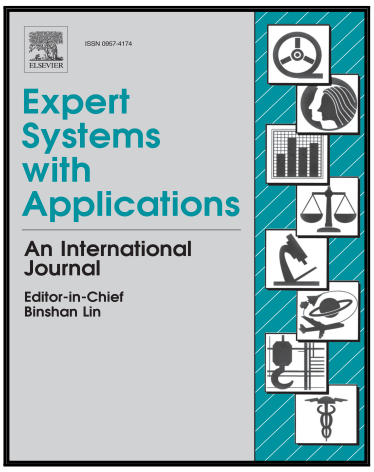

Automatic Selection of Molecular Descriptors using Random Forest: Application to Drug Discovery

Gaspar Cano, Jose Garcia-Rodriguez, Alberto Garcia-Garcia, Horacio Perez-Sanchez, Jón Atli Benediktsson, Anil Thapa, Alastair Barr

PII: S0957-4174(16)30681-9
DOI: 10.1016/j.eswa.2016.12.008
Reference: ESWA 11017

To appear in: *Expert Systems With Applications*

Received date: 5 July 2016
Revised date: 5 December 2016
Accepted date: 6 December 2016

Please cite this article as: Gaspar Cano, Jose Garcia-Rodriguez, Alberto Garcia-Garcia, Horacio Perez-Sanchez, Jón Atli Benediktsson, Anil Thapa, Alastair Barr, Automatic Selection of Molecular Descriptors using Random Forest: Application to Drug Discovery, *Expert Systems With Applications* (2016), doi: 10.1016/j.eswa.2016.12.008

Highlights

- Random Forest based approach to improve the selection of molecular descriptors
- Automatic features selection improves drug discovering methods accuracy
- Reduction of complexity and time requirements allows to explore larger datasets

Automatic Selection of Molecular Descriptors using Random Forest: Application to Drug Discovery

Gaspar Cano[a], Jose Garcia-Rodriguez[a,*], Alberto Garcia-Garcia[a], Horacio Perez-Sanchez[b], Jón Atli Benediktsson[c], Anil Thapa[d], Alastair Barr[e]

[a] *Department of Computers Technology, University of Alicante, Po Box 99, 03080, Alicante, Spain*
[b] *Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), Universidad Catolica San Antonio de Murcia (UCAM), Guadalupe 30107, Murcia, Spain*
[c] *Faculty of Electrical and Computer Engineering, University of Iceland, Smundargata 2, 101 Reykjavk, Iceland*
[d] *Computing Services, University of Iceland, Smundargata 2, 101 Reykjavk, Iceland*
[e] *Faculty of Science and Technology, University of Westminster, 115 New Cavendish Street, London W1W 6UW, United Kingdom*

Abstract

The optimal selection of chemical features (molecular descriptors) is an essential pre-processing step for the efficient application of computational intelligence techniques in virtual screening for identification of bioactive molecules in drug discovery. The selection of molecular descriptors has key influence in the accuracy of affinity prediction. In order to improve this prediction, we examined a Random Forest (RF)-based approach to automatically select molecular descriptors of training data for ligands of kinases, nuclear hormone receptors, and other enzymes. The reduction of features to use during prediction dramatically reduces the computing time over existing approaches and consequently permits the exploration of much larger sets of experimental data. To test the validity of the method, we compared the results of our approach with the ones obtained using manual feature selection in our previous study (Perez-Sanchez et al., 2014).The main novelty of this work in the field of drug discovery is the use of RF in two different ways: feature ranking and dimensionality reduction,

*Corresponding author: Tel.: +34 610488989; fax: +34 965903681
Email addresses: gcano@dtic.ua.es (Gaspar Cano), jgr@ua.es (Jose Garcia-Rodriguez), agarcia@dtic.ua.es (Alberto Garcia-Garcia), hperez@ucam.edu (Horacio Perez-Sanchez), benedikt@hi.is (Jón Atli Benediktsson), anilth@hi.is (Anil Thapa), A.Barr1@westminster.ac.uk (Alastair Barr)

and classification using the automatically selected feature subset. Our RF-based method outperforms classification results provided by Support Vector Machine (SVM) and Neural Networks (NN) approaches.

Keywords: Random Forest, Drug Discovery, Molecular Descriptors, Computational Chemistry

1. Introduction

Virtual screening methods are widely used nowadays in the drug discovery process (Zhao et al., 2013; Ma et al., 2011; Yan et al., 2014; London et al., 2014), where they provide with predictions about which ligands from large compound databases might bind to certain protein targets. Using this approach, it is possible to reduce the number of compounds that need to be tested experimentally in small labs or even when using High Throughput Screening infrastructures (Bajorath, 2002; Gong et al., 2010; Polgar & M Keseru, 2011; Tidten-Luksch et al., 2012; Mueller et al., 2012). Within virtual screening methods, one can find both Structure Based (SBVS) and Ligand Based (LBVS) methods. SBVS methods exploit information about the protein target and co-crystallized ligands (when available), while LBVS methods only exploit information about known ligands. Both SBVS and LBVS methods use different forms of scoring functions for affinity prediction and can complement high-throughput screening techniques; however, accurate prediction of binding affinity by any virtual screening method is a very challenging task. Use of modern computational intelligence techniques that do not impose a pre-determined scoring function has generated interest as a mean to improve prediction accuracy (Ain et al., 2015; Ballester & Mitchell, 2010). Selection of chemical characteristics (molecular descriptors) with greater discriminatory power has the potential to improve scoring predictions of which compounds will be good candidates, i.e., bioactive.

