2A2-03

状況分解による多視点からの遺伝子間関係発見支援

Supporting Multiaspect Genes Relation Discovery using Situation Decomposition

山川 宏[*1] 山口 睦世[*2] 仲尾 由雄[*1]
Hiroshi Yamakawa Mutsuyo Yamaguchi Yoshio Nakao

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

The situation decomposition extracts multiple situations, each of which is a combination of an attribute set and a case set, from relational database. We propose that we use extracted situations as aspects for the case comparison. Compact representation is extracted using singular value decomposition, for the gene group that appeared during the midembryogenesis of rat's kidney. And the situation decomposition method was applied to that data. In the results, top's 5 situations about ETMIC criterion were the combinations of a gene set and an attribute set which were different in each other. It means that the situation decomposition can offer some different aspects. The relations between genes were investigated by using gene-attribute network of top's two aspects. Comparing genes becomes easy in each aspect, since the attributes which emphasize the differences between genes are selected.

1. はじめに

本発表では，ETMIC 状況分解 [7] で得られる状況という概念を，事例比較の視点として利用する方法を提案する．

状況分解は関係データから複数の状況を抽出する解析手法で，各状況は部分事例集合と部分属性集合により構成される．抽出される状況は，一種の情報量基準である ETMIC 値の，部分空間毎の事例選択変化に対する極大点探索により得られる．

従来研究において，状況という概念は，予測のための部分情報 [7] やマルチモジュール認識行動システムの各モジュール [6] として用いられてきた．しかし，本報告では，事例比較の視点つまり,「状況を，相互に比較しやすい部分事例集合と，その視点となる部分属性集合の組合せ」としての利用を提案する．

次章で，状況を抽出する手法である状況分解と，状況を事例比較の視点として利用する妥当性を議論する．3 章で，特定の遺伝子集団に対して状況を用いた遺伝子比較を行うための具体的手段を説明する．4 章で，ラットの腎臓で胚発生中期に発現量が増大する遺伝子集団に対し，提案手法を適用する．

2. 事例比較の視点としての状況という概念

2.1 状況分解 – 状況の自動抽出 –

状況分解 [7] は，教師無し学習手法の一種であり，関係データから複数の状況という概念を取り出す．関係データは事例 (ここでは，イベント/レコード/ケースなどと同義) の集合であり，各事例は属性 (ここでは，特徴量/フィールドなどと同義) に対応した値を持つ．図 1 では，左側の白枠四角がその入力データを表す．

状況分解は，この入力データから，図中のいくつかの網線による四角で示すような，複数の状況と呼ばれる概念を抽出する．状況 $J = \{A, C\}$ とは，横軸の全属性の集合からその部分集合 A を選択し，同時に，縦軸の全事例の集合からその部分集合 C を選択したものである．状況として抽出しうる組合せは非常に多数存在する．そこから，状況に対する評価を利用し，部分空間毎に事例選択変化に対する極大点を探索する．

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*

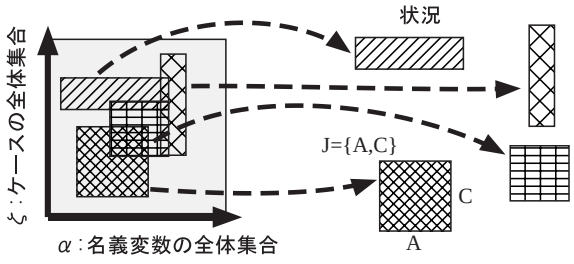

図 1: 状況分解のイメージ図
事例／変数それぞれの部分集合の選択を表現しているが図での表示が困難なため，便宜的に連続領域で示す．

状況に対する評価基準として，ETMIC 基準がある，次式で与えられる．

$$E(A,C) = n_C \left(\min_i \left(I_{X_A^{+i};X_i}(C) \right) - \max_j \left(I_{X_A;X_j}(C) \right) \right) \quad (1)$$

ここで，n_C は選択事例数，$I_{X_A^{+i};X_i}(C)$ と $I_{X_A;X_j}(C)$ はそれぞれ部分事例集合 C における二つの部分空間の相互情報量，さらに，X_A^{+i} は，部分属性集合 A に属性 i を追加した部分空間，X_i は，属性 i による部分空間，X_A は，部分属性集合 A による部分空間，X_j は，属性 j による部分空間，をあらわす．

