プロテオミクスにおけるいくつかの問題： タンパク質の幾何と代数

（問題１）タンパク質の立体構造の分類

タンパク質はアミノ酸が 1 次元的につらなるひも状分子ですが、生体内では適当に折りたたまれ、一意的な立体構造をとります。これらの立体構造はいくつかの球状のかたまり(ドメイン)から構成されており、各ドメインの大まかな折れたたみ構造(フォールド)に着目すると、タンパク質を階層的なグループに分けることができます。

タンパク質の機能を予測したり解析するためには、このような分類が重要な役割を果たします。しかし、分類データベースによって分類の仕方が異なり、分類の自動化も出来ていません。

問題を難しくする要因[1]：

(a) タンパク質を、構造や機能の類似性するものをまとめてグループ(タンパク質ファミリー)に分けることは比較的容易にできる。しかし、階層のおおまかな構造を定めるそれらグループ同士の間の関係については、判断する人間の主観に左右される。また、互いに無関係な多くの小クラスターが散在することにもなる。

(b) 構造上の特徴だけからドメインを定義するのは難しい。一般には、最初にタンパク質をドメインに分割して、その後分類する。しかし、分割の仕方が分類結果を左右する。分類データベースの間の不一致の大きな原因は、このドメイン定義の違いにある。

(c) 局所的な構造の類似性と全体的な構造の類似性のどちらに重点をおくかで、分類結果が異なる。(例えば、タンパク質はおおきく all α、all β、α/β、そしてα+βの４つに分類される。しかし、α/βとα+βの区別はそれほど自明ではない。)

(d) フォールドと機能の相関は強いとはいえないので、構造の類似性と機能の類似性のどちらに重点をおくかで、分類結果が異なる。

(e) 分布に偏りがあり、ひとつのタンパク質ファミリーのみからなるフォールドの扱いが問題となる。大半のフォールドはひとつのファミリーしか含まないが、少数のフォールドは多くのファミリーを含む。(ほとんどのファミリーは、1000 程度のフォールドに属する。)

解決への取り組み：

立体構造の客観的な記述に向けて、いくつかの取り組みがなされています。[2]はタンパク質の立体構造を行列(距離行列)で表現し、その固有ベクトルを用いて、立体構造の分布を３次元座標で表示しています。[3]は、結び目理論の Vassiliev 不変量を応用してタンパク質を分類しています。また、[4]は全てのタンパク質の立体構造を記述するための標準構成部品の集合を用

意し、タンパク質の分類問題を標準構成部品による近似の問題に置き換えています。

参考文献：

[1] 木下賢吾,中村春木: 7章タンパク質の構造から見た生物情報,バイオインフォマティクス(美宅成樹,榊佳之編). 東京化学同人(2003).

[2] J.Hou, G.E.Sims, C.Zhang, S.H.Kim: A global representation of the protein fold space. Proc Natl Acad Sci 2003,100:2386-2390.

[3] P.Rogen, B.Fain: Automatic classification of protein structure by using Gauss integrals. Proc Natl Acad Sci 2003,100:119-124.

[4] W.R.Tayler: A 'periodic' table for protein structre. Nature 2002, 416:657-660.

（問題２）アミノ酸配列からのタンパク質立体構造の予測

自動化により大量処理が可能となった DNA 配列の解析とは異なり、タンパク質の立体構造を実験的に同定するのは時間と手間のかかる作業です。そのため、アミノ酸配列の情報があふれかえる一方で、立体構造が解明されているタンパク質の数はそれほど多くはありません(2006 年 1 月末現在で、約 3 万 5 千個)。

しかし、タンパク質の機能は立体構造と密接に関係しており、医薬品をコンピュータ上で設計するにも立体構造に関する情報は不可欠です。そこで、立体構造の予測が盛んに研究されています。また、CASP(Critical Assessment of Techniques for Protein Structure Prediction)と呼ばれる予測コンテストが隔年で開催され、研究者が予測技術を競う場となっています[1]。

