Vol.2012-BIO-29 No.20
2012/6/29

タンパク質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

下田 雄大[1]　石田 貴士[1]　秋山 泰[1]

概要：タンパク質ドッキング計算では，高速フーリエ変換 (Fast Fourier Transform，FFT) を応用した高速計算法が知られているが，FFT を用いる場合，評価関数が畳込み和の形式に限定され，設計の自由度が低くなるという欠点がある．そこで，本研究ではより複雑な評価関数を用いることを想定し，FFT を用いない実空間上でのドッキング計算を考える．FFT を利用しないことで生じる計算コストの増大に対し，高スコアの複合体構造の偏在を利用してヒューリスティックに高スコアの複合体構造のみを階層的に探索することで計算結果を変えずに計算時間を短縮するための手法を提案する．

キーワード：タンパク質間相互作用，タンパク質ドッキング，ヒューリスティック探索

An efficient score calculation method without using FFT in protein docking prediction

Shimoda Takehiro[1]　Ishida Takashi[1]　Akiyama Yutaka[1]

Abstract: In protein-protein docking, a fast calculation method using fast Fourier transform (FFT) is well known, but the form of the evaluation function is limited to the sum of convolution. In this study, we developed an efficient docking calculation method without using FFT in order to use various evaluation functions. Against the increase of computational cost, we proposed the heuristic method that hierarchically searches only high-score complex structures using the locality of high-score complex structures.

Keywords: Protein-Protein Interaction, Protein-Protein Docking, Heuristic search

1. はじめに

タンパク質同士の相互作用は，生命現象の中心的な役割を果たしており，相互作用のペアや，その複合体構造の理解は病因の解明や創薬において重要である．タンパク質ドッキングは複合体構造の予測のみならず，タンパク質間相互作用の予測にも利用されており [1], [2]，重要性を増している．そのため現在までに ZDOCK[3]，MEGADOCK[2] などの数多くのドッキング予測プログラムが開発されてきた．タンパク質ドッキング計算では，高速フーリエ変換 (Fast Fourier Transform，FFT) を応用した高速なドッキングスコア計算法が知られているが，FFT を用いる場合，評価関数が畳込み和の形式に限定され設計の自由度が低くなるという欠点がある．また，その FFT の利用による計算速度の向上は，計算オーダのレベルでの議論は行われてきたが，実際の計算時間でどの程度の差があるのか，ヒューリスティクスを入れた場合どうなるかについては検証が行われていない．

本研究では将来予想されるより複雑な評価関数を用いたドッキング計算に対応することを想定し，FFT を用いない実空間上でのドッキング計算を考える．我々のグループで開発された MEGADOCK をベースとし，まず FFT を用いた場合と用いない場合の実際の計算時間の差について確認を行い，更に高スコアの複合体構造の偏在を利用して

[1] 東京工業大学 大学院情報理工学研究科 計算工学専攻
Graduate School of Information Science and Engineering, Tokyo Institute of Technology

ヒューリスティックに探索することで計算結果を変えずに計算時間を短縮し，効率的な計算を行う手法を提案する．

2. タンパク質ドッキング計算

ドッキング計算を行う対象のタンパク質ペアについて，一方を**レセプター**，他方を**リガンド**とする．まず，レセプターとリガンドを 3 次元のグリッド (格子) として表現し各グリッドにタンパク質の特性を表すスコアを与える (図 1)．その後リガンドを様々な角度に回転させ，レセプターの周囲を平行移動させて多数の複合体を生成する．各平行移動パターンの複合体構造について，重なったグリッドのスコアの積の総和を取り各複合体構造の評価スコア，ドッキングスコアとする．

