A Cloud-Based Workflow Approach for Optimizing Molecular Docking Simulations of Fully-Flexible Receptor Models and Multiple Ligands

Renata De Paris*†, Duncan A. D. Ruiz* and Osmar Norberto de Souza‡
*Grupo de Pesquisa em Inteligência de Negócio - GPIN
Faculdade de Informática, PUCRS, Porto Alegre, RS, Brasil
E-mail: renata.paris@acad.pucrs.br and duncan.ruiz@pucrs.br
†School of Computing Science
Newcastle University, Newcastle upon Tyne, UK
‡Laboratório de Bioinformática, Modelagem e Simulação de Biossistemas - LABIO
Faculdade de Informática, Porto Alegre, RS, Brasil
E-mail: osmar.norberto@pucrs.br

Abstract—**The use of conformations achieved from Molecular Dynamics (MD) simulations in docking experiments is the most accurate approach to simulate the natural interactions between receptor and ligands at molecular environments. However, such simulations are computational costly and their overall execution may become unfeasible due to the large quantities of structural information needed to represent a Fully-Flexible Receptor (FFR) model. The problem is even more challenging when FFR models are used to perform docking-based virtual screening in a large database of ligands. This study aims at developing a cloud-based workflow to efficiently optimize docking experiments between a FFR model and multiple ligands in two strategic ways: by discarding groups of unpromising MD conformations for specific ligands at docking execution time, and by exploiting on-demand resources from the Microsoft Azure cloud platform. The proposed environment is built on e-Science Central, which is a powerful cloud-based workflow enactment system designed to handle scientific high-throughput tasks. As a result, we expect to reduce the number of docking experiments per ligand without affecting the quality of the produced models and, therefore, considerably decreasing the time consumed by docking experiments.**

Keywords—Scientific Workflow, Cloud Computing, Molecular Docking Simulations, Fully-Flexible Receptor Model.

I. Introduction

With the advances in X-ray crystallography and nuclear magnetic resonance spectroscopy, the number of protein structures available in biological databases have been constantly increasing. This is encouraging the pharmaceutical industry since it expects to benefit from the wealth of data on new targets and hence produce newer and more efficient drugs [1]. However, these advances have an effect on the costs involved in for the discovery of novel potential drugs, especially those to perform molecular docking-based virtual screening. Molecular docking simulation, which is the main step of Rational Drug Design (RDD) [2], predicts a binding mode and affinity of a given small molecule for a target receptor mimicking the *in-vitro* and *in-vivo* screening by using docking software [1], [3] such as AutoDock4.2 [4].

Most studies in the field of molecular docking simulations have only focused on applying the simple “lock-and-key” mechanism. In this approach, experiments are performed between rigid receptors and flexible ligands since almost all docking software are only capable of treating the rotatable bonds of small molecules. The main challenge faced by docking software is to consider the explicit flexibility of receptors. One reason behind this limitation is the significant computational complexity that is required to include all degrees of conformation freedom in the target protein. Proteins are inherently flexible systems and have an intrinsic ability to undergo functionally relevant conformation transitions under native state conditions on a wide range of scales, both in time and space. Moreover, realistic docking simulations need to take into account the molecular flexibility, for both receptor and ligand, to address their explicit plasticity during their interactions [5]. One of the most affordable and accurate method for identifying alternative binding forms of proteins at different timescales is Molecular Dynamics (MD) simulation [3], [5]. The result of a MD simulation is a series of instant conformations of the protein receptor along the simulation time scale. The set of snapshots or conformations produced by MD simulations is called a Fully-Flexible Receptor (FFR) model [6]. The main challenge faced by incorporating a FFR model in docking experiments is the high computational cost of performing and analyzing each of the receptor conformation and ligand interactions.

