A Framework for Inverse Virtual Screening

Large-Scale Protein Targets Identification

R. Vasseur[1, 2], S. Baud[1], L. A. Steffenel[1], X. Vigouroux[2], L. Martiny[1], M. Krajecki[1], M. Dauchez[1]
1-UFR Sciences Exactes et Naturelles, University of Reims (URCA), Reims, FRANCE
2-Bull SAS, Education & Research, Echirolles, FRANCE

romain.vasseur@etudiant.univ-reims.fr
stephanie.baud@univ-reims.fr
luiz-angelo.steffenel@univ-reims.fr
xavier.vigouroux@bull.net
laurent.martiny@univ-reims.fr
michael.krajecki@univ-reims.fr
manuel.dauchez@univ-reims.fr

***Abstract*—Molecular docking are widely used computational technics that allow studying structure-based interactions complexes between biological objects at the molecular scale. The purpose of the current work is to develop a framework that allows performing inverse virtual screening to test at a large scale a chemical ligand docking on a large dataset of proteins, which has several applications in the field of drug research. We developed different strategies to distribute the docking procedure, as a way to efficiently exploit the computational performance of multi-core and multi-machine (cluster) environments. This tool has been tested on 24 protein-ligand complexes taken from the Kellenberger dataset to show its ability to reproduce experimentally determined structures and binding affinities.**

Keywords—Protein-Ligand docking; inverse docking; ranking methods; distributed computations; HPC experiments.

I. INTRODUCTION

In the field of drug discovery or drug design, molecular docking is focused on protein-ligand complexes to study how the chemical ligand that is a drug will bind the target protein receptor. The prediction of the binding mode of a ligand into a protein target cavity, the structure of the complex and the estimation of the binding affinity between both partners is crucial to find new therapeutic compounds to cure life threatening diseases. Molecular docking represents a virtual alternative to costly and time-consuming systematic wet biological experiments such as High Throughput Screening (HTS) processes and/or Nuclear Magnetic Resonance (NMR)-based screening. Then, it is called Virtual Ligand Screening (VLS) or *in silico* ligand screening and has become a method of choice for rational drug design, hits identification and hits to leads optimization [1][2][3]. At present, several applications are available for virtual screening, such as PLANTS [4], DOCK Blaster [5], GOLD [6], AutoDock [7][8], FlexX [9], Glide HTVS [10], ICM [11] and LigMatch [12].

VLS tries to predict probable bindings of a huge number of ligands (to the order of millions) to a unique target receptor and is linked to multiple ligand dockings. Such methods require knowledge of the three dimensional structure of a receptor alone or associated with its experimental ligand. Many chemical databases and libraries provide millions of compounds, among which we can cite some public and free ones such as the PDBbind database [13] or the ZINC database [14], some with fees access as the Cambridge Structural Database [15] and several private pharmaceutical collections. Protein structures are obtained from the Research Collaboratory for Structural Biology (RCSB) Protein Data Bank (PDB) [16], an open source database that collects all public experimental data on tridimensional biological structures. For a large number of proteins, X-ray crystallography and NMR provide experimental structural data. In November 2013, the number of protein structures publicly available in the Protein Data Bank is over 85,000 the number of nucleic acids structures is about 2,500 and the number of structures of nucleic acids-protein complexes is about 4,000. The total number of structures available in the PDB increased on average by 6,500 structures per year during the last decade [16]. Yet, it is important to highlight that these statistics do not include the large number of proprietary structures as described above held by pharmaceutical companies that dispose of their own private structures databanks. To use non-resolved structures for a protein of interest, 3D prediction models can be built *de novo* [17] or based on partially known fragments by homology modelling [18][19].

The purpose of the current work is to develop a new virtual screening tool that allows performing large-scale structure-based inverse docking. The main idea of this approach is to test at a large scale a chemical ligand on a large dataset of proteins. In the fields of drug design and structural biology, inverse docking methodology would find several applications. It can be used to search for additional uses of new drugs, by searching for interactions with protein groups outside the usual research field. Inverse docking can also be used to identify potential side effects of new drugs or to help choosing the less harmful treatment for a disease. Several problems arise when performing inverse docking, as we are no longer targeting a single protein but thousands. One of the main concerns is the computation time, which represents a clear obstacle when dealing with a large number of different proteins. For instance, even with

 ISBN: 978-1-61208-335-3

the use of multicore processing we shall not restrain the inverse docking to a single computer but rely on multiple computational environments such as clusters and grids. In order to effectively use wide computational resources, however, we cannot simply launch a batch of docking computations but we must rethink docking in terms of task distribution, of pipelining, as well as load balance and fault tolerance. Recently, in number of works, several implementations to performed massively parallel ligand screening are reported in the literature with Message Passing Interface (MPI or openMP) only [20] or combined with multi-threading programming [21], with cloud-computing to treat Full Flexible Receptors (FFR) models [22][23] or even with FPGAs or GPUs accelerators [24].

In this work, docking simulations were performed with the AutoDock4.2 software [25] and we developed a set of Python scripts to reverse the docking process. We also developed a Python framework embedding different strategies to distribute the docking procedure, as a way to efficiently exploit the computational performance of multi-core and multi-machine (cluster) environments. Data presented in this paper result from the testing described hereafter. The experiment was conducted to compare the docked poses obtained with our tool for a set of chemical ligands on their experimental target to the determined structure of the complex obtained by X-ray crystallography. The rest of this paper is structured as follows: Section II presents the different strategies we developed to decompose the docking computation, the description of the test set and methods we used to generate and to rank the docking poses. In Section III, docking poses given by these strategies are compared to the native ones (X-ray structures). Finally, all results are afterward discussed in Section IV.

II. METHODS

A. Parallel Decomposition

To obtain a better implication of the computational resources, we must imperatively improve task parallelism when conducting large-scale inverse docking. If decomposing a docking job in parallel task may trigger a better utilization of the computational resources through pipelining and load balance, it also contributes to the fault tolerance aspects since only a small segment of the execution is lost in the case of a computer crash or execution failure. For this, we developed two methods to decompose the docking computation and improve tasks distribution and fault tolerance.

