相同グループ法による系統プロファイリングを用いた植物遺伝子機能の大規模推定

●佐藤 直樹　◆藤原 誠
東京大学大学院総合文化研究科

＜研究の目的と進め方＞

ポストゲノム研究の重要な課題の一つは，ゲノム上に推定された約半数にも及ぶ機能未知タンパク質の機能確定である。既知タンパク質や既知ドメインとの類似性による従来からの解析に加え，ゲノムならではの解析として系統プロファイリングがある。これは，類似形質をもつ生物群に共通に存在する遺伝子は，その共通形質に関連している可能性があることに基づく方法論である。この問題は一般的なものであるが，最適なターゲットとして，大量水平移動によって遺伝子が導入された場合，すなわちシアノバクテリアの細胞内共生による葉緑体の誕生及びそれを持つ真核光合成生物の創成を考えた。細胞内共生説は一般に確実と考えられているが，実際にはどんなシアノバクテリアが共生体となったのかなど，肝心な点はよくわかっておらず，また，既知の光合成明反応関連タンパク質遺伝子の他にどれだけの遺伝子が共生体から真核宿主に導入されたのかもわかっていない。

平成 18 年度の本特定領域研究では，相同グループの形成と系統プロファイリングのためのソフトウェアとそれを用いて作成された Gclust データベースを発表した。さらにこれを用いて，植物，紅藻，珪藻，緑藻，シアノバクテリア 10 種の他，多種の非光合成生物も含む最新ゲノムデータから，光合成生物に共通なタンパク質を推定した。

本研究では，相同グループ法による系統プロファイリングの方法論の有効性を確立して，データを公開するとともに，これによって系統特異的に保存されているタンパク質群を推定し，生物群特異的機能に関与するタンパク質群を一括して同定することを目的とする。特に，計算と実験を融合した研究を行うことを特色とする。最終的には植物以外にも拡張して解析を進める。

＜ 2007 年度の研究の当初計画＞

(1) Gclust ソフトウェアの改良：昨年度のバージョン 3.5.3 で使われているアルゴリズムは，タンパク質グループごとに異なる閾値を自動設定して，生物学的に最も自然な相同グループを形成するやり方をとっているが，依然としてパラメータ設定に依存する部分が多く，十分とはいえない。根本的に異なるアルゴリズムを現在検討中で，これの実装により，高機能化，高速化をはかる。(2) Gclust サーバーの利便性の改良：機能が明らかなタンパク質を与えてそれと類似のプロファイルを持つタンパク質を検索する機能を追加する。また，生物群を事前に特定せずに，保存されているタンパク質を発見できるアルゴリズムを検討する。これを使って，今まで隠れていた大規模水平移動の発見を目指す。(3) 系統プロファイリングの適用によって推定される新規機能遺伝子の機能アノテーションとそれを用いた Gclust クラスタと GO (Gene Ontology) とのリンクの構築　(4) データセットの拡張：植物では，イネ，ヒメツリガネゴケなど，また，これまでほとんど入れていなかった菌類を 10 種類程度追加する。さらに，動物ゲノムに関しても，順次追加していく。その上で，植物・菌類・動物のさまざまなグループ特異的なタンパク質群を推定する。(5) 相同グループを利用した微生物ゲノムの距離構造の解析：これまでの遺伝子の並び方を利用したゲノム比較を拡張するため，Gclust 相同グループを単位としたタンパク質遺伝子の距離関係に基づく新たなゲノム比較の手法を開発する。

＜ 2007 年度の成果＞

(1) Gclust ソフトウェアの改良： 当初計画した根本的改変には至らなかったが，下記 1-1~1-3 のような改良を加え，大量データにも耐えるバージョン 3.5.5w を作った。また 1-4 のような検討を行った。1-1. バージョン 3.5.3 では，ゲノムの数が多くなると，クラスタが分割してしまう場合や類似クラスタが大きなクラスタを形成することがあった。類似したタンパク質グループを合併するか分離するかの判断を，Jensen-Shannon distance measure を用いて処理することその他により，これらの問題を改善した。1-2. Gclust に入力する総当たり BLASTP の結果を取り込むファイル形式として，blastall の -m 8 オプションで出力される表も利用できるように改変した。これは blastall の通常の出力形式が，バージョンによりたびたび変更されることに対応したものである。1-3. その他，細かいパラメータの調整や評価関数の変更を行うことで，実際の大型データの処理が適切になるよう，微調整を行った。1-4. Gclust データベースの特徴を詳しく考察するため，NCBI で公開されている COG(Cluster of Orthologous Groups of proteins) との比較を行った。タンパク質ファミリーを与えるという点では，両者とも類似の性質をもつと考えられたため，Gcluust と COG がどの程度似ているのか検討した。Gclust クラスタは COG のサブクラスタとなっていることが多いことがわかった。このほかの Gclust の大きな特徴として，真核原核にまたがったデータベースを作ることができることがあげられる。

