MOL2NET, International Conference Series on Multidisciplinary Sciences
http://sciforum.net/conference/mol2net-03

A new simplex machine-learning approach for analysis of structural chemical diversification processes. Comparison with other molecular modeling methods

Abir SARRAJ-LAABIDI[1,2], Asma HAMMAMI-SEMMAR[3], Nabil SEMMAR[1,*]

Affiliations:

1 Université de Tunis El Manar, Institut Pasteur de Tunis, Laboratory of Bioinformatics, Biomathematics and Biostatistics (BIMS), 1002, Tunis, Tunisia

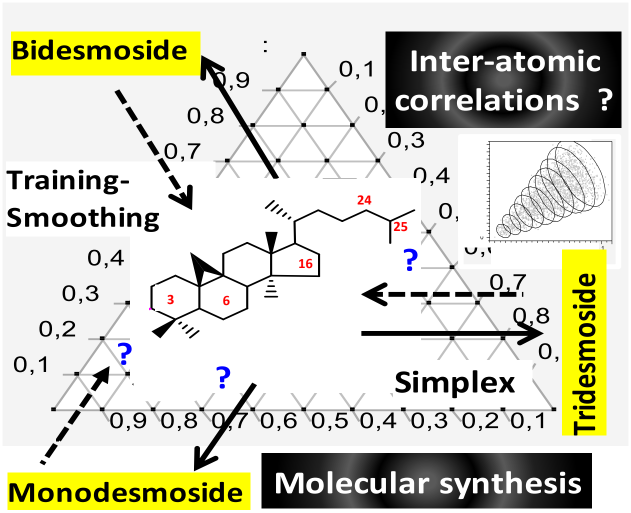
2 Université de Tunis El Manar, Faculté des Sciences de Tunis, Campus Universitaire, 2092 Tunis, Tunisia

3 Université de Carthage, National Institute of Applied Sciences and Technology (INSAT), 1080, Tunis, Tunisia

* **For correspondence**: nabilsemmar5@gmail.com

Keywords:
Computational chemistry, cycloartane, glycosylation, machine-learning, molecular ramification, saponins, simulation.

Graphical Abstract



Abstract.
Metabolism represents highly organized system characterized by strong regulations satisfying the mass conservation principle. In this work, a new simplex-based simulation approach was developed to learn scaffold information on metabolic processes controlling molecular diversity from a wide set of observed chemical structures. This approach is based on iterative in silico combinations of molecular profiles using Scheffé's mixture design. It was illustrated by cycloartane-based saponins of *Astragalus* genus containing one, two or three ramification chains with variable relative glycosylation levels. Competing and sequential glycosylation processes of different carbons were highlighted by the machine-learning simplex method. Comparisons between this simplex approach and other molecular modeling approaches were made to highlight advantages and limits of the new one.

Introduction

Molecules represent highly organized systems governed by inter-atomic links and interactions. Beyond carbon-carbon links forming the molecular backbone (aglycone or skeleton), chemical substitutions represent inter-molecular and inter-atomic processes responsible for multiway structural diversification in a wide metabolic system. Such a diversification is strongly governed by mass conservation principle under which a whole resource is shared between different component of systems leading to negative and positive correlations between elements belonging to different and same regulation ways, respectively. To highlight such variation trends with their shape, a new simplex-based machine-learning approach was developed and applied on chemical structural system after classifying molecules into different groups and decomposing them into different constitutive parts (components) [1].

Method

By reference to mass conservation, at inter-molecular scale, the biosynthesis of a specific molecule is carried out at the expense of others molecules. At intra-molecular scale, the chemical substitutions represent a limited resource for which a set of carbons (constituting a molecule) could enter in competition. Such inter- and intra- molecular dependences in a mixture system are governed by simplex rule: in a simplex space with *q-1* dimension, *q* components (*q* separated groups) vary the one relatively to the other under unit sum constraint representing limited resource in the whole system (**Fig.1**).

