応用分子進化学

タンパク質の機能解析を例として

藤 博幸 (京都大学化学研究所)
レクチャーノート作成: 桑田和正 (京都大学情報学研究科)

この講演では分子進化学からのバイオインフォマティクスへのアプローチを紹介する。ここでは種の進化を離れ、生体システムの進化、特にそれを支えるタンパク質の進化の情報をもとにしてタンパク質の機能や構造の解析を行う。

まず始めに、最近の分子生命科学の状況について講演者の視点から簡単に解説する。次に、タンパク質の機能解析を「生化学的機能」と「生物学的機能」の2つに分け、各々について話を展開する。また、どのような観点から上の2つの機能を分類するかは、イントロダクションの中で述べる。

1 イントロダクション

近年の分子生命科学の進展とバイオインフォマティクスの扱うデータの変遷について紹介しよう。

まず、セントラル・ドグマを確認しておこう。遺伝情報はDNAの塩基配列に書き込まれている。これが転写 (transcription) されてRNAに写され、そこから翻訳 (translation) されタンパク質が生成される。

DNA及びRNAの塩基配列のデータは4つのアルファベットからなる (向きを持った) 文字列として表現される。またタンパク質のアミノ酸配列は20個のアルファベットからなる (向きを持った) 文字列として表現される。これらのデータを情報科学的に処理するときには文字列の形のみで扱われる。

文字列として転写と翻訳を見ると、転写での対応は1対1であり、翻訳では3つの文字のセットがひとつのアミノ酸に変換される。

これらの話は前の講演でより詳しく扱われているのでそちらを参照されたい。

タンパク質はひも状の分子配列をしているが、機能しているときには折れたたまっていて立体構造をとっている。この構造の座標データ自体もバイオインフォマティクスの解析対象である。これらのデータは各原子の3次元のデカルト座標で表される。

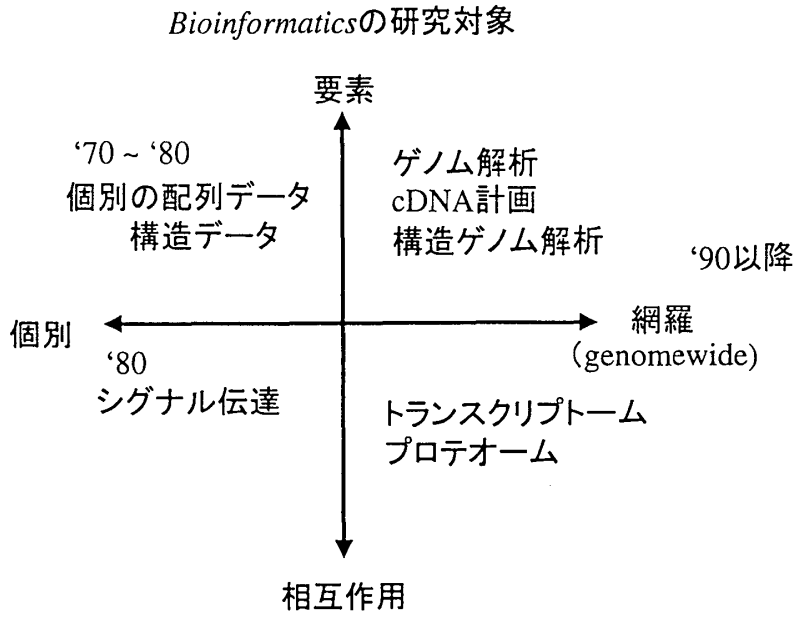

図 1: バイオインフォマティクスの研究対象の変遷

バイオインフォマティクスの研究対象の変遷

バイオインフォマティクスの研究対象がどのように変遷してきたのか振り返ってみよう(図1)。1970年代から1980年代にかけて核酸の塩基配列の研究は非常な進歩を見せた。塩基配列が決まればそこからタンパク質の配列も予測できる。配列データの急激な増加にともないデータベースが作成され、それらを解析する技術も同時にでき上がっていった。この頃には配列を決めること自体が目的であり、特定の生命現象に関与する遺伝子を同定することが分子生物学の主流であった。

