Molecular Descriptors Selection in Support Vector Machine Classification of Protein-ligand Binding Affinity with Applications to Molecular Docking

Chen-En Hsieh*[1], Grace Shiahuy Chen†[1], Chia-Chen Lin‡[3], and Yaw-Ling Lin§ [1,2]

[1]Department of Applied Chemistry, Providence University, Taichung, Taiwan
[2]Dept of Computer Science and Information Engineering, Providence University
[3]Dept of Computer Science and Information Management, Providence University

Abstract

In this paper, we propose algorithms for biomolecular docking sites selection problem by support vector machines with selective feature reduction. The proposed method can reduce the number of various amino acid features before constructing SVM prediction models. Given frame boxes with features and analyze the important features by correlation coefficients to LE values. The algorithm will ranking these possible candidate locations on the receptor before AutoDock examinations.

The method is implemented upon the widely employed automated molecular docking simulation software package, AutoDock. Experiments are set up to test upon Japanese encephalitis related biomolecules in virology research. The proposed affinity estimation algorithm is about 4 folds faster with 2% LE value lost in this experiments. Hadoop MapReduce frameworks are used in our experiments to parallelize the underlying massive computation works corresponding to ligand-receptor pairs examined under the experiment.

Keywords: bioinformatics, algorithm, molecular docking, drug design, AutoDock, Hadoop, MapReduce, machine learning, SVM, aaindex.

1 Introduction

While designing new drugs still poses great challenges, 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 [1]. Computer-aided drug design techniques, especially the molecular docking simulation, can now be effective in reducing costs and speeding up drug discovery [2, 4].

Progress in functional genomics and structural studies on biological macromolecules are producing more and more potential therapeutic targets, but also increases the importance of small molecule docking and virtual screening of candidate compounds algorithms. [13, 3]. 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 bio-molecule, 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 102,720 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

*g1026011@pu.edu.tw

†grace@pu.edu.tw

‡mhlin3@pu.edu.tw

§Corresponding author. yllin@pu.edu.tw. This work is partly supported by grants from the Ministry of Science and Technology of Taiwan, Republic of China (MOST 103-2632-E-126-001-MY3.)

drug ligand by calculating the energy of interaction between them, and then search the optimal binding sites in most stable state.

Amino acid index (AA-index [12]) is propose the different physicochemical value and biological properties value of 20 amino acids by numerical values. AA-index now have 544 different index set.

Machine learning is a "field of study that gives computers the ability to learn without being explicitly programmed" defined by Arthur Samuel in 1959 [18]. Support vector machines(SVM), artificial neural networks(ANN), representation learning, and inductive logic programming are some of many machine learning algorithms. LIBSVM (a library for support vector machines) [5] is a SVM package since year 2000. LIBSVM is to help users to apply their application by SVM easier. The package provide One-class to multiclass and probability present functions.

The enormous computational time needed for massive molecular dynamics simulations of protein-ligand conformations obtained by molecular docking is a serious problem. To coordinate and utilize the underlying computational mechanism, we adopt the standard cloud computing platform infrastructure to alleviate the underlying computation tasks. Hadoop [20] is a software framework intended to support data-intensive distributed applications. It is able to process petabytes of data with thousands of nodes. Hadoop supports MapReduce programming model [19] for writing applications that process large data set in parallel on Cloud Computing environment. Recently, Hadoop has been applied in various domains in bioinformatics [17].

2 Method and Materials

In our previous works [11, 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 we fully cover the receptor surface by boxes. The second step is to analyze these boxes by prediction model and rank these boxes. In this experiment we fully cover the receptor surface and take top 5 boxes for finding the good docking position and not perform all boxes resources for reduce the computation cost.

