nMDSによるタンパク質結合予測

Protein binding predictionusing non-metric multidimensional scaling method

物理学専攻　茂木大夢

1. 研究目的

本研究は、タンパク質の構造を網羅的に解析し、タンパク質の結合面を予測することである。具体的にはデータベースから取得したタンパク質複合体を、nMDSを用いて重ね合わせ、結合部位の予測を行っている。タンパク質結合のメカニズムは現時点でよくわかっていないが、結合を理解することにより生命システムの解明や、創薬などに貢献できると考えられている。

2. 結論

詳しい内容は修士論文や本紙の次章より先に譲るとして、結論は以下の通りである。

- 全てがうまくいったわけではないが、タンパク質の重なりの評価を行うことができた。これは先行研究では行われていなかったことである。
- ABAC database(本研究で用いたデータベースのこと)に存在するサンプルを複数個重ねて、結合部位を調べてみた。その結果、1clvについては複数箇所で結合部位を持っていることがわかった。
- 重ね合わせたドメインは座標軸で見ると結合部位として選ばれやすい部分があることがわかった。

3. 背景

タンパク質は生物を構成する重要な生体高分子であり、実に多様な現象を引き起こしている。というのもタンパク質は数万種類以上存在し、それぞれで機能が異なるからだ。タンパク質と一口に言っても、RNA代謝、DNA代謝、遺伝子発現制御、アミノ酸や核酸の生合成… etcなど、その機能の数は数え切れない。

近年では学問の進歩により、これらのタンパク質の多様な機能の要因が、タンパク質の立体構造の特異性にあることがわかってきた。生体反応における反応はタンパク質の立体的特異性による結合により、代謝を促進したり阻害したりする。

構造と機能は「鍵と鍵穴」の関係に例えられる。結合することにより、タンパク質の機能を特異的に阻害したり、逆に向上させることが行われている。だが最近の研究では、このモデルに当てはまるものばかりでないことがわかってきた。あるタンパク質は自身に柔らかい部分が存在し、結合する際にその部分の立体構造を変化させていることがわかってきたのだ。本研究はそのタンパク質の歪みを考慮した解析法として「nMDS」を提案して、解析してみたものである。

4. 研究概要

AB、A'Cの二つのタンパク質複合体(ただし、AとA'は相同性の高いドメイン。B,Cは相同性の低いドメイン)を多数集めてきて、AA'の部分で重ね合わせを行い、結合部位の位置を統計的に分析しているものとなる。(なお相同性が高いとは、SCOPの分類上、familyであることを意味し、”相同性が低い”とはSCOP上の異なったsuperfamilyか

ら習得したことを意味している)。なおこれらのタンパク質のデータはABAC database[5]を参考にしている。

5. 先行研究の疑問点

本研究は、大枠は本研究と同じであるが先行研究では、タンパク質を5つに分類して分析していた。先行研究には以下の疑問点がある。

・タンパク質の重ね合わせがうまくできているか？

先行研究では重ね合わせができているか否か、という指標は用意されていなかった。もしかしたらたんぱく質がA, A'の部分で重なっていない可能性があり、その原因次第では、先行研究の結果に信憑性がなくなってしまう。

・結合する際のタンパク質の歪みが考慮されているか？

たんぱく質は結合する際に構造を変化させることは、背景で述べた。全くこの歪みを考慮しない予測より、実際のたんぱく質結合の挙動に近い挙動をする予測をしてみるべきである。

以上の2点を解決するために、本研究ではもっと詳しく、精密に重ね合わせを行い、網羅的に分析を行ってみた結果である。これら二つの問題を解決するために、我々はnMDSを提案する。その際に評価できるであろう点は以下の通りである。

1. 原子間距離の順位により並べ直す事で、タンパク質の揺らぎを表現できる可能性がある。

順位では並べ直す手法であれば、原子間距離のとりうる値に揺らぎがある。nMDSでは、極端な例えだが、順位さえ正確であれば、どのような原子間距離をとっても良いことになる。この原子間距離の値の揺らぎが、タンパク質の結合面の揺らぎを表現している可能性がある。

