様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２８ 年 １０ 月 ２５ 日現在

機関番号：３８００５
研究種目：若手研究(B)
研究期間：2014～2015
課題番号：２６７３０１５２

研究課題名（和文）Development of an intelligent dynamic docking pipeline for improving molecular docking simulations

研究課題名（英文）Development of an intelligent dynamic docking pipeline for improving molecular docking simulations.

研究代表者
Hsin KunYi (Hsin, Kun-Yi)

沖縄科学技術大学院大学・その他の研究科・研究員

研究者番号：６０６０４１５５

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：我々は、薬理学に基づく予測パイプラインネットワークを開発した。分子経路に関与するタンパク質に対する試験化合物の結合可能性を正確・効率的に予測できる。予測スコアと実験的結合親和性は、R>0.8と良好なパフォーマンスを示した。更に既知のバイオアッセイ結果と比較することにより、様々なキナーゼ阻害剤の選択性を予測した。この結果も高い一貫性を示している。インパクトの高い学術誌に発表し、システム生物学研究所や東京大学医科学研究所とのプロジェクトでも、創薬分子を同定するためこのツールを使用した。さらに、創薬関係者をはじめとした社会全体へ貢献できるよう、オープンアクセスでsystemsDockを公開した。

研究成果の概要（英文）：In order to precisely and efficiently predict the binding potentials of test compounds against proteins involved in a molecular pathway, we have developed a network pharmacology-based prediction pipeline. By assessing the correlations between the prediction scores and the experimental binding affinities, our prediction method shown a good performance in predicting the binding potentials (R >0.8).
Additionally, we predicted the selectivity of various kinase inhibitors by comparing with known bioassay results, showing a good consistency. The relevant research results have been published on high-impact journals. We have also applied it to several joined projects helping collaborators, including those in Systems Biology Institute (SBI, Tokyo) and The University of Tokyo (IMSUT), to identify druggable molecules. A publicly accessible website called “systemsDock” (http://systemsdock.unit.oist.jp/) has been published, dedicating our achievements to the community of drug discovery.

研究分野：Bioinformatics

キーワード：Docking Simulation　Molecular Dynamics　Network Pharmacology　Machine Learning　Molecular Interaction　Drug Discovery

１．研究開始当初の背景

Molecular docking simulation is an important tool used in the discovery of lead compounds for drug design. However the generally unreliable results obtained by the currently available docking tools may be ascribed to several methodological defects: (a) the scoring functions are over-simplified (e.g. use of point charges) in order to calculate protein-ligand binding potential rapidly, (b) training sets only provide reliable information for particular protein families[1], (c) protein flexibility and solvent-related terms are only taken into account in a very primitive way.

With the massively parallel computing power now available within our university (OIST), I proposed to develop an Intelligent Dynamic Docking Pipeline (IDDP) which applies machine learning algorithms and will incorporate molecular dynamics and hybridized quantum mechanics/molecular mechanics (QM/MM). The development of IDDP will provide a major advance in the quality and reliability of such docking simulations. The proposed development builds on an existing prototype platform[2] which uses multiple docking programs together with in-house machine learning algorithms and has already resulted in significantly improved docking simulations (the black bar in Figure 1).

Molecular docking algorithms have been developed over the last decade and are now widely used in industry and academia. Docking packages commonly used in the pharmaceutical industry, include Surflex, LigandFit, Glide, GOLD, FlexX, eHiTS and AutoDock. The prediction reliability is however still limited[3] (the white bars in Figure 1) and to address this, many approaches have been developed. For example consensus docking to select a correct binding mode[4] and rescoring to re-rank the docked poses; both gave modest improvements.

Our in-house tests, however, show that the best Pearson correlation coefficient measured between the predicted and experimental binding affinity was only up to 0.45. Such poor reliability leads to unnecessary experimental testing and increased costs in the drug discovery process. Thus, correctly predicting the binding energy of a given protein-ligand complex continues being one of the most important and difficult issues in the application of the docking simulation.

Molecular Dynamics (MD) simulations can provide an accurate description of ligand binding that takes into account the flexibility of both the protein and the ligand. Detailed simulations (50 to 100 ns) of protein-ligand complexes can, however, take days of processor time and this has been a barrier to the routine use of the technique in ligand-docking studies. New docking protocols are being developed that allow faster sampling times. Our preliminary studies using the Desmond package have shown that it will be possible to incorporate MD into the IDDP to screen a large number of docked ligands in a reasonable time.