To improve the scoring of small molecules, it is necessary to carefully select the predictor variables which must help to decide among the different chosen input features (Guyon & Elisseeff, 2003). The set of features that describes

small molecules can be arbitrarily large, so that in most cases a pre-selection stage is required. The input variables (predictors) for a dataset are a fixed number of features, in our domain: the molecular descriptors. The values of these predictors can be binary, categorical, or continuous and represent the set of the system input data. The feature selection process consists of two main stages: acquisition of data (filtering, suitability, scaling) and feature selection. First, we should ask an important question: What are the most relevant features for our application domain? As we are working with standardized databases, we avoid steps for filtering, scaling, or deciding the suitability of this data. We will focus on the selection of features. There are different motivations for doing so, but we will seek to obtain a number of benefits (Guyon et al., 2006). In particular, we hope to get some of the following benefits:

- Reduction of the data to be processed.
- Reduction of features, reducing the cost of continued storage.
- Improved performance, improved processing speed can lead to an improvement in prediction accuracy.
- Improved display, improved representation helps the understanding of the problem.
- Reduced training time, smaller data subset decreases training time.
- Reduction of noise in the data, removing irrelevant or redundant features.

A proper selection of the set of molecular descriptors (predictors) is essential to optimize the prediction and automatic selection of these descriptors. This is a clear objective of automatic versus manual selection (ad hoc) methods. What are the most important variables in the classification models? This problem is common in many research domains. Usually, it is solved using the variable that best explains our model and adapts to the domain in which we work. For some domains, the segmentation criteria are simple or are constructed around artificial variables (dummy). These are the mechanisms that are adopted by a domain

expert and sometimes it is a multidisciplinary task. The use of computational intelligence techniques allows us to select these variables in an automatic way by quantifying their relative importance.

Once the idea of the relevance of the selected features is introduced, those not selected, or which have been left out, should be irrelevant or redundant. Therefore, the order of relevance allows us to extract a minimal subset of features that are enough to make an optimal prediction. In RF, the classification method is based on the use of decision trees on multiple samples of a dataset. RF has the ability to select a reduced set of candidates among a large number of input variables in our model (predictors) by finding linear relationships between them, this is what makes this method very interesting for this purpose.

In this paper we applied Random Forest as a feature selector but also as a classifier. We used public datasets to test the classification performance of the method. The main contribution of the paper is the automatic selection of a ranked and reduced subset of features to feed the classifier, enabling the system to obtain a good accuracy while dramatically reducing the computational cost thus allowing the system to explore large datasets. Our RF-based method outperforms manual selection of descriptors and improves classification results over SVM or NN approaches.

The rest of the paper is organized as follows: Section 2 describes the methodology, including the description of the public datasets employed to test the selection of variables. In addition, a computational intelligence method is introduced (RF). In Section 3, a set of experiments with RF to fit and model the automatic feature selection are presented. At last, in Section 4, a discussion of the results is presented and, finally, conclusions are drawn and some future works are listed.

2. Methodology

This section describes the pipeline, datasets, and methods we used to improve the selection of molecular descriptors. To apply the computational intelligence technique Random Forest to the selection of molecular descriptors, the

model was trained with different datasets that have been widely used by different virtual screening techniques. Automatic selection of variables was compared with data obtained by the manual selection (ad hoc) of combinations of these descriptors as tested in our previous study (Perez-Sanchez et al., 2014).

2.1. Method Pipeline

We propose a two stages method based on RF: in a first stage we trained the RF with databases of known active (drugs) and inactive compounds, to help to define the best descriptors for scoring/classification by providing the most relevant information in the classification step (Figure 1, 1-3) and improving the results of our previous work (Perez-Sanchez et al., 2014). This selection drastically reduces the computational complexity and time allowing to focus the computational effort on the proposed candidates which will permit to accelerate biomedical research. In a second stage, after the automatic selection of these molecular descriptors, we applied again a RF-based approach. This time RF is used as a classifier to determine the goodness of the selection to provide a prediction of a molecules activity (Figure 1, 4-6). Figure 1 shows the data flow from feature selection of the dataset to the classification step where the best results are measured in terms of AUC (Area Under the Curve) for each dataset. Accurate feature selection has the potential to improve system performance, processing speed, and can lead to an improvement in prediction accuracy.

2.2. Ligand Databases and Molecular Properties

In order to test our method, we compared results with our previous work using manual feature selection (Perez-Sanchez et al., 2014) employing standard VS benchmark tests, such as the Directory of Useful Decoys (DUD) (Huang et al., 2006), where VS methods's efficiency to discriminate ligands that are known to bind to a given target, from non-binders or decoys, is checked. Input data for each molecule of each set contains information about its molecular structure and whether it is active or not. We focused on three diverse DUD datasets (details are shown in Table 1)that cover kinases, nuclear hormone receptors and, other

