Design of a Molecular Modeling Lab

BIOPHARMAC Project

Inês J Sousa, Miguel X Fernandes

Contents

Abbreviations/ Constants/ Biological data

3D: Three-dimensional.
ADMET: Absorption, Distribution, Metabolism, Excretion, Toxicity.
ANN: Artificial Neural Network.
AURKA: Aurora kinase A.
DUD: Directory of Useful Decoys.
eHiTS: electronic High Throughput Screening.
FN: False Negative.
FP: False Positive.
GI_{50}: Concentration that causes 50% growth inhibition.
IC_{50}: Concentration that causes 50% inhibition.
K_D: Dissociation constant.
LC_{50}: Concentration that kills 50% of a sample.
MW: Molecular Weight.
MASC: Multiple Active Site Corrections for Docking and Virtual Screening.
MRE: Mean Ranking Error.
nHAcc: Number of hydrogen bond acceptors.
nHDon: Number of hydrogen bond donors.
NPV: Negative Predictive Value.
OOB: Out-of-Bag.
PPV: Positive Predictive Value.
Q: Accuracy.
QSAR: Quantitative structure-activity relationship.
QSPR: Quantitative structure-property relationship.
R^2: Squared correlation coefficient.
R^2_{CV}: Squared cross-validation correlation coefficient.
RF: Random Forest.
RMSD: Root mean square deviation.
RMSE: Root Mean Squared Error.
SE: Sensibility.
SEE: Standard Error of Estimate.
SP: Specificity.
SVM: Support Vector Machine.
TGI: Concentration that causes total growth inhibition.
TN: True Negative.
TP: True Positive.
VIF: Variance Inflation Factor.
ΔG^0: Standard Gibbs free energy.

1. Introduction

The term molecular modeling may be defined as the mathematical description of molecules, generally using computational resources. Computational molecular modeling tackles numerical computation of molecular structures and molecular interactions (though chemical reactions are not generally addressed in molecular modeling studies applied to drug design). Molecular modeling is at a development stage that can be automated for implementation on a PC, though human input remains a must for results' interpretation and selection.

A major problem is the overwhelming amount of in available as almost every aspect of molecular structure and interaction can be described in a qualitative or approximate quantitative way. Molecular modeling is an extremely useful way to investigate topics that are too expensive or too time consuming to be investigated experimentally. It also helps scientists make predictions before running the actual experiments.

When using molecular modeling to get an answer for a pharmacological/clinical question, the immediate issue concerns to the software use. Then, we need to have some information regarding the quality (where we usually establish accuracy and efficiency) of the answer.

An example of prior information or knowledge needed follows

- What do you want to know?
- What is the desired accuracy in the results?
- How much time and computer power will it take?
- Which approximations are being made in the model systems?
- Which will be the prediction meaningfulness of the results?

When we get the answers to these questions, we can run calculations. Now we have to determine which piece of software is suitable, what it costs and how to use it.

During this report we present a series of computational methods and their typical uses.

2. Methodology

2.1. Docking

This methodology presents several applications, being the major application the identification of new active compounds for a particular target protein. Docking can also be used as a reliable and fast filter in high-throughput virtual screening, thereby providing a pool of ideas for novel lead structures. This methodology is also applied to identify a molecular target for a set of compounds and to correlate docking results with the biological activity when synthesis and experimental testing have been already performed. Docking studies the interactions between the compounds and the target under study to understand the mechanism of action of the compounds.

The data required to perform the docking studies is the tridimensional structure of compounds and targets. In case of experimental tests have been performed is possible to verify if there is a correlation between experimental results of biological activity and docking results.

From the docking studies is possible to obtain interaction energies between the compounds and targets, these energy values can be correlated with biological activity values obtained experimentally. Through this methodology is also possible to analyze the interactions established between compounds and targets, analyzing by this way the type of interactions involved and the conformations of compounds which give better interaction energies. When docking is used to perform virtual screening is possible to identify potential leads compounds from a large dataset based on the ranking of interaction energies obtained.

