課題名（タイトル）:

大規模タンパク質間ネットワーク推定に関する研究

利用者氏名：○宮野　悟*, 秋山　泰, 石田　貴士**, 内古閑　伸之***, 大上　雅史**, 佐藤　智之**
松崎　由理****

**所属：* 社会知創成事業 次世代計算科学研究開発プログラム
次世代生命体統合シミュレーション研究推進グループ データ解析融合研究開発チーム
** 東京工業大学　大学院情報理工学研究科
*** 中央大学　理工学部物理学科**

報告内容

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

本課題では、次世代生命体統合シミュレーションプロジェクトの委託研究として、大規模タンパク質間ネットワーク推定を行っている。

本研究では、生命体の理解に向けて重要な鍵の一つとなる「タンパク質間相互作用ネットワーク」の推定を可能とすることを目的として、大規模データ解析の手法に基づき、超高速に候補タンパク質間の相互作用の可能性を調べる計算アルゴリズムを開発するとともに、大規模並列計算機上で性能測定などを行う。

これまでに，小規模な系を例題としてシステム生物学的な研究上の課題に対して本システムを適用して、新たなタンパク質相互作用の候補の提案などを実施してきた（Matsuzaki, *et al.*, *Journal of Bioinformatics and Computational Biology*, 2009 他）。その際には、約 100 のタンパク質立体構造を対象に、全タンパク質の組み合わせについ相互作用可能性の有無を総当たりで予測した。この問題の規模は 100×100 程度であった。また，データ解析融合チームが掲げる「肺がんと薬」のテーマに関係の深いヒトの EGFR シグナル伝達系を例題として、500×500＝250,000 規模の問題で予備実験を行った。昨年度までに，データ解析融合チームの宮野研究室と連携して、遺伝子ネットワーク推定結果から得た探索対象タンパク質群約 1,000 構造を対象に、1,000×1,000＝1,000,000 規模の問題を対象に本課題で開発したタンパク質相互作用予測プログラムを適用し、その性能評価を行った。

大規模計算における性能評価を行うため、RICC への移植を行い、計算機実験および性能評価を実施している。

２．　具体的な利用内容、計算方法

本研究では、以下の手順で網羅的なタンパク質間相互作用予測を行う。タンパク質ドッキング計算には独自で開発したソフト「MEGADOCK」（大上他、情処論 TOM, 2010）を利用する。

網羅的タンパク質―タンパク質ドッキング

対象とするタンパク質の全組み合わせについて、1 対 1 のドッキング計算を行い、3,600 個の複合体候補構造（デコイ）を作成する。

ドッキング計算は、形状相補性に関する項 G と静電的相互作用の項 E の計算からなるドッキングスコアによってデコイを評価する。それぞれのタンパク質を 1 辺が 1.2Å の三次元ボクセル空間上に配置し、タンパク質の内部、表面、その他の種別によって，各ボクセル上に異なるスコアを代入する。形状相補性の項 G には、我々が新規に提案した「real Pairwise Shape Complementarity（rPSC）」スコア（大上他、情処論 TOM, 2010）を用いる。既存手法と比べて計算時間の面で有利なモデルとなっている。また、静電的相互作用を、アミノ酸残基ごとに CHARMM19 に基づいて原子に電荷を与え、ボクセルに分割してボクセル電荷 $q(l, m, n)$ を決定し、静電的相互作用の項 $E_R(l, m, n)$, $E_L(l, m, n)$ を決める（対象とするタンパク質ペアの一方をレセプターR、もう一方をリガンドL とする）。ドッキング予測の良さを表す評価値であるドッキングスコア S を

$$R(l,m,n) = G_R(l,m,n) + iE_R(l,m,n)$$
$$L(l,m,n) = G_L(l,m,n) + iwE_L(l,m,n)$$
$$S(\alpha,\beta,\gamma) = \mathcal{R}\left[\sum_{l=1}^{N}\sum_{m=1}^{N}\sum_{n=1}^{N} R(l,m,n)L(l + \alpha,m + \beta,n + \gamma)\right]$$

