Vol.2009-BIO-18 No.7
2009/9/18

事前知識を用いた遺伝子発現量の部分空間クラスタリング

大 村 蓉 子[†1]　瀬々 潤[†1]

時期，周期特異的に機能する遺伝子の発現量は，観測条件全体で相関があるのではなく，特定の時期及び周期でのみ高い相関を示すことがある．このような遺伝子群発見のため，Biclustering を始めとする部分空間クラスタリングが研究されている．しかし，既存手法にはオーバーフィットによる生成されたクラスタ信頼性の低さや，多量のクラスタ生成による結果解釈の困難さに問題点がある．そこで，近年のゲノム解析で蓄積されている遺伝子機能を用い，類似の機能を有する遺伝子から部分空間クラスタを生成し，その後，他の遺伝子へクラスタを拡張し，さらに重複を除去する事で，既知の機能に即したクラスタを発見する．遺伝子オントロジーを用いて既存手法と比較したところ，Biclustering に比べ，既知の遺伝子機能に即した遺伝子，条件群を発見できた．

Subspace clustering of gene expression data with prior knowledge

Yoko Omura[†1] and Jun Sese[†1]

The subspace clustering such as Biclustering has been researched for finding genes activated on specific conditions or specific cell cycles, whose gene expression levels are highly correlated only under the active conditions. However, existing methods have problems of the lack of cluster reliability caused from over-fitting and the difficulty to interpret the clusters because of the generation of numerous clusters. To overcome these problems, we generate a new method, which contains four steps: selection of functional similar genes using biological priori knowledge, generation of subspace clusters from the selected genes, enhancement of the clusters with non-selected genes, and removal of redundant clusters. Our test shows that clusters generated by our method are more correlated with Gene Ontology annotations than Biclustering.

†1 お茶の水女子大学 大学院人間文化創成科学研究科
Ochanomizu University Graduate School of Humanities and Sciences.

1. はじめに

近年，マイクロアレイ実験が安価になってきたことにより，大量のマイクロアレイを用い，様々な環境下において遺伝子発現量観測を行う実験が増加している．このような実験では，全ての環境下において遺伝子発現が相関しているとは限らないため，特定の環境下のみで発現量の相関が見られる遺伝子やその条件の抽出技術が求められている．このような，時期，条件特異的に働く遺伝子群とその条件を知るために，部分空間クラスタリングが導入され，既存の手法として Biclustering[1)]，pCluster[2)] がある．しかしこれらの手法は，クラスタ生成に大量のメモリや計算時間を必要とすることや，全体的に発現の低いクラスタが生成される，クラスタ間に重複が多いことなどから，クラスタの重要性の正しい判断を妨げ，また既知の生物学的知識に沿わないクラスタを生成してしまう可能性があるという問題点がある．他方，ゲノム解析の進展により遺伝子機能に関する知識の蓄積が進んでおり，発現量解析では無知識から遺伝子機能を求めるのではなく，既に分かっている遺伝子機能に関する情報を利用することで，より既存の知識に即したクラスタの発見が可能であると考えられる．そこで本研究では，予め類似機能を有する遺伝子集合を網羅的データから抽出し，このデータから pCluster を生成．その後発現パターンの類似している他の遺伝子を追加してクラスタを拡張することにより，既存の知識に即したクラスタを抽出し，さらにクラスタ間の重複を，クラスタの併合，削除により減らすことで重要なクラスタを選別することができる手法を提案する．提案手法により，Biclustering と比べて，既知の知識に沿うクラスタを得ることができるだけでなく事前に決定した特定の遺伝子発現量データからクラスタを生成することで，メモリ使用量も軽減でき，高速に結果を得ることができた．

2. 関連研究

マイクロアレイによる実験で得られた遺伝子発現量データの解析方法として類似した発現パターンを示す遺伝子群や観測条件群をグループ化するクラスタリングが用いられている．クラスタに分割することで，クラスタ内の各遺伝子にどのような共通点があるのかを知ることができ，遺伝子の分類や機能の推定に役立っている．多くのクラスタリング手法では，全ての遺伝子あるいは条件で相関が高い事が要求されるが，このような手法では時期，条件特異的に働く遺伝子の抽出が難しい．そこで，時期，条件特異的に働く遺伝子群とその条件を求められるよう，近年部分空間クラスタリングが着目されている[1)–4)]．これらの手法は，特定の観測条件群でのみ発現パターンが類似している遺伝子群を見つけることもでき，