2.2. Quantitative Structure-Activity Relationship (QSAR)

QSAR studies are applied to determine which structural properties are important to the biological activity and can be used in order to improve the biological activity of the compounds based on the molecular descriptors that are correlated with compounds structure. This methodology can be applied in cases where the compounds haven´t yet been synthesized to perform a screening in chemical data banks or virtual libraries. QSAR is also used to establish correlations between structural properties and electronic properties of potential candidates to drugs and to predict the ADMET properties (absorption, distribution, metabolism, elimination and toxicity) or oral biodisponibility of compounds.

To apply this methodology is necessary several data as 3D structure of compounds, and experimental property values. To perform QSAR study is appropriate to use a set with a minimum of 30 compounds.

From QSAR study is obtained a QSAR model which provides a relationship between the property under study and molecular descriptors that influences the property, which are the most relevant descriptors and by this way is possible to know which structural characteristics is necessary increase or decrease to improve the property of compounds.

2.3. Artificial Neural Networks (ANN)

This methodology, as QSAR, is applied to establish quantitative relationships between structure and biological activity, and can be used in more complex models once this methodology is nonlinear.

To perform these studies is necessary 3D structure of compounds, and experimental property values. The set of compounds used should be a minimum of 30 compounds.

Through ANN study is obtained a nonlinear model that establishes a relationship between structure and activity. By the sensibility analysis is possible to find the significance of each descriptor for the model, and determine by this way which molecular descriptors have a higher influence in the biological activity.

2.4. Support Vector Machine (SVM)

Support vector machine is a computer algorithm that learns using a set of compounds to assign labels to objects.

The obtained information of this methodology is a classification of compounds in inactive or active for a specific target.

The SVM models are trained by using known active compounds and putative inactive compounds extracted from compound families that contain inactive compounds. To use this methodology is necessary a set of more than 20 structures, half active with known activity (experimental property values) and half inactive. Then, the obtained model is used to assign the compounds with unknown activity as active or inactive.

This methodology is used when we don't have the 3D structure of the target (protein) and only if we want to know if the compound is active or not, but this doesn´t give us the activity value.

2.5. Random Forest (RF)

Random forest (RF) is a computer algorithm that learns by assigning labels to objects.

From this methodology is obtained a classification of the compounds in inactive or active for a specific target.

To use this methodology we need active compounds and putative inactive compounds extracted from compound families and the compounds that you want to know if are active or not. Random forests method is applied to a minimum of 25 compounds.

This methodology is used when we don't have the 3D structure of the target (protein) and only if we want to know if the compound is active or not, but this doesn´t give us the value of activity.

2.6 Pharmacophore

Pharmacophore is the spatial arrangement of features that is essential for a molecule to interact with a specific target receptor. The candidate pharmacophore is computed by multiple flexible alignments of the input ligands.

At the final is obtained a file with pharmacophore model. This file contains the pharmacophore characteristics and the distance between them. This information can be used to perform virtual screening in a database. To run this method is necessary up to 32 structures.

The employed method is a ligand-based method that doesn´t require the structure of the target receptor. Instead, the input is a set of structures of drug-like molecules that are known to bind to the receptor. This methodology is applied to a broad range of different active structures to a specific target and the aim is to know which is the main characteristic of this set of structures.

2.7. Virtual Screening (VS)

Virtual screening (VS) has been extensively explored for facilitating lead discovery and for identifying agents of desirable pharmacokinetic and toxicological properties.

Using this methodology is obtained a small set of compounds, and these compounds are considered potential inhibitors or substrates.

To use virtual screening is necessary a pharmacophore model and a database where will be applied the model. This method is used, in a first approach, when the database has a huge number of compounds and the objective is to know which compounds are active or inactive for a specific target.

3. Results Processing and Statistics Application

3.1. Docking

3.1.1. Energy Distribution:

When the docking is applied as a screening to find from a set of compounds which are the best inhibitors of a specific target, the obtained interaction energies can be ranked according to their energies and represented in a graphic. From this representation is possible to identify the ligands with the best energies. Based on the accentuation of the slope for the negative section of the plot is possible to conclude if the interactions are favorable or unfavorable for the target under study.

3.1.2. Dissociation Constant:

The obtained energy values are the standard Gibbs free energy (ΔG^0), from this can be calculated the dissociation constant (K_D). Standard Gibbs free energy provides information about ligand-receptor complex stability, lower ΔG^0 indicates higher stability complexes. To obtain K_D values is used the following equations:

$$K_D = \frac{[R][L]}{[RL]} \qquad \Delta G^{\circ} = RT \ln K_D$$

K_D higher than micromolar indicates that the studies shouldn't be pursued.

