MEDWIN PUBLISHERS
Committed to create value for Researchers

Medicinal & Analytical Chemistry International Journal

Molecular Docking: Considerations of a Low Cost and Suitable Methodology and Some Successful Applications

Batista VS and Nascimento-Júnior NM*

Department of Organic Chemistry, São Paulo State University, Brazil

***Corresponding author:** Nascimento-Júnior NM, Laboratory of Medicinal Chemistry, Organic Synthesis and Molecular Modeling (LaQMedSOMM), Department of Organic Chemistry, Institute of Chemistry, São Paulo State University – UNESP, Rua Professor Francisco Degni, 55, Jardim Quitandinha, 14800-060, Araraquara, São Paulo, Brazil, Tel: +55 21 3301-9897; Email: nailton.monteiro@unesp.br

Opinion

Volume 2 Issue 3

Received Date: September 25, 2018

Published Date: October 04, 2018

Abstract

Molecular docking can be a powerful tool, given the growing amount of available solved protein crystallographic structures and always improving computational tools but it can also develop in a more complex issue than one would initially imagine. The reasoning behind this statement comes from the simple fact that even if a good amount of data is readily available, care should be taken to ensure that the starting point of any docking job suits the purpose of the study at hand. There are several docking methodologies which can be useful in one or another work flow and to many researchers outside the field this can be elusive as there are several parameters that must be addressed beforehand. In the present study, the authors give an overview about our molecular docking protocol and analysis of the results, highlighting key aspects that must be addressed in each step of the process and setting the reader's mind to focus on what kind of information have to be considered before starting a docking job. Our focus is towards freeware software and web servers, resulting in a virtually free methodology that can be easily applied by newcomers to molecular docking. Also, examples of successful docking applications are given.

Keywords: Molecular Docking; Virtual Screening; Protein Obtention; Homology Modeling; Protein Energy Minimization

Abbreviations: SBDD: Structure-based drug design, PDB: Protein Data Bank; DSV: Discovery Studio Visualizer; GA: genetic algorithm; nAChRs: nicotinic acetylcholine receptors; NMR: Nuclear Magnetic Resonance; NPD: Natural Product Database; FDA: Food and Drug Administration HTS: High Throughput Screening; SPR: Surface Plasmon Resonance; LE: Ligand Efficiency; LLE: Lipophilic Ligand Efficiency; SAR: Structure Activity Relationship.

Introduction

Structure-based drug design (SBDD) is receiving increasing attention, both in industry and academia, as structural genomics, computational tools and spectroscopic techniques become more sophisticated. This is a result of the increasing number of protein tridimensional structures elucidated and publicly available [1]. In this context, molecular docking is a wide spread method useful in applications such as virtual screening, lead optimization and to provide understanding of the intricate aspects of intermolecular recognition. This strategy focus on the prediction of the structure of protein-ligand and/or protein-protein complexes through two key steps: conformational sampling of the system and scoring the resulting complexes by a scoring function [2]. There are many different strategies to address both sampling and scoring in molecular docking, as well as several commercial and academic software packages to perform the calculations [3]. A side from the docking packages themselves, there are also other essential steps involved in the development of a docking methodology that relate to the protein structures used, ligand databases and docking parameters [4,5]. Such complexity has to be properly addressed in a successful docking job by careful consideration of the chosen methodology and its limitations.

The first and most relevant concern when envisaging a molecular docking study is, ultimately, why you are using it. There are a number of computational tools just as good (or more suitable) as docking for a given problem and none of them should be disregarded in the planning of a computational methodology. The molecular modeler must be aware of the strengths and weaknesses of a given methodology, as well as the applicability of such approach into the setting of its own problem, and the ideal scenario would be the one where the researcher has all this in mind beforehand.

In the present manuscript we propose and discuss a docking methodology, its basis and key steps, focusing on software packages and web servers freely available to the academic community or subsidized by some funding agencies (e.g. CAPES, BR). The proposed methodology can be successfully applied to solve docking problems, as it will be show further (applications section). As well as some other examples of successful docking applications.

Software Packages and Web Servers

The following software packages and web servers recommended are shown below:

a. Database for protein obtention: Protein Data Bank (PDB) [6];
b. Database for ligand obtention: ChEMBL [7];
c. Ramachandran plot: RAMPAGE [8];
d. Homology modeling: Swiss Model [9];
e. 3D structure visualizer and/or builder: Discovery Studio Visualizer (DSV) [10] or PyMol [11];
f. Ligands ionization state prediction: Marvin Sketch [12];
g. Ligands/protein energy minimization: MOPAC2016 [13];
h. Docking package: CCDC GOLD Suite [14-17].

Discussed Methodology

The methodology described in this paper is summarized in Figure 1.

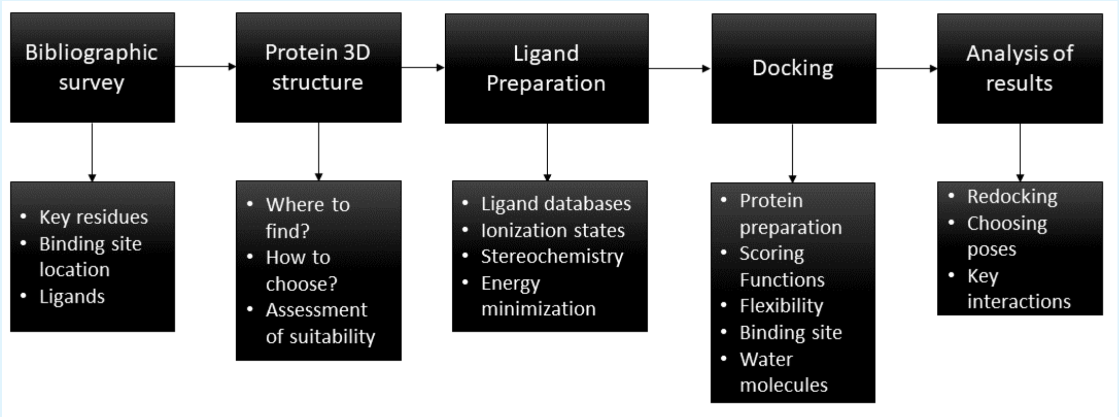


Figure 1: Summary of the discussed methodology.

Copyright© Batista VS and Nascimento-Júnior NM.

