平成 22 年度 RICC 利用報告書

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　：○秋山泰　石田貴士　内古閑伸之
大上雅史　佐藤智之　藤原康広　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

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。現在では，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、その性能評価を進めている。

開発中のプログラムについて、2010 年 6 月までは、東京工業大学の「TSUBAME 1.2」システム上で予備実験等を行ってきたが、8,000 コアを目指した、より大規模な性能評価を行うため、2010 年 7 月より 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。

i.　<u>網羅的タンパク質―タンパク質ドッキング</u>

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。

ii. <u>デコイのリランキング</u>

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。

iii. <u>デコイのクラスタリング</u>

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。

iv. <u>タンパク質間相互作用（PPI）評価と判定</u>

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 i. ドッキングと iii. 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

i. タンパク質間相互作用予測システムの性能評価および改良

MEGADOCK のタンパク質間相互作用評価法の改善、ランキング方法の改良と共に、タンパク質の柔軟性を考慮する目的で、主鎖レベルとアミノ酸残基側鎖レベルでの前処理を行い、ドッキング精度の向上を図った。

① <u>アンサンブルドッキング計算</u>

主鎖レベルについては、構造アンサンブルの作成およびアンサンブル間での相互作用予測を試みた。　具体的には、16 種のタンパク質立体構造について分子動力学計算によって各々10 程度の構造をサン

プリングし、それらの間の網羅的ドッキング計算（16×10）×(16×10)= 25,600 件　を行った。代表構造の抽出方法を 4 種類適用して比較するため、合計 102,400 規模の計算を行った。この結果、これまで良い精度を得られていない、非結合（unbound）状態の結晶構造を用いたタンパク質間相互作用予測精度が向上した（F 値による評価で 0.08 から 0.40 に上昇）。

② アラニン変換による前処理評価

アミノ酸側鎖レベルの柔軟性を考慮した前処理では、側鎖をアミノ酸残基のなかで比較的短いアラニン残基に変換し、ドッキング結果を評価した。実際に 44 ペアのタンパク質複合体について、結合状態と非結合状態のタンパク質立体構造を用い、かつ、各々の構造でアミノ酸残基側鎖のアラニン変換をする場合としない場合（44×4=176 件のドッキング計算）との結果を比較した。相互作用プロファイルを用いたクラスタ解析により各条件について各ペアから得られたデコイを 10 グループに分割したところ、エネルギースコアが最も低く、メンバ数の最も多いグループが最も正解に近かった件数は、結合状態のタンパク質ペアを用いた場合より、前処理を行った場合の方が約 7%(44 ペア中 3 ペア）増加した。

ii. 計算効率向上

(a)我々の開発した形状相補性スコア(rPSC)モデルのパラメータ最適化、(b)各タンパク質ペアのドッキング計算高速化のためのチューニング、(c)FFT ライブラリ(FFTW)のスレッド機能と SIMD 機能を利用したノード内並列の実装と性能計測を行った。以上の(a)，(b)，(c)の改良により、計算速度を従来の約 2 倍にすることができた。

さらに、(d) OpenMP によるノード内の回転角度並列の設計（MPI と OpenMP による二段階並列化）を進めているところである。

iii. 実データへの応用

細菌走化性シグナル伝達系、ヒト EGFR シグナル伝達系について、MEGADOCK による網羅的 PPI 予測を行い、従来一般的なベンチマークデータで得た精度と同程度の精度(F 値 0.36 程度)を得られることを確認した。また、肺がん薬剤に関連するタンパク質と EGFR 系のタンパク質との相互作用予測も進行中である。

4． 今後の計画・展望

i. タンパク質間相互作用予測システムの開発

① アンサンブルドッキング計算

構造アンサンブル群を用いたドッキングにおいて、サンプリング法を相互作用面に着目することで評価し、タンパク質間相互作用予測精度を向上させる。

② アラニン変換による前処理評価

これまでの相互作用プロファイルを用いたクラスタ解析について、デコイのグループ分割法について問題があるため、相互作用部位予測に適した方法とパラメータの探索を行う。

③ ポストドッキング解析

ドッキング過程で得られた複合体集合から適切な代表構造をより効率良く抽出するために、複合体のアミノ酸残基間相互作用に着目したポストドッキング手法を開発する。

ii. 実データへの応用

今年度の EGFR 系を対象とした PPI 相互作用予測結果について、相互作用プロファイル解析などを利用して大規模な未知相互作用探索を進める。

5． 一般利用で演算時間を使い切れなかった理由

申請時には全体の約 1.4%程度の演算時間を見積

った。2011 年 2 月末時点において、既に一般利用の下限である計算資源 1%以上の演算時間は使用している。3 月末日までに見積もりどおりの 1.4%の演算時間に達する可能性はあるが、これまでに混雑状況によって投入したジョブが実際に同時実行される数が少なくなることが頻繁にあったため、現時点では不確実である。この混雑状況は事前の予測の範囲をこえており、本研究ではやむなく東京工業大学の環境での実施と組み合わせて計算を行った。

平成 22 年度　RICC 利用研究成果リスト

私たちは 2010 年度の途中から一般利用させていただきました．以下の論文は，査読後の確認計算等に RICC を使用させていただきました．査読後の論文内容の変更は困難でしたので，RICC 利用に関する記載はありませんが，RICC での再計算と不可分な成果であるため、今年度の報告に加えさせていただきます． 次年度以降は RICC 利用を明記いたします。

【論文、学会報告・雑誌などの論文発表】

Akiyama, Y, Matsuzaki, Y, Uchikoga, N, and Ohue, M, Exhaustive protein-protein interaction network prediction by using MEGADOCK. *BioSupercomputing Newsletter*, Vol. 3, 8, (2010).