図1にあるように、現在の分子生物学を2つの座標軸で把握してみよう。ひとつは、要素と相互作用。タンパク質の例をとると、要素とは単独のタンパク質を指す。もちろん個々のタンパク質は各々機能を持つ。しかし、高次の機能は他のDNAやタンパク質と合わさって相互に作用しながら現れてくると考えられる。こう考えると1970年代から1980年代の配列決定の研究は個々の要素を決定する研究であったと言える。ところがこれだけでは本来の興味であった生命現象を記述しきれないことが段々と分かってきた。

そのような流れを受けて、1980年代から要素決定の研究から分子間の相互作用へと分子生物学者の興味はシフトしていった。例としてシグナル伝達という現象が挙げられる。脳からホルモンが分泌されて対象の器官に到達したとき、ホルモンの情報は受容体と呼ばれる部分に結合し、ホルモンの情報を別の形にして細胞内に伝える。この伝達のメカニズムに興味が持たれていた。

ここで図1の横軸、個別と網羅に着目しよう。1980年代までの研究は個々の研究者が自分の興味に従って研究を進めていたといえる。例えば発がん機構なら、その機構を調べ

たいという動機の下で、そこで必要となるデータのみの収集、解析を行っていた。しかし1990年代に入り、いわゆるゲノムプロジェクトが進展してきた。このプロジェクトは生物が持っている遺伝情報を全て塩基配列として決定することを目指したものである。それまで個々の研究者が個別の興味で行っていた配列決定を網羅的に扱おうとした、とも言える。

しかし前にも注意したように、要素を確定するだけでなく要素間の相互作用を考えねばシステムとしての生命の理解には至らない、と考えられてきた。そうした中でトランスクリプトームやプロテオームの研究が現れてきたと見ることができる。これらは、大まかには、各々転写現象および翻訳現象を時間的空間的にどのように発現するか総体的に見る研究である。すなわち、個別現象としてのセントラルドグマにおける、転写と翻訳のプロセスを網羅的に捉えたものといえよう。セントラルドグマでのDNAを網羅的に見たものがゲノムであることに対応していると見ることができる。

上に見たように、バイオインフォマティクスの対象は質的、量的に変遷している。質的には、要素の解析から相互作用の解析への発展、量的には、個別の解析から網羅的な解析への発展が見てとれる。ここで網羅と言うとき、ある生物の有する遺伝情報が尽くされているという意味を込めていることが重要である。単に量が多いということではない。

また、研究者の興味は相互作用や網羅的なものにシフトしているわけだが、要素解析や個別解析の重要性が減じたわけではない。例えばタンパク質のフォールディング機構の解析は、アミノ酸の配列と立体構造との関係の理解につながる重要な問題である。しかしこの研究は網羅的でもないし相互作用の解析とも言いにくい。実際、要素と相互作用、個別と網羅は相補的な関係にある。例として、相互作用の研究における個別と網羅の比較をしてみよう。個別の解析を行うと、相互作用が分子レベルでどのような機構によって生じているのかが分かるが、ネットワーク中で起きている現象を全体像として捉えることはできない。逆に網羅的な解析を行うと、全体で何が起こっているかは分かるが、分子レベルでの機構は分からない。

これらの研究内容の変遷と分類を踏まえ、以下タンパク質の機能解析の話に移ることにしよう。

2　タンパク質の生化学的機能解析

タンパク質のバイオインフォマティクスの目的は、アミノ酸の配列、タンパク質の立体構造、そしてタンパク質の機能の3者の間の関係を明らかにすることである。そこでは前提として、アミノ酸の配列に立体構造の情報や機能に関する情報は全て含まれているとしている。

アミノ酸配列についての理解が深まれば与えられた配列からタンパク質の機能や立体構造をに関する情報を得ることができる。逆に、ある機能や立体構造を持ったタンパク質をデザインすることも可能になる。

しかし、アミノ酸配列からの構造・機能の予測は現時点ではまだ困難である。今行われ

ている一番実用的な方法は、相同なアミノ酸の配列から構造や機能を推測する、いわゆる相同配列の比較解析である。ここで、相同配列とは進化的な起源を共有する配列のことをいう。

一般に相同な配列は、同じ、あるいは類似した機能を持っている。従って、タンパク質の配列の解析の目的のひとつは、どれとどれが相同であるかを調べることとも言える[1]。相同タンパク質の形成には2つの要因がある。ひとつは種分化に伴うもの (orthologue) で、もうひとつは遺伝子重複によるもの (paralogue) である。遺伝子重複によって複製された遺伝子の多くは機能を失い偽遺伝子化するが、まれに新しい機能を獲得し、それが集団に広まることがある。遺伝子重複による相同配列はこのようにしてでき上がる。