The starting point of any docking study should be an extensive research of the existing scientific literature regarding the desired molecular target. Any information about the properties of the system in study is of great importance. The researcher should be concerned with retrieving information like key residues for binding of ligands, where is the binding site located, what class of molecules binds to this specific binding site, presence or absence of water molecules into this site, what is the effect of the binding site environment on the ligand's ionizable moieties, existence of induced-fit effects and so on. A good amount of the mentioned information can be obtained just by analyzing the crystal structure of the desired target but care should be taken to avoid overlooking information that is not straightforward obtainable from this analysis. The presence of water molecules, for instance, can only be evaluated if the crystallographic resolution of the desired structure is good enough for this purpose.

The docking package used in this work is GOLD, available in the CCDC GOLD Suite, which have been extensively tested and has shown excellent performance for pose prediction and virtual screening. Its four native scoring functions, GoldScore [14], CHEMPLP [15], ChemScore [16], and ASP [17] were ranked 1th, 2th, 5th and 6th, respectively, in a study by Li, et al. [18] comparing several scoring functions from popular docking software like MOE [19] and GLIDE [20]. In addition, the CCDC GOLD Suite is subsidized by CAPES Brazilian funding agency to academic research institutions, which alongside with the use of freeware software packages and web servers, making the overall software cost for using this methodology virtually zero. GOLD uses a stochastic approach for conformational sampling, implemented through genetic algorithm (GA), alongside with the formerly mentioned scoring functions to score the ligand's predicted binding modes and has several available options for tuning the calculation as needed.

Obtention of the Targeted 3D Structure

Initially, a suitable tridimensional structure must be obtained. This can be done by either retrieving the data from a database, usually the PDB, or by computational modelling approaches, like homology modelling, and this choice depends on what information is already available for the chosen target. Homology modelling demands more effort and has more room for inaccuracies but this approach has been successfully applied in many cases [21]. On the other hand, by using an already elucidated crystal structure, the advantage of experimental evidence supports the observed conformation of the system. Nevertheless, the choice of the starting structure is of major importance because changes in the position of the residues comprising the binding site can fool docking engines [22]. The evaluation of the suitability of a crystal structure is done by analyzing its Ramachandran plot, which can be done free of charge using RAMPAGE web server. It is important that the binding site residues are in favored or allowed regions of the graph, provided that the reason why a given residue is inside the allowed region instead of the favorable region is explainable, like to accommodate the presence of a ligand. The structure's resolution should be as good as possible. Also, crystallographic parameters contained in the PDB file header, like R-factor, should not be overlooked. A review by Wlodawer, et al. [23] that covers the basics of protein crystallography for those outside the field is recommended to better understand the implications of such parameters on the quality of the crystallographic data.

It is possible that the researcher has to choose from a variety of available crystal structures for the desired target either to start the docking or to build a homology model. In such cases, the most important thing to keep in mind is what your problem requires. For instance, when faced with several crystal structures for building a homology model the recommendation is that the template and target are of the same family of proteins and with at least 30% sequence identity between them but if the relevant residues in the homology model are properly positioned it is possible to find a situation where identities smaller than 30% are acceptable [24]. Additionally, it is possible to build homology models using pieces of different crystallized proteins to complete the gaps from each other. A homology model can be built online using the SwissModel web server. As for selecting among relevant crystal structures to a straightforward docking approach, the ideal scenario would be the one where you have a co-crystallized ligand to use as a reference. For example, when working with neuronal nicotinic acetylcholine receptors (nAChRs) it is advisable to start from a structure containing nicotine, a classic agonist, because the observed interactions can guide what are the relevant motifs for molecular recognition. Selecting a structure with a known co-crystallized ligand is also advantageous because it is easier to identify the region which should be treated as the binding site.

Copyright© Batista VS and Nascimento-Júnior NM.

Setting up Ligands for Docking

With the appropriated crystal structure at hand, the next step is setting up ligands to dock into the desired binding site. Their tridimensional structures should be drawn using a software such as DSV or PyMol, the first being available free of charge. It is possible to obtain published ligands online on the ChEMBL database but it is important to always check the original paper for biological activity and absolute stereochemistry data. In docking approaches the absolute stereochemistry is mandatory for a good level of confidence in the obtained results. Thus, ligands acting in racemic form or with absolute stereochemistry poorly defined by the original authors should be avoided. If unavoidable, it is possible to set the docking software to score all possible stereoisomeric forms of any molecule by building and docking them into the binding site and then sorting the best option based on the calculated score for each one.

It is important that, when building a database of ligands, one takes into account what the problem to be solved requires. First of all the researcher has to be sure that the selected ligands all bind into the same binding site in order to compare the differences in their binding modes. The number of ligands needed depends on the desired application. Docking-based virtual screening, for instance, requires a good amount of molecules while the understanding of the binding motifs for a given molecule might only need one or two. When the interest residues in the protocol's ability to diferentiate between active and inactive compounds, it is best to use a focused database, meaning that the actives and inactives ought to have some degree of similarity. If the researcher is interested in using docking in a 3D-QSAR study; the bioassays for the selected ligands must all have been performed through the same in vitro protocol in order for the bioactivities to be comparable. These examples are only a few of the possible criteria that one has to take into account for the selection of ligands suitable to the study.

Another matter to be addressed when building ligands for docking is their ionization state. There are many ligand-receptor interactions that only occur due to the existence of an ionized moiety in the ligand. It is possible to predict the ionization state of a given molecule using ChemAxon'sMarvinSketch software which is freely available for academic research. It is important to notice that the binding site's micro environment can play a significant role in the ionization state of ligands so the calculation should only be used when there is no information regarding the active micro species for a given molecule.

The ligand's starting conformation should be reasonably low in energy so an energetic minimization step is advised. This can be done by quantum mechanics, semi-empirical, empirical methods or molecular mechanics. The choice of method is not so important, provided that the conformation is realistic, since the software is going to perform a stochastic search of the ligand's conformational space. We recommend the semi-empirical approach because of it is easy of use, presents good accuracy, has low computational cost and because the state-of-the-art algorithm MOPAC's PM7 is freely available to the academic community and also implementable in many third party software, like Mercury [25], which is a part of the CCDC GOLD Suite.