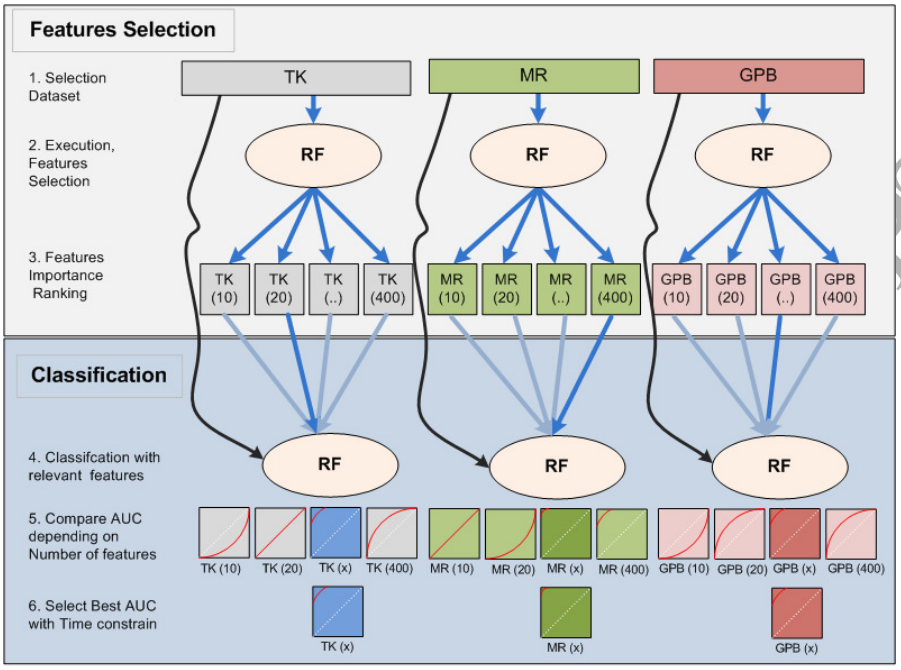

Figure 1: Data flow for automatic feature selection. In the feature selection step we feed the RF with a large set of different features from three public datasets (Table 1). The RF provided as a result a ranking of the features with the highest discriminative power for each dataset. In the classification step we train the RF using different sets of data represented by features obtained in previous selection step. The idea is to find the minimum set of features that achieves a good classification rate. We use the AUC for this purpose.

proteins such as TK, which corresponds to thymidine kinase (from PDB 1KIM (Champness et al., 1998)), MR, which corresponds to mineralocorticoid receptor (from PDB 2AA2 (Bledsoe et al., 2005)), and GPB, which corresponds to the enzyme glycogen phosphorylase (from PDB 1A8I (Gregoriou et al., 1998)).

Next, using the ChemoPy package (Dong-Sheng Cao, 2013) we calculated, for all ligands of the TK, MR and GPB sets, a set of diverse molecular properties derived from the set of constitutional, CPSA (charged partial surface area) and

Protein	PDB Code	Resolution (Å)	Ligands	Decoys
GPB	1A8I	1.8	52	1851
MR	2AA2	1.9	15	535
TK	1KIM	2.1	22	785

Table 1: Number of active (ligands) and inactive compounds (decoys) for each of the ligand datasets used in this study and obtained from DUD.

fragment/fingerprint-based descriptors, as described in (Perez-Sanchez et al., 2014).

2.3. Computational Intelligence Methods

The use of computational intelligence methods will allow us to provide a sufficient subset of features. Since the early 50s, computational intelligence research has focused on finding relationships between data and analyse these relationships (James, 2013). These problems are found in a wide variety of application domains: engineering, robotics or pattern recognition (Fukunaga, 1990), systems that recognize writing (Lee, 1999), voice (Huang et al., 2001), pictures (Young, 1994), sequencing genes (Liew et al., 2005), illness diagnostic (Berner & Lande, 2007) or spam rejection (Blanzieri & Bryl, 2008) are good examples.

Given a number of training data samples together with an expected output, the computational intelligence processes allow us to find the relationship between the pattern and the expected result, using that training data. The goal is to predict the unknown output for new data, e.g., test data. Training data is used for the optimal selection of these parameters, and different algorithms are used from a broad range of computational intelligence techniques. A classifier is a function that assigns to an unlabeled sample a label or class. A sample of several predefined categories or classes is classified. Classification models can be constructed using a variety of algorithms (Michie et al., 1994).

2.3.1. Random Forest

Random Forest (Breiman, 2001) (Figure 2) is a supervised learning method that can be applied to solve classification or regression problems. It is composed by a combination of tree predictors such that each tree depends on the values of a random vector independently and with the same layout for each of the generated vectors. Many disciplines use Random Forest: Accident analysis (Harb R, 2009), mechanical engineering (Longjun et al., 2011), financial engineering (Lariviere & Van den Poel, 2005; Xie et al., 2009), language models (Xu & Jelinek, 2007) or biology (Ding & Zhang, 2008). during the expansion of forest.

In Random Forest (Hastie, 2009), each individual tree is explored in a particular way:

1. Given a set of training data N, n random samples with repetition (Bootstrap) are taken as training set.
2. For each node of the tree, M input variables are determined, and $m << M$, variables are selected for each node. The most important variable randomly chosen is used as a node. The value of m remains constant
3. Each tree is developed to its maximum expansion.

The error of the set of trees depends on two factors:

- Correlation between any two trees in the forest, avoiding the use of a subset of variables randomly chosen data resampling (Bootstrap).
- A strong classifier, the importance of each tree in the forest, shows that with a low value of this error, the increase of these classifiers decreases the forest error.

2.3.2. Error Estimation

The OOB (out-of-bag) error is defined to estimate the classification or regression error in RF (James, 2013). It estimates a selection of the input observations based on Bagging (Breiman, 1996), (resampling of a random subset of predictors to be replaced in each tree). On average, each tree Bagging uses two-thirds of the

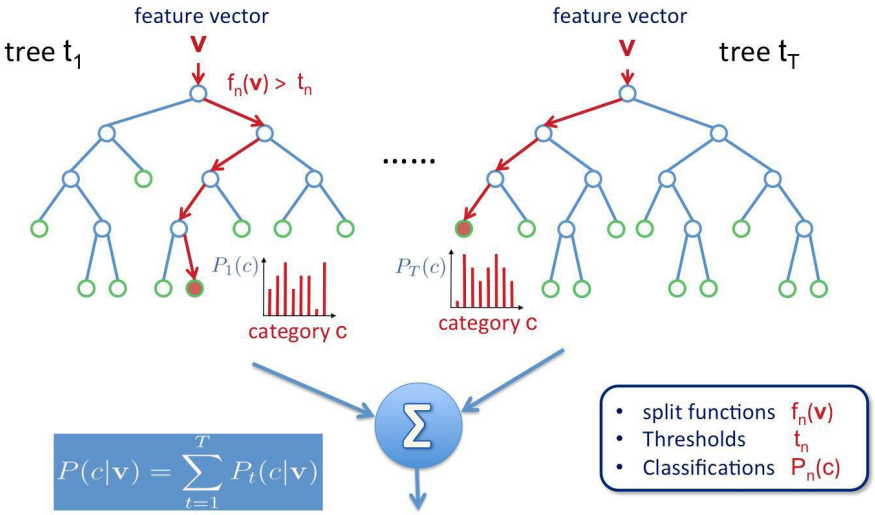

Figure 2: Random Forest is "a collection of classifiers that are structured as trees t_n where $F_n(v)$ are independent and identically distributed random vectors and each tree produces a vote of the most popular class for an input x (predictor)". The random vectors $P_n(c)$ represent a set of random numbers that determine the construction of each tree (Tae-Kyun, 2006).

