

論文の内容の要旨

応用生命工学専攻

平成 11 年度博士課程入学

氏 名 門田 幸二

指導教官名 清水謙多郎

論文題目 cDNA マイクロアレイを用いた遺伝子発現解析手法の開発

1. はじめに

DNA マイクロアレイは、目的試料の相対的な遺伝子発現の変化を対照試料に対する比として、各試料をそれぞれの波長の異なる蛍光色素 (Cy3, Cy5) でラベルし、競合的ハイブリダイゼーションによって、1 回の実験で数万種類の遺伝子発現の変化を観察することができる技術である。最近ではその応用範囲を癌の分類などの臨床応用へと広げつつある。

癌の分類は、従来、組織学的・免疫学的手法により行われていたが、同じ型に分類されていた症例に対し、同じ薬を投与しても異なる臨床経過をたどるなどの問題があった。このためマイクロアレイを用いた遺伝子発現レベルでの癌のサブタイプの発見やその分類が試みられてきた。これまでにメラノーマや急性白血病など様々な疾患に対して、そのサブタイプの発見やサブタイプ間で発現が大きく異なる遺伝子群を用いて新たな症例を分類する試みがなされるなど、実際の臨床応用に期待もてる報告がなされている。しかしながら、悪性度の高い群と低い群の分類を行う場合に、実際は悪性度が高い症例であるにもかかわらず悪性度が低いと誤分類してしまうなど方法論的にまだ未成熟である。このような誤りが起こる他の原因としては、サンプル調製からデータ取得までの一連の過程における誤差の蓄積に起因する再現性の問題や、ある 2 群の分類に用いる遺伝子群の抽出法や用いる遺伝子数の問題などが複合的に絡み合っているものと考えられる。

癌発生のメカニズムに関して、複数の遺伝子異常が順次、積み重なることによって良性腫瘍が次第に悪化していき、前癌性病変(腺腫)を経て癌に発展していくという多段階

発癌説がある。その多くの知見が大腸癌における分子生物学的研究の成果によるものであるが、大腸癌は頻度的にも多く検討に必要な臨床材料が得られやすいこと、癌の病理学的そして遺伝的前駆体とされる“腺腫”の存在があることなどがその大きな理由である。

DNA マイクロアレイを用いて腺腫と癌の発現プロファイルの観点からみた分類を行う場合、悪性度では癌の方が高く、また少なくとも腺腫の一部は癌に進展していくことから、より悪性度の高い病態である癌を正しく癌と分類できること（感度 100%）が実際の臨床応用を目指す上で非常に重要である。また腺腫に関しては、単に腺腫であると分類するよりむしろどの程度癌の病態に近いかを評価することにより、遺伝子発現プロファイル的には癌に非常に近い腺腫のハイリスク群をスクリーニングできる可能性が考えられる。

そこで本研究では、1) 再現性の高いマイクロアレイデータを抽出する統計処理法の開発、2) 悪性度の異なる 2 つの臨床サンプル群に対してより悪性な状態を特徴づける遺伝子群の抽出法、及びその遺伝子群を用いて臨床応用を目指した高精度な分類システムの開発を目的とした。

2. マイクロアレイデータの統計処理法の開発

マウス胎児及び成体の主要組織（脳、心臓など）を含む計 49 組織を目的試料として、またマウス胎児 17.5 日目 whole body を対照試料として DNA マイクロアレイ実験が行われた。解析のための DNA チップとしては、理化学研究所（以下理研と略す）のマウス全 cDNA 収集プロジェクトによって得られた 18,816 個の cDNA が載せられたチップが用いられた。一つの目的試料に対して 2 回実験が行われ、チップ上の各遺伝子に相当するスポットの目的試料（Cy3 で標識）及び対照試料（Cy5 で標識）のシグナル強度、バックグラウンドシグナル強度、そしてフラグ情報が含まれる出力ファイルが 2 つ生成された。一連の実験は理研のスタッフによって行われた。

本研究で開発した統計処理法は、できるだけ再現性が高くかつ取り扱えるデータ数の減少を抑えたマイクロアレイデータのフィルタリングを行うことを目的としており、前述の 2 つのファイルを入力として、以下の三つの主要な手順を経てフィルタリングを行うものである： 1) 人手により、フラグが立てられたスポットを除去する、2) シグナル強度が Cy3 及び Cy5 両方のチャンネルにおいて、 $(\mu_{bg} + x\sigma_{bg})$ より低い値をもつスポットを除去する（ここで、 μ_{bg} 及び σ_{bg} は、それぞれバックグラウンドシグナル強度の平均と標準偏差を示す）、3) 2 回の実験をそれぞれ縦軸、横軸として発現比 $\log_2(\text{Cy3/Cy5})$ のプロットをとり、その最小二乗近似直線 $\pm y\sigma$ を越えるスポットを除去する（ここで、 σ は最小二乗近似直線からのスポットの距離の標準偏差を示す）。ここで、 x はバックグラウンドに埋もれている程度を示す閾値を、そして y は 2 回の実験のバラツキの程度を示す閾値を決定するためのパラメータであり、 N （統計処理前の遺伝子数に対する各手順後に残った遺伝子数の割合） $\times R$ （2 回の実験におけるフラグの立っていない発現比の相関係数）のスコアが最大値をとるような値を採用する。これら一連の処理を経た 2 回の実験データは正規化され、さらにそれをもとに 2 回の発現比の平均値が計算される。