Protein Preparation

Once the ligands have been built, the preparation of the protein structure for docking is started. The docking calculation can be performed using different software packages, like Auto DockVina [26], GLIDE and GOLD, each one having its own particularities. Although the choice of which software to use is up to the researcher, the focus of this work will be directed towards GOLD because of the reasons formerly explained. Nevertheless, the guidelines presented here apply for most docking engines.

In GOLD, the initial step for protein preparation is the addition of hydrogen atoms, which are not explicit in structures obtained through X-Ray crystallography. If the initial structure already has hydrogens, like the ones obtained through Nuclear Magnetic Resonance (NMR), this step can be safely ignored. The addition of such atoms is automatic upon input of the required command by the user. It is important to notice that some amino acid residues, like histidine (His, H), have more than one possible tautomeric state and this should be reflected in the docking calculation by using the most likely tautomer for the system under evaluation. If no information about tautomerization of binding site residues is available, it is reasonable to test for all of them and select the one yielding better results.

The next step is the removal of water molecules and co-crystallized ligands. All waters inside the binding site must be removed even if they are going to be used in the calculation. The removed waters can be saved to a separate coordinate file which can be used further down the workflow to define its position. The co-crystallized

Copyright© Batista VS and Nascimento-Júnior NM.

ligands located inside the binding site must also be removed before docking and the ones outside have no influence in the calculation since the software only considers a small predefined section of the whole protein. Ligands left in the binding site are going to be treated as an additional protein chain so cofactors can be used in calculations.

Setting up the Docking

After the protein and ligand's are prepared, the docking job itself can begin. GOLD has configuration templates for some classes of proteins, like kinases and cytochromes belonging the P450 super family, although for most docking jobs the user will have to set the desired configurations manually.

In GOLD the molecular modeler has the option to perform rigid or flexible docking. Rigid docking has the advantage of experimental evidence to support the obtained results but lacks the freedom of movement that would better represent the real biomolecular system. The opposite applies to flexible docking: the flexibility of the true biological system is mimicked but as the amino acids side chains move, the uncertainties revolving their positions increase and the output conformation might not be a good representative of the reality, thus increasing the number of false positives (i.e. ligands that appear to dock well but in reality do not bind). Also, the computational time required for a flexible docking greatly increases in comparison with the rigid approach. GOLD can make up to ten side chains flexible for a given protein and has also the option for introducing localized backbone movements through rotation of the improper torsion defined by the atom sequence Cα-N-C-Cα. The software has two other options to introduce protein flexibility, namely soft potentials and ensemble docking. The first approach uses alternative Lennard-Jones potentials in the external Vander Waals contributions to the fitness function in order to allow shorter contacts, while the second uses two or more superimposed forms of the same protein in the same docking job to consider different protein conformations. Protein flexibility should be introduced preferably when there is experimental evidence supporting it.

The center and size of the search radius – the binding site –, must be defined by the user. GOLD allows setting an atom, point, ligand or list of atoms as the center of the active site. The binding site radius should cover the size of the largest ligand to be docked. As mentioned before, if there is a co-crystallized ligand complexed with the biomacro molecule, it is possible to obtain information of where to define this binding site. The user has to be assured that the ligand is indeed in the orthosteric site and not inside an allosteric one, or the opposite depending on the site of interest.

After the definition of the binding site, the user has to load the ligands into the software and define how many GA runs will be performed for each of them. Every GA run results in a different binding mode with a score associated to the predicted ligand-protein interactions. If one docks the same ligand originally bound to the active site, it is possible to set its original conformation as reference, i.e. redocking, which is going to be addressed further down this manuscript. Ligands can be treated as rigid or flexible during the calculation, with flexibility defined as free rotation of rotatable bonds. There are other options for ligand flexibility such as flipping ring corners, amide bonds, pyramidal nitrogen, planar R-NR1R2 and also protonated carboxylic acids. There is also an option to detect possible ligand's internal hydrogen bonds. A ligand should always be treated as flexible, the only exception being cases where the evaluation of the interactions of a specific conformation is desired.

Water molecules in a docking job using GOLD can be treated as fixed, free to rotate, free to translate within a 2 Å radius or both the latter. It is also possible to let the software decide whether or not the water should be bound or displaced by the ligand during the job. One should be cautious when deciding if a water molecule should be included since many ligand-protein interactions are mediated by water. As mentioned earlier, only structures having adequate resolution (≥ 2.7 Å) [23] give useful information about the presence of water molecules inside the binding site. Information about whether or not to include a water molecule in the calculation can be obtained by checking the interactions observed in homologous crystallized proteins or through molecular dynamics studies.

The docking can be run using four different scoring functions in GOLD, namely ChemPLP, GoldScore, ChemScore and ASP. Each one of them has its strengths and weaknesses and their applicability to a given problem has to be evaluated. Guidelines for choosing among them will be described in the next section of this work. The software allows rescoring the solutions of a job using different scoring functions and this can be used to compensate the deficits from the original run. It is possible to enforce diversity by setting the software to generate diverse solutions for the same ligand. Also, the

Copyright© Batista VS and Nascimento-Júnior NM.

procedure can be programmed to terminate early if a set of solutions are all within a predefined RMSD threshold from each other and this is particularly useful when working with easy flexible ligands, i.e. ligands with little conformational freedom. All these options have impact on the speed and efficiency of the search. The GA parameters can be optimized to balance the ration between speed and efficiency as needed.

GOLD has several additional parameters that can be tuned to address specific situations, like docking with metalloproteins and complexed metal ions, soft potentials for works involving discrete backbone and side chain movements, covalent docking and others that will not be addressed here. As stated before, extensive research is useful to decide whether or not to introduce them into the calculation. Detailed information about those parameters and how to use them is available on the software's user guide.

Analysis of the Docking Results

A big concern regarding molecular docking is the validity of the results. As stated before, the applicability of a scoring function and the chosen docking configuration has to be assessed beforehand. This is usually done by evaluating their ability to differentiate between active and inactive molecules from a database, e.g. ROC curve analysis [27], by its capacity to align a set of ligands in a way that the variations in their bioactivities is explained, e.g. CoMFA [28], or by redocking the co-crystallized ligand into the binding site to compare the calculated coordinates and interactions with the experimental data. This work will focus on redocking, as it is the simplest of the validation techniques, and information about how the validation protocol was performed for the two first examples is given elsewhere [29,30].

