社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－6 (3)
2006／9／15

極大２部クリーク列挙法による遺伝子発現モジュールの抽出

岡田吉史，藤渕航，ホートン・ポール
独立行政法人 産業技術総合研究所 生命情報科学研究センター

遺伝子発現データから，ある特定の条件下において類似の発現動態を示す"遺伝子発現モジュール"を抽出する方法として，Biclustering が注目されつつある．本研究では，極大２部クリークの全列挙法に基づく Biclustering 法（BiModule）を開発した．我々は，これを人工データおよび *S. Cerevisiae* の遺伝子発現データに適用し，既存の代表的な Biclustering 法によるモジュール抽出結果との比較実験を行った．本報告では，BiModule が他手法に比較して，人工的に埋め込まれた Bicluster をより高い精度で検出し，さらには，Gene Ontology における機能アノテーションや既知のタンパク質相互作用を良く反映したモジュールを抽出可能であることを示す．

A biclustering method for finding gene expression modules based on a maximal biclique enumeration

Yoshifumi Okada, Wataru Fujibuchi, Paul Horton
Computational Biology Research Center, International Institute of Advanced Industrial Science and Technology (AIST)

In recent years, biclustering methods have been suggested to discover gene expression modules with shared expression behavior under certain experimental conditions. In this report, we propose a new biclustering method, BiModule, based on a maximal biclique enumeration algorithm. Comparative experiments to existing salient biclustering methods are performed to test the validity of biclusters extracted by BiModule using synthetic data and real expression data. We show that BiModule provides high performance compared to the other methods in extracting artificially-embedded modules as well as modules strongly related to GO annotations and protein-protein interactions.

1 はじめに

最近になって，遺伝子発現データにおける局所的なパタン（発現モジュール）を同定する Biclustering 法が注目されつつある．従来のクラスタリング法では，全ての実験条件（以下，サンプル）にわたって類似の発現パタンを示す遺伝子群をクラスタとみなすのに対し，Biclustering 法では特定のサンプルで類似の発現動態を示す遺伝子サブセットをモジュール（Bicluster）として抽出する．これまで，遺伝子の発現モジュールを発見するための Biclustering 法がいくつか提案されている[1-7]．それらは，特定のサンプルで共発現する遺伝子を組合せ論的に探索する NP 完全な問題を扱うため，準最適な Bicluster をグリーディあるいは確率論的に探索する近似アルゴリズムに基づく．これに対し，我々は，Biclustering を２部グラフからの極大完全グラフ（極大２部クリーク）の抽出問題としてとらえ，準最適解としてではなく，全ての可能な Bicluster を列挙することで遺伝子発現モジュールを抽出する方法（BiModule）を開発した．Bicluster は，遺伝子の頂点集合とサンプルの頂点集合間に枝が張られた２部グラフ上のクリークとみなせる．BiModule は，極大２部クリークと等価なことで知られる飽和アイテム集合（以下，飽和集合）の高速列挙アルゴリズムが実装されている．これにより，全ての可能な Bicluster を現実的な時間内で列挙できるだけでなく，特定の評価関数やクラスタ妥当性指標を用いて，目的に沿った Bicluster を選別できるようになる．また，BiModule では，一定の発現レベルを持つ Bicluster だけでなく，発現変化の傾向をもつ Bicluster を生成する工夫もなされている．

本研究の目的は，人工データおよび実際の発現データを用いて，BiModule と既存手法の性能を統計的に比較評価することである．したがって，抽出された Bicluster に関する生物学的意味についてはここでは議論しない．まず，人工データを用いた実験では，人為的に埋め込まれた正解モジュールの抽出精度を評価する．また，実際の発現データを用いた実

