統計数理(2006)
第 54 巻 第 2 号 315–331

特集「予測と発見」
[総合報告]

グラフマイニングとその統計的モデリングへの応用

鷲尾 隆[1,2] ・樋口 知之[1] ・井元 清哉[3] ・玉田 嘉紀[4] ・佐藤 健[5] ・元田 浩[2]

(受付 2006 年 4 月 5 日; 改訂 2006 年 6 月 12 日)

要　　旨

本報では，はじめにデータマイニング分野において近年盛んに研究されているグラフマイニングの背景，研究経緯，関連研究等を概観し，次に部分グラフのクラス，部分グラフ同型問題，正準ラベル，マイニングの基準など，グラフマイニングを理解する上で重要な幾つかの基礎概念を説明する．更に，そこで必要とされる多頻度アイテム集合や多頻度グラフの探索原理と代表的手法について解説する．最後に，マイクロアレイ遺伝子発現プロフィールデータから遺伝子発現関係をベイジアンネットワークで同定した結果に，更にグラフマイニングを適用して各遺伝子発現の依存関係に関する知見を得る解析を報告し，統計的モデリングへの応用可能性について述べる．

キーワード：グラフマイニング，部分グラフ同型問題，Downward Closure Property，多頻度グラフ，ベイジアンネットワーク．

1. はじめに

一般にデータマイニングは，膨大な表形式ないしはトランザクション形式データから，データが内蔵する有用な部分的特徴を発掘する手法であると考えられている．更に近年では，対象とされるデータが半構造テキストや木，記号系列，グラフ，論理関係など，複雑な構造を持つものにまで拡大されつつある．その中でも特にグラフは，数学における基本的な研究対象構造であり，言語論理とも強い関係をもつ．また，統計数理や機械学習においてもグラフィカルモデリングやベイジアンネットワーク，ニューラルネットワークなど，グラフ構造を有するモデルが頻繁に用いられる．更に生物学や化学，材料化学，社会通信ネットワークなど様々な実分野においてもグラフ構造データは幅広く見られる．しかし一方で，膨大なグラフ構造データから特徴的部分構造を見つける問題の多くが，本質的計算困難性を有することが知られている．たとえば，ある大きなグラフに別のより小さなグラフが部分グラフとして含まれているか否かを調べる問題は NP-完全である(Garey and Johnson, 1979)．このような背景から，近年，グラフ構造データを対象として有用な知識を効率的に発掘するグラフマイニング手法が盛んに研究

[1] 統計数理研究所：〒106–8569 東京都港区南麻布 4–6–7
[2] 大阪大学 産業科学研究所：〒567–0047 大阪府茨木市美穂ヶ丘 8–1
[3] 東京大学医科学研究所：〒108–8639 東京都港区白金台 4–6–1
[4] 統計数理研究所(現 株式会社ジーエヌアイ創薬解析センター：〒814–0001 福岡市早良区百道浜 3–8–33–608)
[5] 国立情報学研究所：〒101–8430 東京都千代田区一ツ橋 2–1–2

されるようになって来た．

グラフマイニング研究の発端は，大規模なグラフから何らかの基準の下で特徴的な部分グラフを発掘するヒューリスティック探索に関する，1990 年代半ばの S UBDUE（Cook and Holder, 1994）及び GBI（Yoshida et al., 1994）の研究がある．1999 年になって，多頻度部分グラフの完全探索を目指す WARMR が発表された（Dehaspe and Toivonen, 1999）．2000 年にはトランザクション形式データに関するデータマイニングアルゴリズムである Apriori アルゴリズムがグラフ理論によって拡張され，高速に多頻度部分グラフの完全探索を行う AGM が発表された（Inokuchi et al., 2000）．これらの先駆的研究の後，グラフマイニング研究は急速に盛んになった．

より効率的に多頻度部分グラフを完全探索する手法としては，部分グラフ同型問題を効率的に解くために新たなグラフ不変量を導入した FSG（Kuramochi and Karypis, 2001），部分グラフ同型の効率的深さ優先探索のために DFS 木を用いる gSpan（Yan and Han, 2002），グラフデータ中の多頻度連結部分グラフを探索する AcGM（Inokuchi et al., 2002），自由木探索を拡張して疎グラフデータ中から非常に高速に多頻度グラフを発掘する Gaston（Nijssen and Kok, 2004）などが提案されている．一方，発掘しようとする部分グラフの種類を拡張する手法としては，帰納推論データベースの枠組みを用いて与えられた単調性条件を満足する部分パスの集合（バージョン空間）をグラフデータから発掘する MolFea（de Raedt and Kramer, 2001），多頻度閉部分グラフを探索する CloseGraph（Yan and Han, 2003），与えられた単調性条件を満足する部分自由木の集合をグラフデータから発掘する FreeTreeMiner（Ruckert and Kramer, 2004），1 枚の大きな疎グラフ中から互いに重ならない多頻度連結部分グラフを発掘する SiGraM（Kuramochi and Karypis, 2004），グラフデータ中から極大な多頻度連結部分グラフを発掘する SPIN（Huan et al., 2004），階層的（Taxonomy）なラベル付けを有するグラフデータ中の多頻度連結部分グラフを探索する Generalized AcGM（Inokuchi, 2004），部分グラフ同型探索に様々な制約を導入してグラフに限らずデータ中に埋め込まれた多頻度の部分パスや部分木の発掘を可能にした B-AGM（Inokuchi et al., 2005）など，様々なものが提案されている．これらの内，比較的時期の早い手法については文献（Washio and Motoda, 2003）が詳しい．この他にもグラフデータから種々の部分グラフを発掘する手法が提案されており，ライデン大学のホームページ “Homepage for Mining Structured Data” で最新の手法を含めた紹介や比較を見ることができる（Nijssen, 2005）．

本報では，以下にグラフマイニングを理解する上で重要な部分グラフ，同型問題，グラフ不変量，マイニング基準といった基礎を説明する．その後，グラフマイニングで必要とされる探索原理とそれを用いる代表的手法について解説する．最後に遺伝子発現の因果関係に関するベイジアンネットワークを用いた統計的モデリングへの応用可能性について述べる．

2. グラフマイニングの基礎

グラフマイニングの背景には，グラフ理論や探索理論に関する豊富な研究が存在している．ここでは，多くのグラフマイニング手法を理解する上で必要となる幾つかの基礎原理を，閉路や並行路を含むラベル付き無向グラフの場合について説明する．有向グラフやラベル無しグラフについては説明を省略するが，同様な原理が成り立つ．

2.1 一般部分グラフと誘導部分グラフ

グラフ $G(V, E, f, \ell)$ は，頂点の集合 V，頂点ペアを結ぶ辺の集合 E，辺による頂点の接続関係を表す関数 $f : E \to V \times V$，頂点や辺のラベル付け関数 ℓ の 4 項組で表される．頂点のラベル集合を L_v，辺のラベル集合を L_e とした時，ℓ は更に $\ell_v : V \to L_v$ 及び $\ell_e : E \to L_e$ の 2

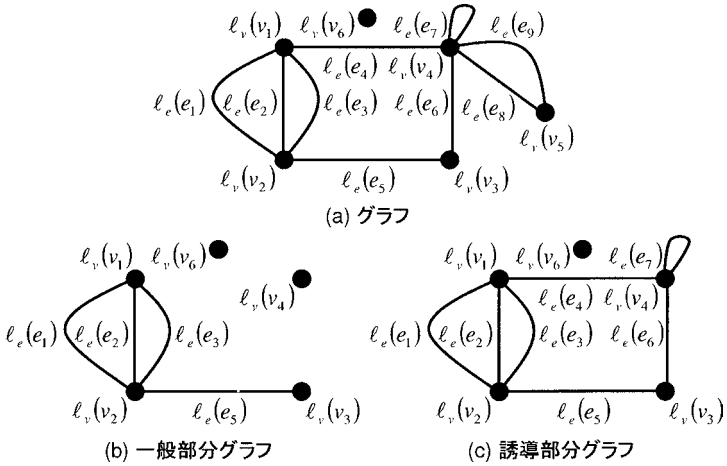

図 1. グラフと部分グラフの例.