observations, the remaining third will not be used in the comments off-exchange (OOB). So, you can predict the response to the i-th observation using each tree that will produce $B/3$ predictions for the observation i. In order to obtain a single prediction for the i-th element, we forecast based on the average of these responses (for regression) or by majority vote (for classification). This leads to a single OOB prediction for the i-th observation, which can be obtained in this way for each of the n observations. The sum of the OOB error and the average importance of all OOB trees determine the total and the relative importance of selected variables.

2.3.3. Importance of Variables

In Random Forest, a ranking of the contribution of each variable is determined to predict the output variable (Hastie, 2009), establishing a relative importance between them. This value is calculated using two different measures.

The first measure is the MDA (Mean Decrease Accuracy), which is based on the contribution of each variable to the prediction error (MSE for regression) and the percentage of misclassifications (for classification). The second measure of importance, the MDG (Mean Decrease Gini) from the Gini index, is the criterion used to select each partition in the construction of the trees. If a decrease of the error attributed to a variable occurs, its contribution will be lower for all trees.

For each tree t, we consider the error associated with a sample as OOB_t, $errOOB_t$ denoted as the error of a single tree t OOB_t sample. Randomly permuting the values of X_j in OOB_t to get a permuted sample and calculate their $errOOB_{tj}$, OOB_{tj} as predictor error on the permuted sample t. Thus express the importance of variables (VI) as:

$$VI(X_j) = \frac{1}{\text{ntree}} \sum_t (errOOB_{tj} - errOOB_t)\,.$$

A large value of VI indicates the importance of the predictor. By similarity, in the context of classification Bagging, we add the contribution of the Gini index and the decrease in each partition on a given as average for all predictor trees.

The Gini index measures the classification error committed in node t yet being this leaf, the class assigned randomly an instance, following the distribution of elements in each class in t. The Gini index for a node t can be calculated as:

$$i(t) = \sum_{i \neq j}^{c} P_i P_j = 1 - \sum_{j}^{c} P_j^2\,,$$

where c is the number of classes and P_i is the estimated probability of class i for instances that reach the node. Therefore, the Gini index and information gain are measures based on the impurity of each node.

3. Random Forest: Model Estimation

In any model of computational intelligence it is important to establish and determine the parameters that will enable us to adjust this model. In RF, the adequate number of trees must be determined, as well as how many predictors are used in the construction of each tree node. A reasonable strategy for accomplishing this is to set different values and evaluate the prediction error condition.

The model behavior is influenced by two parameters: the number of trees and the number of partitions to be made (splits). In this section, the influence and the optimal values for these parameters are analyzed. Experiments were developed using the RF implementation in the R package (R Core Team, 2013).

3.1. Number of Trees

Among the main parameters that can be set in RF, we can find the *ntree*, which sets the number of trees used in the model. We note that as the size of the tree grows in terms of number of nodes, their training accuracy improves until it stabilizes. For the three datasets, it can be estimated that the resulting error OOB is quite low for all cases. With a value of 300 trees *ntree*, the error remains stable. However, for a small number of trees it can be observed that this leads to an overfitting model on the training data in all the tested datasets (Figure 3).

3.2. Number of Splits

The other main parameter is *mtry*, which represents the number of input variables to be used in each node.

To construct each forest tree in RF, whenever a tree is divided it is considered a random sample of m predictors chosen from the complete set of p input predictors (molecular descriptors). These splits can choose only m predictors, usually the square root of the number of input predictors for classification and a third part of these predictors are used for regression.

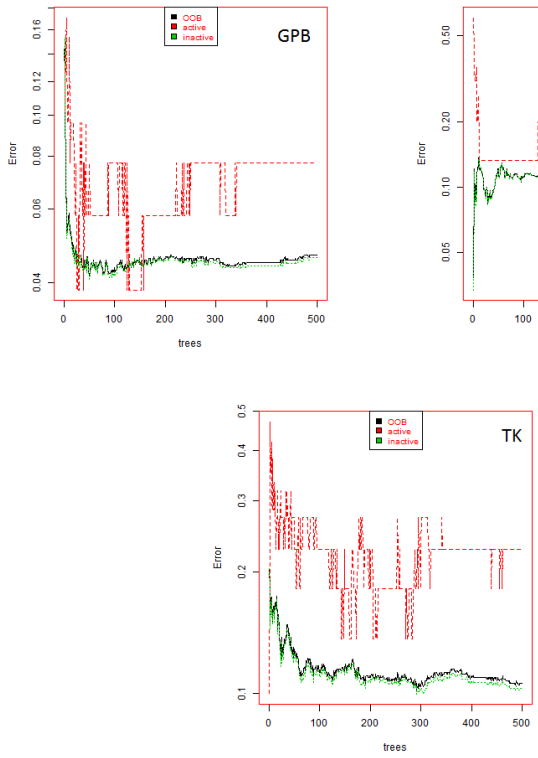

Figure 3: OOB error (black line), active misclassify (red line) and inactive misclassify (green line) vs. number of trees for the dataset GBP, MR and TK.

As we can see in the graph that estimates the minimum OOB error, the lowest error occurs when *mtry* takes values between 17 and 34 for GPB and MR data sets. A minimum value close to 0.013 is reached in the case of MR. We can set the value of *mtry* as the square root of the number of predictors, by default (Figure 4). We may also use a previous resampling featuring RF packet (TuneRF), estimating an optimal value for minimizing the OOB *mtry* error for each dataset.