どのような観測条件でどの遺伝子が働いているのかを発見できる．

Biclustering[1] は，クラスタの平均 2 乗誤差が閾値以下のクラスタの内，最も大きさの大きいものから順にクラスタを出力する．この手法は遺伝子発現量解析に部分空間クラスタリングを導入した走りとも言え，閾値の設定が比較的容易であり，また互いに重複の少ないクラスタを生成することができるが，網羅的な遺伝子発現量からのクラスタ抽出を行おうとすると，出力するクラスタの大きさを最適化するために実行時間がかかる事，全体として発現の低いクラスタが多数生成されることから，既知の生物学的知識に沿わないクラスタが生成されることがある．本提案手法は，既知の遺伝子情報を用いることで，既存の知識に即したクラスタを生成できること，網羅的な発現情報に対しても，十分高速に計算できる点で有意性がある．

pCluster[2] は，1 乗誤差が一定の範囲内に収まる部分空間クラスタを効率よく列挙する方法である．値間の類似度が非常に高く，観測条件特異性の高いクラスタを得ることが出来る．しかし，生成するクラスタ数が増えると大量のメモリを消費すること，クラスタ間に重複が多く重要性の判断が困難なことがこの手法の問題点である．提案手法は，予め事前知識を用いて生成したデータセットに対し，pCluster を適用する事で，メモリの少ない環境下でも網羅的遺伝子データに対し解析が出来ること，重複したクラスタの内，どのクラスタが重要であるかを選別でき，結果の解釈が容易であることの 2 点において優位性がある．

一度 Biclustering で求めたクラスタを基に，クラスタ境界を変更することで，より良質なクラスタ生成を目指す手法として，FLOC[3] がある．FLOC では Biclustering よりも大きい，または 2 乗誤差の少ないクラスタを生成することが可能であるが，Biclustering と比べてさらにメモリと時間を必要とし，ノイズを許容する可能性のある手法であり，また大量のクラスタが生成される点は解消されていない．本研究では，確率的に有意な重複を持つクラスタを併合することで，クラスタ数を絞り，結果の解釈を容易にしている．

3. 手　　法

本研究では，ある類似した機能を持つ発現量データにおける pCluster の実行結果を基に，それ以外の遺伝子に対してクラスタを拡張し，同じ機能を持つ遺伝子群を確実に同定する方法を提案する．本手法の全体像を**図 1** に示す．以下の節で，それぞれの詳細を説明する．

3.1 pCluster と bicluster

定義 1　観測した遺伝子発現量の集合を $\mathcal{D}$ とし，観測した全遺伝子集合と全観測条件集合をそれぞれ $X(\mathcal{D})$, $A(\mathcal{D})$ とする．遺伝子 $x \in X(\mathcal{D})$ の条件 $a \in A(\mathcal{D})$ における遺伝子発現量の観測値を d_{xa} とする．また，$x \in X(\mathcal{D})$, $a \in A(\mathcal{D})$, $X \subseteq X(\mathcal{D})$，$A \subseteq A(\mathcal{D})$ に対し，以下の値 d_{xA}, d_{Xa}, d_{XA} を定義する．

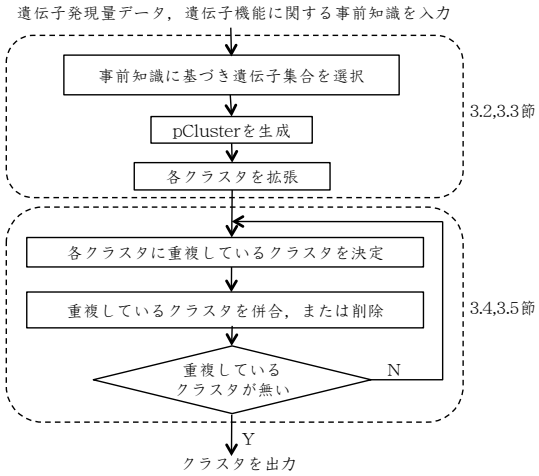


図 1　提案手法の全体の流れ
Fig. 1　Overview of proposed algorithm.

$$d_{xA} = \frac{1}{|A|}\sum_{a\in A} d_{xa}$$

$$d_{Xa} = \frac{1}{|X|}\sum_{x\in X} d_{xa}$$

$$d_{XA} = \frac{1}{|A||X|}\sum_{a\in A, x\in X} d_{xa}$$

d_{xA}, d_{Xa}, d_{XA} はそれぞれ遺伝子 x の条件集合 A における遺伝子発現量の平均値，条件 a における遺伝子集合 X の遺伝子発現量の平均値，遺伝子集合 X と条件集合 A における遺伝子発現量の平均値を表している．

Biclustering では，クラスタ内に含まれる値の 2 乗誤差が最も小さくなるクラスタを見つける．

定義 2　(クラスタの 2 乗誤差) 遺伝子集合 X と条件集合 A に対し，遺伝子発現量の 2 乗

Vol.2009-BIO-18 No.7
2009/9/18

誤差 $r(X, A)$ を

$$r(X, A) = \frac{\sum_{a \in A, x \in X} (d_{xa} - d_{xA} - d_{Xa} + d_{XA})^2}{|X||A|}.$$