Redocking analysis is very straightforward: one only need to compare the atomic positions of the redocked ligand with the original conformation extracted from the PDB file via RMSD. This can be done easily in GOLD itself by selecting the ligand removed from the binding site in the previous steps as reference. Usually, when the RMSD is low there will be good agreement between the interactions observed for the reference and the ones calculated by the software. More flexible ligands are likely to have higher RMSD values but as long as the pharmacophoric groups are correctly aligned it is safe to assume good accuracy. Most docking jobs are interested in evaluating interactions for ligands other than the reference, i.e. cross-docking. This limits the applicability of validation through redocking because ligands completely different from the reference might not be well predicted by the computational method. In parallel, the binding site can undergo conformational changes in response to the presence of a ligand, thus reducing the validity of cross-docking situations even if the redocking was successful. In any case, the choice of scoring function based on redocking is performed by assessing which one yields lowest RMSD for the top ranked pose of the redocked ligand.

The results of a docking run are a set of predicted binding modes (poses) for each ligand ranked by the score function according to the interactions between the ligands and the protein. Being so, the expectation is that the top ranked pose represents the conformation and interactions of the true biological system. This is not always true as there will be occasions where the score for the two top ranked binding modes is similar but the poses themselves are distinct. A useful criteria to select which one of them is representative is to evaluate the predicted interactions for both and select the pose that has the most interactions involving key residues, i.e. residues known to be of importance to molecular recognition. We recommend this type of analysis when the difference in score between the top and second ranked binding modes is below 0.5 units. The selected docking poses can be extracted to a PDB file to be analyzed in a structure visualizer such as DSV or PyMol. For analysis of the results DSV is very suitable, first because it is freely available and second because of its automated options for ligand-protein interactions.

Usually docking algorithms fail to correlate the calculated score with the experimental binding affinity and thus it is important to notice that this score can only be used to estimate the affinity of a compound in relation to others [31]. This is the reason why when analyzing the results of a docking job one should consider interactions with key residues as well as the overall orientation of the docked ligands when compared to the reference.

Successful Docking Applications

The first example herein presented of successful molecular docking application is a study by Okada-Junior, et al. [32] targeting malaria disease through the cytochrome *bc*$_1$ complex of *P. falciparum*, a protozoa from the Plasmodium spp associated with severe malaria [33]. Atovaquone (1) is a first line treatment for malaria disease but resistance has been reported since the early 1990s, thus making the search for new treatments of

Copyright© Batista VS and Nascimento-Júnior NM.

paramount importance [34]. The drug is an analogue of ubiquinone and acts by competitively inhibiting the endogenous substrate at the cytochrome bc1 enzyme complex, thus hampering cellular respiration due to reduction of electron transfer in the mitochondria and ultimately promoting the collapse of the mitochondrial membrane potential [35]. In the mentioned work, a new series of phtalimidic derivatives were synthesized and tested for *in vitro* inhibition of *P. falciparum* from the 3D7 strain (chloroquine sensitive), resulting in two promising compounds with low micromolar inhibitory concentration (IC_{50}) of 4.2 and 6.8 μM, respectively named 2 and 3 in this work (10 and 16 in the original work). Compound 2 was also tested for inhibition of *P. falciparum* K1 strain, which is resistant to the antimalarial drugs chloroquine, pyrimethamine and sulfadoxine, displaying approximately the same inhibition as for the sensitive strain (IC_{50} = 4.3 μM). Aiming to understand the molecular basis of such results, a docking study using 2 was performed. The methodology follow the guidelines presented in this paper: both Atovaquone and 2's 3D structures where built in DSV and optimized through the semi-empirical algorithm PM6 [36] with MOPAC and the docking was performed in GOLD using ChemPLP scoring function. No 3D structure was available for *P. falciparum* cytochrome *bc*$_1$.The more suitable protein structure was found to be the *bc*$_1$ cytochrome from *S. cerevisiae* complexed with stigmatellin at 1.9 Å resolution (PDB ID 3CX5) [37]. Atovaquone binds to the complex III of S. cerevisiae with high affinity (Ki = 5nM), which led to its development as a nonpathogenic surrogate model for studying bioactive compounds targeting the parasite through this mechanism. A crystallographic structure of the same protein complex crystallized with atovaquone exists (PDB ID: 4PD4) [38] but it has lower resolution and lacks the Rieske protein, which is essential for molecular recognition and thus was discarded for docking. The binding mode of atovaquone in 4PD4 was used as reference after confirmation that the difference in atomic positions of binding site residues was minimal. Also, it is important to mention that atovaquone was modelled in its ionized form due to studies claiming that this is the active microspecies. Docking of atovaquone into 3CX5's binding site resulted in interactions very similar to the ones observed for the reference structure, yielding RMSD= 0.4 and PLPscore= 63.3, indicating good accuracy of the computational approach. Compound 2 was then docked using the validated conditions and its calculated score was 63.6. The binding modes of atovaquone and 2 are presented in Figure 2. Molecular structures for 1, 2, and 3 are presented in Figure 3.

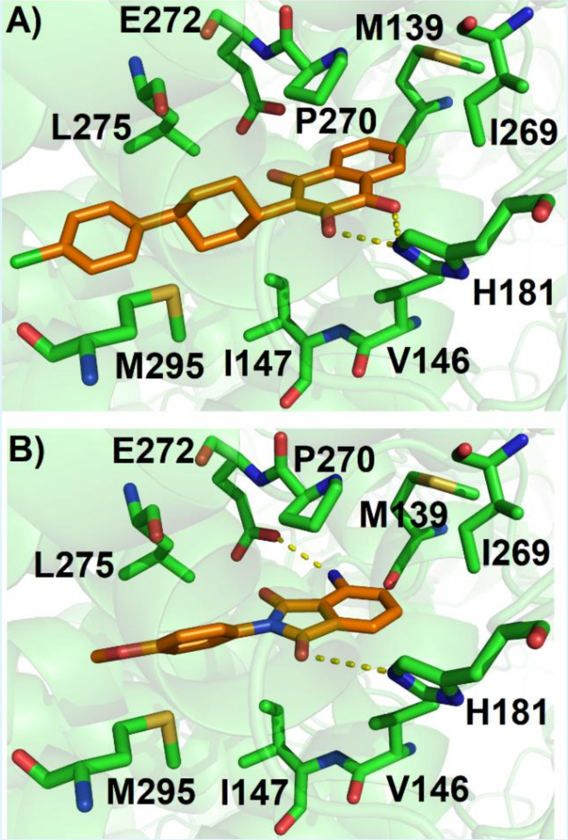


Figure 2: A) Atovaquone's binding mode and B) 2's calculated binding mode. Hydrogens where omitted for clarity. Dashed yellow lines represent hydrogen bonds. Figure generated with PyMol.