状況分解の振る舞いを説明するために，図 2 にデータ分布の例を示す．3 次元変数空間中の各変数 X, Y, Z の変域を $[0.0, 1.0]$ とし，事例は 1,000 個である．平面 A$(X + Z = 1)$ 上に 500 事例が，平面 B$(Y + Z = 1)$ 上に 500 事例が，それぞれ一様に分布する．現状の状況分解は各変数を離散化して名義変数として扱うため，分布に関する線形性は考慮しない．

状況分解は，内部に関係を持つ部分状況を抽出しようとするため，図 2 に示すように平面 A 上の事例と変数 X, Z 及び，平面 B 上の事例と変数 Y, Z の選択を行う．図 1 で言えば，平面 A の抽出は名義変数の全体集合 α から部分変数 X, Z を選択し，事例の全体集合 ζ から平面 A 上の 500 事例を選択する．

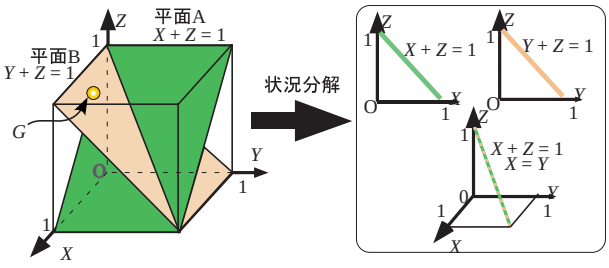

図 2: 状況分解の例題

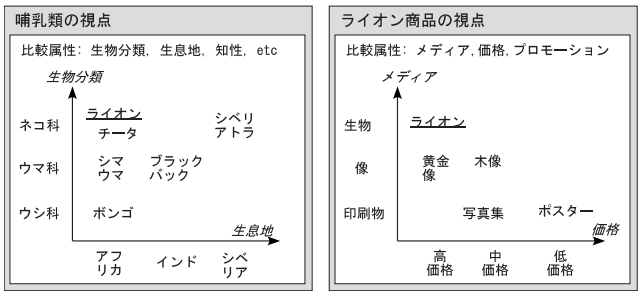

図 3: 実物のライオンに対する二つの視点 (状況)

2.2 事例比較の視点としての状況

本節では，ETMIC 状況分解で抽出される状況が，ユーザが事例比較に利用する視点として妥当であることを議論する．

既に述べた，図 2 の例では，平面 A 上の 500 事例は X 軸と Z 軸の視点から状況内の事例を比較しやすく，平面 B 上の 500 事例は Y 軸と Z 軸の視点から状況内の事例を比較しやすい．

図 3 では，別の例として，実物のライオンを一つの事例とし，そに関する様々な事例が混在するときに，異なる二つの視点で把握する場合を考察する．哺乳類の視点では，比較属性[*1]は，生物分類，生息地，知性などがあり，事例としては，チータ，シマウマなどが含まれる．一方，同様にライオンを含んでも，ライオン商品の視点では，比較属性はメディア，価格，プロモーションなどであり，事例としては黄金像，写真集などが含まれる．このような例から考えると，事例比較に適した視点の性質として以下の三つがあると考えた．(a) 視点内の事例数が多い性質，(b) 表現空間において事例が適度にまとまりつつ広がって分布 (例えば，クラスター状) する性質，(c) 非選択属性に対して汎化できる性質．

状況分解に用いる，状況の評価基準 (ETMIC 基準, 式 1) は，上記 3 つの性質を反映している．性質 (a) としては，選択事例数 n_C を乗ずることにより，状況内の事例数を増大させている．性質 (b) としては，属性削除に関する相互情報量の項 $\min_i(I_{X_A^{+i};X_i}(C))$ により，状況の選択属性に関して状況内の事例間での共変化関係を増大させている．性質 (c) としては，属性追加に関する相互情報量の項 $\max_j(I_{X_A;X_j}(C))$ により，選択していない属性からの独立性が高くさせる．

上記の議論に基づき，ETMIC 値は，様々な視点を選択する基準の候補として妥当であると判断した．

*1 本来，ある視点において，固定属性と比較属性がある．固定属性は視点内で一定値をもち，視点自体の特定には有用である．比較属性は視点内で値が変化し，視点内の事例間の比較に有用である．そのため今回は比較属性のみについて議論している．例えば，哺乳類の視点においては汗腺を持つことなどは固定属性である．