Matsuzaki, Y, Ohue, M, Uchikoga, N, Ishida, T, and Akiyama, Y, Computer prediction of protein-protein interaction network using MEGADOCK -- application to systems biology. *TSUBAME e-Science Journal*, No. 2, 34-37, (2010).

Uchikoga, N, Hirokawa, T, Analysis of protein-protein docking decoys using interaction fingerprints: application to the reconstruction of CaM-ligand complexes. *BMC Bioinformatics*, Vol. 11, 236 (2010).

Shirota, M, Ishida, T, and Kinoshita, K, Absolute quality evaluation of protein model structures using statistical potentials with respect to the native and reference states. *Proteins: Structure, Function, and Bioinformatics* (2011), (accepted)

大上雅史，松崎由理，松崎裕介，佐藤智之，秋山泰, MEGADOCK ：立体構造情報からの網羅的タンパク質間相互作用予測とそのシステム生物学への応用，情報処理学会論文誌 数理モデル化と応用(TOM), Vol. 3, No. 3, 91-106 (2010).

【国際会議などの予稿集、proceeding】

Kusuma, W A, Akiyama, Y, Design and simulation of hybrid de novo DNA sequence assembly for large eukaryotic genome. *The 2010 International Conferences on Parallel and Distributed Processing Techniques and Applications (PDPTA 2010)* (2010).

【国際会議、学会などでの口頭発表】

Ohue, M, Matsuzaki, Y, Matsuzaki,Y, Sato, T, Akiyama, Y, *In silico* prediction of PPI network with structure-based all-to-all docking. *InCoB2010 - the 9th International Conference on Bioinformatics*, (2010).

Akiyama, Y, Matsuzaki, Y, Uchikoga, N, and Ohue, M, Exhaustive protein-protein interaction network prediction by MEGADOCK. *Asia Hub for e-Drug Discovery (AHeDD) Symposium 2010*, (2010).

Matsuzaki, Y, A computational screening system of protein-protein interactions: connecting protein structural information to biological pathway estimation. *ETHZ - Tokyo Tech Workshop：Computing with GPUs, Cells, and Multicores*, (2010).

Uchikoga, N, Hirokawa, T, and Akiyama, Y, Searching near-native decoys from various types of protein complexes by cluster analysis with Interaction FingerPrint. *The 48th Annual Meeting of the Biophysical Society of Japan*, (2010).

Ishida, T, Protein disorder prediction based on the evolutionary conservation. *Gordon Research Conference "Intrinsically Disordered Proteins",* (2010)

Ishida, T, GPU Acceleration of de Novo Protein Tertiary Structure Prediction, *ETHZ - Tokyo Tech Workshop：Computing with GPUs, Cells, and Multicores,* (2010).

大上雅史, 松崎由理, 松崎祐介, 佐藤智之, 秋山泰, MEGADOCK : 立体構造情報からの網羅的タンパク質間相互作用予測とそのシステム生物学への応用, 情報報処理学会研究報告 数理モデル化と問題解決（MPS）, Vol.2010-MPS-78 No.3, pp. 1-9. [2010-5-21, 第 78 回　数理モデル化と問題解決 研究発表会, 群馬大学, 群馬県前橋市] (2010).

ポスター発表

Matsuzaki, Y, Ohue, M, Matsuzaki, Y, Sato, T, and Akiyama, Y, A computational screening system of protein-protein interactions: connecting protein structural information to biological pathway estimation. *11th International Conference on Systems Biology (ICSB)*, (2010).

Matsuzaki, Y., Uchikoga, N., Ohue, M., Ishida, T. and Akiyama, Y., Exhaustive protein-protein interaction prediction with MEGADOCK and its application to signaling pathway estimation. *Asia Hub for e-Drug Discovery (AHeDD) Symposium 2010*, (2010).

Uchikoga, N, Hirokawa, T, and Akiyama, Y, Post-docking analysis with Interaction FingerPrints. *Asia Hub for e-Drug Discovery (AHeDD) Symposium 2010,* (2010).

Ohue, M, Matsuzaki, Y and Akiyama Y, Development of a Protein-RNA Interaction Prediction Method Based on a Docking Calculation. *The 2010 Annual Conference of the Japanese Society for Bioinformatics (JSBi2010)*, (2010).

Uchikoga, N, Hirokawa, T, and Akiyama, Y, Cluster analysis in post-docking process for rigid-body docking problem by using Interaction Fingerprints. *CBRC2010* (2010).

Hotta, S, Toshimoto, K, Ikeda, K, Kusama, M, Maeda, K, Sugiyama, Y, and Akiyama, Y, A web-based system for drug clearance pathway prediction. *Asia Hub for e-Drug Discovery (AHeDD) Symoposium 2010*, (2010).

Owatari, Y, Sekijima, M, and Akiyama, Y, Binding analysis of neuraminidase inhibitors by fragment molecular orbital calculation, *Asia Hub for e-Drug Discovery (AHeDD) Symoposium 2010*, (2010).

松崎由理, 内古閑伸之, 大上雅史, 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シグナル伝達系への応用, 第 3 回 バイオスーパーコンピューティングシンポジウム, (2011).

大上雅史, 松崎由理, 内古閑伸之, 石田貴士, 秋山 泰, タンパク質と RNA の立体構造に基づいた網羅的計算による相互作用予測, 2011 年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(HPCS2011), (2011).

内古閑伸之, 広川貴次, 秋山泰, 相互作用プロファイルを用いたタンパク質間ドッキング問題における前処理の評価, 第 3 回バイオスーパーコンピューティングシンポジウム, (2011).

内古閑伸之, 広川貴次, 秋山泰, 相互作用プロファイルによるタンパク質複合体予測のポストドッキング解析, 第 38 回構造活性相関シンポジウム, (2010).