Structural Affinity Measurements of Biomolecular Surface
for Docking Sites Selection

Chen-En Hsieha, Grace Shiahuy Chena, Jie-Shan Yehb and Yaw-Ling Linabc,*

a Department of Applied Chemistry, Providence University, Taiwan.
E-mail: g1026011@pu.edu.tw; grace@pu.edu.tw
b Dept. of Computer Science and Information Management, Providence Univ.
E-mail: jsyeh@pu.edu.tw
c Dept. of Computer Science and Information Engineering, Providence Univ.
*Corresponding Author: yllin@pu.edu.tw

ABSTRACT
Identification and optimization of the best ligand-receptor docking sites based on the structures of biomolecules are common scientific challenges. Given the ligand-receptor docking pair, our proposed Affinity-Estimation algorithm scans through the surface area of the receptor and estimates the potentials of these possible candidate sites above. Later on, the algorithm reexamines these potential sites with finer grains inquisition by utilizing the AutoDock [13] automated docking tool. The proposed algorithm is especially suitable if the ligand size is relatively smaller than the receptor.

The proposed algorithm is used to perform experiments upon Japanese encephalitis related biomolecules in virology research. The proposed affinity estimation algorithm is about 2 folds faster within similar binding free energies for these experiments in average. The proposed method exploits the MapReduce paradigm as a management and parallelizing tool to parallelize the underlying massive computation works corresponding to ligand-receptor pairs examined under the experiment.

Keyword: bioinformatics, molecular docking, drug design, AutoDock, MapReduce

1. Introduction
Molecular docking is a common method for searching new potential drugs. Improvement of the results of docking can be achieved by molecular dynamics simulations of protein-ligand complexes. The large amount of structural investigations on medically relevant proteins reflects the general recognition that the structure of a potential drug target is very precious knowledge; the molecular docking simulation, can now be effective in reducing costs and speeding up drug discovery [1, 2, 5]. Progress in functional genomics and structural studies on biological macromolecules

1
are producing more and more potential therapeutic targets. Hence it also increases the importance of effectiveness and efficiency of algorithms for small molecule docking and virtual screening of candidate compounds [3, 11]. Usually, the first step in the molecular docking is to find the position of the space and the conformation matched. In molecular docking, the receptor is possibly a biological protein or biomolecule, and the ligand is possibly a different protein, medicine or compound. Molecular docking simulation is often used as a method for virtual screen by setting a protein to match a group of compounds, and report the final best compound [2].

Protein structures play critical roles in vital biological functions [7]. To date, there are more than 107,754 protein structures [1] determined by the advances in X-ray crystallography and NMR spectroscopy to date, molecular biologists these days proceed in the direction of analyzing and classifying these protein structures in order to discover the interaction with ligand and receptors. Molecular docking simulation is a method for computer-aided drug design (CADD). It simulate the interaction between a protein receptor and a drug ligand by calculating the energy of interaction between them, and then search the optimal binding sites in most stable state.

$$V = W_{vdw} \sum_{i,j} \left( \frac{A_{ij}}{r_{ij}^2} - \frac{B_{ij}}{r_{ij}^6} \right) + W_{ele} \sum_{i,j} \frac{q_i q_j}{r_{ij}} +$$
$$W_{hbond} \sum_{i,j} E(t) \left( \frac{C_{ij}}{r_{ij}^2} - \frac{D_{ij}}{r_{ij}^6} \right) + W_{sol} \sum_{i,j} (S_i V_j + S_j V_i) e^{-\frac{r_{ij}}{r_{0i}}}$$

The enormous computational time needed for massive molecular dynamics simulations of protein-ligand conformations obtained by molecular docking is a serious problem. To alleviate the underlying computation tasks, we adopt the standard cloud computing platform infrastructure and utilize the underlying computational mechanism. Recently, Hadoop [17] has been applied in various domains in bioinformatics [15]. Hadoop is a software framework intended to support data-intensive distributed applications. It is able to process petabytes of data with thousands of nodes. It supports MapReduce programming model [16] for writing applications that process large data set in parallel on cloud computing environment.

