3A1-02

同期発現する遺伝子集団の階層的発見

Hierarchical coexpression analysis across many microarray data set

山川 宏[*1] 丸橋 弘治[*1] 仲尾 由雄[*1]
Hiroshi Yamakawa Koji Maruhashi Yoshio Nakao

[*1](株) 富士通研究所
FUJITSU LABORATORIES LTD.

Now that a large volume of gene expression data are available for analyses. We propose a method for analyzing gene expression data obtained with various experimental conditions. It extracts groups of genes whose expression levels are synchronizingly changed through a specific context (i.e., a set of samples) by recurrently executing gene clustering based on the correlation of gene expression levels and context extraction based on the distribution of the expression level of a gene cluster. We applied this method to a set of gene expression data consisting of 22,215 genes on 1,435 microarray data and compared the resulting gene clusters with a KEGG pathway (Regulation of actin cytoskeleton). As a result, we found some tissue-specific gene clusters compose a sub-network in that pathway.

1. はじめに

近年の DNA アレイ技術の進歩により，同時に大量の遺伝子発現状態を測定することが可能になった．そこで，発現状態のの関係性を利用することで創薬開発の効率化等が期待されている．例えば，既知の疾患マーカー遺伝子と同期的に変化する別の遺伝子を特定できれば、疾患原因の特定の重要な手がかりとなり得る．

この種の解析技術は，モデルベースのアプローチと，マイニング的なアプローチに分けられるであろう．発現に強い制約を与える，モデルベース・アプローチには，ベイジアンネットワークによる生成モデルを発現データにフィッティングする研究などがある．しかしながら，適用範囲は酵母菌などのように発現情報が時系列情報として得られる場合に限定されがちである．一方で，モデルを特定しない，汎用の多変量解析手法 (主成分分析，相関分析，クラスタリングなど) を用いたマイング・アプローチは，適用範囲は限定されない．しかし，解析結果についてそれ以上に踏み込んだ解析を行うことは困難である．そこで，２つの実験結果を組み合わせた試み等もある [2].

近年は，公開された遺伝子発現データが増大しつつあるため，多くのデータを積極的に活用する，実験サンプル横断的な遺伝子間関係の抽出も注目され始めている．例えば，人において同期発現する遺伝子群の発見や [3]，種を超えて同期発現する遺伝子群の抽出の解析 [4] などがある．この場合には，マイニング・アプローチを取らざるを得ず，主な解析手法は相関分析とクラスタリングであり，やはり，それ以上突っ込んだ解析には至っていない．

そこで，我々が研究を進めている状況分解技術 [6, 5][*1]を改良し，横断的マイニング・アプローチに適用する．つまり，所与の実験サンプルの部分集合をコンテストと呼ぶことにし，特定のコンテクストにおける発現関係と同期発現する遺伝子群を探索する．候補と成り得る，コンテクストは膨大なので，コンテクスト選択を適切に行えば，所与の実験サンプル全体に対する分析のみでは得られない多くの情報を得られる．

なお，バイアスを用いて，大規模データに対し現実的な計算量で探索を可能にした点は，提案技術の重要な特徴である．

連絡先: 山川宏，(株) 富士通研究所 IT コア研究所，
〒 211-8588 川崎市中原区上小田中 4-1-1，tel:044-754-2658, fax:044-754-2693, e-mail:*ymkw@jp.fujitsu.com*

*1 従来状況分解技術の共変関係を，2 項間の相関関係に置き換え．

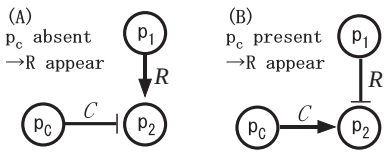

図 1: 遺伝子発現におけるコンテクストの役割
遺伝子 p_1,p_2 の関係 R の表出は，コンテクストを担う遺伝子 p_c により制御される． →: 促進リンク， –| 抑制リンク

