Brenda G. Medina
Department of Computer Science
University of Texas at El Paso
El Paso, TX 79968
bmedina3@miners.utep.edu

Mentors: Michela Taufer and Trilce Estrada
Department of Computer and Information Sciences
University of Delaware
Newark, DE 19716
taufer@cis.udel.edu, trilce.cs@gmail.com

Towards Automatic Protein-Ligand Docking: Data Clustering

ABSTRACT

The development of novel pharmaceutical drugs relies on the improvement of methods for drug design. Due to the cost of wet lab experiments, computational methods are increasingly being used to aid in this process. Docking is a commonly used process and it consists of simulating protein-ligand bindings and providing details about the simulated interaction. The drawback about docking is the proportionally inverse relation between accuracy and time of the simulation algorithm. Thus, an automatic protein-ligand docking procedure is needed to alleviate this drawback. Clustering is used in order to find correlations among features such as the distance and the lowest energy complex; correlations which will be exploited when developing the automatic protein-ligand docking procedure. This paper presents preliminary results of clustering the docking/simulation results data aimed at producing an automatic protein-ligand docking procedure.

1. INTRODUCTION AND LITERATURE REVIEWS

As the world around us moves to higher levels of modernization, our lifestyles are affected, but one aspect that remains constant is the concern for our health. Pharmaceutical companies strive to find better drugs and thus need better methods for drug-designing. The traditional wet lab procedure for drug engineering can be expensive in terms of people, resources, time, and money. As computers become more powerful, efficient, and accurate, it becomes natural to take advantage of these computers by applying computational methods to the science realm, in particular to drug design.

There are two different types of computational approaches in drug design: *De Novo Design* and Docking [3]. De Novo design is aimed at building a molecule from scratch to fit the binding site of the protein, while docking refers to the attempt of 'binding' several known molecules into the protein and determining which of the molecules, or ligands, fits best [3]. The latter approach can be aided by computational methods; by developing simulations of the bindings between the protein and the ligands, and rating each docked complex. Hence, the process of docking has two components: an algorithm used for the protein-ligand binding simulation and a scoring function to rate the docked protein-ligand complex [3]. The scoring function is used due to the following: in actual wet lab experiments, most compounds do not bind to the protein but in a simulation, an ideal case is considered, that is, in the computational methods, every ligand has the ability to bind to the protein; thus it is critical to evaluate and score the docked complexes [2]. The algorithm used for the binding simulations can fall into one of

the following categories: (1) Quantum Mechanics, (2) Molecular Mechanics, or (3) Molecular dynamics [1]. Each of the algorithm categories mentioned above differ only in accuracy and time-cost [1].

Some of the major setbacks with docking is the tradeoff between time and accuracy. Some docking methods are very accurate but are time-expensive while other methods sacrifice accuracy for time-efficiency. Some computational attempts have been made, such as using a grid framework for the simulations, to alleviate time cost [1]. In addition, considering a greater number of features can increment both the complexity and accuracy of the method. As an example, consider rigid versus flexible proteins and/or ligands; considering a flexible ligand or protein increments the number of simulations needed to obtain significant results and thus increments simulation time significantly [3].Thus, the development of an automatic protein-ligand docking procedure is needed. This procedure will consider features of the ligands, and based on these features, will determine which ligand, along with its conformation and rotation, is best-fitted for the protein, that is, without the need to run the simulation. Clustering is used in order to find correlations among features such as the distance and the lowest energy complex. This paper presents preliminary results of clustering the docking-results data aimed at producing an automatic protein-ligand docking procedure.

In section 2, an explanation of some concepts will be presented and the technical approach described; in section 3, some limitations of clustering will be described, section 4 will cover the obtained results, and in section 5 the results will be discussed. This paper will conclude with the discussion of future work and some acknowledgments in sections 6 and 7 respectively.

2. METHODOLOGY

2.1. Concept Description

2.1.1 Simulations

The simulations were performed on two complexes: 1hvi and 1hvj. **Figure 1** shows the superposition of the two complexes. The protein used was an HIV protein responsible for the DNA defragmentation which aids the HIV virus to mutate itself in order to avoid recognition by the body's immunological system. **Figure 2** show the structure of the HIV protein.