3.1.3. Docking Results Correction:

The docking results can be corrected in order to increase the accuracy of ligands results and decrease the error values, avoiding for this way the occurrence of biases. To perform this correction, is calculated the standard deviation and the average of docking results for a ligand (*i*) that interact with all targets (*j*):

$$\mu_i = \sum_j (S_{ij})/N \text{.....} j = 1, N$$
$$(\sigma_i)^2 = \sum_j (S_{ij} - \mu_i)^2 / (N\text{-}1) \text{.....} j = 1, N$$

The correction of docking energies is made using the following equation:

$$S_{ij}' = (S_{ij} - \mu_i) / \sigma_i$$

Where S_{ij}' is the corrected energy for compound *i* in the active site *j* and S_{ij} is the docking energy, the corrected energy is named as multiple active site correction energy (MASC).

3.1.4. Mean Ranking Error (MRE):

MRE is calculated using the obtained docking energies S_{ij} for several ligands (*i*) in the active site (*j*) through the equation:

$$Err_j = (S_{jmelhor} - S_{jj}) / (S_{jmelhor} - S_{jpior})$$

S_{jj} is the energy to cognate ligand (*j*) in the active site (*j*), $S_{jmelhor}$ corresponds to the best energy for any ligand in the site (*j*) and S_{jpior} is the worst energy for any ligand. The difference between the best and the worst energy for any ligand provide the range of energies for all ligands in the active site and ($S_{jmelhor} - S_{jj}$) is the difference between the best obtained energy for the active site and the energy for the cognate ligand. This error can also be calculated using the average value Err_j in relation to all active sites (*j*). Error value of 0.0 means that energy values are perfect and a 0.5 value indicates that is an average value.

3.2. Targets Preferences:

Sometimes, the analysis of docking results shows that the targets under study have preference of interaction with some compounds. Based on this is possible to identify the scaffolds for these compounds that targets interact preferably.

3.2.1. Specificity Study:

To perform this study is necessary to select one enzyme family that acts either in healthy either in cancer cells (for example kinase family). Then, constant dissociation values of these kinases are compared with one kinase that only acts in cancer cells (for example Aurora kinase A (AURKA)). For this comparison are selected the ligands that show best docking energies interacting with the kinase that only acts in cancer cells, the docking is performed between these ligands and kinases that acts either in healthy either in cancer cells. The next step is the calculation of the ratio between constant dissociation values of kinases that acts in healthy and cancer cells and the kinase that acts specifically in cancer cells, determining for this way the specificity of compounds in relation to studied enzymes.

This type of studies is important to predict the occurrence of undesirable effects, the compounds should have higher specificity in relation to the desired target compared with the remaining targets.

3.3. Enrichment Test:

To perform this test was constructed a database of ligands to each target under study. The database is constituted by 1000 decoys and 20 ligands (known ligands and the best ligands determined by docking in relation to each target). The decoys were selected from a DUD (*Directory of Useful Decoys*) database based on their properties (molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, number of rotatable bonds and log *P*) which should be similar to the known ligands.

The docking is performed between each target and the compounds of database and the compounds were ranked according to their interaction energies, the results are represented graphically and the results are better when the line of results distribution is much more to the left of the graph.

3.4. Correlation between experimental and computational data:

When are available the experimental results of activity for the compounds studied in docking, is possible to verify the occurrence of correlation between experimental and computational data. Is very difficult to obtain correlation between these two types of data because computationally is only studied the interaction between the compounds and targets, and experimentally there are several factors that influence the results as compounds solubility and the diffusion to inside the cell. When are obtained correlations for the docking between the compounds and a specific target this can indicate that probably the compounds inhibit that target.

4. Quantitative Structure-Activity Relationship (QSAR)

4.1. Heuristic Method (HM):

Heuristic method is one of the methods used in QSAR and has two purposes, the molecular descriptors selection and to perform Multiple Linear Regression (MLR). This method checks if the values for each descriptor are available for each structure and if there is a variation in descriptors values. HM performs descriptors elimination discarding that satisfies one of the following conditions: (a) the descriptor isn't available for all structures, (b) descriptor value is constant for all structures. After this all possible one-parameter regression equations for each descriptor are calculated.

