厚生労働科学研究費補助金

創薬基盤推進研究事業

（トキシコゲノミクス研究）

トキシコゲノミクスのための遺伝子ネットワーク解析法の開発

平成19年度　総括研究報告書

主任研究者　奥野　恭史

平成20（2008）年　4月

目　次

厚生労働科学研究費補助金（創薬基盤推進研究事業）
総括研究報告書

トキシコゲノミクスのための遺伝子ネットワーク解析法の開発

主任研究者　奥野　恭史　京都大学薬学研究科 准教授

研究要旨
本研究は、化合物による生体系への影響を、薬物作用遺伝子群や毒性関連遺伝子群の遺伝子発現ネットワーク（分子ネットワーク）の変動として解析する高精度な薬物安全評価アルゴリズムの開発と実用化を目的としている。すなわち、化合物を作用させた各種細胞のDNA マイクロアレイ実験による網羅的遺伝子発現データから、バイオインフォマティクス手法によって薬物毒性特有の遺伝子発現ネットワークを構築し、薬物毒性を反映する遺伝子ネットワークのパターンとして薬物安全性を評価するトキシコゲノミクス計算法の確立を目指す。一般に遺伝子発現ネットワーク解析手法は、生命現象と分子メカニズムを繋げる有力なバイオインフォマティクス手法であり、本手法をトキシコゲノミクスへ適用することにより毒性と遺伝子発現パターンの相関を鮮明にし、予測精度の向上と毒性分子メカニズムの解明という成果をもたらすものと予想される。近年、マイクロアレイ解析のトキシコゲノミクスや薬理ゲノミクスへの適用に対する期待が国内外を問わず高まって来ているが、トキシコゲノミクスや薬理ゲノミクスへのネットワーク解析手法の適用・成功事例は今のところ世界的に皆無である。従って、本研究は、ネットワーク解析法のトキシコゲノミクスへの応用という試み自身の独創性を有しており、薬物の毒性評価とその毒性発現の分子メカニズムに関する知見をも同時に提供する強い特色を有している。最終年度にあたる平成19年度の研究進捗も、計画通り極めて順調に進捗した。具体的には、ヒト肝癌由来細胞株であるHepG2細胞を用いて、肝毒性を有するインスリン抵抗性改善薬トログリタゾン(TGZ)および、同じチアゾリジン系抗糖尿病薬物であるが肝毒性が少ないとされるピオグリタゾン(PGZ)のそれぞれの薬効ならびに肝毒性を評価するために、マイクロアレイデータから遺伝子ネットワークの構築を行った。その結果、トログリタゾンの薬物細胞毒性に特異的な遺伝子ネットワーククラスターを見いだし、WST-1検証実験を行うことにより、毒性関連遺伝子群の同定に成功した。本研究によって開発される高精度な毒性予測手法は、医薬品開発における早期毒性予測による医薬品開発期間・コストの軽減化と、国民における医薬品使用の安全性の向上を実現するものと期待できる。

A．研究目的

ポストゲノム時代の今日、生命科学研究の関心は遺伝子・蛋白質の単一の機能解明にとどまらず、これらを統合した生命システム全体の機能解明へと移行しつつある。医薬の観点においても、化合物と直接作用する単一のタンパク（遺伝子）の機能変化からその薬理活性の全てを語り尽くそうとする従来の考え方では不十分であり、化合物と特定のタンパク質との直接的な作用が周辺や下流の遺伝子たちにどのような影響を及ぼすのかという化合物と生体系との作用を多種多様な生命システムの変動とみなす新しい概念の導入および解析法の開発が必須である。そこで、本研究は、薬物標的分子や毒性原因遺伝子などの単一遺伝子（タンパク質）を対象にした従来の解析手法から逸脱し、化合物による生体系への影響を薬物作用遺伝子群や毒性関連遺伝子群の遺伝子発現ネットワークの変動として解析する高精度な薬物安全評価アルゴリズムの開発と実用化を目的としている。すなわち、薬物曝露におけるマイクロアレイ実験による網羅的遺伝子発現データから、バイオインフォマティクス手法によって薬物毒性特有の遺伝子発現ネットワークを構築し、薬物毒性を反映する遺伝子ネットワークのパターンとして薬物安全性を評価するトキシコゲノミクス計算法の確立を目指す。一般に遺伝子発現ネットワーク解析手法は、生命現象と分子メカニズムを繋げる有力なバイオインフォマティクス手法であり、本手法をトキシコゲノミクスへ適用することにより毒性と遺伝子発現パターンの相関を鮮明にし、予測精度の向上と毒性分子メカニズムの解明という成果をもたらすものと予想される。研究計画としては、初年度（平成17年度）にはアルゴリズム開発のための実験データを収集し、2 年次（平成18年度）にはアルゴリズムの実質的開発と追加マイクロアレイ実験を行い、最終年度となる平成19年度には、先の2年間で開発検討行ってきたネットワーク解析手法の有効性の評価とそれによる毒性メカニズムの解明を目指す。最終的には、マイクロアレイデータを入力とした時に、薬理作用遺伝子ネットワーク、毒性作用遺伝子ネットワークを自動抽出し、その薬理効果・毒性作用を予測するシステムとする。本研究によって開発される高精度な毒性予測手法は、医薬品開発における早期毒性予測による医薬品開発期間・コストの軽減化と、国民における医薬品使用の安全性の向上を実現するものと期待できる。

B．研究方法

1．チアゾリジン系薬物作用遺伝子ネットワークの構築

トログリタゾンとピオグリタゾンのマイクロアレイデータから、各濃度の時系列遺伝子発現プロファイルにおいて発現量が３倍以上変動した遺伝子を抽出した。全濃度での時系列における発現変動パターンに基づいてクラスタリングした後に濃度毎に分割する事により、各薬剤及び各濃度の遺伝子ネットワークと薬剤及び濃度間での遺伝子ネットワークの構築を行った。具体的には、遺伝子の経時的な発現パターンが薬物刺激によってどの様な変動を受けるのかを分類する為、３倍以上変動している遺伝子群の発現変動プロファイルについてk-means 法で120種類のクラスターに分類した。次に、各クラスター内の代表的な遺伝子変動パターンを得るため、各クラスターのベクトル重心