The first strategy to distribute docking computations aims at the reduction of the exploring space through the multiplication of the number of small 3D boxes. For instance, the "single grid" used in a blind docking experiment and describing the whole protein volume is arbitrary split into several grids. Each grid is a sub-volume of points covering a piece of the protein. Assuming a regular decomposition, we define a geometrical Arbitrary Cutting method (AC) as 12-part decomposition scheme, i.e., 3x2x2 (3 on the longest axis of the protein). We also tested multiple space cuttings of the whole-space to find a suitable decomposition ratio in prior experiment and the 12-part scheme showed better quality docking results than other geometrical cuttings into multiple subspaces as *n*-part schemes where n = 8 (2x2x2), 27 (3x3x3) or 64 (4x4x4) [26]. Indeed, a large number of 3D boxes may improve parallelism but the number of subspaces is also dependent on the size and shape of the protein. So, having too small 3D boxes may limit the movement of the ligand and impact the success of the ligand docking. Hence, the choice of decomposition must be carefully tuned and the number of generated chunks must be precisely balanced. Moreover, the several subspaces are overlapping each other to explore the entire protein surface and overcome the presence of the 3D boxes edges. Indeed, one of the constraints imposed by AutoDock is that the ligand cannot bind outside of the box. The overlapping is inherently dependent on the ligand size, so in our experiments we set two ranges for the partial overlapping: a third of the juxtaposed boxes if the ligand size is inferior to it, or the size of the ligand if the ligand is larger than that.

This decomposition strategy is simple to implement and the subspace grids can be easily generated from the coordinates of the protein. By multiplying the number of 3D boxes we can deploy the docking over different processors in order to be computed in parallel. One drawback of this strategy, however, is that it does not check the protein surface for cavities (which are potential docking sites), and may therefore "cut" right in the middle of a potential cavity, making it less interesting. Another drawback of this method is that only ligands inside the grid can be evaluated. Indeed, any atom of the ligand outside the 3D box will not be treated and will eliminate the pose of the conformer during the sampling process, which may prevent the detection of potential bindings when part of the ligand crosses the boundaries of the 3D box. So, to overcome boundaries problems, we also use a more rational knowledge-based method.

This second method to perform space cutting consists in predicting upstream pockets and cavities on the surface receptor with additional programs and carry out dockings only on these pockets [27][28]. For this Pocket Search method (PS), we used the Fpocket program [29] that screens pockets and cavities using a geometrical algorithm based on Voronoï tessellations. The second version of the software (Fpocket2) is compatible with a multiprocessing parallel use. Only pockets that show a long side superior to a third of the whole protein longest side and inferior to the half of the whole protein longest side are conserved as to limit the number of generated jobs and to avoid multi-exploration of the same space. One advantage of the pocket strategy is to refocuse the docking algorithm exploration zones only on predictive biological sites of interest (potential binding sites). As only these interesting zones are included in the docking procedure they can drastically improve the overall inverse docking performance. At the opposite side, the pocket search is a predictive method and as such it may exclude some potential zones, which should not be overpassed by the AC method described above.

 ISBN: 978-1-61208-335-3

B. Preparation of the Test Set

The test set used in this study is constructed from the Rognan's group [30] set of 100 protein-ligand complexes. To be able to perform accurate High Definition (HD) docking only proteins structures with a long side inferior to 60 Angstroms are conserved. Twenty-four complexes have passed this process and are included in the final test set (see TABLE I). Molecular weights of ligand molecules range from 114 to 659 Daltons, number of atoms in the ligand range from 10 to 52 and number of rotatable single bonds (rotors) in ligand molecules range from 0 to 23. All ligands molecules bind to their target protein non-covalently. Structures files and coordinates of all the complexes are downloaded from the Structural Chemogenomics Group website [30]. For the convenience of computation, each complex file was split into a protein molecule file in PDB format and a ligand molecule file, which is saved in Mol2 format. All preparation settings are available in the work from Kellenberger *et al.* [31]. The program automatically generates all docking parameters files and each complex is then subjected to an exhaustive conformational sampling procedure with AutoDock.

C. Conformationnal Sampling Procedure

The AutoDock program (version 4.2) is used to generate an ensemble of docked conformations for each ligand molecule. This program utilizes a Lamarckian Genetic Algorithm (LGA) for conformational sampling [32]. Each LGA run outputs a single docked conformation as a final result. For the AC method and the PS method 50 individual LGA runs are performed to generate 50 docked conformations for each ligand. All AutoDock docking experiments were performed with the default parameters of the Lamarckian algorithm for initial population size (*ga_pop_size = 150*), maximal number of energy evaluation (*ga_num_evals = 2500000*) and maximal number of generations (*ga_num_generations = 27000)*. The protein structure is kept fixed during docking.

D. Ranking the Best Ligand Pose

AutoDock needs to compute an affinity grid for each atomic type to pre-evaluate the binding energy. The affinity grid is contained in a 3D box that frames the protein surface. The binding energy is evaluated with a tri-linear interpolation of the eight-grid points affinity value surrounding each atom of the ligand. For the scoring step, computation time will only depend of the number of atoms in the ligand and will be independent of the protein volume. The free energy of binding ΔG is computed with the AutoDock4 scoring function (AD4) [33]. The AD4 scoring function is composed by several energy terms of classical physics force fields. The free energy of bonding is expressed by the sum of molecular mechanics components such as a dispersion-repulsion term, a term for the hydrogen bonding, a term for the electrostatics contribution, a term descripting the energy associated to bond lengths, bond angles and associated restriction entropy loss and a term for the desolvation energy (equation (1)).

$$\Delta G = \Delta G_{vdw} + \Delta G_{hbond} + \Delta G_{elec} + \Delta G_{tor} + \Delta G_{solv} \quad (1)$$

TABLE I. THE 24 EXPERIMENTAL PROTEIN-LIGAND COMPLEXES

PDB code	Res. (Å)	Protein	Ligand
1azm	2.0	Carbonic Anhydrase I	5-Acetamido-1,3,4-Thiadiazole-2-Sulfonamide
1cbs	1.8	Cellular Retinoic-Acid-Binding Protein Type II	Retinoic Acid
1ebp	2.1	Epididymal retinoic acid binding protein	Retinoic Acid
1fkg	2.0	Fk506 Binding Protein	(1R)-1,3-Diphenyl-1-Propyl(2S)-1-(3,3-Dimethyl-1,2-Dioxopentyl)-2-Piperidinecarboxylate (Rotamase Inhibitor)
1fki	2.2	Fk506 Binding Protein	(21S)-1-Aza-4,4-Dimethyl-6,19-Dioxa-2,3,7,20-Tetraoxobicyclo Pentacosane
1glp	1.9	Glutathione S-Transferase Yfyf	Glutathione Sulfonic Acid
1glq	1.8	Glutathione S-Transferase Yfyf	S-(P-Nitrobenzyl) Glutathione
1hfc	1.5	Fibroblast Collagenase	(N-(2-Hydroxymatemethylene-4-Methyl-Pentoyl)Phenylalanyl)Methyl Amine
1icn	1.7	Intestinal Fatty Acid Binding Protein	Oleate (Oleic Acid)
1lic	1.6	Adipocyte Lipid-Binding Protein	Hexadecanesulfonic Acid
1lmo	1.8	Mucopeptide N-Acetylmuramylhydrolase	Di-N-Acetylglucosamine
1mcr	2.7	Immunoglobulin delta Light Chain Dimer	N-Acetyl-L-His-D-Pro-Oh
1mmq	1.9	Matrilysin	Hydroxamate Inhibitor
1mup	2.4	Major Urinary Protein Complex	2-(Sec-Butyl) Thiazoline
1nco	1.8	Holo-Neocarzinostatin	Apo-Carzinostatin chromophore
1poc	2.0	Phospholipase A2	1-O-Octyl-2-Heptylphosphonyl-SN-Glycero-3-Phosphoenolamine
1rob	1.6	Ribonuclease A	Cytidylic Acid
1srj	1.8	Streptavidin	Naphthyl-Haba
1stp	2.6	Streptavidin	Biotin
1tng	1.8	Trypsin	Aminomethylcyclohexane
1tnl	1.9	Trypsin	Tranylcypromine
1ukz	1.9	Uridylate Kinase	Adenosine-5'-Diphosphate
3ptb	1.7	*beta*-Trypsin	Benzyldiamine
8gch	1.6	*gamma*-Chymotrypsin	Gly-Ala-Trp (peptide)