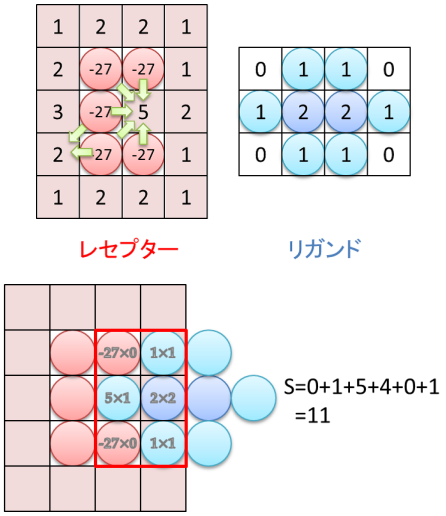

図 1 上:グリッドの rPSC スコアの付与例
下:ドッキングスコアの計算例

現在，MEGADOCK では形状相補性モデルに rPSC[2]，静電相互作用モデルに CHARMM19 による分子電荷に基づいた計算を用いている．MEGADOCK ではリガンドを 3,600 通りのパターンで回転させ，レセプターとリガンドの 3 次元グリッドの 1 辺の長さを N として $(2N)^3$ 通りの平行移動を行う．よって，1 つのタンパク質ペアについてドッキングスコア計算される複合体構造の数は，$3{,}600 \times 8N^3$ となる．ドッキング計算の計算量は，単純に積和計算を行うと $O(N^6)$ だが，ドッキング計算に高速フーリエ変換を利用することで，計算量を $O(N^3 \log N)$ に削減する事が可能である．

3. FFT を用いない実空間での評価スコア計算の高速化

FFT を用いないドッキング計算に FFT には適用できないヒューリスティックな手法を取り入れることで，計算時間の短縮を行い，効率的な計算の実現を目指す．

新たな計算量の削減方法を提案する前に，FFT 計算の際に必要となっているグリッドサイズの調整の問題について述べる．FFT を用いる場合，グリッドが立方体で，かつ立方体の 1 辺のグリッドの数が，FFT の底を基数とする冪乗数である必要がある．そのためスコアが 0 のグリッドを足して拡張することで，グリッドサイズの調整が行なわれている．しかし，FFT を用いない場合，必要最小限の直方体として扱うことで FFT に比べスコア計算を行う複合体構造の数と大きさを減らすことが可能である．

FFT を利用する場合，全ての平行移動パターンの複合体構造のドッキングスコア計算をまとめて行うが，実際には出力として使用するのはごく一部である．FFT を用いない場合は計算する構造を柔軟に選択することができる．平行移動パターンの空間上の高ドッキングスコア複合体構造の分布には偏りがある．図 2 の左上に平行移動パターン空間上の高スコア構造の分布例 (3 次元の分布から切り取った 2 次元の図) を示した．青色のマスに対応するのが高スコアの複合体構造である．中心部の複合体構造はレセプターとリガンドの実体が重なりペナルティを受けるためスコアが低い．また，外縁部の複合体構造はレセプターとリガンドの重なるグリッドが少ないためスコアが低い．この偏りを利用して，低ドッキングスコアの複合体構造の計算をできる限り避けることを考える．まず平行移動パターンを k 個飛ばしで動かして一部の複合体のみを計算し，その中でドッキングスコアが閾値 T を超えた複合体構造のみ平行移動パターン空間上で隣接する複合体構造を計算することを階層的に繰り返す．その様子を図 2 に段階的に示した．低スコアの複合体の計算を省略できていることがわかる．