問題を難しくする要因：

(a) アミノ酸は 20 種もあり、可能なアミノ酸配列の数は宇宙にある原子の数より多い
(b) タンパク質は、多くの原子から構成される多自由度系であり、取り得る構造の数が多い
(c) アミノ酸が一つ置き換わるだけで、タンパク質全体の構造が大きく変動する場合がある
(d) 多くの相異なるタンパク質が、類似の立体構造構造を形成する
(e) タンパク質には、他の分子と相互作用する際に、大きな構造変化を起こすものがある
(f) 単独では特定の立体構造をとらないタンパク質が存在する

解決への取り組み：

タンパク質の立体構造を予測する方法は、立体構造が既に判明しているタンパク質との類似性をもとに予測を行う比較モデリング[2]と、アミノ酸配列の情報のみを用いて予測を行う ab

initio(de novo とも呼ばれる)モデリング[3]の、二つに大別されます。

自然界には数百万のタンパク質が存在すると考えられていますが、その大部分のタンパク質の立体構造は、高々数千程度の基本構造(フォールド)で記述されると考えられています。従って実用的には、それらを全て実験で同定してしまえば構造予測の問題は比較モデリングで解決されます。しかし、それではタンパク質が立体構造を自発的に構成していく基本原理が解明されたとはいえません。

一方、ab initio 法ではタンパク質の立体構造は系の自由エネルギーが最小の構造に対応すると解釈され、アミノ酸や水分子の間の相互作用関数を考慮して最適な構造を探します。しかし、アミノ酸の数が 100 超えるタンパク質ではポテンシャルエネルギー曲面は非常に複雑で、グローバルミニマムを見つけることは至難の業です。

ところで、そもそもタンパク質の立体構造を指定するには、どんな情報がどのくらい必要なのでしょうか。[4]では類似構造をとるタンパク質のいくつかで共通するアミノ酸について、それらが同時に出現する確率を考えています。それによれば、その確率分布の情報だけを用いて、配列に類似性はないが類似の構造をとる新規のタンパク質を設計できるようです。立体構造は、比較的少数のアミノ酸-アミノ酸相互作用によって決定されているのかもしれません。残りのアミノ酸については、例えば、最終的な立体構造ではなく畳み込みの過程で生じる中間体の安定化に寄与しているとの報告[5]があります。

参考文献：

[1] 渕上壮太郎： CASP6 参加報告． PRC ニュースレター,No.2005.18 (2005). http://prc.protein.osaka-u.ac.jp/prc/prc2005/2005-18.html.

[2] 藤博幸,富井健太郎: 4 章 相同性検索技術の基礎,バイオインフォマティクス (美宅成樹,榊佳之編). 東京化学同人(2003).

[3] O.Schueler-Furman, et al.: Progress in Modeling of Protein Structures and Interactions. Science 2005, 310:638-642.

[4] M.Socolich,et al.:Evolutionary information for specifying a protein fold. Nature 2005, 437:512-518.

[5] C.Roodveldt, A.Aharoni, D.S.Tawfik: Directed evolution of proteins for heterologous expression and stability. Curr Opin Struct Biol 2005, 15:50-56.

（問題３）タンパク質-タンパク質相互作用の分類および予測

細胞内で行われている各種の生体プロセスは、タンパク質が相互作用することによって遂行

されています。従って、生体機能を解明するには、まずタンパク質の相互作用を調べる必要があります。

一般に、タンパク質は複合体を形成することで相互作用を行います。つまり、物理的に接触することで、タンパク質の分子の動き(熱的ゆらぎ)が他のタンパク質に伝わり、相互作用が実現されます。リボソーム(50 個のタンパク質と 3 個の RNA から成る)のように、長期にわたって存在するものもありますが、大半の複合体は一時的に形成されます。

複合体の種類については、複合体を構成する各タンパク質のアミノ酸配列が 30%以上一致するものを同一タイプと見なすと、現在のところ約二千種類が知られています。全体では一万種類程度に分類できるのではないかとの予測もあります[1]。しかし、タンパク質複合体は不安定なため実験的に立体構造を解析することは難しく、コンピュータを用いた構造予測に大きな期待が寄せられています。[2]