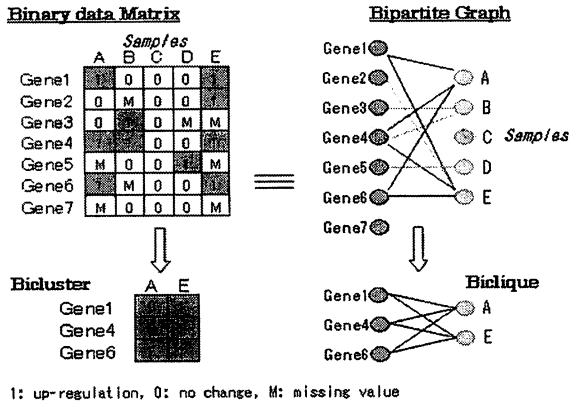

図1：Biclusterと2部クリークの関係

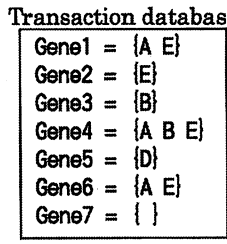

図2：トランザクションデータベース

験では，*Saccharomyces Cerevisiae*(*S. Cerevisiae*)のデータセットを用いて，抽出されたBiclusterがGOアノテーションおよびタンパク質間相互作用をどの程度良く反映しているかを評価する．

本稿の構成は以下のとおりである．まず，2章ではBiclusterと2部クリークとの関係について述べ，3章では極大2部クリークと等価な飽和集合の定義について述べる．4章ではBiModuleのアルゴリズムについて説明し，5章で他手法との比較評価実験の方法と結果，6章を本研究のまとめとする．

2 Biclusterと2部クリーク

Biclusteringへの入力は，一般的に用いられる発現データ行列，すなわち行と列がそれぞれ遺伝子とサンプル，各セルが発現値をとる行列である．既存のBiclustering法のいくつかは，発現値を離散化し，それらを単なる記号データとして扱う[1, 3, 7]．特に，発現値を2値化したデータ行列のBiclusteringは，2部グラフから2部クリークを抽出することと同義になる．図1は，2値データ行列におけるBiclusterと，2部グラフから抽出した2部クリークの関係を示している．ここでのBiclusterは，特定のサンプル集合で1をとる遺伝子の集合で構成される部分行列であり，これは2部グラフにおける遺伝子の頂点集合とサンプルの頂点集合の間で任意の2点間に辺を持つ2部クリークに対応する．他のクリークに包含されない極大2部クリークを列挙することで，極大のBiclusterを列挙できる．極大2部クリークを多項式時間で列挙するアルゴリズムがいくつか提案されている[8,9]．本研究では，実装プログラムが公開されているLCM (Linear time Closed itemset Miner)を用いる[9,10]．これはトランザクションデータベースから飽和集合と呼ばれるパタンを列挙するプログラムとして実装されている．しかしながら，飽和集合の列挙問題は，極大2部クリークを見つけ出す問題と等価であるため[11]，飽和集合を列挙すれば，極大のBiclusterを見つけ出すことができる．次章では，飽和集合について説明する．

3 飽和集合と極大2部クリーク

全アイテムの集合を$I=\{1,\ldots,n\}$，その部分集合をアイテム集合（itemset）と呼ぶ．Dは全トランザクションの集合であり，トランザクションデータベースと呼ぶ．ここで各トランザクションTはIの部分集合である．あるアイテム集合Pを含むDのトランザクションをPの出現という．Pが出現している集合をPの出現集合，その大きさをPの頻度と呼ぶ．ここでアイテム集合Pが，「Pと同じ頻度を持ち，かつ，Pを含むような集合が存在しない」ときPを飽和集合と呼ぶ．与えられた定数θ以上の頻度を持つ飽和集合を頻出飽和集合という．θはサポート値とよばれる．例として，図1の2値データ行列から作成したトランザクションデータ（図2）を用いて飽和集合を説明する．図2の各行が，1つのトランザクション（遺伝子），トランザクションに含まれる個々のアルファベットがアイテム（サンプル）に対応する．アイテム名は，数字や記号など，アイテムを識別できるものとする．アイテムAは，アイテム集合{A}と同じ頻度かつ{A}を含む集合{A,E}が存在するので飽和集合ではない．アイテム集合{A E}は，同頻度かつ，これを包含する集合がないので飽和集合である．

頂点がトランザクションとアイテムに対応し，トランザクションがアイテムを含むときにそれらの間に枝を張ると，2部グラフの極大2部クリークと飽和集合が1対1に対応する．図2の{A, E}のようなサンプルの飽和集合が見つかれば，それらに共通して応答する遺伝子（Gene1, 4, 6）を特定でき，極大2部クリーク，すなわちBiclusterを列挙できる．