2.1 Molecular Docking Simulation

AutoDock

AutoDock is a automated molecular docking simulation software package since 1990. AutoDock uses Genetic Algorithm (GA) [8, 9], Lamarckian Genetic Algorithm (LGA) [16] in finding the lowest energy, and the used the Amber molecular force field scoring function [6]. AutoDock is now version 4.2, include two program with autodock and autogrid. Autogrid pre-calculates a set of grids which describing the target protein and autodock performs the docking of the ligand to these grids. The force field evaluation function of AutoDock4.2 is described as the following: [15]

$$V = W_{vdw} \sum_{i,j} \left(\frac{A_{ij}}{r^{12}} - \frac{B_{ij}}{r^{6}} \right) + W_{\text{elec}} \sum_{i,j} \frac{q_i q_j}{\varepsilon(r_{ij}) r_{ij}} + W_{\text{hbond}} \sum_{i,j} E(t) \left(\frac{C_{ij}}{r^{12}} - \frac{D_{ij}}{r^{10}} \right) + W_{\text{sol}} \sum_{i,j} (S_i V_j + S_j V_i) e^{\left(\frac{-r_{ij}^2}{2\sigma^2} \right)}$$

2.2 Feature ranking algorithm

Our previous works [10] for the ligand/receptor docking sites problem produces feasible results when the ligand size is comparatively much smaller than the receptor molecules, we use the Cover-frame [10] algorithm to fully cover the surface of the receptor and perform all boxes to autodock. Thus, in this paper we propose different methods to reduce the boxes to save some resource without perform all boxes.

Before the SVM start, we have to set all 544 AA-index and the LE value to each boxes as their label. The rationale of the label is the ligand will be attracted by some environment, label and analyze these boxes to find out the good boxes. While the label finished, we can start the work. First, compute the correlation coefficient with LE value(or quality class value). Second, ranking these correlation coefficient value by Absolute value. Third, take top k features to generate prediction model, in order to find the best combination of the k features (note: the final features will $\leq k$)

```
FEATURE-RANKING($B, k$)
    Input: $B$: Boxes with labels.
           $k$ : Number of features to be analyzed.
    Output: $F$ : A list of chosen features.
    Let $C$ be the correlation coefficients list for boxes in $B$ related to the corresponded $LE$ values.
    Get the ranking order list of $C$ according to the absolute coefficients.
    Let $F$ be the top most $k$ features in $C$.
    return the feature ranking list $F$
```

Figure 1: The FEATURE-RANKING algorithm.

3 Experiments

In order to select the important relevant features, we choose the 8 most relevant features ranked by their corresponding correlation coefficients to LE values (or quality class value). Note that the prediction model spends more time when the number of selected features get higher than 8. The experiment spends about 6 machine-hours with 197 boxes for 10 receptors; the total computation time more than 1,166 machine-hours (about 48 machine-days with compute all 83 boxes for 6 receptor in total.) [11]. The total ligand-receptor pairs are made by 10 pairs consisted of 1 ligand (ACETYLCHOLINE) of small molecular, and 10 receptors (1QU6, 1ZIW, 2A0Z, 2BR8, 2C9T, 2Z62, 2Z63, 2Z64, 2Z7X, 2Z80) [14], toll-like receptor (TLR), double-stranded RNA-activated protein kinase (PKR), acetylcholine binding protein (ACHBP) are member of receptors [14].

3.1 Testing environment and data source

The experiments are performed on one NFS server and four IBM blade server in the Providence University Cloud Computation Laboratory. Each server is equipped with two Quad-Core Intel Xeon 2.26GHz CPU, 24G RAM, and 296G disk under the Ubuntu version 12.04 with the virtualization platform KVM/QEMU. Under the current system environment, we create 32 virtual machines by KVM; each virtual machine is set to one core CPU, 1G RAM, and 10G disk running under the O.S. Ubuntu version 12.04 with Hadoop version 1.21 MapReduce platform. Each virtual machine is responsible for one map operation and one reduce operation. The total number of the map/reduce operations is up to 32 respectively.

