Application of exhaustive protein-protein interaction prediction system by using protein docking to signal transduction pathways

Yuri Matsuzaki[1†] Masahito Ohue[1,2] Nobuyuki Uchikoga[3]
Takashi Ishida[1] Yutaka Akiyama[1]
[†]y_matsuzaki@bi.cs.titech.ac.jp

1. Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, Japan
2. Department of Physics, Chuo University, Tokyo, Japan
3. Physics Department, Chuo University, Tokyo, Japan

Protein-protein interaction networks compose crucial part of cell regulation processes. We have developed a high troughput protein-protein docking system "MEGADOCK" which calculate high-throughput protein-protein interaction network prediction based on all-to-all rigid-body docking of protein tertiary structures. The system accepts a set of proteins tertiary structures data as input and outputs the possible interacting pairs from all the combinations of the input proteins. This docking based method has an advantage in providing not only predictions of protein pairs that add new relations for understanding biological pathway, but also predicted complex structures of the two proteins which give insight on the novel interaction mechanisms. Although such docking based calculation requires massive computational resources, recent advancements on the computational sciences have made such large-scale calculation feasible with massively parallel computers.

We have implemented MEGADOCK to compute number of docking problems by thread and process parallelization. Upon docking, ligand protein structure is rotated to 3600 different angles and then the system conducts a translational search of favourable binding site. Calculations of 3600 angles are distributed by thread parallelization implemented with OpenMP. Because the calculations for each pair are independent, we can parallelize an all-to-all exhaustive PPI prediction task on hundreds or thousands of CPU cores. This part was implemented using MPI library. A user can specify the number of receptor and ligand protein data to be assigned to a single processor after taking the memory capacity into consideration. We have tested this data parallelization using up to 24,576 nodes and showed it was sufficiently scalable.

Prediction of the relevant PPIs is performed according to the affinity scores calculated by the post-processing of all the docking results. Each docking calculation outputs several thousands of decoys that yield high docking score. We defined affinity scores based on the highest docking score and distribution of the docking scores of the high scoring decoys. It represents how improbable to find that highest score among the score distribution if the given pair of proteins is just a random pair that do not actually interact.

We performed pathway reconstruction of two canonical signal transduction pathways using MEGADOCK: bacterial chemotaxis (13 proteins, collected 89 structure data including multiple structures for each protein species) and human EGFR signal transduction pathway (49 proteins, 497 structures). The F-measure value of prediciton was 0.464 when applied to chemotaxis system and 0.385 when applied to human EGFR system.