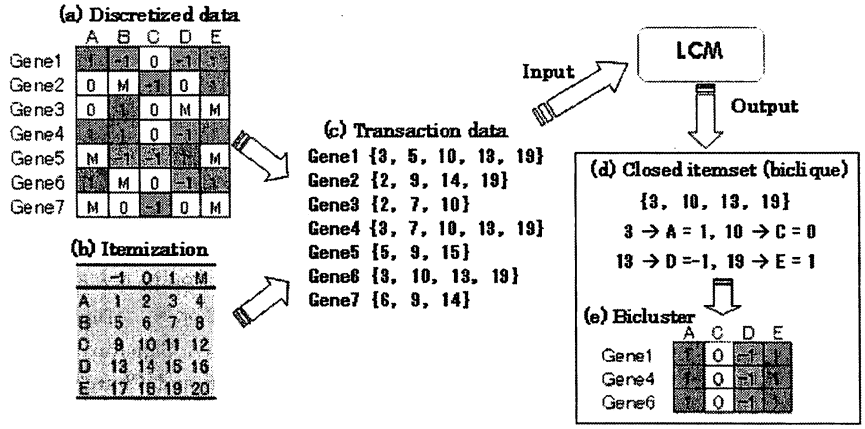

図3：BiModuleの処理手順

4 方法

図3は，BiModuleで行う処理の概略図である．まず，与えられた発現データを正規化し，それらを多段階の階級に離散化する(4.1節)．次に離散データをトランザクションデータに変換する．ここでは，2章で述べたような一定の値をとるBiclusterだけではなく，サンプルにわたって遺伝子が発現傾向を示すBiclusterを抽出できるような変換操作を行う（4.2節）．続いて，LCMアルゴリズムを用いて飽和集合をなすサンプルを列挙する（4.3節）．ここで，他の飽和集合に包含される飽和集合を除外する．最終的に残った飽和集合に関して応答する遺伝子セットを求めてBiclusterを列挙する（4.4節）．以下，それぞれの処理の詳細を説明し，最後にBiModuleにおけるパラメータ設定や実装について述べる（4.5節）．

4.1 データの正規化と離散化

まず，各行に対して発現値が平均0，分散1になるように正規化する．次に，それらの正規化データをいくつかの階級に離散化する．図3aでは簡単のため，階級数を3（-1, 0, 1）とした場合を例示している．ここでデータ行列内のMは欠損値を表す．なお，本研究では7階級（-3, -2, -1, 0, 1, 2, 3）で離散化を行っている．各階級の値を階級値と呼ぶ．階級幅は，データの最大値と最小値の差を階級数で割った値とする．しかしこのとき，最大値や最小値が極端な値（外れ値）をとる場合には，データが多く分布する値域において不当に大きな階級幅がとられてしまい，データに対する離散化が適切に行われない．そこで，離散化の前処理として外れ値の置換処理を行った．まず，反復切断補正法により，正規化データの±2SDまたは±3SD以上のデータを繰り返し除外し，最終的な最大値・最小値を得る．除外された外れ値については，+2(3)SD以上，-2(3)SD以下の値をそれぞれ最終的な最大値，最小値に置き換える．離散化は，発現データに対する正規化，反復切断補正法および外れ値の最大値・最小値への置換の後に行う．

4.2 トランザクションデータの作成

サンプルにわたり遺伝子が発現変化を示すBiclusterを生成可能にするため，Itemizationという処理を行う（図3b）．これは，各サンプルの全ての階級値にIDをつけ，どのサンプルでどの階級値を持つかをアイテム化する処理である．ここで，行はサンプル名，列は階級値および欠損値である．例えば，アイテム7は，サンプルBで階級値が1であることを表している．続いて，各遺伝子をトランザクション，各サンプルの階級値に付されたIDをアイテムとして，データ行列をもとにトランザクションデータベースを作成する（図3c）．このとき，欠損値（M）を示すアイテムはトランザクションデータベースには含めない．Itemization処理によって，多段階の発現レベルをトランザクションデータに反映できるため，一定の値をとるBicluster だけでなく，発現変化を示すBiclusterの抽出が可能となる．