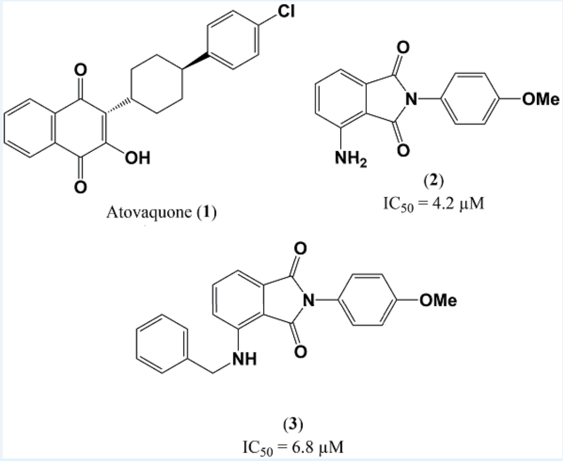


Figure 3: Molecular structures of compounds 1, 2 and 3.

Copyright© Batista VS and Nascimento-Júnior NM.

The computational model predicted the occurrence of hydrogen bonds involving the phtalimidic moiety and the side chains of E272 and H181, as well as close contacts with the side chains of residues I147, V146, P271 and with backbone atoms of W142. The methoxyphenyl moiety takes part in hydrophobic interactions with the side chains of M295 and L275. The hydrogen bond involving carbonylic oxygen from the ligand and the side chain of residue H181 is observed for the reference. The interaction of 2 with E272 is not present in the crystal structure of atovaquone but exists in other inhibitor's complexes, like stigmatellin and 5-n-heptyl-6-hydroxy-4,7-Dioxobenzothiazole (HDBT), although in some cases being mediated by a water molecule. The ligand's phthalimidic ring mimics atovaquone's hydrophobic contacts with the side chains of V146 and P271. The methoxyphenyl moiety is positioned close to where atovaquone's cyclohexyl moiety is, simulating the reference's hydrophobic contacts with residues L275 and M295. These results alongside with an enzymatic assay measuring how compound 2 affects the activity of the *bc*$_1$ complex decylubiquinol-cytochrome c oxido reductase, which resulted in 74% inhibition at 70 μM, indicate that the proposed phtalimidic derivative indeed acts through this mechanism of action and can be a promising lead to further development in malaria drug discovery research. This real life example gives insight on how docking can be useful in the understanding of relevant interaction motifs and mechanism of action of newly synthesized compounds.

The second example was developed by Kayode, et al. [39] which used molecular docking as a tool for the identification of inhibitors of mesotrypsin through virtual screening. Mesotrypsin (PRSS3) is one of the digestive enzymes produced and secreted by the human pancreas and it's over expression has been observed in several different types of cancer. It also is related to metastasis in some types of cancer, like prostate and pancreatic. This enzyme is very different from other human trypsins because it has almost total resistance to biological trypsin inhibitors [40]. In the mentioned work the researchers performed an ensemble docking using three crystal structures of mesotrypsin, with PDB ID's 3P92 [41], 3P95 [41] and 1HW4 [42] using Glide XP algorithm to screen the Natural Product Database (NPD) and the Food and Drug Administration (FDA) approved Drug Database. Ligands where prepared automatically through Schrodinger's LigPrep module. Molecules having docking scores lower than -10 kcal/mol versus one or more receptors were selected and visually inspected for hydrogen bonds with residues D189, S190, G192, R193 and G216 from the binding pocket, thus resulting in 28 promising candidates. Among those, 12 were readily commercially available and therefore were obtained and evaluated for inhibitory activity towards mesotrypsin with two of them indeed displaying inhibition in the low micromolar range, namely diminazene (4) (Ki = 3.66 μM) and hydroxystilbamidine (5) (Ki = 10.57 μM). Diminazene was successfully crystallized with mesotrypsin (PDB ID: 5TP0) and its experimental binding mode matched the calculated one with RMSD = 0.614 over the well-defined electron density region, thus validating the proposed methodology. The second benzamidine moiety from 4, which is located within the solvent channel, presented poor definition of the electron density in that region of the crystal structure, indicating that the drug is able to adopt multiple conformations within this channel and in agreement with other published crystal structures for this molecule complexed with bovine trypsin [43]. The experimental binding mode of the well-defined electron density region is shown in Figure 4. Molecular structures for 4 and 5 are presented in Figure 5.

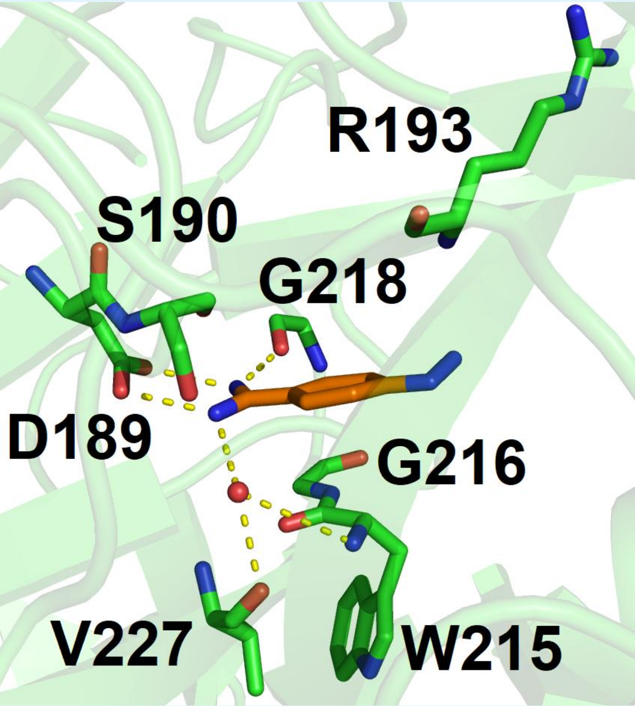
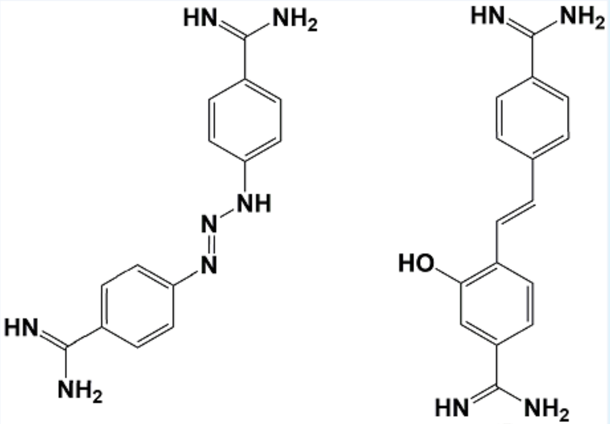


