多次元系統プロファイリングによる植物遺伝子機能の大規模推定

●佐藤 直樹
東京大学大学院総合文化研究科

＜研究の目的と進め方＞

多数のゲノムが解読されてきた今，多数の生物のゲノム情報を比較することによって，これまで個別のゲノムの解析ではわからなかった情報を得ることができるようになった。一方で，多数のゲノムにまたがる比較には計算機の大容量化と計算速度の高速化が必要であるが，昨今の計算機の進歩はこれを十分に満たしている。こうした状況を見越して，私は数年前からゲノムにコードされた全タンパク質を多数のゲノムにわたって比較する手法の開発と，それを利用した知識発見，および実験による実証を進めてきた。これまでの３年間，この比較ゲノム班において，Gclust ソフトウェアの開発とデータベースの公開を行うとともに，系統プロファイリングによって推定された遺伝子機能の実験的検証にもつとめてきた。今後の２年間では，これまでに開発した計算手法をさらに改良して，より多数の生物の比較に利用できるようにすること，光合成に関わるタンパク質の完全なセットの推定，さらに光合成だけでなくさまざまな植物群に特異的なタンパク質の機能発見を進める。研究は，今まで通り，計算と実験の２本立てで，両者のフィードバックを行いながら推進していく。特にこれからの２年間で実現したいこととしては，従来の系統プロファイリングからワンランクアップした新しい手法として，生物種，配列情報，生物形質情報の３者の間の教師なしプロファイリング（多次元系統プロファイリング）を開発し，事前の知識なしに，生物形質と配列情報の連関を浮かび上がらせる手法の開発にチャレンジしたい。本研究で提案している方法は，原理的にはどんな生物に関しても適用できる手法であるので，動物，微生物も含めた大きなデータベースを作り上げ，一般に利用できるようにもしたい。

＜2008 年度の研究の当初計画＞

計算機による研究１．クラスタリングと系統プロファイリングの手法の改良：より大規模なデータの処理を的確に処理できるような Gclust ソフトウェアの改良を進める。**２．プロファイリング対象とする生物種の拡大：**これまでの 95 種から 100 を超える多数の生物種のタンパク質についてのクラスタリングを行い，これに基づいて，系統プロファイリングを行う。結果は，ウェブを通じて公開する。**３．教師なしの多次元プロファイリング手法の開発：**（1）配列情報と組み合わせてプロファイリングを行うために，生物形質情報データを収集，整備する。主に光合成生物を中心としてパイロット的に手法を検討する。（2）生物種，配列情報，形質情報という３つの軸を組み合わせた相関解析による教師なし系統プロファイリングを行い，これまでに知られていない生物群特異的な遺伝子機能の発見を行う。（3）より多くの生物を含むデータの解析を行い，生物がもつ基本タンパク質セットを推定する。**４．クラスター情報に基づくバクテリアのシンテニー領域の推定：**近いパラログも含めたオルソログの近接関係を，多次元尺度法により推定し，バクテリアゲノムの構造的進化を解析する。

実験による解析：主として光合成生物を中心として，機能未知タンパク質の新規機能の解析を進める。（1）シアノバクテリアに由来する植物・藻類の葉緑体タンパク質のうち，機能がわからないものについて，シロイヌナズナのタグライン，コケのノックアウト，クラミドモナスの RNAi などを用いて解析する。（2）被子植物特異的あるいはその他系統特異的な酵素で遺伝子の同定がされていないものについて遺伝子を探し，機能を特定する。特に私がもともと専門としていた脂質代謝系の酵素などを中心として，これまで欠けていた遺伝子を見つけ出す。

＜2008 年度の成果＞

計算機による研究　１．クラスタリングと系統プロファイリングの手法の改良：これまで，Gclust というクラスタリングソフトウェア（version 3.5.5z4）と，系統プロファイル検索ウェブサーバー（下記）を開発してきた。まず，Gclust ソフトウェアに関しては，今後のデータサイズの大型化に対応するべく高速化を図るため，並列化を行った．このソフトウェアの中核となるタンパク質グループごとのホモログ探索は，８個ないし 128 個の cpu を用いた並列化により，cpu 数にほぼ比例した高速化ができ，全所要時間も半減した．系統プロファイル検索サーバーに関しては，植物の相同遺伝子プロモータ比較サーバーを名大の山本氏らとともに開発した．