問題を難しくする要因：

(a) タンパク質の相互作用ネットワークなら、細胞中の全てのタンパク質について、それらがどのタンパク質と相互作用するかを同定すれば得られる。しかし、複合体を特定するには、それらの相互作用が「いつ」そして「どこで」行われるかも調べなければならない。

(b) タンパク質の中には、複合体を形成する際に構造変化を起こすものがある[3]。

(c) タンパク質の中には、環境によって立体構造が変化するものがある。一般には、立体構造が異なれば、異なる相手と異なる相互作用を起こす。しかし、同一の相手と異なる相互作用を起こす場合もある。

(d) 同一のタンパク質が、様々な相互作用に関与している。それらは、同じ複合体の形成にかかわる場合もあれば、別々の複合体にかかわる場合もある。

(e)タンパク質のなかには、普段はほどけていて、相互作用する時にのみ折りたたまれるものがある。

(f) 類似のタンパク質は、類似の複合体を形成する。しかし、立体構造が分かっている複合体の数が少なく、相同性を利用した構造予測にはあまり期待できない。

解決に向けての取り組み：

複合体の形成においては、隣接する分子が接触する領域が広いほど結合が安定します。そこで、構成要素となるタンパク質の立体構造が分かっていれば、表面に露出しているアミノ酸の空間的な相補性と化学的な親和性を考慮して、複合体の立体構造をある程度予測することができます(ドッキング法)。例えば[4]は、パターン認識の手法を応用して、フーリエ変換によりタンパク質形状の相補性の判定を行っています。しかし、研究は端緒についたばかりであり、今後十

年での進展が期待されています[5]。

参考文献：

[1] P.Aloy, R.B.Russell: Ten thousand interactions for the molecular biologist. Nat Biotechnol 2001, 22:1317-1321.

[2] A.Sali, R.Glaeser, T.Earnest, W.Baumeister: From words to literature in structural proteomics. Nature 2003, 422 13:216-225.

[3] C.-S.Goh, D.Milburn, M.Gerstein: Conformational changes associated with protein-protein interactions. Curr Opin Struct Biol 2004, 14:104-109.

[4] E.Katchalski-Katzir, et al.: Molecular surface recognition: determination of geometric fit between protein and their ligands by correlation techniques. Proc Natl Acad Sci 1992, 89:2195-2199.

[5] P.Aloy, M.Pichaud, R.B.Russell: Protein complexes: structure prediction challenges for the 21 st century. Curr Opin Struct Biol 2005, 15:15-22.

（問題４）タンパク質の離散微分幾何

コンピュータを用いてタンパク質の立体構造の分類や予測を行うには、バックボーンの類似性を数値化する必要があります。また、タンパク質間相互作用の分類や予測には、二つのタンパク質分子の表面形状の凹凸が、どの程度フィットしているかの数値化が必要です。

現在は、代表点に関する根平均二乗(Root Mean Square, RMS)が広く用いられています。しかし、例外値の寄与が大きいため、類似度が低いと精度が落ちます[1]。また、これは代表点の分布に関する評価であり、形状の相似性を表している訳ではありません。例えば、RMS ではバックボーンを一カ所で折り曲げたものと、元のバックボーンとの類似性を検出できません。

一般にコンピュータ上では、バックボーンは空間折れ線で、タンパク質分子は多面体で近似されます。そこで、微分幾何学的アプローチを用いて、空間折れ線のねじれ具合や多面体の表面の凹凸を数値化することが期待されています。

タンパク質分子の表面について：

分子の表面を定義するにはいくつかの方法があります。よく用いられるのは、分子の構成原子の各々を適当な大きさの剛体球で近似し、それらの和集合の表面を別の剛体球(水分子)を接触させながら転がす方法です[2,3]。球の分子側の球面が描く面は接触表面(Molecular Surface)、球の中心が描く面は溶媒露出表面(Solvent Accessible Surface)と呼ばれます(図 1)。分子間相