A major weakness in all current scoring functions is the use of fixed point charges in determining electrostatic binding energy. A quantum mechanical description of ligand interactions allows a more realistic description to account for dipole-dipole and charge-transfer interactions8. We have recently obtained the QM/MM module from Schrodingers Small-Molecule Drug Discovery Suite which will be used to calculate electrostatic interaction energies in the final steps of docking simulation. The IDDP incorporating both MD and QM/MM will in the first instance be extensively tested against the PDBbind database of well-characterized protein-ligand complexes. This will allow tuning of the machine learning algorithms and should result in a significant improvement in the correlation between experimental and predicted binding energies (dashed-line bar in Figure 1). This will open the door to a “systems pharmacology” approach by for example testing lead compounds for off-target binding (e.g. by determining the binding (specificity) of a particular inhibitor with the proteins in the influenza pathway namely FluMap[5]).

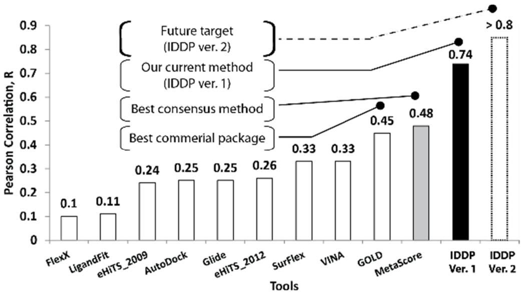

Figure 1. Performance of various docking simulations using PDBbind version 2007 benchmark (1300 complexes). Values are the correlations between the calculated docking scores and the corresponding experimental affinities. White bars are the nine commonly used docking programs. Gray bar is a consensus method combining the features of seven docking tools. Black bar is the IDDP version 1, and the dashed bar is the target for IDDP version 2.

２．研究の目的

Molecular docking has been heavily used in rational drug design for decades but the reliability still remains unsatisfactory because of unfavorable scoring functions and methodological defects. We aim to further develop our Intelligent Dynamic Docking Pipeline program package for application in "systems pharmacology" that will provide a step change in accuracy and reliability by including machine learning, quantum mechanics / molecular mechanics (QM/MM) and molecular dynamics approaches.

３．研究の方法

We aim to expand on our existing programs to develop an Intelligent Dynamic Docking Pipeline for molecular docking which will include machine learning, molecular dynamics (MD) and hybridized quantum mechanics / molecular mechanics (QM/MM). The steps in our existing and planned program package are shown in Figure 2. The current version can select protein targets (steps 1 and 2), run docking simulations using multiple docking tools (step 3) and analyze results using machine learning algorithms (steps 4 and 5) and already provides a major improvement in prediction reliability (see Figure 1). The incorporation of MD and QM/MM (steps 6 and 7) in this workflow will provide more accurate values for binding energies. Our university HPC equipped with > 3,000 CPUs provides the resource for these CPU-intensive simulations. Works of the system development are described as following:

1) **Machine learning systems**: We will apply multiple docking tools (e.g. eHiTS, GOLD and AutoDock VINA) together with the machine learning algorithms to reduce the error caused by individual tools. We have initially developed the prototype of a learning model showing an outperformance compared with other methods (the black bar in Figure 1) but further optimization is needed. We will optimize two machine learning systems using the algorithms of Random Forest (used in step 4, Fig 2) and multinomial logistic regression (step 5, Fig 2). We will also experiment with other algorithms (e.g. Decision Tree and SOMs).

2) **Application of MD**: for speed, the conformational change and energy minimization of the overall complex is neglected in a conventional docking simulation. Docking tools are also commonly not good at handling the presence of water molecules or metals in the binding interaction. The application of MD is expected to overcome those limitations. Different protocols for running MD simulations (e.g. time, temperature, incorporation of solvent etc.) will be tested for screening ligand complexes using the Desmond package.

3) **Correlations** between the predicted and experimental binding affinities: we will use the PDBbind[9] database (~3000 complexes) for the validation. Many of the proposed scoring functions were built and validated using PDBbind, thus we can conduct a direct comparison with those functions. Our target is to achieve a correlation coefficient > 0.8.