4.3 サンプルに関する飽和集合列挙

LCMへの入力は，トランザクションデータファイルとサポート値である．サポート値はBiclusterを構成する遺伝子の最小数である．出力は，図3dに示されるようなアイテムの飽和集合である．ここでは，{3, 10, 13, 19}が飽和集合として出力されている．これをサンプル名に直すには，Itemization表を参照する．すると，このサンプル集合は{A, C, D, E}であり，それらの階級値は{1, 0, -1, 1}であることがわかる．

以下，LCMアルゴリズムの概要を説明する．アルゴリズムに関する詳しい説明や性能については，文献[9]を参照されたい．

LCMでは，prefix保存飽和拡張という手法を用いて，飽和集合の探索を高速化する．飽和集合Pに対し，そのコアアイテム core(P_i)を $clo(P_{\leq i})=P$ が成り立つような最小添え字を持つアイテムとする．ここで，$clo(P_{\leq i})$は，i以下の添え字を持つPの閉包（出現集合の共通部分）である．このとき，P'がPのprefix保存飽和拡張であるとは，あるアイテム $e > core(P_i)$ に対して $P'= core(P\cup\{e\})$, $P'_{\leq e-1}=P_{\leq e-1}$ が成り立つことである．P_0 を，全てのトランザクションに含まれる集合の中で極大なものとする．ただし，P_0 は空集合にもなりえる．P_0 でない任意の飽和集合は，自身より頻出度が大きい飽和集合のprefix保存飽和拡張であり，さらに，そのような飽和集合は唯一に決まる．したがって，P_0 から出発して再帰的にprefix保存飽和拡張を生成することで，全ての頻出飽和集合を深さ優先で発見できる．

LCMを実行することにより，発見されたサンプルの飽和集合が逐次的に列挙される．

4.4 フィルタリング処理とBicluster出力

LCMに入力するデータサイズが大きい場合には，非常に膨大な数の飽和集合が生成される．しかしながら，そのほとんどは他の飽和集合に包含されているか，あるいは大部分が重複したものである．そこで，出力された飽和集合を以下のスコアSに基づいてソートした後で，上位の飽和集合に80%以上包含される飽和集合を除外する．

$$S = A \times \log_2(g) \times \log_2(s)$$

ここで，Aは飽和集合のサンプルが持つ階級値の絶対値の平均値，gは飽和集合のサンプルに応答する遺伝子数，sは飽和集合に含まれるサンプル数である．これにより，サイズがある程度大きく，他に大部分を包含されないBiclusterを取り出せる．また，階級値の絶対値の平均を考慮することで，平均的に発現レベルの低いBiclusterを排除できる．最終的に残った飽和集合に関して応答する遺伝子セットを求めてBiclusterを列挙する（図3e）．

4.5 実装とパラメータ設定

BiModuleは，LCMによる飽和集合列挙の部分を除き，全てJava言語で開発されている．なお，LCMはC言語で実装されている．BiModuleの入力は，正規化前の発現データ行列および，抽出するBiclusterサイズに関するパラメータ M_g とMsである．ここで M_g とMsはそれぞれ，出力されるBiclusterの遺伝子数およびサンプル数に関する最小値である．

5 評価方法および結果

5.1 比較に用いたBiclustering手法

Prelicらは，6つの代表的なBiclustering法を，1）人為的に正解モジュールを埋め込んだデータセット，および2）実際の発現データセットに適用し，生成されるBiclusterの妥当性評価を行った[1]．ここで用いられたデータセットおよびBiclustering結果は[12]で公開されている．Prelicらが比較に利用した6つの手法を以下に示す：Divide-and-conquer Algorithm[1]，Iterative Signature Algorithm[2]，SAMBA[3,4]，Cheng and Church's Algorithm[5] Order Preserving Submatrix Algorithm[6]，xMotif[7]．以下本稿では，それぞれBiMax，ISA，SAMBA，CC，OPSM，xMotifと呼ぶ．本研究では，Prelicらによって提供されている人工データと実際の発現データにBiModuleと上記の6手法を適用した結果を比較する[12]．なお，BiModuleに入力したパラメータは M_g=7，Ms=4である．