 ISBN: 978-1-61208-335-3

The best ligand poses obtained by AC and PS methods are discriminated using the best energy of binding for each method with the AD4 function. In addition, the localization of best energy docked poses is compared to the experimental pose with the measurement of the Euclidian Distance (ED) between the two ligands geometrical mass centers. When ligands are in the same binding cavity as the experimental one and the ED is lower than 2.5 Angstroms, the ligand pose is considered similar to the crystallographic pose and is called X-pose. When ligands are partially docked in the experimental cavity or able to dock in a juxtaposed cavity and ED is included between 2.5 and 8.5 Angstroms, the ligand pose is called J-pose (for Juxtaposed-pose). Beyond this value, we checked that any ligand is localized in another binding area than the experimental structure. In this case, the wrong ligand pose is called W-pose. (All of these ligands poses were checked by hand and visualized with VMD [34]). Thus, ligand pair Root Mean Square Deviation (RMSD) computation evaluates the shift between the binding conformation of the best-docked ligands and the crystallographic conformation. The RMSD corresponds to the measure of the average distance between atomic positions of two structures expressed in Angstroms as it shows in equation (2).

(2)

$$RMSD(v,w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix}-w_{ix})^2 + (v_{iy}-w_{iy})^2 + (v_{iz}-w_{iz})^2)}$$

III. RESULTS

As described above, our methodology was tested on 24 experimental protein-ligand complexes available in the PDB. Both AC and PS methods were used individually and in a combined procedure to evaluate their ability to re-dock an experimental ligand on its native protein target receptor.

For the PS method, the experiment shows that for this size of proteins (see II.A), the Fpocket algorithm found at the most five or six different well-sized pockets. TABLE II gives the volume of the three first pockets found for each experimental complex. For all proteins of the set (100%), one pocket at least is detected, for nineteen proteins in the set (19/24, 79%) two pockets are detected and for 14/24 (58%) three pockets are detected. If structures displaying at least 4 pockets are selected, the ratio of the set falls down to 9/24 (37.5%) and decreases even more when considering a higher number of pockets. Thus, it appears that for each protein-ligand complex selecting only the first pocket found by the Fpocket algorithm is enough to consider the whole set; the results point that selecting at most the three first pockets should refine the search. In addition, the number of jobs launched partly depends on the number of pockets that will be explored. Thus, the number of jobs launched is precisely defined for each complex.

A fixed number of jobs can be very interesting to monitor the speed-up and the scalability of the program over a variant number of available cores. In theory, the optimal load balance should be reached if the number of available cores is superior or equal to the number of launched jobs. So, to optimize the computation time we should set the best ratio jobs/cores and to do this a fixed number of jobs is necessary. For example, this set of complexes generates a pool of maximum 360 jobs (24 complexes x (12 AC method boxes + 3 pockets boxes from the PS method at the most)). So, the best energy structure of the ensemble of the twelve boxes is conserved for the AC method and the best energy structure of each of the first three pockets is conserved for the PS method. Finally, four docked poses at the most are obtained for each complex, which will be compared with the experimental ligand pose of the crystallographic ligand-protein complex. Previously, we define that the re-docking is successful if an X-pose or a J-pose were obtained for the ligand (see II.D).

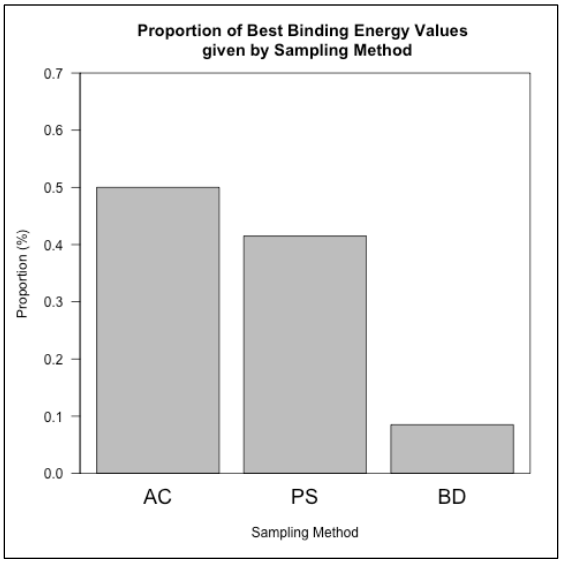

Figure 1. Proportion of Best Binding Energy Values given by the Sampling Method (AC: 62.5%, PS: 29%, BD: 8.5%).

Firstly, the results for AC and PS methods are compared with the corresponding Blind Docking experiment (BD). Blind Docking was introduced to detect possible binding sites and ligands binding modes by scanning the entire surface of protein targets [35][36]. This represents the "naïve" approach to dock ligands on unknown targets but is barely parallelizable. In fact, for each complex the AutoDock software will launch only one infrangible docking task with the whole volume to explore. Depending on the shape of each receptor, a large number of runs/generations is required in order to systematically cover the entire protein surface and consequently to obtain good docking results.