$V_{CL=n}=(v_{t=0h},v_{t=2h}^{C=0},\cdots,v_{t=24h}^{C=0},v_{t=2h}^{C=3},\cdots,v_{t=24h}^{C=3},v_{t=2h}^{C=100},\cdots,v_{t=24h}^{C=100})$ を計算し、濃度C=(vehicle;0 μM, 3 μM, 100 μM)毎のベクトル $V_{CL=n}^{C=m}$, $(n=1,\cdots,120,m=0,3,100)$ に分割した。このとき、クラスター重心は薬物毎に計算した。さらに、クラスターを固定し、各濃度間(C=0 vs C=3, C=0 vs C=100, C=3 vs C=100) での相関係数 $CC_{CL=n}^{C=(m_i,m_j)}=Cov(V_{CL=n}^{C=m_i},V_{CL=n}^{C=m_j})$, $(i\neq j)$ を計算し、次に、濃度を固定し各クラスター重心の相関係数 $CC_{CL=(n_k,n_l)}^{C=(m_i,m_j)}=Cov(V_{CL=n_k}^{C=m_i},V_{CL=n_l}^{C=m_j})$, $(k\neq l)$ を計算した。続いて、各クラスターの代表発現パターンに基づいて、「120クラスター」×「3濃度」のクラスター間の相対距離を主座標分析法により座標化し、遺伝子ネットワーク及び変動パターンの変化を図示した。

2.チアゾリジン系薬物作用遺伝子ネットワーク解析による毒性評価

上記、3倍以上変動した遺伝子の全てのペアについて相関係数(ピアソンの積率相関係数)を算出した。ここで、閾値以上(p>0.97)の相関係数をもつ遺伝子ペアについて、リンクを繋げることにより各薬物、各濃度におけるrelevant networkを構築した。それぞれの遺伝子ネットワークの構造特徴を比較することによって毒性評価ができるかどうかの検討を行った。

3.毒性特異的遺伝子ネットワーククラスターの実験検証

[細胞培養]: 10% (v/v) fetal bovine serum (GIBCO)、100 U/mL penicillin (GIBCO) および100 μg/mL streptomycin (GIBCO) を含有するDulbecco's Modified Eagle Medium (GIBCO) を用い、コラーゲンコートディッシュにて37℃、5% CO2の条件で細胞培養を行った。

[WST-1細胞増殖測定]: WST-1を基質とした脱水素酵素の活性を指標に、細胞増殖を測定した。HepG2細胞1x104 cellsをvehicle (1% dimethylsulfoxide)及び各濃度 (3 μM, 100 μM) のトログリタゾンもしくはピオグリタゾンで刺激し、各０，２，６，１２，２４時間後に飽和WST-1溶液 (Roche) を加えた。37℃、5 % CO2の条件下で2時間反応させた後、還元型WST-1 (Formazan) の増加量を測定するため、マイクロプレートリーダーで450nmの吸光度を測定した。さらに、バックグラウンドシグナルとして650nmの吸光度を測定した。

[遺伝子過剰発現]:トログリタゾンの薬物細胞毒性に特異的な遺伝子ネットワーククラスターに含まれている遺伝子を発現ベクター (pCMV-sport6) に組み込み、それらのベクターを単独もしくは複数を組み合わせてHepG2細胞に遺伝子導入した。遺伝子導入から２４時間と４８時間後に導入遺伝子 (群) の細胞増殖における効果を調べるため、WST-1法を用いて解析を行った。

(倫理面への配慮)

本研究は計算機アルゴリズムの開発と培養細胞を用いたインビトロ実験のみであり、倫理面に関する問題は一切無い。

C. 研究結果

1.チアゾリジン系薬物作用遺伝子ネットワークの構築

Figure 1において抽出された合計1,428個の遺伝子において遺伝子ネットワークの構築を行い、薬物及び薬物濃度間における遺伝子ネットワークの変動を解析した(Figure 2)。「120クラスター」×「3濃度」のクラスター間相関を距離として主座標分析法により座標を決定し、遺伝子ネットワーク及び変動パターン

の変化を図示した。各層でのxy平面内での距離は、クラスター間相関を反映しているため、近接しているクラスター同士は類似した変動パターンを持っており、各層間での移動が大きいクラスターは、濃度間での発現変動量が大きい事を示している。各薬剤において特異的な遺伝子のみを含むクラスターから薬剤特異的ネットワークを、遺伝子が両薬剤で発現するクラスターから共通ネットワークを構成した結果、トログリタゾンもしくはピオグリタゾンに特異的な変動を示すクラスター群と、トログリタゾンとピオグリタゾンで共通した変動を示すクラスター群が同定された(Figure 2)。

これらのクラスターからトログリタゾンによる細胞毒性に大きく寄与するクラスターを選定するため、前記の1,428個の遺伝子に対する主成分分析を行い(Figure 3)、その結果をクラスターに反映させた(Table 1)。トログリタゾンおよびピオグリタゾンの全濃度の経時的発現変動に対して主成分分析を行い、3変数に情報を集約したところ、第二主成分においてトログリタゾン毒性濃度である100μMで刺激した際のベクトルとそれ以外の濃度で刺激した際のベクトルが逆の方向性を示した。従って、第二主成分はトログリタゾンによる毒性を示していると判断された(Figure 3)。次に、第二主成分スコアの上位および下位それぞれ１００位以内に含まれる遺伝子を多く含むクラスターを選別した(Table 1)。その結果からトログリタゾンの毒性発現に寄与すると予測された遺伝子を多く含み、トログリタゾンの毒性発現に特異的な発現変動パターンを提示するクラスターとして、クラスター２０(CL20)及び１２(CL12)などが選定された。トログリタゾン特異的なCL20及びCL12はvehicle及び3μMでは殆ど変動が見られないが、100μMにおいてネットワーク上での位置が大きく変動していた(Figure 4)。

２.チアゾリジン系薬物作用遺伝子ネットワーク解析による毒性評価

チアゾリジン系薬物作用ネットワークにおいて、Vehicleでは$\gamma \approx 1.61$でのスケールフリー性が確認されたのに対し、薬物投与時ではネットワークのランダム性が増大し（γが減少し）、ネットワーク全体の特性が変化を受けていることが明らかと成った。特に、トログリタゾン毒性濃度においては、高いエッジ数を持つノードが増大し、ランダムネットワーク性が増大することが確認された。また、トログリタゾン特異的な遺伝子ネットワーククラスタがネットワークパラメータ変動の原因となっていること示された。これにより本研究で開発したネットワーク解析手法が、薬物毒性による遺伝子ネットワーク撹乱の指標となることが示唆された。(Figure 5)