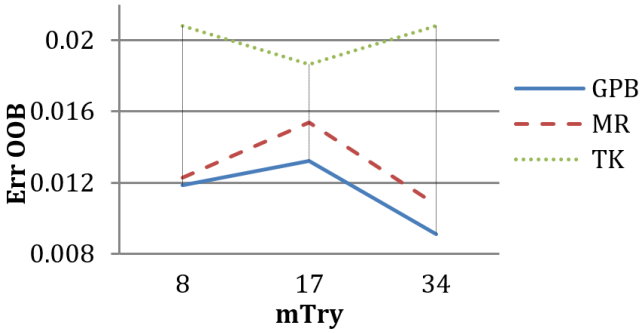

Figure 4: Relationship between OOB error and *mtry*.

3.3. Automatic Selection and Ranking of Features

The relative importance of the variables within each dataset determines the automatic selection of molecular descriptors used. In our experiments we can observe the input and differentiate these descriptors from the dataset.

For different molecular datasets and for each descriptor, we can observe the importance of the contribution to predict the model and determine the sensitivity with respect to the prediction of the final activity (Figure 5 and Table 2).

4. Results and Discussion

Random Forest selects automatically the molecular descriptors which allow to improve the goodness of the fitting process, considering that this selection of features depends on the dataset. We developed a set of experiments to test the validity of our method with an automatic selection of molecular descriptors. Furthermore, we compared it with the manual method (ad hoc) used in our previous work ((Perez-Sanchez et al., 2014)).

The selection of descriptors was performed according to the dataset, using Random Forest for the selection of variables, and then using RF, SVM and a MultiLayer Perceptron (NNET) for the classification of the previous selection. The AUC determines the goodness of the fitting for the prediction of the activity.

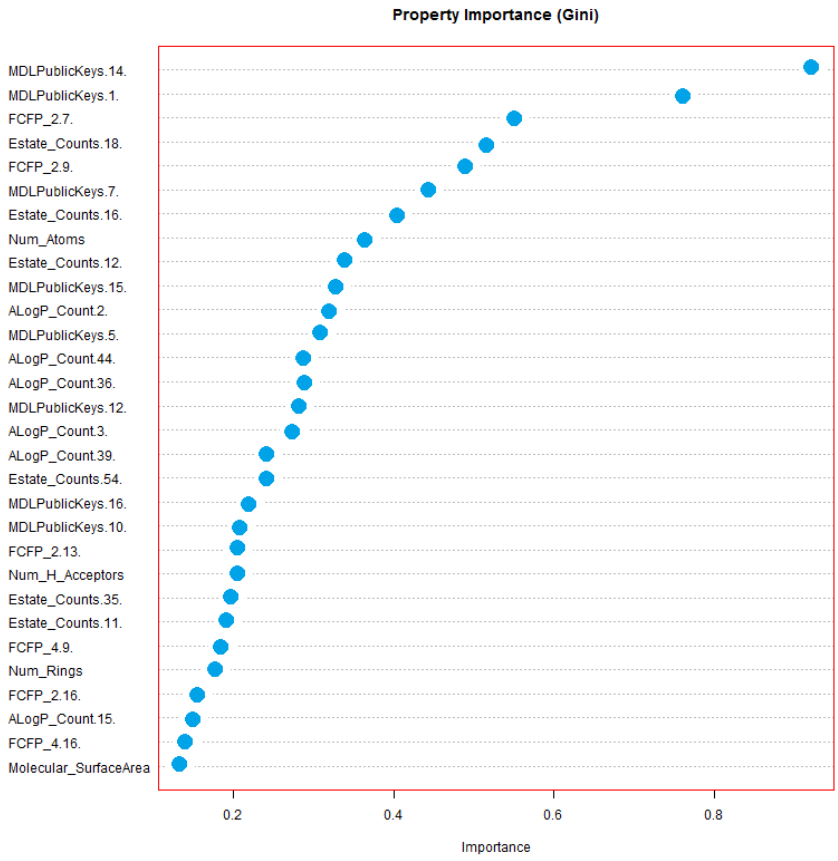

Figure 5: Relative importance of the predictors for the dataset MR.

Order	TK	MR	GPB
1	FCFP_2.12	MDLPublicKeys.14	Estate_Keys.13
2	ALogP_Count.48	Estate_Counts.18	ALogP_Count.56
3	MDLPublicKeys.12	MDLPublicKeys.1	ALogP_Count.8
4	Estate_Keys.34	Estate_Counts.16	Estate_Keys.34.
5	ECFP_4.5	MDLPublicKeys.7	Estate_Counts.34
6	ALogP_Count.56	ALogP_Count.3	Estate_Counts.13
7	Estate_Keys.9	Num_Rings	MDLPublicKeys.1
8	FCFP_4.12	MDLPublicKeys.15	ECFP_4.12
9	ALogP_Count.72	MDLPublicKeys.5	MDLPublicKeys.15
10	ECFP_6.1	FCFP_2.9	Num_H_Donors

Table 2: Top 10 molecular descriptors for dataset (ordered by relative importance.

In general terms, we observe that the number of significant variables (relative importance) predicting the final activity varies with the dataset. But in all cases with less than 10 features we obtain results over 0.9. In the worst case, the use of more than 80 features for TK does not improve the AUC. Furthermore, employing an accurate number of features saves time in the training stage and accelerates the whole process.

On the one hand, we show the results of the different classifiers depending on the feature subset size. From the experiments we observed that RF outperforms SVM and NNET in the three tested datasets. Another important conclusion that can be extracted from Figure 6 is that RF presents a decent results with only 4 features that is the minimum number that we have tested. On the contrary, SVM needs more than 10 features to obtain results over 0.9 AUC. RF shows a good stability and offers better results with a higher number of features but results with a few number of features are really good and demonstrate the good performance of RF to find the features with higher influence in the classification results. Unstable behaviour in SVM and NNET results could come from their inability to deal with datasets with high-dimensional data with a low

number of observations.