２．プロファイリング対象とする生物種の拡大：シアノバクテリアのカバー数を 25 種から 34 種に増やし，アノテーション情報も付加した CyanoClust サーバーを構築し，公開した。クラスタリングのための大きな計算に用いている東京大学医科研のスーパーコンピュータが，秋から機種入れ替えのため利用できなかったため，95 種データセットに関する生物種の拡大はこれからになる。

３．教師なしの多次元プロファイリング手法の開発：これについては準備中である。

４．クラスター情報に基づくバクテリアのシンテニー領域の推定：近縁種の間では，ゲノム上におけるオルソログ遺伝子の存在位置が大まかに保存されており，保存関係をもとにゲノムのコア領域をいくつかに区分できることを，多次元尺度法を用いた解析から明らかにした（投稿中）。これに基づき，シアノバクテリアゲノムを７個の仮想リンケージグループ（VLG）に分類すると，これらが複製開始点との間で一定の距離関係を保っていること，VLG と遺伝子機能との大まかな関係が見られることなどを明らかにした。ゲノム進化において組み換え等により遺伝子の存在位置が変わる場合，保存遺伝子はでたらめに移動するのではなく，複製開始点や他の VLG との一定の距離関係を保ちながら移動すると考えられる。

実験による解析：（1）シアノバクテリアと植物・藻類に共通に存在する約 50 個のタンパク質に関して，シアノバクテリアと植物の両方における機能解析により，これらが確かに植物の葉緑体に局在し，光合成に関連する機能を持つ chloroplast proteins of endosymbiont origin（CPRE）であることを証明した。ちなみに葉緑体局在というのは，PSORT や TargetP などのソフトウェアで推定可能でないものもまだあり，我々の研究室で発見した PEND タンパク質もその一つである。ターゲット配列予測に頼らない葉緑体タンパク質の推定は，依然として意味がある。さて，シアノバクテリアと植物の遺伝子破壊株について，光合成関連の測定を行った結果，調べた 37 の相同グループのうちで，これま

でまだ一方でしか解析ができていない 10 グループ（点線上に記載）を別にして，両者ともに何らかの影響が見られたものは 15 グループにのぼった（図１）。ただし必ずしも表現型は同じではなかった（図２）。これは，本研究のねらいである，系統的に保存されている遺伝子がシアノバクテリアと植物の両者で光合成に重要な役割を果たしていることを実証できたと同時に，厳密な機能は両者で異なる可能性を示唆するものであり，細胞内共生による植物の進化におけるシアノバクテリア由来遺伝子の意義について，新しい評価ができた。なお，どちらの生物でも表現型が観察できなかった４つの CPRE の中には，すでに機能が別のグループによって報告されているものもあり，それによれば，強光など特別な条件で表現型が表れるという。今回の実験はあくまでも標準的な培養条件でのもので，ストレス条件などで調べればさらに多くの遺伝子破壊株で表現型が見られる可能性がある。

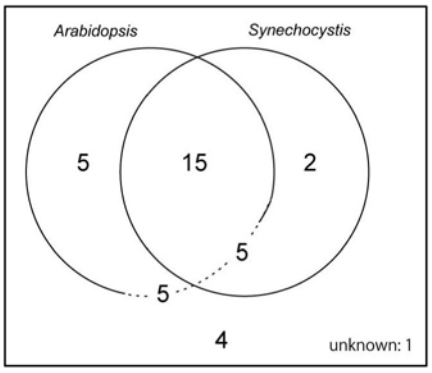

図１：シロイヌナズナとシアノバクテリアの対応する CPRE の破壊株において表現型が見られたものの関係

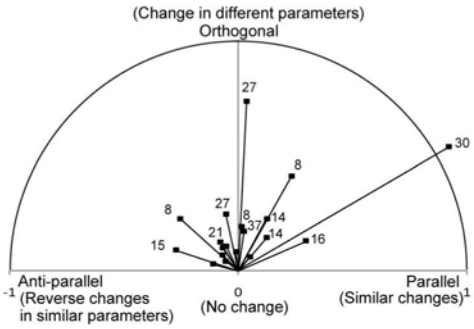

図２：シロイヌナズナとシアノバクテリアの両方で破壊株が得らた CPRE についての PAM 蛍光測定の結果の相関。PAM で得られる６種類のパラメータをベクトルとして，両生物での結果のベクトルのなす角度と絶対値の相乗平均を新たにベクトルとしてプロットした。両ベクトルがそろっていると右向きになり，異なっていると直交成分が出てくる。