３.毒性特異的遺伝子ネットワーククラスターの実験検証

毒性発現と共に遺伝子発現が増大するCL 20 (Figure 6)に属する遺伝子群が、細胞増殖の抑制に寄与しているかどうかを明らかにするために、これらの遺伝子を過剰発現させた際の細胞増殖に対する影響を観測した (Figure 7a)。CL20に含まれているMAFF (V-maf musculoaponeurotic fibrosarcoma oncogene homolog F)、DDIT3 (DNA-damage-inducible transcript 3)、IFRD1 (Interferon-related developmental regulator 1)、CG018 (Hypothetical gene CG018)の４遺伝子の発現ベクターを単独もしくは同時にHepG2細胞に遺伝子導入した際の細胞増殖への影響をWST-1法により解析した結果、４種類の遺伝子の発現ベクターを同時にHepG2細胞に遺伝子導入した際に最も強い細胞増殖の抑制が遺伝子導入後２４時間と４８時間のいずれにおいても検出

された。MAFF、IFRD1、CG018単独の過剰発現ではコントロールのMockと比較してHepG2細胞の増殖に有意な差は検出されなかったが、DDIT3単独の過剰発現においては遺伝子導入後２４時間においてのみHepG2細胞の増殖が有意に抑制された。以上のことから、今回、トログリタゾンによる肝毒性特異的な遺伝子ネットワーククラスターとして選定されたCL20（MAFF, DDIT3, IFRD1, CG018の４遺伝子）に含まれている遺伝子群が、本検証実験系においても協調的に細胞増殖を抑制していることが示唆された。

また、トログリタゾン100μMで発現抑制されるCL25（Figure 6）から CDKN2C (cyclin-dependent kinase inhibitor 2C), SHCBP1 (SHC SH2-domain binding protein 1), KBTBD7 (Kelch repeat and BTB (POZ) domain containing 7), DTL (denticleless homolog), SKP2 (S-phase kinase-associated protein 2), FEN1 (flap structure-specific endonuclease 1) , RFC5 (replication factor C (activator 1) 5, 36.5kDa) の7つの遺伝子を選別した。トログリタゾン毒性によるこれらの遺伝子の発現抑制の細胞増殖への関与を明らかにするため、これらの遺伝子をHepG2細胞において強制的に発現させた状態でトログリタゾン 100 μMによる刺激を行い、WST-1による細胞増殖を観測した。その結果、48時間後において、SHCBP1 または SKP2導入細胞での有意な細胞増殖が確認された($p < 0.01$)。(Figure 7b) さらに、特筆すべき点として、7遺伝子全てを導入した細胞において、24 時間後および48時間後にトログリタゾン毒性に抵抗してWST-1 Assay の増大が確認された。また、対照実験として、Plasmid のみの導入を行った条件下で、SKP2ないしRFC5導入細胞が細胞増殖の傾向を示す事が発見された($p < 0.01$)(Figure 7c)。以上のことから、トログリタゾンによる肝毒性特異的な遺伝子ネットワーククラスタとして選定されたCL25に含まれている遺伝子群 (CDKN2C, DTL, FEN1, KBTBD7, SHCBP1, SKP2, RFC5の7遺伝子）が、本検証実験系においても協調的に細胞増殖を制御し、トログリタゾン毒性により阻害されていることが示唆された。

D．考察

HepG2細胞とトログリタゾン及びピオグリタゾンを用いた遺伝子ネットワーク解析を行い、主成分分析結果を反映させた結果、トログリタゾンの薬物細胞毒性に特異的なクラスターが同定された。主成分分析のみでは、高変動を示す遺伝子のみが選別され、それ以外の重要な働きを持つ遺伝子群は見落とされる傾向にあるが、本手法では、毒性特異的な主成分の上位／下位の遺伝子を多く含む薬物特異的変動クラスターを選別することにより、薬物毒性特異的に変動したクラスター（遺伝子群）として同定することが可能となった。これらのクラスターにはアポトーシスの感受性に関わる遺伝子や細胞周期に関わる遺伝子に加えて、細胞毒性とは関連がまったく報告されていない酵素類や細胞外マトリックス因子も含まれており、大変興味深い。主成分分析第二主成分における遺伝子順位や遺伝子発現変動の度合いから、CL20及びCL25に注目し。CL20およびCL25に含まれる遺伝子群がシステムとして薬物毒性に寄与しているかどうかを、遺伝子群の過剰発現による実験により検証を行った。

CL20 に属する4遺伝子に対するWST-1アッセイの結果から、DDIT3はトログリタゾン毒性と関連して、細胞増殖を阻害する事が示された。さらに、DDIT3単独ではなく、4つの遺伝子を同時に強制発現させたところ、細胞増殖はより強く阻害された。DDIT3はGADD153 としても知られており、PPAR-gamma転写パスウェイが誘導するNSCLC

細胞の増殖抑制の候補要因である。これまでに、DDTT3とMAFF, IFRD1, CG018が協調的に作用するという知見は報告されていない。すなわち、これら３つ遺伝子がトログリタゾンの毒性に関連する新規関連遺伝子であることが示唆された。また、CL25に属する遺伝子においては、FEN1, RFC5、DTL1 は Cell Cycle プロセスへの関与と、PCNA(Proliferating Cell Nuclear Antigen)との関連性が報告されている。本手法により得られたクラスターにおいて、PCNAは主成分分析第二主成分においてCL25と同様の傾向を示すCL12 に属していることは興味深い事実である。FEN1は、複製や修復時に生成するDNAの分岐に重要な役割を果たしている。また、FEN1はインビトロでPCNAにより活性化されることが知られている。DLT１はDNA合成、細胞周期、細胞分裂, 増殖、分化などにおいて重要な役割を担うと共に、PCNAと共に、DNA損傷後のCDT1分解に関与することが報告されている。RFC5はreplication factor C complexのサブユニットであり、出芽酵母ではDNA複製および有糸分裂(mitosis) と共役することが報告されている。すなわち、これら複数の遺伝子群が、トログリタゾン毒性による薬物毒性にシステムとして関与している事が示唆された。今回同定した毒性特異的遺伝子ネットワーククラスターの作用機序をFigure 8にまとめる。