 ISBN: 978-1-61208-335-3

For twelve experiments out of the set (12/24, 50%) the best energy score was obtained by the PS method, for 10/24 (41.5%) it was obtained by AC method and only for 2/24 (8.5%), it was obtained by BD experiment (Figure 1). Moreover, the combined results of AC method and PS method give a better energy of docking for 22/24 (91.5%) compared to BD. Furthermore, for 54% of the cases the combined methods gave a RMSD between the experimental structure and the best docking pose lower than 5 Angstroms and a RMSD lower than 10 Angstroms for 23/24 (96%) versus only one for BD (4%) in both case (TABLE III). These results highlight that our methods perform better exploration of the protein surface. Indeed, the ratio (volume/number of runs) explored in the case of our methodology is better optimized than in the case of BD. Both methods ensure a better conformational sampling and a better quality of docking pose than using the BD.

TABLE II. NUMBER OF POCKETS DETECTED FOR EACH PROTEIN AND THEIR VOLUMES

	Pocket 1 (PS1)	**Pocket 2 (PS2)**	**Pocket 3 (PS3)**
PDB	***Volume (Å^3)***	***Volume (Å^3)***	***Volume (Å^3)***
1azm	833	786	244
1cbs	1626	378	557
1ebp	1262	370	616
1fkg	549	**N/A**	**N/A**
1fki	576	756	**N/A**
1glp	1307	370	640
1glq	607	637	686
1hfc	762	683	485
1icn	1655	**N/A**	**N/A**
1lic	978	927	**N/A**
1lmo	1306	143	561
1mcr	676	192	**N/A**
1mmq	409	276	548
1mup	479	583	756
1nco	350	**N/A**	**N/A**
1poc	1016	504	642
1rob	654	576	686
1srj	408	**N/A**	**N/A**
1stp	367	**N/A**	**N/A**
1tng	647	610	**N/A**
1tnl	602	466	512
1ukz	600	1072	**N/A**
3ptb	549	328	529
8gch	765	619	383

The distribution of docked poses depending on the sampling method associated with the best energy is presented in Figure 2. For 18/24 (75%) the sample methods that give the best free energy of binding give also the best docking poses (X-pose or J-pose) distributed as follows: 7/18 (39%) for AC method and 10/18 (55%) for the PS method

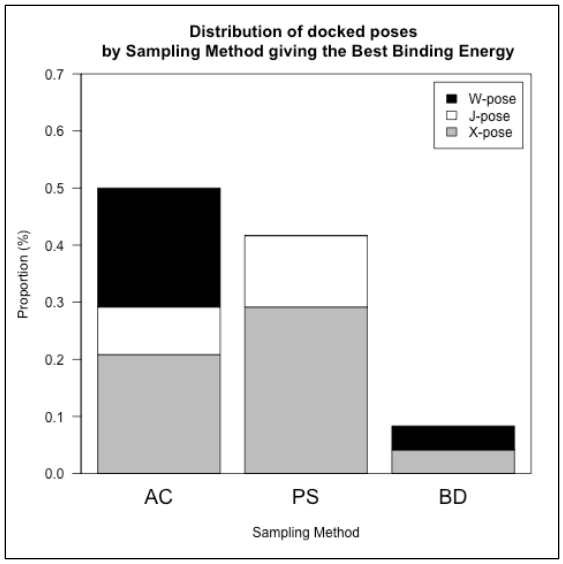

Figure 2. Distribution of docked poses (X-pose in grey, J-pose in white and W-pose in black) by Sampling Method giving the Best Binding Energy.

and 1/18 (6%) for the BD experiment. Among these complexes, the combined method that gives the best free energy of binding gives also the best docking pose for 17 (94.5%) versus only one for BD (5.5%). From TABLE III, we can extract the following correlation: comparing the docked poses at rank 1 of Euclidian distance and rank 1 for the lowest RMSD value, there is a match for 6/24 (25%) in the case where a J-pose is observed and for 14/24 (58.5%) in the case where an X-pose is observed. So, at rank 1 for the two previous criteria, the ligand docked poses (X-poses and J-poses) give the lowest RMSD value for 18/24 (75%). Comparing the docked poses at rank 1 of Euclidian distance and rank 1 and 2 for the lowest RMSD value the proportion reach 22/24 (91.5%). The match ratio is distributed by sampling method as follows: The AC method gives the X-pose for 3 complexes with a mean RMSD value equal to 2.32 Angstroms compared to the experimental structures (1glp, 1mup, 1tnl). The AC method gives also a J-pose for 3 complexes (1hfc, 1icn, 1rob) and an associated RMSD value equal to 7.82 Angstroms compared to the experimental structures. Nevertheless, it is important to mention that for 1hfc and 1icn poses are reverse poses that is to say the ligand acquires a head to tail conformation compared to the experimental one so the RMSD increases. The PS method gives the X-pose for 11 complexes (1azm, 1cbs, 1ebp, 1fkg, 1fki, 1mmq, 1nco, 1stp, 1tng, 3ptb, 8gch). In these cases, the mean RMSD with the experimental structure is 2.93 Angstroms. The PS method gives a J-pose for 4 complexes (1lic, 1mcr, 1poc, 1ukz) and an associated mean RMSD value with the experimental structure of 5.54 Angstroms (Figure 3). The BD method gives an X-pose for 1srj with a RMSD value of 2.47 Angstroms. If the rank 2 for the Euclidian distance is also considered, the PS method is able to replace the ligand for 1srj in an X-pose with 2.35

 ISBN: 978-1-61208-335-3

Angstroms of RMSD. So, the combined method with these evaluation criterions gives the best pose for 22/24, 91.5% of the cases of the total set.