In order to reduce the number of molecular descriptors is applied the following criteria and the elimination of molecular descriptors is performed when: (a) the *F*-test value for each one-parameter with the descriptor is lower than 1.0; (b) the squared correlation coefficient to one-parameter equation is lower than R^2_{min} (0,01); (c) the parameter's *t*-test value is lower than t_1 (0,1); (d) the descriptor is highly intercorrelated with another descriptor and this other descriptor has a higher squared correlation coefficient in one-parameter equation based on these descriptors. All remaining descriptors are ranked according to the correlation coefficient of one-parameter equation.

HM is commonly used in QSAR linear studies and is an excellent tool to perform descriptors selection before the linear and nonlinear models construction. This method presents several advantages as the high speed of processing and the absence of software restrictions related to the dataset size and the unique strategy of variable selection.

This method provides a very fast and good estimation of expected correlation from dataset or gives several models with better regression. HM provides correlations 2 to 5 times more quickly than other methods with similar quality, and the maximum number of parameters involved in the resulting model can be fixed according to the specific situation saving time for this way. The disadvantage of this method is the fact that is limited to linear models.

4.2. Internal and External Validation:

Only the obtained models with squared correlation coefficient for the training set higher than 0.8 are considered. Then, these models are validated internal and externally. The internal validation is performed by cross-validation test, in this technique for each experimental data point, multi-linear regression is recalculated using the same descriptors for the data set without this point. The obtained regression equation is used to predict the value of this data point. The obtained array of predicted data points is linearly correlated with the array of experimental data points and the correlation coefficient of the last correlation is the cross-validated correlation coefficient. The cross-validated correlation coefficient is essentially a characteristic of the predictive power of the correlation equation and its value have to be higher than 0.6 to be considered.

The external validation of the obtained models was performed using the test set. The obtained models are used to predicted the activity of compounds from test set and the squared correlation coefficient obtained for this set indicates if the model is predictive, this value should be higher than 0.5 to be considered.

4.3. Examination of Multi-collinearity:

Multi-co-linearity is analyzed using Tolerance and Variance Inflation Factor (VIF) calculations. Tolerance is a measure of co-linearity and is calculated as $1\text{-}R^2$ for the each variable, small tolerance values to a variable means that this is a perfect linear combination of the independent variables already in the equation and this don´t should be added to the regression equation. All variable involved in the linear relationship have small tolerance.

VIF is a measure of the impact of co-linearity in the variables involved in the regression model. This parameter is calculated by 1/Tolerance and it is higher or equal to 1, when this value is higher than 10 is considered the presence of co-linearity.

http://www.researchconsultation.com/multicollinearity-regression-spss-collinearity-diagnostics-vif.asp

4.4. Statistical Parameters:

t and *F*-test values provide the information about the statistical significance of the model and molecular descriptors involved in the model. The obtained values for these parameters are compared with critical values, if *t*-test values are higher than critical values means that all descriptors involved in the model are statistically significant, if *F*-test value is higher than critical value it indicates that the model is statistically significant.

4.5. Error Parameters:

The Root Mean Square Error (RMSE) is the square root of the variance that is the standard error. The standard error of the estimate (SEE) is the measure of the accuracy of predictions made with a linear regression and it is a standard deviation of the errors. These parameters values should be very close to zero.

5. Artificial Neural Networks (ANN)

The Artificial Neural Networks are applied to obtained nonlinear relationships between structure and activity of compounds. Before the application of the ANN is necessary to select the molecular descriptors that will used, for this way have to be used other method to perform this selection, as Heuristic method.

From the obtained neural networks is possible to analyze their structure, see how many layers they have and how many neurons are in each layer. The obtained results of predicted values from neural networks can be plotted against experimental values for training and test set and for this way is obtained the squared correlation coefficient. As in QSAR the correlation coefficient for training set has to be higher than 0.8 and for test set higher than 0.5.

5.1. Sensibility analysis:

ANN doesn't allow an easy interpretation of descriptors contribution but through sensibility analysis is possible to determine the significance of each descriptor for the model. The sensibility analysis is the ratio between the neural network performance before and after of descriptor removal.

