タンパク質相互作用ネットワークにもとづく立体構造および機能未決定タンパク質の解析優先度決定手法の提案

理学専攻　情報科学コース　岩崎　愛（指導教員：吉田裕亮　副指導教員：由良敬）

1　はじめに

タンパク質は立体構造を形成して機能するため、タンパク質の機能を理解する上でタンパク質の立体構造情報を得ることは重要であり、タンパク質の立体構造を測定する実験が数多く行われてきた。X線結晶解析法などによって現在までに明らかにされたタンパク質の立体構造の情報は、生体高分子立体構造データを集積したデータベースである Protein Data Bank(PDB) [1]に登録されている。

生命科学は現在、個々の分子の特性を研究する時代から、分子の集合の特性を調べる時代に突入した。かつては遺伝子やタンパク質ひとつの構造と機能を研究していたが、現在ではタンパク質などの集合体である細胞小器官や細胞、そして個体そのものの構造と機能を、その構成要素である分子に基づいて研究するようになった。タンパク質の立体構造データがそのような研究に有効に用いられるためには、すべてのタンパク質の立体構造が明らかになっていることが必要である。そこで本研究では、タンパク質立体構造データが細胞小器官などの機能解析を行うために十分な量になっているのか、また十分でない場合はいつ十分になると予想されるか、さらにはどのような順番で実験データを産出すれば、より早く機能解析をすすめることができるようになるかを明らかにすることを目的とした。

2　タンパク質構造既知領域の割合

2.1　対象データと計算手法

NCBI(National Center for Biotechnology Information)が提供する、全タンパク質のアミノ酸配列データベースである Reference Sequence (RefSeq)[2]から取得した全アミノ酸配列において、PDB に登録されているタンパク質のアミノ酸配列と一致もしくは類縁関係にあるタンパク質を、PHMMER[3]を用いて検索した。また、実験的に構造を決定することの難しいタンパク質として、膜タンパク質と天然変性タンパク質が知られていることより、RefSeq のアミノ酸配列から膜貫通領域と天然変性領域をそれぞれ TMHMM[4]と DisEMBL[5]を用いて予測した。これらの情報に基づき、立体構造既知領域の割合を計算した(式 1)。

$$立体構造既知割合 = \frac{構造既知アミノ酸残基数}{全アミノ酸残基数 - 天然変性領域アミノ酸残基数} \quad (式\ 1)$$

2.2　バクテリア由来のタンパク質における計算結果

バクテリアでの計算結果を図 1 に示す。現在、2,700 種超のバクテリアでゲノム塩基配列が決定されており、このゲノムデータに基づき、各バクテリア種がもつ全タンパク質のアミノ酸配列が明らかになっている。これらのアミノ酸配列に対して、上記計算を実行することができた。その結果、全タンパク質のうち 62%（アミノ酸残基単位で測定）が立体構造既知であることが分かった。また、ゲノム全体の天然変性部位の割合が 3.8%であるのに対して、構造未知部分における天然変性部位の割合は 4.7%であり、タンパク質の構造未決定部位には、構造決定部位と比べて天然変性部位と予測される領域が多いことがわかった。