(2) Gclust サーバーの利便性の改良：2-1. 生物群に関する事前情報なしに保存タンパク質グループを発見する方法として，系統プロファイルのクラスタリングを行った。これについては，今後さらに詳しく検討する。2-2. Gclust により推定されるタンパク質クラスターごとに，シロイヌナズナ，イネ，ヒメツリガネゴケの遺伝子上流配列に存在する共通モチーフを検索するサーバーの開発を，名古屋大学小保方・山本氏との共同研究により開始した。

(3) 系統プロファイリングの適用によって推定される新規機能遺伝子の機能アノテーション：3-1. 光合成膜の主要脂質であるジガラクトシルジアシルグリセロール (DGDG) を合成する酵素は，緑色植物では知られていたが，そのホモログが紅藻やシアノバクテリアには存在しないため，紅藻やシアノバクテリアにおける DGDG 生合成酵素は未知のままだった。Gclust を用いて，「紅藻とシアノバクテリアにあって緑色植物など他の生物にはない相同グループ」を選択し，さらに，糖転移酵素モチーフを持つものを探すと 2 個に絞られた。このうちの一つは紅藻の葉緑体ゲノムにコードされていたので，これに着目した。*Synechocystis* におけるホモログである slr1508 を破壊したところ，DGDG を含まない変異体が得られた。このことから，通常の条件では DGDG が必須ではないこと，しかし，光化学系の構築に異常があって光合成活性が低いことなどがわかった。3-2. 光合成生物特異的に保存されているクラスターとして 170 個を推定した。3-3. 共生根粒菌に関して，Nod30b データセットを作成し（東大総合文化青木氏との共同研究），4 種の根粒菌に保存されたクラスタとして 175 個を推定した。3-4. 同様の手法により，これまで未知であったいくつかの遺伝子の候補を絞り込み，現在，実験的検証を進めている。

(4) データセットの拡張 当初予定したとおり，植物を中心にゲノムの追加と更新を行い，「ALL95 データセット」を構築，公開した。これには，真核光合成生物として，被子植物 3 種（シロイ

ヌナズナ，ポプラ，イネ），コケ植物１種（ヒメツリガネゴケ），緑藻２種（*Chlamydomonas, Ostreococcus*），珪藻２種（*Thalassiosira, Phaeodactylum*），紅藻１種（*Cyanidioschyzon*）が含まれ，シアノバクテリア 25 種，光合成細菌 15 種と合わせ，光合成生物の間でのゲノム比較の性能が飛躍的に向上した。この他，光合成をしない細菌 31 種，光合成をしない Bikont 系統の単細胞真核生物９種，Opistokont 系統の真核生物（菌類と動物）９種が含まれ，光合成をするものとしないものの比較も強化された。さらに Bikont に属する非光合成生物である Naegleria，テトラヒメナ，ゾウリムシ，卵菌類などの詳細な系統解析のツールとしても機能することがわかった。さらに，この他に，ヒト，霊長類，マウス，４種の魚類などの脊椎動物を含む 19 種の動物を中心とした「NP28 データセット」も構築し，公開した。これは，動物の比較ゲノム解析をする研究班員との交流により実現したものである。

(5) 相同グループを利用した微生物ゲノムの距離構造の解析：Gclust ソフトウェアを用いて最新 25 種のシアノバクテリアからなるデータセット Cyano25 を構築した。このうち *Nostoc punctiforme* については，完全なゲノム配列が出ていないので，残る 24 種のデータを利用して，相同なタンパク質のゲノム上での存在位置を比較する新たな手法を開発した。まず，個別のタンパク質遺伝子ではなく，それらが所属する相同クラスタを単位とし，24 種のゲノム上での並び方を，metric multidimensional scaling により正規化する。これを元に主成分分析を行い，生物種を超えた linkage のあるクラスタを分類する。この分類を用いて各ゲノム上の遺伝子をマークするというものである。こうして得られる種を超えた linkage（global genomic linkage）は，従来の synteny 概念ではとらえられない緩い近接関係を発見するのに効果的であった。