2. 先行研究の課題点である、「タンパク質の重ね合わせ」をより正確に行える可能性がある。

上記二つの事から、「タンパク質を精度よく重ね合わせることにより、結合面を正しく予測する事が可能になる」可能性がある。

6. nMDSとは

nMDSとは、nom-metric multidimensional scalingの略であり、主には多変量解析の手段として用いられる。nMDSは以下のような特徴を持つ。

1. 順位相関を考慮した、収束アルゴリズムであること

通常のnMDSとは少し違ったアルゴリズムである。詳しくは論文中で述べている。

2. 大規模なデータに応用できる

本来nMDSの当てはめの判定は同様の多変量解析手法のPCAなどと比較し、良い場合があることがわかっている。「良い」とは、正確性を損なうこと無く、情報量の圧縮に成功しているという意味である。タンパク質に適用することにより、果たしてうまくいくかは誰も試していないため、不明だが、やってみる価値はあると考えた。詳しいnMDSについての説明は修士論文中に記載してある。

7. 手順

1. データをダウンロードしてくる

ここでは、論文中で記述されているABAC database[5]というところから、ABA'Cに関するするデータを特定した。

さらにpdb[1]から、特定した名前のタンパク質の情報(pdb ファイル)をダウンロードしてきた。

2. データの構文を解析し、ABACそれぞれに対応する部分を抜き出す

ダウンロードしたデータから、A, A', B, C部分を特定する。alignment表から文字列を抜き出す作業となる。これがpdb ファイルから取ってきたアミノ酸配列文字列の内部に存在しているか否かで判定した。

3. 得られたA, A'のタンパク質をalignment する

alignment とは二つの異なるアミノ酸配列の共通部分を特定する手法である。この情報はABAC databaseに記述があり、それにしたがって両者で共通したタンパク質を抜き出している。

4. 距離行列を作成する

距離行列とは二点間距離を計算し、行列にしたものである。ここでの距離とは、三次元座標上に存在する二点の原子間の距離である。単位はÅである。nMDSでは距離行列を用いて、起動するようになっている。

5. 距離行列からnMDSを使用する

距離行列を用いて、nMDSをここまでで得られた全てのサンプルに使用した。得られた結果をさらに解析した。

8. 結果考察

nMDSが最後まで起動し、出力できたものは全部で500個となった。ABAC databaseに登録されているサンプルは915個であり、そのうち218個のサンプルは結果が出なかったことになる。これらの手法の途中で失敗してしてまった理由は、ファイルのフォーマットが崩れていた、alignmentで失敗してしまった、Segmentation fault でプログラムが途中で強制終了されてしまった、ためである。

正しく重ね合わせが行われているか、否かという判断基準として、RMSDとタンパク質全体のボンドの平均距離変化率を考えた。この二つの指標の詳しい定義は修士論文中で述べている。

nMDSを用いてたんぱく質を重ねてみて、正しく重なったと言えるのは109個という結果になった。先行研究ではカテゴリー分けをし、タンパク質の類似度をbootstrap法のp-valueで計算している。p-valueが低いとランダムに配置されたものよりは、高い確率で類似関係があると言えるというもので、先行研究ではAA'のp-valueは極めて低いものが多かった。だが、これは”うまく重なる”ことを保証しているわけではなく、その上全てのカテゴリーについてp-valueが計算されているわけではなかった。類似度が高くても、正しく重なっているとは限らない。その上、他のカテゴリーではノイズが多かったという結果も出ていた。

一方で本研究ではnMDSを用いて、全てのサンプルについて重ね合わせを行ってみた。本研究でも同様に、全てがうまく重なったわけではないが、先行研究とは異なり、重なりの指標を定義することができた。この点で先行研究よりうまく言ったといえるであろう。

後に成功した109個のサンプルのうち、出現回数の多かったpdbファイルである1clvというタンパク質について詳しく見てみた。以下は1clvが使われたサンプルを複数個重ね合わせてみた結果である。なお以下は、1clvのA'部分のx, y, z軸について見てみた図である。横軸にx,y,z座標の値、縦軸にそのx,y, z軸中に出現した原子数を示している。赤色がA'の原子を表していて、そのうち青色が結合部位として選ばれた部分である。

図 1.