2. 遺伝子群の階層的発見

2.1 コンテクスト依存の関係抽出と計算複雑さ

実験横断的なマイニングでは，個別の実験セットで捉えづらい，普遍的な発現関係を得られるが，全サンプルを対象にしている限りは，それ以上には解析できない．ところが，遺伝子間の発現量の関係は，組織や細胞内部位などの空間的コンテクストや，発達時期や細胞周期などの時間的コンテクストなど，特定のコンテクスト (発現環境) において顕在化するであろう．

そのため，単に与えられたデータ全体に対して発現関係の抽出を行うのではなく，様々なコンテクストにおいて発現関係を抽出すれば，遺伝子発現データから従来以上に多くの情報を引き出せる．具体的には，“コンテクストを実験サンプルの部分集合” とみなし，二つの遺伝子間の発現量相関に着目して，コンテクストに依存した相関関係と，同期発現する遺伝子群を抽出する手法を提案する．

しかしながら，可能なコンテクストの数は，実験サンプル数を N として $O(2^N)$ と膨大であり，網羅的な探索は現実的には不可能である．そこで，次節ではある種のバイアスを導入して探索空間を削減する方法を提案する．

2.2 遺伝子群同期コンテクスト・バイアスと階層的探索

遺伝子発現の生物学的背景を考慮すれば，相関関係に付随するコンテクストは，一定規模を持つ遺伝子群の発現状況に依存する場合も多いだろう．そこで，同期的に発現する任意の遺伝子群における発現/休止状態をコンテクストの分類指標とみなす (遺伝子群同期コンテクスト・バイアス)，を導入する．模式的に，このバイアスが成立する例を示す．図 1 左では，遺伝子 p_c が休止していれば，遺伝子 p_1 が p_2 を促進することで発現に正相関が表れる．図 1 右では，遺伝子 p_c が発現していれ

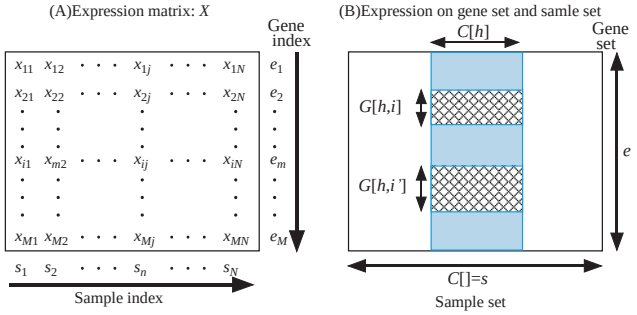

図 2: 遺伝子発現量データの行列表現 (A) と集合表現 (B)

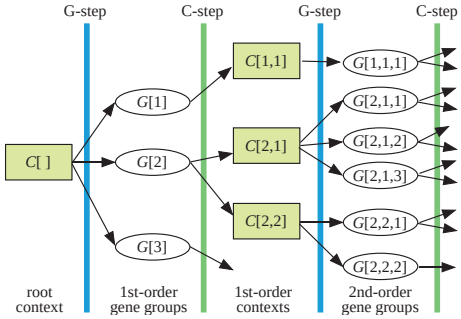

図 3: コンテクスト C と遺伝子群 G の探索階層

ば，遺伝子 p_1 が p_2 を抑制することで発現に逆相関が表れる．

遺伝子群同期コンテクスト・バイアスを用いることで，後述する，同期発現遺伝子群からコンテクストの候補を得る処理 **C-Step**) を実現する．一方，任意のコンテクストにおいて，相関クラスタリング等の従来手法を用いれば同期発現遺伝子群を得る処理 (**G-Step**) を実現できる．そこで，**G-Step** と **C-Step** を交互に繰り返すことで再帰的に，新たなコンテクストと同期遺伝子群を探索する手法を提案する．これにより，コンテクスト依存の発現関係の効率的な発見が可能となった．

3. 階層的な発現関係の発見手法

3.1 遺伝子発現データの取り扱い

解析対象となる発現量行列 X は，図 2 左に示す．遺伝子集合を $p = \{p_i\} : i = \{1, ..., M\}$ としサンプル集合を $s = \{s_j\} : j = \{1, ..., N\}$ とした場合の遺伝子 p_i におけるサンプル s_j の発現量は x_{ij} で，$X = \{x_{ij}\}$ である．