と定義する。(α, β, γ) はリガンドの平行移動ベクトルである。

MEGADOCK ではドッキングスコアを、リガンドを平行移動させながら全空間における畳み込み和として計算する。ある回転角に対してボクセルサイズ N に対して $N \times N \times N$通りの平行移動を行う。その中から最も良いドッキングスコアを持つリガンドの平行移動ベクトルを、その角度におけるドッキング結果として返す。回転角のサンプリングにおいては 3,600 通りの回転パターンで計算を行う。よって、1 つの複合体について計算されるドッキング結合部位は，ボクセルサイズが N のとき、$3{,}600 \times N^3$ 通りとなる。なお、計算時間は単純に畳み込み和をとると $O(N^6)$ だが、離散フーリエ変換 (DFT) と逆離散フーリエ変換(IFT) を用いて、

$$S(\alpha,\beta,\gamma) = \mathrm{IFT}\left[\mathrm{DFT}[R(l,m,n)]^{*}\mathrm{DFT}[L(l,m,n)]\right]$$

とし、高速フーリエ変換(FFT) を用いることで $O(N^3 logN)$ に削減することが可能となる。z^* は z の複素共役を表す。rPSC を用いることにより、既存の代表的ソフトウェア ZDOCK などが用いている複素数による表現モデルに比べて離散関数の虚数部に空きができる分、他の物理化学的相互作用の導入が可能となり、かつ FFT の計算回数の増加を回避し、計算の高速化を図っている。

デコイのリランキング

各タンパク質ペアについて、上位数千個のデコイを ZRANK によりリランキングする。複数の方法を用いてエネルギースコア計算を行う。

デコイのクラスタリング

デコイの構造類似性や相互作用プロファイルを利用したクラスタリングを行う。これまでに RMSD をもとにした構造類似性の評価方法を用いてデコイのクラスタリングを行ってきた。さらに、相互作用に着目した類似性によってデコイをクラスタリングする。この手法では、残基単位での相互作用の有無のパターンについて、Tanimoto index を用いて線形時間で類似度を計算する。

タンパク質間相互作用 (PPI) 評価と判定

デコイのクラスタリング結果を分析し、各タンパク質ペアについてタンパク質相互作用の有無を判定する。

本研究で特に大規模な計算が必要となるのはドッキング計算とクラスタリングである。ドッキングについてはプログラムの並列化が可能である。シングルプロセスによる大量ジョブの実行とともに、並列計算による計算効率を測定し、MEGADOCK の性能を評価する。

3. 結果

並列化と計算効率の向上

OpenMP によるノード内の回転角度並列と、MPI によるレセプターごとのドッキングの並列化を実装し、並列性能を計測した。このハイブリッド並列化したコードを RICC 上でチューニングし、次世代機「京」で大規模並列計算を実行した（2013 年 2 月現在 320,000 コア以上を達成）。

また、ドッキングにおいて計算時間の 7 割以上を占める FFT 計算について、FFTW, FFTE, CSSL の三つのライブラリの FFT 計算エンジンを用いて計算速度を比較した。さらに、スレッド並列の方法について (i)FFT 計算をスレッド並列で計算する方法、(ii)レセプターに対するリガンドの回転角度について並列で計算する方法 の二つを実装して比較した。