項組 $\ell(\ell_v,\ell_e)$ **で表される．例えば，図** 1(a)**に示されるグラフでは，** $V=\{v_1,v_2,v_3,\ v_4,v_5,v_6\}$**，** $E=\{e_1,e_2,e_3,e_4,e_5,e_6,e_7,e_8,e_9\}$ **となる．** E **に含まれる各辺** e_h **は，** V **に含まれる** v_i **と** v_j **を** $f(e_h)=(v_i,v_j)$ **によって関連づける．図** 1 **の場合には，例えば** $f(e_1)=(v_1,v_2)$**，** $f(e_2)=(v_1,v_2)$**，** $f(e_4)=(v_1,v_4)$**，** $f(e_7)=(v_4,v_4)$ **となる．また，** V **に含まれる各頂点** v_i**，** E **に含まれる各辺** e_h **は，それぞれ** ℓ_v **と** ℓ_e **によってラベル** $\ell_v(v_i)$ **と** $\ell_e(e_h)$ **を有する．**

グラフ $G(V,E,f,\ell)$ **の最も一般的な部分構造クラスは“ 一般部分グラフ ”であり，それは**

(1) $V_s\subset V$ and $E_s\subset E$**,**
(2) $\forall e_h\in E_s,\ell_{se}(e_h)=\ell_e(e_h)$ and $v_i,v_j\in V_s$ where $f_s(e_h)=(v_i,v_j)$**,**
(3) $\forall v_i\in V_s,\ell_{sv}(v_i)=\ell_v(v_i)$

を満たす 4 **項組** $G_s(V_s,E_s,f_s,\ell_s)$ **で定義される．図** 1(b)**は，元グラフ図** 1(a)**の頂点** v_5 **及び辺** e_4,e_6,e_7,e_8,e_9 **が無い一般部分グラフの例である．もう** 1 **つの代表的な部分構造クラスは“ 誘導部分グラフ ”であり，一般部分グラフの条件に加えて**

(4) $\forall e_h\in E$**,** $e_h\in E_s$ if $v_i,v_j\in V_s$ where $f(e_h)=(v_i,v_j)$

である 4 **項組** $G_s(V_s,E_s,f_s,\ell_s)$ **によって定義される．図** 1(c)**は，元グラフ図** 1(a)**から頂点** v_5 **を除いた誘導部分グラフの例である．この場合，** v_5 **に直接繋がる** e_8 **と** e_9 **は含まれないが，図** 1(b)**と異なり元グラフ** G **の** v_1,v_3,v_4 **間に存在する** e_4,e_6,e_7 **は含まれる．ここでは，** G_s **が** G **の一般ないしは誘導部分グラフであることを** $G_s\subseteq G$ **と表す.**

2.2 部分グラフ同型問題

グラフマイニングにおいては，多数のグラフ間で共通する部分グラフを探索するため，グラフ理論の“ 部分グラフ同型問題 ”を拡張した定義を用いる．今，グラフ集合 $D=\{G_d(V_d,E_d,f_d,\ell_d)|d=1,\ldots,n\}$ **について，以下のようなあるグラフ** $G_s(V_s,E_s,f_s,\ell_s)$ **及び** V_s **から** V_d $(d=1,\ldots,n)$ **への単射** g_{sd} **を見つける問題を，“ 部分グラフ同型問題 ”という.**

(1) $\forall G_d(V_d,E_d,f_d,\ell_d)\in D,G_s(V_s,E_s,f_s,\ell_s)\subseteq G_d(V_d,E_d,f_d,\ell_d)$,
(2) $\forall d=1,\ldots,n,\forall e_{sh}\in E_s$,
$\exists e_{dh}\in E_d$**,** $f_s(e_{sh})=(v_{si},v_{sj})$ and $f_d(e_{dh})=(g_{sd}(v_{si}),g_{sd}(v_{sj}))$**.**

条件(1)において，$G_s(V_s, E_s, f_s, \ell_s)$ は一般部分グラフ，誘導部分グラフのいずれの定義を取ることも可能である．例えば，図 1 のグラフ(b)と(c)からなる $D = \{(b), (c)\}$ について，部分頂点集合 $V_s = \{v_{s1}, v_{s2}, v_{s3}\}$ と部分辺集合 $E_s = \{e_{s1}, e_{s5}\}$ からなる部分グラフ $G_s(V_s, E_s, f_s, \ell_s)$ と単射 $v_{di} = g_{sd}(v_{si})$ $(i = 1, 2, 3, d = (b), (c))$ は一般部分グラフの場合の上の条件を満たす．即ち，図 1(b) と(c)は G_s を共有し，G_s は D について一般部分グラフ同型である．これに対して，同じく部分辺集合 $E_s = \{e_{s1}, e_{s2}, e_{s3}, e_{s5}\}$ の場合は，誘導部分グラフの場合の上の条件を満たし，G_s は D について誘導部分グラフ同型である．1 つの小さなグラフが 1 つのより大きなグラフの部分かどうかを判定する部分グラフ同型問題は，NP-完全であることが判っている(Garey and Johnson, 1979)．上記の複数グラフ間の部分グラフ同型問題の複雑性は NP-完全より低いことはあり得ない．

2.3 正準ラベル

同型なグラフは等しいグラフ不変量値を持つ．グラフ不変量の例として，グラフに含まれる頂点数や各頂点に接続する辺数(線度)，閉路の数などがある．しかし，不変量値が等しくても同型なグラフとは限らない．これに対して，最も直接的にグラフ構造を表す不変量として，以下の“ 正準ラベル ”がある．同型なグラフは等しい正準ラベルを持ち，正準ラベルが等しければ同型なグラフとなる．

グラフ G の i 番目の頂点 v_i を i 番目の行と列に対応させ，要素によって頂点間の辺の接続関係を表す行列を“ 隣接行列 ”という(Inokuchi et al., 2000, 2003)．i, j 要素は，辺のラベル集合 $\{\ell_e(e_h) | \forall e_h \in E \text{ where } f(e_h) = (v_i, v_j)\}$ で表される．これは厳密には要素が数ではないので行列ではないが，都合上行列と呼ぶ．頂点 v_i と v_j 間に辺が存在しない場合，要素は空ないし 0 とする．以下は図 1 (a)の隣接行列である．

$$\begin{pmatrix}
 & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 \\
v_1 & 0 & \{\ell_e(e_2), \ell_e(e_3), \ell_e(e_4)\} & 0 & \{\ell_e(e_1)\} & 0 & 0 \\
v_2 & \{\ell_e(e_2), \ell_e(e_3), \ell_e(e_4)\} & 0 & \{\ell_e(e_5)\} & 0 & 0 & 0 \\
v_3 & 0 & \{\ell_e(e_5)\} & 0 & \{\ell_e(e_6)\} & 0 & 0 \\
v_4 & \{\ell_e(e_1)\} & 0 & \{\ell_e(e_6)\} & \{\ell_e(e_7)\} & \{\ell_e(e_8), \ell_e(e_9)\} & 0 \\
v_5 & 0 & 0 & 0 & \{\ell_e(e_8), \ell_e(e_9)\} & 0 & 0 \\
v_6 & 0 & 0 & 0 & 0 & 0 & 0
\end{pmatrix}.$$

そして，無向グラフ G に関して，その $n \times n$ 隣接行列の各 i, j 要素 $x_{i,j}$ から次のようなコード表現を定義する．

$$code(G) = x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3} \cdots x_{1,n} \cdots x_{n-1,n}x_{n,n}.$$