Figure 4: Representation of the binding mode observed for diminazene (4) complexed with PRSS3 (PDB ID: 5TP0). Hydrogens where omitted for clarity. Dashed yellow lines represent hydrogen bonds. Figure generated with PyMol.

Copyright© Batista VS and Nascimento-Júnior NM.



(4)
Ki = 3.66 μM

(5)
Ki = 10.57 μM

Figure 5: Molecular structures of compounds 4 and 5.

The interaction with R193, not shown in Figure 4 due to poor definition of the electron density in this region, is one of the most critical interactions for this enzyme as this residue is located on a region of the binding site where almost all other trypsin and chymotrypsin family serine proteases have a highly conserved glycine. Therefore, further optimization of this interaction may be a strategy to address selectivity for mesotrypsin over other trypsins. Thus, this molecule presents itself as an excellent starting point for selective mesotrypsin inhibitors, as there is currently almost none [44]. Looking at this example the power of docking as a tool for virtual screening becomes clear. Then, that docking can be useful to drive the design of selective inhibitors towards a target over other related proteins.

The final application discussed here is a work by Crawford, et al. [45] which used molecular docking to enhance the affinity of a lead compound targeting mitogen-activated protein kinase, kinase, kinase and kinase 4 (MAP4K4, a.k.a. HGK). MAP4K4 expression and function are linked to focal adhesion dynamics regulation, embryotic development, insulin sensitivity, systemic inflammation, lung inflammation, atherosclerosis andtype-2 diabetes [46]. It has been recently reported that this enzyme is involved in lung adenocarcinoma maintenance through regulation of the MAPK/ERK pathway, acting via inhibition of protein phosphatase 2 [47]. Initially the researchers used high throughput screening (HTS) to search a fragment library, using surface plasmon resonance (SPR), aiming to identify a hit for further development and they reported the progression for an oxazole fragment, herein named 6 (1 in the original work), with Kd = 220 μM and ligand efficiency (LE) of 0.42. Based on the comparison between the fragment and other kinase inhibitors they postulated that the oxazole acted competitively at the enzyme's ATP site and thus performed a molecular docking targeting this site. The docking was performed with Glide SP algorithm through Schrodinger's Maestro interface, with LigPrep module being the method for ligand preparation and using an unpublished in-house MAP4K4 structure complexed with one of the HTS hits as 3D coordinates (resolution of 2.35 Å). Binding poses where evaluated based on Glide's docking score and by assessing the formation of hydrogen bonds to hinge residues E106 and C108 and other favorable intermolecular interactions with binding site residues. The docking software predicted the oxazole ring to be taking part in a conventional hydrogen bond with the backbone nitrogen of residue C108 and also a non-classic hydrogen bond with the carbonyl oxygen from E106's backbone. Aiming to maximize these hydrogen bonds they synthesized biaryl compounds which had complementary hydrogen bond donor/acceptor characteristics. Restrictions where applied when designing these fragments as to keep them inside an ideal drug-like space, thus permitting only fragment growths leading to molecules with molecular weight below 350 Da and cLog P below 3.5. This step led to the identification of a quinazoline 7 (8 in the original work) having 55-fold increase in potency (Kd = 4.3 μM and IC50 = 0.189 μM) that still had comparable LE to the lead fragment 6(LE = 0.41) and also having good lipophilic ligand efficiency (LLE) of 2.4. Then a structure activity relationship (SAR) study was conducted using the identified scaffold as template for exploring the effects of different substituted aryl groups. *Ortho* and *para* substitutions resulted in decreased potency while introduction of halogens in *meta* position yielded more potent molecules. From this series, molecules 8 and 9 (19 and 22 in the original publication) displayed the best enhancement, with $IC_{50}$of 0.058 and 0.077, LE of 0.55 and 0.54 and LLE of 3.94 and 3.03, respectively. The X-ray co-crystal structure of compound 9 was solved (PDB ID: 4OBO) and revealed the molecular basis for the observed results from this series. First, the amino quinazoline core indeed addressed the desired hydrogen bonds initially observed for 6 and an additional non-classic hydrogen bond exists between this moiety and the backbone carbonyl oxygen of C108. Substitutions in *para* position where detrimental due to a salt bridge between K54 and D171 that blocked further expansions in this position and *ortho* position was not well tolerated due to unfavorable torsion strains arising from binding into a nearly planar

Copyright© Batista VS and Nascimento-Júnior NM.

binding pocket. The side chain of residue Y36, which is part of the P-loop, moved down to interact via T-stacking with the quinazoline core. The halogen atom in *meta* position takes part in a favorable interaction with the gatekeeper residue M105. From the data obtained in this step it became clear that although potency was enhanced, the molecules where deviating from the ideal drug-like space initially proposed for cLogP and thus to increase polarity a nitrogen walk around the quinazoline core was undertaken, replacing carbon for nitrogen atoms in positions 5, 7 and 8. The results of these modifications showed that Y36 had major impacts on the ligand's potencies due to the formerly mentioned T-stacking being unfavorable for the substitutions in positions 7 and 8, mainly because of repulsion involving the pi system and the aromatic lone pairs. Substitution in position 5, producing compound 10 (26 in the original work), indeed enhanced both potency and LLE (IC_{50} = 0.017 µM and LLE = 5.33), driving the series back into the desired drug-like space. Then, a new series of close analogues of 10 where synthesized, culminated in compound, 11 (29 in the original work) having comparable potency of 0.017 µM and superior LEE of 6.34, as well as favorable *in vivo* pharmacokinetic properties. The crystal structure for MAP4K4 complexed with 11 was also solved (PDB ID: 4OBP) and displayed virtually the same interactions, having only an additional hydrogen bond with the side chain of K79. The binding modes of both crystallized molecules are displayed in Figure 6. Molecular structures for compounds 6 through 11 are displayed in Figure 7.

