順位情報を用いた遺伝子発現情報のクラスタリング

甲藤恭子　(指導教員：瀬々潤)

1　はじめに

遺伝子発現量データの解析では、遺伝子やサンプルをグループ化するために、クラスタリングが頻繁に使われる。同一のグループに属する遺伝子は同じ機能や関連する機能を持つ期待が持て、分析や実験の対象候補が得られる。また、同一のグループに属するサンプルは類似した細胞状態を有している可能性が高い。ただ、このクラスタリングを行う際には次のような問題がある。クラスタリングの前処理として、使用機材や実験条件が異なっても比較を可能にするために観測値の補正が行われるが、この補正は恣意性が強く主観が含まれてしまう。そこで、異なったマイクロアレイデータからであっても、信頼性の高いクラスタリングができるよう、遺伝子発現量を順位情報に変換した上でクラスタリングする手法を取る。

2　順位情報に変換する利点と問題点

遺伝子の発現量採取法として、マイクロアレイが一般的であるが、マイクロアレイは複数のメーカから販売され、さらに、各社別々の方式で遺伝子発現量の採取を行っているため、感度が異なり、メーカごとに異なった発現量の補正が必要となる。図 1 の両端に絶対的な値の異なる発現量例を示した。その点、順位情報であれば、絶対的な値は欠損してしまうものの、メーカごとに補正を行う必要はなくなる。

その一方で、問題点も存在する。発現量の高い遺伝子が非常に少なく、発現量の少ない遺伝子が非常に多いという偏った分布をしている。さらに、マイクロアレイを始めとする遺伝子発現量の観測技術は発現量の少ない遺伝子を正確に定量する事が難しい。このような発現量情報に対し、数量を順位に変換した場合、発現量の低い遺伝子の採取誤差を増幅してしまう可能性が高い。このため、順位情報に変換した場合は、この問題点を克服する必要がある。

本研究では、一般に細胞の状態変移は、いくつかの遺伝子がマスターキーとなり大きく発現量が変化して行われている可能性が高いことに着目した解決策を取る。図 1 の中央に例示した。順位が高い所での順位の入れ替わりは重要視し、順位の低い所では重要度を下げる事で、高順位優先の距離空間を作成し、発現量が低い際の観測誤差を克服する。また、発現量観測の際には欠損値もある点も考慮する。

3　関連研究

順位情報同士の距離としては、Spearman の順位相関係数や Kendall tau などが知られている。これらを利用したクラスタリングとして神嶌と藤木 [1] や Busse ら [2] があるが、いずれもすべての順位において同一の重みがつけられており、本研究で目的とする順位別の重み考慮はできないため、本研究を行う意義がある。

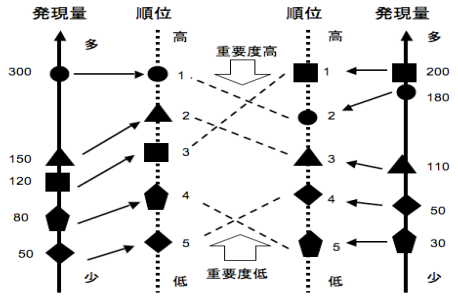

図 1: 数量情報を順位情報に変換して比較。順位が高いものを重要視

4　手法

ここでは、[1] の手法を改良して順位の高いものを優先する手法を提案する。

手順:

I. 各サンプルの順位情報を k 個のグループ $G_1, \cdots, G_k$ にランダムに分割する。

II. グループ $n = 1, \cdots, k$ について、遺伝子 g が $g'(g, g' \in G_n)$ より上位に来る確率 $\Pr[g, g']$ を求める。

III. サンプルの順位 O がグループ n に属する確率 $P(O, n)$ を求める。最も $P(O, n)$ が大きくなるクラスタ n を見つける。この時サンプル O はグループ k に属することにする。(この作業により、新しいグループ $n \in \{1, ..., k\}$ が生成される)

IV. II, III をあらかじめ定めた回数だけ行う。

[1] では、II. で各クラスタの中心順序を求めていたが、この操作が重く、また一意に中心順序が求まるとは限らないため、本研究では中心順序作成を回避している。これに伴い III の操作も変更している。

ここで、$\Pr[g, g'](n)$ と $P(O, n)$ の計算について考えよう。クラスタ k で g が g' より上位に現れる回数を $|g, g'|$ と表すと、単純には、

$\Pr[g, g'](k) = |g, g'|/(|g, g'| + |g', g|)$ であるが、確率が非ゼロになるように事前分布を導入し、以下の式で計算を行う。

$\Pr[g, g'](k) = (|g, g'| + 0.5)/(|g, g'| + |g', g| + 1)$　また、$P(O, k)$ は順序 O の中の各順序ペアがどれくらい起こりうるかを、全てのペアに渡って掛け合わせた確率で類推する。以下の式で表される。

$$P(O, n) = \prod_{g \succ g' \in O} \Pr[g, g'](n) \qquad (1)$$

本来各ペアの順序は互いに依存し独立ではないため、確率の積は正しく全体の起こる確率を表していないが、複合確率を求めるのは容易ではないため、近似的にこの確率を利用する。