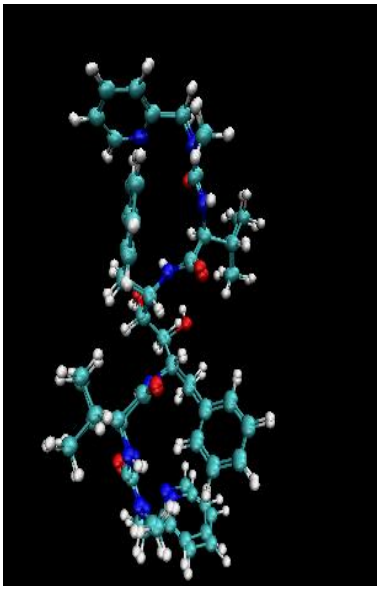


Figure 1: 1hvi and 1hvj complex superposition

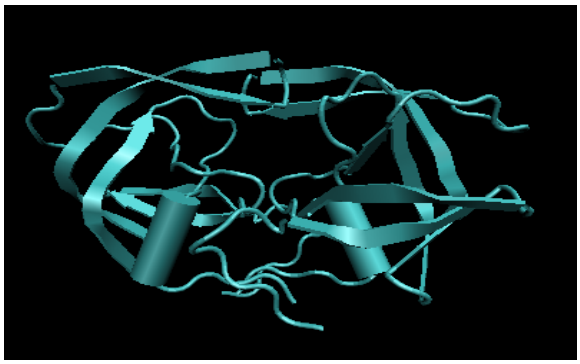


Figure 2: HIV protein structure

The simulation process begins with a protein and a set of ligands. For each ligand, many conformations are generated randomly and for each of those conformations, many more rotations are generated randomly. After which, each of the generated ligands is docked into the protein. The simulation returns the set of ligands generated in the order of increasing protein-ligand energy. The goal of the simulation is to find the ligand, conformation and rotation, which minimizes the energy of the protein-ligand complex, energy which is associated with the most stable complex hence, the binding most likely to occur in nature. **Figure 3** summarizes this process.

Chemistry at HARvard Molecular Mechanics (CHARMM) was used to perform the simulations. CHARMM is a program to simulate biological macromolecules such as proteins [4].

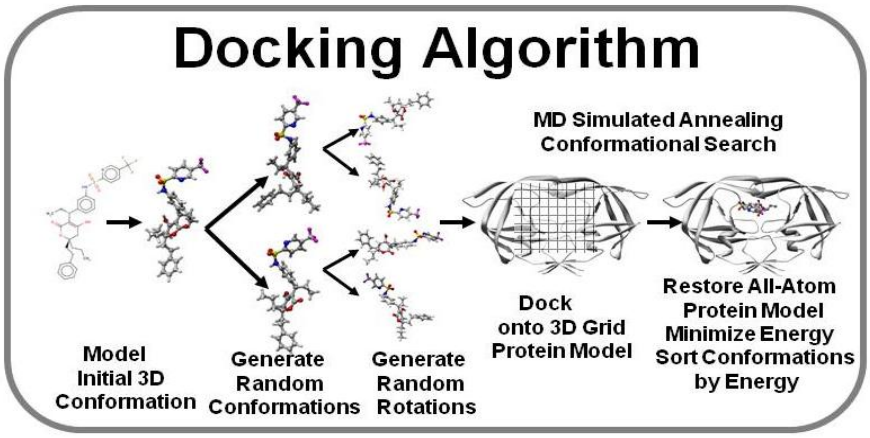


Figure 3: Illustration of the docking algorithm

2.1.2 Clustering algorithm: K-Mean

The simulations results obtained were clustered using an algorithm called K-Mean. This algorithm follows the subsequent steps:

1. Randomly select K centroids from the data points
2. Compute distance from all data points to every centroid
3. Assign cluster membership by minimizing datum-centroid distance
4. Recalculate centroids for each cluster by considering only the data within that cluster
5. Repeat steps 2 , 3, and 4 until no data point switches clusters

Where k is a given parameter, the distance used is independent of the algorithm, and the mode to recalculate the centroids in step 4 is independent of the algorithm.
This algorithm was chosen for its simplicity and due to time constraints.

2.2 Technical Approach

The clustering algorithm was applied to 300 data points for each of the two complexes and k = 7 was used which was selected randomly. The type of distance used was the root-mean-square deviation (RMSD) which is calculated as follows:

Let v and w be data points obtained through the simulations, that is, let them be ligands, and let v and w have n molecules which are represented by 3 dimensional vectors. Hence, v and w are sets of n 3-D vectors; $v = \{ (v_{1x}, v_{1y}, v_{1z}), (v_{2x}, v_{2y}, v_{2z}), \ldots, (v_{nx}, v_{ny}, v_{nz}) \}$ and
$w = \{ (w_{1x}, w_{1y}, w_{1z}), (w_{2x}, w_{2y}, w_{2z}), \ldots, (w_{nx}, w_{ny}, w_{nz}) \}$. Then the RMSD between v and w is computed as follows:

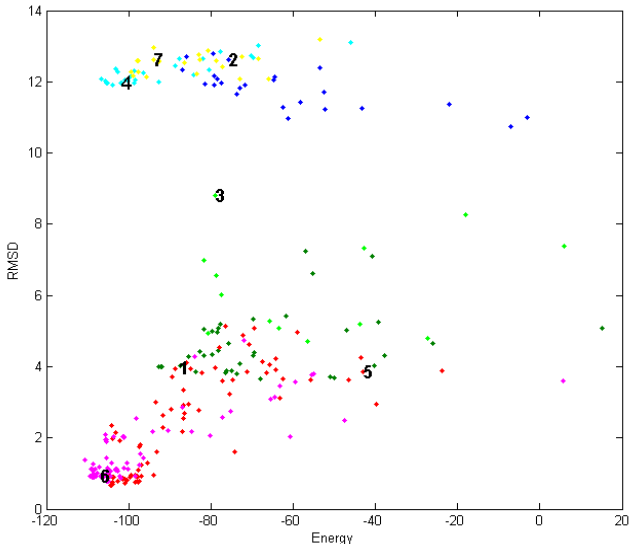
$$\text{RMSD}(v, w) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2}$$