5.2 人工データを用いた評価

人為的にモジュールを埋め込んだデータを用いて，BiModuleおよび既存手法がどの程度正確にそれらのモジュールを抽出できるかを，「遺伝子適合スコア」と呼ばれる指標に基づいて評価する．ここでは，以下の3つのタイプのモジュールが埋め込まれたデータ行列を扱う．

・Constantモジュールを含むデータ行列
・Coherentモジュールを含むデータ行列
・Overlappingモジュールを含むデータ行列

Constantモジュールとは全てのセルが同一の値を持つモジュールであり，Coherentモジュールとはサンプルにわたって値が変化し，全ての遺伝子が同じ発現傾向を示すモジュールである．ConstantおよびCoherentタイプではともに，100遺伝子×50サンプルのデータ行列内に，10遺伝子×5サンプルのモジュールが10個含まれている．なお，10個のモジュールは互いに重複しないものとする．

Overlappingモジュールは，データ行列内の対角線上に並ぶ10個のCoherentモジュールが重複度dで重なり合っているモジュールである．d=0はモジュール間で全く重なりが無い場合，d=10は隣接する2つのモジュールが遺伝子側に10，サンプル側に10の重複（すなわち10×10=100セル分）を持つことを意味する．

抽出されたBiclusterの妥当性を評価するため，以下の遺伝子適合スコアを用いる．

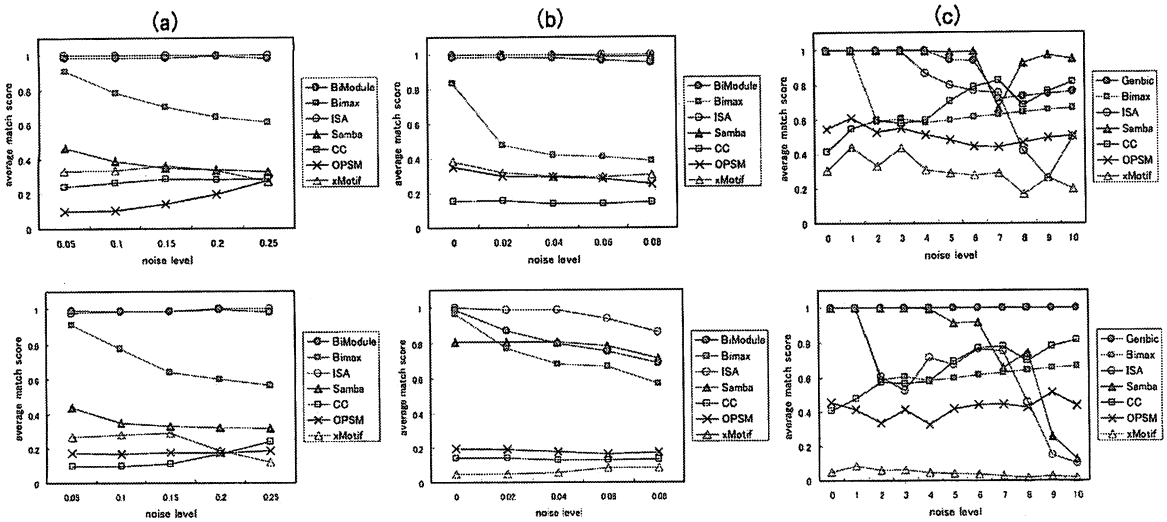

図4：遺伝子適合スコア

$$GMS_G(M_1,M_2)=\frac{1}{|M_1|}\sum_{(G_1,S_1)\in M_1}\max_{(G_2,S_2)\in M_2}\frac{|G_1\cap G_2|}{|G_1\cup G_2|}$$