本手法を用いて前述の 49 組織に対する実験データを解析した結果、統計処理前後の相

関係数の値が平均約 0.073 上昇し、特に処理前の値が低いデータほどより効果的であった。同じ組織由来のデータが発現プロファイルの類似度を指標にしてクラスタリングを行った際に同じ末端枝を形成するか否かも再現性を示す指標として用いられるが、その形成個数も本手法により 44 個から 47 個に増加した。また、上記評価基準をもとに、従来の 3 つの統計処理法と比較した結果、本手法が総合的に優れていることが確認された^{1,2)}。

3. 臨床サンプルの悪性度分類システムの開発

大腸癌 28 例（うち、臓器転移を起こしていない原発巣（CA と略す）16 例、臓器転移を起こした原発巣及び転移巣 12 例）及び腺腫（AD と略す）7 例の計 35 例を目的組織として、また大腸正常組織を対照試料としてマイクロアレイ実験が行われた。DNA チップとしては、Research Genetics 社製の 20,784 個のヒトクローンが載せられたチップを用いた。一連の実験は、理研の設備を用いて横浜市立大学のスタッフにより行われ、腫瘍サンプルごとに 2 つの出力ファイルが生成された。これらのファイルに前述の統計処理を適用し、最終的に縦軸が 20,784 遺伝子、そして横軸が 35 例の臨床サンプルからなる遺伝子発現行列を作成した。

全 35 例のうち、AD7 例及び CA16 例をトレーニングセットとして用い、計 23 例からなる遺伝子発現行列をもとに悪性（CA）状態を特徴づける遺伝子群の抽出を行った。残りの 12 例は評価のためのテストセットとして用いた。悪性状態を特徴づける遺伝子の抽出は、母集団(20,784 遺伝子)全てを用いたときの発現行列と j 番目の遺伝子 ($j=1,2,\dots,20784$) 発現プロファイルを除く残りの遺伝子発現行列に対し、下記の条件を満たす j 番目の遺伝子を候補として抽出した：

$$\overline{r_{ik}^{jth}}_{(i,k \in CA, i \neq k)} < \overline{r_{ik}^{all}} \quad \text{かつ} \quad \overline{r_{ik}^{jth}}_{(i \in AD, k \in CA)} > \overline{r_{ik}^{all}}$$

ここで、 $\overline{r_{ik}}$ はサンプル集合 i と k の発現比の総当りの相関係数の平均値、 $\overline{r_{ik}^{all}}$ は母集団を用いたときの $\overline{r_{ik}}$ 、 $\overline{r_{ik}^{jth}}$ は、母集団から j 番目の遺伝子を除いたときの $\overline{r_{ik}}$ を示す。本手法の有効性を評価するために、Leave-One-Out Cross-Validation（LOOCV と略す）を用いて CA の 16 例のうち一つを未知サンプル x とし、残りをトレーニングセットとして候補遺伝子の抽出を行い、以下の式により x を分類した：

$$S = \overline{r_{xk}^{all}}_{(k \in CA)} - \overline{r_{xk}^{all}}_{(k \in AD)}$$

評価は、 $S > 0$ ならば x は癌であるとし、 $S < 0$ ならば癌ではないとして、16 例各々を未知サンプルとして行った。

解析の結果、16 例全てについて正しく癌と判定できることが確認された。また、それらのスコアの有意性は、対応する候補遺伝子数（1728~2096 個）を母集団からランダムに 1000 回抽出して得られたスコアの 1%有意水準を越えており、開発した遺伝子抽出法の有効性が示された。候補遺伝子のさらなる絞込みのために、16 例全ての LOOCV によ

り候補として抽出された 335 個を用いて、全 35 例についてその分類精度を評価した。その結果、テストセットの 12 例を含む 28 例の癌サンプルを正しく癌であると分類することができた。また腺腫 7 例中 3 例が $S > 0$ (癌) と分類されたが、うち 2 症例は他の腺腫に比してその腫瘍サイズが大きい (共に 25mm) ことが確認された。腫瘍サイズは腫瘍切除の指標としても用いられていることから、本手法が腺腫の中での高悪性群のスクリーニングにも有用であることが示唆された。

4. まとめ

本研究では、DNA マイクロアレイ実験により得られる遺伝子発現データを効率的に処理し臨床応用を目指す上で有用な二つの手法を開発した。一つは、2 回行われた重複実験データを入力として再現性 (R) と最終的な遺伝子数の割合 (N) の両方を考慮し、 $N \times R$ が最大値をとる閾値を採用して統計処理を行うものであり、その有効性が本手法を適用する前のデータ及び他の方法を適用した結果との比較により示された。もう一つは、悪性度の異なる 2 群の臨床サンプルの発現データを入力として、高悪性群に特徴的な遺伝子群を抽出し悪性度分類を行うものであり、高悪性群を感度 100% で分類できることが確認された。本手法は、従来の病理学的分類の追試のみならず、生存率などの観点から見た予後の予測などへの応用が原理的に可能であり、様々な臨床診断への応用が期待される。

参考文献

- 1) 門田幸二、岡崎康司、清水謙多郎、林崎良英. マイクロアレイを用いた発現データベース構築とデータマイニング., *Mol. Med.*, **37**, 1186-1194, 2000.
- 2) Kadota, K., Miki, R., Bono, H., Shimizu, K., Okazaki, Y., Hayashizaki, Y., PRIM (Preprocessing Implementation for Microarray): An efficient method for processing cDNA microarray data., *Physiol. genomics*, **4**, 183-188, 2001.