無向グラフの隣接行列の対角対称性より，上三角部分の要素のみで表される．下三角部分の要素も使って，同様なコードを有向グラフの場合にも定義できる．各要素にその内容の辞書順に番号を振った時，あるグラフに一意に対応する正準ラベルは，コード上の要素の並びで番号を並べて得られる数字が最小(あるいは最大)のコードとして定義される．そして，そのコードに対応する隣接行列を“ 正準形 ”と呼ぶ．正準ラベルと正準形によって，グラフ表現の多様性や同型問題の探索空間は著しく削減される．

2.4 マイニング基準

データマイニングでは，ある基準を満たすデータ部分に着目して部分的特徴を発掘する．代表的基準のクラスとして，“ 上半束 ”と呼ばれる集合族に対して定義される “Downward Closure Property (DCP)” に基づくものがある．“ 上半束 ”を成す集合の概念は広い．L を“ 結合演算 ” $\cup$ を持つ“ 集合族 ”とし，L に属する 2 つの集合 a, b の“ 上界 ”を“ 結合 ” $a \cup b$ と定める．こ

こで結合は以下の 2 つの規則に従うものとする．

$$a \cup b = b \cup a \qquad \text{（可換則）}$$
$$a \cup (b \cup c) = (a \cup b) \cup c \quad \text{（結合則）}$$

L に属する任意の 2 つの集合 a,b について $a \cup b$ が L に属する，即ち L が結合について閉じている時，L を“ 上半束 ”という．道路網や通信網，化学分子構造式を表すグラフなど，相互に関係を伴う離散的要素の集まり，即ち離散構造も広義には集合であるが，このような離散構造の集まり L が，要素の関係を壊さないある結合について閉じていれば上半束となる．もちろん，スーパーマーケットにおける顧客の購入商品リストなど，我々が日常でよく扱う有限ブール集合も上半束を構成し得るが，有限ブール集合族は上記に加えて有限でかつ下界 $a \cap b$ を有する“ 束 ”であり，上界と下界の両方について可換則，結合則，吸収則，分配則，相補則を満たす．

一方，“Downward Closure Property (DCP)” は，上記のような広義の集合について，それらの“ 包含関係 ”に基づいて定義される．2 つの集合 a,b の“ 包含関係 ”$a \subseteq b$ を，a,b が“ 順序対 ”をなすこと，即ち，ある 2 項関係 r について $r(a,b) \Leftrightarrow r(b,a)$ である順序関係とする．この包含関係は一般的な定義であるが，特に L が上半束のとき $a \subseteq b$ は $r(a,b) \equiv [a \cup b = b]$ で与えられる．L をその幾つかの要素同士が包含関係を有する“ 順序集合族 ”とし，$a(\in L)$ が L 上に定義されるある性質 P を有するとき $P(a)$ と表すものとする．このとき $a,b(\in L)$ に関する DCP P は以下で与えられる．

$$a \subseteq b \Rightarrow P(b) \rightarrow P(a).$$

上半束に対して良く知られた DCP は，ある商品(アイテム)集合 a がスーパーマーケットの各顧客の購入商品(アイテム)集合 t からなるデータ D に現れる出現頻度(支持度)

$$\sup(a) = \frac{|\{t | t \in D, a \subseteq t\}|}{|D|}$$

がある閾値(最小支持度)minsup 以上である性質

$$P(a) \equiv [\sup(a) \geq \text{minsup}] \tag{2.1}$$

であり，代表的なデータマイニング手法であるバスケット分析で用いられる(Agrwal and Srikant, 1994)．

3. グラフマイニングの探索原理

有限ブール集合やグラフなどの離散構造の集まりであるデータ D を考える．以下では，D において上半束上のある DCP を満たす部分構造を完全探索するマイニング手法を解説する．多くの手法では DCP として，部分構造がある閾値以上の頻度でデータに出現する多頻度性を用いるが，本質的に同じ枠組みでこれ以外の DCP を適用することも可能である．

3.1 多頻度アイテム集合のマイニング

有限ブール集合の集まりであるデータから，多頻度に現れる部分集合を完全探索するマイニングはバスケット分析と呼ばれる(Agrwal and Srikant, 1994)．これはスーパーマーケットにおける売れ筋の商品組み合わせを，多数の顧客の買い物籠(バスケット)の中身を分析して知ることが出来である．あるスーパーマーケットで売っている全商品(アイテム)の集合のように，全アイテムの集合を $I = \{item_1, item_2, \ldots, item_m\}$ とする．そして，ある顧客の 1 回の購入商品(アイテム)群のように，I の空でない部分集合 $t \subseteq I$ をトランザクションと呼ぶ．更にこのようなトランザクションの集合 $D = \{t_d | t_d \subseteq I, d = 1, \ldots, n\}$ を対象データとする．バスケット分析の

1) $F_1 = \{$ Frequent itemsets of cardinality=1$\}$;

2) **for**$(k = 1; F_k \neq \emptyset; k++)$ **do begin**

3) C_{k+1} =apriori-gen(F_k); //New candidates

4) **forall** transactions $t \in D$ **do begin**

5) C'_t =subset(C_{k+1}, t); //Candidates contained in t

6) **forall** candidate $c \in C'_t$ **do**

7) c.count++;

8) **end**

9) $F_{k+1} = \{ c \in C_{k+1} | c.\text{count} \geq$ minimum support $\}$

10) **end**

11) **Answer**= $\bigcup_k F_k$;

図 2. Apriori **アルゴリズム.**

主な目的は，アイテムからなる集合(アイテム集合)$a \subseteq I$ **で，一定頻度以上データ** D **に現れるもの，即ち式(**2.1**)を満たすものを，多頻度アイテム集合として全探索することである．前述したように，多頻度アイテム集合は** DCP **の性質を満たす．**

通常，スーパーマーケットで売っている商品は数千種類以上あるため，全ての可能なアイテムの組み合わせを数え上げる方法で多頻度アイテム集合を求めようとすれば，計算上の組み合わせ爆発に直面してしまう．そのためバスケット分析では，データから多頻度アイテム集合を効率的に全探索する Apriori **アルゴリズムを用いることが多い(**Agrwal and Srikant, 1994**)．このアルゴリズムは比較的単純でメモリ消費量が少なく，かつ効率が高いという特徴を有する．図** 2 **に** Apriori **アルゴリズムの概略を示す．ここで** F_k **は要素数** k **の多頻度アイテム集合の集合，**C_k **は多頻度アイテム集合の候補の集合である．関数** Apriori-gen **は，**join **部と** prune **部からなる．**join **部では要素数** 1 **の多頻度アイテム集合からはじめて，ボトムアップ的に要素数** k **の多頻度アイテム集合から要素数** $k+1$ **の多頻度アイテム集合の候補を作り出す．今，要素数** k **の多頻度アイテム集合は全て見つかっていると仮定する．更に，各集合内のアイテムは予め辞書順にソートされているものとする．この条件の下で** $k-1$ **個の要素が共通な要素数** k **の** 2 **つの多頻度アイテム集合**

$$\begin{aligned} a_k &= \{item_1, item_2, \ldots, item_{k-1}, item_k\} \\ b_k &= \{item_1, item_2, \ldots, item_{k-1}, item'_k\} \end{aligned} \tag{3.1}$$

より，要素数 $k+1$ **の多頻度アイテム集合の候補を上記の** 2 **つのアイテム集合の結合**

$$\begin{aligned} c_{k+1} &= a_k \cup b_k \\ &= \{item_1, \ldots, item_{k-1}, item_k, item'_k\} \end{aligned} \tag{3.2}$$

