2005－BIO－2（5）
2005／10／7

細胞の知識ベース開発と遺伝子発現プロファイルによる細胞種と特徴予測

藤渕　航[1]、Larisa Kiseleva[1]、谷口　丈晃[2]、Paul Horton[1]

[1]産業技術総合研究所生命情報科学研究センター、[2]三菱総合研究所先端科学研究センター

genome と physiome を結ぶためには cellome が必要である。我々の研究は細胞種をターゲットとしてその種類、構造、機能的な特徴をゲノムの情報から推定することに取り組んでいる。まず、細胞に関するゲノムデータは、公共の遺伝子発現データベースを出発点として UniGene を用いた遺伝子名の統一化を行って、どのプラットフォームのデータでも同じフォーマットで処理できるようにした CellMontage データベースを開発した。この遺伝子発現データで独自にアノテーションを行い、ヒト正常細胞 81 種類でその遺伝子発現データ 1,714 プロファイルを分類した。この 81 細胞種を用いて、発現データのみから細胞の判別ができるか試みたところ、同一プラットフォームではランダムな遺伝子 100 個、異種プラットフォームでは 1000 個程度で、細胞の判別ができることがわかった。さらに細胞の画像解析を行い、細胞の持つ特徴パラメータ（例えば、円形率、核の領域の割合など）を抽出して、SVM による判別分析を行ったところ、遺伝子発現情報から細胞の構造的な違いを分類できる可能性を示唆する結果が得られた。

Development of Cell Knowledge Base and Prediction of Cell Types and Characteristics by Gene Expression Profiles

Wataru Fujibuchi[1], Larisa Kiseleva[1], Takeaki Taniguchi[2] and Paul Horton[1]

[1]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology, [2]Research Center for Advanced Science and Technology, Mitsubishi Research Instituite Inc.

The cellome is important for connecting the genome and the physiome. We have studied methods to predict cell types, structures, and functional characteristics by genome information. As the first approach, to obtain the genome information for cells, we used a public gene expression database and developed an integrated database called CellMontage that interconnects different data with UniGene ids for common gene names, which made any data from different platforms available in a uniform way. Using CellMontage we manually annotated gene expression data and classified 1,714 profiles for 81 types of human normal cells. Using the 81 cell types we tried to discriminate cells only by gene expression data and found that it can correctly discriminate cell types with as few as about 100 or 1000 random genes for the profiles from the same platform or different platforms, respectively. Furthermore, by analyzing images of cells, we parameterized some structural characteristics (e.g. roundness, nuclear region ratio, etc.) and performed discriminant analysis by SVM. The results indicate that it may be possible to classify cells with regard to their structural characteristics by gene expression information.

1. はじめに

ゲノム解析が進み、マイクロアレイなどに代表されるような genome からのデータが大量に生産され、細胞内の代謝シミュレーションなどが試みられるようになってきている一方、physiome プロジェクトでは心臓や脳のシミュレーションなど、生体機能の研究が盛んになってきている。しかし、そのミクロとマクロの生物学の間を補うような研究は多くない。より統合的に生体の理解を深めるためのゲノムから細胞そして組織までの研究が必要である。

我々は細胞を中心とした遺伝子発現データベースを開発し、分子レベルでのデータである遺伝子発現データから、形態学的レベルでのデータである細胞の種類や特徴を予測するための研究を行った。分子生物学では分子による還元論的な立場から生命現象を説明するが、一方、古典生物学においては、従来から細胞の形態的観察が行われてきたにも関わらず、両者の接点がそれほど見いだせていない。この論文では、観察者が顕微鏡像からいとも簡単に判別できる細胞の違いを、還元論的に分子のレベルで判別できるかどうかが主題である。

2. CellMontage—Cell Type-oriented Gene Expression Database

現在、登録プロファイル数の最も多いデータベースである GEO[1]データベースでは、プラットフォームが違うために相互に利用不可能なプロファイルが殆どである。そのため、我々は、UniGene をお互いのクロスリファレンス遺伝子とした統一フォーマットを考案し、図 1 の様なパイプラインで自動的に相互利用可能なプロファイルを取り出すことに成功した。同時に、クロスリファレンスを持たないプロファイルは、そのままの遺伝子名（プローブ名）で検索ができるように、データは重複して格納している。

