2．トキシコゲノミクス
1）Percellome Project

菅野　純・北嶋　聡・相﨑健一・五十嵐勝秀・中津則之
高木篤也・種村健太郎・小川幸男・児玉幸夫・関田清司

毒性学（トキシコロジー）は身の回りの物質の毒性（有毒性）を予測し，それらの曝露による被害を未然に防ぐ研究分野である。その精度向上を目的にトキシコゲノミクス研究を開始した。マイクロアレイから細胞1個あたりのmRNAコピー数を得るPercellome法を開発し，遺伝子発現変動を投与量および時間の関数として三次元曲面（surface）データとして可視化し，これを基本に網羅的解析法を独自に開発した。単回投与による肝の初期反応データを中心に延べ約2億超データを採取し，毒性カスケード解明の糸口となる変動遺伝子情報を蓄積しつつある。

はじめに

毒性学（トキシコロジー）は人の健康・安全を確保することを目的とし，外来性の物質（化学物質など）が生体に進入した際の生体反応を記述し理解することにより，身の回りの物質の毒性（有害性）を予測し，それらの曝露による被害を未然に防ぐとともに，曝露された場合の治療法を開発する研究分野である。この有害性予測の精度向上をめざした分子生物学レベルでの毒性研究は，遺伝子，遺伝子発現，タンパク合成，タンパク修飾などのあらゆる段階を対象とする。その際，探索的な科学研究と異なる点として，毒性学には「予期せぬ事態」を見逃がさない網羅性の確保が要求されることが挙げられる。われわれはその第一段階として，遺伝子発現をほぼ全遺伝子についてカバーする高密度cDNAマイクロアレイによる毒性学的トランスクリプトーム研究，すなわちトキシコゲノミクス研究を開始した。

Ⅰ．トキシコゲノミクス研究の目的と最終目標

従来の毒性学は，生物個体が「ブラックボックス」であっても，それに対する「入力」としての化学物質が引き起こす毒性症状をそれからの「出力」として記述し，入出力の関連を体系化することで安全性確保に貢献してきた。われわれのトキシコゲノミクス研究は，ブラックボックスの中身を遺伝子発現カスケードの面から解明することにより生体反応メカニズムに基づいた分子毒性学を構築することを目的としている。そして，究極の目標としてコンピュータ内のバーチャルマウスやバーチャル人間の完成を掲げ，そこへ向う過程で実験動物からヒトへの外挿の精度向上を実現しようとするものである。上述のごとく網羅性を重んじるため，得たデータを既知情報により分類・解

key words
トキシコゲノミクス，分子毒性学，遺伝子発現カスケード，
標準化，Percellome法，三次元多層（Millefeuille surface）データ

析するアプローチ（いわゆるphenotypic anchoring）は最後に回し，トランスクリプトーム情報そのものの中から，生物学的に有意な反応カスケードを「教師なしクラスタリング手法（unsupervised clustering）」などを駆使して抽出するアプローチをとることとした。これはちょうど，電子顕微鏡写真が世に現れた時の状況になぞらえることができる。すなわち，光学顕微鏡では見えない「もの」が新たに見えるようになったわけであるが，それが何であるかは光学顕微鏡像を参照しても簡単にはわからない。電子顕微鏡像を解釈しコンセンサスとしての教科書（図譜など）ができあがって初めて日常的に利用されるようになったわけであり，教科書を完成する作業自体が1つの研究分野をなしたという歴史がある。繰り返しになるが，われわれのめざすトキシコゲノミクスと従来の毒性学との関係は電子顕微鏡と光学顕微鏡の関係にあり，実用化に向けての教科書作成にあたる基礎研究が必要である。

話は前後するが，毒性学研究は，人体実験が可能な一部の状況（成人を対象とした医薬品の臨床試験など）を除いて，ヒトの身代わりとしての実験動物を用いて行われる。上記の目標を達成するために，われわれは動物種としてマウスを選択したが，その理由は，遺伝子情報がヒトに次いで豊富であること，および遺伝子改変マウスが利用可能であり，それから得られる情報（下流カスケードの情報，ノックアウトマウスが示す欠落症状など）が客観的な遺伝子発現カスケードの描出と毒性学的意義づけに大きく貢献することが予見されたことによる。それでも，複数の実験から得られる大量の実験データを蓄積したうえでの横断的な解析が必須となる。そのためには，データの標準化と互換性確保が重要となってくる。そこでわれわれは，マイクロアレイや定量PCRから細胞1個あたりのmRNAコピー数を得るPercellome法を開発した。