E．結論

ヒト肝癌由来細胞株であるHepG2細胞を用いて、肝毒性を有するインスリン抵抗性改善薬トログリタゾン(TGZ)および、同じチアゾリジン系抗糖尿病薬物であるが肝毒性が少ないとされるピオグリタゾン(PGZ)のそれぞれの薬効ならびに肝毒性を評価するために、マイクロアレイデータから遺伝子ネットワークの構築を行った。その結果、トログリタゾンの薬物細胞毒性に特異的な遺伝子ネットワーククラスターを見いだし、WST-1検証実験を行うことにより、毒性関連遺伝子群の同定に成功した。今回開発した遺伝子ネットワーク解析法は、GEM-TRENDシステムとして次のURLよりWeb公開も行っている。http://cgs.pharm.kyoto-u.ac.jp/services/network/

F．健康危険情報

特記事項無し

G．研究発表

１．論文発表

1. Niijima, S. and Okuno, Y. “Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, in press

2. Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimto, R. and Feng, C. “GLIDA: GPCR-Ligand Database for Chemical Genomics Drug Discovery - Database and Tools Update.” *Nucleic Acids Research*, 36, D907-12, 2008

3. Kitajima, M., Minowa, Y., Matsuda, H. and Okuno, Y. “Compound-Transporter Interaction Studies using Canonical Correlation Analysis.” *Chem-Bio Informatics J.*, 7, 24-34, 2007

4. Yamamoto, H., Takematsu, H., Fujinawa, R., Naito, Y., Okuno, Y., Tsujimoto, G., Suzuki, A. and Kozutsumi, Y. “Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for glycan biosynthesis enzyme

gene." *PLoS ONE*, 2, e1232, 2007

5. Ikeda, A., Miyazaki, T., Kakizawa, S., Okuno, Y., Tsuchiya, S., Myomoto, A., Saito, SY., Yamamoto, T., Yamazaki, T., Iino, M., Tsujimoto, G., Watanabe, M. and Takeshima, H. "Abnormal features in mutant cerebellar Purkinje cells lacking junctophilins." *Biochem. Biophys. Res. Commun.*, 363, 835-9, 2007

6. Yamazaki, T., Sasaki, N., Nishi, M., Yamazaki, D., Ikeda, A., Okuno, Y., Komazaki, S., and Takeshima, H. "Augmentation of drug-induced cell death by ER protein BRI3BP." *Biochem. Biophys. Res. Commun.*, 362, 971-5, 2007

7. Naito, Y., Takematsu, H., Koyama, S., Miyake, S., Yamamoto, H., Fujinawa, R., Sugai, M., Okuno, Y., Tsujimoto, G., Yamaji, T., Hashimoto, Y., Itohara, S., Kawasaki, T., Suzuki, A., and Kozutsumi, Y. "Germinal center marker GL7 probes activation-dependent repression of N-glycolylneuraminic acid, a sialic acid species involved in the negative modulation of B cell activation." *Mol. Cell Biol.*, 27, 3008-22, 2007

2．学会発表

1. 日本薬学会128年会　日本薬学会奨励賞受賞講演「バイオ空間とケミカル空間の包括的相関解析とそのインシリコ創薬への研究展開」（2008.3.28）

2. 日本薬学会128年会　日本薬学会・日本学術会議薬学委員会共催シンポジウム「バイオインフォーマティクスの薬学研究・薬学教育への応用と展開」 Bio-Informatics for Pharmaceutical Sciences and Pharmacy Education　「バイオ空間とケミカル空間の包括的相関解析とそのインシリコ創薬への研究展開」（2008.3.26）

3. 第3回三重ゲノム創薬フォーラム「薬学研究におけるアレイインフォマティクス」（2008.2.15）

4. 平成19年度 第2回産業情報交流会「ケミカルゲノミクスに基づくインシリコ化合物探索他」（2007.10.22）

5. 新産業を創る先端科学技術フォーラム2007「ポストゲノム創薬のための新技術」セッション「ケミカルゲノミクスに基づく創薬インフォマティクス」（2007.10.18）

6. 第66回日本癌学会学術総会 International Session「Chemical Genomics for Cancer Research」「Knowledge Discovery and Data Mining in Chemical Genomics」(2007.10.3)

7. 日本ケミカルバイオロジー研究会　第2回年会「ケミカルゲノミクス情報を利用したGPCRのin silicoリガンド探索手法の開発」（2007.5.10）

H. 知的財産権の出願・登録状況

1．特許取得

1. 公開番号WO２００７／００４４７９　A（特開２００７－１１７５２）、「データ処理装置、データ処理プログラム、それを格納したコンピュータ読み取り可能な記録媒体、およびデータ処理方法」、（国際）２００６年6月２８日；２００７年1月１１日公開（国内）２００５年6月３０日出願；２００７年1月１８日公開、

出願人　京都大学、発明者　奥野恭史、辻本豪三、梁智允、種石慶

2. 公開番号WO２００７／１３９０３７ Ａ１；ＰＣＴ／ＪＰ２００７／０６０７３６（特願２００６－１４７４３３）、「ケミカルゲノム情報に基づく、タンパク質－化合物相互作用の予測と化合物ライブラリーの合理的設計」、（国際）２００７年５月２５日出願；２００７年１２月６日公開（国内）２００６年５月２６日出願、出願人　京都大学、発明者　奥野恭史、種石慶、辻本豪三

２．実用新案登録
無し

３．その他
無し

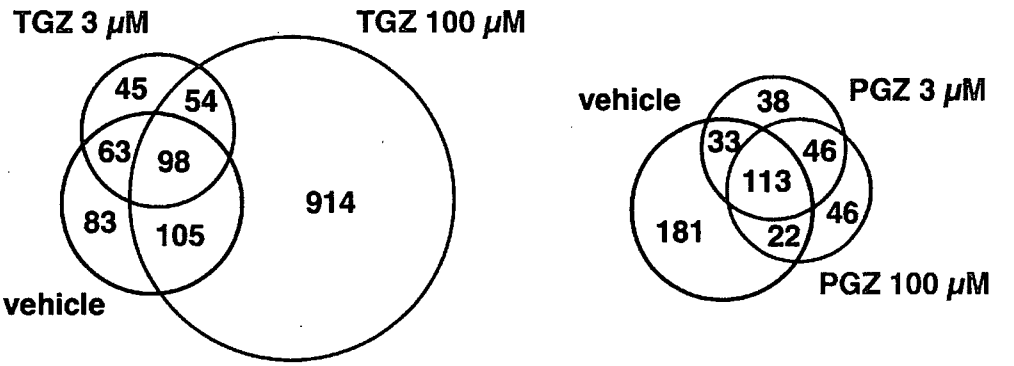

Figure 1. Overlapping sets of genes differently expressed in each concentration of TGZ or PG Z. Numbers in the overlapping region of the Venn diagram represent shared genes. Genes were s elected if the ratio of the relative expression level in time course was larger or smaller th an 3.0.