In this work, molecular clusters consisted of three desmosylation levels making *q*=3 constitutive groups of the simplex system. These molecular clusters were composed by 72 *Astragalus* saponins based on cycloartane with aliphatic lateral chain and characterized by three desmosylation levels representing a control factor with three target modalities (**Fig.1a-c**) and each molecule in these *q* groups is characterized by an individual profile of *p* additive variables with relative values; these additive variables are the relative glycosylation levels of carbons in the molecule. So, in our case, desmosylation (*D*) and glycosylation (*G*) represent two metabolic variability factors at inter-molecular and intra-molecular scales, respectively.

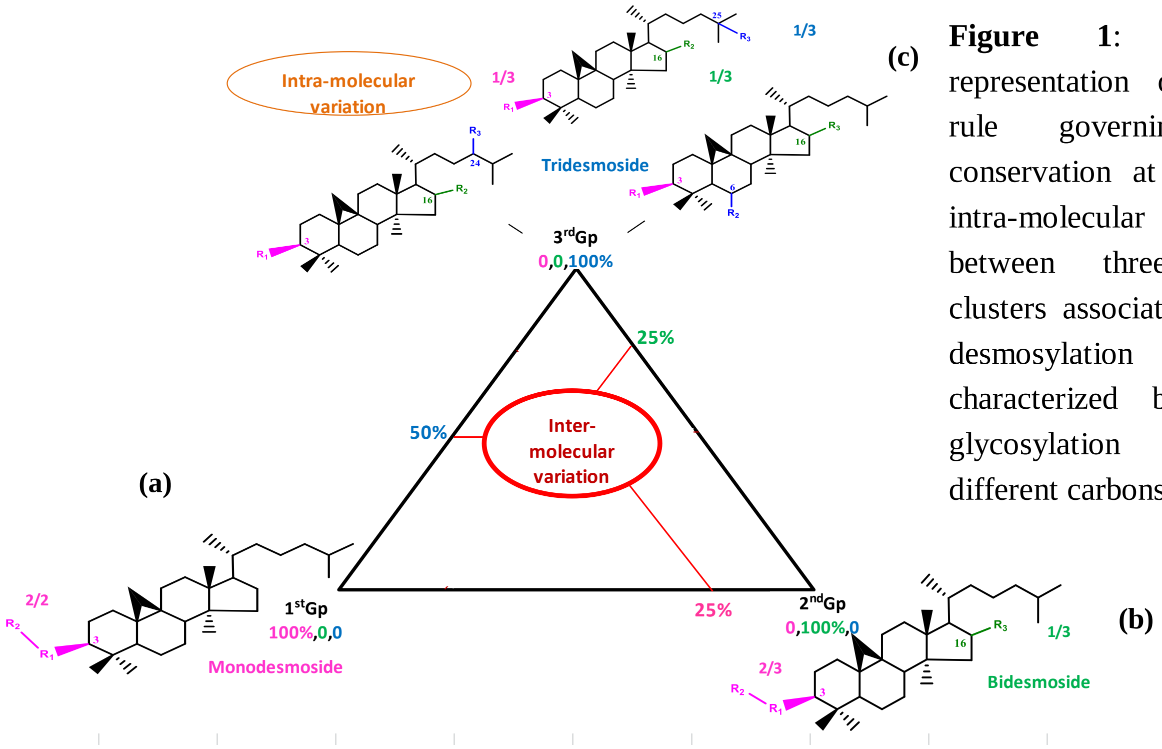


Figure 1: Geometric representation of simplex rule governing mass conservation at inter- and intra-molecular scales between three saponin clusters associated to three desmosylation levels and characterized by relative glycosylation levels of different carbons.

Legend. (**a**) Monodesmoside, (**b**) bidesmoside, (**c**) tridesmoside clusters

After population classification into q (=3) groups (**Fig. 2a**), combinations between clusters were carried out according to a simplex mixture design called Scheffé's matrix [2] (**Fig. 2b**). This matrix combines q groups by randomly varying their weights w_j (j = 1 to q) the ones at the expense of the others. In each combination, the total weight w of the q groups is constant: $w=\sum_{j=1}^{q} w_j = cst$.

So, the Scheffé's matrix consisted of N linear combinations between the q groups. This number of mixtures (N) depends on the two parameters w (number of individuals per mixture) and q (number of clusters):

$$N = \frac{(w+q-1)!}{(\,q-1)!\,w!}$$

In our case, q was equal to 3 clusters and the total weight w was fixed to 10 molecules, so the combinatorial formula gives 66 combinations.