3. 状況を利用した遺伝子比較による解析

我々は，遺伝子と蛋白質の持つ様々な情報の解析支援を目的とした研究環境として，Genesphere[1] を開発している．パスウエイ等の解析を進める過程では，注目する複数の遺伝子とその属性 (分子機能など) を概観することが必要とされる．そこで，Genesphere では，ユーザが注目する属性を決定しながら，遺伝子辞書から抽出した遺伝子間の属性をネットワーク表示する Connection Miner(Cminer) 機能を搭載している．そこで，本報告では，状況が事例比較に適したいくつかの視点を推薦できる能力を，Cminer に組合せた解析を行ったので報告する．

最近は，生物機能に着目した解析ツールとして、GoMiner, GeneSpring（Simplified Ontology）などがあり，GO を用いた解析方法の提案も行われ始めている [2, 5].

状況を利用した遺伝子の比較には，遺伝子の属性体系として Gene Ontology(GO) を，遺伝子データとして LocusLink を用いる．しかし，遺伝子に対する属性の付与は疎であり，このままで，遺伝子比較に適した視点の推薦を行うことは難しい．そこで，以下の手順により，遺伝子集団の解析を行う．

1. Gene Ontology(GO) の推移律を用いた拡張属性の生成
2. 情報量荷重した特異値分解による固有属性抽出
3. 解析対象の遺伝子集団に対するオルソログ拡張
4. 状況分解による遺伝子比較に適した視点の推薦
5. 最大ノルム属性によるネットワーク表示

3.1 GO の推移律を用いた拡張属性の生成

3.11 **Gene Ontology(GO):** GO は遺伝子のオントロジーとしてスタンダードとなりつつあり，多くの研究グループがコンソーシアムに参画している．これは，遺伝子自体を記述した，FlyBase, SGD,LocusLink などの遺伝子データベースに対してメタな関係にある．トップレベルの概念は，”Gene Ontology”であり，その直下に，[b] 生物学的プロセス，[f] 分子機能，[l] 局在位置の 3 カテゴリがある．

3.12 **推移律による拡張属性:** Cminer が利用する，遺伝子データベースである LocusLink では，事例としての遺伝子 c_j 毎に，属性としての GO タームが割り付けられるので，これを属性リスト $L_{org}(c_j)$ で表現する．

しかし，GO タームの割り付けは疎で，GO 階層上の抽象度もまちまちであるため，遺伝子同士の比較が困難である．そこで，各遺伝子の属性要素毎に，階層上での推移律が成立する先祖属性を追加し，これを，拡張属性リスト $L_{ext}(c_j)$ として，上記問題を低減する．

3.2 情報量荷重した特異値分解による固有属性抽出

遺伝子に対する属性の付与が疎であるため，特異値分解 (SVD) を用いて，LocusLink に登録された，ヒト・マウス・ラット遺伝子の属性空間を圧縮し，固有属性の抽出を行う．

3.21 **GO 階層構造に対応する情報量荷重:** 特異値分解は，多くの遺伝子に現れる属性を強調する．一方，拡張属性は GO の階層に基づくため上位の属性ほど頻出する．このため，拡張属性をそのまま利用して特異値分解を行うと，単に上位概念が優先的に選択され，興味深い結果は得られない．

そこで，適度な抽象度 (深さ) の情報を利用するために，属性 i を持つ遺伝子数 m_i の全遺伝子数 m に対する比率を p_i として得られる，情報量 W_i を用いて属性の重み付けを行う．

$$W_i = -\log(p_i) = -\log(\frac{m_i}{m}) \qquad (2)$$

すると，全遺伝子に付与される最上位概念の荷重 $W_i = 0$ となり無視され，特殊化された下位概念ほど強調される．

3.22 **特異値分解 (Singular value decomposition):** 全体で m 個の遺伝子に一度以上付与されている，n 個の拡張属性についての n 行 m 列の行列 G を生成する．拡張属性リスト $L_{ext}(c_j)$ に含まれる拡張属性 i の行列要素 G_{ij} の値として W_i を設定し，それ以外の行列要素は 0 とする．特異値分解は行列 G を，以下の形に分解する．

$$G = USD^T \tag{3}$$

ここで U，D はそれぞれ $U^TU = I_n$，$D^TD = I_m$ を満たすユニタリ行列で，U の列ベクトルを左特異ベクトル，D の列ベクトルを右特異ベクトルと呼ぶ．また S は n 行 m 列の非負の (広義の) 対角行列で，その対角要素を特異値という．