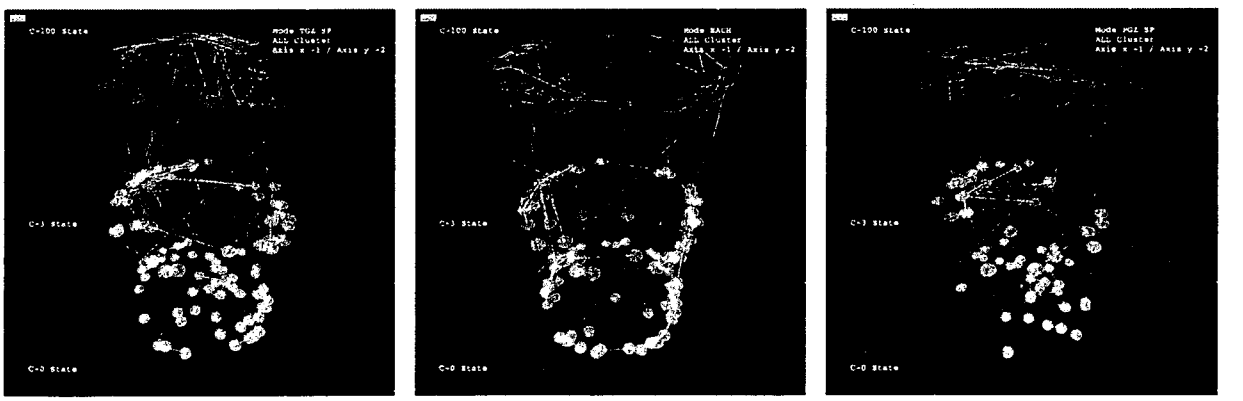

Figure 2. Genes networks by the stimulation of TGZ and PGZ in HepG2 cells. The TGZ-specific n etwork (left panel), PGZ-specific network (right panel) and the network common between TGZ an d PGZ (middle panel) were displayed.

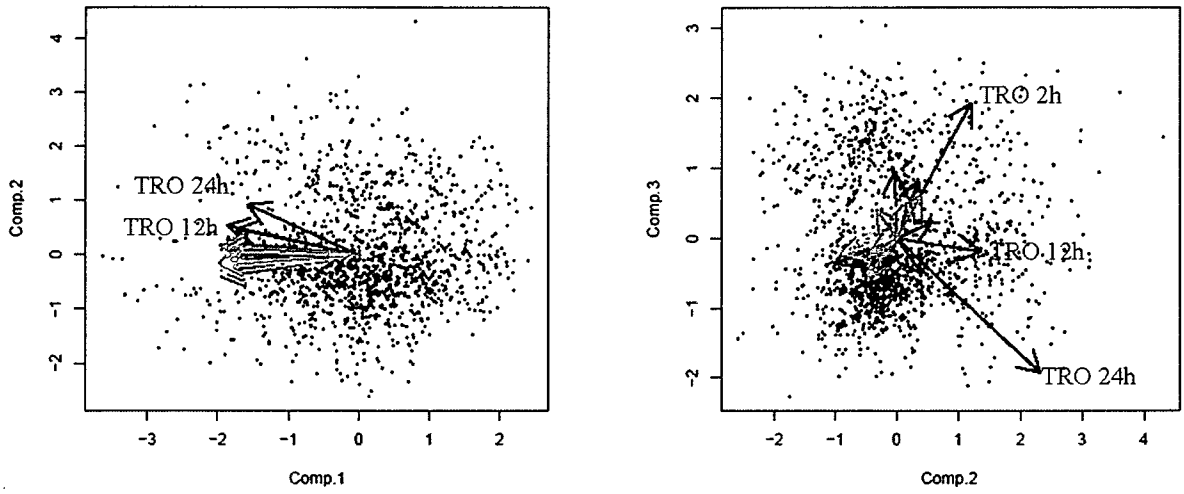

Figure 3. Principal component analysis of gene expression profiles in TGZ- or PGZ-stimulated HepG2 cells. Genes were selected if the ratio of the relative expression level in time course was larger or smaller than 3.0. The axes correspond to principal component (Comp.) 1, Comp. 2 and Comp. 3. Red arrows represent vectors of 100 μM TGZ-stimulated samples, and other color arrows represent samples except for TGZ 100 μM.

Table 1. Relation between the principal component analysis and each cluster of genes network.

Cluster Number	Number of genes		
	PCA (Comp. 2)	TGZ	PGZ
Upper 100 genes of Comp. 2			
39	*18*	*29*	*1*
82	*17*	*3*	*31*
20	*11*	*18*	*0*
59	*10*	*19*	*0*
51	*10*	*11*	*0*
Lower 100 genes of Comp. 2			
12	*30*	*36*	*0*
50	*14*	*1*	*33*
21	*13*	*10*	*26*
25	*12*	*15*	*0*
49	*10*	*11*	*0*
11	*10*	*0*	*25*

Figure 4. CL20 and CL25 in the genes network of TGZ-stimulated HepG2 cells.

Table 2.

	# of cluster	# of genes		
		TGZ	PGZ	TGZ ∩ PGZ
Drug Specific Cluster				
TGZ Specific	52 clusters	1070	107(*)	0
PGZ Specific	29 clusters	58(*)	827	0
Non-specific Cluster	45 clusters	358	601	209