Several attempts have been made to answer the following question: How can the computational efficacy, and prediction accuracy of molecular docking simulations based on FFR models be balanced to perform practical virtual screening in a large library of ligands? Previous studies have reported efforts in strategically selecting a small number of MD conformations before starting the docking experiments [7] [8]. Landon et al. [9] represented 90% of the conformational flexibility within ligand-binding site of a 160 ns MD trajectory of the H5N1 into reduced ensembles of 10 apo and 5 holo structures by using clustering algorithms. Recently, researchers have showed an increased interest in reducing the number of MD structures during the docking experiments. That, however, raises another question: How to select an accurate representation of the receptor and to generate a Reduced Fully-Flexible Receptor

978-1-4673-9560-1/15 $31.00 © 2015 IEEE
DOI 10.1109/CloudCom.2015.43

(RFFR) model tailored to specific ligands? A way to answer this question was presented by Machado et al. [10]. They developed FReDoWs, a scientific workflow that automates docking experiments and performs a selective docking execution by identifying the best Free Energy Binding (FEB) values, which are achieved from exhaustive docking experiments between the FFR model and a ligand with similar structure. Despite the high accuracy showed by its results, FReDoWs may become time-consuming as the number of dissimilar small molecules and receptor snapshots increase. Another way to answer the second question is to make use of wFReDoW [11], our previous work. In that work, we developed a HPC environment with tens of nodes of the Amazon Elastic Compute Cloud to reduce both the dimensionality of FFR models and the overall docking execution time by using a strategically approach to identify promising snapshots. On average, wFReDoW was able to reduce from days to hours the overall time execution, maintaining over 95% of accuracy in a 3.1 ns MD trajectory of InhA enzyme from *Mycobacterium tuberculosis* [12]. Nevertheless, wFReDoW neither supports docking-based virtual screening of FFR models, nor it is be able to scale on demand high performance computing facility due to the limitations of the MPI cluster model.

This PhD thesis intends to address the obstacles described above by using a set of computational techniques to develop a cloud-based workflow approach to efficiently handle molecular docking simulations of FFR models and multiple ligands. In particular, this work will address the time-consuming computations by balancing the volume of molecules to be docked based on the available cost budget and applying the Self-adaptive Multiple Instances Pattern (P-SaMI) [13]. P-SaMI is a data-flow pattern that strategically discards unpromising docking solutions of a specific ligand in real-time docking experiments. This environment is deployed on e-Science Central (e-SC), a cloud-based workflow enactment system [14] which supports: a) scaling the application out onto cloud resources depending on the volume of data and the available financial budget; b) managing the data and services required to optimize docking experiments of FFR models based on the P-SaMI data-flow; and c) providing a web interface for users to configure the environment, analyze docking results and trace the data before, during and after the experiments. As a result, it is expected to considerably reduce the total time spent in docking-based virtual screening of FFR models so that, new ever-growing virtual libraries of ligands can be exploited more effectively and, therefore, helping to improve the whole RDD process.

II. Proposed Methodology

This section describes the computational techniques selected to identify groups of promising MD conformations for specific ligands with scalable high throughput screening deployed on the cloud. As preliminary studies, the clustering of snapshots generated to be used as input data to the cloud-based workflow is described in De Paris et al. [8] and the application of this clustering in a case-study approach is presented by Quevedo et al. [15].

A. Generating RFFR Models Tailored for Specific Ligands

According to Amaro et al. [16] better theories and more efficient computational schemes are still needed to allow the selection of most relevant conformations to ligand binding in a predictive manner. In this thesis it is proposed to make use of a partitioning of snapshots achieved with the k-means algorithm and submit it in P-SaMI, a data-flow pattern to perform massively parallel docking experiments of FFR models [13]. P-SaMI is the approach applied to discard clusters with unpromising receptor conformations, taking into account docking results from processed snapshots at docking execution time. To ensure the quality of P-SaMI assessments, it is important to have high affinity between conformations belonging to the same partition. For this reason, the similarity measures used by the k-means algorithm is a set of essential properties extracted from the substrate-binding cavity of every MD conformation. Thus, if a receptor conformation achieves a good docking result - FEB value significantly negative - for a unique ligand, it is possible to consider that other conformations belonging to the same group will also interact favorably for this ligand. One RFFR model is composed by the set of processed snapshots of the FFR model for one ligand.

