Rule-based modeling 法による受容体型チロシンキナーゼシグナル伝達系の数理モデル化と超並列計算機を利用したパラメータ推定

中茎 隆[1], 木村 周平[2], 仲 隆[3], 畠山 眞里子[1]

[1]理化学研究所 ゲノム科学総合研究センター 情報伝達システムバイオロジー研究チーム
[2]鳥取大学 工学部 知能情報工学科 [3]九州産業大学 情報科学部 知能情報学科
t-nakaku@gsc.riken.jp, kimura@ike.tottori-u.ac.jp,
naka@is.kyusan-u.ac.jp, marikoh@gsc.riken.jp

概要：受容体型チロシンキナーゼシグナル伝達系の詳細かつ大規模な数理モデルを Rule-based modeling 法により構築することを考える。このメソッドの導入によりタンパク質内の複数ドメインとその結合を考慮することができる反面、モデルの規模が著しく増大する。また、シミュレーションの際に必要なモデルパラメータの推定問題も大規模になるため超並列計算機の利用が必要不可欠となる。本発表ではベンチマークモデルを作成し、モデルの規模と並列度、計算コストの関係を調査する。

1 はじめに

細胞内シグナル伝達系では、タンパク質相互作用の微小なダイナミクスの変化が細胞のフェノタイプを決定したり、分子内の１つのアミノ酸置換が疾病の発生率や薬剤の感受性を左右するといった分子とネットワークが密接に関係している特徴を持つ。受容体型チロシンキナーゼシグナル伝達系はヒトの多くの癌に関与することが知られており、これまで多くの実験的、数理的な研究結果が報告されている[1]。本研究では、このシグナル伝達系を対象とし、これまで分かっているタンパク質相互作用、機能情報、細胞内局在情報を Rule-based modeling 法[2]を用いて包括的にモデル化し、リアリティのあるシミュレーション解析とその実用化応用を目指す。特に、大規模モデルのパラメータ推定問題は重要であるため、本発表ではテストモデルを作成し、モデルの次元数と計算コストに関するベンチマークテストを行う。

2 Rule-based modeling 法

本研究では、一分子内の複数の結合ドメインを考慮したBioNetGen[2]のアルゴリズムを用いたRule-basedのモデリングを行う。BioNetGenは一つのタンパク質内の複数ドメインを考慮し、１：Xのタンパク質相互作用の速度論モデルを構築できるアルゴリズムで国内外の評価・汎用性が高い。これにより、受容体やシグナルタンパク質の各々のリン酸化部位に結合する複数種類のタンパク質相互作用を仮定でき、詳細なモデル化が可能になる。モデルにn種類の分子が関わる場合，モデルはn本の連立微分方程式で構成される。そのため１回のシミュレーションに必要な時間計算量は，概ね分子種数nに比例する。BioNetGenではプリセットとして与えた分子種からその結合ドメインを考慮した複数の複合分子種を生成するが、文献[2]の例題モデルを用いた我々のベンチマークテストの結果、プリセットとしてn種類の分子を定義した際に、概ね$O(n^3)$でモデルの複合分子種数が増加することがわかった。

3 パラメータ推定問題

シミュレーションを実行するためには全ての速度定数や分子種の初期濃度といったパラメータが必要となる。それらのパラメータの一部は文献などの値が利用可能であり，他の一部については生化学実験により測定可能である。一方、測定できない残りのパラメータについては生化学実験によって測定可能な量から推定する必要がある。パラメータ推定は，生化学実験によって測定されたデータに合うようにパラメータを調整する問題であり，関数最適化問題として定式化される。この問題は非線形高次元の関数最適化問題となるため，非常に多くの計算量を必要とする。次世代スーパーコンピュータに実装される最適アルゴリズムの開発の準備として、木村らによって提案された進化的アルゴリズム GLSDC[3]を用いて並列度を上げた時の全実行時間（演算＋通信＋待ち時間）、通信パターン、使用メモリ容量、使用ディスク容量の見積もりを行った。現在使用している最適化アルゴリズムはマスタースレーブモデルに基づいて並列化されており，主にマスターノードでは全体の統括を，スレーブノードではローカルサーチの演算

を行っている。この最適化アルゴリズムは同時並列的に計算できるタスク数よりも CPU 数が充分に少ない環境での実行，および同時並列的に計算するタスクの計算量がほぼ同等であることを前提としてデザインされている。ところが本研究において我々がターゲットするモデルでは，計算量が非常に多いため，現実的な時間内で計算を終了させるには並列度を充分に上げる必要がある。