として生成する．但し辞書順に $item_1 < item_2 < \cdots < item_k < item'_k$ **である．多頻度アイテム集合の候補生成には集合の結合しか用いないため，上半束上の探索となる．仮に** c_{k+1} **が多頻度アイテム集合であるなら，その** DCP **より** a_k **や** b_k **も多頻度アイテム集合である．仮定より要素数** k **の多頻度アイテム集合は全て見つかっているので，**a_k **と** b_k **も必ず既に見つかっている．**

逆に言えば，今見つかっている要素数 k の多頻度アイテム集合の内，式 (3.1) を満たす集合の全ての組み合わせについて式 (3.2) の結合をとれば，漏れなく多頻度アイテム集合候補を得ることができる．

prune 部では候補を絞り込むために，上記候補について要素数 k の各部分集合が全て多頻度アイテム集合かをチェックする．これは DCP より多頻度アイテム集合であるためには，その部分集合は全て多頻度アイテム集合でなければならないからである．図 2 の関数 subset では，多頻度アイテム集合候補の内でトランザクション t に含まれるもの全てを列挙する．このようにして要素数 $k+1$ の多頻度アイテム集合の候補全てについて 1 回のデータスキャンで支持度を計算し，最小支持度を越えるものを多頻度アイテム集合とする．更に $k=k+1$ と k を更新して上記を繰り返す．DCP より要素数が増えるとアイテム集合の支持度は減少し，minsup 以上の多頻度アイテム集合は存在しなくなり，上記アルゴリズムは停止する．データに存在する最も要素数の多い多頻度アイテム集合の要素数を $k_{\max}$ とすると，Apriori アルゴリズムにより高々 $k_{\max}+1$ 回のデータ D のスキャンで，効率的に全ての多頻度アイテム集合を求めることができる．

3.2 多頻度グラフのマイニング

ラベル付きの頂点及び辺からなる多数の無向グラフ $G_d(V_d, E_d, f_d, \ell_d)$ の集まりであるデータ $D=\{G_d(V_d,E_d,f_d,\ell_d)|d=1,\ldots,n\}$ について，そこに多頻度に現れる誘導部分グラフないしは連結誘導部分グラフを完全探索するマイニングを考える．あるグラフを G_s とし，D 中で G_s を含むすべてのグラフの集合を $D_s=\{G_d|G_d\in D, G_s\subseteq G_d\}$ としたとき，G_s は D_s について部分グラフ同型である．多頻度グラフのマイニングは，$|D_s|$ が一定数以上である部分グラフ同型な誘導部分グラフないしは連結誘導部分グラフ G_s をすべて探索する問題である．前節と同様に D における G_s の支持度を

$$\sup(G_s)=\frac{|D_s|}{|D|}$$

とした時に，多頻度性 $\sup(G_s)\geq \text{minsup}$ はグラフからなる上半束の上で DCP となる．

グラフのコード表現上で，隣接行列の $(k-1)\times(k-1)$ 左上部分行列に対応する部分グラフ構造が共通する大きさが k の 2 つの多頻度誘導部分グラフ G_{a_k}, G_{b_k} から，大きさが $k+1$ の多頻度誘導部分グラフ候補 $G_{c_{k+1}}$ を導出する結合を定義する．

$$\begin{aligned} code(G_{a_k}) &= x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3}\cdots x_{1,k}\cdots x_{k-1,k}x_{k,k} \\ code(G_{b_k}) &= x_{1,1}x_{1,2}x_{2,2}x_{1,3}x_{2,3}x_{3,3}\cdots x'_{1,k}\cdots x'_{k-1,k}x'_{k,k} \end{aligned} \tag{3.3}$$

$$\begin{aligned} code(G_{c_{k+1}}) &= code(G_{a_k})\cup code(G_{b_k}) \\ &= x_{1,1}x_{1,2}x_{2,2}\cdots x_{1,k}\cdots x_{k-1,k}x_{k,k}x'_{1,k}\cdots x'_{k-1,k}z_{k,k+1}x'_{k,k} \end{aligned} \tag{3.4}$$

但し $code(G_{a_k})$ は正準形に対応し，$code(G_{a_k})\leq code(G_{b_k})$ とする．これは同一のグラフを表すコード間の結合や同じグラフコードの組み合わせの異なる順序の結合といった冗長な結合を避けるためである．$code(G_{c_{k+1}})$ は G_{b_k} の隣接行列の k 行目ないし k 列目に対応するコード部分を G_{a_k} のコードに繋げ，かつ最後の $x'_{k,k}$ の前に新たな要素 $z_{k,k+1}$ を挿入したものである．このようにして得られた $code(G_{c_{k+1}})$ に対応する隣接行列をグラフ $G_{c_{k+1}}$ の“ 正規形 ”という．各コードに対応する隣接行列 $A(G_{a_k}), A(G_{b_k}), A(G_{c_{k+1}})$ は以下のようになる．

$$A(G_{a_k})=\begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 \\ \boldsymbol{x}_2^T & x_{k,k}\end{pmatrix},\quad A(G_{b_k})=\begin{pmatrix} X_{k-1} & \boldsymbol{x}'_1 \\ \boldsymbol{x}'^T_2 & x'_{k,k}\end{pmatrix},$$

$$A(G_{c_{k+1}}) = \begin{pmatrix} X_{k-1} & \boldsymbol{x}_1 & \boldsymbol{x}'_1 \\ \boldsymbol{x}_2^T & x_{k,k} & z_{k,k+1} \\ \boldsymbol{x}_2'^T & z_{k+1,k} & x'_{k,k} \end{pmatrix}.$$

ここで X_{k-1} は G_{a_k} と G_{b_k} に共通する大きさ $k-1$ のグラフを表す隣接行列であり，$\boldsymbol{x}_i$ と $\boldsymbol{x}'_i (i=1,2)$ は $(k-1)\times 1$ の列ベクトルである．$z_{k,k+1}$ と $z_{k+1,k}$ の要素は G_{a_k} と G_{b_k} の k 番目の頂点間の辺ラベルを表す．2 つの値は無向グラフの場合には対称性より同一であるが，元の $A(G_{a_k})$，$A(G_{b_k})$ からは決まらず 2 つの場合が考えられる．1 つは結合して得られるグラフ $G_{c_{k+1}}$ の k 番目と $k+1$ 番目の頂点の間にラベル $\{\ell_e(e_i)|f_{c_{k+1}}(e_j)=(v_k,v_{k+1})\}$ を持つ辺を付加する場合，もう 1 つはそれらの頂点間に辺を付加しない場合である．これによって $z_{k,k+1}$ と $z_{k+1,k}$ が “$\{\ell_e(e_i)|f_{c_{k+1}}(e_j)=(v_k,v_{k+1})\}$” か “0” である複数通りの隣接行列が生成される．

多頻度連結誘導部分グラフを完全探索する場合には，上記と若干異なる正準形や結合の定義を用いる．ある大きさ k のグラフを表す隣接行列の中で，$(k-1)\times(k-1)$ 左上部分行列が連結部分グラフを表すもののうち，最小のコードを持つものを正準形とする．そして，$code(G_{a_k})$ が連結部分グラフの正準形に対応するコードで，$code(G_{b_k})$ は式(3.3)を満たし，かつ G_{b_k} が非連結部分グラフであるか，または連結部分グラフなら $code(G_{a_k}) \le code(G_{b_k})$ を満たす場合のみ，式(3.4)によって結合を行う．G_{b_k} が連結部分グラフの場合には G_{a_k}, G_{b_k} 双方が連結部分グラフであり，両者の順番を入れ替えた冗長な結合を避けるためにコードの大小関係の制約を課す．G_{b_k} が非連結部分グラフの場合には，制約より順番を入れ替えた結合はできないためコードの大小関係の制約は課さない．この結合により得られたコードにおいて，連結部分グラフとなるように $z_{k,k+1}$ と $z_{k+1,k}$ の値を決めることによって，多頻度連結誘導部分グラフ候補を漏らさず生成可能である．