2. Method and Materials

In our previous works [10], we proposed methods to distinguish the surface/inside atoms of the receptor by selecting a suitable distance maximizing the standard deviation of corresponding neighboring degrees of the molecules. With various considerations and different set ups of the underlying parametric spaces, the searching
space for the docking simulation problem can be significantly reduced.

The main idea of the molecular docking algorithm to estimate the binding free energy, the LE (lowest energy) value, between two molecules is the following. The first step concerns with adjusting positions between two molecules by probing the suitable parameter combining with ways of searching the underlying parametric space. The second step is to utilize the scoring function procedure to fine-tune the final result at the attempted position. In this experiment we fully cover the receptor surface for finding the good docking position and then perform more resources for refinement the result.

2.1 Structural Frame algorithm
Our previous works for the ligand/receptor docking sites problem produces feasible results when two molecules possesses relatively similar sizes. On the other hands, when the ligand size is comparatively much smaller than the receptor molecules, it is relatively harder to find a reasonably good docking position by trying out just one box frame upon the surface of the receptor. Thus, in this paper we propose different methods for small ligand.

The idea is to survey the surface of the receptor and evaluate the potentials of these possible candidate locations on the receptor. Once the evaluated candidate positions have been decided, the second phase of the strategy is to measure these ligand docking positions with more in-depth explorations. Thus, the first step of our two-stage algorithm is selecting the candidate frame boxes on surfaces of the receptor. See Fig.2 and Fig.3 for an example.

The first two steps of Structural Frame algorithm is to detect surface atoms of given receptor by the $GetSurface(R, \rho)$ algorithm [10]. The step 3 of Structural Frame algorithm is to set up candidate frame boxes that covers all receptor surface atoms by repetitively selecting the frame box with maximum neighboring boxes so far and then eliminating the chosen box with all its neighbor boxes until all surface atoms are covered. The final step simply reports the chosen candidate surface covering box positions list $C$.

2.2 Affinity Estimation algorithm
The rationales behind the affinity estimation algorithm shown in Fig.1 is to survey the surface of the receptor and evaluate the potentials of these possible candidate locations on the receptor.
The second step of affinity estimation is to measure these ligand docking positions by probing these positions with smaller sets of frame boxes parameterized by smaller $m_p$. It follows that the step 3 of affinity estimation is to set up potentially more promising ligand docking sites lies on top fewer portion parameterized by the $\rho_D$ of $C_P$ list.

Finally, step 4 and 5, the algorithm further refines the results by using *autogrid* and *autodock* on smaller searching spaces parameterized by smaller $m_D$ with probe number set at $P_2$. The *refinement* phase will prepare `.gpf` and `.dpf` for each box $b$ in $D$ by `preparegpf.py` and `preparedpf.py`, and then start *autogrid* and *autodock*, and return the best $(\text{pos}, E)$.

---

$AE(L, R, s, \rho, r, m_1, m_2, t, p_1, p_2)$

*Input:* $L, R$: the molecule atoms of the ligand and receptor.
$s$: The magnification factor of the ligand size
$\rho$: The surface atom ratio
$r$: The resolution of these grids
$m_1, m_2$: Ligand magnification factor in PROBE phase and DIG phase
$t$: Grid percentage used in DIG phase
$p_1, p_2$: The probe number for PROBE phase and DIG phase

*Output:* $\text{pos}$: The resulting position (conformation) of the ligand
$E$: The lowest energy of the result

1. Let $C \leftarrow \text{COVER-FRAME}(L, R, s, \rho)$ be the candidate surface covering box positions list
2. Prepare for each $b \in C$ by the box sized $m_1$ using `preparegpf.py` and `preparedpf.py`, and then start *autogrid* and *autodock*; the results are placed within
3. Let $D$ be box list with the top $t$ percent of the $C$ with the LE in PROBE phase.
4. Prepare for each $b \in D$ by the box sized $m_2$ using `preparegpf.py` and `preparedpf.py`, and then start *autogrid* and *autodock*
5. Obtain the best position $\text{pos}$ with lowest energy $E$ from results placed within $k$ `.dlg` files; return $(\text{pos}, E)$.