疎行列の性質を利用して大規模な行列に対応できる特異値分解 (SVD) ツールとして，計算ソフトウエアである MATLAB のライブラリを利用する．

3.3 解析対象の遺伝子集団に対するオルソログ拡張

解析対象の遺伝子集団の Accession No.(GenBank 用) を Kidney Developement Gene Expression Database (*http://organogenesis.ucsd.edu/*) から取得し，これを LocusID (LocusLink 用) に変換する．

次に、解析対象の遺伝子集団においては，付与された属性が不十分な問題を補うために，相同な遺伝子を同一視する処理としてオルソログ拡張を実行する．具体的には，ヒト，マウス，ラットの３種類の生物種についての，相同な遺伝子について，その属性を相互に追加する (つまり OR 論理として拡張を行う)．

ここで使用したオルソログデータは HomoloGene よりヒト，マウス，ラットについて抽出した 13952 行 (エントリー)，3 列 (生物種) の LocusID の行列である (欠損有り)．

3.4 遺伝子集団に対する状況分解

解析対象の遺伝子集団について，固有属性 (上位 20 次元) における状況分解を行ない，複数の状況 $J = \{A, C\}$ を抽出する．これにより，遺伝子集団内の遺伝子を比較するのに適した遺伝子の部分集合 C と，固有属性集合 A を推薦する．

各固有属性は属性の線形和であるから，固有属性集合 A の推薦は元の属性 (GO ターム) を重み付けして選択することに相当する．

3.5 最大ノルム属性によるネットワーク表示

状況 $J = \{A, C\}$ が選択した固有属性集合 A に含まれる各固有属性のインデックスを a とし，属性のインデックスを i とし，特異値分解で得た行列 U の要素を U_{ai} とする．すると，この状況における拡張属性 i の行列 U に関するノルムは

$$|U_i| = \sqrt{\sum_{a \in A} {U_{ai}}^2} \tag{4}$$

である．次に，遺伝子 c_j 毎に，最大ノルム属性 $i_{max}(c_j)$ を次式により決定する．

$$i_{max}(c_j) = \arg \max_{i \in L_{ext}(c_j)} |U_i|$$

これは，遺伝子 c_j 毎に，拡張属性リスト $L_{ext}(c_j)$ に含まれる属性中で，最大ノルムを持つ属性である．

Cminer による遺伝子-属性ネットワーク画面では，状況が選択した遺伝子およびそれ毎の最大ノルム属性を表示した．さらに，関連する遺伝子と属性間のリンクを表示した．

4. 解析結果 – ラット 腎発生の遺伝子発現 –

本章では，Stuart[4] らが遺伝子発現解析から得た実験結果の一部を利用し，遺伝子比較に適した視点を抽出し解析する．

Stuart らは，DNA アレイを用いて，ラットの腎発生中の遺伝子発現を得て，それに対して時系列パターンのクラスタ解析を行った．その解析により，腎発生中に発現する 8,740 個の遺伝子は，5 つの典型的な発現時系列パターンに分類された．

我々は、その中から，胚発生中期 (midembryogenesis) でピークに達する、2 番目の遺伝子集団 (Group2) について、提案手法を適用する．AccessionID では 168 個ある Group2 の遺伝子のうち，LocusID に変換された遺伝子が 66 個であり，その中で属性をもつ以下の 24 個の遺伝子を扱う．ここで，各遺伝子は，**シンボル**を太字で，次に名前を表示する．

AGR: Agrin, **Calm1**: Calmodulin 1 (phosphorylase kinase, delta), **Col1a1**: collagen, type 1, alpha 1, **Dcn**: decorin, **ENP1MR**: Epithelial membrane protein 1, **Erbb2**: Avian erythroblastosis viral (v-erb-B2) oncogene homologue 2 **Erp29**: endoplasmic retuclum protein 29, **Galr3**: galanin receptor 3, **ID125A**: Inhibitor of DNA binding 1, helix-loop-helix protein, **ILGFBPA**: Insulin-like growth factor binding protein 2, **Lamc1**: laminin, gamma 1, **Lbp**: lipopolysaccharide binding protein, **Mmp2**: matrix metalloproteinase 2, **Mtap6**: microtubule-associated protein 6, **Nkaa1b**: ATPase, Na+K+ transporting, alpha 1, **Npr1**: natriuretic peptide receptor 1, **Phb**: Prohibitin, **Pkcb**: protein kinase C, beta 1, **Pmp22**: peripheral myelin protein 22, **SOMATO**: somatostatin receptor 5, **Serpina1**: serine (or cysteine) proteinase inhibitor, clade A, member 1, **Sm22**: Transgelin (Smooth muscle 22 protein), **Sparc**: Secreted acidic cystein-rich glycoprotein (osteonectin), **Ucp2**: Uncoupling protein 2, mitochondrial.