上記いずれの結合の場合でも図 2 に類似したアルゴリズムにより，頂点 1 個の多頻度誘導部分グラフからはじめて逐次より頂点数の多い多頻度(連結)誘導部分グラフをボトムアップに完全探索することが可能である．はじめの F_1 に相当するものは頂点 1 個の多頻度誘導部分グラフである．F_k は頂点数 k の多頻度(連結)誘導部分グラフの集合，C_k は多頻度(連結)誘導部分グラフの候補の集合である．関数 Apriori-gen の join 部では頂点数 1 の多頻度(連結)誘導部分グラフからはじめて，上記 2 つの多頻度(連結)誘導部分グラフの結合によってボトムアップ的に頂点数 k の多頻度(連結)誘導部分グラフから頂点数 $k+1$ の多頻度(連結)誘導部分グラフの候補を作り出す．多頻度(連結)誘導部分グラフの候補生成にはグラフの結合しか用いないため上半束上の探索となる．仮に $G_{c_{k+1}}$ が多頻度(連結)誘導部分グラフであるなら，その DCP より G_{a_k} や G_{b_k} も多頻度(連結)誘導部分グラフである．頂点数 k の多頻度(連結)誘導部分グラフが既に全て見つかっているならば，G_{a_k} と G_{b_k} も必ず既に見つかっている．即ち，今見つかっている頂点数 k の多頻度(連結)誘導部分グラフの内，式 (3.3)を満たすグラフの全ての組み合わせについて式 (3.4)の結合をとれば，漏れなく多頻度(連結)誘導部分グラフ候補を得ることができる．

関数 Apriori-gen の prune 部では候補を絞り込むために，上記候補について頂点数 k の各誘導部分グラフが全て多頻度(連結)誘導部分グラフかをチェックする．これは DCP より多頻度(連結)誘導部分グラフであるためには，その部分グラフは全て多頻度(連結)誘導部分グラフでなければならないからである．具体的には $A(G_{c_{k+1}})$ について，それから i 行 i 列 $(i=1,\ldots,k-1)$ を除去した $k\times k$ の各部分行列が，全て既に探索された多頻度(連結)誘導部分グラフを表す場合のみを多頻度(連結)誘導部分グラフ候補として残す．関数 subset では，多頻度(連結)誘導部分グラフ候補の内でグラフ G_d の(連結)誘導部分グラフであるもの全てを列挙する．ここではデータに対する各候補のマッチングと頻度計算の計算コストを大幅に削減するため，あらかじ

め前処理でデータ中のグラフを全て正規形の隣接行列で表しておく．また，同一の多頻度(連結)誘導部分グラフ候補が複数の正規形で表現される場合があるため，その中でも正準形であるものに各正規形の頻度を全て合計する．このようにして頂点数 $k+1$ の多頻度(連結)誘導部分グラフの候補全てについて 1 回のデータスキャンで支持度を計算し，最小支持度を越えるものを多頻度(連結)誘導部分グラフとする．更に多頻度アイテム集合のマイニングの場合と同様に，k を更新して上記を繰り返す．DCP より頂点数が増えると部分グラフの支持度は減少するので，minsup 以上の多頻度(連結)誘導部分グラフは存在しなくなり，上記アルゴリズムは停止する．データに存在する最も要素数の多い多頻度(連結)誘導部分グラフの頂点数を $k_{\max}$ とすると，高々$k_{\max}+1$ 回のデータ D のスキャンで，全ての多頻度(連結)誘導部分グラフが得られる．

4. 統計的モデリングへの応用

多数の観測変数の関係をモデル化する統計的方法には，ベイジアンネットワークに代表されるようにグラフ構造を有するモデルを用いるものが多い．このような統計的モデル化手法にグラフマイニングを組み合わせることで，様々な解析ができる可能性がある．ここではその例として，マイクロアレイ遺伝子発現プロフィールデータ(以下マイクロアレイデータ)から遺伝子発現関係をベイジアンネットワークで同定した結果に，更にグラフマイニングを適用して各遺伝子発現の依存関係に関する知見を得る解析を報告する．これはベイジアンネットワークによるモデル化の後処理としてグラフマイニングを適用し，対象とする多変数間の依存関係をより明確に把握する試みである．

4.1 遺伝子発現データ

細胞内の DNA 鎖には多数の遺伝子(gene)がコードされている．これらの遺伝子群について，基準とする細胞状態(リファレンス細胞)での遺伝子発現状態に対して，興味がある別の細胞状態(サンプル細胞)での遺伝子発現状態の相対的な活性の違いを調べる測定手段が，マイクロアレイ分析である．cDNA マイクロアレイ分析に用いるスライドグラスには，多数のスポットが格子状に並んでいる．1 つのスポットが 1 つの遺伝子の発現状態を表し，あるスポットが赤に発色している場合，そのスポットに対応する遺伝子 $gene_j$ はリファレンス細胞よりもサンプル細胞でより多くの mRNA を合成していることを示し，緑だと逆，黄色だと両細胞で同程度 mRNA を合成していることを示す．

興味あるサンプル細胞が n 個ある場合にそれぞれについてマイクロアレイ分析を行うと，図 3 に示すように n 個のマイクロアレイデータが得られる．i 番目のマイクロアレイの $gene_j$ に対応するスポットの正規化された Cy3 色素の発光強度を G_{ij}，同じく Cy5 の発光強度を R_{ij} とすると，スポットの発光を表す数値は $x_{ij}=\log_2(R_{ij}/G_{ij})$ で与えられる．p 個のスポットを持つ i 番目のマイクロアレイは，p 次元ベクトル $x_i=(x_{i1},\ldots,x_{ip})$ で表され，n 個のマイクロアレイデータ全体の集合は $\{x_1,\ldots,x_n\}$ で表される．

4.2 ベイジアンネットワークノンパラメトリック回帰モデル

一般にある遺伝子の活性は他の遺伝子の活性によって影響を受ける．ある遺伝子 $gene_j$ の活性に直接影響する親遺伝子を $gene_1^{(j)},\ldots,gene_{q_j}^{(j)}$ とした時，その影響を各遺伝子及びその各マイクロアレイ測定毎に固有の誤差分散を許す以下のノンパラメトリック加法回帰モデル(nonparametric additive regression model)で表す(Hastie and Tibshirani, 1990; Imoto et al., 2002a)．

$$x_{ij}=m_{j1}(p_{i1}^{(j)})+\cdots+m_{jq_j}(p_{iq_j}^{(j)})+\epsilon_{ij}.$$

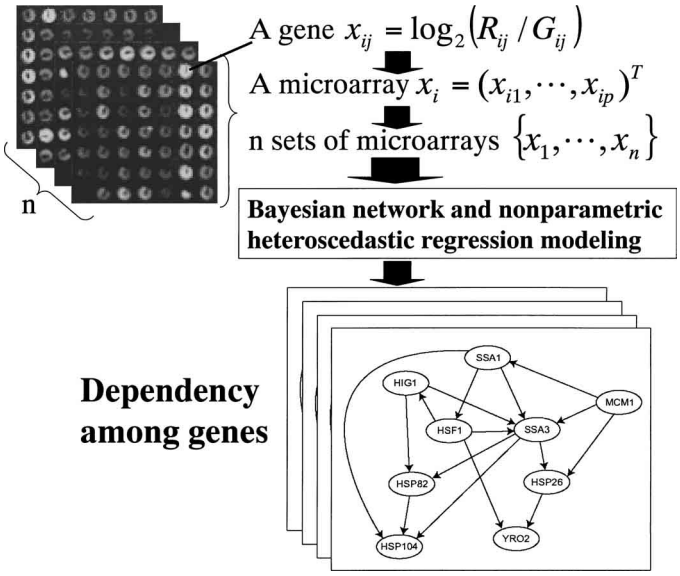

図 3. マイクロアレイデータと遺伝子依存関係モデリング.