On the other hand. we presented the same data but comparing the performance of each classifier with the different datasets (Figure 7). All methods work fine with GPB with a low number of features. Datasets MR and TK present a more erratic behavior with SVM and NNET while RF works fine for all cases offering best results with GPB. While classifiers work fine with a large number of features, achieving results close to 1.0 AUC with MR and GPB, results with TK are slightly worse. The only dataset where SVM and NNET outperform RF using a large number of features, which means almost no feature selection, is MR.

The main conclusion of this study is that RF outperforms SVM and NNET using a minimum subset of relevant features (obtained with RF) producing considerably good results and saving time and resources compared with the other classifiers.

From the results obtained using this technique for variable selection, we can retrain the model with databases of known active or inactive compounds (Table 3). This information can be used to improve predictions and contribute to improved performance and acceleration in the discovery of new drugs using virtual screening techniques.

Descriptor	TK		MR		GPB	
Ad Hoc	NNET_EE246	0.94	NNET_EstCt	0.87	NNET_EAE246	0.96
Ad Hoc	SVM_AE246	0.95	SVM_EstKy	0.98	SVM_AlCnt	0.98
	BINDUSRF	0.70	BINDSURF	0.70	BINDSURF	0.68
Auto	C_RF_SVM	0.94	C_RF_ SVM	0.99	C_RF_SVM	0.99
Auto	C_RF_NNET	0.94	C_RF_NNET	0.99	C_RF_NNET	0.98
Auto	C_RF_RF	0.95	C_RF_RF	0.98	C_RF_RF	0.99

Table 3: Top values obtained for the AUC of the ROC curves for the DUD data sets TK, MK, GPB and BINDSURF processed by NNET, SVM using a manual selection of descriptor (Perez-Snchez et al., 2014) against automatic selection processed by RF.

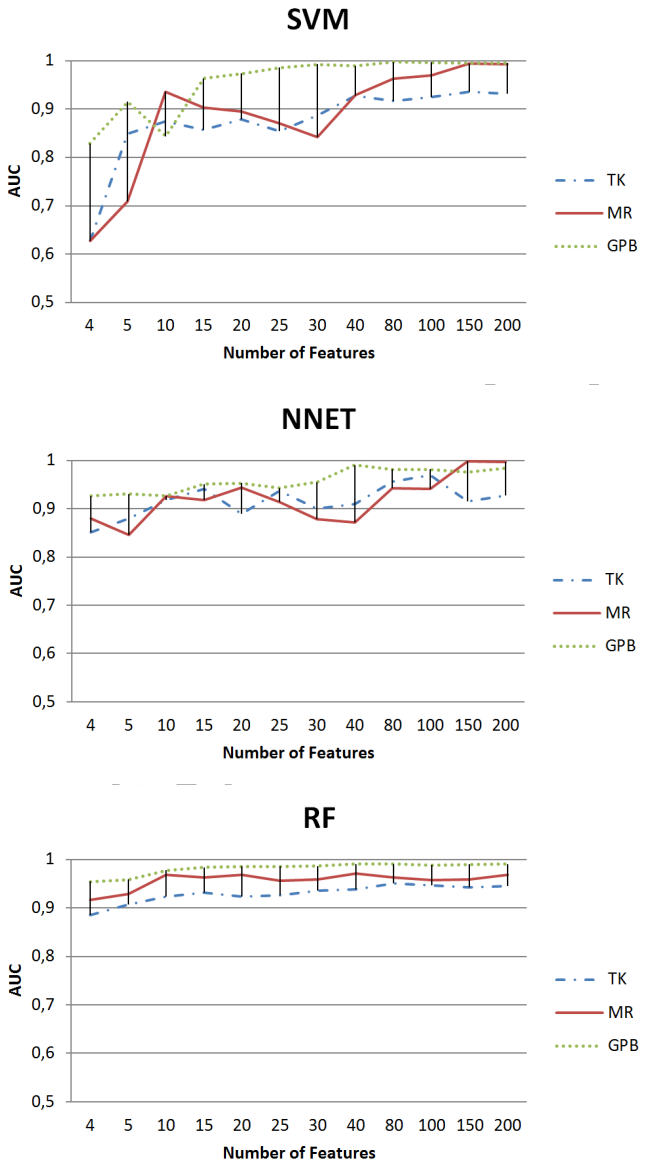

Figure 6: AUC vs Number of features (ordered by relative importance with RF) using SVM, NNET and RF as classifiers and applied to datasets TK, MR and GPB. Classifiers perspective.