The Affinity Estimation algorithm.

---

**2.3 Detect surface atoms**

The idea is to partition atoms of a receptor into *surface points* and *inside points* so that it is possible to wisely place the corresponding ligand into suitable places nearing to surfaces of a receptor. The inside atoms are those atoms with sufficient number of neighboring atoms, while surface atoms are those having fewer neighboring atoms. However, the neighboring relation is defined by a suitable distance. Setting the cut-off distance by an extremely small meaning that every atom is isolated, while an overly large distance resulting every atom pair being neighbor to each other. The trick is to pick the suitable distance that produces the most meaningful neighboring numbers.
2.4 Fixed radius analyze

In order to tuning the best ratio of the surface point, we took some experiment and noticed that the surface point often concentrated to the center of the 3D shape in the $\sigma$ based neighbor strategy [10]. Then we start an different strategy, the idea is based on the limited bond length. The fixed radius analyze shows in Fig.4 and Fig.5.

In the same ratio we divide the surface point and the inside point, the $\sigma$ based method [10] will lost some atom which is the surface point shows in Fig.2. The fixed radius
method are more dispersion to the surface point shows in Fig.3.

3. Experiments

In order to tuning the parameters ($\rho_D$, $m_P$, $m_D$), we use the MapReduce framework to reduce the time of waiting each result. The whole experiments spend a fare amount of time. In order to tune in for each parameter, the experiment spends about 6 hours for 32 nodes; the total computation time more than 1,166 machine-hours (about 48 machine-days in total.) The total ligand-receptor pairs are made by 18 pairs consisted of 3 ligand (GLYCEROL, N-ACETYI-D-GLUCOSAMINE, BETA-D-MANNOSE) of small molecular, and 6 receptors (1QU6, 1ZIW, 2A0Z, 2Z62, 2Z7X, 2Z80) [12], toll-like receptor (TLR), double-stranded RNA-activated protein kinase (PKR) are member of pattern recognition receptors [12]. These ligands are mostly consist of small molecules, while the receptors are mostly macromolecules (proteins). In performing these experiments set up, we fixed resolution $r = 0.5$ and probes number $p$ ranged from 256,000 to 1,024,000.

Fig. 4: Average LE value with different $r$, shows the LE value is better at $r = 10$–25Å. While use the big $r$ will cause the problem like Fig.2
Fig. 5: Different $r$ with their $\sigma$ value, shows the $r$ value versus $\sigma$ value. The maximum $\sigma$ value often at large $r$ value but receive worse LE value.

Fig. 6: Average LE value with different surface point choosing strategy. Performance of the AE algorithm under mixed parametric settings ($\rho$, $m_P$, $m_D$) versus the Global-grid [10] method and the current method (change the surface point detect strategy). These results are tested upon 18 different ligand-receptor pairs. Points with the same color are linked with different setting by various probe numbers.

4. Concluding Remarks

In this paper, we propose a two-stage algorithm, namely the affinity estimation algorithm shown in Fig.1. The algorithm surveys the surface of the receptor and evaluates the potentials of these possible candidate locations on the receptor. Note that these candidate locations are boxes covering the surface of the receptor is determined by the Structural-Frame algorithm.

In the estimation stage, these candidate docking positions are evaluated by probing positions with smaller sets of frame boxes. In the second refinement stage, the algorithm further refines these positions by using larger searching spaces and finer grains inquisition with more probe numbers in performing the later AutoDock examinations.

It is shown in the experiments that we constantly obtain better lowest energies comparing with the Global-grid [10] methods. In average, the proposed affinity estimation algorithm can be easily parallelized in the cloud computing frame work.

5. Acknowledgment
This research was partially supported by the Ministry of Science and Technology under the Grants MOST 103-2221-E-126-009- and MOST 103-2632-E-126-001-MY3.

REFERENCES


biotechnology, 20(4), 429-436.