生化学的機能と生物学的機能

ここまで機能という言葉を使ってきたが、以下これを生化学的機能と生物学的機能の2つに区別する。前者の例としては酵素活性やリガンド結合能[2]があり、記憶、発生、行動などの高次の生命現象は後者に属する。生化学的機能は個別のタンパク質 (前節の言葉では「要素」) の属性として捉えられるのに対し、生物学的機能はタンパク質のネットワーク (前節の言葉では「相互作用」) から現れるより高次の現象にあたる。

より具体的に、ペプチド性ホルモンのシグナル伝達における受容体のエクソン・シャフリングを、2つの機能の視点から考えてみよう。ホルモンは細胞膜を抜けられないので、受容体を通じて担っている信号を細胞内に伝える。受容体は2つのタンパクの複合体からなり、ホルモンが結合することで、2つの相対的な位置関係が変化する。その結果、受容体の細胞膜の内側にある部分で変質が生じ、信号が伝わる。

ここで、ホルモンとそれに対応する受容体が各々2種類あるとしよう。この2つの受容体にエクソン・シャフリングが起こると、2つの機能が部分的に入れ替わることがある。結合可能なホルモンは変化しないのに結合時に内部に伝える信号が変わる、といったことが起こり得るのである。結果として個々のタンパク質の変化だけでなく、相互作用のネットワーク自体を変化させる。ここにエクソン・シャフリングの重要性がある。前に定義した言葉を用いると、エクソン・シャフリングは生化学的レベルだけでなく、生物学的レベルでの機能の多様化をもたらす要因になる、と言える。

生化学的機能は単独のタンパク質の性質を見れば良いので、これは配列からの予測は可能であろう。一方生物学的機能は複数のタンパク質が構成するネットワークの機能を見なければならないので、個々の配列からの情報だけでは難しい。従って、多くの場合、配列からの予測対象になるのは生化学的機能である。

以下、先に提示した相同配列の比較解析について詳しく見ることにしよう。特に、これを利用して酵素活性やリガンド結合能のような生化学的機能がどのように理解できるか、例を通じて紹介する。

---

[1] Hubbard, T.J.P. *Current Opinion in Str. Biol.* **7**, 190-193 (1997).

[2] リガンドとは化学の分野で言う配位子とは異なり、酵素、受容体、輸送タンパク質などに結合する因子をさす。リガンド結合能とは、タンパク質の有するそのような因子を結合できる能力である。

相同配列の比較解析

相同配列の比較解析を3つのステップに分けて解説しよう。

まず、比較のために何らかの方法で相同配列を集めなければいけない。どれが相同であるか分かっている場合にはデータベースから直接データを集めればよい。しかし、どれが自分の持っている配列と祖先を共有しているか分からない場合は、配列を比較しながらデータベースを検索して、その中から相同と思しきものを集める操作が必要になる。

次に、集めてきた相同配列に対しマルチプル・アラインメントを作成する。

最後に、得られたマルチプル・アラインメントから構造／機能の情報を収集する。

実は、このステップは分子系統解析で行うものと全く同じである。最後に構造／機能の情報を収集する代わりに、分子系統樹を作成するわけである。

Step 1: データベース検索　まずデータベース検索から見ていこう。はじめに自分が機能／構造を知りたい配列 (問い合わせ配列) を用意する。そしてそれに類似した配列をデータベース中から検索する。特に検索の結果得られた配列の中で機能や構造が既知のものがあれば、それらと同様の構造あるいは機能を有するものと推測できる。これが最も単純な比較解析である。

検索の方法、類似度の測り方について、もう少し詳しく見てみよう。

データベース検索では、まず、問い合わせ配列とデータベース中の各配列をペア毎に比較し、2配列間の類似度が最大になるようアミノ酸の対応づけを行う。この対応づけをアラインメントと呼ぶ。その上で、データベース中の配列と問い合わせ配列の類似度は、各々アラインメントを作成したときの実現可能な類似の程度で定める。

進化の過程でアミノ酸配列に突然変異が起こるとき、配列の置換だけではなく、挿入／欠失が発生する。特に欠失した場所と内容の情報は配列に残らないため、必要に応じてアミノ酸配列中の欠失部分と思しき場所に適切な長さの空記号を補うことも考えなければならない。アラインメント作成時には、それら全ての要素を考慮に入れた上で、類似度を最大化したい。