更に順位の低い順序の重みを下げるために、g の順

位に依存した変数 r を導入し、以下の式を考える。

$$P(O,n) = \prod_{g \succ g' \in O} \big((1-r)\mathrm{Pr}[g,g'](n) + 0.5r\big) \qquad (2)$$

ここで g の順位を $O(g)$ とし、$r = \frac{O(g)}{|O|}$ としよう。積和の要素は順位が若ければ (1) と同一となり、高くなると 0.5 に近づくため、g と g' の順序が $P(O,k)$ に反映されにくくなる。

5 結果の比較と考察

本研究では、提案手法 (重み有りと無し) との比較として、順位情報を Kendall tau および元の数値のユークリッド距離を用いて階層的クラスタリングを行った。データは長嶋ら [3] のものを、上位 100 遺伝子に限定したものである。

表 1: 重み付き確率を用いないクラスタリング結果

クラスタ 1	control_for_AG1478, HRG_45min, HRG_U0126_5min,HRG_U0126_15min, HRG_U0126_30min
クラスタ 2	control_for_U0126, HRG_10min, HRG_15min, HRG_30min, HRG_90min, HRG_U0126_10min, HRG_U0126_45min, HRG_U0126_60min
クラスタ 3	HRG_5min, HRG_60min, HRG_U0126_90min

表 2: 重み付き確率を用いたクラスタリング結果

クラスタ 1	HRG_45min,HRG_60min,HRG_90min
クラスタ 2	control_for_AG1478,HRG_U0126_30min, HRG_U0126_45min,HRG_U0126_90min
クラスタ 3	control_for_U0126,HRG_5min,HRG_10min, HRG_15min,HRG_30min, HRG_U0126_5min, HRG_U0126_10min,HRG_U0126_15min, HRG_U0126_60min

表 1 は順位データを重みを用いない式 (1) の手法で、表 2 は重みを考慮した式 (2) の手法でそれぞれクラスタリングを行い、16 個のサンプルが 3 つのグループに分かれるようにした。

順位情報同士の距離を求める方法として、Kendall tau によりサンプル間の距離行列を作成し、その後 Ward 法によりクラスタリングを行ったものを図 2 に示す。Kendall tau を計測する際、発現量に欠損値が含まれるため、どちらか片方に欠損値を含む場合、その値を除いた遺伝子で距離を計測した。図 3 に、ユークリッド距離と ward 法によるクラスタリング結果を示す。図 2 と図 3 に関しては、生成されるクラスタの数が 4 つになるような閾値を与え分割した。

今回の提案手法の結果として表 2 に注目すると、クラスタ 1 の要素として HRG_45min,HRG_60min,HRG_90min が含まれている。これらは HRG 刺激から時間がたったサンプルであり、刺激によって変化した細胞の様子をとらえられたと考えられる。これは、行った全ての結果中で今回提案する重み付きの順位クラスタリングの結果に唯一の特徴であった。

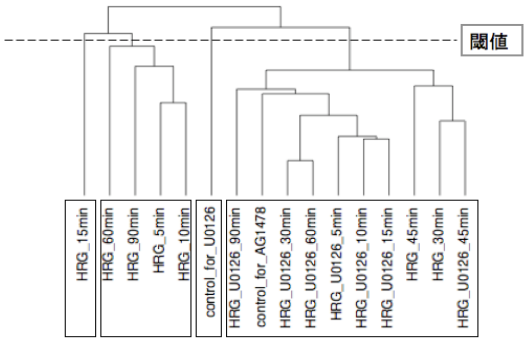

図 2: Kendall tau によるクラスタリング結果

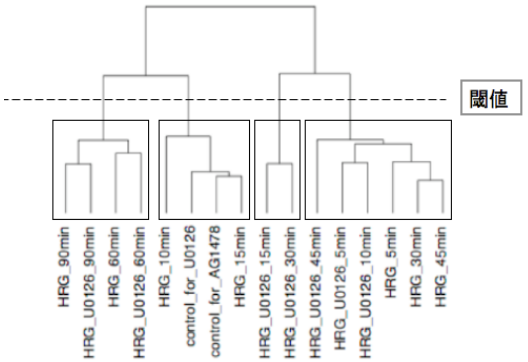

図 3: ユークリッド距離によるクラスタリング結果

6 今後の課題

今後は、順位データを使ったクラスタリングを更に改良したい。また今回データとして扱った遺伝子発現情報のデータに限らない、別のデータを使って比較を行うことで、今後の応用についても検討したい。

参考文献

[1] Toshihiro Kamishima and Jun Fujiki: Clustering Orders. Proc. of the 6th Int. Conf. on Discovery Science, 2003.

[2] Ludwig M.Busse, *et al.*: Cluster Analysis of Heterogenious Rank Data. Proc. of th 24th Int. Conf. on Machine Learning, 2007.

[3] Takeshi Nagasima *et al.*:Quantitative Transcriptional Control of ErbB Receptor Signaling Undergoes Graded to Biphasic Response for Cell Differentiation. J. Biol. Chem., Vol. 282, Issue 6, 4045-4056, 2007.