4) **Case study** of target identification for kinase inhibitors: We will apply our program package to screen a set of test compounds (leads/drugs) against proteins involved in the MAPK pathway (e.g. B-Raf, MEK and ERK), to identify the potential on-/off-target binding responsible for drug effects and toxicity. The MAPK pathway and bioassay validation data

will be provided through the collaborations with the Systems Biology Institute (SBI, Tokyo) and The Institute of Medical Science, The University of Tokyo.

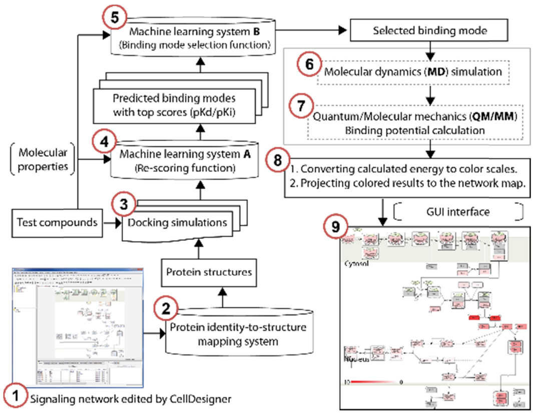
5
Machine learning system B
(Binding mode selection function)
Selected binding mode
6
Molecular dynamics (MD) simulation
7
Quantum/Molecular mechanics (QM/MM)
Binding potential calculation
Predicted binding modes
with top scores (pKd/pKi)
4
Machine learning system A
(Re-scoring function)
Molecular properties
8
1. Converting calculated energy to color scales.
2. Projecting colored results to the network map.
GUI interface
9
3
Test compounds
Docking simulations
Protein structures
2
Protein identity-to-structure
mapping system
Cytosol
Nucleus
1
Signaling network edited by CellDesigner

Figure 2. Schema of the Intelligent Dynamic Docking Pipeline for a systems pharmacology study. (1) a signaling network is firstly curated. (2) the identities of those proteins involved in the network are retrieved for looking up the corresponding protein structures in 3D through a built protein identity-to-structure mapping system. (3) multiple docking tools are applied to generate binding modes. (4) the docking scores of each generated binding mode are calculated by the machine learning system A. (5) the most predictive binding mode is then identified by the machine learning system B. (6) the application of MD is to refine the selected docking pose after docking simulation. (7) hybridized QM/MM is applied to calculate the binding potential. (8) and (9) finally, the calculated binding potentials are converted into a white-to-red color scale through a GUI interface to visualize the binding strength, and are projected on the network map for a comprehensive inspection.

4．研究成果

In order to precisely and efficiently predict the binding potentials of test compounds against proteins involved in a molecular pathway, we have developed a network pharmacology-based prediction pipeline. It is mainly composed of a high-precision scoring function for molecular simulation with a well-designed machine learning model. This pipeline enables researchers to predictively screen a large number of small molecules over a complex molecular pathway, allowing comprehensively identifying the on-/off-targets.

For prediction validation, we tested our method using PDBbind dataset, containing about three thousand protein-ligand complexes. By assessing the correlations between the prediction scores and the experimental binding affinities, it shown a good performance in predicting the binding potentials. The correlations have been improved to >0.8 (Figure 3). Additionally, we predicted the selectivity of various kinase inhibitors by comparing with known bioassay results, showing a good consistency (Figure 4).

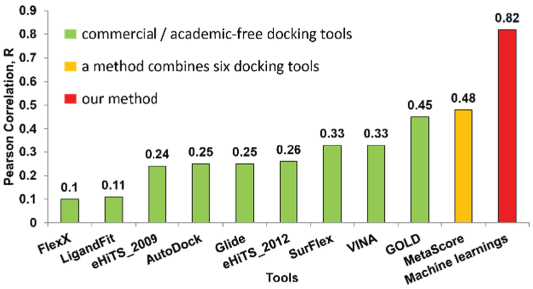

Figure 3. Performance of our method in docking simulation compared with others.