From the N (=66) combinations between the q (=3) groups, N average relative glycosylation profiles are calculated in a response matrix (**Fig.2c**). Then, the mixture design and its response matrix are iterated K (=30) times by bootstrap technique to take into account the molecular variability between and within the three clusters (**Fig.2d**). Finally, the K resulting response matrices were averaged to get a single matrix containing N smoothed molecular profiles (**Fig.2e**).

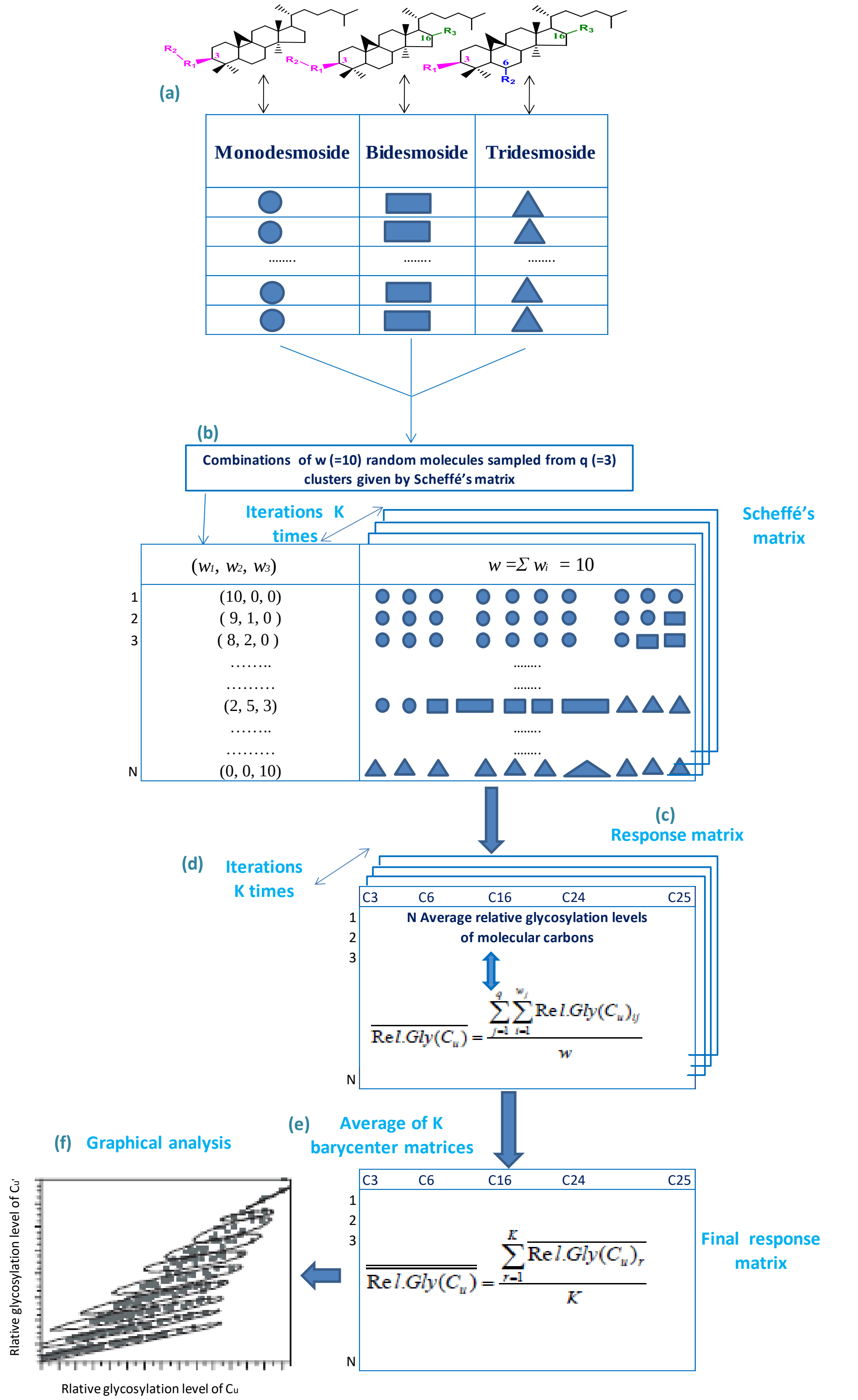

(a)
Monodesmoside
Bidesmoside
Tridesmoside
(b)
Combinations of w (=10) random molecules sampled from q (=3) clusters given by Scheffé's matrix
Iterations K times
Scheffé's matrix
(w1, w2, w3)
w =Σ wi = 10
1 (10, 0, 0)
2 (9, 1, 0)
3 (8, 2, 0)
(2, 5, 3)
N (0, 0, 10)
(c) Response matrix
(d) Iterations K times
C3 C6 C16 C24 C25
N Average relative glycosylation levels of molecular carbons
$\overline{\mathrm{Rel}.Gly(C_u)} = \frac{\sum_{j=1}^{q}\sum_{i=1}^{w_j} \mathrm{Rel}.Gly(C_u)_{ij}}{w}$
(e) Average of K barycenter matrices
$\overline{\overline{\mathrm{Rel}.Gly(C_u)}} = \frac{\sum_{r=1}^{K} \overline{\mathrm{Rel}.Gly(C_u)}_r}{K}$
Final response matrix
(f) Graphical analysis
Rlative glycosylation level of Cu'
Rlative glycosylation level of Cu

Figure 2: Different methodological steps of the simplex approach applied to the Astragalus saponin system stratified according to three degrees of desmosylation.

From the finial matrix of the *N* smoothed profiles, scatter plots were used to visualize relationships between substituted carbons (**Fig.2f**). For each smoothed plot crossing relative glycosylation degrees of two given carbons, the three states of desmosylation *j* (*j* = 1, 2 or 3) were separately analyzed by projecting their weights on the corresponding