ではどのようにアラインメント間の類似度を評価するのか。まず、置換に関する評価を定める。これはアミノ酸どうしの近さを見れば良い。実際、評価のためのスコアテーブルが作成されている。すなわち、仮想的に設定した2つのアミノ酸AとBの近さはスコアテーブルのAの列とBの行の交わる部分のスコアで与えられる。このスコアはアミノ酸の構造と対応しており、例えば小型親水のアミノ酸同士や疎水のアミノ酸同士は置換しやすいし、疎水のアミノ酸から親水へは置換しにくい。欠失に対する評価には、経験的にアフィン・ギャップ・ペナルティが採用されている。すなわち、長さLの欠失のスコア$g(L)$は$g(L) = \alpha + \beta(L-1)$ で定める (α, βは定数)。挿入はもう一方の配列での欠失とみなせば、これらで全ての変異を考慮したことになる。

このスコアリングのもとに、全ての可能なアラインメントを発生させてスコア最大のものを見つけたい。しかし、配列が長くなれば、可能なアラインメントの個数は莫大になる。実際、初等的な組み合わせ論とStirlingの公式から、可能なアラインメントの総数はアミノ酸配列の長さに対する指数のオーダーで増大することが分かる。従って全てを数え上

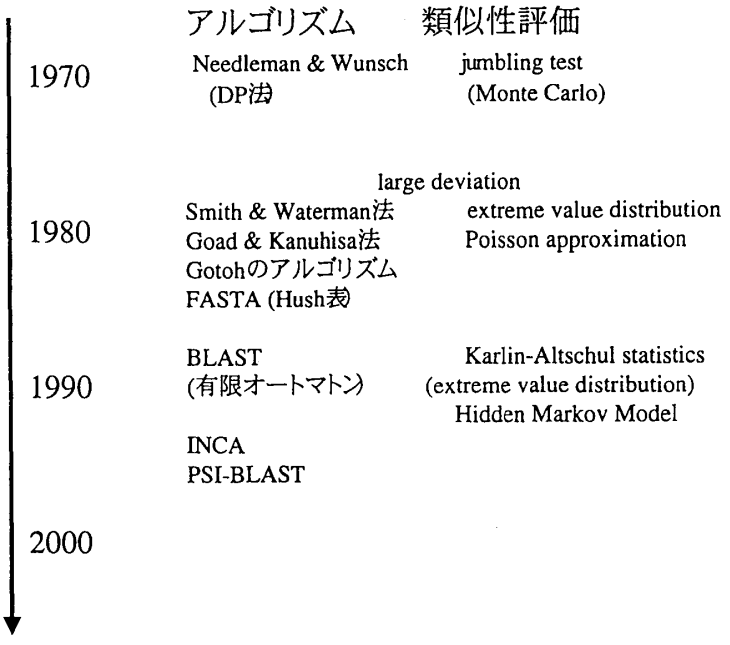

図 2: データベース検索技術の変遷

げることは困難である。この問題を回避するため、分子生物学では動的計画法 (Dynamic Programming Algorithm) [3] をベースにした方法が用いられている。

この方法では、まず2次元空間の縦と横に、例えば左上が各々配列の開始点になるように2つの配列を並べる。その上で左上から右下に流れる2次元空間内でのパスを定める問題に最適アラインメント決定の問題を帰着する。

こうして作成したペア毎のアラインメントに対し、有意な類似性の認められるものを相同配列とみなす。次のステップでそれらのマルチプル・アラインメントを作成する。

データベース検索に要求されることとして、ひとつは処理速度が挙げられる。実際にデータベース検索を行う場合には莫大な量のデータから類似配列を見つけなければならない。そのための高速アルゴリズムの開発が必要とされている。

また、比較解析による構造／機能予測の観点からは、検出感度の向上が求められる。類似性の高い配列に構造／機能が未知のものしかない場合には、配列の類似度が低いものまで範疇に入れて構造／機能が既知のものを探す必要が生じる。しかし、類似度が低い場合には誤って似ているとされている可能性が生じる。それを避けるためには、相同配列か否かを評価するための検出感度に充分な質が必要になる。類似性評価は、基本的には離散確率過程の理論をもとにして作られている。