ここで，M_1，M_2はモジュール（またはBicluster）のセット，G，Sはそれぞれ1つのモジュールに含まれる遺伝子セットおよびサンプルセットである．このスコアは，M_1の各モジュールについて，M_2に含まれるモジュールとの最大適合スコアの平均値を表わしている．今，M_{opt}を人為的に埋め込まれたモジュールセット（正解セット），Bをある Biclustering 法から出力された Bicluster セットとすると，$GMS_G(B, M_{opt})$は出力されたBiclusterと正解モジュールセットとの一致度を表し，$GMS_G(M_{opt}, B)$は正解セットをもれなく抽出する再現度を意味する．どちらのスコアも$M_{opt}=B$のときに最大値1をとる．

ここでは，各 Biclustering 法の「データに含まれるノイズに対する感度」と「ノイズを含まないが重複しているモジュールの抽出精度」の2つの観点から評価した結果を示す．図4は，ノイズを含む(a)Constant および(b)Coherent モジュール，(c)ノイズを含まない Overlapping モジュールに関する遺伝子適合スコアの一致度（上段）と再現度（下段）を示している．(a)と(b)については，正規分布から得られるランダムな値をノイズとして，元のデータ行列に付加した．図(a)，(b)および(c)は異なるノイズレベル（標準偏差）および重複度（d=0,1,...,10）において，それぞれ10個の異なるデータ行列を生成し，それらに対するスコアの平均値をプロットしている．

Constant モジュール（図4a）では，BiModule と ISA はノイズの増加の影響をほとんど受けず，一致度と再現度はともに98%以上を維持している．続くBiMaxはノイズの増加とともに正解モジュールの抽出精度が，一致度，再現度ともに 91%（noise level=0.05）から 60%程度（noise level=0.25）に減少している．SAMBA，CC，OPSM，xMotif に関しては全てのノイズレベルにわたり，一致度，再現度ともに50%以下であった．Coherent モジュール（図4b）では，一致度については，BiModule，ISA，そしてSAMBAが全てのノイズレベルにわたって95%以上の精度を示した．BiMax は，Constant モジュールの場合よりもノイズの影響を強く受けている．再現度については，BiModule，BiMax，ISA，SAMBA が上位グループを形成しているが，中でも ISA が最も高く安定した精度を示した．Overlapping モジュール（図4c）では，一致度，再現度ともに BiModule が良好な精度を示している．特に，再現度に関して言えば，重複度の増加に関わらず全てのモジュールをもれなく抽出できている．SAMBA は一致度に関しては良好な成績だが，再現性では重複度の増加とともに急激な抽出漏れが見受けられる．ノイズに最も頑健であったISAは，重なりが少ない（d=1 or 3）場合はどちらのスコアも 100%を示すが，重複度が増すにつれ抽出精度は減少し，d=9 では一致度，再現度ともに 20%代まで落ち込んでしまう．

全体をとおして，BiModule が最も安定かつ高い精度でモジュールの抽出を行っている．これは，BiModule における以下の操作が有効に働いていると思われる．BiModule は，4章で述べた Itemization によって全てのセル値が同一な Bicluster（Constant モジュール）だけでなく，サンプルにわたる発現変化の傾向を持つ Bicluster（Coherent モジュール）を抽出できる．また，離散化により，微弱なノイズを含む値や外れ値の影響を取り除くため，ノイズに

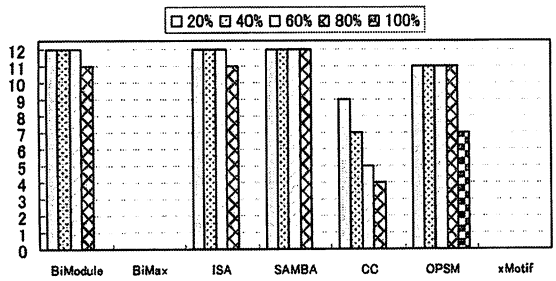

図5：有意なGOカテゴリ(p<0.001)を含むBicluster数

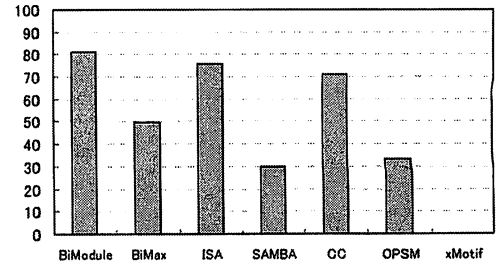

図6：タンパク質間相互作用に関わる遺伝子ペアが有意に含まれるBiclusterの割合

対する感度を抑えることができる．さらに，BiModuleでは，Biclusterに対してサンプル数と遺伝子数のバランスを考慮したフィルタリングを行うため，CCやxMotifのようにどちらか一方の次元のみに拡張したBiclusterは抽出されない．