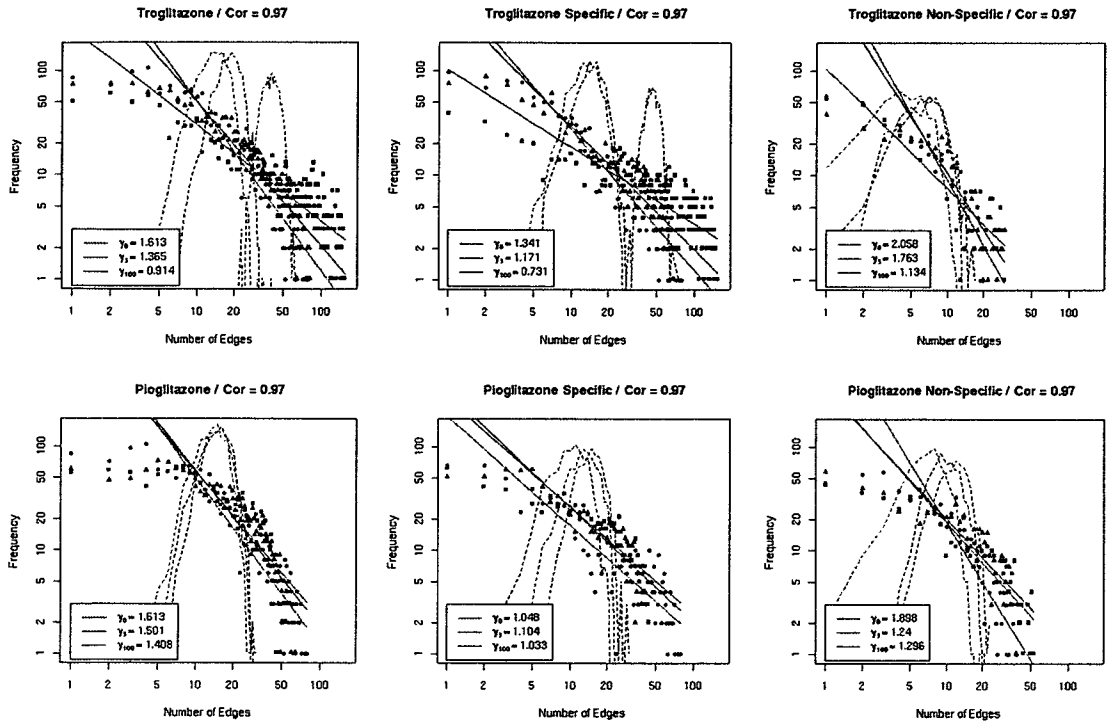

Figure 5. Distribution of the number of links for TGZ and PGZ networks. (x axis, log10 scale) compared with number if links (y axis, log10 scale) in the vehicle network(green round) , 3 μM sti mulated networks (blue triangle) and 100 μM stimulated (red square). The each color line depict the least squares fit of the data to a linear line in the log-log plot. The each dashed-line depicts the distr ibution of random networks.

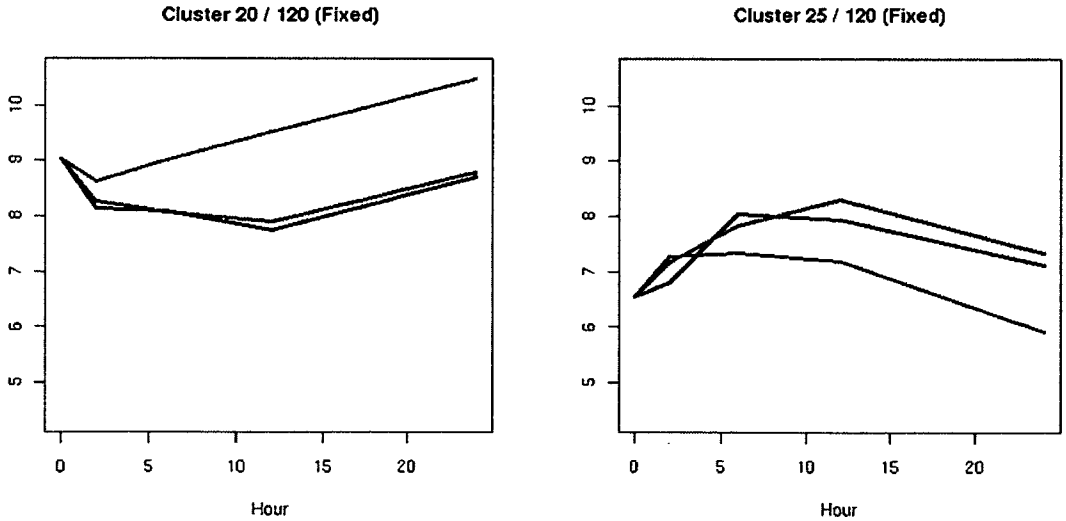

Figure 6. The gene expression patterns of CL20 and CL25.
Black line: Vehicle, blue line: 3 μM, Red line: 100 μM

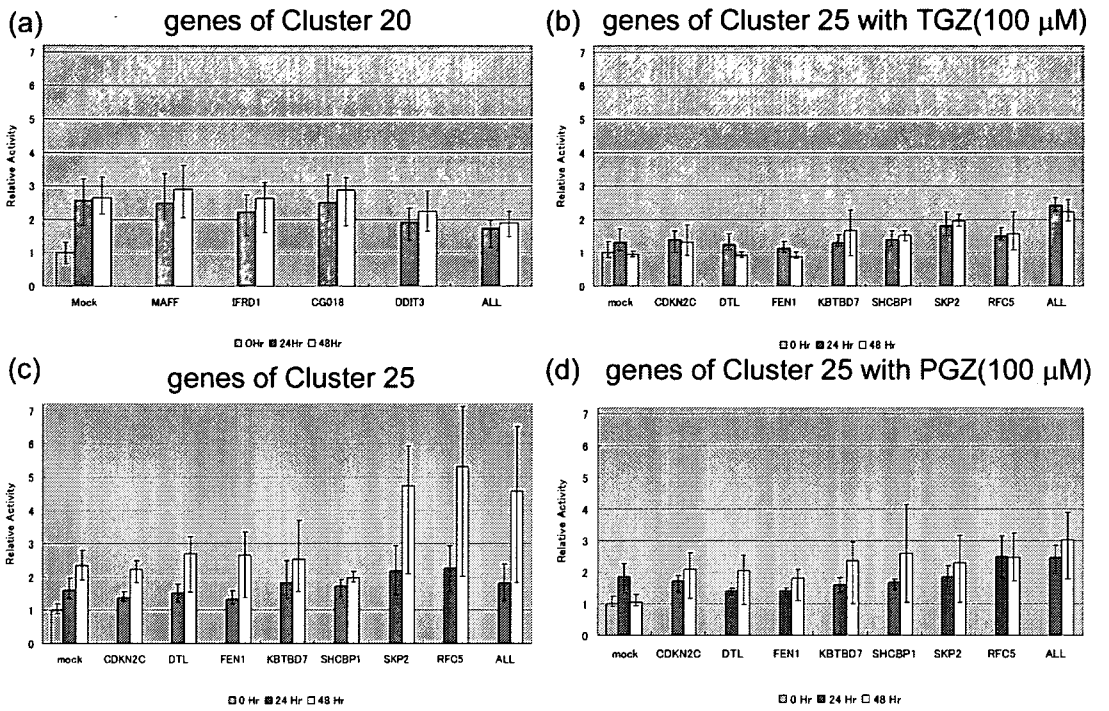

Figure 7. WST-1 assay of CL20 and CL25.