また、類似性評価の手法は実質的に統計検定と同じことを行っている、ということが指摘されている。以下、統計検定の言葉で類似性評価を見てみよう。

[3] Needleman and Wunsch (1970)

まず帰無仮説を「2つの配列がアミノ酸組成に従ってランダムにマッチングしている」と設定する。randomnessについては、前に与えたスコアのつけ方に応じてtest statisticを決定する。その上で、マッチングの分布をextreme value distributionなどで求める。そこで実際に構成されたアラインメントに対し、帰無仮説が棄却された場合に、2つの配列が相同だと定める。すなわち、実際のアラインメントが充分に「ランダムに発生したとは言いがたい」場合に、統計的に有意とするわけである。

Step 2: マルチプル・アラインメント　次に、データベース検索で見つけてきた相同配列に対して、多数個でアラインメントを同時に作成する必要がある。ここでは、マルチプル・アラインメント作成の方法を幾つか紹介する。

まず、多次元の動的計画法を紹介しよう。ペア毎の場合には、2つの配列を各々座標軸に沿って並べた2次元の空間を考える。その上で漸化式を計算して2次元空間内のパスを与えることで、アラインメントを決定した。このアイデアをN個の配列で行おうとすると、N次元の空間のパスを定める問題になる。しかし、単純にそのままでは計算機のメモリ的にも計算時間的にも実用的とはいえない。改善のため幾つかの研究がなされているが、いまだ実用に耐え得るレベルには至っていない。

現在主に使われているのは、次に示すプログレッシブ・アラインメント法である。これにも幾つかの種類があるが、ここではtree-based法を紹介する。

まず、比較しようとする配列をペア毎に比較して系統樹を作成する。これをguide treeという。guide treeに沿って、配列の近い順にアラインメントを(ペア毎に)作成する。そこで新たに作られたアラインメントを1つの配列とみなして動的計画法が適用できる[4]。そこで得られたアラインメントを用いて、guide treeで次に近い部分とペア毎に(配列、またはアラインメントの族を)比較し、次のアラインメントを順次作成してゆく。

この方法だと、各ステップでは最適なアラインメントを作成している。しかし全体で最適なマルチプル・アラインメントを作成する保証はない。guide treeを用いることで、最適ではなくともそれに近いアラインメントが得られると期待した方法と言える。

この手法の欠点は、途中に作成したアラインメントの段階で間違った欠失が入ると最後まで修正が効かないことである。その改善のための手法が幾つか提案されているが、繰り返し計算が必要なものが多く、またプログレッシブ・アラインメントほど一般的ではない。

Step 3: マルチプル・アラインメントからの構造／機能情報の収集　Step2で得られたマルチプル・アラインメントを用いて、どのように構造／機能を収集するのか。これには幾つかの方法が用いられているが、進化の情報を用いるとどんなことが分かるか明らかにするのが、本講演の要である。具体的には、分子系統解析による方法とevolutionary traceの方法を紹介する。

[4] 2つの配列間の動的計画法を改良して、配列(またはアラインメント)の族2つに対し、2次元空間内のパスを決める方法で2つの族の間のアラインメントが作成できる。

分子系統解析による機能情報の収集

分子系統樹をクラスター分析の観点から見ると、ドラッグデザインなどに応用できる。ここでは、「エンドセリン受容体アンタゴニストの設計」という具体的な問題でどのように分子系統樹が利用されるのか、例を通じて見てみよう[5]。

エンドセリンはペプチド性ホルモンの一種である。強力な血管収縮活性を示すため、医薬の分野で注目されている。エンドセリンの細胞への作用は、前記のように受容体を通じてなされる。このエンドセリン受容体はGPCRと呼ばれるタンパク質のファミリーに入っている。現在の薬品の30-40%はGPCRを作用対象としており、ゲノム創薬の立場からは非常に重要な族である。

この研究ではまずエンドセリン自体に人工的に突然変異を起こして、どこが結合の際に活性部位になるのかを調べた。結果、置換すると結合活性の極端に落ちる部位を特定した。そこで結合部位に良く類似した分子IRL1722を合成し、エンドセリン受容体に作用させたのであるが、充分な活性は得られなかった。