TABLE III. EVALUATION CRITERIONS OF THE SAMPLING METHODS

	Energy (kcal/mol)		**RMSD (Angstroms)**				**Gravity Centers Euclidian Distance (Angstroms)**					
	Rank 1		***Rank 1***		***Rank 2***		***Rank 1***			***Rank 2***		
PDB	*Method*	*Value*	*Method*	*Value*	*Method*	*Value*	*Method*	*Value*	*Pose*	*Method*	*Value*	*Pose*
1azm	AC	-5.15	PS1	1.95	**N/A**	**N/A**	PS1	1.12	X-pose	**N/A**	**N/A**	**N/A**
1cbs	PS2	-6.84	PS2	2.24	AC	8.86	PS2	1.17	X-pose	PS1	1.59	X-pose
1ebp	PS2	-8.68	PS2	2.00	AC	2.73	PS2	0.77	X-pose	PS1	1.23	X-pose
1fkg	PS1	-5.96	PS1	5.49	AC	8.22	PS1	1.43	X-pose	AC	3.98	J-pose
1fki	PS1	-10.49	PS1	0.60	PS2	1.75	PS1	0.59	X-pose	PS2	1.00	X-pose
1glp	AC	-4.46	AC	2.71	PS1	5.42	AC	0.76	X-pose	PS1	2.74	X-pose
1glq	BD	-3.66	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
1hfc	AC	-4.78	AC	8.75	**N/A**	**N/A**	AC	5.43	J-pose	**N/A**	**N/A**	**N/A**
1icn	AC	-3.97	AC	8.80	**N/A**	**N/A**	PS1	3.49	J-pose	AC	3.66	J-pose
1lic	PS1	-4.65	PS1	5.75	**N/A**	**N/A**	PS1	3.63	J-pose	AC	4.23	J-pose
1lmo	AC	-3.26	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**	**N/A**
1mcr	AC	-4.03	PS2	4.41	**N/A**	**N/A**	PS2	2.81	J-pose	**N/A**	**N/A**	**N/A**
1mmq	AC	-6.31	AC	3.97	PS1	4.16	PS1	0.79	X-pose	AC	1.59	X-pose
1mup	AC	-4.23	AC	2.59	PS1	4.04	AC	1.55	X-pose	PS1	2.02	X-pose
1nco	PS1	-7.19	PS1	7.83	**N/A**	**N/A**	PS1	2.10	X-pose	AC	8.22	J-pose
1poc	PS1	-1.91	PS1	6.71	**N/A**	**N/A**	PS1	3.95	J-pose	**N/A**	**N/A**	**N/A**
1rob	PS2	-5.29	AC	5.91	PS1	9.89	AC	5.32	J-pose	PS2	8.05	J-pose
1srj	BD	-7.48	PS1	2.35	BD	2.47	BD	0.45	X-pose	PS1	1.23	X-pose
1stp	PS1	-6.10	PS1	1.34	AC	2.42	PS1	0.37	X-pose	AC	0.55	X-pose
1tng	PS1	-5.87	PS1	1.05	AC	1.53	PS1	0.63	X-pose	AC	0.83	X-pose
1tnl	AC	-5.96	AC	1.68	PS1	2.44	AC	0.35	X-pose	PS1	0.41	X-pose
1ukz	AC	-6.74	PS1	5.31	**N/A**	**N/A**	PS1	3.39	J-pose	**N/A**	**N/A**	**N/A**
3ptb	AC	-5.52	PS1	1.52	AC	2.07	PS1	0.19	X-pose	AC	0.23	X-pose
8gch	AC	-5.00	PS1	4.32	**N/A**	**N/A**	PS1	1.03	X-pose	**N/A**	**N/A**	**N/A**

a. **N/A: Non Applicable** data – RMSD or Euclidian Distance > 10 Angstroms

 ISBN: 978-1-61208-335-3

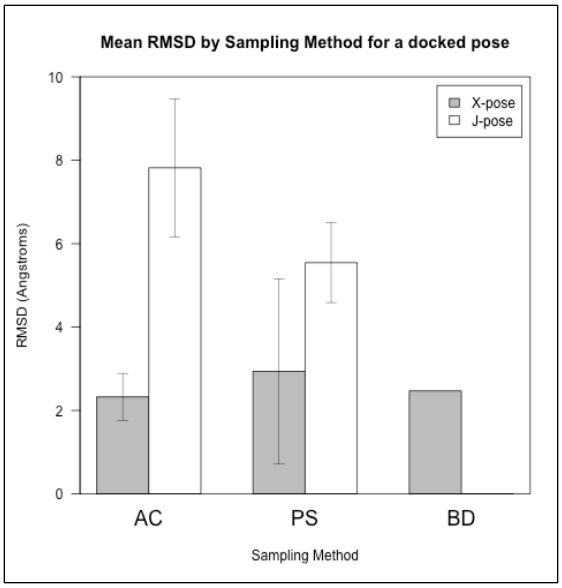

Figure 3. Mean RMSD (in Angstroms) for an X-pose (in grey) and a J-pose (in white) by Sampling Method at rank 1 of Euclidian distance and rank 1 and 2 of RMSD.

Figure 4 shows the results obtained with the AC method for the experimental complex 1stp. An X-pose with an Euclidian distance between ligands geometrical mass centers of 0.55 Angstroms (rank 1) with a RMSD value of 2.42 Angstroms (rank 1) is observed. As we can see on Figure 4 and Figure 5 with two different types of protein representations, the re-docked ligand reached successfully the experimental cavity of binding and adopts a similar conformation compared to X-ray structure. On Figure 4, the New Cartoon style represents only the secondary structure of the backbone skeleton of the protein whereas on Figure 5, all amino-acids side chains are included to build the protein surface thanks to the MSMS algorithm. The local structure of side chains creates reliefs and since some of them display specific chemical properties, they can arrange themselves in binding cavities. The ligand pose and conformation in the binding site will be related to the cavity geometry. As we can see in Figure 4, a good ligand pose implies a chemical conformation that precisely place the chemical groups implied in Hydrogen bonds in an appropriate range of distance (around 2.0 Angstroms). Hydrogen bonds are strong dipole-dipole interactions between electro-negative atoms, and according to local chemical composing they are partially in charge of ligand docking in a binding pocket. For ligands from seven complexes, there is a match between RMSD and mass centers distance but not between both and the best binding energy. In all cases the pose giving the best energy is localized in different from cavities that the crystallographic ones. These results can be explained by several settings of using decomposing method (Figure 5). For 1azm, the best energy is obtained with the box-11 of the AC method (-5.15 kcal/mol) whereas best RMSD with an X-pose is obtained by the PS method (PS1). The AC pose is localized in a different cavity from the crystallographic one. The box-11 dimensions do not allow to include the crystallographic area and they do not permit to refind the experimental pose. On the other side, the PS1 box dimensions do not allow to refind the AC pose cavity neither. The experimental cavity (S1) is included in an another AC box, box-7. The ligand pose obtained with this box is localized in the same cavity as the previous AC box (S2) and presents a better energy than PS1 pose. If we set the dimensions of a tuned box able to include the two binding sites S1 and S2, the ligand pose obtained binds into S1 with even better energy of -5.26 kcal/mol. Finally, to maximise the number of energy evaluations and the conformationnal sampling, we carried out a 256 runs on the previous tuned box and anew the crystallographic cavity is obtained with a poorer energy compared to S1 of -4.49 kcal/mol. So, just the box boundaries presence is not enough to conclude, 1azm complex may wrong prepared or this case shows the limits of the AutoDock force-field.