Basically, P-SaMI procedures are to split clusters of snapshots into batches before starting the experiments and to define batches of un/promising snapshots. As shown in Figure 1, P-SaMI is based on docking results from snapshots that have been docked (processed) previously, where a priority and status is defined for every batch to prioritize the execution of those snapshots that are in batches with good docking results. Each batch ranges the priority from 1 to 3 (1-high, 2- medium and 3-low) and the status may be classified as “A” (Active), “F” (Finalized) and “D” (Discarded). A batch with non-promising snapshots gives low priority and, eventually, is discarded by P-SaMI. Thus, if a substantial amount of conformations from batch L presents favourable interaction with a ligand, it is assumed that L contains promising snapshots; otherwise, it can be discarded prior to the end of whole execution as L contains unpromising snapshots. P-SaMI defines the status and priority for batch L by evaluating estimated sampling average $\overline{esa}_{ij}$ according to equation (1):

$$\overline{esa}_{ij} = \frac{(\overline{sum}_{ij} + (0.4985 * \#F_{ij} * (2 * \overline{sa}_{ij} - \overline{sd}_{ij})))}{\#L_{ij}} \quad (1)$$

where i denotes the i-cluster, j identifies the j-batch belonging to the cluster i, F is the set of remaining snapshots to be processed from batch L_{ij}, $\overline{sum}$ is the score results of the processed snapshots from L_{ij}, $\overline{sa}$ and $\overline{sd}$ are the average and the standard deviation of score results from batch L_{ij}.

To ensure the quality of the clustering of snapshots, we conducted a set of experiments using a 20ns MD trajectory of the enzyme Enoyl-Reductase or InhA-NADH complex from *Mycobacterium tuberculosis* (PDB ID: 1ENY) [17]. In that simulation, data were collected at every 1ps, resulting in a FFR model with 20,000 instantaneous receptor structures. Firstly, the clustering of snapshots was generated considering similarity measures extracted from the following substrate-binding cavity properties: the accessible surface area and volume; the heavy atoms; and the pairwise RMSD values relative to the first snapshot of the FFR model. De Paris et al. [8] applied Davies-Bouldin, Dunn’s and Gap statistic clustering validity criteria to identify the ideal partitioning. Moreover,

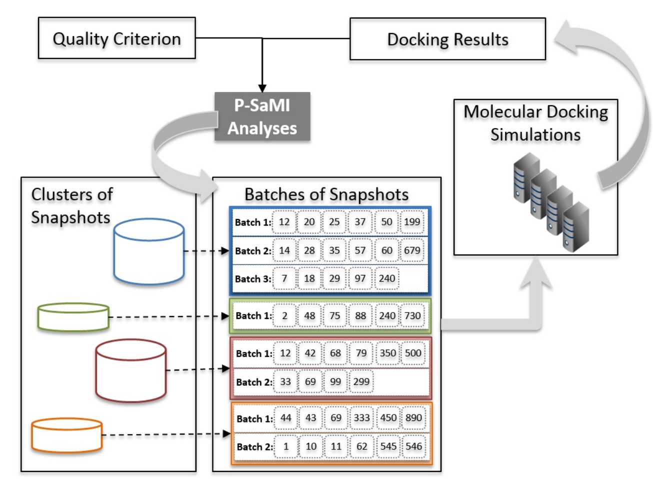
Quality Criterion
Docking Results
P-SaMI Analyses
Molecular Docking Simulations
Clusters of Snapshots
Batches of Snapshots
Batch 1: 12 20 25 37 50 199
Batch 2: 14 28 35 57 60 679
Batch 3: 7 18 29 97 240
Batch 1: 2 48 75 88 240 730
Batch 1: 12 42 68 79 350 500
Batch 2: 33 69 99 299
Batch 1: 44 43 69 333 450 890
Batch 2: 1 10 11 62 545 546

Fig. 1. Model of a P-SaMI data-flow execution. By spreading the clusters of snapshots into batches, P-SaMI makes decision on small chunks of processed snapshots and access alternative samples of snapshots inside the same cluster.