3. LIMITATIONS

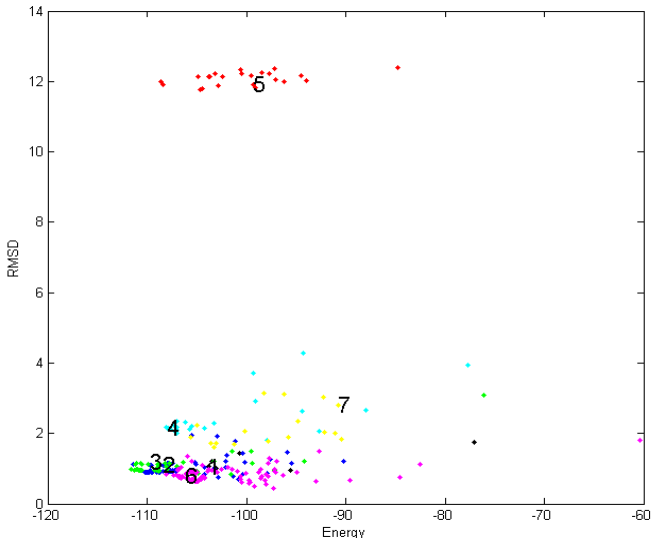
Some of the challenges encountered when using the k-Mean algorithm are the following: (1) finding the optimized k value (2) determining the optimum method to compute the centroids, and (3) avoiding result fluctuations due to the randomness involved. Every clustering algorithm has its challenges thus it is impossible to have an ideal scenario.

4. RESULTS

Graph 1 depicts the clustering results for complex 1hvi and **Graph 2** depicts the clustering results for complex 1hvj.



Graph 1: Results for the 1hvi complex with k =7



Graph 2: Results for the 1hvj complex with k =7

The optimal ligand needs to be known before the clustering is performed in order to be able to evaluate the clustering results. On **Graph 1** and **Graph 2**, the x axis represents the energy of the ligand-protein complex, and the y-axis represents the distance between that datum and the already-known optimal ligand. The integers, 1 through 7, on the graphs represent the centroids of the seven clusters, and a different color was assigned to each cluster.

5. DISCUSSION

The results obtained for both complexes support the initial hypothesis in which it was predicted that the largest cluster would have the lowest energy data points. This, in turn, indicates that the clustering method produces meaningful and significant results and thus is a reliable method.

Much analysis is still be done. For example, the simulation begins with a specified ligand conformation, thus it is still to be determined whether this initial conformation plays a key role in the results obtained or if it is indeed the clustering algorithm which is responsible for the results.

6. FUTURE WORK

There exists much space for result improvement and change in the above execution. The following are considered possibilities for future work:

- Compare results when using different:
  - K values in the case of the K-Mean algorithm
  - Types of distances
  - Methods of computing the centroids in the case of the K-Mean algorithm
  - Clustering methods
- Take into account force fields and energies when clustering as opposed to only complexes
- Improve efficiency by calculating only the distances needed at the time of clustering instead of all distances between every data point

7. ACKNOWLEDGEMENTS

- CRA-W Distributed Mentorship program (CRA-W DMP)
- University of Delaware (UDel)
- Grant Funding:
  - NSF OCI #0506429, DAPLDS - a Dynamically Adaptive Protein-Ligand Docking System based on Multi-Scale Modeling

REFERENCES

[1] Abramson, David, Celine Amoreira, et al. *A Flexible Grid Framework for Automatic Protein-Ligand Docking.* Proceedings of the 2nd IEEE international conference on e-science and grid computing, 2006. Monash University (Australia), University of Zurich (Switzerland), San Diego supercomputing center (USA).

[2] Godden, Jeffrey W, Florence L. Stahura, and Jurgen Bajorath. *Statistical Analysis of Computational Docking of Large Compound Data Bases to Distinct Protein Binding Sites.* Journal of Computational Chemistry, vol 20. John Wiley & Sons, Inc, 1999. MDS Panlabs, University of Washington.

[3] Kavraki, Lydia E. *Protein-Ligand Docking, Including Flexible Receptor-Flexible Ligand Docking.* Connexions module (2007): m11456: http://cnx.org/content/m11456/1.10/

[4] Taufer, M, M. Crowley, D. Price, A.A. Chien, C.L. Brooks III. *Study of a Highly Accurate and Fast Protein-Ligand Docking Algorithm Based on Molecular Dynamics. ipdps*, p. 188, 18th International Parallel and Distributed Processing Symposium (IPDPS'04) - Workshop 9, 2004