この結果、FFTW の回転角度によるスレッド並列を実装したものが、最も高速であることがわかった。ただし、タンパク質の大きさ(FFT の点数が 2 のべき乗とな

るもの)によっては、CSSL2 が他の FFT ライブラリより高速であった。

相互作用プロファイルを用いたタンパク質間相互作用予測システムの改良

これまでに、タンパク質複合体を残基間の相互作用パターンとして捉えることで、複数の複合体の類似度を容易に比較することのできる、相互作用プロファイル(Interaction FingerPrint＝IFP)を導入してポストドッキング手法を開発してきた。IFP を用いることで、ドッキング過程で得られた多数の候補複合体構造（デコイ構造）をクラスタ解析することが容易になり、見通しよく正解に類似したデコイ構造を探索することができるようになった。しかし、本質的な問題としてドッキング過程で正解に類似したデコイ構造を得ることができない場合が少なからずあり、十分な結果を得ることができなかった。この問題は剛体ドッキング過程において探索空間を網羅できていないことが原因と考えられる。

そこで、クラスタ分割することで得られた各デコイ構造グループから代表 IFP を作成することで探索空間を絞り込んで再ドッキングを行う手法を開発し、RICC 上でパラメータサーベイを行った。その結果、最初のドッキング課程で得られたどのデコイ構造よりも正解に近いデコイ構造を得ることができた。

実データ応用（１）ヒト EGFR 系への適用

東京大学医科学研究所 宮野研究室より提供された肺がんに関連する遺伝子のリストをもとに対応するタンパク質の構造データを収集し、ドッキングに必要な条件を満たすデータを選択した 1,921 構造データを用いて、網羅的ドッキング(約 400 万件)を行った。主な計算は理研 AICS の「京」で行い、RICC では後処理を中心に計算を行った。

現在、これまでに相互作用の報告のないタンパク質ペアのうち MEGADOCK による相互作用評価値の高い 16 件について、宮野研究室の推定したがん関連ネットワーク 256 件との対照を行っている。これまでのところ、MEGADOCK がスクリーニングした 16 件の未知相互作用のうち 7 件が、推定されたがん関連のネットワークにマッピングされた。これら 7 件について、発現データや局在データをもとに相互作用の可能性を検討し、提案する相互作用候補の絞り込みを行っている。

実データ応用（２）ヒトアポトーシス系への適用

MEGADOCK による PPI 予測結果に対し、既知相互作用表面データをもとにしたタンパク質間相互作用予測を組み合わせることにより、予測の正確さを向上する手法を開発した。この手法をヒトのアポトーシス系に適用して検証した。

4. 今後の展望

相互作用プロファイル解析などのポストドッキング解析について改良をすすめ、さらに未知相互作用の探索を続ける。また、今後は肺がん以外の疾患パスウェイ解析に対象を広げることも検討する。

平成 24 年度　RICC 利用研究成果リスト

【論文、学会報告・雑誌などの論文発表】

1. Masahito Ohue, Yuri Matsuzaki, Takashi Ishida, Yutaka Akiyama, Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis., *Lecture Note in Bioinformatics,* **7632:** 178-187, Springer Heidelberg, Nov. 2012.

【国際会議、学会などでの口頭発表】

1. Yutaka Akiyama, Large-scale protein-protein interaction network prediction by an exhaustive rigid docking system MEGADOCK., *4th Biosupercomputing symposium*, Tokyo, Japan, Dec. 2012.
2. Masahito Ohue**,** Yuri Matsuzaki, Takashi Ishida, Yutaka Akiyama, Improvement of the protein-protein docking prediction by introducing a simple hydrophobic interaction model: an application to interaction pathway analysis., *The 7th IAPR International Conference on Pattern Recognition in Bioinformatics (PRIB2012),* Tokyo, Japan, Nov. 2012.
3. 秋山泰, 大規模タンパク質ネットワーク推定とその応用, *ISLiM ソフトウェア研究開発報告会,* 東京, Jan. 2013.