ここで，$p_{ik}^{(j)}$ は i 番目のマイクロアレイで測定された $gene_j$ の k 番目の親遺伝子の発現値であり，ϵ_{ij} は平均 0，分散 σ_j を持つ独立正規分布に従う．また，$m_{jk}(\cdot)$ はB－スプライン関数で構成される平滑化関数である(Eilers and Marx, 1996**)．この時，ベイジアンネットワークノンパラメトリック回帰モデル(**Bayesian network and nonparametric regression model**)は以下で表される．**

$$f(\boldsymbol{x}_i;\Theta_G)=\prod_{j=1}^{p} f_j(x_{ij}|\boldsymbol{p}_{ij};\Theta_j).$$

ここで，$f_j(x_{ij}|\boldsymbol{p}_{ij};\Theta_j)$ は $\boldsymbol{p}_{ij}=[p_{i1}^{(j)},\ldots,p_{iq_j}^{(j)}]^t$ が与えられた時の平均 $m_{j1}(p_{i1}^{(j)})+\cdots+m_{jq_j}(p_{iq_j}^{(j)})$，分散 σ_j を持つ正規分布確率密度関数であり，Θ_G は各 m_{jk},σ_{ij} を含むパラメータベクトルである．ただし，親遺伝子が存在しない $gene_j$ には，$i=1,\ldots,n$ に亘る x_{ij} の平均 μ_j，分散 σ_j^2 を用いる．

全遺伝子に関する n 個のマイクロアレイ測定データ全体に亘るこのモデル(グラフ構造)の事後確率は

$$\pi(G)\int\prod_{i=1}^{n} f(x_i;\Theta_G)\pi(\Theta_G|\lambda)d\Theta_G$$

で与えられ，原理的にはこれが最大となるモデルを求めればよい．ここで $\pi(G)$ は生物学的な背景知識から得られるネットワーク G の事前確率分布，$\pi(\Theta_G|\lambda)$ は λ をハイパーパラメータとする Θ_G に関する事前確率分布である．$\pi(\Theta_G|\lambda)$ としてはパラメータの多次元正規分布を用いる．また $\pi(G)$ は，ある遺伝子がその各親遺伝子によって統制されていることが明らかか否かに関する生物学的な背景知識から容易に計算可能である．更に遺伝子間の統制パターン候補が複数ある場合には，それぞれに対応する $\pi(G)$ の中から最も事後確率の高いモデルを探索することも可能である．しかし，Θ_G が高次元であるため，一般にこの積分の直接計算は容易で

はない．そこで，積分のラプラス近似に基づく $BNRC$ (Bayesian network and Nonparametric Regression Criterion)の計算によって，モデルの対数事後確率を評価する方法を用いる(Imoto et al., 2002b)．最大事後確率モデル(MAP 解)を完全探索することは計算量的に困難なため，ここでは最良優先探索(Greedy 探索)が用いられる(Imoto et al., 2004)．ある遺伝子間の統制パターン候補とそれに対応する $\pi(G)$ の下で，それを初期パターンとして各遺伝子間の影響パスの付加，除去，方向の反転を逐次適用して，より事後確率が大きいモデルを探索していく．探索の袋小路に至るとバックトラックして更に事後確率の大きいモデルを探し続け，規定の r 個のモデルを探し終えて停止する．従って図 3 の下方に示すように，探索中途を含め規定個数の多数のモデルが得られる．各 $gene_i$ から $gene_j$ への影響パスの強さは，そのパスを除去した場合としない場合のモデルの対数事後確率の差異 $\Delta BNRC^{ij}$ の大きさによって評価される．対数事後確率差が大きいほど，影響の大きなパスであると考えられる．

4.3 バスケット分析による頻出主要部分ネットワーク抽出

探索によって得られる多数のモデルはそれぞれ遺伝子間依存性に関する異なるネットワーク候補を表す．各ネットワークにおいて $\Delta BNRC^{ij}$ の大きい影響パスは，実際の遺伝子発現の間の依存関係を説明するために必要である可能性が高いが，最良優先探索の過程においてたまたまいくつかの影響パスの $\Delta BNRC^{ij}$ が大きく評価されてしまう可能性もある．従って，本当に必要な可能性の高い影響パス及びそれらが繋がったネットワークは，$\Delta BNRC^{ij}$ の大きさに加えて探索途中の多くのネットワークに安定して見られる構造であると考えられる．そこで，探索過程で導かれるモデルの集合 $\{G_1, \ldots, G_r\}$ から，それらにある最小支持度以上頻出する主要部分ネットワークを抽出することを考える．

ネットワーク G_h の各頂点は $gene_i$，各辺は $\Delta BNRC^{ij}$ の値を持つ影響パスに対応する．G_h 中のある $gene_i$ に対応する頂点は 1 つしかないため，各頂点は各遺伝子の固有名で識別される．また，各辺もある遺伝子 $gene_i$ から別の $gene_j$ へのパスとして 2 つの遺伝子固有名の順序対 $gene_i \rightarrow gene_j$ で識別される．従ってネットワーク G_h は，順序対 $gene_i \rightarrow gene_j$ をアイテムとした時にその集合で表現可能である．今，このようなアイテム集合からある一定値 $\Delta BNRC^{\min}$ 以上の影響の大きさを持つ辺 $gene_i \rightarrow gene_j$ のみを残し，$\tilde{G}_h (\subseteq G_h)$ を得る．これによって主要な影響パスのみを含むモデルの集合 $\{\tilde{G}_1, \ldots, \tilde{G}_r\}$ が得られる．これから探索過程で導かれるモデル集合の頻出主要部分ネットワークは，最小支持度 minsup 以上の多頻度アイテム集合としてバスケット分析を適用することによって導出できる．

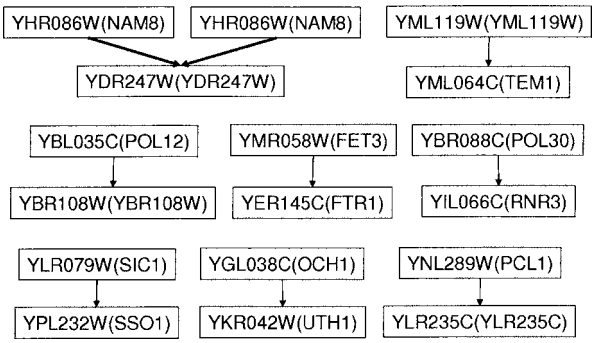

図 4. 遺伝子間依存性の頻出部分ネットワーク(支持度 92.3%)．

あるマイクロアレイデータ $\{x_1,\ldots,x_n\}$ に最良優先探索を適用して，その途中で得られた 5000 個のベイジアンネットワークノンパラメトリック回帰モデル $\{G_1,\ldots,G_{5000}\}$ を得た．各モデルは 801 個の頂点(遺伝子)とその間に約 2600 個の辺(影響パス)を有する．これらのモデルに含まれる全ての辺に亘る $\Delta BNRC^{ij}$ の最大値と最小値の差の 20% を $\Delta BNRC^{\min}$ として，影響が小さい下位の辺を除去し $\{\tilde{G}_1,\ldots,\tilde{G}_{5000}\}$ を得た．この除去により，各モデルは平均 184 個の頂点とその間の平均 115 個の辺に縮約された．そしてこれに $\mathrm{minsup}=0.9$ の下でバスケット分析を適用し，殆どのモデル主要部分に共通して現れる頻出主要部分ネットワークを導出した．図 4 に導出された最大の頻出主要部分ネットワークを示す．これは 1 枚の連結したネットワークではなく 17 遺伝子からなるほぼ孤立した複数の影響パスで構成されるネットワークであるが，全体の 92.3% のモデルに共通して共起が見られた．多くが 2 つの遺伝子間の影響関係であるが，これらが同時に作用していると考えられる．