なお，Stuart らは，論文 [4] において，上記 Group2 の遺伝子集団は，主に細胞外マトリクス (extracellular matrix) に関与するタンパク質により特徴づけられると述べている．

4.1 実験設定

特異値分解では，LocusLink に含まれるヒト・マウス・ラットをあわせた 24,489 個の遺伝子を扱う．全 13,557 個中で利用されている属性は 5,325 個である．解析対象の遺伝子集団に対する状況分解には，特異値分解結果の上位 20 個の固有属性を用いる．なお，状況分解では，固有属性毎に，その最大値と最小値の間を含むように 5 等分することで離散化した．

4.2 状況分解の結果一覧

ETMIC 状況分解を用いて抽出された視点 (状況) を，ETMIC 値により評価づけし，上位の 5 つの視点を $\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \mathcal{E}$ とし，表 1 に示す．視点毎に，ETMIC 値，選択固有属性, 選択遺伝子数を示した．また，右半分には，遺伝子選択の異なり数 (24 個中) のクロス表を示した．

このように，各々の視点が異なる固有属性と遺伝子を選択することで，解析対象遺伝子の異なる側面を捉えやすくなる．

表 1: 状況分解で得られた上位 5 つの視点 (状況)

順位	ETMIC	固有属性	遺伝子数		遺伝子異なり数のクロス表				
					$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$	$\mathcal{D}$	$\mathcal{E}$
1	8.21	4 6	18	$\mathcal{A}$	0	7	11	8	8
2	8.02	9 13	17	$\mathcal{B}$	7	0	12	11	9
3	7.45	15 18	13	$\mathcal{C}$	11	12	0	15	11
4	7.25	10 13	14	$\mathcal{D}$	8	11	15	0	14
5	6.93	10 11	14	$\mathcal{E}$	8	9	11	14	0

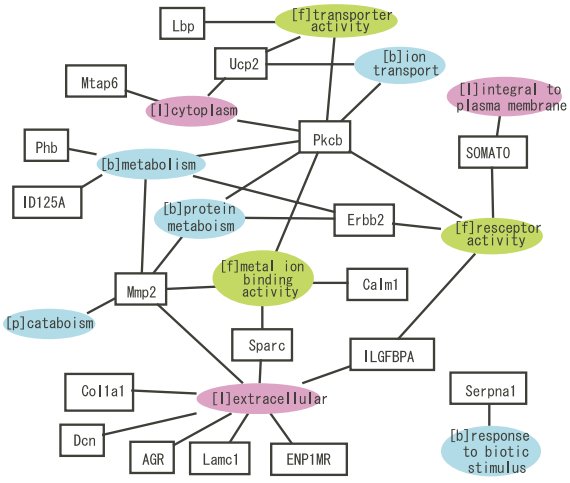

図 4: 最上位の視点における遺伝子-属性ネットワーク
四角は遺伝子，楕円は属性を表す．[p]=生物学的プロセス,
[f]=分子機能,[l] 局在位置. 固有属性=(4 6) 遺伝子数=18.

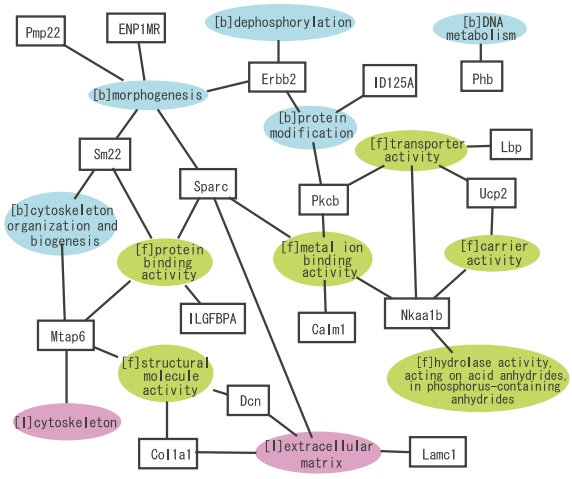

図 5: ２番目の視点における遺伝子-属性ネットワーク
四角は遺伝子，楕円は属性を表す．[p]=生物学的プロセス,
[f]=分子機能,[l] 局在位置. 固有属性=(9 13) 遺伝子数=17.