【その他（ポスター発表）】

1. Yuri Matsuzaki, Masahito Ohue, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, Protein-protein interaction network prediction by using rigid-body docking tool MEGADOCK., *Biophysical Society 57th Annual Meeting,* Philadelphia, USA, Feb. 2013.
2. Yuri Matsuzaki, Masahito Ohue, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, Application of exhaustive protein-protein interaction prediction system by using protein docking to signal transduction pathways., *International Society for Computational Biology, ISCB-Asia 2012*, Shenzhen, China, Dec. 2012.
3. Masahito Ohue, Yuri Matsuzaki, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, MEGADOCK: a high-speed protein-protein interaction prediction system by all-to-all physical docking., *International Society for Computational Biology, ISCB-Asia 2012*, Shenzhen, China, Dec. 2012.
4. Yuri Matsuzaki, Masahito Ohue, Takashi Ishida, Yutaka Akiyama, Application of exhaustive protein-protein interaction prediction system by using protein docking to signal transduction pathways., *20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2012)*, Long Beach, USA, Jul. 2012.
5. Masahito Ohue, Yuri Matsuzaki, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, MEGADOCK: a rapid screening system of protein-protein interactions by exhaustive docking., *20th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB2012)*, Long Beach, USA, Jul. 2012.
6. Yuri Matsuzaki, Masahito Ohue, Takashi Ishida, Yutaka Akiyama, Application of exhaustive protein-protein interaction prediction system by using protein docking to signal transduction pathways., *3DSIG: The 8th structural bioinformatics and computational biophysics meeting*, Long Beach,

USA, Jul. 2012.

7. Masahito Ohue, Yuri Matsuzaki, Nobuyuki Uchikoga, Takashi Ishida, Yutaka Akiyama, MEGADOCK: a rapid screening system of protein-protein interactions by exhaustive docking., *3DSIG: The 8th structural bioinformatics and computational biophysics meeting*, Long Beach, USA, Jul. 2012.
8. 秋山泰, 松崎由理, 内古閑伸之, 石田貴士, 大上雅史, 網羅的タンパク質ドッキング解析プログラム MEGADOCK の開発と応用, *ISLiM ソフトウェア研究開発報告会,* 東京, Jan. 2013.
9. 大上雅史, 松崎由理, 石田貴士, 秋山泰, MEGADOCK: 大規模タンパク質間相互作用予測システムとその応用, *ハイパフォーマンスコンピューティングと計算科学シンポジウム (HPCS2013)*, Jan. 2013.

【その他（研究会報告）】

1. 山本航平, 大上雅史, 石田貴士, 秋山 泰, 構造情報に基づくタンパク質間相互作用ネットワーク予測精度の改善., *情報処理学会研究報告バイオ情報学(BIO)*, **2012-BIO-32**(14):1-7, Dec. 2012. [第 32 回バイオ情報学研究会, 京都]
2. 大上雅史, 松崎由理, 石田貴士, 秋山 泰, MEGADOCK を用いたタンパク質間相互作用予測のヒトアポトーシスパスウェイ解析への応用., *情報処理学会研究報告バイオ情報学(BIO)*, **2012-BIO-32**(13):1-8, Dec. 2012. [第 32 回バイオ情報学研究会, 京都]
3. 大上雅史, 松崎由理, 石田貴士, 秋山 泰, 簡易疎水性相互作用モデルによるタンパク質間ドッキング予測の高精度化., 情報処理学会研究報告バイオ情報学(BIO), **2012-BIO-29**(21):1-3, Jun. 2012. [第 29 回バイオ情報学研究会, 沖縄]
4. 下田雄大, 石田貴士, 秋山 泰, タンパク質間ドッキング予測における実空間での効率的な評価スコア計算方法の研究., 情報処理学会研究報告バイオ情報学(BIO), **2012-BIO-29**(20):1-3, Jun. 2012. [第 29 回バイオ情報学研究会, 沖縄]
5. 藤原隆之, 松崎由理, 石田貴士, 秋山 泰, タンパク質間ドッキング予測における機械学習を用いた目的関数の動的調整., 情報処理学会研究報告バイオ情報学(BIO), **2012-BIO-29**(19):1-3, Jun. 2012. [第 29 回バイオ情報学研究会, 沖縄]