解くべき最適化問題の次元数をN_U、並列度をN_Cとすると、全実行時間の理論予測は次式で与えられる。

$$O(N_U^3 \times \mathrm{ceil}(3N_U/N_C)) \quad \cdots\cdot(1)$$

そこで、並列計算機を用いて実測実験を行い(1)式の検証を行った。実験として以下の 2 種類のテストを行った（図 1、2)。図 1 より、(1)式に従って、次元数N_Uの増加は全実行時間に$O(N_U^4)$でインパクトを与えることが確認できた。一方で、図 2 が示すように現在のアルゴリズムでは大規模問題に対しては並列度N_Cの効果が顕著には現れないことが分かった。これは、並列的に計算できるタスクの計算量、通信パターンにバラツキが存在するためである。以上のように，大規模なネットワーク解析では従来の最適化アルゴリズムが前提としていた条件とは異なる条件でのパラメータ推定を必要とするため，我々がこれまで使用してきたアルゴリズムを単に使用しただけではその探索性能を充分には発揮できないと予想される。本研究では、現在使用している最適アルゴリズムの改良を行う予定である。一方で、使用メモリ容量、使用ディスク容量に関しては、ベンチマークテストの結果、並列度や次元数の増加は大きなインパクトを与えないことが確認できた。

4　まとめ

膜受容体の動態と、そのシグナル伝達系を数理モデルによって解析することにより、リアリティのあるシミュレーション解析とその実用化応用が可能になる。このためには以上で述べてきたように超並列計算機が必要不可欠である。特に本研究で対象とする膜受容体ファミリーはイレッサ（肺がん治療薬）やハーセプチン（乳がん治療薬）などの標的となっており、受容体のわずかな違いが分子動態および細胞ネットワークへどう影響するかが理論的なモデルで示されれば、今後の薬剤開発の指針となる可能性が高い。

謝辞

ベンチマークテストは Riken Super Combined Cluster System (RSCC)を用いて実行された。また、理化学研究所次世代スーパーコンピュータ開発実施本部の杉原崇憲博士には並列計算環境について貴重なアドバイスを頂いたことに深謝します。

参考文献

[1] 畠山ら「RTK シグナル伝達系のシステムバイオロジー -数理解析の基礎と創薬への応用」実験医学 16, 2530-2535, 2006

[2] J. R. Faeder et al. “Rule-based modeling of biochemical networks” Complexity, 10, 22-41, 2005

[3] S. Kimura et al. “High Dimensional Function Optimization using a new Genetic Local Search suitable for Parallel Computers” Proc. of the 2003 Int. Conf. on Systems, Man, and Cybernetics, 335-342, 2003

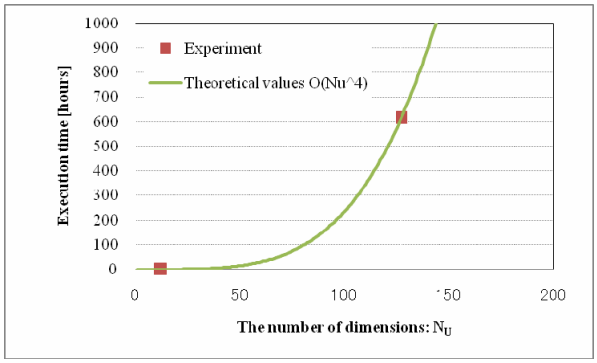

図 1 次元数N_Uに対する全実行時間の関係：N_C=32 の最適化問題に対して、次元数N_U（=12, 127）と全実行時間の関係を測定

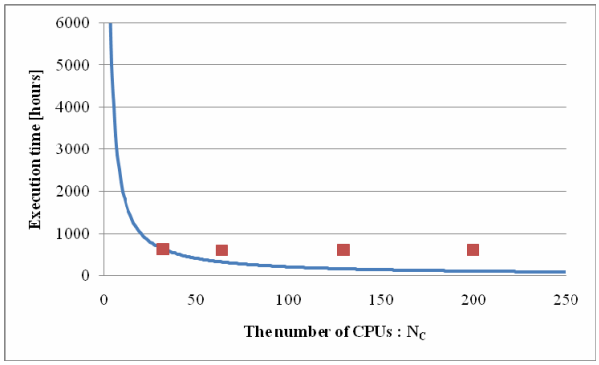

図 2 並列度N_Cに対する全実行時間：N_U=127 の最適化問題に対して、並列度N_C（= 200, 130, 64, 32CPU）と全実行時間の関係を測定