Gclust について

東京大学大学院総合文化研究科広域科学専攻生命環境科学系
佐藤直樹
E-mail: naokisat@bio.c.u-tokyo.ac.jp
2006 年 4 月 25 日

Gclust データベースと Gclust ソフトウェアについて

ゲノム情報が氾濫している現在，研究対象としている遺伝子のホモログがどの生物にあるのか, BLAST でサーチしてもなかなかわからない, ということがあります。主な理由は，ウェブサイトのデータベースにはヒトやマウスあるいはイネなどの大量の EST 情報が載っていて，大部分が実は同一の遺伝子のものだったりするのですが，サーチをしたときに上位にずっと並んでしまい, 肝心の見つけたいものが下位に追いやられて見えなくなったり，埋没してしまいどこにあるか探せない，というようなことでしょう。それぞれのゲノムにあるそれぞれのホモログを簡単に探し出すことができれば，実験研究者にとっては非常に便利です。Gclust というのはこうした目的のために開発されたソフトウェアとデータベースです。G は genome のつもりです。clust は clustering の略で，ゲノムにコードされた全タンパク質の中で似ているもの同士を集めることを指しています。

さて，Gclust データベースの作成では，予め選択したゲノムにコードされる全推定タンパク質配列の総当たり BLASTP を行い，その結果を Gclust ソフトウェアによって処理することで，類似タンパク質クラスターを生成します。この際に他のデータベースでよく使われるやり方は，双方向ベストヒットを選択することによって「オーソログ」を決めるというものですが，本来 ortholog という遺伝学用語はきちんとした系統解析の結果，系統的に対応することが確認された場合にだけ用いる言葉であるという用語上の問題点と，類似タンパク質の個数は必ずしもゲノムあたり 1 個ではなく，必ずしも正しい類似タンパク質が双方向ベストヒットによって得られるとは限らないという実際上の問題点があります。Gclust は，こういう問題点を避ける（克服するとはいえないかもしれませんが）ために，類似タンパク質はすべて集めてしまいます。しかし，むやみに似たものを集めていくと，本来全然無関係のものまで一緒のクラスターに入るおそれがあります。このため，ここでは詳しく述べませんが，さまざまな基準を設けて，安心して類似タンパク質と考えられるものだけからなるクラスターを推定できるようにしました。それでもクラスターを作る際の自由度があるため，条件の設定により，大きめのクラスターと小さめのクラスターができます。Gclust データベースでは，同一のゲノムセットについて，二通りのクラスタリング結果を提供している場合もあるのは，このためです。使用者が便利と思う方を利用すれば結構です。

系統プロファイリング

Gclust データベースの特徴は，単に類似タンパク質クラスターを提供するばかりでなく，クラスタリングの結果に基づく系統プロファイリングを提供していることです。系統プロファイリングというのは，ある特定の系統の生物だけに保存されている遺伝子（タンパク質）を探し出すことですが，こうした遺伝子の中には，それらの特定の生物だけに共通した特定の機能（代謝経路，形態形成，細胞分化など）と関連しているものが含まれているはずです。こうした遺伝子の場合，他にホモログが無いわけですから，機能のアノテーションができず, hypothetical protein などと書かれているでしょう。アノテーションは元々他の生物で機能がわかっている場合にそれをコピーして作ることが多いので，このようなことになってしまいます。運がよければ，Pfam などの検索によって機能ドメインがわかることもあるかもしれませんが，Gclust で作られるクラスターの多くは，Pfam ドメインがついていないものです。このようにして，ある生物群特異的な遺伝子（タンパク質）が見つかると，それらの遺伝子を破壊するなり，遺伝子発現を調べるなりして，それらの生物群特有の機能に関連があるのかどうかを調べることができます。このように系統プロファイリングは機能ゲノム解析と組み合わせることによって，大きな力を発揮します。私たちの研究室では，シアノバクテリアの共生によって伝えられたと考えられる植物核ゲノムにコードされた葉緑体タンパク質を大量に発掘する研究をしています。おそらくこのようなタンパク質は 1,000 個ほどもあると推定されますが，それは Gclust によって初めて可能になりました。これまでも相同性によってシアノバクテリアと植物にある類似のタンパク質を探す試みはありましたが，あくまでも単発的な 1:1 の比較で，今回のような完全な物ではありませんでした。また，ターゲティング予測を利用した葉緑体タンパク質の推定はありますが，私たちのやり方では，ターゲティング予測を使わずに大量の葉緑体タンパク質を見つけだすことができました。もちろん共生体起源のものに限られますが。

このような系統プロファイリングの方法は，もっと広く応用できるはずです。今回公開する Gclust データベースでは，シアノバクテリアを中心とするいくつかのデータセットの他に，バクテリアのデータ，全生物のデータを加えています。後２者では，生物群ごとの選択だけになっていますが，基本的に希望の生物群に保存されたタンパク質のクラスターを選び出すことができます。その際，希望以外の生物群について，「存在しない」とするか「どっちでもよい」とするかの２通りが考えられます。それも選ぶことができます。多くの場合，特定の生物群に保存されていても，それ以外の生物にも散発的に存在することもあります。従って，あまり厳しく他の生物には存在しないという条件を課さない方がよい結果が得られるかもしれません。これは，目的とする生物群とその性質によって異なるでしょうから，それぞれ試してみることをお勧めします。一度やってみて思った通りの選択ができない場合，選択条件を見直してみることが大切です。