互作用を考える場合は表面形状の相補性を表現できる前者が、分子の立体構造の安定性を考える場合は後者が用いられることが多いようです。

これらの曲面はコンピュータ上では、数種の基本図形の組み合わせ、あるいは、多角形のメッシュ等により近似表現されます。球状タンパク質については、球調和関数を用いた表現も提案されています。

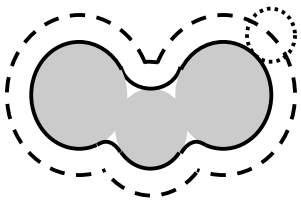


図 1. 分子表面の断面図。実線が接触表面、点線が溶媒露出表面を表す。

解決に向けての取り組み：

[4]は、バックボーンを空間折れ線とみなしてその曲率とトージョンを定義し、タンパク質の立体構造の類似性を定量化しています。また[5]は、バックボーンを四面体の連鎖で近似することで、その微分構造を定義しています。

表面形状に関しては、米国の BioGemetry プロジェクト[6]で様々な角度から研究が行われています。例えば[7]では、剛体球の和集合として表される物体について、その幾何学的量を計算するための様々なアルゴリズムが提案されています。[8]は、タンパク質複合体における分子の接触面をボロノイ多面体を用いて定義し、その幾何学的特徴を考察しています。

その他には、表面形状の記述に R.Forman の離散モース理論を用いた研究[9]もあります。また、生物学に特化していませんが、ドイツの MATEON には、離散微分幾何プロジェクト[10]があります。

参考文献：

[1] P.Koehl: Protein structure similarities. Curr Opin Struct Biol 2001, 11:348-353.

[2] M.G.Rossmann, E.Arnold (Fds.): Chapter 22: Molecular Geometry and Features. International Tables for Crystallography, Volume F: Crystallography of Biological Macromolecules. Dordrecht: Kluwer Academic Publishers (2001).

[3] M.L.Connolly: Molecular Surfaces: A Review. Network Science. http://www.netsci.org/Science/Compchem/feature14.html.

[4] S.Rackovsky, H.A.Scheraga: Differential Geometry and Polymer Conformation. 1. Macromolecules 1978,11:1168-1174.

[5] N.Morikawa, Discrete differential geometry of proteins: a new method for encoding three-dimensional structures of proteins. ArXiv: math.CO/0506082, 2005.
[6] The BioGemetry project. http://biogeometry.duke.edu
[7] H.Edelsbrunner: The Union of Balls and Its Dual Shape. Discrete Compt Geom 1995, 13:415-440.
[8] Y.-H.A.Ban, H.Edelsbrunner, J.Rudolph: Interface surfaces for protein-protein complexes. Proc 8th Intl Conf Res Comput Mol Bio 2004, 205-212.
[9] F.Cazals, F.Chazal, T.Lewiner: Molecular Shape Analysis Based upon the Morse-Smale Complex and the Connolly Function. Proc 19th ACM Sympo on Comput Geom 2003, 351-360.
[10] DFG Research Center MATHEON “Mathematics for key technologies”. Project F1: Discrete Differential Geometry. http://www.math.tu-berlin.de/geometrie/ddg.

（問題５）タンパク質複合体の表現

自動車や飛行機の設計をはじめとして、科学や産業の様々な分野でコンピュータによるシミュレーションが行われています。そこでよく問題となるのが、部品の形状やその可動域をモデル化することの難しさです[1]。これは、タンパク質相互作用の解析においても同様です。

例えば、タンパク質Ｂと相互作用を行う(つまり、複合体を形成する)二つのタンパク質ＡとＣを考えます(図2)。このとき、ＡとＣの形状を見れば、二つの相互作用が同時に起こらないことは明白です。あるいは、相互作用を行う二組のタンパク質Ａ、ＢとＣ、Ｄを考えます(図３)。この場合も、二組のタンパク質が同じ複合体を形成することは、それらの形状から一目瞭然です。しかし、コンピュータ上で座標データからこれらを推定するには、手間も時間もかかります。