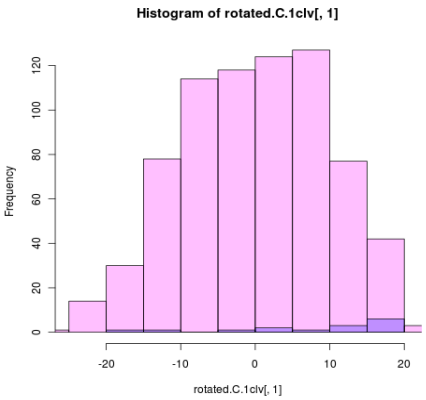

図 2.

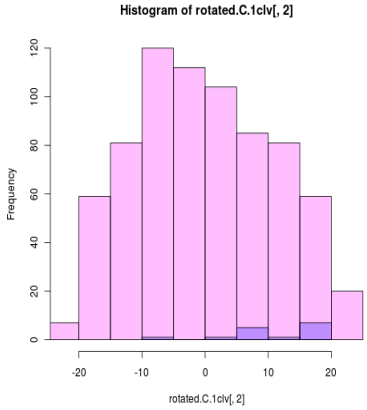

図 3.

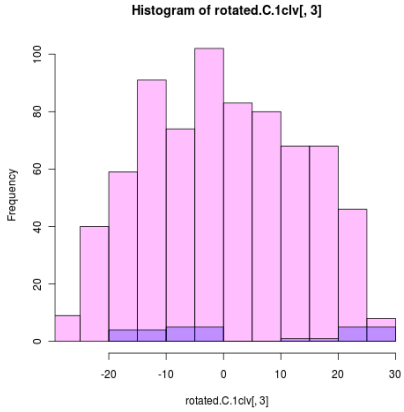

図 1.　1clv の A’ドメインの x 座標と結合部位の x 座標の出現回数

図 2.　1clv の A’ドメインの y 座標と結合部位の y 座標の出現回数

図 3.　1clv の A’ドメインの z 座標と結合部位の z 座標の出現回数

また、結合部位を詳しく見てみると、1clv の A’のドメイン部分は、全く別の結合部位が使われて、他のドメインと結合していることがわかった。

参考論文

[1] PDB, http://www.pdb.org/

[2]http://www.wwpdb.org/documentation/format33/v3.3.html

[3]Taguchi, Y.H., Oono, Y.: Relational patterns of gene expression via non-metric multidimensional scaling analysis, Bioinformatics, vol.21, pp.730-40, (2005).

[4] 田口　善弘, 大野　克嗣, 横山　和成:「非計量多次元尺度構成法への期待と新しい視点」統計数理　(2001) 第 49 巻　第 1号 133–153

[5] ABAC database http://scoppi.biotec.tu-dresden.de/abac/

[6] SCOP: Structural Classification of Proteins, http://scop.mrc-lmb.cam.ac.uk/scop/

[7] Beauty is in the eye of the beholder: proteins can recognize binding sites of homologous proteins in more than one way. http://www.ncbi.nlm.nih.gov/pubmed/20585553, PloS Comput Biol. 2010 Jun 17;vol.6, no.6 : e1000821

[8] R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0 URL http://www.R-project.org (2009)

[9]Hiromu Mogi, Y-h. Taguchi , Protein binding prediction using non-metric multidimensional scaling method , IPSJ SIG Technical Report, 2011, vol.2011-BIO-25, no.41, pp.1-2.

[10]Hiromu Mogi, Y-h. Taguchi, Protein binding prediction using non-metric multidimensional scaling method, 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops(BIBM), vol.2, pp.950-952.