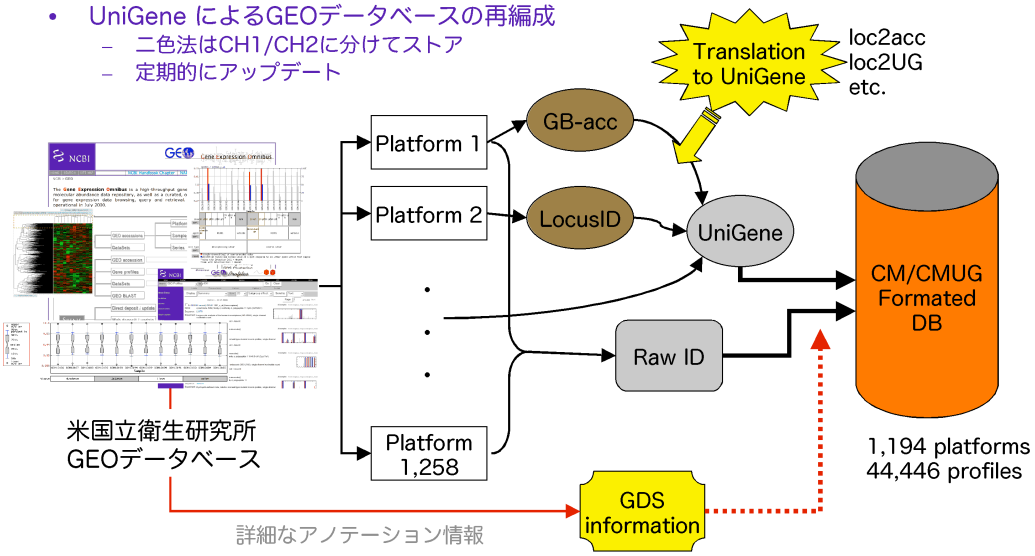

図 1. 公共データベースから統一フォーマットによる遺伝子発現データベース CellMontage 作成パイプライン

これにより、現在では 324 のプラットフォームから、30,497 のプロファイルが相互比較可能である。これは、全プロファイル約 45,000 プロファイルの 65%以上に当たるプロファイルが UniGene を用いて解析可能であることに相当する。

さらに、これとは別に独自に論文を読むなどして、GEO に含まれるヒトの正常細胞を分類し、92 種の細胞に当たる 42 プラットフォームから 1,844 プロファイルを整理することができた。このうち、UniGene を持つ CellMontage フォーマットに変換できたものは、81 種 40 プラットフォームの 1,714 プロファイルであり、これ以降の解析に用い

た。

3. 遺伝子発現プロファイルによる細胞種判別

まず、先ほどの 81 種の細胞を用いた細胞種の判別が可能か、相関係数を用いた相関序列解析を行った。相関係数計算には、異なるプラットフォームからの数値データの正規化などの前処理を避けるため、スピアマン順位相関係数(SRC 係数)を用いた[2]遺伝子の発現順序のみに注目して相関を求めた。さらに、使用する遺伝子の数を、16, 32, 64, …, 4096, 8192 と 10 段階に増やしながら、それぞれランダムな遺伝子を 20 回ずつ選択して SRC 係数を計算した。SRC 係数の総計算回数は、これらのパラメータに加えて、1,714 プロファイルの総当たりで行うため、3 億回近い計算になる。そこで、通常のプログラムよりも 4 倍以上高速な計算のできるアルゴリズム RaPiDS(Rapid Profile Data Search)[3]を開発し、実際の計算に用いた。

これを用いて、各プラットフォーム毎に相関係数を計算し、これを 1 標本 t 検定：

$$t = \rho\sqrt{n-2}\Big/\sqrt{1-\rho^2}$$

によって自由度 n-2 の確率に直す。この確率でプロファイルに序列を付け、自分と同種の細胞が何番目に来るかをプロットした。全てのプラットフォームのうち、最も多くの正常細胞プロファイル 297 枚を含んでいた GPL96(Affymetrix HG-U133A)に対して様々なプロファイルで SRC 係数を計算して序列を求めた。結果を図 2 に示す。