4.3 視点毎のネットワーク表示による解析

提案手法を用いたことで，視点毎に異なる遺伝子と異なる属性をもつ遺伝子-属性ネットワークが抽出された．今回は，紙面の都合上により上位２視点の遺伝子間関係を俯瞰する．

4.31 **最上位視点の遺伝子-属性ネットワーク表示:** 固有属性として 4 軸，6 軸を選択し，18 個の遺伝子を選択した最上位視点のネットワーク表示を図 4 に示す．最上位視点では，細胞質 (cytoplasm) と細胞外 (extracellular) の間を，主に代謝 (metabolism) と金属イオン結合活性 (metal ion binding activity) により結合している．この視点は，細胞質と細胞外に関する遺伝子の対比という視点において、代謝現象と金属イオン結合活性が介在していると解釈できる．

4.32 **２番目の視点の遺伝子-属性ネットワーク表示:** 固有属性として 9 軸，13 軸を選択し，17 個の遺伝子を選択した，第 2 位視点のネットワーク表示を図 5 に示す．この視点では，形態形成 (morphogenesis) という生命現象の関連遺伝子に対置して，細胞骨格 (cytoskeleton) 関連遺伝子や細胞外マトリクス (extracellular matrix) 関連遺伝子、金属イオン結合活性を持つ遺伝子が存在するという見方を提供している．

以上のように，各々の視点では，選択した遺伝子間の違いを際だたせるように属性選択が行われた．

5. おわりに

ラットの腎臓で胚発生中期に特異的に発現する遺伝子集団に対し，特異値分解により属性の前処理を行った後，ETMIC 状況分解で抽出した状況を視点と見なし，視点毎の遺伝子-属性ネットワーク表現により遺伝子間関係を俯瞰した．上位二つの視点を調べたところ，遺伝子同士の違いを強調する属性が選択され，遺伝子間の比較に有用な複数の視点となっている．これより，状況分解技術を事例比較の視点の推薦技術として利用できる可能性を示せた．

今後は，状況の汎化能力を用いた属性の補間や，GO 上のカテゴリの積極的な使い分けなどを行いたい．また，ETMIC 値は客観的指標であるため，ユーザの領域知識や視点を組合せることができない [3]．そこで，主観的指標からの結果評価や，同時にそれらとの統合についても検討したい．

謝辞 本研究の推進にあたり，ネットワーク表示の生物学的解釈に御助力頂いた (株) 富士通研究所の丸橋 弘治氏, 実験環境構築に御助力頂いた富士通 (株) の山下辰博氏に感謝する．

参考文献

[1] 富士通 (株). Genesphere. http://venus.netlaboratory.com/material/messe/xminer.

[2] 廣澤桂, 長谷川義和, 豊田哲郎, 小長谷明彦. シロイヌナズナマイクロアレイデータを用いた GSCope パスウェイランキングの検証. 第 26 回日本分子生物学会年会, pp. 3PA–075, 2003.

[3] 大崎美穂, 佐藤芳紀, 北口真也, 横井英人, 山口高平. 医療データセットを用いたルールの興味深さ指標の評価. 人工知能学会 第 64 回知識ベースシステム研究会, pp. 87–94, 2004.

[4] Robert O. Stuart, Kevin T. Bush, and Sanjay K. Nigam. Changes in global gene expression patterns during development and maturation of the rat kidney. In *Proc Natl Acad Sci U S A.*, Vol. 8;98, pp. 5649–54, 2001.

[5] 瀧浩平, 竹中要一, 松田秀雄. 遺伝子の発現プロファイルと Gene Ontology による注釈情報を統合した遺伝子制御ネットワークの推定手法. 第 26 回日本分子生物学会年会, pp. 3PA–54, 2003.

[6] 山川宏, 宮本裕司, 馬場孝之, 岡田浩之. 自律学習における CITTA による部分知識の結合. 第 10 回マルチ・エージェントと協調計算ワークショップ (MACC2001), December 2001.

[7] 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, 2004. (掲載予定).