一方，pCluster では次の 1 乗誤差の値を利用する．

定義 3 (pCluster) δ, nr, nc をユーザ定義の値とする．クラスタ $\mathcal{D}' = X \times A$(ここで $X \subseteq X(\mathcal{D})$ かつ $|X| \geq nr$ 及び $A \subseteq A(\mathcal{D})$ かつ $|A| \geq nc$) に対し，任意の $x, y \in X$ 及び $a, b \in A$ が

$$|(d_{xa} - d_{xb}) - (d_{ya} - d_{yb})| \leq \delta$$

を満たすとき，$\mathcal{D}'$ が pCluster であると呼ぶ．pClustering は，全 pCluster を列挙する問題である．

では，実際にこれらの既存手法ではどのようなクラスタリング結果を得ることができるのだろうか？**表 1** を例に示す．表 1 は遺伝子発現量データ，g1,. . .,g10 の 10 個の遺伝子の a,. . .,l の 12 の観測条件における発現量を数値で示している．この例では簡単のため，発現量は整数としておく． 閾値 $nc = 3$，$nr = 3$，$\delta = 1$ と設定し，pCluster を使ってクラスタリングを行うと，**表 2** に示す 12 個のクラスタが生成される．各行が生成された各クラスタを示しており，genes が遺伝子，conditions が観測条件を表している．同じ遺伝子が複数のクラスタに含まれ，重複しているクラスタが多数生成されており，各クラスタの違いを明確にするのは非常に難しい．一方 $\rho = 1.0$ で Biclustering によってクラスタを生成した場合の結果が**表 3** である．pCluster と比べると数も少なく重複もほぼ無いが，このように閾値を低く設定した場合あまり大きいクラスタの出力が見込めず，また外れ値も比較的多く含んでしまう．さらにここで重要なのは，ここで使用しているデータ（表 1）そのものが非常に小さいということである．実際にはずっと大きなデータにおいて，これらの手法が適用されることが求められているにもかかわらず，これらの方法はデータが大きくなるにつれて生成クラスタ数が増えて違いが分かりにくくなる．さらに莫大なメモリの消費に加え，発現類似性に関してクラスタリングする上で類似度の低いクラスタが多く生成され，クラスタとして信頼性に欠けてくる可能性もあることから，実験者が気軽に使用出来るツールとしては難がある．そこで，これから本文ではこれらの問題点を改善する，クラスタリングの改良手法を提案する．

表 1 遺伝子発現量データ例
Table 1 An example of gene expression data.

	a	b	c	d	e	f	g	h	i	j	k	l
g1	6	31	33	6	39	21	12	48	10	28	41	31
g2	80	5	8	35	14	28	20	23	55	3	16	6
g3	28	12	15	12	21	10	7	30	18	10	23	10
g4	7	40	42	4	48	2	1	58	7	38	5	38
g5	10	20	30	40	50	60	50	40	30	20	4	18
g6	2	12	14	24	26	27	3	60	36	44	41	10
g7	10	37	59	80	24	24	24	50	24	24	42	35
g8	6	31	35	6	41	21	12	50	10	29	42	29
g9	4	14	16	11	46	32	23	62	38	19	20	12
g10	10	20	22	31	32	40	1	68	44	10	15	18
g11	23	33	2	6	4	20	55	82	57	44	60	31

表 2 表 1 から生成された pCluster
Table 2 pClusters extracted from Table 1.

genes	conditions	genes	conditions
g6,g9,g10	a,b,c,h,i,l	g6,g9,g10,g11	a,b,h,i,l
g5,g6,g9,g10,g11	a,b,l	g1,g2,g3	b,c,e,h,j,k
g2,g3,g8	b,c,e,h,j,k	g1,g2,g3,g4	b,c,e,h,j
g3,g4,g6,g9,g10	b,c,l	g3,g4,g8	b,h,j,l
g2,g3,g4,g8	b,h,j	g1,g2,g3,g8	b,j,k
g1,g2,g3,g8	c,e,h,j,k	g1,g2,g3,g4,g8	c,e,h

表 3 表 1 から生成された bicluster
Table 3 Biclusters extracted from Table 1.

genes	conditions
g1,g3,g8,g10	b,c,e,l
g1,g7,g8	a,f,h,k
g2,g3,g4,g5	b,h,j
g6,g9,g10	a,f,i

3.2 類似した機能を持つ遺伝子発現量データ

クラスタを生成する上で，実験者が各クラスタの示す意味を理解し，解釈し易い結果を簡単に得られるかどうかは非常に重要である．そこでまず，本手法では以下を定義する．

定義 4 (事前知識から選択されたデータ集合) 予め事前知識から選択した遺伝子集合を $X_S \subset X(\mathcal{D})$ とする．この時，遺伝子集合 X_S 及び条件集合 $A(\mathcal{D})$ から成るデータ集合を $\mathcal{D}_S$ とする．