サーチの仕方

Gclust データベースのウェブサイトでは，目的とするクラスターを取得するためのさまざまな手段を提供しています。まず方針として，系統プロファイリングをねらうのか，特定のタンパク質についての相同タンパク質情報を得たいのか，目的によって分かれます。あらかじめ調べたいタンパク質があるときは，まずそのタンパク質配列を準備します。メニューで適当なデータセットを選び，それから，上にある BLAST というリンクによって BLAST サーチを行います。通常の BLAST の結果が出力されますが，そこには選んだデータセットに含まれるタンパク質だけが表示されています。通常，トップの項目を選びます。これははじめに自分で検索にかけたタンパク質そのものであることもあると思いますが，もともとデータセットに含まれない生物の遺伝子やタンパク質で検索した場合には，一番似たものが選ばれてきます。このタンパク質の名称をクリックすると，それを含むクラスターが表示されます。あらかじめそのタンパク質の名称や機能などがわかっている場合には，それらを利用してサーチをすることも可能です。ただし，データベースにちょうどある言葉でサーチすればよいのですが，微妙に違っていてうまく出てこないこともあります。そうした場合は，配列からサーチするのが最も簡単でしょう。ひとたびクラスターが表示されれば，あとは，その配列を取得するか，CLUSTALW によってアラインメントを作ることが可能です。あまり配列数が多いか配列が長いときには CLUSTALW のサービスは提供されず，代わりに配列セットをダウンロードして，自分でアラインメントを計算することになります。これは，サーバーの負荷を軽減するための措置ですので，ご理解ください。

系統プロファイリングからはいる場合，まず検索に最も適したデータセットを選択する必要があります。次に，生物種ごとに選択をするか，または，生物群ごとの選択にするか決めます。生物種ごとの選択はうまくやらないと大事なものを逃すおそれが多分にあります。生物群ごとの選択では，選択(yes)，非選択(no)，選択に無関係(any)という選び方ができます。Yes の場合，その生物群の過半数に保存されているタンパク質を選びます。No であっても過半数未満というだけで，ある程度の生物種には保存されていることもありますので，注意してください。全くゼロという選択はここには準備していません。実際上，前にも述べたように，本来なさそうなものが案外あちこちにあるものなので，絶対的にゼロという指定はあまり適切ではないでしょう。どうしてもある特定の生物に関して保存されていないことを指定したければ，生物種ごとの指定のパネルを利用してください。ただしこの場合，かなり煩雑な操作になります。このような指定をした上で，検索ボタンを押すと結果が表示されます。クラスターの大きさが大きい順に表示されており，通常，１番のクラスターはほとんどすべての生物で保存されている大きなクラスターです。上位１０くらいのクラスターは，多数のパラログを含み，系統プロファイリングという観点からはあまり利用価値はありません。

既知の特定のタンパク質のファミリーやスーパーファミリーの配列をまとめて取得して系統解析をしようという場合にも，Gclust データベースは便利です。実際，私自身も系統樹を作るときにはたいていこうしています。普通のウェブサイトで BLAST サーチをやっ

て配列を集めると,仕分けが大変ですし,EST 由来の部分配列も多く混じっています。Gclust データベースの場合，生物種が限定されますが，その代わり，データセットに含まれる生物種にあるほとんどすべてのホモログが一括して取得できます。場合によって，複数のクラスターに分割していることもありますが，その場合には，クラスタリングの度合いの異なるデータセットを調べてみるのも有効です。また，表示されたクラスターの下に Related と表示がある場合，クラスターとしては別になっているが，よく似た配列のクラスターが示されています。これらを手がかりとして，相同タンパク質を集めることができます。さらに別の問題として，マルチドメインタンパク質があります。ドメイン構造としては相同であっても，クラスターの他のメンバーには含まれないドメインを持つようなタンパク質は同じクラスターに入っていません。特に多数の複合したドメインからなるタンパク質はクラスタリングの妨げになるため，計算作業の際にあらかじめ別にしてあります。本来ならば，こうしたものはドメイン構造が異なるので，クラスターに含めるのはおかしいのですが，以下の２つの場合には何らかのやり方で取り上げる必要があります。第一は，類似の複雑な構造を持ったマルチドメインタンパク質がいくつもある場合。この場合には，マルチドメインタンパク質の個別のデータでドメイン表示を比較する必要があります。Gclust そのものの出力にはこれができるようになっているのですが，ウェブサイトにする際には，複雑になるため，この表示をしていません。第二は，計算上はマルチドメインになったが，実際はほんの少しの問題で,他のクラスターメンバーと大差ない場合です。この場合には，手作業で仲間はずれになっていたものを追加する必要があります。いずれにしても，配列をダウンロードした上で，自分で作業をすることになります。

自分でやってみようと思ったら

ウェブサイトには Gclust ソフトウェアと周辺で使うさまざまなスクリプト等をおいています。ダウンロードは自由ですが，私の著作者としての権利は放棄していませんので，ご注意ください。また，ソフトウェアは小さなものですが，自分でゲノムデータを処理する場合には，膨大なメモリが必要になります。どの程度の作業をするのかによって必要なシステムが異なりますので，ご相談ください。ちなみに ALL145 のデータは，東京大学医科学研究所のヒトゲノム解析センターのスパコンを使って２週間くらいかけて作りました。

文献（ソフトウェアと系統プロファイリング）

N. Sato, M, Ishikawa, M. Fujiwara and K. Sonoike (2005)
Mass identification of chloroplast proteins of endosymbiont origin by phylogenetic profiling based on organism-optimized homologous protein groups
Genome Informatics **16**: 56-68.