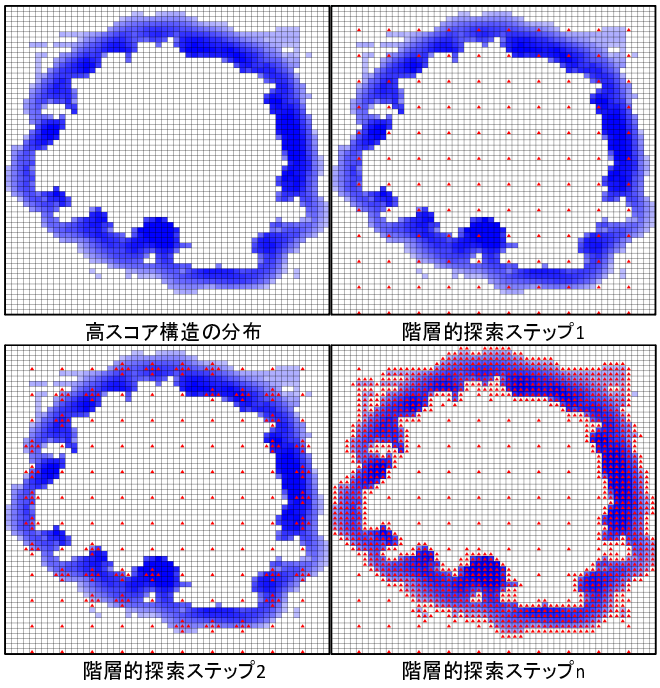

図 2 左上：平行移動パターン空間上の複合体構造のスコア分布
右上，左下，右下：階層的探索ステップ 1，2，n

4. 実験結果

実験には，タンパク質ドッキング予測に広く用いられている Protein-Protein Docking Benchmark 2.0[4] より，残基数が幅広く分布するように選出した 7 つのタンパク質ペア (1GCQ, 1AY7, 1ACB, 1CGI, 1PPE, 2PCC, 1SBB) を用いた．

計算時間の比較を図 3 に示す．グラフには 7 タンパク質ペアにかかった計算時間の合計を示す．ヒューリスティックを取り入れた階層的探索については，パラメータを変化させ，出力結果が変化しないという条件の中で最も高速なものの計算時間を示した．

(1) 単純な FFT 不使用
FFT を用いない場合，FFT を用いた場合に比べて平均で約 600 倍の計算時間がかかっている．単純な積和計算は $O(N^6)$，FFT を利用すると $O(N^3 \log N)$ というオーダの違いを反映して，大きなタンパク質になるほど計算時間の比率も大きくなっていた．

(2) グリッド削減
グリッド削減によってスコア計算を行う複合体構造のサイズとそれ自体の数を減少させ，計算時間が平均で約 7.0 倍短縮された．

(3) グリッド削減＋階層的探索
階層的探索を組み合わせることでスコア計算を行う複合体構造を絞り込み，計算時間がさらに平均で約 17 倍短縮された．最終的に FFT 利用との計算時間の差を約 5 倍にまで縮めることに成功した．

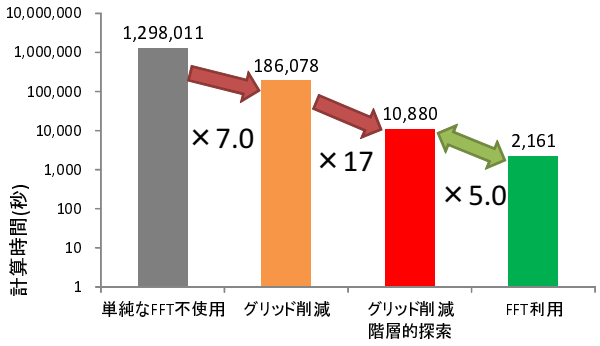

図 3 手法別の計算時間比較
(7 タンパク質ペアのドッキング計算時間の合計)

5. おわりに

FFT を利用した場合のドッキング計算の高速化について，ドッキング計算の条件を揃えた場合，FFT の利用によりドッキング計算が比較的小さいタンパク質でも数百倍高速化されることを確認した．また，その一方で余分なグリッドの削減と高スコア構造の階層的探索を組み合わせて実験を行い，約 100 倍の高速化を達成し，効率の面で現実的でないと考えられていた FFT を用いないドッキング計算を，実行可能な範囲で実現できることを示した．

この FFT を用いないスコア計算方法は，新たな物理化学的相互作用の導入を試行錯誤する場合や，複雑な多体の関係のモデル，条件分岐を伴うスコアモデルなどを FFT では扱えない問題を扱いたい場合に有用である．

今後の課題は，FFT を用いないドッキング計算のさらなる効率化と，FFT では扱えなかった評価関数のドッキング計算への適用によるドッキングの精度の向上である．

参考文献

[1] Wass M.N., Fuentes G., Pons C., Pazos F., Valencia A., Towards the prediction of protein interaction partners using physical docking, *Molecular Systems Biology*, 7, 469, 2011.
[2] 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, 3(3): 91-106, 2010.
[3] Chen R., Li L. and Weng Z., ZDOCK: An Initial-stage Protein-Docking Algorithm, *Proteins*, 52(1): 80–87, 2003.
[4] Mintseris J., Wiehe K., Pierce B., Anderson R., Chen R., Janin J. and Weng Z., Protein-Protein Docking Benchmark 2.0: an update, *Proteins*, 60(2): 214–216, 2005.