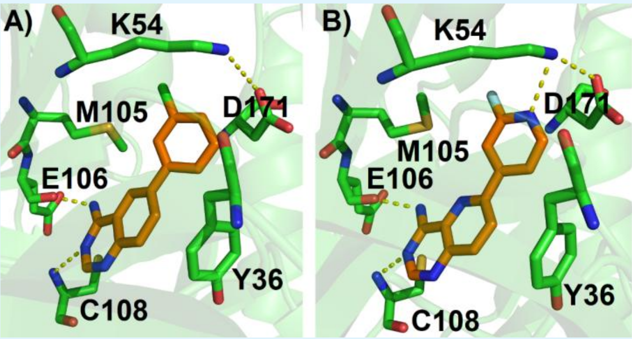


Figure 6: A) 9 complexed with MAP4K4 (PDB ID: 4OPO) and B) 11 complexed with the same enzyme (PDB ID: 4OBP). Hydrogens where omitted for clarity. Dashed yellow lines represent hydrogen bonds. Figure generated with PyMol.

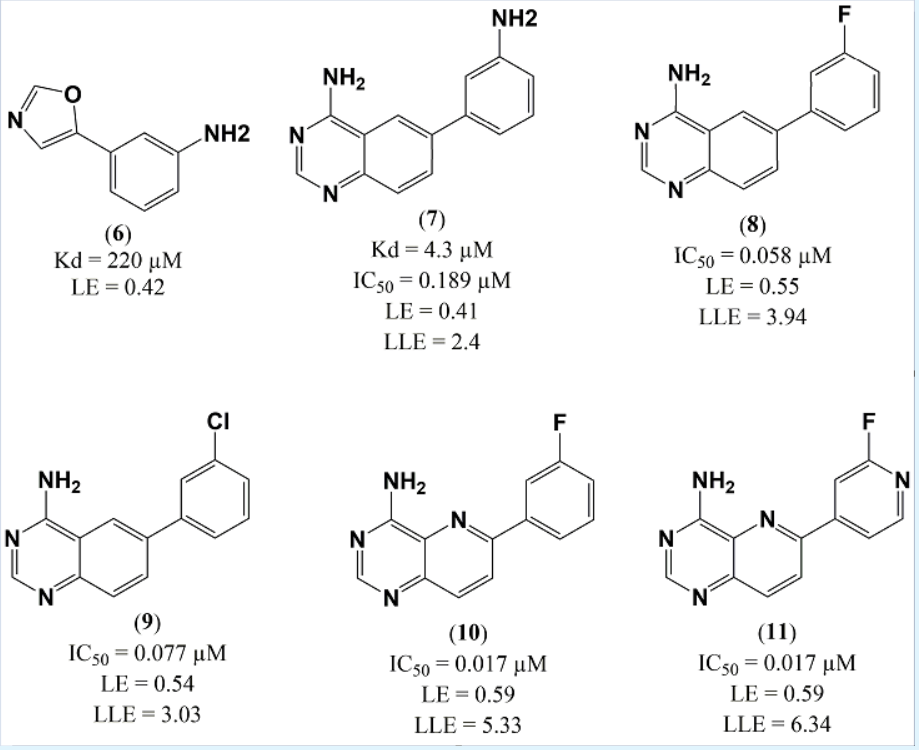


Figure 7: Molecular structures of compounds 6 through 11.

This example illustrates how interactions predicted by molecular docking can be useful to drive synthetic efforts towards enhancing potency of a lead compound.

Conclusion

A low cost docking methodology was discussed with the goal of introducing key aspects of molecular docking to newcomers to the field, like how to choose a starting crystal structure, build ligands for docking correctly, how to interpret the results and other aspects. The discussed methodology was successfully applied and published in combination with other results. In addition, other examples illustrating the applicability of molecular docking were also discussed. Those examples show the importance of extensive bibliographic survey about the system to be modelled. This methodology can serve as a guide and definitely has room for modifications, which we strongly advise in order to properly address the docking problem to be solved.

Acknowledgments

The authors would like to thank São Paulo Research Foundation (FAPESP, process number: 2018/00187-7), for the financial support. Also, we thank Guilherme Gonçalves Costa (IQ/Unicamp) for English language revision.

References

1. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Docking and Scoring in Virtual Screening for Drug Discovery Methods and Applications. Nat Rev Drug Discov 3(11): 935-949.
2. Meng XY, Zhang HX, Mezei M, Cui M (2011) Molecular Docking: A Powerful Approach for Structure-based Drug Discovery. Curr Comput Aided Drug Des 7(2): 146-157.
3. Pagadala NS, Syed K, Tuszynski J (2017) Software for molecular docking: a review. Biophys Rev 9(2): 91-102.
4. Elokely KM, Doerksen RJ (2013) Docking Challenge: Protein sampling and Molecular Docking Performance. J Chem Inf Model 53(8): 1934-1945.
5. Huang N, Shoichet BK, Irwin JJ (2006) Benchmarking Sets for Molecular Docking. J Med Chem 49(23): 6789-6801.
6. https://www.rcsb.org/
7. https://www.ebi.ac.uk/chembl/
8. http://mordred.bioc.cam.ac.uk/~rapper/rampage.php
9. https://swissmodel.expasy.org/
10. Dassault Systèmes BIOVIA , Discovery Studio Modeling Environment, San Diego, Release 2017.
11. The PyMOL Molecular Graphics System , Version 0.99, Schrödinger, LLC, 2002.
12. MarvinSketch (version 6.2.2, calculation module developed by ChemAxon, http://www.chemaxon.com/products/marvin/marvinsketch/, 2014.
13. Stewart JP MOPAC (2016) Stewart Computational Chemistry, Colorado Springs, CO, USA.
14. Jones G (1997) Development and validation of a genetic algorithm for flexible docking. J Mol Biol 267(3): 727-748.
15. Korb O, Stutzle T, Exner TE (2009) Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS. J Chem Inf Model 49(1): 84-96.
16. Baxter CA, Murray CW, Clark DE, Westhead DR, Eldridge MD (1998) Flexible Docking Using Tabu Search and an Empirical Estimate of Binding Affinity. Proteins 33(3): 367-382.
17. Mooij WT, Verdonk ML (2005) General Targeted Statistical Potentials for Protein-Ligand Interactions. Proteins 61(2): 272-287.
18. Li Y (2014) Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. J Chem Inf Model 54(6): 1717-1736.
19. Molecular Operating Environment (MOE) (2013) Chemical Computing Group ULC, 1010Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2018.
20. Glide Schrödinger (2018) LLC, New York.
21. Vyas VK, Ukawala RD, Ghate M, Chintha C (2012) Homology Modeling a Fast Tool for Drug Discovery:

Copyright© Batista VS and Nascimento-Júnior NM.