3.2 階層的な探索

コンテクストから遺伝子群を得る (**G-Step**) と，同期発現遺伝子群からコンテクストを得る (**C-Step**) を交互に繰り返えして，図 3 のように階層的に発現関係を探索する．

探索履歴リスト $h = g_1, c_1, g_2, ...$ とし，履歴 h における遺伝子群を $G[h]$，コンテクストを $C[h]$，とする．g_i は遺伝子群を，c_j はコンテクスト指定するインデックスであり，両者とも階層内で同一の親ノードをもつ兄弟ノード間の順序を示す．

G-Step では，コンテクスト $C[h]$ に含まれる発現情報から処理 $\mathcal{P}_G$ により遺伝子群集合 $\mathcal{G}[h]$ を得る．

$$\mathcal{G}[h] = \{G[h,1], G[h,2], ..\} = \mathcal{P}_G(C[h]) \quad (1)$$

ここで，遺伝子群集合 $\mathcal{G}[h]$ における i 番目ら遺伝子群を指定する履歴リストは，$h, g_i : (i = 1, 2, ...)$ とする．

C-Step では，コンテクスト $C[h]$ と遺伝子群 $G[h, g_i]$ を用いて，処理 $\mathcal{P}_C$ により発現情報を処理し，コンテクスト集合 $\mathcal{C}[h, g_i]$ を得る．

$$\mathcal{C}[h, g_i] = \{C[h, g_i, 1], C[h, g_i, 2], ..\} = \mathcal{P}_S(C[h], G[h, g_i]) \quad (2)$$

最後に，得られたコンテクスト毎に $C[h] := C[h, g_i, c_j]$ と代入して，再び **G-Step** に戻る．

実際の動作としては，まず，最上位コンテクストであるサンプルの全体集合 $(C[] = s)$ に対し **G-Step** を適用し，そこで発見した遺伝子群を $\mathcal{G}[] = \{G[1], G[2], ..\}$ とする．次に **C-Step** では，i 番目の遺伝子群 $G[g_i]$ を用いて，コンテクスト集合 $\mathcal{C}[g_i] = \{C[g_i, 1], C[g_i, 2], ..\}$ を抽出する．

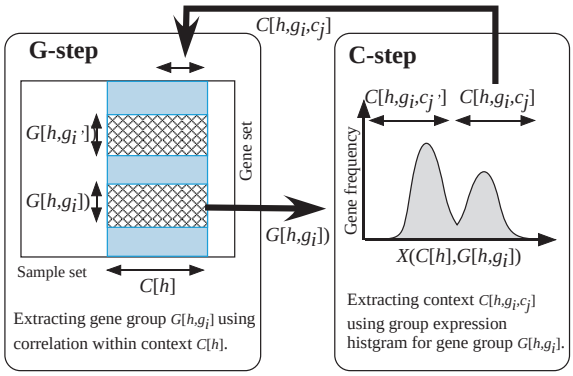

図 4: G-step と S-step の繰り返しによる再帰的探索処理

3.21 **G-step** の処理: 本ステップではまず，コンテクスト $C[h]$ における全ての遺伝子ペア (p_α, p_β) についての相関係数 (式 3 参照) である発現相関行列 $R(C[h]) = \{r_{\alpha\beta}(C[h])\}$ を得る (G-step_1).

$$r_{\alpha\beta}(C[h]) = \frac{\sum_{j \in C[h]} (x_{\alpha j} - \bar{x}_\alpha)(x_{\beta j} - \bar{x}_\beta)}{\sqrt{\sum_{j \in C[h]} (x_{\alpha j} - \bar{x}_\alpha)^2} \sqrt{\sum_{j \in C[h]} (x_{\beta j} - \bar{x}_\beta)^2}} \quad (3)$$