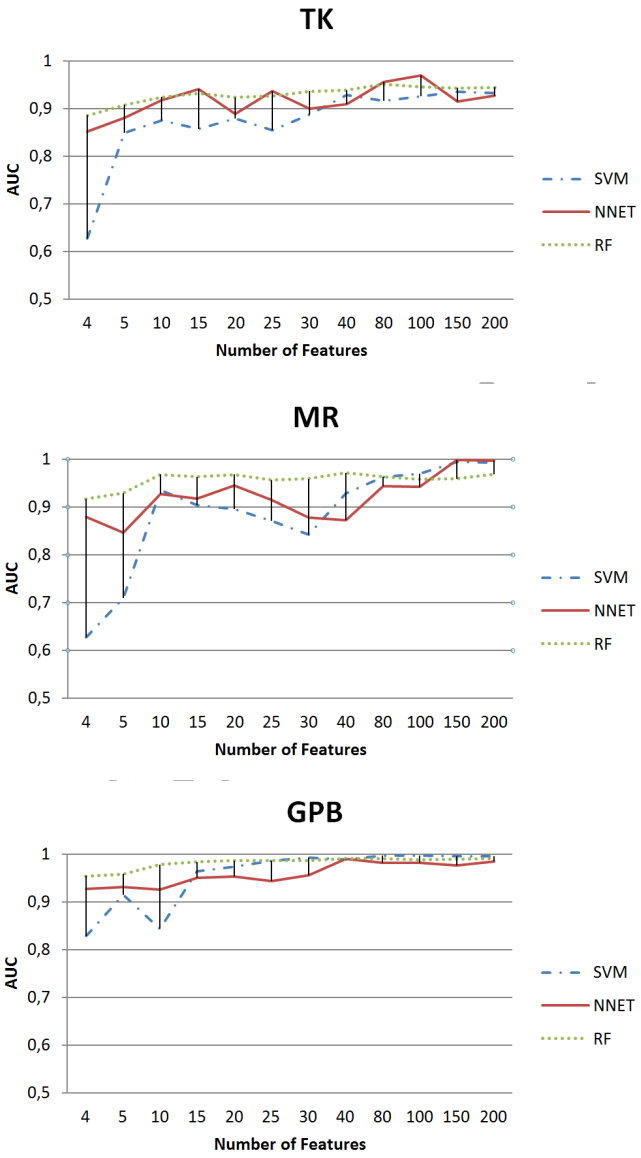

Figure 7: AUC vs Number of features (ordered by relative importance with RF) using SVM, NNET and RF as classifiers and applied to datasets TK, MR and GPB. Datasets perspective.

4.1. Discussion

We have presented aspects of the problem of automatic feature selection. This paper covers the challenges of feature selection through computational intelligence methods. In addition, we proposed a solution and an alternative to traditional manual selection of features (ad hoc), which requires a very precise knowledge of the scope of the domain, and sometimes the involvement of multiple disciplines or experts in the problem to predict.

The use of Random Forest eases the selection of molecular descriptors of the dataset, ensuring the best possible prediction of activity in an automated way. The use of this method for classification (the final prediction for the activity) improves the goodness of the fit.

Support Vector Machine is an effective classification method, but it does not directly obtain the feature importance. There have been some attempts to combine it with feature selection strategies but none of them improved Random Forest results for this task. Compared with SVM or neural networks, RF is able to estimate feature importance during training for little additional time. It is faster to train and has fewer parameters. The use of cross validation is unnecessary. Data does not need to be rescaled, transformed, or modified. It is resistant to outliers and is able to automatically handle missing values. And more importantly, it works better with large databases and a large number of features. Furthermore, RF is applicable to high-dimensional data with a low number of observations.

On the other hand, it can be extremely sensitive to small perturbations in the data: a slight change can result in a drastically different tree. Overfitting can be observed for some datasets with noisy classification/regression tasks. Finally, feature selection performed with Random Forest is sometimes difficult for humans to interpret.

5. Conclusions

In this work, we have proven the power of automatic selection of characteristics (molecular descriptors) using Random Forest, thus avoiding the manual selection of descriptors (ad hoc). The improvement on the prediction of the activity is explained by improving the goodness of the fitting and its value is expressed by the AUC of the Receiver Operating Charasteristic (ROC) curves.

We used RF for two purposes: feature ranking and dimensionality reduction, and classification using the automatically selected feature subset.

We have demonstrated empirically the abilitys of RF to determine the most relevant features by comparing the results with our previous work (Perez-Sanchez et al., 2014) that used ad-hoc feature selection and comparing RF with other relevant classifiers like SVM and Multilayer Perceptron. The use of Random Forest not only improves the accuracy of the classification methods selecting the most relevant features but also reduces the computational cost. This reduction combined with the use of parallel architectures allows the exploration of larger datasets in less time. Our RF-based method outperforms classification results provided by SVM and NN approaches.

However, it should be mentioned that the computational intelligence approaches could be used only when there are datasets available with active and inactive compounds. Given the good results obtained in terms of accuracy and computational resources reduction, it is concluded that this methodology can be used to improve the drug design and discovery, therefore helping considerably in biomedical research.

Future works include the automation of the choice of a learning algorithm depending of the characteristics of a given prediction problem, data source, and prediction performance. We also work on the creation of metaclassifiers that combine predictions of different classifiers. Despite the fact that our virtual screening method has already been parallelized, we are working on the GPU implementation of the whole pipeline. Finally, we are considering the application of this study to solve Quantitative Structure-Activity Relationship (QSAR)

problems.

Acknowledgements

This work was partially supported by the Fundación Séneca del Centro de Coordinación de la Investigación de la Región de Murcia under Project 18946/JLI/13. This work has been funded by the Nils Coordinated Mobility under grant 012-ABEL-CM-2014A, in part financed by the European Regional Development Fund (ERDF).

References

Ain, Q. U., Aleksandrova, A., Roessler, F. D., & Ballester, P. J. (2015). Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, *5*, 405–424.

Bajorath, J. (2002). Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, *1*, 882–894.

Ballester, P. J., & Mitchell, J. B. O. (2010). A machine learning approach to predicting proteinligand binding affinity with applications to molecular docking. *Bioinformatics*, *26*, 1169–1175.

Berner, E. S., & Lande, T. J. (2007). Clinical decision support systems: Theory and practice. chapter Overview of Clinical Decision Support Systems. (pp. 3–22). New York, NY: Springer New York.

Blanzieri, E., & Bryl, A. (2008). A survey of learning-based techniques of email spam filtering. *Artif. Intell. Rev.*, *29*, 63–92.

Bledsoe, R. K., Madauss, K. P., Holt, J. A., Apolito, C. J., Lambert, M. H., Pearce, K. H., Stanley, T. B., Stewart, E. L., Trump, R. P., Willson, T. M. et al. (2005). A ligand-mediated hydrogen bond network required for the

activation of the mineralocorticoid receptor. *Journal of Biological Chemistry*, *280*, 31283–31293.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32.

Champness, J. N., Bennett, M. S., Wien, F., Visse, R., Summers, W. C., Herdewijn, P., De Clercq, E., Ostrowski, T., Jarvest, R. L., & Sanderson, M. R. (1998). Exploring the active site of herpes simplex virus type-1 thymidine kinase by x-ray crystallography of complexes with aciclovir and other ligands. *Proteins: Structure, Function, and Bioinformatics*, *32*, 350–361.

