Université Libre de Bruxelles

Rational drug deisgn and PKPD modeling

CHIM-F-4001

RDD Report

Author:
Charlotte Nachtegael

Supervisor:
Martine Prevost

May 2017

1 Introduction

During this work, we study the modeling of the protein-ligand docking of the reverse transcriptase with a non-nucleoside inhibitor, the nevirapine. We start from existing structures, ensuring that we can compare the results obtained with the simulations with high confidence to the reference. We worked with the structure 1VRT on PDB (`http://www.rcsb.org/pdb/explore.do?structureId=1vrt`), regrouping the reverse transcriptase, the DNA and the nevirapine.

2 Methods

2.1 Docking program

We use for this work the AutoDock (ADT) program [1]. ADT is based on a structure-based ligand docking method. The receptor is considered as a rigid structure where a molecular docking will be performed with the ligand with some degree of freedom. The objective function is calculated with the help of a grid limiting the search space to the presumed binding site, where pre-calculation of different energies such as electrostatic or Van Der Waals energies is performed in order to reduce the computation time. The aim of the algorithm is to optimize the objective function. Different algorithms can be used to perform this task. Parameters such as the size of the grid or the choice of the algorithm and its parameter for the optimization can influence the quality of the results.

2.2 Choice of study

We decided to not change any parameters of the grid and rather study the Lamarckian genetic algorithm (LGA) for the optimization, as well as its parameters [2]. The LGA creates a population of different random solutions for the docking and optimizes them with random moves until a stop criterion is encountered such as a specific number of iterations without improvement of the objective function. The fittest individuals are then selected for reproduction to replenish the population after the death of the non-selected individuals. The individuals are then submitted

to a mutation with a specific probability. The new population is improved with the local search algorithm again and the process is repeated until the stop criterion of the LGA is reached. Moreover, this algorithm is combined with a Multi-Deme principle, meaning that five populations evolves as explained for the LGA algorithm in parallel and after some iterations, migrations between the populations are allowed in order to bring some randomness and variety in the different populations to escape the local minimas. Migration here consists in exchanging the information of the least fittest individual with the information of fittest individual of the fittest population.

2.3 Docking process

The first part consists to clean the structures from the water molecules and obtain the receptor and the ligand in separate files. Hydrogen bonds are added to the structure. Actually, as we need to calculate a electrostatic term, we need charges. So by adding these polar hydrogens, we can attribute to each atoms charges and report the charges of the hydrogens to the heavy atoms they were bonded to, merging them in the process. The receptor is considered as a rigid structure, so no change is applied to the structure. As for the ligand, we have to determine the rotatable bonds in order to maybe limit the degrees of freedom of the ligand. Here however, the nevirapine had only one rotatable bond, so we did not have to touch to the degrees of freedom. Afterwards we define the binding site with the grid. The grid is a volume containing pseudo-atoms virtually representing the atoms of the ligand. Properties are attributed to these atoms (all the atoms have a charge of +1 for example) and are used to calculate the energies between the pseudo-atoms and the surrounding atoms of the receptor. These values are then extrapolated to calculate the objective function during the docking of the ligand. The parameters of the docking algorithm are then determined. We tried to both increase and decrease each parameter individually in order to evaluate their influence on the quality of the results (Table 1).

Parameters	**Default**	**Increase**	**Decrease**
Number of runs	10	100	1
Number of generations	27,000	270,000	2,700
Population size	150	1500	15
Number of best individuals kept	1	5	0
Crossover rate	0.8	0.9	0.4
Mutation rate	0.02	0.2	0.01
Mean Cauchy	0	0.5	N/A
Variance Cauchy	1	2	0.5
Number of generations for selection of worst	10	20	5

Table 1: Parameters changed during our simulations

2.4 Analysis

The structures obtained are ranked and clustered together according to their RMSD. We use the RMSD between the structure and the reference structure, the RMSD between the structure the best structure of the cluster, as well as the docked energy of the clusters to compare the different results obtained with the different changes in the parameter.

3 Results

3.1 Default run

The best docked energy is of -8.63 and the worst is -8.61. The RMSD-reference is in the range of 0.57 to 0.64. The lowest RMSD observed were the furthest from the best structure of the cluster with 0.11 and 0.18. We only have the first cluster in the results.

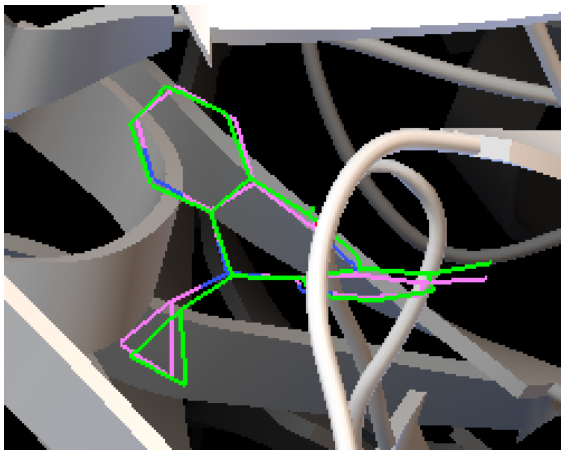

Figure 1: Best structure of the default LGA analysis (green) in the binding site superimposed with the structure of reference (pink).

3.2 Number of runs

The consequence of increasing or decreasing the number of runs is simply impacting on the number of results we obtain. When we increased the number of runs, we found a similar range for the docking energy than with the default number of runs (-8.63 to -8.61), with a notable exception at -8.53 with a exceptional RMSD-reference of 0.38. In general, we obtain several times better RMSD reference than what we observed with the default run. With only one run, the chance that the result would be better than the default is thin.