(a) HepG2 cells were transfected with the expression vector(s) for mock, MAFF, DDIT3, IFRD1, CG018, or four genes and assayed at the indicated times. Data are shown as the mean ± SEM (n = 8). Data are represented as fold of the value at 0 h, (*, $p < 0.05$ for DDIT3 at 24 h and four genes at 24 and 48 h vs mock).

(b, c, d) HepG2 cells were transfected with the expression vector(s) for mock, CDKN2C, DTL, FEN1, KBTBD7, SHCBP1, SKP2, RFC5 or seven genes and assayed at the indicated times. (b) Troglitazone stimulated HepG2 cell, (c) control(Plusmid only) (d) Pioglitazone stimulated HepG2 cell. Data are shown as the mean ± SEM (n = 8). Data are represented as fold of the value at 0 h, (b: $p < 0.01$ for SHCBP1 or SKP2 at 48 h and seven genes at 24 and 48 h vs mock. c: $p < 0.01$ for SKP2 or RFC5 at 48 h).

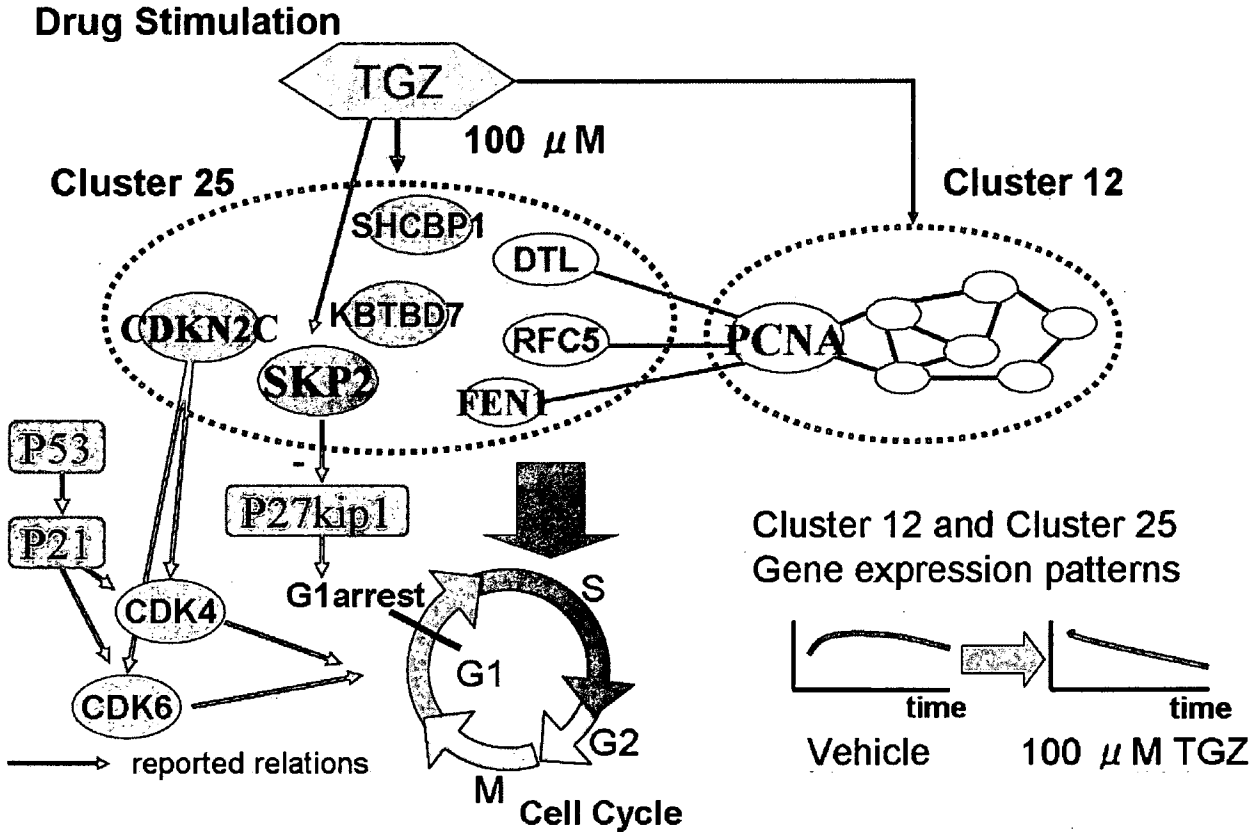

Figure 8. Troglitazone toxicities and cell cycle arrest mechanism.

研究成果の刊行に関する一覧表

書籍

著者氏名	書籍全体の編集者名	書籍名	出版社名	出版地	出版年
奥野恭史ほか	石渡信一・桂勲・桐野豊・美宅成樹	生物物理学ハンドブック	㈱朝倉書店	日本	2007

雑誌

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
Niijima, S. and Okuno, Y.	Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection.	**IEEE/ACM Transactions on Computational Biology and Bioinformatics**	-	-	in press
Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimto, R. and Feng, C.	GLIDA: GPCR-Ligand Database for Chemical Genomics Drug Discovery - Database and Tools Update.	**Nucleic Acids Research**	36	D907-12	2008
Kitajima, M., Minowa, Y., Matsuda, H. and Okuno, Y.	Compound-Transporter Interaction Studies using Canonical Correlation Analysis.	**Chem-Bio Informatics J.**	7	24-34	2007
Yamamoto, H., Takematsu, H., Fujinawa, R., Naito, Y., Okuno, Y., Tsujimoto, G., Suzuki, A. and Kozutsumi, Y.	Correlation index-based responsible-enzyme gene screening (CIRES), a novel DNA microarray-based method for glycan biosynthesis enzyme gene.	**PLoS ONE**	2	e1232	2007
Ikeda, A., Miyazaki, T., Kakizawa, S., Okuno, Y., Tsuchiya, S., Myomoto, A., Saito, SY., Yamamoto, T., Yamazaki, T., Iino, M., Tsujimoto, G., Watanabe, M. and Takeshima, H.	Abnormal features in mutant cerebellar Purkinje cells lacking junctophilins.	**Biochem. Biophys. Res. Commun.**	363	835-9	2007
Yamazaki, T., Sasaki, N., Nishi, M., Yamazaki, D., Ikeda, A., Okuno, Y., Komazaki, S., and Takeshima, H.	Augmentation of drug-induced cell death by ER protein BRI3BP.	**Biochem. Biophys. Res. Commun.**	362	971-5	2007

