ABSTRACT

This work describes the design and implementation of a controlled genetic algorithm for a molecular docking process between protein-ligand complexes. This algorithm was designed to keep a range of values for a selection pressure, and genotypic and phenotypic diversity, and aimed to prevent premature convergence or being trapped in a local minimum energy. During the evolution process, these parameters are checked in each generation and when one of them is out of the allowed interval, a controlled correction is applied to the population by replacement a number of individuals in a random mode. The best one is never replaced with this strategy. The controlled correction can be applied in a finite number of times to allow that this algorithm stops when the diversity is too small. Additionally, this designed genetic algorithm allows a refinement in the search of the best protein-ligand complex. This algorithm is divided into three zones, where each one develops a specific number of energy computations and a particular step for the angle rotation of the rotatable bonds present in the ligand. Transition from one zone to another needs a specific conversion operation of the values inside the chromosome. Finally, for the validation of this algorithm, 10 simulations were performed for two complexes: 1AI5 and 2TMN, and the obtained values showed an average RMSD value lower than 2.5 Å.

1. INTRODUCTION

Computational processes used to find energetically favorable complex between two types of molecules is called molecular docking. Depending on the input of these molecules, docking tools can be classified into two types. The first one refers to macromolecular docking, in which two macromolecules such as proteins or DNA are put to interact between them. The second one refers to docking of small molecules, in which a macromolecule binds to a small molecule. In the case of drug design, this process is known as docking between one protein and a ligand [1].

Molecular docking is a useful tool in drug design because it allows making a cheap and efficient search of protein-ligand complexes in a large virtual space. On the other hand, artificial intelligence techniques have become useful tools in solving such problems. In recent years, many approaches have been developed to solve the problem of molecular docking. In this sense, we can remark simulated annealing, tabu search, genetic algorithms, among others. All these techniques have obtained satisfactory results in simulations of protein-ligand interactions [2,3,4].

2. METHODS

Genetic algorithm has as inputs the plain text files that contain both three dimensional structure of the protein and the ligand, step of the rotation angles for the rigid ligand in the spatial axes, minimum rotation of rotatable bonds for three zones of the algorithm, and the bias for the selection pressure, genotypic and phenotypic diversity. Moreover, this algorithm contains percentage of crossover and mutation, list of amino acids that are part of the active site of the protein, size of the bounding box and number of energy computations for each zone of the algorithm.

Results from the algorithm developed include plain text files of the best protein-ligand complexes and their energy values in kcal/mol. The energy is calculated using the AMBER force field [5].

In a semi-flexible molecular docking process, the protein remains rigid in the space while the tridimensional structure of the ligand changes according to the rotatable bonds that this holds. Therefore, the designed chromosome includes a bit for each rotatable bond that the ligand possesses.

In this study, the molecular docking process was performed using a three-dimensional grid, called bounding box, which contains all atoms of the amino acids belonging to the active site of the protein. Because of the position of the ligand inside the active site of the protein should be optimized, it is necessary to include three bits in the chromosome. This indicates the three-dimensional grid point with coordinates \((x, y, z)\) in which the centroid of the ligand will be located. Another important aspect is the orientation of the ligand into the bounding box, because the ligand is a rigid structure. This orientation can be changed...
by rotating all the atoms with respect to the three spatial axes. For this reason, it must be included three additional bits into the chromosome to indicate the rotation of the rigid structure of the ligand in the x, y and z axes, respectively.

Finally, in each generation the selection pressure, the genotypic and the phenotypic diversity are calculated. If the obtained value for any of these parameters is lower than the established threshold for each one, a controlled correction process replaces a number n of parents with n new individuals. This action can be performed by a finite number of times, defined by the user, to allow that the algorithm will stop for characteristic convergence criteria. The genotypic diversity is measured as the variance of the sum of the bits of the chromosome for each individual and the phenotypic diversity is measured as the variance of the fitness from individuals.

3. SIMULATIONS AND RESULTS

Accuracy of complex structures obtained in the simulations can be assessed by computations of the Root Mean Square Deviation (RMSD), given by (1), between the resulting structure and the crystal one. For each complex, 10 simulations were performed and the average of the RMSD was calculated. The generation size was set to 400 individuals, and the number of energy computations was set to 2000, 2000 and 5000 for each zone. The crossover rate was set to 0.8 and a mutation rate of 0.5 was used, because once the adjustment of the chromosomes is done, it is necessary a high percentage of mutation, which it allows new individuals to obtain new rotations with the new minimum rotation of the rotatable bonds. The setting values were the following: 2, for the selection pressure, and 0.2 for both the genotypic and phenotypic diversity.

As an example, in Table 1 are shown RMSD computations for two protein-ligand complexes.

\[ \text{RMSD} = \sqrt{\frac{\sum_{i=1}^{N_{\text{at}}} d_i^2}{N_{\text{at}}}} \]  

Table 1. Preliminary results for 10 simulations of two protein-ligand complexes

<table>
<thead>
<tr>
<th>PDB id</th>
<th>protein-ligand complex</th>
<th>1AI5</th>
<th>2TMN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average ligand RMSD (Å)</td>
<td>2.49</td>
<td>2.07</td>
<td></td>
</tr>
<tr>
<td>Standard deviation RMSD</td>
<td>0.24</td>
<td>0.18</td>
<td></td>
</tr>
<tr>
<td>Average complex RMSD (Å)</td>
<td>0.14</td>
<td>0.12</td>
<td></td>
</tr>
<tr>
<td>Standard deviation RMSD</td>
<td>0.01</td>
<td>0.01</td>
<td></td>
</tr>
</tbody>
</table>

4. CONCLUSIONS

From the results we can see that the controlled genetic algorithm designed allows simulation of semi-flexible molecular docking successfully. Moreover, the algorithm allowed the searching for a more stable complex because at the end of a zone of the genetic algorithm. Individuals who survived are refined through adjustment of the chromosomes, crossover and the mutation, where the most important part in the process of refinement for the new individuals generated is the last operation, because that is the one that allows the conversion of the chromosomes in the different zones. The control action over the genetic algorithm was a good strategy to reduce the probability of convergence in a local minimum or a premature convergence.

The application of a controlled correction with a 30 percent of replacement of the actual population is a good strategy to avoid the convergence to a local minimum. However, if the new individuals have a small fitness value compared with the average fitness of the actual population, the algorithm will converge to the previous solution, the one found before the corrective action.

Finally, preliminary results suggest that a controlled genetic algorithm applied to molecular docking simulations is an approach that should be explored, aimed to achieve better results and optimize the design of computer-aided drug design.

5. REFERENCES