4.4 グラフマイニングによる頻出主要部分ネットワーク抽出

前節と同じマイクロアレイデータから得られた 5000 個のベイジアンネットワークノンパラメトリック回帰モデルにグラフマイニングを適用し，頻出主要部分ネットワークを抽出する．同じくモデルに含まれる全ての辺に亘る $\Delta BNRC^{ij}$ の最大値と最小値の差の 20% を $\Delta BNRC^{\min}$ として，それより影響が小さい辺を除去し主要な影響を有する辺のみを対象とした．ただし，ここではネットワーク G_h 中の個々の遺伝子間の依存関係ではなく，遺伝子作用のプロセス間の依存関係を分析する．各遺伝子の機能は一般に Gene Ontology (GO)Term と呼ばれる記述子によって簡潔に表される(Gene Ontology Consortium, 2000, 2005)．GO では 1 つの遺伝子に対して Process, Function, Component の 3 つの側面から GO Term と呼ばれる記述を割り当てている．ここでは，遺伝子が如何なるプロセスで作用するかを表す 33 種類の Process GO Term に着目し，データ中の各遺伝子固有名をそれらの Process GO Term に付け替える．こうすると G_h の中に同じラベルを有する頂点が複数存在するようになるため，頂点や辺はラベルによっては完全には識別できなくなる．このようなネットワーク構造をアイテム集合として扱うことは困難であり，頂点や辺のラベル以外にもトポロジー情報を含むグラフとして扱う必要がある．そこで，ここでは 3.2 節で述べた原理に基づき多頻度連結誘導部分グラフをマイニングす

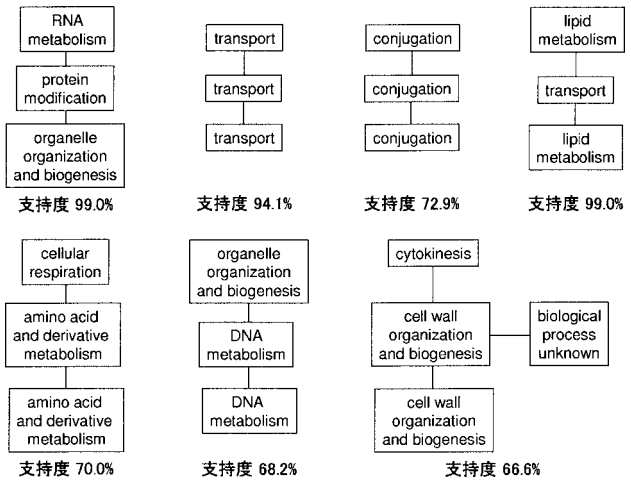

図 5. 2/3 以上のモデルに現れる遺伝子作用プロセス間依存性の主要部分ネットワーク.

る AcGM 手法(Inokuchi et al., 2002)を適用した.

ベイジアンネットワークノンパラメトリック回帰モデルは，遺伝子間の影響の方向性を含む有向グラフで表される．ただし影響の異なる方向性を含んでいても測定データを同様に説明できる等価なモデルが複数存在する場合も多く，モデリングにおいて影響の有無の評価に比較すれば方向性に関する評価の信憑性は低い．また，AcGM を含め現状公開されている多くのグラフマイニングツールが無向グラフ解析の機能のみを有するものが殆どであることから，本報ではネットワークの各辺の方向性を無視し，無向グラフの多頻度連結誘導部分グラフを導出した．2/3 以上のネットワークに共通して見られた 3 遺伝子以上からなる主要部分ネットワークを図 5 に示す．これは大半のモデルに共通して見られる最大の大きさの主要部分ネットワークであると考えられる．個々のモデルが 801 個の頂点から成る大規模ネットワークであるにもかかわらず，遺伝子作用プロセス間で強い影響からなる部分ネットワークは非常に小規模なものに限ら

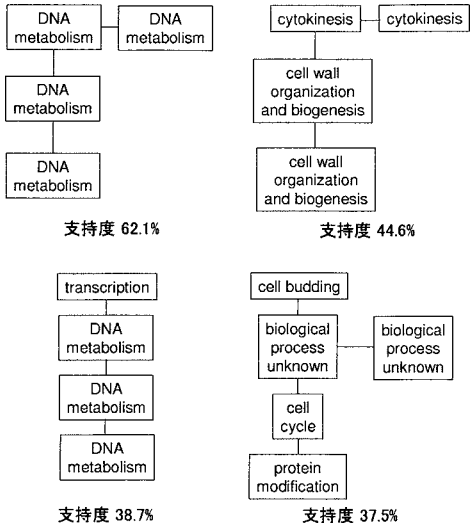

図 6. 1/3 以上のモデルに現れる遺伝子作用プロセス間依存性の主要部分ネットワーク.

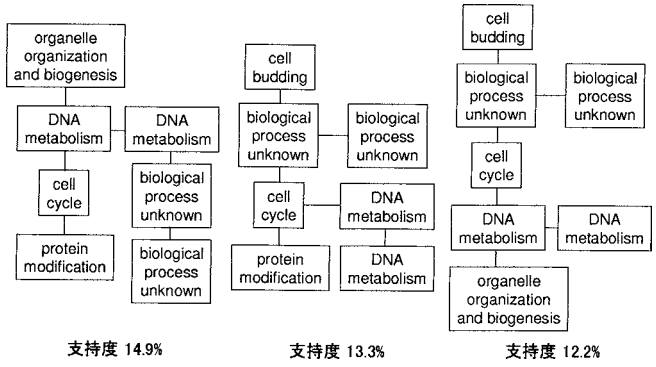

図 7. 10%以上のモデルに現れる遺伝子作用プロセス間依存性の特徴的部分ネットワーク.

れることが分かる．図 6 は 1/3 以上のモデルに共通する頻出主要部分ネットワーク，図 7 は全体の 10%以上に共通する頻出部分ネットワークである．このようにより大きな部分ネットワークの中には一部に特徴的に現れるものが存在するが，ネットワーク全体の大きさから見ると特徴的部分の大きさは限られていると言える．このことから，前節のバスケット分析による結果と同様に，遺伝子種類間の強い依存関係についても安定した大きな構造は見られないと見なすことができる．しかしながら，DNA metabolism 同士の関係や DNA metabolism と organelle organization and biogenesis の関係，cell cycle と protein modification の関係などには多くに共通した安定性が見られ，本解析によって結びつきの強い遺伝子作用プロセスを把握できると考えられる．

5. おわりに

本報では，構造化データに関する代表的マイニング手法であるグラフマイニング手法の現状を概観し，更にグラフマイニングの基礎概念，原理，代表的手法について解説した．そして，多変数間の依存関係を表現する大量のベイジアンネットワークノンパラメトリック回帰モデルに安定に見られる部分ネットワークを同定するために，グラフマイニングを適用した．その対象として，遺伝子発現状態を表すマイクロアレイ測定データを取り上げた．これはベイジアンネットワークによるモデル化の後処理としてグラフマイニングを適用し，対象とする多変数間の依存関係を把握するものであるが，モデリングの過程そのものにグラフマイニングを用いて特徴的な変数間依存関係を同定する方法など，今後探求すべき課題は多いと考えられる．

謝　辞

本研究は，部分的に情報・システム研究機構，新領域融合研究センター，機能と帰納プロジェクトの研究費補助を受けた．また本研究は，東京大学医科学研究所ヒトゲノム解析センター・スーパーコンピュータシステムの計算機を利用して行った．

参 考 文 献

Agrwal, R. and Srikant, R. (1994). First algorithms for mining association rules, *Proceedings of the 20th VLDB Conference*, 487–499.