Naito, Y., Takematsu, H., Koyama, S., Miyake, S., Yamamoto, H., Fujinawa, R., Sugai, M., Okuno, Y., Tsujimoto, G., Yamaji, T., Hashimoto, Y., Itohara, S., Kawasaki, T., Suzuki, A., and Kozutsumi, Y.	Germinal center marker GL7 probes activation-dependent repression of N-glycolylneuraminic acid, a sialic acid species involved in the negative modulation of B cell activation.	**Mol. Cell Biol.**	27(8)	3008-22	2007

その他

著者氏名	執筆タイトル名	掲載誌名	巻号	ページ	出版年
奥野　恭史ほか	ケミカル・バイオ情報に基づく創薬インフォマティクス研究	Pharma VISION NEWS	No. 9	13-16	2007
薬事日報　2008. 3. 21　掲載記事					
薬事日報　2007. 5. 1　掲載記事					

研究成果の刊行物・別刷

Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection

Satoshi Niijima and Yasushi Okuno

Abstract— **Until recently, numerous feature selection techniques have been proposed and found wide applications in genomics and proteomics. For instance, feature/gene selection has proven to be a powerful tool for biomarker discovery from microarray and mass spectrometry data. While supervised feature selection has been explored extensively, there are only a few unsupervised methods that can be applied to exploratory data analysis in which class information is unavailable. In this paper, we address the problem of unsupervised feature selection. First, we extend Laplacian linear discriminant analysis (LLDA) to unsupervised cases. Second, we propose an efficient algorithm for computing LLDA. Finally, a new unsupervised feature selection algorithm, called LLDA-based Recursive Feature Elimination (LLDA-RFE), is presented. We apply LLDA-RFE to several public datasets of cancer microarrays and compare its performance with those of state-of-the-art unsupervised methods, Laplacian score and SVD-entropy, and of a supervised filter method, Fisher score. Our results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. It performs even better than Fisher score for some of the datasets, despite the fact that LLDA-RFE is fully unsupervised.**

Index Terms— **Feature selection, linear discriminant analysis, graph Laplacian, microarray data analysis.**

I. Introduction

In recent years, feature/gene selection methods have been widely used in genomics and proteomics to handle a deluge of data produced by high-throughput technologies such as microarray and mass spectrometry. In microarray studies, for instance, a small fraction of genes typically exhibit significant differential expression among tens of thousands of genes whose expression levels are measured simultaneously. Thus, it is of great importance to identify genes relevant to a biological phenomenon of interest and to characterize their expression profiles. Gene selection can be useful for multiple purposes: to save computational costs of subsequent analysis by reducing the number of genes; to improve the prediction performance of classifiers by using discriminative genes only; and to identify informative genes for further investigation of their biological relevance. Specifically, gene selection has proven to be a powerful tool for biomarker discovery, i.e. searching for potential marker genes contributing to classification of cancer subtypes or prediction of clinical outcomes, which leads to more reliable diagnosis and better treatments of cancer.

To date, numerous techniques for feature selection have been developed [12] and also been applied successfully to the analysis of biological data with many features. In contrast to supervised feature selection, however, unsupervised feature selection has not yet been explored extensively. Indeed, there have been only a few unsupervised methods proposed until recently [7], [14], [28], [30]. Unsupervised feature selection is of great use particularly for class discovery, where class information is unavailable. For instance, clustering is usually performed to find clusters in microarray samples on the basis of the expression profiles of all genes, but the clusters so obtained can be obscured by the large number of irrelevant genes. Therefore, unsupervised feature selection is essential to the exploratory analysis of biological data. Moreover, even when class labels are provided by external knowledge, but may be unreliable or mislabeled, overfitting can be alleviated by performing feature selection in an unsupervised manner. It is obviously more challenging to identify features that reveal underlying cluster structures in the samples than to find those exhibiting similar patterns across all the samples.

To address this problem, we propose a new unsupervised feature selection method, called Laplacian linear discriminant analysis-based Recursive Feature Elimination (LLDA-RFE). LLDA-RFE is closely related to Laplacian score [15], which is also based on graph Laplacian and can be applied in an unsupervised manner. The major difference is that, whereas Laplacian score is a univariate approach, LLDA-RFE is multivariate, allowing for selecting features that contribute to discrimination in combination with other features. Recently, Wolf and Shashua [30] proposed the $Q - \alpha$ algorithm, taking advantage of the spectral properties of the graph Laplacian of features. While the $Q - \alpha$ algorithm has an interesting property that the sparsity of features naturally emerges, it does not scale well to the feature size. Also, it involves iterative computations on a matrix of the feature size in a least-squares optimization process to ensure a local maximum solution. In contrast, our proposed algorithm for LLDA-RFE is computationally tractable and has a global maximum solution. It is shown that LLDA includes the maximum margin criterion (MMC) [18] as a supervised case. Although LLDA-RFE is a natural extension of MMC-RFE, the proposed algorithm need not reduce dimensionality before applying LLDA, unlike the MMC-RFE algorithm proposed previously [23].

We compare the performance of LLDA-RFE with those of state-of-the-art unsupervised feature selection methods, Laplacian score and SVD-entropy [28], on seven public datasets of cancer microarrays. The performances of these methods are evaluated by their capability of identifying discriminative genes without using class information. We also compare the performance between LLDA-RFE and a supervised filter method, Fisher score [8], [15]. Experimental results demonstrate that LLDA-RFE outperforms Laplacian score and shows favorable performance against SVD-entropy. Despite the fact that LLDA-RFE is fully unsupervised, it performs even better than Fisher score for some of the datasets.

The rest of this paper is organized as follows: In Section II, we give outlines of LDA and the MMC. We then introduce LLDA and extend it to unsupervised cases in Section III. An efficient algorithm for LLDA is also proposed. We present the LLDA-RFE

Manuscript received XXX XX, 2007; revised XXX XX, 2007.

S. Niijima and Y. Okuno are with the Department of PharmacoInformatics, Frontier Education Center, Graduate School of Pharmaceutical Sciences, Kyoto University, 46-29 Yoshida Shimoadachi-cho, Sakyo-ku, Kyoto 606-8501, Japan.