Ⅱ．Percellome法：細胞1個あたりのmRNAコピー数として発現値を得る方法

原理は，サンプルの細胞数に対して標準化するという単純なものである。具体的には，サンプル破砕液のDNA量から細胞数を求め，外部標準mRNA（スパイクRNA）を細胞1個あたり決まった分子数だけその破砕液に添加し，そしてRNA抽出，測定に移る。スパイクRNAの発現値が細胞1個あたり何コピーに由来するかが既知であることを利用し，サンプル中のすべてのRNA測定値を，細胞1個あたりのコピー数に換算する。スパイクRNAは，5種類の枯草菌遺伝子のmRNAを濃度を公比3で振って混合したカクテル（dose-graded spike cocktail：GSC）として用意した。これにより，各サンプルに細胞1個あたりのコピー数が既知の標準点5点が導入され，これらをつなげば標準直線が得られる。この直線を用いて実際に測定したmRNAの発現値（マイクロアレイの場合，数万種）をコピー数に換算する[1-3]（**図❶**参照）。

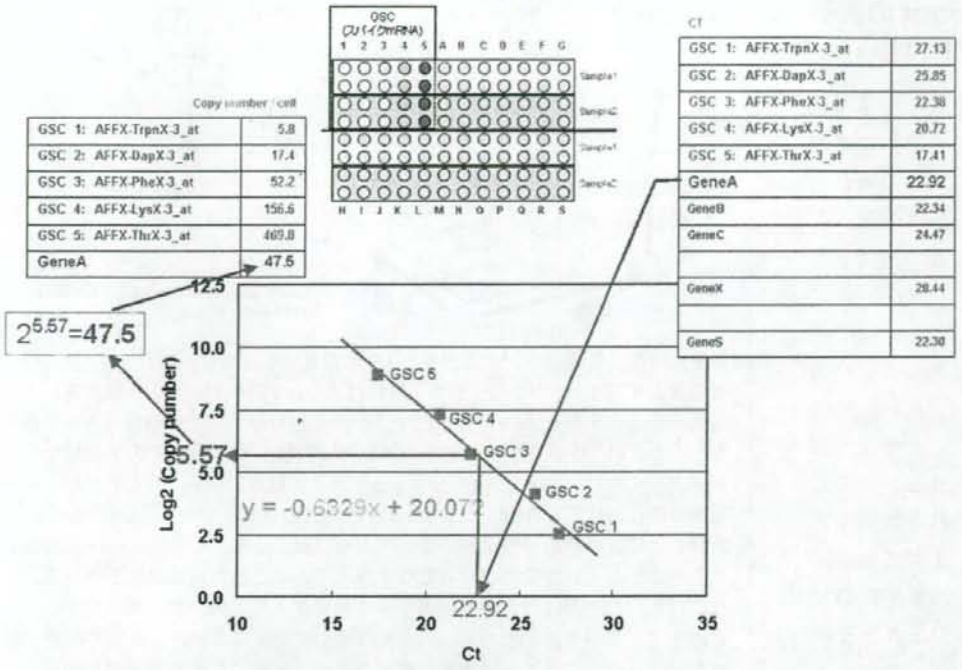
このような測定に用いるマイクロアレイには十分な定量性・直線性が備わっている必要がある。その検証にはLBM標準サンプル〔肝（L）と脳（B）を100：0，75：25，50：50，25：75および0：100に混合した5サンプルからなるセット〕を用いた。現在までに，Affymetrix社GeneChip®（マウス，ラット，ヒト，ゼノパス），Agilent社の単色44kアレイ（マウス，ラット，ヒト），およびCodelink社（マウス，ラット，ヒト）マイクロアレイについて，定量性・直線性を確認し，Percellome測定が可能な状態となっている。

GSCの添加を正確に行うために高精度を要求されるDNA定量については，手作業プロトコルおよび自動ロボット（PerkinElmer社 JANUS）のプロトコルを準備した。GSCのストックも含め共同研究ベースで供給可能である（連絡先：kanno@nihs.go.jp）。

Ⅲ．Percellome法の定量的リアルタイムPCR（Q-PCR）を含む他のプラットホームへの適用

Percellome法は，GSCの受け入れ条件を整えることにより，様々なプラットホームに適用可能である。その1つとして最も定量性が高いとされるリアルタイムPCR（ABI PRISM 7900 HT・96ウェルプレート）への適用例を示す（**図❶**）。現行のQ-