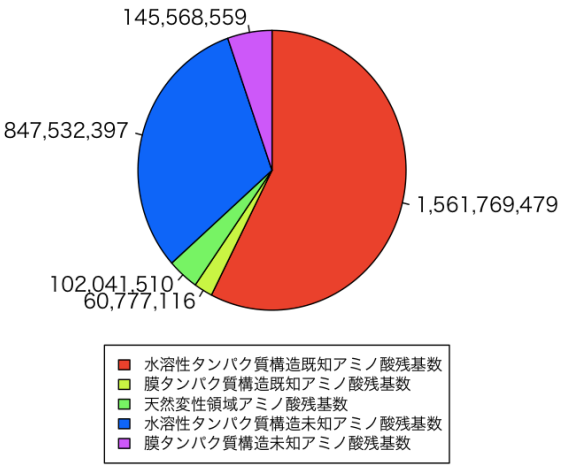

図 1：バクテリアにおける計算結果

3　タンパク質構造既知領域割合の時系列変化

3.1　対象データと計算手法

タンパク質の立体構造データベース PDB は毎週更新されていることより、上記の対応データは時系列にしたがって変化する。PDB の更新は、新規測定データの追加であることから、今までの増分を外挿することで、実験技術のめざましい進歩がない限りは、未来の増分を推定することができる。PDB には各データがいつ登録されたかが記載されているので、その情報を用いて、2000 年まで年ごとにデータをさかのぼりながら、第 2 節の計算を繰り返し、対応関係を蓄積した。

3.2　バクテリア由来のタンパク質における計算結果

年時変化のグラフを図 3 に示す。このグラフから、バクテリアの水溶性タンパク質の立体構造情報は、2008 年あたりから増加率が鈍化していることがわかった。2008 年からの増加率が維持されると仮定して、グラフを未来に向かって外挿すると、2032 年にバクテリアのもつすべての水溶性タンパク質の立体構造がわかることになる。膜タンパク質は、立体構造データの増加率が、水溶性タンパク質の増加率よりも急であることも明らかになった。構造を決定することが技術的に難しい膜タンパク質は、近年の技術革新により、構造決定の速度が上がっていることが考えられる。

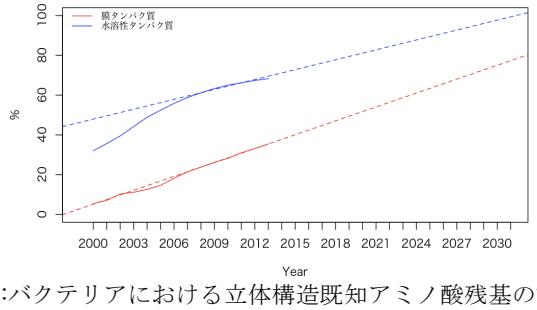

図 2:バクテリアにおける立体構造既知アミノ酸残基の割合

4　最大非連結グラフとなるタンパク質構造未決定領域の探索

4.1　対象データと計算手法

解析の優先度が高いタンパク質として、構造未決定領域の中で類縁配列が多いタンパク質が考えられる。そのようなタンパク質の構造が決定されることで、図 3 のグラフの傾きを大きくすることができる。そこで、タンパク質構造未決定である領域のアミノ酸配列を FASTA 形式ファイルで格納し、タンパク質構造未決定領域同士の配列類似度を、PHMMER を用いて計算した。各配列をノード、類縁関係をエッジとしてネットワークを描き、ノード数の大きな非連結グラフを探した。グラフデータベースである Neo4j[6]を使用しグラフを作成し、リレーション(エッジ)を探索することで、グラフを色分けするためのラベルを各ノードに振った（図 4)。各ラベルのノード数を求めることで独立集合のノード数を求めた。この非連結グラフのうちの 1 つのタンパク質の立体構造を決定することで、同じ色となったアミノ酸配列のタンパク質立体構造も、計算科学的手法で明らかにすることができる。

図 3：Neo4j によるグラフ

4.2 ヒトタンパク質における計算結果

ヒトのタンパク質構造未決定部位 73,179 領域のノード中、最大の非連結グラフとして、3,615 個のノードが存在するグラフが存在した。このグラフのタンパク質は、zinc finger protein であることが分かった。第 2 番目に大きなグラフとして、2,277 個のノードが存在するが存在し、このグラフのタンパク質は killer cell imunoglobuilin-like receptor であった。最大の非連結グラフのアミノ酸配列をもつタンパク質の立体構造が一つでも決定されれば、未決定領域 107 残基×3,615 個=386,805 アミノ酸残基の構造が決定されることになり、これはヒトタンパク質の全残基数の 0.8%に相当する。