$\mathcal{D}_S$ から閾値 δ, nc, nr を用いて得られる pCluster の集合を $\mathbf{D}_C = \{\mathcal{D}_{C1}, \mathcal{D}_{C2}, \ldots \mathcal{D}_{Cn}\}$ とする．このとき, これらの生成クラスタを基準に X_S のクラスタリングでは見つけることができなかった，類似機能を持つ遺伝子群を $\mathcal{D}$ から見つける事が目的であるため，拡張の条件を緩和できるよう，ある程度の小ささであるクラスタを生成することにする．これらのクラスタをもとのデータの遺伝子に拡張していくことで，その機能に関与している機能を持つ遺伝子を含む同じ機能を持つ遺伝子群を発見する事ができると考えられる．続いて, そのクラスタの拡張について具体的に説明する．

Vol.2009-BIO-18 No.7
2009/9/18

3.3 クラスタの拡張と事前知識の利用

部分空間クラスタリングにおける問題点として，探索空間が大きいため，生成されたクラスタがデータにオーバーフィットし，既存の知識に沿わない結果が生成されることが挙げられる．一方，遺伝子オントロジー[5] をはじめとする生物学的知識の蓄積により，機能が類似した遺伝子に関する情報が得られるようになっている．本節ではこの事前知識を基に部分空間クラスタを作成することで，クラスタのオーバーフィットを避け，事前知識を補完する部分空間クラスタを作成する．

まず，クラスタ $\mathcal{D}_C$ に対して新たな操作「拡張」を定義する．

定義 5 データベース $\mathcal{D}$ の部分集合 $\mathcal{D}_S$ から得られる pCluster$\mathcal{D}_C$ を考える．遺伝子 $g \notin X(\mathcal{D}_C)$ を $\mathcal{D}_C$ に追加し，遺伝子集合 $X(\mathcal{D}_C) \cup \{g\}$，条件集合 $A(\mathcal{D}_C)$ から成るクラスタを作成する．この操作を g による $\mathcal{D}_C$ の遺伝子の拡張と呼ぶ．同様に，条件 $a \in A(\mathcal{D})$ を加えることを条件の拡張と呼ぶ．

遺伝子 g による拡張の際，g の発現が $\mathcal{D}_C$ 内のものと関連が薄い場合，拡張によりクラスタが壊れてしまう．ここでは，適切な g を選ぶため，次の定義を行う．

定義 6 遺伝子 g により $\mathcal{D}_C$ を拡張した時，クラスタ内の値の 2 乗誤差が最も小さくなる遺伝子を

$$g_{min} = argmin_{g \notin X(\mathcal{D}_C)} r(X(\mathcal{D}_C) \cup \{g\}, A(\mathcal{D}_C))$$

と定義する．

$r(X(\mathcal{D}_C) \cup \{g_{min}\}, A(\mathcal{D}_C)) \leq \rho$ ならば $X(\mathcal{D}_C)$ を $X(\mathcal{D}_C) \cup \{g_{min}\}$ に拡張する．$r(X(\mathcal{D}_C) \cup \{g_{min}\}, A(\mathcal{D}_C)) \leq \rho$ である限り，この走査を繰り返し，$\mathcal{D}_C$ を g_{min} で拡張していく．

以上の拡張を，各クラスタについて行う．本手法では，最初に与えられた大規模な遺伝子発現量データから，直接 pCluster を求めた場合と比べて，確実に類似した遺伝子を見つけることができ，且つ類似した機能を持つ遺伝子の発現量データからのみクラスタを求めるので，計算に消費されるメモリを軽減出来るという利点がある．

実際に表 1 を $\mathcal{D}$ として本手法を適用してみよう．例として g1,. . .,g11 という 11 個の遺伝子からなる遺伝子集合 $X(\mathcal{D})$ の中で，$X_S = \{$g4, g5, g6, g9$\}$ が事前知識より同一の機能を有すると仮定する．X_S 及び 12 の観測条件からなるデータ集合 $\mathcal{D}_S$ から pCluster を生成．閾値は先の例と同様 $nc = 3$，$nr = 3$，$\delta = 1$ と設定する．生成される pCluster$\mathbf{D}_C = \{\mathcal{D}_{C1}, \mathcal{D}_{C2}\}$ が **表 4** である．これらのクラスタは表 2 にそのまま含まれているが，注目する遺伝子を絞り込んだため，数は少ない．次にこの $\mathbf{D}_C$ をもとに実際に拡張をおこなっていく．ここで閾値 $\rho = 1.0$ とする．$\mathcal{D}_{C1}$ に対して $X(\mathcal{D}) - X_{C1}$ の遺伝子を 1 つずつクラスタに加えたときの平均 2 乗誤差を求めると，$g_{min} =$ g10，g11 で $r(X_{C1} \cup \{g10\}, A_{C1}) = r(X_{C1} \cup \{g11\}, A_{C1}) = 0.0$ となる．これは ρ を下回っているので，クラスタに追加することができる．最小値を複数取る場合は，その全てを追加する．$\mathcal{D}_{C1}$ を更新し，$X_{C1} =$ {g5，g6，g9，g10，g11} となる．続いてそれ以外の遺伝子について平均 2 乗誤差を求めていくと $g_{min} =$ g8 で $r(X_{C1} \cup \{g8\}, A_{C1}) = 6.94$ であった．これは ρ を満たしていないため，追加はできない．よってこれ以上追加出来る遺伝子はないことが分かり，このクラスタの拡張は完了したことになる．同様にして $\mathcal{D}_{C2}$ も拡張した結果が**表 5** である．平均 2 乗誤差を用いた拡張により，pCluster や bicluster では得る事のできなかったクラスタも得ることができ，また生成数が少ないこと，既知の情報を用いていることから，各クラスタの解釈や機能の予測を容易にすることができたと考える．