points. Then, the values of equal weight were delimited by confidence ellipses. With w = 10, 11 weight values (from w_j = 0 to w_j = 10) are concerned resulting in eleven ellipses of weights for each desmosylation state. The succession of the eleven ellipses provides a trajectory indicating how the glycosylation degrees of the two considered carbons varied the one in relation to the other for the formation of considered desmosylation level.

Results

Graphical analysis of smoothed plots revealed competitive, sequential or cooperative processes between C3, C6, C16, C24 and C25 carbons for glycosylation (**Fig.3**):

A competitive process was highlighted between carbon C3 and carbons C6, C16, C24 and C25 for glycosylation (**Fig.3a-d**). This was highlighted by an increase in glycosylation level of C3 at the expense of the other carbons for monodesmoside formation. This leads to monodesmosylated saponins with relatively well-glycosylated C3. In bidesmosides system, the level of C16 glycosylation showed a net increase independently of the carbons C24 and C25 states and with a possible alternative glycosylation way from C6 (**Fig.3e-g**). This indicated flexible role of C16 in the synthesis of bidesmoside saponins. On the other hand, positively inclined ellipses in the three plots (**Fig.3e-g**) indicated that 16-glycoslation played a preparatory role for next glycosylations concerning C6, C24 and C25 leading to tridesmoside system. These sequential glycosylation processes of carbons were highlighted in tridesmoside saponins showing high levels of glycosylation for the carbons C6, C24 and C25 vs intermediate and stable for C16 (**Fig.3h-j**). These simulation results indicate that carbons C6, C24 and C25 have played open roles for glycosylation for the synthesis of tridesmoside saponins.