ここで，$C[h]$ に含まれる，サンプル数を $|C[h]|$ として，$\bar{x}_\alpha = (1/|C[h]|) \sum_{j \in C[h]} x_{\alpha j}$，$\bar{x}_\beta = (1/|C[h]|) \sum_{j \in C[h]} x_{\beta j}$ である．

次に，発現相関行列 $R(C[h])$ 内で，相関係数が相関しきい値 θ より大きい遺伝子ペア $(r_{\alpha\beta}(C[h]) > \theta)$ を結合して得られるクラスター (同期発現遺伝子群)$G[h, g_i]$ を複数個得る．その遺伝子群の集合を $\mathcal{G}[h]$ とする (G-step_2)．なお，遺伝子群インデックス g_i は遺伝子群のサイズ順に付与するものとする．

以上で，図 4 左に示す，式 1 に相当する処理 $\mathcal{P}_G$ を実現する．

3.22 **C-step** の処理: 本ステップでは，集合 $\mathcal{G}[h, g_i]$ 内の，遺伝子群 $G[h, g_i] = G[h, g_i]$ 毎に，処理 $\mathcal{P}_C$ を適用する．

まず，遺伝子集合 $G[h, g_i]$ とコンテクスト $C[h]$ で指定される部分発現行列 $X(C[h], G[h, g_i])$ に対し，特異値分解 (SVD) を適用した．ここで得た，サンプルに関する第一固有ベクトルを遺伝子群代表発現量 $\tilde{X}(C[h], G[h, g_i])$ とする (C-step_1).

次に図 4 右に示すように，$|C[h]|$ 個のサンプルのに関する遺伝子群代表発現量 $\tilde{X}(C[h], G[h, g_i])$ の分布を解析する．ここで，2.2 節で述べた遺伝子群同期コンテクスト・バイアスに基づき，発現分布に複数のピークがあれば，コンテク

表 1: 抽出した 7 つのコンテクスト

Context	#	一定規模以上の遺伝子群の数			
	samples	100	10	5	all
$C[]$	1435	4	17	39	616
$C[3,1]$	687	2	13	29	597
$C[3,1,2,1]$	161	3	19	37	528
$C[3,1,5,1]$	469	5	16	35	633
$C[3,2]$	748	6	14	31	515
$C[3,2,1,1]$	153	4	17	41	583
$C[3,2,5,1]$	77	3	15	38	621

スト $C[h,g_i]$ を分割して新たなコンテクスト集合 $\mathcal{C}[h,g_i] = \{C[h,g_i,1], C[h,g_i,2], ..\}$ を得る (C-step_2).

さらに，集合 $\mathcal{C}[h,g_i]$ から，着目すべきコンテクスト $C[h,g_i,i]$ の選択する．新たなコンテクストにおいては，それ以前に発見していない発現関係を抽出できることが望ましいが，その推定は容易ではないので，分布の分離性が良く，事例数 $|C[h,g_i,i]|$ が大きなコンテクストを優先的に選択する (C-step_3)．以上より，式 2 における処理 $\mathcal{P}_G$ を実現する．

4. 階層的コンテクストの抽出実験

本実験では，NCBI 提供の発現データである Gene Expression Omnibus (GEO) [*2]から，Affymetrix GeneChip Human Genome U133 Array Set HG-U133A を利用した 1,435 個の実験サンプル上における 22,215 個の遺伝子を解析対象とした．

4.1 規格化 – Quantile Normalization –

ここでの実験横断的マイニングでは，全ての実験サンプルを同等に比較するアプローチを取っている．これは，各実験セット内で相関関係を抽出し，後で複数の実験セットを統合する Homin ら [3] のアプローチと異なる．本アプローチは，個別の実験では捉えづらい，普遍的な関係を捉えうる可能性がある反面，実験を超えた発現量の比較に危うさが伴う．つまり複数サンプルの比較では，生物学的な原因以外に由来する様々な雑音を除去する必要がある．しかし，サンプルを超えて発現値が一定となるマーカー遺伝子などが存在しないため，現在のところ決定的な規格化手法は存在しない．

本実験で利用するサンプル数が比較的多いため，それでも高速に実行可能な，“Quantile Normalization” という規格化手法を用いた．これは，サンプル間で発現量の分布 (ヒストグラム) 同型であるという仮定の基での発現量調整である [1].