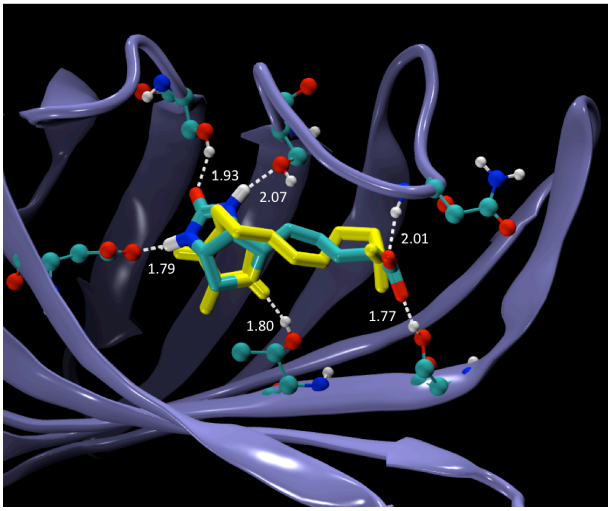

Figure 4. 1STP -- Streptavidin (New Cartoon, in purple)/Biotin (Licorice, X-ray in cyan, X-pose in yellow) protein-ligand complex stabilized by hydrogen bonds in the binding site.

Crystallographic pose refinding may be precluded by boxes boundaries but it is also impacted by protein shape specifications. In fact, for 1ukz, the cavity is closer to a funnel with a long and slight pipe that sinks into the protein structure. The experimental ligand is housed at the bottom of the pipe in a burried area in the protein core. Fpocket detects the left large extremity as part as a full binding pocket (PS2) and the hidden area as an another binding pocket (PS1). The AC box (giving the best energy) only takes in the funnel cavity and does not include the burried site (like PS2 does) and inversely PS1 includes the crystallographic cavity but does not take in the large surface cavity. It explains why there is no match between the AC method that gives the better energy and the PS1 X-pose.

 ISBN: 978-1-61208-335-3

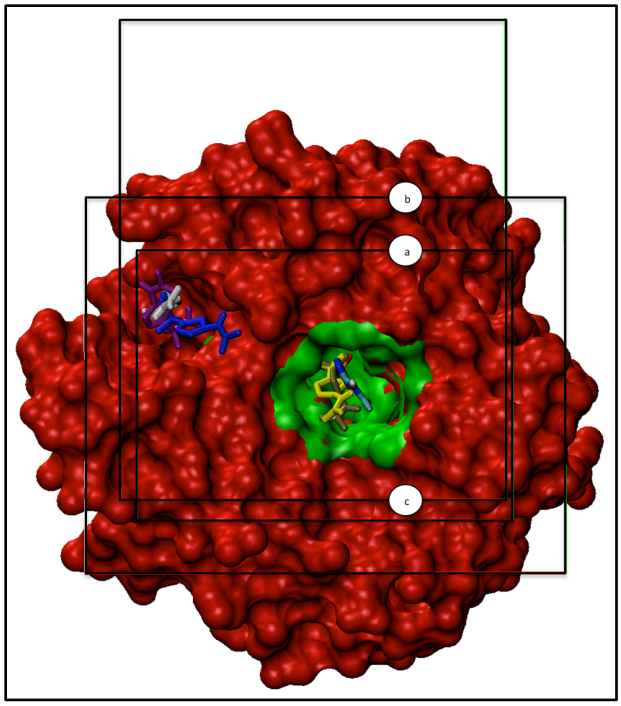

Figure 5. 1azm ligands in the crystallographic cavity (MSMS, in green): X-ray pose (in cyan), PS1 pose (in tan), Tuned box-256R pose (in yellow) and 1azm ligands in another cavity: AC box-11 pose (in blue), AC box-7 pose (in purple), Tuned box-50R pose (in purple) bound on the whole Carbonic Anhydrase I protein receptor with a: PS1 pocket box, b: Tuned box, c: AC box-7

Failed dockings can be explained by protein shape specifications but also by ligand chemical structure. Some ligands as 1lmo or 1rob are very exposed in large valleys at the protein surface, which are correctly identified by the Fpocket program as a binding pocket but the docking program could fail to place correctly the ligand on a planar surface. Else, the chemical nature of ligands could increase the docking process weakness: 1lmo ligand is a big flexible di-saccharide and 1rob ligand is an ADN nucleoside both containing –ose residues hard to treat with the Autodock force field.

Only for 1glq in the test set, the best energy value is given by the blind docking experiment (-3.66 kcal/mol). The ligand pose is neither in the crystallographic cavity neither in any pocket cavity and binds on a relative open cavity. However, 1glq and 1glp are two crystallographic structures of the same protein with about the same degree of resolution complexed with two similar ligands (see TABLE I). For all that, Fpocket is not able to find precisely the same pockets in 1glq so the boundaries are not exactly at the same place and do not allow to retrieve the experimental pose with the PS method. The AC method does but the energy of binding is worse than for the pose obtained by the blind docking. Nevertheless, if we launch multiple blind docking experiments, this artefact binding mode should not be retrieved several times.

1stp and 1srj are two crystallographic structures of the same protein (see TABLE I) with a large variation in resolution neatness. In fact, if a structure with a higher resolution than 2.0 Angstroms is available it is assumed that a structure with a lower resolution degree is a worse structure. In this case the two structures do not show remarkable difference of structure of the binding site. For the binding site in 1stp, the protein-ligand complex is the well known Streptavidin/Biotin complex in which the protein have a β-barrel secondary structure. This complex is one of a strongest non-covalent interactions known in nature. It is used extensively in molecular biology as a marker. The ligand fits perfectly in the binding site and the interactions are stabilized through a complex network of Hygrogens bonds. For 1stp, the experimental ligand was well replaced by the PS1 method (0.37 Angstroms) and AC method (0.55 Angstroms) with the best binding energy equal to -6.10 kcal/mol for PS1. For 1srj the ligand is Naphtyl-Haba docked in the same cavity as Biotin. It was well re-placed by the blind docking experiment (0.45 Angstroms) and PS1 method (1.23 Angstroms) with the best binding energy equal to -7.48 kcal/mol for BD. This results could be explained by the asymetric shape of the protein that confers a geometry less spherical than a regular globular protein. Consequently, the long axis of the protein takes a high value and imposes the same grid spacing as the others proteins. But in this case, the surface to explore included in the blind docking box is less important and the majority of the grid points are not on the protein surface. So, the ratio volume/runs is very high and the algorithm explore much more precisely the binding pocket and leads to the best energy pose with the maximum goodness.

IV. DISCUSSION/CONCLUSION

In order to be able to treat many hundred proteins computations on High Performance Computing (HPC) architectures, we developed a set of methods to parallelize the treatment of each protein, as well as to distribute the tasks among a given set of machines as a way to speed up the overall execution of the inverse docking. For this, we developed a framework that can embed the AC and the PS method to explore as best as possible the protein surface and rationally dock the ligand into the binding cavity.