そこで、エンドセリン受容体に着目し、その系統樹を作成した。結果、サブスタンスP受容体がエンドセリン受容体の比較的近縁になることを発見した。このサブスタンスP受容体には、結合しその働きを阻害するアンタゴニストが多く知られている。系統樹で近縁であることから、これらアンタゴニストでエンドセリン受容体に結合するものがあると推測される。そのアイデアに沿って、エンドセリン受容体に結合するものを3種類発見した。その中で特に阻害活性の強かったCGP49941に着目し、これとIRL1722をもとにしてエンドセリン受容体に非常に強く結合するものを作ることができた。

このように、系統樹をクラスター分析的に用いることにより分子進化の問題以外にも応用することができる。この考え方をより系統的に進めたものとして、evolutionary traceがある。この手法は、モチーフ(の拡張されたもの)及び立体構造を利用することに特徴がある。

モチーフとevolutionary trace

まず、配列解析で用いられるモチーフの概念を、レトロウイルス・プロテアーゼの解析[6]を通じて紹介しよう。レトロウイルスは逆転写現象の原因と考えられているウイルスで、有名なものでは、エイズウイルスはレトロウイルスの仲間である。

レトロウイルスの逆転写酵素の配列中にプロテアーゼドメインがある。このプロテアーゼの働くメカニズムに興味があった。レトロウイルスがライフサイクルを進める上で重要な働きをしていることが分かっていたためである。

はじめに、プロテアーゼのドメイン配列を抜き出してデータベース検索を行った。類似した既存のプロテアーゼの存在を期待したのだが、レトロウイルス以外では見つけられなかった。

そこで、とりあえず検索にかかったレトロウイルスのプロテアーゼに対してマルチプル・アラインメントを作成した。そうすると、マルチプル・アラインメントのある部分で、

[5] Th Fruh et al. *Bioorganic & Medical Chemistry lett.* **6**, 2323 (1996).
[6] Toh et al. *EM B O J.* **4**, 1267 (1985), Toh et al. *Nature* **315**, 691 (1985).

全体的に強く一致する箇所 (モチーフという) があると分かった。このモチーフ部分について、各サイトで出現しやすいアミノ酸を調べて、コンセンサス配列を構築した。

次に、この部分に着目し、これと同じ配列を持つものを再びデータベース検索で探した。そうすると、コンセンサス配列は酸性プロテアーゼ (ペプシン・レニンなど) の活性中心の部分に類似していることが分かった。そのことから、レトロウイルスのプロテアーゼは酸性プロテアーゼだと予測された。

この予測を受けて、既知酸性プロテアーゼの立体構造をもとにしたHIVプロテアーゼのホモロジー・モデリング[7]、酸性プロテアーゼ阻害剤 (ペプスタチン) によるHIVプロテアーゼの阻害[8] などの研究がなされた。また、X線結晶構造解析により、HIVプロテアーゼと酸性プロテアーゼが同じ機能を示す立体構造を持っていることが実際に確認できた[9]。

このように、アラインメントから、機能的／構造的に重要な (変化しにくい) 箇所を発見できる。つまり、マルチプル・アラインメントの各配列中で保存している部分 (モチーフ) がそれらの箇所に対応している。

これは配列中には複数箇所に離れて現れうるが、配列をもとにしたタンパク質が立体構造を取ったときに、対応する部分は近くに集まると考えられる。集まる場所の候補としては、酵素の活性中心あるいは他のタンパク質と複合的に働く場合のインターフェース部分が挙げられる。また、他のケースとして、疎水コアがある。体内は水に近い環境にあるため、疎水性のアミノ酸はタンパク質が立体構造を取るとき親水性のアミノ酸に囲まれて中に押し込められる。その結果、疎水部分は非常に緊密に配置され、変異が発生すると同様の構造の維持が困難になる。このようなタンパク質が立体構造をとった時に、その内部に形成される疎水アミノ酸の集合した部分を疎水コアと呼ぶ。

それではevolutionary traceの説明に入る。

まず、端点までの長さが全ての配列で等しい有根系統樹があるとしよう。時間軸に対し垂直に線を引き、系統樹との交点を定める。そして、各交点に対しそこを起源にもつ配列を1つのグループとして、配列を幾つかのサブファミリーに分ける。このサブファミリー内の配列はお互いに「近い」ものが集まっていると言える。ここで各サブファミリー毎に、その中で保存しているアラインメント中のサイトを検出する。得られた各サブファミリーで保存されたサイトのうち、全てのサブファミリーで保存されているか、あるいはサブファミリー間では異なったアミノ酸に変化しているものを選択する。こうして得たものをtrace残基という。さらにここから立体構造の中でtrace残基に対応する部分を色分けして[10]、機能推定を行う。

このようにして取り出されたtrace残基のうち、各サブファミリーに特異的な保存が見られるものは、サブファミリーの間での機能の違いを反映していると考えられる。例えば同

[7] L.H. Pearl and W.R. Taylor, *Nature* **329**, 351 (1987).