表 4 事前知識から生成された pCluster
Table 4 pClusters extracted from data consisting of g4,g5,g6 and g9.

genes	conditions
g5,g6,g9	a,b,l
g4,g6,g9	a,b,l

表 5 拡張後のクラスタ
Table 5 Enhanced clusters.

genes	conditions
g5,g6,g9,10,g11	a,b,l
g1,g2,g3,g4,g6,g8,g9,g10	a,b,l

3.4 クラスタの重複を避ける

pCluser における問題点として，多くのクラスタが重複して生成されることが挙げられる．これは，求められた pCluster $\mathcal{D}'$ に対し，$\mathcal{D}'$ の部分集合や $\mathcal{D}'$ と重複を多く持つデータも pCluster の条件を満たすことが多いからである．この問題点は，我々の事前知識上のクラスタに基づく拡張したクラスタ生成においても同様である．しかし，遺伝子発現量解析の際には，このようなクラスタの内最も多くの遺伝子や条件をカバーするクラスタのみが必要とされ，他の小さなクラスタは解析の対象とはならないため，結果から除外したい．本節では，このように重複したクラスタから，真に必要なクラスタを抜き出すため，クラスタ間の重複に着目し，より大きなクラスタで表すことの出来るクラスタを削除する．複数のクラスタが重複する場合，大きさの小さなクラスタより，大きいクラスタの方が偶然抽出できる確率も低く，生物学的な発見に繋がる可能性が高い．ここでは，あるクラスタ $\mathcal{D}_C$ より，クラスタの大きさ $|X(\mathcal{D}_C)||A(\mathcal{D}_C)|$ が小さいクラスタがどのように重複するか，場合を分けて考える．そして，重複度合いが大きい場合は大きさの小さいクラスタを削除する．クラスタ $\mathcal{D}_C$, $\mathcal{D}'_C$ 間の重複度合いの判定を定義する．まず，$C(\mathcal{D}_C)$ をクラスタ $\mathcal{D}_C$ 内の観測値の集合とする．$|C(\mathcal{D}_C)| = |X(\mathcal{D}_C)||A(\mathcal{D}_C)|$ である．$\mathcal{D}_C$, $\mathcal{D}'_C$ 間で重複している観測値の数は

$|C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|$ である．ここで，2 つのクラスタ間で与えられた重複が起こる割合を

$$p(\alpha, \beta, \gamma) = {}_{\beta}C_{\gamma} k^{\gamma}(1-k)^{\beta-\gamma}$$
$$\text{ここで } k = \frac{\alpha}{|X(\mathcal{D})||A(\mathcal{D})|}$$

と定義する．これは 2 項検定の式であり，ここでは，$\alpha = |C(\mathcal{D}_C)|$, $\beta = |C(\mathcal{D}'_C)|$, $\gamma = |C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|$ となる．

設定した ω を閾値とし，重複の度合いを判定する．$\mathcal{D}'_C$ は，$|C(\mathcal{D}'_C)| < |C(\mathcal{D}_C)|$ なるクラスタとする．

(1)　$p(|C(\mathcal{D}_C)|, |C(\mathcal{D}'_C)|, |C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|) \leq \omega$ なるクラスタ $\mathcal{D}_C$ が存在する．

有意に重複が認められる場合を示しているため，$\mathcal{D}_C$ を基に，$\mathcal{D}'_C$ に含まれて $\mathcal{D}_C$ に含まれない遺伝子，及び条件について拡張を行う．

図 2 は，表 2 で示した pCluster の中のある 2 クラスタの重複の様子を示したものである．これらのクラスタは多くの部分空間が重複しているが，また遺伝子 g8 の観測条件 b における発現量が外れ値を取るため，異なるクラスタとして生成されているが，他の観測値は非常に類似している．このクラスタの重複度合いを計算してみると $p = 2.537E-10$ となり，$\omega = 1.0E-4$ を下回るので，有意に重複が多いことが分かり，2 つのクラスタを併合する．併合の手順については，次節で説明する．

(2)　$p(|C(\mathcal{D}_C)|, |C(\mathcal{D}'_C)|, |C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|) \leq \omega$ なる $\mathcal{D}_C$ が存在せず，
$p(|C(\mathcal{D}_C)|, |C(\mathcal{D}'_C)|, |C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|) > \omega$ なるクラスタ $\mathcal{D}_C$ が複数存在する．