3.3 Number of generations

This determine how long will the LGA last. Increasing this parameter does not yield better results than the default run with respect to the docking energy, and even yield worst results with the RMSD-reference with 0.58 to 0.67. However, decreasing the number of generations yield the same range of docked energy, but their best RMSD-reference is lower than the default run with 0.55. This phenomenon could be linked to the fact that sometimes good results could be lost due to the genetic operators. So when the number of the generations decreased, you decrease also the probability to lose good patterns over time.

3.4 Population size

Increasing the size of the population allows a greater diversity in the solutions yielded for each generation. On the other hand, decreasing that same parameter could lead to a faster convergence of the population. With the increase, we obtained increased docking energy compared to the default run (up until -8.09), but the RMSD-reference is much lower with 0.358. We can also observe for the first time the presence of other clusters in the results. We also observed a "upside-down" structure (Fig 2) with a RMSD of 0.47 and a energy docking of -8.44. The decrease of the population size yielded us extremely similar results of energy docking always equal to -8.63 and a RMSD-reference of 0.61 or 0.62. So increasing the population size is preferred in order to introduce as much randomness as possible to escape the local minimas.

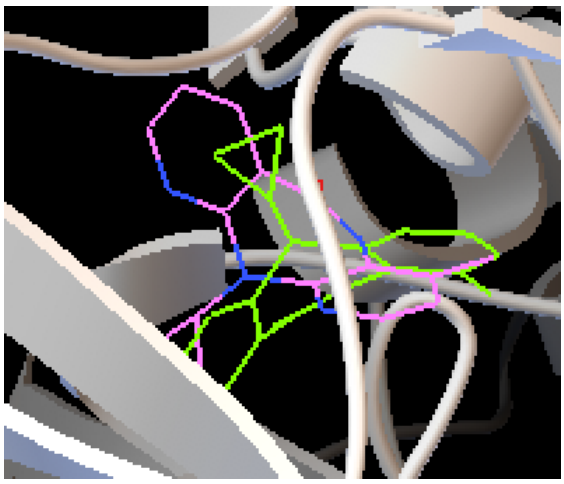

Figure 2: Upside down structure observed (green) in the binding site superimposed with the structure of reference (pink).

3.5 Number of best individuals kept

This is an elitism factor. It determines how many of the best individuals you keep as it is for the next generation. When increasing this factor, you promote the convergence towards one specific pattern that yielded good results, meaning that it is more difficult to escape the local minimas. We thusly observed similar results as observed when decreasing the population size. When no individuals is kept for the

next generation, the results are of really poor quality for both the docking energy and the RMSD-reference. Indeed, you do not keep the best individual in order to improve in a iterated way, obtaining each generation completely new individuals.

3.6 Crossover rate

Crossover rate is the expected number of pairs in the population that will exchange genetic material. Increasing the rate yielded similar docking energy to the default run, but with better RMSD-reference between 0.54 to 0.65. When decreasing this parameter we did not observe differences from the default run.

3.7 Mutation rate

The mutation rate increase the probability to mutate the children obtained after the reproduction between individuals kept for the new population. Increasing the parameter yielded both really good and really bad energy docking compared to the default run. We also observed an increased number of results from the second cluster. Decreasing the parameter yielded similar results to the increase of the number of best individual kept and to the decrease of the population size. Because of the lack of randomness insert to escape the local minimas, the results converged towards them and stagnate there much more easily.

3.8 Mean Cauchy and Variance Cauchy

The cauchy variable corresponds to the distribution from which we sample random numbers. Changing these parameters did not give different results from the default run. This could be due to the fact that we did not choose values different enough to observe a difference or this could be due to the fact that it is linked to the draw of a random number, which would be difficult to compare between runs.

3.9 Number of generations for the selection of the worst

The worst individuals are selected as below a specific threshold which is calculated according to a specific number of previous generations. Decreasing the parameter means that you take into account only the close past, meaning that the threshold will be very stringent as the population improves iteratively, while increasing means that the past were the results were worst are also taken into account given a more lax threshold. The decrease of the parameters yielded similar docked energies, but better RMSD-reference in general compared to the default run (0.55 to 0.6). The increase of the parameter gave no real differences compared to the default run.

4 Conclusion

The choice of the parameters will influence the results, some more than others. This choice will be led by many different motivations, such as the strategy of the algorithm, the computation time desired,... By superimposing the structures obtained with the reference, we could observe that we obtained really close results to the reference, except for the rare instances where we discovered a "upside-down" structure.

References

[1] Michel F. Sanner, Arthur J. Olson, and Jean-Claude Spehner. Reduced surface: An efficient way to compute molecular surfaces. *Biopolymers*, 38 (3):305–320, 1996. ISSN 1097-0282. doi: 10.1002/(SICI)1097-0282(199603)38:3⟨305::AID-BIP4⟩3.0.CO;2-Y. URL `http://dx.doi.org/10.1002/(SICI)1097-0282(199603)38:3<305::AID-BIP4>3.0.CO;2-Y`.

[2] Jan Fuhrmann, Alexander Rurainski, Hans-Peter Lenhof, and Dirk Neumann. A new lamarckian genetic algorithm for flexible ligand-receptor docking. *Journal of Computational Chemistry*, 31(9):1911–1918, 2010. ISSN 1096-987X. doi: 10.1002/jcc.21478. URL `http://dx.doi.org/10.1002/jcc.21478`.