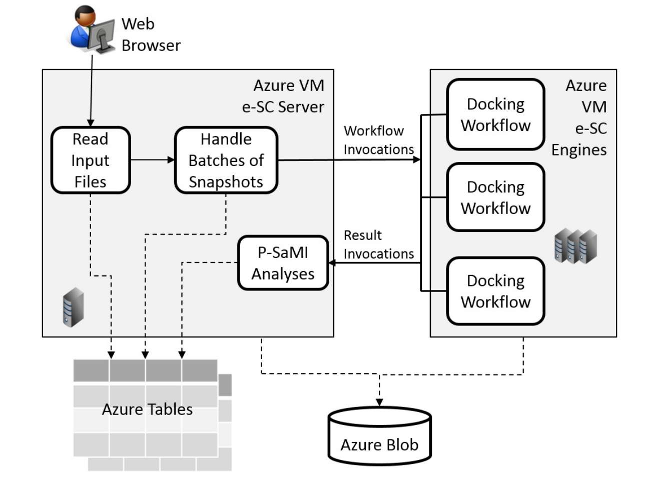

Fig. 2. The design of the proposed cloud-based scenario workflow to optimize docking-based virtual screening by considering FFR models. Data and control flows are monitored by e-SC which is also responsible for scaling Azure VMs.

to validate the accuracy of the best partitioning solution, exhaustive docking experiments between the FFR model and a set of 20 ligands for which the binding mode is known were performed and analyzed. Although the target cavity of FFR models should be known in advance, statistic assessments in terms of FEB variance values show that the best clustering solution accurately identified protein structural changes that occurs inside the binding cavity for specific ligands.

Further experiments were performed to validate the accuracy of the clustering of snapshots described above. The triclosan (TCL from PDB ID: 2B35) and isoniazid (INH-NAD from PDB ID: 2NV6) ligands, two known ligands of the InhA structure, were applied to assess the quality of the RFFR models produced by using the clustering of snapshots as input for P-SaMI on wFReDoW [11]. The high performance was confirmed running wFReDoW for INH-NAD, where P-SaMI discarded 68% of MD conformations of the FFR model. Indeed, 100% of the top 10 and 96% of the top 100 snapshots with the best FEB values were processed. These experiments, which are reported in Quevedo et al. [15], were valuable to demonstrate the successful running of P-SaMI when a clustering of snapshots generated by similar features in the substrate-binding cavity are used as input data.

B. The Workflow Scenario Applied for Performing Docking Experiments between FFR Models and Multiple Ligands

In this study e-Science Central (e-SC) [14] is the workflow enactment system used to orchestrate the large scale docking applications and reduce the overall time taken to perform docking-based virtual screening of FFR models based on the P-SaMI data-flow. Each workflow in e-SC is composed of blocks connected into a direct acyclic graph. The invocation of a workflow enables a sequence of blocks that run when all input ports have input data buffered and ready to use [18]. A block is a thin layer of an API that allows users to customize their own algorithms as well as to access input data and properties to generate result invocations. This enactment system is particularly useful to improve the performance of large scientific experiments in the cloud. Recently, Cala et al. [18] designed a scalable system based on e-SC for reducing the time taken to generate QSAR models in the Azure cloud platform. By assessing the scalability of prediction of chemical activity experiments they demonstrated efficiency of 90% with 200 worker nodes on Azure. Another advantage of applying e-SC system in this study is its web interface which provides communication between the users and the system before, during and after the run time. This allows us to concentrate our efforts on deploying the system without worrying about how to create easy-to-use graphical interfaces.

The proposed cloud-based workflow for optimizing molecular docking simulations between FFR models and multiple ligands is shown in Figure 2. Figure 3 details the Docking Workflow which performs molecular docking simulations using AutoDock4.2 [4]. The workflow execution starts when the user introduces all input data needed to configure the P-SaMI and AutoDock. For P-SaMI data-flow the inputs are the clustering of snapshots, the best and worst FEB values assigned by a specific ligand, the percentage of snapshots to be placed in each batch, and the percentage of processed snapshots per batch to start the analyses. The ligand and the FFR model paths, the set of receptor conformations as well as the autogrid and autodock parameters are docking inputs that also need to be configured by a user before starting the experiments.