4.2 階層的に抽出した同期遺伝子群

規格化後のデータに対し，**G-Step** での相関しきい値 $\theta = 0.8$ として，3. 章で述べた手法を用いて，階層的に同期遺伝子群の抽出を行った．現状は，手動で C-step_2,C-step_3 を行っており，表 1 に示す，7 コンテクストに着目して，発現関係と同期遺伝子群を抽出した (解析した範囲は探索階層の一部である).

まず，ルートコンテクスト $C[]$ には図 1 に示すように，616 群の遺伝子群を含むが，その中から遺伝子数が 130 個と大きく，サンプルの発現分布が分割しやすい遺伝子群 3($G[3]$) を用いて，下位コンテクスト ($C[3,1], C[3.2]$) を抽出し，さら深いに階層のコンテクストについても同様の処理を行った．

ここで，階層的な探索で得られる新たなコンテクストにより増加する，情報を見積もる．まず，各コンテクストにおいて相関しきい値 $\theta = 0.8$ 以上の遺伝子ペアを数え上げ，コンテクスト間

*2 URL: *http://www.ncbi.nlm.nih.gov/geo/*

表 2: 高相関遺伝子ペアのコンテクスト間の重なり

Context $C[\cdot]$	$\emptyset$	3,1	3,1, 2,1	3,1, 5,1	3,2	3,2, 1,1	3,2, 5,1
$\emptyset$	1457	484	924	1387	868	1018	1415
3,1	484	1430	320	490	266	516	515
3,1,2,1	924	320	1299	1051	1079	641	963
3,1,5,1	1387	490	1051	2614	1399	1058	1562
3,2	868	266	1079	1399	2844	611	952
3,2,1,1	1018	516	641	1058	611	1209	1041
3,2,5,1	1415	515	963	1562	952	1041	1725

表 3: 遺伝子群に対応付けられた KEGG パスウエイ (一部)

Gene group	# genes	KEGG pathway (a)	(b)	(c)	(d)	(e)	(f)
# genes in pathway		130	247	209	196	163	105
annotated # genes		53	44	37	35	35	31
G[1]	170	.	.	.	2	.	18
G[2]	130	2	.	5	8	5	.
G[3]	130	.	.	**9**	2	4	.
G[4]	117	.	.	**4**	1	5	.
G[10]	29	18	.	.	.	.	.
G[3,1,1]	465	31	.	1	3	1	15
G[3,1,2]	148	.	3	**7**	2	2	.
G[3,2,1]	387	24	4	6	1	9	.
G[3,2,2]	217	.	3	2	3	1	2
G[3,2,3]	211	.	1	.	3	.	14
G[3,2,4]	182	.	5	1	2	.	.

表中の数字はすべて，遺伝子の数をあらわす．(a)Oxidative phosphorylation (hsa00190), (b)MAPK signaling pathway (hsa04010), (c)Regulation of actin cytoskeleton (hsa04810), (d)Calcium signaling pathway (hsa04020), (3)Focal adhesion (hsa04510), (f)Cell cycle (hsa04110)

での高発現遺ペアを共有する数を表 2 に示した (LocusID で比較)．表 2 内の，7 コンテクスト間の 21 通りの類似度を Jaccard 係数 $J(,)$ で見積もると，その平均値は 0.36 である．6 つの親子関係にあるコンテクスト間 ($C[]$ と $C[3,2]$ など) の異なりは大きく，平均値は 0.19 である．特に同一遺伝子群から分割したコンテクストの異なりは大きい ($J(C[3,1], C[3,2]) = 0.07$).