Ding, Y.-S., & Zhang, T.-L. (2008). Using chou's pseudo amino acid composition to predict subcellular localization of apoptosis proteins: An approach with immune genetic algorithm-based ensemble classifier. *Pattern Recogn. Lett.*, *29*, 1887–1892.

Dong-Sheng Cao, Q.-N. H. Y.-Z. L., Qing-Song Xu (2013). Chemopy: freely available python package for computational biology and chemoinformatics. *Bioinformatics*, *doi:10.1093/bioinformatics/btt105*, 092–1094.

Fukunaga, K. (1990). *Introduction to Statistical Pattern Recognition (2Nd Ed.)*. San Diego, CA, USA: Academic Press Professional, Inc.

Gong, L.-L., Fang, L.-H., Peng, J.-H., Liu, A.-L., & Du, G.-H. (2010). Integration of virtual screening with high-throughput screening for the identification of novel Rho-kinase I inhibitors. *Journal of biotechnology*, *145*, 295–303.

Gregoriou, M., Noble, M. E., Watson, K. A., Garman, E. F., Johnson, L. N., Krulle, T. M., Fuetene, C. D. L., Fleet, G. W., & Oikonomakos, N. G. (1998). The structure of a glycogen phosphorylase glucopyranose spirohydantoin complex at 1.8 å resolution and 100 k: The role of the water structure and its contribution to binding. *Protein Science*, *7*, 915–927.

Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2006). *Feature Extraction: Foundations and Applications (Studies in Fuzziness and Soft Computing)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.

Harb R, R. E. S. X., Yan X (2009). Exploring precrash maneuvers using classification trees and random forests. *Accid Anal Prev. 2009*, *41*, 98–107.

Hastie, T. R. F. J., Trevor (2009). *The Elements of Statistical Learning*. Springer.

Huang, N., Shoichet, B. K., & Irwin, J. J. (2006). Benchmarking Sets for Molecular Docking. *J. Med. Chem.*, *49*, 6789–6801.

Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. (1st ed.). Upper Saddle River, NJ, USA: Prentice Hall PTR.

James, W. D. H. T. T. R., G. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag New York, Inc.

Lariviere, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert Syst. Appl.*, *29*, 472–484.

Lee, S.-W. (1999). Advances in handwriting recognition. *World Scientific*, *34*.

Liew, A. W.-C., Yan, H., & Yang, M. (2005). Pattern recognition techniques for the emerging field of bioinformatics: A review. *Pattern Recogn.*, *38*, 2055–2073.

London, N., Miller, R., Krishnan, S., Uchida, K., Irwin, J., Eidam, O., Gibold, L., Cimerman?i?, P., Bonnet, R., Shoichet, B., & Taunton, J. (2014). Covalent docking of large libraries for the discovery of chemical probes. *Nature chemical biology*, *10*, 1066–1072.

Longjun, D., Xibing, L., Ming, X., & Qiyue, L. (2011). Comparisons of random forest and support vector machine for predicting blasting vibration characteristic parameters. *Procedia Engineering*, *26*, 1772 – 1781. {ISMSSE2011}.

Ma, D.-L., Chan, D.-H., & Leung, C.-H. (2011). Molecular docking for virtual screening of natural product databases. *Chemical Science*, *2*, 1656–1665.

Michie, D., Spiegelhalter, D. J., Taylor, C. C., & Campbell, J. (Eds.) (1994). *Machine Learning, Neural and Statistical Classification*. Upper Saddle River, NJ, USA: Ellis Horwood.

Mueller, R., Dawson, E. S., Niswender, C. M., Butkiewicz, M., Hopkins, C. R., Weaver, C. D., Lindsley, C. W., Conn, P. J., & Meiler, J. (2012). Iterative experimental and virtual high-throughput screening identifies metabotropic glutamate receptor subtype 4 positive allosteric modulators. *Journal of Molecular Modeling*, *18*, 4437–4446.

Perez-Sanchez, H. E., Cano, G., & Garcia-Rodriguez, J. (2014). Improving drug discovery using hybrid softcomputing methods. *Appl. Soft Comput.*, *20*, 119–126.

Polgar, T., & M Keseru, G. (2011). Integration of virtual and high throughput screening in lead discovery settings. *Combinatorial Chemistry &# 38; High Throughput Screening*, *14*, 889–897.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

Tidten-Luksch, N., Grimaldi, R., Torrie, L. S., Frearson, J. A., Hunter, W. N., & Brenk, R. (2012). IspE inhibitors identified by a combination of in silico and in vitro high-throughput screening. *Plos One*, *7*, e35792.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Syst. Appl.*, *36*, 5445–5449.

Xu, P., & Jelinek, F. (2007). Random forests and the data sparseness problem in language modeling. *Computer Speech & Language*, *21*, 105–152.

Yan, C., Liu, D., Li, L., Wempe, M., Guin, S., Khanna, M., Meier, J., Hoffman, B., Owens, C., Wysoczynski, C., Nitz, M., Knabe, W., Ahmed, M., Brautigan, D., Paschal, B., Schwartz, M., Jones, D., Ross, D., Meroueh, S., & Theodorescu, D. (2014). Discovery and characterization of small molecules that target the gtpase ral. *Nature*, *515*, 443–447.

Young, T. Y. (Ed.) (1994). *Handbook of Pattern Recognition and Image Processing (Vol. 2): Computer Vision*. Orlando, FL, USA: Academic Press, Inc.

Zhao, S., Kumar, R., Sakai, A., Vetting, M., Wood, B., Brown, S., Bonanno, J., Hillerich, B., Seidel, R., Babbitt, P., Almo, S., Sweedler, J., Gerlt, J., Cronan, J., & Jacobson, M. (2013). Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*, *502*, 698–702.