
18. Wang, Y., Xiao, J., Suzek, T. O., Zhang, J., Wang, J., Bryant, S. H.: PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res. 37(Web Server issue): W623–633 (2009)
19. Melville, J.L., Burke, E.K., Hirst, J.D.: Machine learning in virtual screening. Comb. Chem. High Throughput Screen. 12(4), 332–343 (2009)
20. Mitchell, J.B.: Informatics, machine learning and computational medicinal chemistry. Future Med. Chem. 3(4), 451–467 (2011)
21. Varnek, A., Baskin, I.: Machine learning methods for property prediction in chemoinformatics: Quo Vadis?. J. Chem. Inf. Model. 52(6),1413–1437 (2012)
22. Periwal, V., Kishtapuram, S., Open Source Drug Discovery Consortium, Scaria, V.: Computational models for in-vitro anti-tubercular activity of molecules based on high-throughput chemical biology screening datasets. BMC Pharmacol. 12: 1 (2012)
23. Jamal, S., Periwal, V., Open Source Drug Discovery Consortium, Scaria, V.: Predictive modeling of anti-malarial molecules inhibiting apicoplast formation. BMC Bioinformatics. 14: 55 (2013)
24. Lee, Y., Jana, S., Acharya, G., Lee, C. H.: Computational analysis and predictive modeling of polymorph descriptors. Chem. Cent. J. 7(1): 23 (2013)
25. ChemAxon - cheminformatics platforms and desktop applications, `http://www.chemaxon.com`
26. Berthold, M.R., Cebron, N., Dill, F., Gabriel, T.R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B.: KNIME: The Konstanz Information Miner. In: Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007), Springer (2007)
27. Kotsiantis, S. B.: Supervised Machine Learning: A Review of Classification Techniques. Informatica 31: 249–268 (2007)
28. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
29. Fukunishi, Y.: Structure-based drug screening and ligand-based drug screening with machine learning. Comb. Chem. High Throughput Screen. 12(4): 397–408 (2009)
30. Hsu, C.W., Lin, C.J.: A Comparison of Methods for Multiclass Support Vector Machines. IEEE Transactions on Neural Networks 13(2): 415–425 (2002)
31. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2(3): 27:1–27:27 (2011)
32. Breiman, L.: Random Forests. Machine Learning. 45(1): 5-32 (2001)