図 2 の左側のグラフでは、確認の意味で、自己プラットフォーム(GPL96)からのプロファイルでの相関序列解析の結果である。同じ細胞種が 2 プロファイル以上存在する細胞種 55 個での遺伝子数と相関順位のプロットである。左端の Y 軸上は期待順位を示している。明らかに、使用するプローブの個数が増えると、順位が良くなり、グラフが下の方へ収束している傾向が見える。特に、大部分の細胞種で、遺伝子の数で 128 個までに序列の改善がほぼ終了し、それ以上遺伝子を増やしても大きなな改善は見られない。全遺伝子を使用した場合に、placenta と cillilary ganglia 以外の細胞種では 297 枚中の 10 位以内に見つけ出されることがわかる。

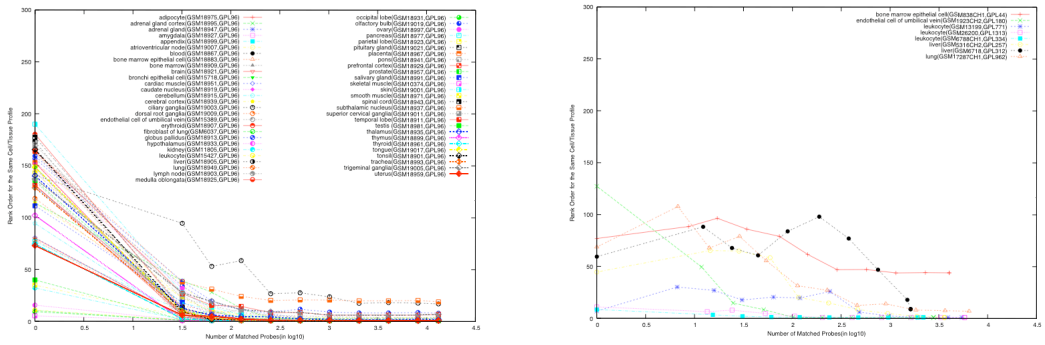

図 2. 同一プラットフォーム 55 種の細胞種（左）と異なるプラットフォームに 8 種（右）の細胞種による相関序列解析

一方、右図は affymetrix 以外のプラットフォームからの 8 データを質問プロファイルとして､同じく affymetirx のデータベース GPL96 に対して序列解析をした結果である。注意する点は、この 8 個には一色法だけでなく、二色法マイクロアレイも含み、その場合には正常細胞データに当たる片方のチャネルのデータのみを利用している。これを見ると 7 個の細胞については、遺伝子の数が 1000 個を超えると、同一プラットフォームの場合と同じように 10 位以内の順位に収束し、改善が少なくなる。残念ながら、一つは 50 位ほどで止まったままである。これは bone marrow epithelial cell(dual channel)であり、質問データの由来を調べてみると、体細胞ではなく細胞株によるもので bone

marrow stromal cell(骨髄間質細胞)が正しい由来であり、未分化な形質を持ち、上位から smooth muscle cell や bronchial epithelial cell にマッチしていた。これが、GPL96 の骨髄由来上皮細胞とは少しのずれがあるのかも知れない。また、GPL96 側には、他に血球細胞などの contamination があることも記述されていた。

4. 細胞画像からの特徴パラメータ測定と SVM による予測

細胞が異なればその構造や機能が異なるはずである。これをゲノム情報から予測できないか、そのためには何か必要かを理解することを試みた。まず、構造の違いを数値化するために、正常細胞のプロファイルが存在する 81 種の細胞種の顕微鏡写真を集め、その特徴パラメータの抽出を試みた。細胞写真は、東北大学医学部の協力のもとで収集したものと、インターネットからのものとの二種を用いた。このように写真の出典が様々なため、倍率や解像度が不明なものを含んでいる。そこで、細胞サイズの絶対値に関するパラメータは測定対象から外し、次の３つのパラメータに絞った測定をした。

(1)細胞円形度　(2)細胞縦横比　(3)核／細胞面積比

測定のために、図 3 の様な半自動的に細胞や核の外周や形態を認識するソフトウェアを開発し、測定の円滑化を図った。今回は、予備的な研究に当たり、20 種の細胞種から測定された値について、平均値を算出し、それより上位にあるか、下位にあるかで細胞を２群に分けた。この中から、均質なプロファイルを含むプラットフォームである GPL96 のみを用いて、16 細胞種 58 プロファイルを入力データとし、細胞種毎の Leave-one-out 交差検定による SVM 予測を行った。