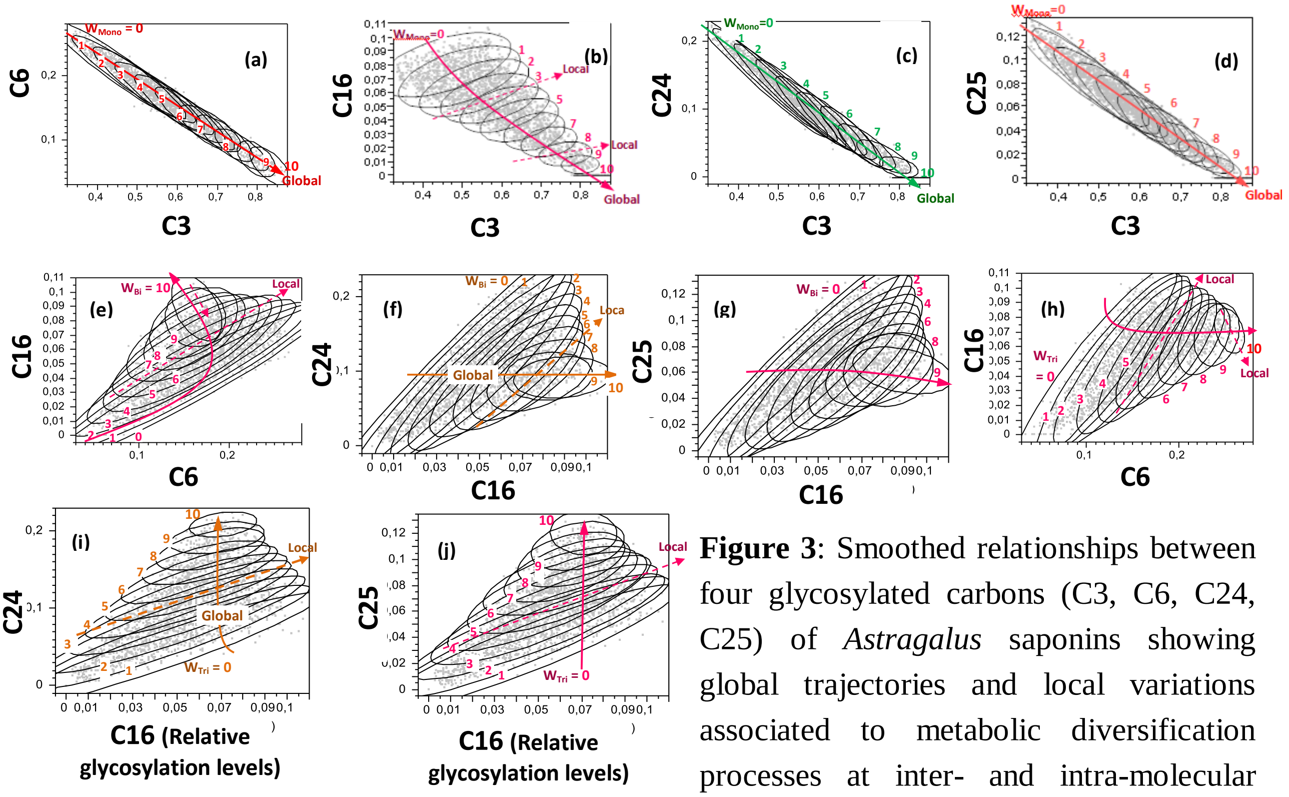


Figure 3: Smoothed relationships between four glycosylated carbons (C3, C6, C24, C25) of *Astragalus* saponins showing global trajectories and local variations associated to metabolic diversification processes at inter- and intra-molecular scales, respectively.

Discussion

Application of the new simplex approach helped to highlight laws governing structural variability of saponins based on cycloartane with aliphatic lateral chain. This shows original objective of simplex approach compared to others computational methods (ab initio, DFT (Density Functional Theory), docking, molecular dynamic…). In fact, the aim of most molecular modeling methods is the description and prediction of physical and chemical proprieties of single or paired molecules [3-5]. In ab initio and DFT methods, these proprieties are approached by extensive combinatorial calculations of inter-atomic interactions or by iterative calculations of electronic density, respectively. Using docking and molecular dynamic, predictions of molecular structures are carried out using by energy calculations; this implies calculations of all the possible binding energies between two molecules or resolution of the motion equation, respectively (**Fig. 4**). However, simplex approach is appropriate for analysis of self-regulation processes of multiple diversification poles of complex systems using machine-learning applied at both inter- and intra-molecular scales. Double scale analysis attributes advantage to simplex approach compared to other computational methods which focus only on intra-molecular scale (inter-atomic scale) (**Fig. 4**). Another advantage of simplex approach concerns big data analysis with possibility of working on unlimited number of molecules belonging to different groups by opposition to other methods conceived for analysis of single or paired molecules at once. Finally, the advantage of other methods turns out in the use of different types of variables (like as electronic proprieties, atomic proprieties….) which authorize their application in extended fields. In contrast the new simplex approach can only be applied in specific fields because in this method the variables must be of the same nature.