Each experiment starts by reading a clustering of snapshots, the ligand path and P-SaMI parameters. Based on P-SaMI parameters the Read Input Files workflow splits every cluster into batches where its snapshots are stored with a specific batch number, status and initial priority in two Azure Tables: P-SaMI and Snapshots. The P-SaMI table stores information used to identify promising snapshots taking into account docking results already obtained from the processing batches whereas the Snapshots table stores docking results of every MD conformation and ligand processed. The P-SaMI table sorts the data stored by the batch priority to speed up processing snapshots with high priority, which in turn are also considered promising snapshots for a specific ligand. The

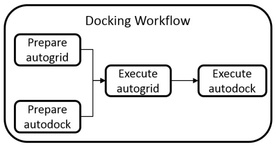

Fig. 3. Molecular docking simulation process for a MD conformation and a small molecule based on AutoDock4.2 [4] features.

Handle Batches of Snapshots workflow takes a percentage of snapshots from batches that are in the top of P-SaMI table and invokes Docking Wokflows to be launched on Azure virtual machines. Figure 3 shows the main docking steps performed by the Docking Workflow, where the docking parameters and the receptor conformation and ligand paths are inputs to the workflow. Invocation results the workflow are sent to the input port the P-SaMI Analyses wokflow. This workflow assesses docking results by using equation (1) to update the status and priority of batches analyzed on P-SaMI table. The input data and the three main workflows are stored on the e-SC Server which controls enactment of the whole application. Azure Blob stores all data files needed to run the cloud-based workflow, including input and output docking files. In order to avoid unnecessary storage charges, temporary files used by AutoDock4.2 are deleted after each docking execution.

III. Conclusion

This paper presents an outline of the significant problems in performing docking-based virtual screening of FFR models and some solutions that have been made to reduce its high computational cost. In addition, it includes a description of the proposed methodology and the preliminary experiments for validating parts of this methodology. The preliminary results confirmed the high accuracy achieved by running P-SaMI when a clustering of snapshots with high affinity in their partitions is used as input data. As future steps, we intend to develop the cloud-based workflow in e-SC and to move it to the Microsoft Azure cloud platform to improve performance, while the number of docking tasks increases. The experiments will start by reducing to hours or days the overall elapsed time of 1.5 years taken when a FFR model of 20,000 receptor snapshots and a set of 20 known ligand are performed on a single processor. After that, we intend to enlarge such experiments by using a similar FFR model and sets of ligands extracted from large database, such as ZINC [19].

Acknowledgment

We thank Paul Watson for the opportunity to work with his research team and Jacek Cala for his valuable assistance on e-Science Central. This work was performed with the support from National Council for Scientific and Technological Development (CNPq-Brasil) and FAPERGS.

References

[1] V. Garg, S. Arora, and C. Gupta, "Cloud computing approaches to accelerate drug discovery value chain," *Combinatorial chemistry & high throughput screening*, vol. 14, no. 10, pp. 861–871, 2011.

[2] I. D. Kuntz, "Structure-Based Strategies for Drug Design and Discovery," *Science*, vol. 257, no. 5073, pp. 1078–1082, 1992.

[3] H. Alonso, A. A. Bliznyuk, and J. E. Gready, "Combining docking and molecular dynamic simulations in drug design," *Medicinal research reviews*, vol. 26, no. 5, pp. 531–568, 2006.

[4] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility," *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–2791, 2009.

[5] P. Cozzini, G. E. Kellogg, F. Spyrakis, D. J. Abraham, G. Costantino, A. Emerson, F. Fanelli, H. Gohlke, L. A. Kuhn, G. M. Morris, M. Orozco, T. A. Pertinhez, M. Rizzi, and C. A. Sotriffer, "Target flexibility: An emerging consideration in drug discovery and design," *Journal of Medicinal Chemistry*, vol. 51, no. 20, pp. 6237–6255, 2008.