図❶　定量PCRによるPercellome法の概要



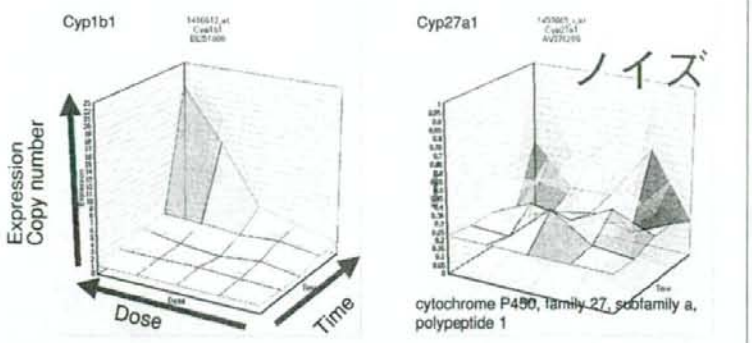
96ウェルプレートにサンプルをduplicateで分注し，図中1～5のウェルにおいてGSCの5種類のスパイクmRNAを増幅する。A～Sにおいてコピー数未知の測定対象を増幅する。プライマーの増幅効率は可能なかぎり揃えておく。CT値をX軸に，コピー数の対数をY軸に，GSCの5点をプロットすると，各サンプルごとに検量線が得られる。これを用いてGeneAのCT値が22.92であったら，コピー数は47.5/細胞と求まる。GeneChip®においても基本的に同様の方法で約45,000プローブセットのPercellome値を得ている。（グラビア頁参照）

PCR絶対定量法では，遺伝子ごとに検量線が必要であり，多数のサンプルについて多数の遺伝子の発現を定量するには不向きである。Percellome Q-PCRでは，マイクロアレイと同様の原理を用いる。すなわち，サンプル破砕液にその細胞数に比例する量のGSCを添加する。それらのCt値は既知コピー数に対応することを利用してPCRプレートごとの検量線を得て，それを参照することで測定したい遺伝子のCt値を細胞1個あたりのmRNAコピー数に換算する。これにより，GAPDHやActinなどのハウスキーピング遺伝子が変動してしまう際の問題，少数の遺伝子を検討する際にGlobal normalization法が適応しにくい問題などが解決される。同一サンプルをこのPercellome Q-PCRとAffymetrix GeneChip®とで測定し比較したところ，測定したプローブセットの9割程度に整合性が確認され，GeneChip®とPercellome Q-PCRとの間でのコピー数の換算式が得られている。この他に，Agilent社製の単色マイクロアレイとCodeLinkアレイにGSCを測定可能なカスタムアレイを用意し終え，LBMサンプルのデータなどをもとに，これらとの間の換算式も得つつある。なお，Percellome法は，Affymetrix社の新しいエクソンアレイの定量性・直線性の検討にも適応可能である。Affymetrix社のHuman Exon 1.0 ST Arrayと従来型の発現アレイHuman Genome U133 plus 2について，性質の異なるヒト癌細胞株2株から調製したLBM様標準サンプル（100：0，75：25，50：50，25：75および0：100混合5サンプル）による比較を行っている。

Ⅳ．Percellome Projectの実験プロトコル

体内に侵入してきた化学物質などを第一に感知するのは，多くの場合タンパク質であり，それからの次の影響が遺伝子発現に波及した場合にmRNAの変動として現れ，それが次のタンパク質を誘導し，次のmRNA変動を招く，と模式図的には考えられる。このような初期応答を観測する目的から，まず成獣の肝を対象とした単回経口投与実験プロジェクトを開始した。mRNA合成のスピードと動物実験の手技上の現実的限界を考慮し，単回強制経口投与の2，4，8および24時間後にサンプリングを行うプロトコルを設定した。また，用量依存性を考慮し，投与量を溶媒対象(0)，×1，×$\sqrt{10}$，および×10とした4群を設定した。すなわち，1つの化合物について，4（時点）×4（用量）の16群，各群3匹，合計48匹とした。マウスはノックアウトマウスを併用する前提からC57BL/6（SLC）を，週齢は肝の酵素が安定する12週齢，性別は性周期のない雄とした。また後述するが，肝，腎，肺，心，脳など，ほとんどの臓器に明瞭な日内変動が認められるため，マウスを10週齢時点で搬入し，明暗12時間サイクルを厳守した環境で2週間馴化した後，明サイクル2時間目に投与，以後，各予定時刻の前後20～30分以内にサンプリングを完了することとした。1匹につき，麻酔（エーテル），脱血（腋窩動脈より）の後の2～3分以内としている（実施要領を執筆中）。マイクロアレイはAffymetrix社GeneChip® Mouse430 2.0（初期は430A）を用いた。サンプルはプールせず，個体ごとに測定した。