In cases that were applied QSAR method, ANN results can be compared with QSAR results, in order to understand which is the best method, linear or nonlinear, to establish the relationship between structure and activity.

6. Machine Learning Methods

6.1. Support Vector Machine (SVM)

6.1.1. Parameters of model validation:

After obtaining SVM results, the parameters for model validation are calculated. These parameters are: Accuracy, Sensibility, Specificity, Positive Predicted Value and Negative Predicted Value. This calculation is performed for each kernel obtained in each type of classification.

Accuracy (Q) is defined as a percentage of corrected classified cases of the set under study and is calculated using the following equation

$$Q = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

Sensibility (SE) is the probability of the model identify correctly the active set of compounds, this parameter is calculated based on the ratio between the number of active compounds classify correctly and the total number of active compounds.

$$SE = \frac{TP}{TP + FN} \times 100$$

Specificity (SP) is the probability of the model identify correctly the inactive set of compounds and it is calculated as the ratio between the number of inactive compounds classify correctly and the total number of inactive compounds.

$$SP = \frac{TN}{TN + FP} \times 100$$

The Positive Predicted Value (PPV) is the percentage of positive matches.

$$PPV = \frac{TP}{TP + FP} \times 100$$

The Negative Predicted Value (NPV) corresponds to the negative matches.

$$NPV = \frac{TN}{TN + FN} \times 100$$

The best model is chosen based on these parameters, is selected the model that presents the high accuracy, sensibility, specificity, positive predicted value and negative predicted value, and lower number of false negatives.

6.2. Random Forest (RF)

6.2.1. Parameters of model validation:

In RF Method is used the same parameters of validation as SVM. Additionally is used the *Out-of-Bag* (OOB) parameter that is provided by R-package in the application of RF Method. The OOB percentage indicates us about the prediction ability of the model, lower values of OOB means the better is the prediction ability of the model.

6.2.2. Variables Significance:

After the application of RF method is possible to construct a graphical representation of the influence of each descriptor in the mean decrease of accuracy and in the mean decrease of Gini. Based on this representation is possible to conclude about variables significance, higher decrease values caused by a specific variable more significant is this descriptor for the model.

The two methods, SVM and RF, can be compared in order to determine which is the most appropriate method in each case. This comparison is performed based on validation parameters values and number of false negatives for training and test sets.

7. Pharmacophore

7.1. Features:

The analysis of the obtained pharmacophore is performed by the identification of the features that constitute the pharmacophore and the distances between them. In these type of studies are obtained several pharmacophore models that are ranked according to their score, higher score values indicates higher similarity with pharmacophore used as reference. These results shows with features the compounds should have to be considered as inhibitors, substrates of the enzymes under study. Using the distances between the features is possible create simple or complex structures with the minimum requirements, it is also possible to perform virtual screening using the distances between the features.

8. Virtual Screening

8.1. Distribution:

After perform the virtual screening, we can plot the results distribution representing the number of hits against RMSD values. Based on this is possible to conclude about the type of the distribution, if it is a normal distribution or not, this is made using the calculated R^2. It is a normal distribution if R^2 is higher than 0.9, if R^2 is lower than 0.9 isn´t a normal distribution.

8.2. Filters Application:

In virtual screening are applied two filters, in the application of first filter only the structures that show RMSD higher than 0.1 are considered. To these structures are applied a second filter, the structures that don´t obey Lipinski rules are eliminated. These filters allow obtaining a very small set of structures in comparison with the non utilization of these two filters. The obtained structures show a high structural diversity.

9. Programs

Program	Application	Availability
eHiTS	Docking	1800 USD/year
CODESSA	QSAR	Free academic use
Vega ZZ	Quantum-chemical and Thermodynamic molecular descriptors calculation	Free
E-Dragon (Web server)	Molecular Descriptors Calculation	Free
FlexScreen	Docking	Free
HyperChem	Design and Optimization of molecular structures	995 USD
ChemBioOffice Standard 2010 Suite	Design of molecular structures	710 USD
Statistica	ANN and SVM	1050 USD
R-package	RF	Free
VLife	VS	??
PharmaGist (Web server)	Pharmacophore	Free