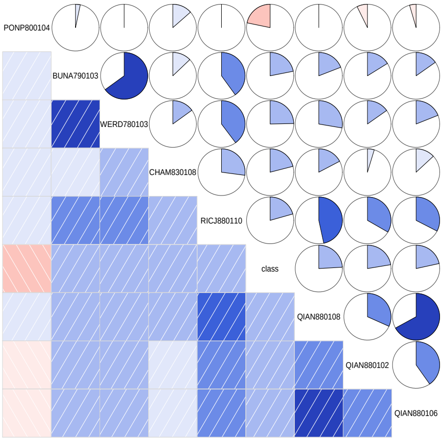

Figure 2: The figure shows the correlation coefficient of the top 8 AA-index features and quality class.

The Protein structure data sources are gathered from the Protein Data Bank [1], which maintains this single archive of macromolecule structural data freely and publicly available to the global community. These protein structure data are downloaded from the wwPDB's ftp server (ftp://ftp.wwpdb.org/), where the number of protein structure data is 116,258. We download 1 ligand and 10 receptor as 10 set data are treated as the testing data for our experiments.

3.2 Comparison prediction model with different features

As shown in Figure 4, the proposed feature ranking algorithm is about 4 folds faster to reduce 197 boxes into 50 boxes (each receptor take 5 boxes choose by prediction model) with the cost of about 2% LE value inferior in the experiment. The Figure 5 shows the comparison by different feature selection method will lead to different result. In this case, take top 8 (absolute value) features for

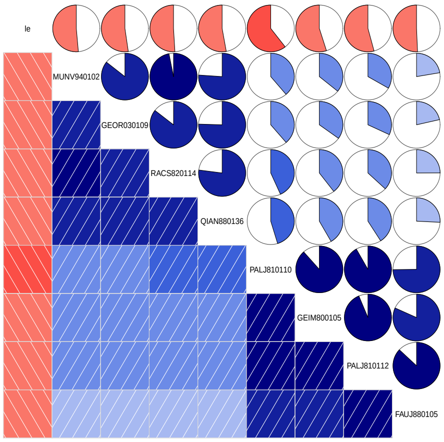

Figure 3: The figure shows the correlation coefficient of the top 8 AA-index features and LE.

prediction model makes the better result.

4 Concluding Remarks

In this paper, we propose a feature ranking algorithm. The algorithm first compute the features by correlation coefficient to LE value, and then rank this features by absolute value. The last is to take k features to generate the prediction model. This algorithm are helpful to abandoned some boxes with lower chance to docking successful. In the future, we will perform more experiments to find better algorithms for average results in shorter computation time.

5 Acknowledgment

This research is partially supported by the ministry of science and technology (MOST) of Taiwan, R.O.C, under the grant MOST 103-2632-E-126-001-MY3.

References

[1] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

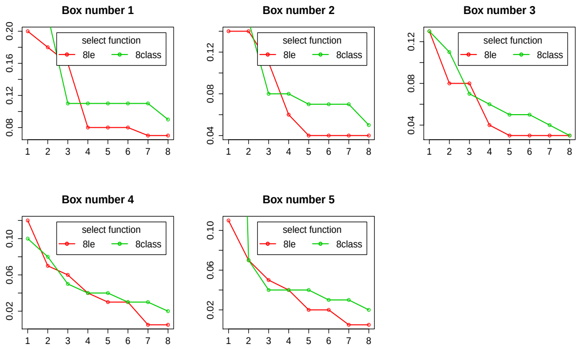

Figure 4: Comparison of top 8 features by feature selection by **LE** value versus **class** value. The x-axis indicates number of features chosen for the prediction model; the y-axis is the ratio of the LE comparing to the best box (a ratio from 1 to 0).

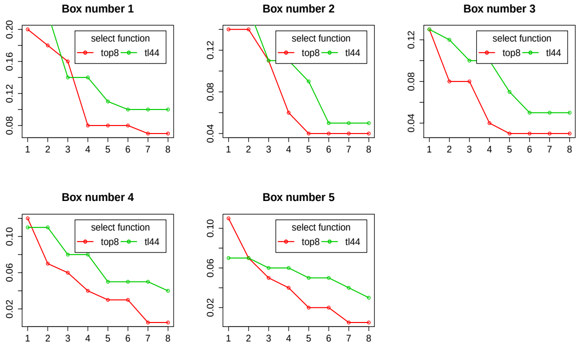

Figure 5: Comparison of top 8 features versus top 4-positive plus 4-negative features by feature selection by **LE** value. The x-axis indicates number of features chosen for the prediction model; the y-axis is the ratio of the LE comparing to the best box (a ratio from 1 to 0).