(6) その他：色素体の起源となったシアノバクテリアの探索
シアノバクテリアは大きく２つのグループに分かれ，系統１は *Prochlorococcus* や *Synechococcus* を含み，系統２は，*Anabaena* などを含んでいた。色素体は単系統で系統２から分岐することがわかった（著書 0612221115）が，これをさらに裏付けるため，生物種の数をさらに増やしたデータを用いて計算を行っている。計算はまだ続いているが，以前の結果をほぼ裏付けるような結果が得られている。

＜国内外での成果の位置づけ＞

Gclust サーバーについては，植物関連や脂質関連の学会の他，シアノバクテリア，コケなどの個別の生物群の研究者の集まりの場で紹介し，次第に利用者が増えてきている。Gclust サーバーは汎用の遺伝子情報サーバーではないので，使用頻度が非常に高いとはいえないものの，研究室内からのアクセスを除いて，一日平均 80 件程度のアクセスがあり，その半数は外国からのものである。脂質研究では，論文 0707231610 や 0801251334 や学会発表を通じて，比較ゲノムの有効性が次第に評価されてきており，それがサーバーの使用率にも反映していると見られる。

紅藻とシアノバクテリアに共通した DGDG 合成酵素の発見は，他の研究者が別のアプローチで時間を掛けて行った研究とほぼ同時に出版され，我々のアプローチの有効性が証明された。

＜達成できなかったこと、予想外の困難、その理由＞

(2) Gclust サーバーの利便性の改良のうち，「機能が明らかなタンパク質を与えてそれと類似のプロファイルを持つタンパク質を検索する機能を追加する」という点は取りやめた。これは，ALL95, NP28 など 70 万配列以上のクラスタリングを適切に進めるためのソフトウェアの開発と微調整（上述）に時間がかかったためもあるが，このやり方自体をさらに大きな枠組みで考え直すことがよいと判断されたためである。

(3) 系統プロファイリングの適用によって推定される新規機能遺伝子の機能アノテーションとそれを用いた Gclust クラスタと GO とのリンクの構築の項目のうちで，GO との連携に関する部分は，現在残されている。実際に GO を取り込むためには，Gclust ソフトウェアによる処理の最初から GO データを入れる必要があり，適当なパーサーを開発することと，Gclust の仕様の変更が必要になる。また，どのような形で関連づけを行うのが最適であるのか，方針を決めかねているのも理由であった。幸い，（５）の genomic linkage 解析でも特に詳しく研究しているシアノバクテリアの解析から，特に linkage と機能の間の関連が考えられたので，linkage と関連させながら機能との対応関係を整理するのがよいと考え，近々実際の作業に着手することにしている。

＜今後の課題＞

相同タンパク質のプロファイルから系統特異的なタンパク質グループを発見する手法の開発が，一つのテーマである。また，生物のさまざまな情報とタンパク質クラスタをあわせてクラスタリングするという新たな手法を来年度に計画している。また，すでに推定できている機能遺伝子候補についての実験的検証を強化することで，本手法の有効性を実証していく。

＜成果公表リスト＞

1) 論文／プロシーディング

0612221038: Terasawa, K., et al., The mitochondrial genome of the moss *Physcomitrella patens* sheds new light on mitochondrial evolution in land plants, *Mol. Biol. Evol.*, 24, 699-709 (2007).

0707231610: Sato, N. and Moriyama, T. 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 (2007).

0801251157: Nozaki, H. et al., A 100%-complete sequence reveals unusually simple genomic features in the hot spring red alga *Cyanidioschyzon merolae*. *BMC Biology* 5, 28 (2007).

0801251334: Sakurai, I. et al., Digalactosyldiacylglycerol is required for stabilization of the oxygen-evolving complex in photosystem II. *Plant Physiol.*, 145, 1361-1370 (2007).

2) データベース／ソフトウェア

0606210932: Gclust データベース公開サーバー http://gclust.c.u-tokyo.ac.jp/　このエントリーは 2006 年度野茂のであるが，サーバーを 2007 年度に２度更新し，新規データセットを公開した。

3) 著書

0612221115: Sato, N., Origin and Evolution of Plastids: Genomic View on the Unification and Diversity of Plastids, The Structure and Function of Plastids, pp75-102, Springer, 2006.

0707231629: 東京大学光合成教育研究会　光合成の科学 (2007).