（2）根粒菌に共通して存在する機能未知タンパク質群の推定も行い，比較ゲノム領域の研究の柱の一つである共生と生物間相互作用の研究にも多少とも貢献するよう努めた。このほか，植物脂質の生合成に関わるタンパク質遺伝子について，比較ゲノム情報を整理し，生物種をまたいだ実験解析を可能にした。実際にいくつかの遺伝子について，機能解析を進めているが，まだ報告する段階ではない。また，植物では，葉緑体の分裂にペプチドグリカン関連遺伝子が働いていることが示されているが，ペプチドグリカン合成系について，シアノバクテリアと植物における対応関係をまとめた。これをもとに新たな植物遺伝子の機能解析を始めた。

＜国内外での成果の位置づけ＞

本研究に関連して，系統プロファイルを利用して遺伝子機能を発見しようとする方法論が一昨年あたりから出てきていることがあげられる。我々はこの可能性を 2002 年の段階で予見し，それに向けたソフトウェアの開発と実験の方法論を構築してきた。実験を含むため時間はかかったが，世界的に受け入れる素地が出てきたことは，本研究をアピールしていく上で好ましいと考えている。

＜達成できなかったこと、予想外の困難、その理由＞

教師なし多次元プロファイリング手法の開発については，もとになる生物情報データの整理が遅れているが，その理由の一つは，あまり生物として研究されていない生物のゲノムが多くなってきていることである。他の生物でもそうであるが，これまで実験によく用いられている材料と，ゲノムが決められる材料とが，近縁であっても異なる種である場合が多いので，系統プロファイルの取り扱いの上では，こうした場合にも，両者のデータをうまく利用できる工夫が必要である。つまり，ただ単にゲノムデータと生物情報データをリンクするのではなく，生物種で作る空間にこれらの情報をマップするというような全く新しい考え方が必要ではないかと思われる。これには系統樹をうまく利用する工夫が必要となる。

＜今後の課題＞

今後の研究では，多次元空間へのマッピングを利用した新しい多次元系統プロファイリングの手法の開発と実際のテストデータベースの作成が一つの課題である。また，植物とシアノバクテリアの対応する遺伝子の平行した解析の有効性が実証できたので，これをさらに推定できるすべての遺伝子に拡張して，解析を進めたい。これについては，人数と費用がかかるので，他機関との共同研究などを検討する。光合成関係以外の遺伝子の機能推定についても，実験を進めていく。

＜成果公表リスト＞

1）論文／プロシーディング（査読付きのものに限る）

0707231610: Genomic and biochemical analysis of lipid biosynthesis in the unicellular rhodophyte *Cyanidioschyzon merolae*: lack of plastidic desaturation pathway results in mixed pathway of galactolipid synthesis. *Eukaryotic Cell* 6, 1006-1017

0901152244: The assembly of the FtsZ ring at the mid-chloroplast division site depends on a balance between the activities of AtMinE1 and ARC11/AtMinD1 *Plant Cell Physiol* 49:345-361

Gclust ソフトウェアの論文など関連論文は投稿中または印刷中。

2）データベース／ソフトウェア

公開データベース：

0606210932: Gclust サーバー：http://gclust.c.u-tokyo.ac.jp/

0707251539（名古屋大学小保方研究室で登録）：PPDB: http://ppdb.gene.nagoya-u.ac.jp/cgi-bin/index.cgi

0901152201: CyanoClust サーバー：
http://cyanoclust.c.u-tokyo.ac.jp/

ホームページ：http://nsato4.c.u-tokyo.ac.jp/

共同研究：共生窒素固定細菌に共通するタンパク質の推定について，青木誠一郎博士（東大総合文化）と協力して進めている。また，Gclust で得られるクラスタに含まれるタンパク質の遺伝子のうちでシロイヌナズナ，イネ，ヒメツリガネゴケのプロモータの比較ができるように，山本義治博士（名大）の PPDB と連携し，Homolog gene search を作った（上の URL）。