また，全遺伝子を選んだコンテクスト ($C[]$) で抽出される関係に含まれる遺伝子数は 2,552 で，全体の 11%でしかないが，深さ 2 のコンテクスト ($C[3,1], C[3,2]$) を追加する，4,976(22%) に増大し，深さ 3 のコンテクスト ($C[3,1,2,1], C[3,1,5,1], C[3,2,1,1], C[3,2,5,1]$) を追加すると，5,390(24%) に増大する．以上より，本手法により関係を付与しうる遺伝子数が増大することを示した．

4.3 KEGG パスウエイ上での遺伝子群の解釈

前節で得られた遺伝子群は，DNA アレイの発現データに基づくため，mRNA の転写のレベルに関わる．一方，細胞レベルの分子間相互作用ネットワークの知識であるパスウエイ情報が，KEGG データベースには約 300 個が蓄積公開されている．DNA の発現量と蛋白質量は必ずしも比例しないが，本節では，転写レベルの情報を分子間相互作用のレベルに対応付けることで，生物学的意味の解釈を試みる．

3 コンテクスト $C[], C[3,1], C[3,2]$ 内の主な遺伝子群と，KEGG パスウエイの遺伝子群との重なりを表 3 に示した (重なり数の上位 6 パスウエイ)．重なりが大きなパスウエイの一つである (c) アクチン細胞骨格の制御 (Regulation of actin cytoskeleton: hsa04810) に着目して生物学的解釈を行う．パ

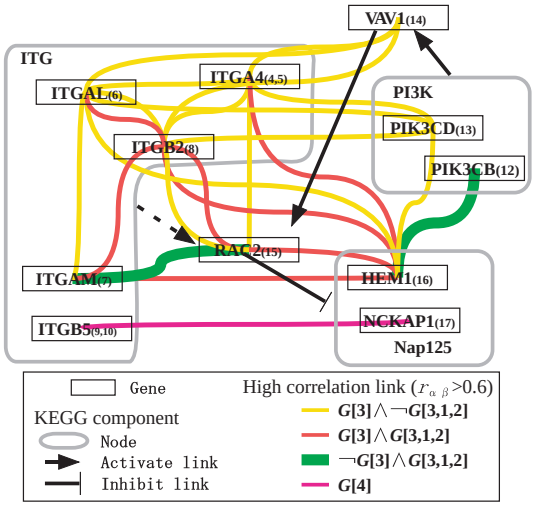

図 5: パスウエイ (c) への 3 遺伝子群のマッピング (抜粋)

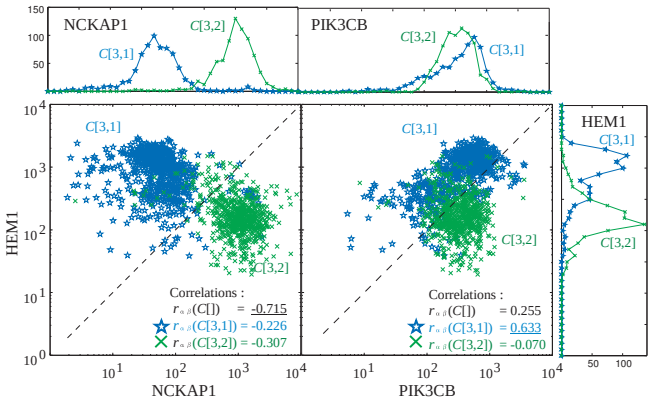

図 6: コンテクスト $C[3,1](☆)$ と $C[3,2](\times)$ での発現状況
3 遺伝子の発現量についてのヒストグラムと散布図

スウエイ (c) は 209 個の遺伝子を含み，そのうち 37 個が上記 3 コンテクスト内の遺伝子群に対応付けられる．そして，対応遺伝子を 9 個含む $G[3]$ により分割された下位コンテクストで $G[3,1,2]$ の 7 遺伝子が目立っている．

図 5 に示すように，遺伝子群 G[3] はパスウエイ (c) 上の"*PIK3CD* → *VAV1* → *RAC2* −| *HEM1*"という相互作用パスに集中して現れていた．パスを構成する各ノードには、複数の遺伝子が対応するが、遺伝子群 $G[3]$ には、その一部のみが抽出されていた．$G[3]$ に含まれるそれらの遺伝子は、いずれも白血球に関する働きが文献で報告されていた．例えば、免疫との関与が知られている *PIK3CD* は白血球での高発現が、*HEM1* は造血幹細胞由来の細胞（白血球を含む）でのみ発現することが報告されていた．ここで遺伝子群 $G[4]$ に含まれる *NCKAP1* は，*HEM1* パスウエイ上で同一ノードを占めるにも関わらず，逆相関の関係にある点が興味深い (図 6 左参照)．