Our results show that the methods we are developing perform better volume exploration with a better ratio volume/runs than a classical blind docking experiment. In fact, to perform an accurate high definition docking we have to deal with coherent grid spacing. By default, AutoDock builds affinity grids with a spacing of 0.375 Angstroms that corresponds to a quarter of the bond between two atoms of Carbone. We defined a spacing interval between 0.375 and 0.450 in which we consider the accuracy of the simulation as a HD docking. The main drawback of this method is that AutoDock is able to build and also explore a 3D box of 126 x 126 x 126 points at the most. So, only a protein whose long axis is lower than 60 Angstroms can fit into the grid box.

 ISBN: 978-1-61208-335-3

Considering this kind of protein for a blind docking experiment, the AutoDock program is also limited in the number of simulations runs, that is to say in the number of times the initial LGA is reinterred (256 runs max.). So, AC method considers the BD box volume cut into 12 sub-boxes with a partial overlap. Each sub-box is explored by the LGA with 50 runs of simulations that roughly correspond to the half of the ratio volume/runs for the BD. Whereas the ratio is more difficult to precisely estimate, it is even better with PS method, which explains the effectiveness of the program to perform better exploration and to obtain better docking quality results than BD experiments.

As many docking programs [37][38], we have shown that our framework is a successful tool to re-place correctly the ligand into the active site of the target receptor in a non-covalent manner. Furthermore, it is also able to predict accurate ligands bindings independently of active site knowledge [39]. For this, we evaluated a good docking pose using three criteria: free energy of binding, Euclidian distance between mass centers and RMSD of the re-docked ligand with respect to the crystallographic ligand. Combinations of these criteria are able to discriminate right docking poses from experimental data. The combination between the binding energy and the RMSD (rank1 and 2) is able to discriminate 66.5% of the test set and the one between the mass center distance (rank1 and 2) and the RMSD (rank1 and 2) is able to discriminate 91.5% of the test set. On the other side, the ratio is 75% for the combination of binding energy and center of mass distances (rank1 and 2) and 71% for the combination of the triad. This is explained by the nature of the evaluation criteria. RMSD and mass centers distances are implicitly correlated because they both describe a space position. Mass centers distances describe a space position for the entire ligand whereas RMSD describe a space position for each atom of the ligand, both always in respect to the experimental structure. In fact the RMSD reflects the ligand structure in a local environment, its capacity to adapt itself to the binding cavity. Consequently, taking into account the numbers of atoms implied both in the binding site and in the ligand structure and the number of torsions available for the ligand, the probability to obtain a low RMSD in a different cavity than the crystallographic binding site is close to zero. This is well shown in TABLE III, for 8 cases out of 9 if the RMSD is higher than 10 Angstroms the corresponding mass center distance is higher than 10 Angstroms too (N/A data). That explains the good ratio for these criteria combination. On the other side, the space position adopted by the ligand in the binding site translated by the RMSD value impacts the chemical match between chemical groups able to make non-covalent interactions (Hydrogen bonds, van der Waals forces and electrostatics) with atoms in the binding cavity. These forces represent a major contribution into the energy function that is used to evaluate the free energy of binding (see II.D). So, the ratio of the combination of RMSD and energy of binding can be explained partially by this relationship.

Nonetheless, in this experiment we have shown that we reproduce ligands experimental poses with our framework. As the references are experimental data, we dispose of comparison elements (RMSD and mass centers distances). The results obtained in this study (distances determining X-pose and J-pose and associated RMSD) are good enough to validate the method for detecting workable binding sites. To identify already known binding sites or new ones the aim of this program is to perform predictive experiments on large sets of proteins for a given ligand of interest. For these, we will only dispose of the free energy of binding to discriminate good docking poses. For 7/24 there is no match between the binding energy and the geometric criterions. In some remarkable cases we have shown previously, only the free binding energy computation does not allow to retrieve similar poses to the crystallographic ones. That is demonstrating that the evaluation of the binding energy is not an absolute reference. To reduce the unsuccessful ratio we have to reinforce the ranking evaluation process by adding other scoring methods able to make up rare cases of force field failures. However, in most of the cases we have seen that the PS method strongly performs to detect druggable cavities on a protein receptor. In fact some proteins present multiple binding sites well described in enzymology allosteric phenomena especially. The advantage of using multiple pockets search is to identify well differentiated multiple sites on the fly during a unique docking simulation. That allows us to consider ligand repositioning experiments and also second targets and off-targets hunting. In addition the AC method is able to overcome the PS method failures with adding search completeness and not excluding planar binding surfaces such as protein-protein binding area in particular. So, we demonstrate that the combination of the two methods is an accurate strategy to identify new protein targets for a given ligand.

We developed an effective tool to perform large-scale inverse virtual screening works on both HPC hardware and personal computer able to identify proteins targets for a chemical ligand of interest. Originally developed for and with AutoDock4.2, the framework will embed a version with AutoDock Vina [40] as docking engine that supports multithreading natively but does not allow fine-grain control of algorithm parameters contrary to the previous AutoDock software.

ACKNOWLEDGMENT

The experiments presented on this paper were carried out on the ROMEO Computing Center (https://romeo.univ-reims.fr) and the Multi-scale Molecular Modelling Platform (https://p3m.univ-reims.fr). Romain Vasseur is founded by a Bull SAS CIFRE grant. We thank Romulo Gadelha de Moura Lima, for his great contribution to format the code.

REFERENCES

[1] R. Abagyan and M. Totrov, “High-throughput docking for lead generation,” Curr. Opin. Chem. Biol., vol. 5, no 4, 2001, pp. 375-382.

[2] D. Giganti *et al.*, "Comparative Evaluation of 3D Virtual Ligand Screening Methods: Impact of the Molecular Alignment on Enrichment", J. Chem. Inf. Model., vol. 50, no 6, 2010, pp. 992-1004.

[3] G. Klebe, "Virtual ligand screening: strategies, perspectives and limitations," Drug Discov. Today, vol. 11, no 13-14, 2006, pp. 580-594.

[4] O. Korb, T. Stützle, and T. E. Exner, "Empirical Scoring Functions for Advanced Protein−Ligand Docking with PLANTS," J. Chem. Inf. Model., vol. 49, no 1, 2009, pp. 84-96.

[5] J. J. Irwin *et al*., "Automated Docking Screens: A Feasibility Study," J. Med. Chem., vol. 52, no 18, 2009, pp. 5712-5720.

[6] G. Jones, P. Willett, and R. C. Glen, "Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation," J. Mol. Biol., vol. 245, 1995, pp. 43-53.

[7] S. Cosconati *et al*., "Virtual screening with AutoDock: theory and practice," Expert Opin. Drug Discov., vol. 5, no 6, 2010, pp. 597-607.