The relevant research results have been published on high-impact journals, including Nucleic Acids Research, Scientific Reports and IEEE. We have also applied it to several joined projects helping collaborators, including those in Systems Biology Institute (SBI, Tokyo) and The University of Tokyo (IMSUT), to identify druggable molecules. We have also developed a publicly accessible website called “systemsDock” for investigating "systems pharmacology" of a given compound, sharing the screening facility to researchers and dedicating our achievements to the community of drug discovery. The website is now available at http://systemsdock.unit.oist.jp/ (Figure 5).

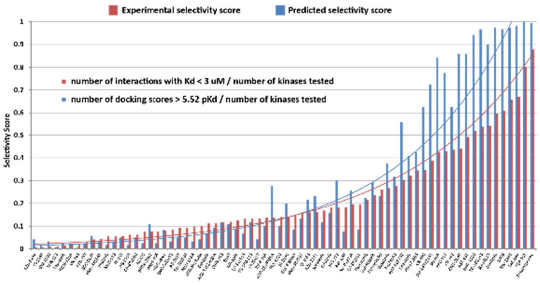

Figure 4. A comparison was conducted using the screening approach proposed in this study (blue bars) and bioassay results[6](red bars). The calculation of a predicted selectivity score is "S = number of kinases docked with score p*K*d >5.52/total number of kinases tested", whereas the experimental selectivity scores is "S = number of kinases found to bind with *K*d <3 μM/number of kinases tested”. A compound with a lower selectivity score indicates that it actively interacts with a small number of target proteins, implying a lower potential for off-target effects. Trendlines are the 2nd order polynomial regression functions. In most cases, screening accurately predicted the actual calculated binding constants; however, in some cases, screening predicted significantly higher binding constants than experimental data revealed, while no significant underestimates were observed.

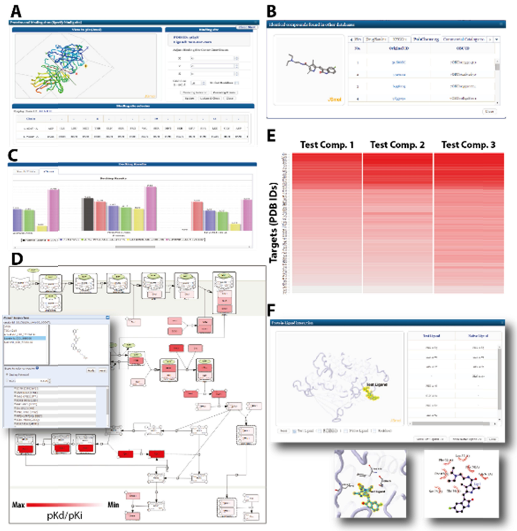

Figure 5. Screenshots of the systemsDock web interface (http://systemsdock.unit.oist.jp). (A) Interactive functions for binding site specifications are accessed by clicking on the displayed protein structure or amino acids listed in the sequence table to define the location of the preferred binding site. Users can adjust x-y-z coordinates to refine the location. (B) Links are provided for the test compound in external databases, as well as to visualize the compound in 3D. (C) Prediction results are furnished in an interactive histogram. Docking scores for each compound are grouped by proteins. By clicking on one of the bars, molecular binding interactions can be graphically shown in 2D/3D for structure-based investigation as shown in (F). (D and E) Visualizing results through a pathway map provided by the user or using a heat map. Colors of proteins are displayed as white-to-red scales or as white and red according to the docking scores. Click on a colored node (i.e. protein) to display binding interactions in 2D/3D as shown in (F). (F) Visualizing protein?ligand binding interactions of the test compound or native ligand in 2D/3D. Protein residues involved in the binding interaction are automatically identified. For reference, those that interacted with a native ligand, if available, are also listed. Clicking on any of the residue entries listed allows users to center and display the specified residue for closer inspection.

＜引用文献＞

① Warren, G. L. *et al.* A critical assessment of docking programs and scoring functions. *Journal of medicinal chemistry* **49**, 5912-5931 (2006).

② Hsin, K.-Y., Ghosh, S. & Kitano, H. Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology. *PloS one* **8**, e83922 (2013).

③ Plewczynski, D., Łaźniewski, M., Augustyniak, R. & Ginalski, K. Can we trust docking results? Evaluation of seven commonly used programs on PDBbind database. *Journal of computational chemistry* **32**, 742-755 (2011).