[8] R.F. Nutt et al. *Proc. Natl. Acad. Sci. USA* **85**, 7129 (1988), P.L. Darke et al. *J. Biol . Chem.* **264**, 2307 (1989).

[9] M. Miller et al. *Science* **246**, 1149 (1989).

[10] この時の考察対象は相同タンパク質のファミリーであるから、全て同様の立体構造を持つと想定できる

じ触媒として機能するにしても基質になるものが違う場合がある。また、trace残基のうち全体で保存しているもの(モチーフに同じ)は、全てに共通の重要な機能を持った箇所と考えられる。得られた部分配列の比較により、そのような性質を調べることがevolutionary traceの考え方である。

ここで、最初に系統樹に与えた分割をどの位置で取るかについて少し言及しよう。系統樹内での機能の分岐時期が明らかに分かっているのであれば、その直後で分割を考えれば良い。しかしそうでない場合は、実際に幾つかの位置での分割を取り、そこでの結果を比較することになる。古い時期だとサブファミリーのサイズが大きくなり、見つかるtrace残基が少なくなる。一方、新しい時期での分割の方が種の分化が進んでおり、従ってそこから定まるサブファミリーのサイズは小さくなる。そのときtrace残基は増えるが、目的と関係ないノイズが混入してくる。

例えば、分割を時間的に新しくしていった場合、trace残基に対応する部分が立体構造のある面に集中してきたとする。さらに、ある時期を越えると他の面にも現れてきたとしよう。そのときは、後者をノイズと見て、最初の特徴的な面に何らかのconstraintが掛っていると見る。例えば、その面が他のタンパク質と相互作用を起こすときのインターフェースになっていると考えるわけである。ちなみに、有根系統樹の分割を、根の上で行ったとき(つまり、実質分割がない時)の、trace残基がモチーフになっていると考えられる。

系統樹の時間発展のどの時期での分割が良いのかは現時点では理論的に定められてはいない。目測で適当と思しきところを定めている。この部分については改良の取り組みがなされている。

また、系統樹の作成にも幾つかの方法があり、一般には系統樹の末端までの枝の長さは不均一になる(進化速度の違い)。ここでは説明を簡単にするため特殊な系統樹での時間軸に垂直な線での分割を考えたが、系統樹の分割方法についても多少の考察が必要である。

上に挙げた例のように、立体構造の表面でクラスターを形成している場合、他のタンパク質との相互作用と関わる部分だと考えられる。従って、evolutionary traceの結果がタンパク質のあいだの相互作用を(網羅的ではなく個別に)調べる際の手助けになることが示唆される。

実験研究者が、実験によってタンパク質の機能や構造について調べる場合、アミノ酸配列に人工的に突然変異を発生させてその結果から機能や構造を調べる。一方、ここで紹介した計算機科学からのアプローチでは、相同配列の形成に伴う様々な突然変異を自然の行った実験とみなす。そこで相同配列の解析を行い、進化的情報を得ることによって機能や構造を調べるわけである。

3　タンパク質の生物学的機能解析

イントロダクションで見たように、生化学的機能はタンパク質それ自身の属性であるから、配列から直接解析することができた。ここでは少し手法を変え、ゲノムを用いて生物学的機能が配列から解析できることを見てみよう。

生物学的機能はの解析は、換言すれば相互作用のネットワークの予測と言える。すなわち、網羅的にどのタンパク質とどのタンパク質が相互作用しているか予測し、そこからネットワークを再構築することが課題になる。実際には多数の手法が提案されているがここでは以下の3つに絞って紹介する。