5.3 遺伝子発現データを用いた評価

ここではGasch ら[13]によって提供されている *S. Cerevisiae* の遺伝子発現データセットを用いる．これは，さまざまなストレス条件下で測定された173サンプルに関する2993遺伝子の発現データを含む．我々は，このデータセットに個々の手法を適用し，得られたBiclusterが，Gene Ontology (GO)アノテーションと既知のタンパク質間相互作用をどの程度反映しているかについての評価を行う．なお，BiModuleに入力したパラメータは，M_g=80, Ms=5である．通常，Biclusteringのアルゴリズムによって，生成されるBicluster数は異なる．そこで，出力Bicluster数が100を超える場合には，それらのサイズが大きい順に100個を最適なBiclusterとみなして実験を行った．結果として，各手法が生成したBicluster数は，BiModule (20)，BiMax (100)，ISA (67)，SAMBA (100)，CC (100)，OPSM (12)，xMotif (100) であった．

5.3.1 Gene Ontology(GO)アノテーション

Bicluster内の遺伝子について統計的に有意に多く含まれるGOカテゴリを調べるため，フリーで利用可能なGO解析ツールBiNGOを用いた[14]．図5は，有意と判定されたGOカテゴリ (p<0.001) を含むBiclusterの個数を，それらのカテゴリを含む割合別 (20, 40, 60, 80, 100%) に示したものである．ただし，ここではBicluster数が最も少ないOPSMに合わせ，各手法においてBiclusterのサイズが大きい順に12個とりだした場合を示してある．図から，BiModule, ISA, SAMBA, OPSMの4手法が，GOアノテーションを良く反映したBiclusterを構成していることがわかる．BiModule, ISAについては，11個のBiclusterにおいて80%以上，SAMBAについては，全てのBiclusterで80%以上の遺伝子が有意なGOカテゴリを有している．これらの3手法は，人工データからのモジュール抽出においても好成績を示していた．一方，OPSMは，人工データに対してはノイズの影響を受けやすいだけでなく，重複したモジュールの抽出精度も比較的低かった．しかし現実のデータにおいては，11個のBiclusterにおいて80%以上，7個のBiclusterにおいて全ての遺伝子が有意なGOカテゴリを持っていた．CCに関しては，上記の4手法に比較して有意なGOカテゴリを含むBiclusterが少なく，それらのカテゴリを持つ遺伝子の割合も低かった．これは，CCで得られるサイズの大きいBiclusterは，発現値が0のセルを多く含むため，生物学的に意味のある遺伝子が含まれにくいことが考えられる．BiMaxとxMotifでは，有意なGOカテゴリを持つBiclusterは存在しなかった．BiMaxでは発現データを2値データに離散化してしまうため，データの持つ情報が失われ，生物学的に関連性の無いBiclusterが多数生成されている可能性がある．また，xMotifでは，得られるBiclusterのほとんどが1つの遺伝子しか含まなかったため，有意なGOカテゴリは見つからなかった．

5.3.2 タンパク質間相互作用

Biclusterとタンパク質間相互作用との関連性を調べるため，DIP データベース[15]で提供される相互作用データを用いた．このデータには，14,742組のタンパク質間相互作用が，遺伝子のペアで記述されている．与えられたBiclusterがどの程度タンパク質間の相互作用を反映しているかを見積もるスコアとして，「Bicluster 内において相互作用していない遺伝子ペアの割合」を用いる．このスコアは，ランダムに選ばれた遺伝子グループよりも，Bicluster内における遺伝子群に関して有意に小さい値をとることが期待される．そこで，各手法で生成された個々のBiclusterと同じサイズのランダムな遺伝子グループを1000セットずつ作成し，それぞれのBiclusterに