5 タンパク質間相互作用ネットワークを利用した機能未知タンパク質の機能推定

5.1 対象データと計算手法

多くのタンパク質は、細胞内でお互いに相互作用することで、機能を実現する。よって、タンパク質間相互作用ネットワーク(PPI)は、タンパク質の機能を解析する上で重要な手がかりをもたらす。PPI データから、解析優先度の高いタンパク質を見つけだす手法を開発しその有効性を検討するため、具体的にヒトのオートファジータンパク質を解析した。ヒトのオートファジーに関与することが推定されているタンパク質(1,466 個)[7]の PPI ネットワークを調べたところ、相互作用が一つもない孤立タンパク質が存在することがわかった(図 5 の a ノード)。IntAct[8]から取得した全タンパク質の PPI データに基づいて、オートファジータンパク質と相互作用をもつすべてのタンパク質(6,560 個)についての PPI を描き、孤立タンパク質(図 5 の a ノード)を他のオートファジータンパク質(図 5 の c で囲まれたノード)に接続する上で重要なタンパク質(図 5 のｂノード)を探索した。

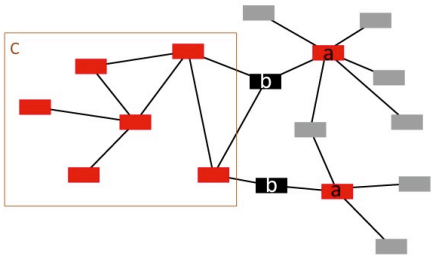

図 4:赤ノードがオートファジータンパク質

5.2 結果

図 5 のｂノードのようなタンパク質の中でも、式 2 を満たす重要なタンパク質として 5 個のタンパク質が見つかった。

$$\frac{\text{隣接するオートファジー関連タンパク質の数}}{\text{隣接するタンパク質の数}} \geq 0.30 \quad (\text{式 } 2)$$

これらのタンパク質は、ヒトのオートファジータンパク質と同定されていないタンパク質であるが、孤立しているオートファジータンパク質と、オートファジータンパク質の PPI を接続するタンパク質であるため、オートファジーに関係するタンパク質である可能性が高い。つまりこの 5 個のタンパク質について解析することで、細胞内での真のオートファジーPPI を見いだすことができると考えられる。

6 まとめ

バクテリアとヒトでの計算の結果、今現在、PDB の立体構造情報は十分ではなく、全バクテリアタンパク質の 62%、ヒト全タンパク質の 47%しかわかっていない。また、タンパク質間相互作用は、タンパク質の機能や構造に重要な関係があり、PPI を使用することで次に解析すべきタンパク質について提案することができた。PPI は非常に巨大なネットワークとなるため、どのような形式で表現するかが、今後の課題となる。

参考文献

[1] Berman, H., Henrick, K. and Nakamura, H. (2003) “Announcing the worldwide Protein Data Bank” *Nature Structural Biology* **10:** 980.

[2] NCBI Resource Coordinators (2016) “Database Resources of the National Center for Biotechnology Information” *Nucleic Acids Research* **45**: D1-D11.

[3] Robert D. Finn et al (2015) “HMMER web server: 2015 update” *Nucleic Acids Research* **43:** W30-W38.

[4] A. Krogh, B. Larsson, G. von Heijne, and E. L. L. Sonnhammer (2001) “Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes” *Journal of Molecular Biology*, 305(3):567-580

[5] R. Linding, L.J. Jensen, F. Diella, P. Bork, T.J. Gibson and R.B. Russell (2003) “Protein disorder prediction: implications for structural proteomics” *Structure* **11:** Issue 11, 4

[6] Neo Technology, Inc. Neo4j Version 3.1. https://neo4j.com/

[7] Homma, K., Suzuki, K., and Sugawara, H. (2011) “The Autophagy Database: an all-inclusive information resource on autophagy that provides nourishment for research” *Nucleic Acids Research* **39:** D986-D990.

[8] Orchard S et al (2013) “The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases” *Nucleic Acids Research* **42:** D358-D363.