Cook, J. and Holder, L. (1994). Substructure discovery using minimum description length and background knowledge, *Journal of Artificial Intelligence Research*, **1**, 231–255.

de Raedt, L. and Kramer, S. (2001). The levelwise version space algorithm and its application to molecular fragment finding, *Proceedings of IJCAI01*: *Seventeenth International Joint Conference on Artificial Intelligence*, **2**, 853–859.

Dehaspe, L. and Toivonen, H. (1999). Discovery of frequent datalog patterns, *Data Mining and Knowledge Discovery*, **3**(1), 7–36.

Eilers, P. H. C. and Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion), *Statistical Science*, **11**, 89–121.

Garey, M. and Johnson, D. (1979). *Computers and Intractability*: *A Guide to the Theory of NP-Completeness*, W. H. Freeman and Company, New York.

Gene Ontology Consortium (2000). Gene Ontology: Tool for the unification of biology, *Nature Genetics*, **25**, 25–29.

Gene Ontology Consortium (2005). Saccharomyces Genome Database, http://www.yeastgenome.

org/GOContents.shtml.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, Chapman & Hall, Florida.

Huan, L., Wang, W. and Prins, J. (2004). SPIN: Mining maximal frequent subgraphs from graph databases, *Proceedings of the 2004 Conference on Knowledge Discovery and Data Mining (SIGKDD2004)*, 581–586.

Imoto, S., Goto, T. and Miyano, S. (2002a). Estimation of genetic networks and functional structures between genes by using Bayesian network and nonparametric regression, *Proceedings of Pacific Symposium on Biocomputing*, **7**, 175–186.

Imoto, S., Kim, S., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S. and Miyano, S. (2002b). Bayesian network and nonparametric heteroscedastic regression for nonlinear modeling of genetic network, *Proceedings of 1st IEEE Computer Society Bioinformatics Conference*, 219–227.

Imoto, S., Higuchi, T., Goto, T., Tashiro, K., Kuhara, S. and Miyano, S. (2004). Combining microarrays and biological knowledge for estimating gene networks via bayesian networks, *Journal of Bioinformatics and Computational Biology*, **2**(1), 77–98.

Inokuchi, A. (2004). Mining generalized substructures from a set of labeled graphs, *Proceedings of Fourth IEEE International Conference on Data Mining (ICDM2004)*, 415–418.

Inokuchi, A., Washio, T. and Motoda, H. (2000). An apriori-based algorithm for mining frequent substructures from graph data, *Proceedings of PKDD2000: Principles of Data Mining and Knowledge Discovery, 4th Europian Conference, Lecture notes in Artificial Intelligence 1910* (ed. Jan Zytkow), 13–23, Springer, London.

Inokuchi, A., Washio, T., Nishimura, K. and Motoda, H. (2002). A fast algorithm for mining frequent connected subgraphs, IBM Technical Research Report: RT0448, IBM Tokyo Research Laboratory.

Inokuchi, A., Washio, T. and Motoda, H. (2003). Complete mining of frequent patterns from graphs: Mining graph data, *Machine Learning*, **50**, 321–354.

Inokuchi, A., Washio, T. and Motoda, H. (2005). A general framework for mining frequent subgraphs from labeled graphs, *Journal of Fundamenta Informatiae, Special Issue on Advances in Mining Graphs, Trees and Sequence*, **66**(1–2), 53–82.

Kuramochi, M. and Karypis, G. (2001). Frequent subgraph discovery, *Proceedings of ICDM'01: 1st IEEE International Conference on Data Mining*, 313–320.

Kuramochi, M. and Karypis, G. (2004). Finding frequent patterns in a large sparse graph, *Proceedings of the 2004 SIAM Data Mining Conference*, http://www.siam.org/meetings/sdm04/proceedings/index.htm.

Nijssen, S. (2005). Homepage for Mining Structured Data, http://hms.liacs.nl/index.html.

Nijssen, S. and Kok, J. N. (2004). A quickstart in frequent structure mining can make a difference, *Proceedings of the 2004 International Conference Knowledge Discovery and Data Mining (SIGKDD'04)*, 647–652.

Ruckert, U. and Kramer, S. (2004). Frequent free tree discovery in graph data, *Proceedings of ACM Symposium on Applied Computing (SAC2004)*, *Special Track on Data Mining*, 564–570.

Washio, T. and Motoda, M. (2003). State of the art of graph-based data mining, *ACM, SIGKDD Explorations*, **5**(1), 59–68.

Yan, X. and Han, J. (2002). gSpan: Graph-based substructure pattern mining, *Proceedings of ICDM'02: 2nd IEEE International Conference on Data Mining*, 721–724.

Yan, X. and Han, J. (2003). CloseGraph: Mining closed frequent graph patterns, *Proceedings of the 2003 Conference on Knowledge Discovery and Data Mining (SIGKDD2003)*, 286–295.

Yoshida, H., Motoda, K. and Indurkhya, N. (1994). Graph-based induction as a unified learning framework, *Journal of Applied Intelligence*, **4**, 297–328.

330 Proceedings of the Institute of Statistical Mathematics Vol. 54, No. 2, 315–331 (2006)

Graph Mining and Its Application to Statistical Modeling

Takashi Washio[1,2], Tomoyuki Higuchi[1], Seiya Imoto[3], Yoshinori Tamada[1], Ken Sato[4] and Hiroshi Motoda[2]

[1]The Institute of Statistical Mathematics
[2]The Institute of Scientific and Industrial Research, Osaka Univrsity
[3]Human Genome Center, Institute of Medical Science, University of Tokyo
[4]National Institute of Informatics

This report introduces graph mining techniques actively explored in a recent datamining study, and demonstrates its application to gene network analysis in conjunction with statistical modeling. The study on graph mining was initiated in the mid 1990's, and became widely explored after 2000 upon the proposal of its complete search algorithm. Graph mining is used to find characteristic substructures shared by some graphs in a given massive graph data. In particular, the exhaustive search of frequent subgraphs widely seen in the data is a representative task of graph mining. As this task contains subgraph isomorphism problems, which are known to be NP-complete, its high computational complexity is essential. Accordingly, the development of a practical fast algorithm for graph mining is a key issue in the study.

A property for characterizing the substructures is used to mine characteristic substructures in massive graph data. A naive way to search the characteristic substructures is to check the property on every substructure in the data. However, this approach faces the combinatorial explosion of the substructures in the check. For efficient mining, most graph mining approaches limit the property to a "Downward Closure Property (DCP)." A DCP P is defined as $a \subseteq b \Rightarrow P(b) \rightarrow P(a)$ by using two structures a and b where $P(\cdot)$ means that the property P holds on a structure. Representative DCPs are a frequent itemset and a frequent graph where any of their subitemsets and subgraphs are also frequent. By this definition, if P does not hold on a substructure a, P does not hold on any superstructure of a either. Accordingly, under a set of all substructures of size k and DCP P, candidate substructures of larger size $k+1$ and DCP P are limited to the join of the substructures in the set. This strongly limits the search space of characteristic substructures, and enables practical and fast graph mining.

The graph mining technique was introduced to the post processing of the statistical gene network models obtained from microarray gene expression data. Bayesian network and nonparametric regression models of the gene network were in greedy manner searched in the data. We are interested that subnetworks widely appear over searched networks, since the credibility of such subnetworks are considered to be very high. Basket Analysis and a connected induced subgraph mining AcGM are applied to mine frequent subnetworks over searched networks where each network contains 801 genes. In the results of both Basket Analysis and AcGM, the frequent subnetworks are limited to very small sizes compared with the total size of the gene networks. This indicates that wide varieties of interpretations on a gene network structure are obtained from microarray gene expression data.

In this study, graph mining was applied to the post processing stage of statistical

modeling to extract credible gene subnetworks. Other possibilities such as direct introduction of graph mining to the search process of statistical model structures should be explored in future work.

Key words: Graph mining, data mining, statistical modeling, Bayesian network, gene network.