小さな 1 つのクラスタ $\mathcal{D}'_C$ に複数の大きいクラスタが部分的に重複している状態である．**図 3** は図 2 と同様に表 2 中のある 3 クラスタの重複の様子を示している．ω を上回っており，ユーザの解釈を妨げる重複が存在している可能性があると考えられる．このような場合，複数のクラスタの和集合が，$\mathcal{D}'_C$ に近いものになる場合は，和集合の包含する範囲に値の似ている物があることが予想できるので，小さなクラスタは $\mathcal{D}_C$ の一部であると考え破棄する，もしくは複数のクラスタに部分的に併合したい．この判定を行うため，$p(|C(\mathcal{D}_C)|, |C(\mathcal{D}'_C)|, |C(\mathcal{D}_C) \cap C(\mathcal{D}'_C)|) > \omega$ を満たすクラスタ集合を D_C としたとき，$\bigcup_{\mathcal{D}_C \in D_C} C(\mathcal{D}_C)$ なる集合 C_C を考え，この集合と $\mathcal{D}'_C$ の重複を考える．この判定は，前述の $p(\alpha, \beta, \gamma)$ を用い，$p(|C_C|, |C(\mathcal{D}'_C)|, |C_C \cap C(\mathcal{D}'_C|)$ で計算することが出来る．

以上の考察より，$p(|C_C|, |C(\mathcal{D}'_C)|, |C_C \cap C(\mathcal{D}'_C)|) \leq \omega$ を満たす場合には，$|C(\mathcal{D}'_C)| - |C_C \cap C(\mathcal{D}'_C)| = 0$ ならば $\mathcal{D}'_C$ を破棄，そうでないならば $\mathcal{D}'_C$ を D_C に併合する．

以上の条件以外の場合には，重複が非常に少なく，結果としてユーザの理解を妨げる可能性が低く，独立していると考えられる例である．クラスタの重複が非常に少なく独自の特徴を持っていると言えいずれのクラスタにも併合はしない．

図 2　大きなクラスタの重複 (1)
Fig. 2　An example of a large overlap of clusters(1).

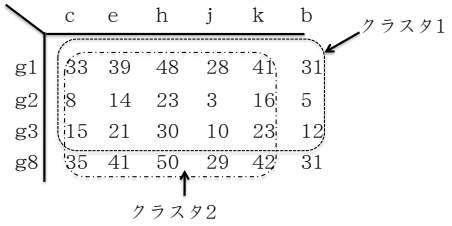
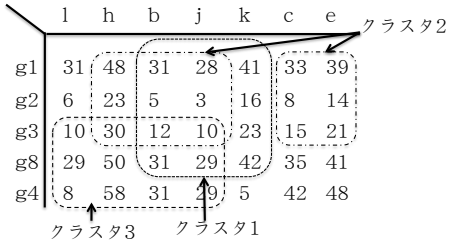


図 3　複数のクラスタの重複 (2)
Fig. 3　An example of several overlaps of clusters(2).



3.5　併合と削除

例として，得られたクラスタ $\mathcal{D}'_C$ には上記の重複の種類 (1) に該当する重複がある場合を考える．つまり，強い関連を持って重複しているクラスタ $\mathcal{D}_C$ を $\mathcal{D}'_C$ によって拡張する．ここで，$\mathcal{D}_C$ に併合できる可能性があるのは，$\mathcal{D}_C$ に含まれていない $g \in X(\mathcal{D}'_C) - (X(\mathcal{D}_C) \cap X(\mathcal{D}'_C))$ の遺伝子集合と $a \in A(\mathcal{D}'_C) - (A(\mathcal{D}_C) \cap A(\mathcal{D}'_C))$ の観測条件集合である．本手法での併合とは，$\mathcal{D}'_C$ を全て $\mathcal{D}_C$ に併合するのではなく指標に基づいて，その中から部分的に 1 遺伝子，1 観測条件単位で併合していく．先の 3.3 で述べた pCluster を拡張する方法と同様に平均 2 乗誤差を指標として遺伝子，条件共に併合を行う．また拡張ではなく，重複したクラスタを減らすことが目的なので，1 遺伝子ずつ併合した場合の平均 2 乗誤差だけでなく 1 属性ずつ併合した場合の平均 2 乗誤差も求めた上で，$r(X(\mathcal{D}_C) \cup \{g\}, A(\mathcal{D}_C)) < r(X(\mathcal{D}_C), A(\mathcal{D}_C) \cup \{a\})$ かつ $r(X(\mathcal{D}_C) \cup \{g\}, A(\mathcal{D}_C)) \leq \rho$ ならば g を $X(\mathcal{D}_C)$ に併合して，$X(\mathcal{D}'_C)$ から削除し，$r(X(\mathcal{D}_C), A(\mathcal{D}_C) \cup \{a\}) < r(X(\mathcal{D}_C) \cup \{g\}, A(\mathcal{D}_C))$ かつ $r(X(\mathcal{D}_C), A(\mathcal{D}_C) \cup \{a\}) \leq \rho$ ならば a を $A(\mathcal{D}_C)$ に併合し，$A(\mathcal{D}'_C)$ から削除する．併合後の $\mathcal{D}_C$ に対して，この処理を最小の 2 乗誤差が ρ を上回る，または $X(\mathcal{D}'_C) - (X(\mathcal{D}_C) \cap X(\mathcal{D}'_C))$，$A(\mathcal{D}'_C) - (A(\mathcal{D}_C) \cap A(\mathcal{D}'_C)) = \phi$ になるまで繰り返す．処理終了後，1 つも $\mathcal{D}'_C$ の遺伝子，そして観測条件を併合できなかった場合でも，$\mathcal{D}'_C$ は削除する．(2) に該当する場合は，この操作を重複する全てのクラスタに対して行えばよい．最小のクラスタまで調べ終わったら，できた新しいクラスタを降順にソートして，また同様の作業を繰り返す．最終的に重複がない，全てのクラスタが独立して

