

遺伝子発現プロファイルデータベース 高速検索システム CellMontage

フジブリ フタル
藤 航

配列解析チーム 研究員



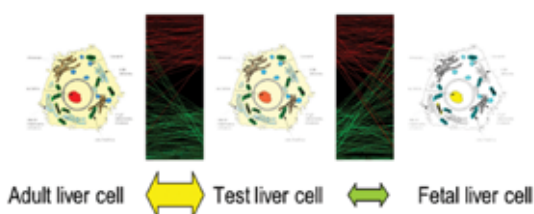
マイクロアレイなど簡便な方法の普及により、遺伝子発現プロファイルデータが年々倍増するようになった現在、遺伝子発現データベースから何らかの情報を引き出す方法論が必要とされています。我々は FASTA や BLAST が「類似している配列」を見つけるのと同じ様に、「類似している細胞」を見つけるシステム作りにこの一年取り組んできました。その結果、生まれたのが CellMontage と呼ばれる対応する遺伝子の発現相関を高速に検索するシステムです。このシステムは、(1) 順位相関係数を使うことで測定プラットフォームが違っていても検索ができ、頑強である、(2) 全遺伝子に着目するためマーカー遺伝子が不要、(3) 開発したアルゴリズムにより高速検索が可能、といった強い利点があります。これまではマイクロアレイなどの測定装置が各社で違うため、わずかな違いに弱く、プラットフォームを超えての情報の相互利用は難しいとされていましたが、この方法を用いるとプラットフォームに依らず、類似細胞が検索の上位に見いだされる事がわかってきました。さらに、マーカー遺伝子法のように全ての細胞種についてマーカー遺伝子を用意しなくても、遺伝子発現の全体像から類似細胞を特定できます。

現在、CellMontage は、米国国立衛生研究所 NCBI の Gene Expression Omnibus データベースとの連携によって、全部で約 3 万件のデータが納められ、ヒトだけで 1 万件以上のデータが検索できるようになっています。このシステムを用いると、検索者の持つ遺伝子発現プロファイルがどの既知の細胞と近いかという事が瞬時にわかるため、図 1 のように未知の細胞や細胞工学によって作られた組織が本当に正しいものかどうかの判定などができます。また、ortholog 遺伝子を使えば、種間比較—例えばヒトとマウスでどの程度細胞が類似しているかなどの測定—も可能になります。さらに将来、改良を重ねる事で、病気の診断もできるようになる可能性を持っています。また、何よりも見逃せないのが高速検索で、独自に開発したアルゴリズムにより、図 2 のように約 1 万遺伝子を使用した検索でも、一秒間に 1,000 枚以上もの遺伝子発現データの類似度を計算でき、自分が持つ遺伝子発現データと近い細胞種を瞬時に調べる事ができます。

もう一つ、このプロジェクトで重要なポイントは、遺伝子発現データをその細胞の情報(アノテーション)と共にデータベースに持つ事です。この観点から我々は標準細胞を多角度から定義するプロジェクトも始めており、例えば、肝細胞や筋肉細胞とはどのような形態(構造)や機能と言った特徴を持ち、その特徴はどのような遺伝子発現プロファイルから予測できるのかと言った問題にチャレンジしています。タンパク質の機能予測の分野では「配列から構造そして機能へ」とよく言われましたが、近い将来には「遺伝子発現プロファイルから細胞構造そして細胞・組織の機能へ」と続く道を研究していくことになるでしょう。

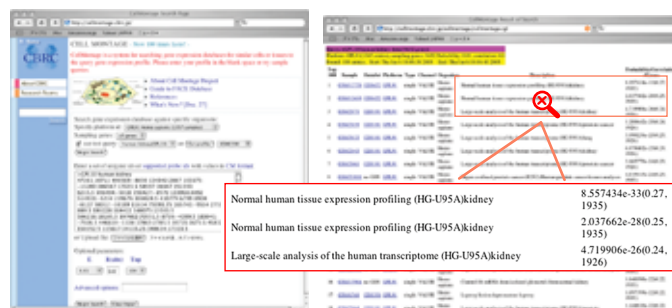
(参考) 特許出願 2004-280257 「遺伝子発現プロファイル検索装置、遺伝子発現プロファイル検索方法およびプログラム」(藤 航、ホートン ボール)。

図1 順位相関によるプロファイルマッチングの原理



ある成人肝臓から測定した遺伝子をその発現量に応じて順番に並べ、データベースに登録してある成人肝細胞または胎児の肝細胞と比較すると、胎児のものに比べ成人のものでは遺伝子の発現順位がより保存されていることがわかります。CellMontageでは、このような違いを瞬時に検出し、類似している順に細胞を出力します。従来のマーカー法などではある特定の遺伝子のみを扱うため、遺伝子間の関連が直接には得られませんでした。順位相関によるマッチング法ではよりダイレクトに遺伝子相互状態のネットワークを理解する事が可能です。さらに遺伝子の発現状態が細胞の構造や機能とどう関連しているのかを調べるため、標準細胞データベースも同時に開発しています。

図2 CellMontage 検索ページ (<http://cellmontage.cbrc.jp/>) (左) と検索結果 (右)



質問プロファイルとして、UniGene 番号とその発現量を入力する(左)と、データベースを高速検索し、結果を返してくれます(右)。ここでは、ヒトの腎臓の cDNAアレイデータを質問プロファイルに用いて、1,345枚のオリゴヌクレオチドアレイデータを 9,432 UniGeneを用いて検索したところ、2秒以内に結果が出力され、上位3つのものは全て腎臓関連データでした。