[2] J. Biesiada, A. Porollo, P. Velayutham, M. Kouril, and J. Meller. Survey of public domain software for docking simulations and virtual screening. *Human Genomics*, 5(5):497, 2011.

[3] Z. Bikadi and E. Hazai. Journal of cheminformatics. *Journal of Cheminformatics*, 1:15, 2009.

[4] N. Cerqueira, J. Ribeiro, P. Fernandes, and M. Ramos. vslaban implementation for virtual high-throughput screening using autodock and vmd. *International Journal of Quantum Chemistry*, 111(6):1208–1212, 2011.

[5] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[6] Y. Duan, C. Wu, S. Chowdhury, M. C. Lee, G. Xiong, W. Zhang, R. Yang, P. Cieplak, R. Luo, T. Lee, et al. A point-charge force field for molecular mechanics simulations of proteins based on condensed-phase quantum mechanical calculations. *Journal of Computational Chemistry*, 24(16):1999–2012, 2003.

[7] M. Gerstein, R. Jansen, T. Johnson, J. Tsai, and W. Krebs. Studying macromolecular motions in a database framework: from structure to sequence. In *Rigidity theory and applications*, pages 401–420. Springer, 2002.

[8] D. E. Goldberg and J. H. Holland. Genetic algorithms and machine learning. *Machine Learning*, 3(2):95–99, 1988.

[9] J. H. Holland. *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence.* U Michigan Press, 1975.

[10] C.-E. Hsieh, G. S. Chen, and Y.-L. Lin. Biomolecular docking sites selection by surface filtration and refinement. *The Proceedings of International Computer Symposium (ICS'2014)*, 3:425–436, Taichung, Taiwan, December 12-14, 2014.

[11] C.-E. Hsieh, S. Chen, P.-S. Hsu, C.-J. Chen, and Y.-L. Lin. Selecting molecular docking sites by neighbor selection and various factors. *The 31th Workshop on Combinatorial Mathematics and Computation Theory (CMCT 2014)*, Taipei, Taiwan, April 25-26, 2014.

[12] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. Aaindex: amino acid index database, progress report 2008. *Nucleic acids research*, 36(suppl 1):D202–D205, 2008.

[13] P. Kolb, R. S. Ferreira, J. J. Irwin, and B. K. Shoichet. Docking and chemoinformatic screens for new ligands and targets. *Current Opinion in Biotechnology*, 20(4):429–436, 2009.

[14] M. S. Lee and Y. Kim. Pattern-recognition receptor signaling initiated from extracellular, membrane, and cytoplasmic space. *Molecules and Cells*, 23(1):1, 2007.

[15] G. Morris, D. Goodsell, M. Pique, W. Lindstrom, R. Huey, S. Forli, W. Hart, S. Halliday, R. Belew, and A. Olson. User guide autodock version 4.2, 2012.

[16] G. M. Morris, D. S. Goodsell, R. S. Halliday, R. Huey, W. E. Hart, R. K. Belew, and A. J. Olson. Automated docking using a lamarckian genetic algorithm and an empirical binding free energy function. *Journal of Computational Chemistry*, 19(14):1639–1662, 1998.

[17] M. C. Schatz. Cloudburst: highly sensitive read mapping with mapreduce. *Bioinformatics*, 25(11):1363–1369, 2009.

[18] P. Simon. *Too Big to Ignore: The Business Case for Big Data.* Wiley and SAS Business Series. Wiley, 2013.

[19] R. C. Taylor. An overview of the hadoop/mapreduce/hbase framework and its current applications in bioinformatics. *BMC Bioinformatics*, 11(Suppl 12):S1, 2010.

[20] T. White. *Hadoop: The Definitive Guide: The Definitive Guide.* O'Reilly Media, 2009.