[8] S. Zhang, K. Kumar, X. Jiang, A. Wallqvist, and J. Reifman, "DOVIS: an implementation for high-throughput virtual screening using AutoDock," BMC Bioinformatics, vol. 9, no 1, 2008, pp. 126.

[9] M. Rarey, B. Kramer, T. Lengauer, and G. Klebe, "A fast flexible docking method using an incremental construction algorithm," J. Mol. Biol., vol. 261, no 3, 1996, pp. 470-489.

[10] R. A. Friesner *et al*., "Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein−Ligand Complexes," J. Med. Chem., vol. 49, no 21, 2006, pp. 6177-6196.

[11] Y. Y. Li, J. An, and S. J. Jones, "A computational approach to finding novel targets for existing drugs," Plos Comput. Biol., vol. 7, no 9, 2011, pp. e1002139.

[12] S. L. Kinnings and R. M. Jackson, "LigMatch: A Multiple Structure-Based Ligand Matching Method for 3D Virtual Screening," J. Chem. Inf. Model., vol. 49, no 9, 2009, pp. 2056-2066.

[13] R. Wang, X. Fang, Y. Lu, and S. Wang, "The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures," J. Med. Chem., vol. 47, no 12, 2004, pp. 2977-2980.

[14] J. J. Irwin and B. K. Shoichet, "ZINC-a free database of commercially available compounds for virtual screening," J. Chem. Inf. Model., vol. 45, no 1, 2005, pp. 177-182.

[15] F. H. Allen, "The Cambridge Structural Database: a quarter of a million crystal structures and rising," Acta Crystallogr. B, vol. 58, no 3, 2002, pp. 380-388.

[16] RCSB Protein Data Bank - http://www.rcsb.org/pdb

[17] P. Bradley, K. M. Misura, and D. Baker, "Toward high-resolution *de novo* structure prediction for small proteins," Science, vol. 309(5742), 2005, pp. 1868-1871.

[18] J. Moult, "A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction," Curr. Opin. Struct. Biol., vol. 15, no 3, 2005, pp. 285-289.

[19] N. Eswar *et al*., "Comparative protein structure modeling using Modeller," Curr. Protoc. Bioinforma., 2006, pp. 5-6.

[20] A. P. Norgan, P. K. Coffman, JP. A. Kocher, D. J. Katzmann, and C. P. Sosa, "Multilevel parallelization of AutoDock 4.2," J. Cheminf., vol. 3(1), no 1, 2011, pp. 1-9.

[21] X. Zhang, S. E Wong, and F. C. Lightstone, "Message passing interface and multithreading hybrid for parallel molecular docking of large databases on petascale high performance computing machines", J. Comp. Chem., vol. 34, no 11, 2013, pp. 915-927.

[22] R. De Paris, F. A. Frantz, O. Norberto de Souza, and D. A. Ruiz, "wFReDoW: A Cloud-Based Web Environment to Handle Molecular Docking Simulations of a Fully Flexible Receptor Model", BioMed Res. Inter., 2013, pp. 1-12.

[23] T. Y. Tsai, K. W. Chang, and C. Y. Chen, "iScreen: world's first cloud computing web server for virtual screening and de novo drug design based on TCM database@Taiwan," J. Comput. Aided Mol. Des., vol 25, 2011, pp. 525-531.

[24] I. Pechan and B. Fehér, "Molecular docking on FPGA and GPU platforms", International Conference on Field Programmable Logic and Applications, 2011.

[25] G. M. Morris *et al*., "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," J. Comput. Chem., vol. 30, no 16, 2009, pp. 2785-2791.

[26] R. Vasseur *et al*., "Parallel strategies for an inverse docking method," Proc. PBio: International Workshop on Parallelism in Bioinformatics, EuroMPI User's Group Meeting (EuroMPI 2013), ACM, Sept. 2013, pp. 253-258.

[27] D. Ghersi and R. Sanchez, "Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites," Proteins Struct. Funct. Bioinforma., vol. 74, no 2, 2009, pp. 417-424.

[28] C. Hetényi and D. van der Spoel, "Toward prediction of functional protein pockets using blind docking and pocket search algorithms," Protein Sci., vol. 20, no 5, 2011, pp. 880-893.

[29] V. Le Guilloux, P. Schmidtke, and P. Tuffery, "Fpocket: An open source platform for ligand pocket detection," BMC Bioinformatics, vol. 10, no 1, 2009, pp. 168.

[30] Structural Chemogenomics Group - http://bioinfo-pharma.u-strasbg.fr

[31] E. Kellenberger, J. Rodrigo, P. Muller and D. Rognan, "Comparative evaluation of eight docking tools for docking and virtual screening accuracy," PROTEINS: Struct. Funct. Bioinf., vol.57, 2004, pp. 225-242.

[32] G. M. Morris *et al*., "Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function," J. Comp. Chem., vol. 19, no 14, 1998, pp. 1639-1662.

[33] R. Huey, G. M. Morris, A. J. Olson, and D. S. Goodsell, "A semiempirical free energy force field with charge-based desolvation," J. Comput. Chem., vol. 28, no 6, 2007, pp. 1145-1152.

[34] W. Humphrey, A. Dalke and K. Schulten, K., "VMD - Visual Molecular Dynamics," J. Molec. Graphics, vol. 14, 1996, pp. 33-38.

[35] C. Hetényi and D. van der Spoel, "Blind docking of drug-sized compounds to proteins with up to a thousand residues," Febs Lett., vol. 580, no 5, 2006, pp. 1447-1450.

[36] B. Iorga, D. Herlem, E. Barré, and C. Guillou, "Acetylcholine nicotinic receptors: finding the putative binding site of allosteric modulators using the blind docking approach," J. Mol. Model., vol. 12, no 3, 2005, pp. 366-372.

[37] I. W. Davis, K. Raha, M. S. Head, and D. Baker, "Blind docking of pharmaceutically relevant compounds using RosettaLigand," Protein Sci., vol. 18, no 9, 2009, pp. 1998-2002.

[38] A. Grosdidier, V. Zoete, and O. Michielin, "Blind docking of 260 protein-ligand complexes with EADock 2.0," J. Comput. Chem., vol. 30, no 13, 2009, pp. 2021-2030.

[39] R. Vasseur *et al*., "Parallel strategies for an inverse docking method," J. Parall. Comput., special issue on Parallelism in Bioinformatics, in press.

[40] O. Trott and A. J. Olson, "AutoDock Vina: Improving the speed and accuracy od docking with a new scoring function, efficient optimization, and multithreading," J. Comput. Chem., vol 31, 2010, pp. 455-461.

 ISBN: 978-1-61208-335-3