a) Conservation of Gene Neighborhood [11] 原核生物において、オペロン(1つのプロモーターによって支配される転写単位)を作るタンパク質同士の相互作用を考えよう。タンパク質間に相互作用がある場合には、遺伝子の順番、あるいはひとつのオペロン内に遺伝子のペアが共に(順番に関係なく)コードされているという現象が保存されている、という報告がある。これを逆に考えよう。まず、あるタンパク質を含むオペロンを各ゲノムから取り出す。そして、各オペロンの中での遺伝子の順番、あるいはどのような遺伝子が同時にコードされているかを調べる。複数個のゲノムについて、オペロン内部の遺伝子順序の保存、あるいは遺伝子のペアがひとつのオペロン内に同時にコードされているという現象がある一定以上の頻度で観察されるとき、それらの遺伝子がコードしているタンパク質は相互作用している可能性があると考える。これを繰り返してネットワークを構築するのが conservation of gene neighborhood 法である。

b) Phylogenetic Profile [12] この方法では、相互作用するタンパク質に対応する遺伝子はゲノムの中で在不在をともにする、と仮定する。N 個のゲノム(1から N までラベリングしておく)を用いて、遺伝子 A と遺伝子 B に対応するタンパク質の相互作用を考える場合を想定しよう。

まず、各々の遺伝子に対応する N 次元のベクトルを定める。具体的には、第 i 成分を、ゲノム i がその遺伝子をコードしていれば1、コードしていなければ0で定める。ここで定まった2つのベクトルに対してその類似度を評価し、「充分に」近い場合は相互作用があるとみなす。類似度の評価には、完全一致、1ビットの違いを許すなどの基準の他に、ユークリッド距離などが使われている。

c) Rosetta Stone [13] ある生物種のゲノムではそれぞれ別の遺伝子としてコードされている遺伝子が別の生物種では融合した1つの遺伝子として存在する状況を考えよう。このとき、もとの材料になった2つの遺伝子に対応するタンパク質は相互作用している場合が多い。この考えの逆を仮定とし、相互作用を推測するのが Rosetta Stone 法である。すなわち、ある遺伝子が別のゲノムにある別々の遺伝子2つの融合であった場合に、その2つの遺伝子に対応するタンパク質は相互作用してるとみなす。遺伝子重複が起きていれば機能が変化している場合があるため、オーソロガスな(遺伝子重複のない)遺伝子同士を比較した方が精度が上がることが知られている。

---

[11] Overbeek et al. *Proc. Natl. Acad. Sci. USA* **96**, 2896-2901 (1999), Bereend, S. et al. *Proc. Natl. Acad. Sci. USA* **99**, 5890-5895 (2002).

[12] Pellegrini,M. et al. *Proc. Natl. Acad. Sci. USA* **96**, 4295-4288 (1999).

[13] Enright, A.J. et al. *Nature* **402**, 86-90 (1999), Marcotte, et al. *Science* **285**, 571-753 (1999).

もちろん、これらの方法には各々欠点がある。

a) では、用いている仮定が相互作用を推測するには充分な適切性がないと言われている。相互作用しているものても、仮定を満たしていない場合がある。逆に、直接相互作用のないものが同じオペロンにコードされていることもある。これらの状況の発生頻度が無視できない程度に高いことが、この方法の精度が落ちる要因として指摘されている。

b) の方法だと、生物にとって極めて重要な遺伝子でベクトルを作ると全て1のベクトルになってしまう。従って、重要な働きを持つタンパク質同士の相互作用を推測する目的には向かない。

c) の場合、相互作用のあるタンパク質の遺伝子がいつでも融合するわけではない。そのため、融合した遺伝子が見つからない場合には相互作用の有無を判定できない。

また、これらの方法を全て組み合わせた方法で相互作用のネットワークが具体的に構成されている。これはまだ予測の段階ではあるが、ネットワークとしてどのような性質を持っているか調べる研究も最近活発になされているようである。例えば、「インターネットのウェブの構造と類似している」「完全にランダムに定めたネットワークと、一様につなげたネットワークの中間に属する」「ある種のフラクタル構造を持つ」などの結果が知られている。

4 まとめ

始めに見たように、バイオインフォマティクスの解析対象は、要素から相互作用に、個別から網羅へと変化してきている。しかし各々は相補的な関係にあり、両方の研究に意味がある。

また、タンパク質の機能解析では、生化学的機能と生物学的機能の2つに機能を分けて考えることが必要だと述べた。前者は個別の要素で定まる機能であり、後者の理解のためには、相互作用を網羅的に捉える必要がある。

本講演では、相同配列の比較あるいはゲノムの比較による構造／機能解析へのアプローチを紹介した。進化的に関係のあるものを対応づけることで、2つの手法から各々生化学的／生物学的機能に関する情報が得られることを示した。