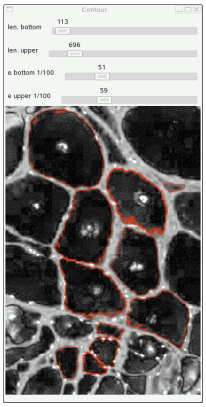

図 3. 細胞パラメータの測定

検定に用いた細胞種は、adipocyte(2)、adrenal gland(2)、bone marrow(2)、kidney(10)、liver(4)、lung(4)、ovary(2)、pancreas(4)、pituitary gland(2)、prostate(2)、salivary gland(2)、skin(2)、testis(10)、thyroid(4)、trachea(2)、uterus(4)であり(括弧はプロファイルの数を示す)、広範囲の異なる細胞種から構成されている。特徴遺伝子の選択は、反復除去法(Recursive Feature Elimination)を使用して、線形カーネルで学習した時に計算される重みから、遺伝子の順位を付け、50%除去を繰り返す、最も基本的な方法をとった。遺伝子を半数に減らしていき、エラー率が最低値になる場合の遺伝子群を採用した。SVM のプログラムは、gist-2.1.1[4]を使用した。その結果を下の表に示す。予測のエラー率の算定には bootstrap estimator である B632 推定法[5]を用いた。

	Positives / 16 cells	Positive samples	Prior	B632 error	Feature genes
Cell roundness	10	32	0.625	0.322	3312
Cell ellipticity	2	6	0.875	0.129	12
Nuc/Cell-area ratio	8	28	0.5	0.179	12

これを見ると、最下段の「核／細胞面積比」のエラー率がわずか 17.9%(予測率)であるにも関わらず、Prior が 50%であることから、予測が非常に良いことがわかる。この時の特徴遺伝子数 12 というのも、少数の特徴量で学習されており興味深い。この特徴遺伝子について、何か細胞の構造と関わりがあるか、現在調査中であるが、前立腺特異遺伝子など腺細胞の遺伝子も幾つか含まれており、細胞質の比が高いことと分泌細胞の働きを関連づけているのかも知れない。

5. おわりに

本研究では、世界中から単に収集しただけでは利用することのできない遺伝子発現データを、統一フォーマットで統合化するパイプラインの作成により成し遂げられ、そのデータを利用した研究である。本稿では、細胞の種類判別解析と細胞の構造予測に焦点を絞った。細胞の種類判別では、例えプラットフォームが違うデータであっても統合的に利用することが可能であるとの結果が得られた。これから、ますますマイクロアレイの測定精度が上がることが見込まれる。そうなれば、どのプラットフォームでも遺伝子の測定値の相関はより上がることは簡単に推測できる。現在の遺伝子発現データでは測定値に多くのノイズを含み、そのために結論があいまいになっているという実情を含む。今後は相互のデータ比較が可能になり、結論にもより信頼性が生まれてくるであろう。そのような時代が来た時に、遺伝子発現データから細胞の構造や機能または代謝などに立脚した「状態」を知ることが現在行われている癌の研究や発生の研究に大きく結びついてくると思われる。単に遺伝子の測定値を癌の転移などのマーカーにするのではなく、それがどのように細胞の機能を変化させるのかを対照とした研究がこれからますます必要となるであろう。

6. 謝辞

細胞の画像を集めるに当たり、ご協力いただいた東北大学医学部の増田高行教授に、心から感謝いたします。

参考文献

[1] Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W, Edgar R., NCBI GEO: mining millions of expression profiles--database and tools. *Nucleic Acids Res*. 2005 Jan 1;**33**(Database issue):D562-6.

[2] 特願 2004-280257: 遺伝子発現プロファイルデータベース高速検索、解析システム

[3] 特願 2005-236198: 遺伝子発現プロファイル比較装置

[4] http://svm.sdsc.edu/

[5] Braga-Neto UM, Dougherty ER: Is cross-validation valid for small-sample microarray classification? Bioinformatics 2004, 20(3):374-380.