いる状態になればクラスタの併合と削除は終了する．

$\rho = 1.0$ と設定し，図 2 で示されたクラスタ $1 = \mathcal{D}'_C$，クラスタ $2 = \mathcal{D}_C$ として本手法を適用すると，$\mathcal{D}_C$ に含まれない $\mathcal{D}'_C$ の遺伝子集合 $X(\mathcal{D}'_C) - (X(\mathcal{D}_C) \cap X(\mathcal{D}'_C)) = \phi$，観測条件集合 $A(\mathcal{D}'_C) - (A(\mathcal{D}_C) \cap A(\mathcal{D}'_C)) = \{\mathrm{l}\}$ が分かるので，1 つずつ併合して平均 2 乗誤差を求める．$r(X(\mathcal{D}_C), A(\mathcal{D}_C) \cup \{l\}) \leq \rho$ となり l は $A(\mathcal{D}_C)$ に併合され，$A(\mathcal{D}'_C)$ から削除する．併合後 $A(\mathcal{D}'_C) - (A(\mathcal{D}_C) \cap A(\mathcal{D}'_C)) = \phi$ となるので処理を終了し，$\mathcal{D}'_C$ を削除する．このように本手法を図 2 も含まれる pCluster 結果表 2 を $\mathbf{D}_C$ と考えて適用すると，**表 6** が得られる．適用前と比べると重複がへり，数も減ったことで，それぞれのクラスタの違いを明確に見ることができると考えられる結果となった．

表 6 表 2 の重複除去後のクラスタ
Table 6 Clusters after removal of overlaps from Table 2.

genes	conditions	genes	conditions
g6,g9,g10,g11	a,b,h,i,l	g3,g4,g6,g9,g10	b,c,l
g1,g2,g3,g8	b,c,e,h,j,k	g3,g4,g8	b,h,j,k

このように，併合，削除を繰り返すことによって pCluster で生じてしまう重複，そして数の多さの問題を軽減でき，またぎりぎり pCluster の差の閾値を超えてしまったために取得できなかったクラスタを生成し直すことができると考えられる．本手法では設定する閾値の数が多く，有意水準，そして平均 2 乗誤差の閾値によっては結果が変わってしまう可能性がある．しかし，これは、Biclustering や pCluster など既存のクラスタリング手法にも同様に言えることであり，実験者の理解を助け，計算量を削減できるという観点から見ると，本手法は有用であると言える．また本研究で用いている小規模なデータはもちろん，大規模データから pCluster を得る場合には重複が増えるため特に効果的である．

4. 結果と考察

実際に酵母遺伝子の実験データを使って本手法による実験，そして既存手法との比較を行った．

4.1 実験データ

本文で実験に用いたのはマイクロアレイによって様々な実験状況下における酵母の遺伝子発現量を観測した 2 つのデータである．Gasch *et al.*[6] は，6,152 の酵母遺伝子に対して様々なストレスを与えた 178 の環境における酵母遺伝子の発現量を観測したデータである．

Spellman *et al.*[7] には 5,767 の遺伝子を 82 の観測条件における発現量をそれぞれ測定したデータを使用する．これは細胞周期によって抑制されている遺伝子を解析するために実験され，遺伝子発現量を観測しものである．

また今回本提案手法に用いるある類似機能として，遺伝子オントロジー[5] の Biological process で転写因子をコードする遺伝子，239 個を利用した．

4.2 実　　験

本提案手法が既存手法と比べて改善することができた点，加えて遺伝子オントロジーですでに知られている遺伝子情報から，得られたクラスタの遺伝子が共通の機能を保持しているかという観点から本手法を評価する．

まず始めに事前知識として，Gasch *et al.* には 239 個の転写因子が，Spellman *et al.* には 220 個の転写因子が含まれていた．

そしてこれらから得た pCluster の一例の概要を表にしたものが**表 7** である．抜き出したデータの遺伝子数が少ない場合は，表 7 のように閾値も小さく設定されることになる．重複除去はクラスタの拡張を行ったあとに実行した．平均 2 乗誤差の閾値は 0.1 である．また，$\omega = e^{-1200}$ で計算した．