また、パスウエイ (c) において、*RAC2* へのパスが存在する *ITG* ノードでも，integrin 類を構成する 2 種類のサブユニットのうち、白血球での働きが報告されている α サブユニット (*ITGA4, ITGAL, ITGAM*) のみが遺伝子群 $G[3]$ に抽出されていた．これらより、遺伝子群 $G[3]$ は、白血球細胞で強く働くタイプの遺伝子に対応すると考えられる．

次に，階層的なコンテクスト探索により，遺伝子群 $G[3]$ の発現レベルが高いコンテクストで相関関係が検出された遺伝子群 $G[3,1,2]$ を見る．新たに，マクロファージの食作用における標的細胞のアポトーシスに関して重要な役割をする *PIK3CB* と *HEM1* の関係が現れた (図 6 右) そして，マクロファージで特によく働く遺伝子群 (*ITGAM, ITGB2, RAC2, HEM1*) は $G[3]$ 同様に含むが，*VAV1* 等のマクロファージ以外でよく働く遺伝子や、*ITGA4, ITGAL* 等の白血球全般で広く発現するような遺伝子は含まなかった．このことは、遺伝子群 $G[3]$ の発現レベルによるコンテクスト抽出により、白血球細胞に関連するコンテクストに絞られた結果、より細かい細胞の種類に依存する遺伝子発現関係が抽出されたものと考えられる．

5. まとめ

実験横断的な遺伝子発現マイニングにおいて，階層的にコンテクスト (実験サンプルの部分集合) を探索することにより，コンテクストに依存した発現関係の抽出を実現した．

実験では，探索空間の一部のコンテクストに着目して発現関係を抽出し，それらがコンテクストを考慮しなければ得られないことを検証した．次に，得られたコンテクスト毎の同期発現遺伝子群と重なりの大きい，KEGG パスウエイを一つ選択し，そこでの同期遺伝子群とコンテクストの関係の一例を調査したところ，組織レベルの分類に関与する階層が現れた．

今後，階層的探索の自動化により網羅的にコンテクストを抽出し，多くのパスウエイでの多様なコンテクストを対応付けしたい．一方，解析信頼性の向上のため，実験セットごとの相関係数を集約した Homin ら [3] の先行研究との比較検討したい．

なお，今後プロテオミックスが主流となり，本技術を適用できるような蛋白質発現データの蓄積が待たれる．また，提案したバイアスを，状況分解における計算量爆発問題に対する，一般的な解決策として利用することも検討したい．

参考文献

[1] B Bolstad. Probe level quantile normalization of high density oligonucleotide array data. http://oz.berkeley.edu/ bolstad/stuff/qnorm.pdf.

[2] Jung Kyoon Choi, Ungsik Yu, Sangsoo Kim, and Ook Joon Yoo. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics*, Vol. 19, pp. i84–i90, 2003.

[3] Homin K. Lee, Amy K. Hsu, Jon Sajdak, Jie Qin, and Paul Pavlidis. Coexpression analysis of human genes across many microarray data sets. *Genome Research*, Vol. 14, pp. 1085–1094, 2004.

[4] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene coexpression network for global discovery of conserved genetic modules. *Science*, Vol. 302, pp. 249–255, 2003.

[5] Hiroshi Yamakawa, et. al. Multi-aspect gene relation analysis. In *Pacific Symposium on Biocomputing PSB2005*, pp. 233–244., January 2005.

[6] 山川宏, 馬場孝之, 岡田浩之. ETMIC 基準を用いた状況分解によるカード分類課題での概念獲得と予測過程. 認知科学, Vol. 11, No. 2, pp. 143–154, 2004.