④ Houston, D. R. & Walkinshaw, M. D. Consensus docking: improving the reliability of docking in a virtual screening context. *Journal of chemical information and modeling* **53**, 384-390 (2013).

⑤ Matsuoka, Y. *et al.* A comprehensive map of the influenza A virus replication cycle. *BMC systems biology* **7**, 1 (2013).

⑥ Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology* **29**, 1046-1051 (2011).

５．主な発表論文等

〔雑誌論文〕（計 3 件）

① **Kun-Yi Hsin**, Yukiko Matsuoka, Yoshiyuki Asai, Kyota Kamiyoshi, Tokiko Watanabe, Yoshihiro Kawaoka and Hiroaki Kitano. "Systemsdock: A Web Server for Network Pharmacology-Based Prediction and Analysis." *Nucleic Acids Research*, (2016). DOI: 10.1093/nar/gkw335.

② Chiba, Shuntaro, Kazuyoshi Ikeda, Takashi Ishida, M Michael Gromiha, Yh Taguchi, Mitsuo Iwadate, Hideaki Umeyama, **Kun-Yi Hsin**, Hiroaki Kitano and Kazuki Yamamoto. "Identification of Potential Inhibitors Based on Compound Proposal Contest: Tyrosine-Protein Kinase Yes as a Target." *Scientific Reports*, (2015). DOI: 10.1038/srep17209.

③ **Kun-Yi Hsin**, Hiroaki Kitano, Yukiko Matsuoka and Samik Ghosh. "Application of Machine Leaning Approaches in Drug Target Identification and Network Pharmacology." *IEEE*, (2015). DOI: 10.1109/ICIIBMS.2015.7439493.

〔学会発表〕（計 8 件）

① **Kun-Yi Hsin**, Development of predictive machine leaning system in target protein identification and network pharmacology (2016). Advances in Systems and Synthetic Biology 2016. Evry, France.

② **Kun-Yi Hsin**, Application of machine leaning approaches in drug target identification and network pharmacology (2015). International Conference on Intelligent Informatics and BioMedical Sciences (ICIIBMS). Okinawa, Japan.

③ **Kun-Yi Hsin**, Yukiko Matsuoka and Hiroaki Kitano. Combining machine learning systems and multiple docking simulations for network pharmacology (2015). 4th Negative Strand Virus Symposium (NSV-J). Okinawa, Japan.

④ **Kun-Yi Hsin**. Discovery of anti-influenza agents and drug targets using pathway map (FluMap) and computational approaches (2015). **東京大学医科学研究所共同研究拠点成果報告会**. Tokyo, Japan.

⑤ **Kun-Yi Hsin**, Molecular Docking Simulations in the Era of Network Pharmacology (2014). **第 42 回構造活性相関シンポジウム (日本薬学会-構造活性相関部会)（招待講演）** Kumamoto, Japan.

⑥ **Kun-Yi Hsin** and Hiroaki Kitano, High-Throughput Virtual Molecular Docking (2014). BioJapan 2014, World Business Forum. Yokohama, Japan.

⑦ **Kun-Yi Hsin**, Samik Ghosh and Hiroaki Kitano. Using machine learning systems and multiple docking simulations for network pharmacology making safer drugs (2014). Advances in Biomolecular Modelling and Simulations using CHARMM. Dublin, Ireland.

⑧ **Kun-Yi Hsin**, Samik Ghosh and Hiroaki Kitano. Molecular Simulations for Network Pharmacology (2014). Symposium EuroQSAR - Understanding Chemical-Biological Interactions. St. Petersburg, Russia.

〔産業財産権〕

出願状況（計１件）

名称：Interaction prediction device, interaction prediction method, and computer program product.
発明者：**Kun-Yi Hsin**, Samik Ghosh and Hiroaki Kitano.
権利者：Inventors, The Systems Biology Institute (SBI) and Okinawa Institute of Science and Technology Graduate University (OIST).
種類：特許
番号：14/407,835
出願年月日：2014 **年** 12 **月** 01 **日**
国内外の別：国内 and **外国**

６．研究組織

(1)研究代表者

Kun-Yi Hsin (Okinawa Institute of Science and Technology Graduate University, Integrated Open Systems Unit, Staff Scientist)
研究者番号： 60604155