表 7 転写因子の発現量データから得られた pCluster
Table 7 pClusters by using transcription factor related genes.

data	nc	nr	δ	生成数	重複除去後の数
Gasch *et al.*	6	4	0.15	40	34
Spellman *et al.*	4	4	0.10	109	96

4.2.1 遺伝子機能に関する考察

事前知識を踏まえて，本手法の生物学的知見との一致性を検証するために，表 7 の結果を拡張することによって得られたクラスタと，遺伝子オントロジー[5] の出芽酵母 3 階層目までのタームとの一致度を利用して比較，検証した．1 つは Biclustering によって求められた bicluster と比較した．bicluster により，多くのクラスタを生成することができるが，解析に長時間を要すること，生成されたクラスタに含まれる遺伝子発現量は，乱数で埋められていくので，次第に実際の発現量では発現パターンの類似性が無い部分空間を抽出してしまう可能性が高まり，クラスタの発現類似性が低くなると考えられる．そこで，Biclustering からは，初期に生成されるクラスタを優先して比較した．Biclustering で利用する平均二乗誤差の閾値は，$\rho = 0.1$ を利用した．また，事前知識として正しい知識を用いた場合と，誤っ

Vol.2009-BIO-18 No.7
2009/9/18

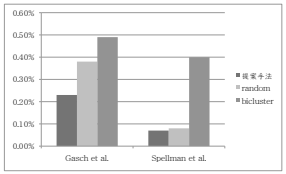


図 4 平均有意水準
Fig. 4 Average significance levels.

た知識を用いた場合の比較として，ランダムに選択した遺伝子を事前知識として用い，同様の解析を行った．**図 4** はそれぞれのデータから各手法によって得られたクラスタを成す遺伝子と遺伝子オントロジーの各タームとの一致度を 2 項検定を使って求めたときの上位 10 クラスタの p 値の平均をグラフにしたものである．ランダムなデータは 10 回の試行によって求めたられた各実験結果の平均値になっている． 図 4 から本手法が既存手法よりも，同じ機能を持つ遺伝子をクラスタとして抽出できていることがわかった．また，ただ小さいデータから得られたクラスタから他の遺伝子へと拡張するのではなく，事前に分かっている知識を用いて本手法を適用することで，クラスタの有意性が高くなっているのだということが示せた．

5. 今後の課題

本論文では遺伝子発現量データにおいて既知の類似機能を持つ遺伝子の発現量のみに着目して pCluster を生成し，その後他の遺伝子との発現パターンの類似性に基づいてクラスタを拡張していくことにより，既知の機能に間接的に関与している機能を持つ遺伝子を同定する方法を提案し，既存手法との比較をおこなった．さらにクラスタ間の重複を減少させて生成クラスタ数をも減らすことで，よりユーザの解釈を助けるような結果を得られることについて述べた．結果，本手法により既存の部分空間クラスタリング手法である Biclustering, そして pCluster のノイズに左右されやすい点やメモリ消費，ユーザの解釈の困難などの当初の問題点を改善することができたと考える．しかし，本提案手法ではまだ既存手法の欠点を補うには足りない部分がある．本提案手法はあくまでも実験データ内にすでに分かっている機能を持つ遺伝子がクラスタを生成できるだけの一定数あることが最低条件であるため，そうでない実験データには適応できない．また，メモリが節約でき，確実に既知の遺伝子情報に沿ってクラスタリングできる一方で，pCluster や biclustering のように網羅的にクラスタリングすることはできない．今後，それらの点についても検討していきたいと考えている．

参 考 文 献

1) Yizong Cheng and George M. Church. Biclustering of Expression Data. *In Proc. of the 8th Int. Conf. Intell. Syst. Mol. Biol.8:93-103,2000.*
2) Haixun Wang, Wei Wang, Jiong Yang, and Philip S. Yu. Clustering by Pattern Similarity in Large Data Sets. *In Proc. of ACM SIGMOD 2002, pages 394-405, 2002.*
3) Jiong Yang, Haixun Wang, Wei Wang, and Philip S Yu. Enhanced Biclustering on Expression Data. *In Proc. of IEEE BIBE 2003, pages 321-327, 2003.*
4) Sungroh Yoon, Christine Nardini, Luca Benini, and Giovanni De Micheli. Enhanced pClustering and Its Applications to Gene Expression Data. *In Proc. of IEEE BIBE 2004, page 275, 2004.*
5) The Gene Ontology Consortium. Gene ontology: tool for the unification of biology. *Nat Genet. 2000 May;25(1):25-9.*
6) Gasch *et al.* Genomic Expression Programs in the Response of Yeast Cells to Enviromental Changes. *Mol Biol Cell. 2000 Dec;11(12):4241-57.*
7) Spellman *et al.* Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol Biol Cell. 1998 Dec;9(12):3273-97.*
8) Hughes *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell. 2000 Jul 7;102(1):109-26.*