Current Perspectives. Drug Discov Today 14(13-14): 673-678.

22. McGovern SL, Shoichet BK (2003) Information Decay in Molecular Docking Screens Against Holo, Apo, and Modeled Conformations of Enzymes. J Med Chem 46(14): 2895-2907.

23. Wlodawer A, Minor W, Dauter Z, Jaskolski M (2008) Protein Crystallography for non-crystallographers, or How to Get the Best (but not more) from Published Macromolecular Structures. FEBS J 275(1): 1-21.

24. Saxena et al. (2013) Fundamentals of Homology Modeling Steps and Comparison among Important Bioinformatic Tools: An Overview. Sci Int 1(7): 237-252.

25. Macrae CF (2008) Mercury CSD 2.0 - new features for the visualization and investigation of crystal structures. J Appl Crystallogr 41(2): 466-470.

26. Trott O, Olson AJ (2010) AutoDockVina: Improving the Speed and Accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 31(2): 455-461.

27. Empereur-mot C, Guillemain H, Latouche A, Zaguray JF, Viallon V, et al. (2015) Predictiveness Curves in Virtual Screening. J Cheminform 7: 52.

28. Cramer RD, Patterson DE, Bunce JD (1988) Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. J Am Chem Soc 110(18): 5959-5967.

29. Hevener KE, Zhao W, Ball DM, Babaoglu K, Qi J, et al. (2009) Validation of molecular Docking Programs for Virtual Screening Against Dihydropteroate Synthase. J Chem Inf Model 49(2): 444-460.

30. Abu-Hammad A, Zalloum WA, Zalloum H, Abu-Sheikha G, Taha MO (2009) Homology Modeling of MCH1 Receptor and Validation by docking/scoring and protein-aligned CoMFA. Eur J Med Chem 44(6): 2583-2596.

31. Vilar S, Constanzi S (2012) Predicting Biological Activities through QSAR Analysis and Docking-based Scoring. Methods Mol Biol 914: 271-284.

32. Okada-Junior CY (2018) Phthalimide Derivatives with Bioactivity against Plasmodiumfalciparum: Synthesis, Evaluation, and Computational Studies Involving bc1 Cytochrome Inhibition. ACS Omega 3(8): 9424-9430.

33. Biamonte MA (2013) Recent Advances in Malaria Drug Discovery. Bioorg Med Chem Lett 23(10): 2829-2843.

34. Vaidya AB, Mather MW (2000) Atovaquone Resistance in Malaria Parasites. Drug Resist Updat 3(5): 283-287.

35. Barton V, Fisher N, Biagini GA, Ward SA, O'Neill PM (2010) Inhibiting Plasmodium Cytochrome *bc*$_1$: a Complex Issue. Curr Opin Chem Biol 14(4): 440-446.

36. Stewart JP (2007) Optimization of Parameters for Semiempirical Methods V: Modification of NDDO Approximations and Application to 70 Elements. J Mol Model 13(12): 1173-1213.

37. Solmar SR, Hunte C (2008) Structure of Complex III with Bound Cytochrome c in Reduced State and Definition of a Minimal Core Interface for Electron Transfer. J Biol Chem 283(25): 17542-17549.

38. Birth D, Kao WC, Hunte C (2014) Structural Analysis of Atovaquone-Inhibited Cytochrome bc1 Complex Reveals the Molecular Basis of Antimalarial Drug Action. Nat Commun 5: 4029.

39. Kayode O, Huang Z, Soares AS, Caulfield TR, Dong Z, et al. (2017) Small Molecule Inhibitors of Mesotrypsinfrom a Structure-based Docking Screen. PLoS One 12(5): e0176694.

40. Salameh MA, Radisky ES (2013) Biochemical and Structural Insights into Mesotrypsin: an Unusual Human Trypsin. Int J Biochem Mol Biol 4(3): 129-139.

41. Salmeh MA, Soares AS, Hockla A, Radisky DC, Radisky ES (2011) The P2' Residue is a Key Determinant of Mesotrypsin Specificity: Engineering High-affinity Inhibitor with Anticancer Activity. Biochem J 440(1): 95-105.

42. Katona G, Berglund GI, Hajdu J, Garf L, Szilagyi L (2002) Crystal structure reveals basis for the inhibitor resistance of human brain trypsin. J Mol Biol 315(5): 1209-1218.

43. Perilo CS, Pereira MT, Santoro MM, Nagem RAP (2010) Structural Binding Evidence of the Trypanocidal Drugs Berenil® and Pentacarinate®

Copyright© Batista VS and Nascimento-Júnior NM.

Active Principles to a Serine Protease Model. Int J Biol Macromol 46(5): 502-511.

44. deVeer SJ, Li CY, Swedberg JE, Schroeder EI, Craik DJ (2018) Engineering potent mesotrypsin inhibitors based on the plant-derived cyclic peptide, sunflower trypsin inhibitor-1. Eur J Med Chem 155: 695-704.

45. Crawford TD, Ndubaku CO, Chen H, Boggs JW, Bravo BJ, et al. (2014) Discovery of Selective 4‑Amino-pyridopyrimidine Inhibitors of MAP4K4 Using Fragment-Based Lead Identification and Optimization. J Med Chem 57(8): 3484-3493.

46. Gao X, Gao C, Liu G, Hu J (2016) MAP4K4: An Emerging Therapeutic Target in Cancer. Cell Bio sci 6: 56.

47. Gao X, Chen G, Gao C, Zhang DH, Kuan SF, et al. (2017) MAP4K4 is a novel MAPK/ERK pathway regulator required for lung adenocarcinoma maintenance. Mol Oncol 11(6): 628-639.

Copyright© Batista VS and Nascimento-Júnior NM.