現在のところ、タンパク質複合体の形状を誰かに説明しようと思うと、３次元 CG 表示を用いて視覚に訴えるほか方法がありません。相互作用のオーバーラップや複合体の因子分解などを簡単に計算できる、効率的な形状の表現方法が期待されるところです。

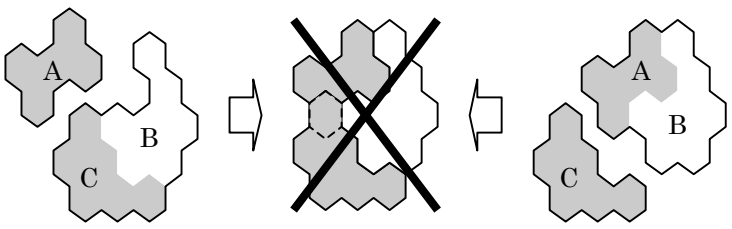
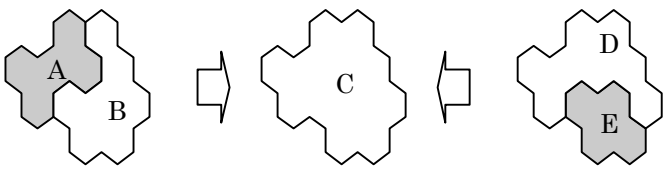


図２ 相互作用オーバーラップの模式図： A(BC) ≠ (AB)C



解決に向けての取り組み：

二つの多面体 A と B に対して、ミンコフスキー和 A+B :={a+b：a∈A, b∈B}あるいは和集合 A∪B の凸閉包を考えれば、多面体の集まりに代数構造を導入できます[2]。しかしその場合、演算を行うと多面体表面の凹凸に関する情報が失われます。その問題を回避するため、[3]は４次元空間のコーンを用いて多面体を近似し、コーンに沿った閉包を考えることで、代数構造を定義しています。そして、複合体形成を加法とするタンパク質の代数を提案しています。

ところで、面積や凸性などの概念と異なり、「形状」にはこれといった数学的な定義がありません。そこで[4]は「形状」の定義として、α形状という概念を提案しています。これを用いれば、複合体の構成因子のつながり具合を表現することができます[5]。

一方、並列コンピュータ言語を用いて、生体の動的な挙動を記述しようとする多くの試みがあります[6,7]。しかし、複合体の静的な形状を記述できるものは見当たりません。例えば[8]は、言語に「場所」の概念を導入し、複合体を細胞の区画の一つとして表現します：タンパク質 A と B から成る複合体 C は、A と B が入っている区画 C と同一視される。これは単なる階層構造であり、形状は全く考慮されません。また[9]は、タンパク質表面の凹凸(活性部位)に文字列を割り当て、形状の相補性を文字列の相補性として表現しています。この場合も、活性部位同士の位置関係は考慮されていません。

参考文献：

[1] M.Bern, et al.: Emerging Challenges in Computational Topology. ArXiv: cs.CG/9909001, 1999.

[2] P.MacMullen: The polytope algebra, Adv.Math 1989, 78:76-130.

[3] N.Morikawa: Toward Sub(Z^N) implementation of protein-protein interactions. 2004. preprint.

[4]H.Edelsbrunner, E.P.Mucke: Three-dimensional alpha shapes. ACM Trans Graphics 1994, 13:43-72.

[5]杉浦厚吉：「形と動きの数理」. 東大全学自由ゼミナール講義ノート(2005). http://www.simplex.t.u-tokyo.ac.jp/~sugihara/lecturenotes/freezemi/sec12.pdf.

[6] A.Regev, E.Shapiro: Cells as computation. Nature 2002, 419:343-343.

[7] V.Danos, C.Laneve: Formal Molecular Biology. Theoret Comput Sci 2004, 325: 69-110.

[8] A.Regev et al.: BioAmbients: An abstraction for biological compartments, Theoret Comput Sci 2004, 325:141-167.

[9] D.Prandi, C.Priami, P.Quaglia: Shape spaces in formal interactions. European Conf on Complex Systems 2005, http://complexsystems.lri.fr/PDF/p217.pdf.