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

理学専攻　情報科学コース　大村蓉子　(指導教員：瀬々潤)

1　はじめに

近年，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，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 [1] 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．

2　関連研究

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，最大のものから順にクラスタを出力する．閾値の設定が容易で，互いに重複の少ないクラスタを生成できるが，網羅的な遺伝子発現量からのクラスタ抽出を行う場合，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事から，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．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が大量のメモリを消費する事，クラスタ間に重複が多く重要性の判断が困難な事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集合に対し，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高速に解析が出来，既存の知識に即したクラスタを生成できること，重複を減少させることにより結果の解釈が容易であることにおいて優位性がある．

3　研究内容

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．図 1 で本研究の大まかな流れを示す．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．

3.1　クラスタの拡張

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 $X_S \subset X(\mathcal{D})$ とし，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

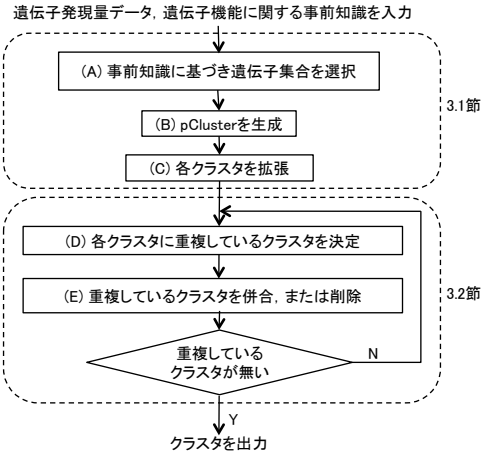

図 1: 提案手法の全体の流れ

$\mathcal{D}_S$ とする (図 1(A))．$\mathcal{D}_S$ から得られる pCluster の集合を $\mathbf{D}_C = \{\mathcal{D}_{C1}, \mathcal{D}_{C2}, \ldots \mathcal{D}_{Cn}\}$ とする (図 1(B))．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，GO [3] 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る．

遺伝子 $g \notin X(\mathcal{D}_C)$ を $\mathcal{D}_C$ に追加しクラスタを拡張し，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．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クラスタが壊れてしまう．ここでは，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 g を選択し，$X(\mathcal{D}_C)$ に追加する．このとき $r(X(\mathcal{D}_C) \cup \{g\}), A(\mathcal{D}_C))$ が閾値 ρ 以下である限り，この走査を繰り返し，$\mathcal{D}_C$ を g で拡張する．これを各クラスタについて行う (図 1(C))．

本手法では，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．

3.2　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，本手法におけるクラスタ生成においても同様である．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，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 $|X(\mathcal{D}_C)| \times |A(\mathcal{D}_C)|$ が

小さいクラスタ $\mathcal{D}'_C$ に注目し，$\mathcal{D}'_C$ が統計的に有意に重複している $\mathcal{D}_C$ を見つける (図 1(D))．クラスタ $\mathcal{D}_C$, $\mathcal{D}'_C$ 間の重複の判定には 2 項検定の式を用いる．求められた p 値が設定した閾値 ω を上回り，有意に重複が認められる場合は，$\mathcal{D}_C$ 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った後．$\mathcal{D}'_C$ を削除する．それ以外の場合には，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，いずれのクラスタにも併合はせず，削除もしない．

重複が存在する場合，3.1 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 $\mathcal{D}'_C$ に含まれて $\mathcal{D}_C$ に含まれない遺伝子 g，そして条件 a を 1 つずつ併合した場合の平均 2 乗誤差を求める．その中で，最小かつ閾値 ρ 以下である 2 乗誤差をとる g，または a を $\mathcal{D}_C$ に併合する．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 を上回るまで繰り返す．処理終了後，1 つも $\mathcal{D}'_C$ の遺伝子，そして条件を併合できなかった場合でも，$\mathcal{D}'_C$ は削除する (図 1(E))．最小のクラスタまで調べ終わったら，また始めから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．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．

4 実験と結果

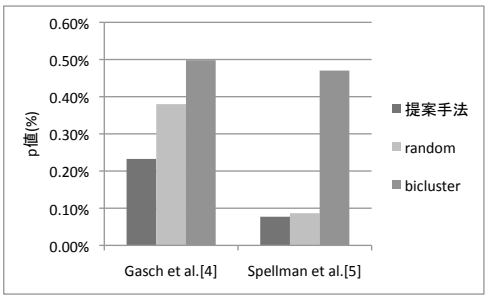

図 2: クラスタが既知の機能に一致する確率

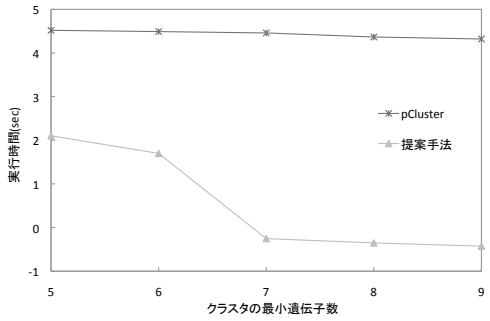

図 3: 実行時間

本文で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 3 つのデータ [4, 5, 6] を用いた．また今回本提案手法に用いるある類似機能として，GO の Biological process で転写因子をコードする遺伝子 239 個を利用した．

図 2 は Biclustering によって求められた bicluster と，事前知識として正しい知識を用いた場合と，誤っ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た結果，そして本提案手法それぞれによって各データ [4] [5] から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られた各実験結果の平均値になっている．グラフ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ことが示せた．

図 3 は Hughes [6] を用いて pCluster と本手法を実行時間について比較した結果である．縦軸は対数目盛を用いた．pCluster がクラスタの大きさに関わらず非常に実行時間が長いのに比べ本手法を用いたところ大幅な時間短縮を実現できたということが分かった．よってメモリの少ない環境下でも本手法によって容易に計算ができることが示せた．

5 まとめ

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，既知の情報に沿うクラスタを生成する方法を提案し，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減らした．その結果，既存手法と比べて生物学的知識に沿うユーザの解釈が容易なクラスタが得られただけでなく，メモリが少ない環境下でも高速に計算することができた．

参考文献

[1] Yizong Cheng, and George M. Church, Biclustering of Expression Data, *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, *Proc. of ACM SIGMOD 2002*, pages 394-405, 2002.

[3] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, *Nat Genet*, 2000 May;25(1):25-9.

[4] Gasch *et al.*, Genomic Expression Programs in the Response of Yeast Cells to Enviromental Changes, *Mol Biol Cell*, 2000 Dec;11(12):4241-57.

[5] Spellman *et al.*, Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. *Mol. Biol. of the Cell*, Vol. 9, Issue 12, pp. 3273-3297, 1998

[6] Hughes *et al.* Functional Discovery via a Compendium of Expression Profiles. *Cell.* 2000 Jul 7;102(1):109-26.