関するスコアがランダムな遺伝子グループよりも有意に小さいかどうかを Z-検定により判定した．図6は，ランダムな遺伝子グループに対し，スコアが有意に小さい Bicluster($p<0.001$)の割合である．BiModule は81%の Bicluster においてランダムな遺伝子グループに対して有意にスコアが小さく，7手法の中で最も良くタンパク質間相互作用を反映していた．これに続き，ISA が75.8%，CC が71%と好成績を示している．CC については，サイズの小さい Bicluster で有意差が見られる傾向にあった．

6. まとめ

本研究では，飽和集合列挙アルゴリズムに基づいて，遺伝子とサンプル間で形成される極大2部クリークを全て列挙し，他に包含されない Bicluster を抽出する BiModule を提案した．これを人工データおよび *S. Cerevisiae* の発現データに適用し，代表的な6つの手法との比較を統計的観点から行った．以下は，本研究で得られた結果のまとめである．

1. 人為的に埋め込まれたモジュールの抽出実験では，全体をとおして BiModule，ISA，SAMBA が好成績であった．特に，BiModule はノイズの影響を受けにくいだけでなく，重複度合いによらず，重なり合ったモジュールを適切に抽出した．概して，BiModule が最も安定して高い精度を示した．
2. *S. Cerevisiae* の発現データを用いた実験では，総合的に見て，BiModule と ISA が GO アノテーションおよびタンパク質間相互作用との関連度が高い Bicluster を生成した．

以上の結果から，BiModule が遺伝子発現モジュールの発見に有効であることが示された．

BiModule は，大きく分けて1）データの正規化と離散化，2）トランザクションデータベースの作成，3）飽和集合列挙，4）フィルタリングの4つの部分に分けられる．しかしながら，本稿では，これらのどの部分がモジュール発見の性能を左右しているかを明らかにしていない．例えば，離散化における階級数，フィルタリングでのスコア関数，あるいは入力パラメータの最小遺伝子数や最小サンプル数によって，モジュール発見精度にどのような違いが見られるかを明確にする必要がある．また，本研究では統計的指標に基づく評価に留まったが，今後は，抽出されたモジュールに含まれる遺伝子の機能や転写制御，あるいは分子間ネットワークといった異なるレベルにおける生物学的知見との照らし合わせから評価を行っていく．

参考文献

1. Prelic, A. et. al. (2006) A Systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics, 22, 1122-1129.
2. Ihmels,J. et al. (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics, 20, 1993–2003.
3. Tanay,A. et al. (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18 (Suppl. 1), S136–S144.
4. Tanay,A. et al. (2004) Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneous genomewide data. Proc. Natl Acad. Sci. USA, 101, 2981–2986.
5. Cheng,Y. and Church,G. (2000) Biclustering of expression data. Proc. Int. Conf. Intell. Syst. Mol. Biol., 93–103.
6. Ben-Dor,A., Chor,B., Karp,R. and Yakhini,Z. (2002) Discovering local structure in gene expression data: the order-preserving sub-matrix problem. Proc. of the 6th Annual Int. Conf. on Computational Biology, ACM Press, New York, NY, USA, 49–57.
7. Murali,T.M. and Kasif,S. (2003) Extracting conserved gene expression motifs from gene expression data. Pac. Symp. Biocomput., 8, 77–88.
8. Tukiyama, S. et al. (1977) A new algorithm for generating all the maximum independent sets, SIAM Journal on Computing, 6, 505-517.
9. Uno, T. et al. (2004a) An efficient algorithm for enumerating closed patterns in transaction databases, Lecture Notes in Artificial Intelligence, 3245, 16-31
10. http://research.nii.ac.jp/~uno/codes-j.html
11. Boros, E. et al. (2002) On the complexity of generating maximal frequent and minimal infrequent sets. Proc. of STACS 2002, Springer, Berlin, 133-141.
12. http://www.tik.ee.ethz.ch/sop/bimax/SupplementMaterials,Biclustering.html
13. Gasch,A.P. et al. (2000) Genomic expression programs in the response of yeast cells to environmental changes. Mol. Biol. Cell, 11, 4241–4257.
14. Maere, S. et al. (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in biological networks. Bioinformatics 21, 3448-3449.
15. http://dip.doe-mbi.ucla.edu/