[6] K. S. Machado, A. T. Winck, D. D. Ruiz, and O. Norberto de Souza, "Mining flexible-receptor docking experiments to select promising protein receptor snapshots," *BMC Genomics*, vol. 11, no. Suppl 5, pp. 1–10, 2010.

[7] S. Tian, H. Sun, P. Pan, D. Li, X. Zhen, Y. Li, and T. Hou, "Assessing an ensemble docking-based virtual screening strategy for kinase targets by considering protein flexibility," *Journal of chemical information and modeling*, vol. 54, no. 10, pp. 2664–2679, 2014.

[8] R. De Paris, C. V. Quevedo, D. D. Ruiz, O. Norberto de Souza, and R. C. Barros, "Clustering molecular dynamics trajectories for optimizing docking experiments," *Computational intelligence and neuroscience*, vol. 2015, 2015.

[9] M. R. Landon, R. E. Amaro, R. Baron, C. H. Ngan, D. Ozonoff, J. Andrew McCammon, and S. Vajda, "Novel druggable hot spots in avian influenza neuraminidase h5n1 revealed by computational solvent mapping of a reduced and representative receptor ensemble," *Chemical biology & drug design*, vol. 71, no. 2, pp. 106–116, 2008.

[10] K. S. Machado, E. K. Schroeder, D. D. Ruiz, E. M. Cohen, and O. Norberto de Souza, "FReDoWS: a method to automate molecular docking simulations with explicit receptor flexibility and snapshots selection," *BMC Genomics*, vol. 12, no. Suppl 4, pp. 2–13, 2011.

[11] R. De Paris, F. A. Frantz, O. Norberto de Souza, and D. D. Ruiz, "wFReDoW: A cloud-based web environment to handle molecular docking simulations of a fully flexible receptor model," *BioMed Research International*, vol. 2013, pp. 1–12, 2013.

[12] E. K. Schroeder, L. A. Basso, D. S. Santos, and O. Norberto de Souza, "Molecular Dynamics Simulation Studies of the Wild-Type, I21V, and I16T Mutants of Isoniazid-Resistant *Mycobacterium tuberculosis* Enoyl Reductase (InhA) in Complex with NADH: Toward the Understanding of NADH-InhA Different Affinities," *Biophysical Journal*, vol. 89, no. 2, pp. 876–884, 2005.

[13] P. Hübler, D. Ruiz, J. E. Ferreira, and O. Norberto de Souza, "P-SaMI: a data-flow pattern to perform massively-parallel molecular docking experiments using a fully-flexible receptor model," in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*. ACM, 2015, pp. 54–57.

[14] H. Hiden, S. Woodman, P. Watson, and J. Cala, "Developing cloud applications using the e-science central platform," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 371, no. 1983, p. 20120085, 2013.

[15] C. V. Quevedo, R. De Paris, D. D. Ruiz, and O. Norberto de Souza, "A strategic solution to optimize molecular docking simulations using fully-flexible receptor models," *Expert Systems With Applications*, vol. 41, no. 16, pp. 7608–7620, 2014.

[16] R. E. Amaro and W. W. Li, "Emerging methods for ensemble-based virtual screening," *Current Topics in Medicinal Chemistry*, vol. 10, no. 1, pp. 2–13, 2010.

[17] F. Gargano, A. L. Costa, and O. Norberto de Souza, "Effect of temperature on enzyme structure and function: a molecular dynamics simulation study," *Annals of the 3rd International Conference of the Brazilian Association for Bioinformatics and Computational Biology, So Paulo, Brazil*, 2007.

[18] J. Cała, H. Hiden, S. Woodman, and P. Watson, "Cloud computing for fast prediction of chemical activity," *Future Generation Computer Systems*, vol. 29, no. 7, pp. 1860–1869, 2013.

[19] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: a free tool to discover chemistry for biology," *Journal of chemical information and modeling*, vol. 52, no. 7, pp. 1757–1768, 2012.