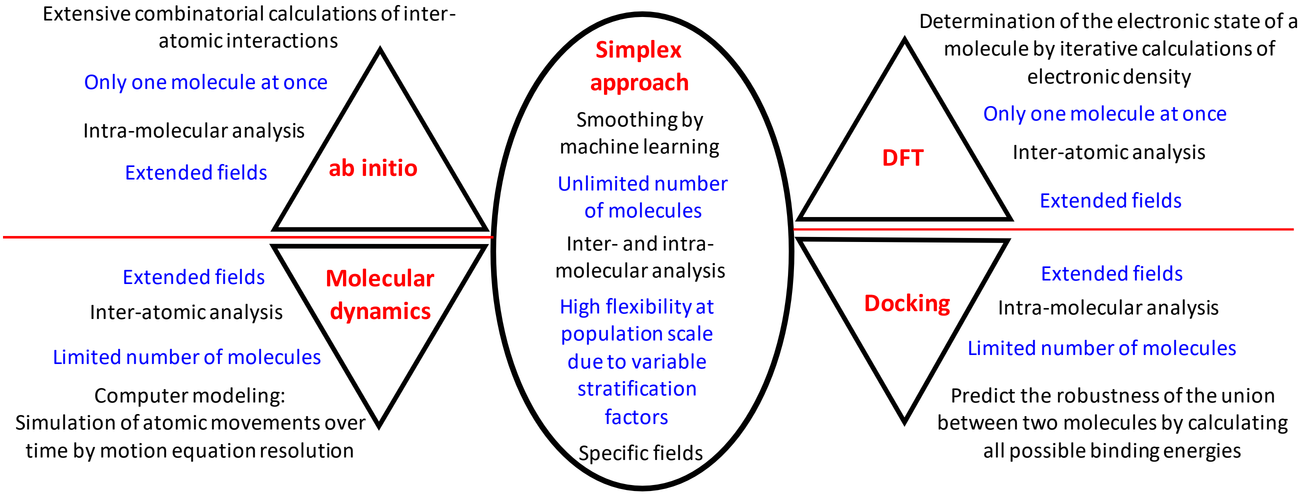

Extensive combinatorial calculations of inter-atomic interactions
Only one molecule at once
Intra-molecular analysis
Extended fields
ab initio
Simplex approach
Smoothing by machine learning
Unlimited number of molecules
Inter- and intra-molecular analysis
High flexibility at population scale due to variable stratification factors
Specific fields
Determination of the electronic state of a molecule by iterative calculations of electronic density
Only one molecule at once
Inter-atomic analysis
Extended fields
DFT
Extended fields
Inter-atomic analysis
Limited number of molecules
Molecular dynamics
Computer modeling: Simulation of atomic movements over time by motion equation resolution
Docking
Extended fields
Intra-molecular analysis
Limited number of molecules
Predict the robustness of the union between two molecules by calculating all possible binding energies

Figure 4: Comparison between the simplex approach and other computational chemistry methods

References

1. Sarraj-Laabidi, A., Messai, H., Hammami-Semmar, A., & Semmar, N. **2017**. Chemometric Analysis of Inter-and Intra-Molecular Diversification Factors by a Machine Learning Simplex Approach. A Review and Research on Astragalus saponins. *Current topics in medicinal chemistry*, *17*(25), 2820-2848.
2. Scheffe, H. **1963.**The simplex-centroid design for experiments with mixtures. *Journal of the Royal Statistical Society. Series B* (Methodological), 235-263.
3. Hoover, W. G., **1986**. *Molecular dynamics*. Vol. 258, Springer-Verlag Berlin Heidelberg, 138p.
4. Koch, W., Holthausen, M. C., **2001**. *A chemist's guide to density functional theory*. Wiley-VCH, New York, 300p.
5. Meng, X.-Y., Zhang, H.-X., Mezei, M., Cui, M., **2011**. Molecular docking: a powerful approach for structure-based drug discovery. *Current Computer-Aided Drug Design* 7, 146-157.