図❷　Percellome Projectにおけるデータの三次元表示（Millefeuille surface data）

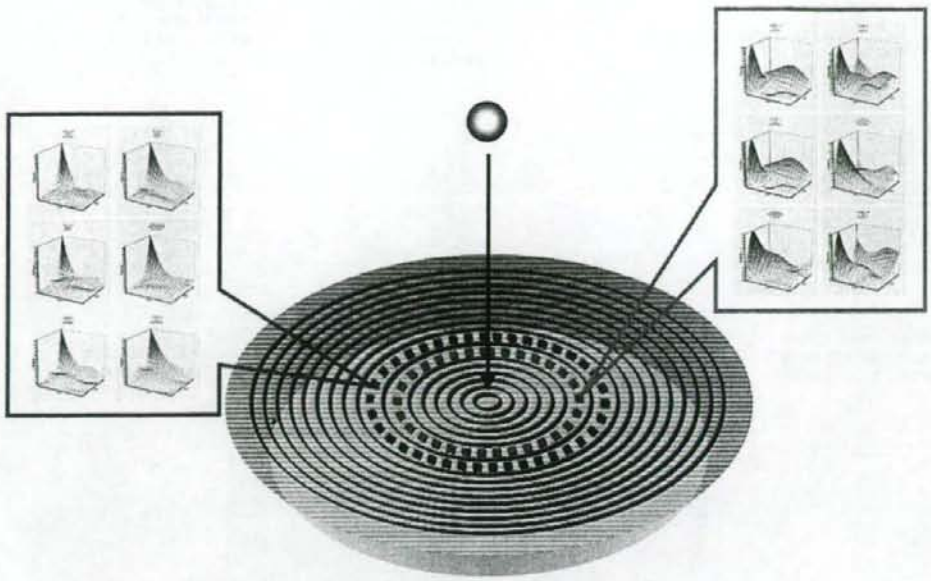


用量，時間，発現値を三次元表示し，1つの遺伝子（プローブセット）につき1枚の曲面を描く。1つの格子点が1群に対応し，3匹のデータから計算される平均値からなる平均surface（濃く色づけ）と±1SDの2つのsurface（薄く半透明表示）の合計3層からなる。ここに示すのは図❹と同一の2, 3, 7, 8-四塩化ジベンゾダイオキシン（0, 1, 3, 10, 30 μg/kg）を単回投与した後，2, 4, 8および24時間後の肝における*CYP1b1*（左）および*CYP27a1*（右）である。*CYP1b1*は，24時間目に用量依存的に誘導がかかっており，SD surfaceも平均surfaceに近接していることから，生物学的に有意なデータであることが用意に判断される。これに対し，*CYP27a1*は，時間経過からも用量からも生物学的に考えにくい上下を繰り返しており，またSD surfaceも分離していることから，生物学的蓋然性のない，おそらくノイズであると直感される。実際に，その発現絶対値はこの場合の公称下限感度である1コピー/細胞を下回っている。　（グラビア頁参照）

Ⅴ．Percellome　Projectデータの構造（Millefeuille surface data）と解析

化学物質単回投与による遺伝子発現変動は，時間および用量に依存するという考えから，x軸＝時間，y軸＝用量，z軸＝mRNAコピー数とする三次元グラフ上に曲面（surface）として投与が誘発する遺伝子発現変動を可視化した（図❷）。1つの遺伝子（GeneChip®ではプローブセット）につき，3匹の平均surfaceと±1標準偏差（SD）surfaceを表示することで，視覚的に反応を捉えると同時に，ノイズあるいはartifactであるか否かの感触を容易に得られるようにした。Percellome Projectではこの三次元surfaceのパターンを基礎に遺伝子発現カスケードの描出に関わる方法論の開発を進めている（図❸）。現在，1つの実験から得られるGeneChip® 48枚のデータを一括処理する能力をもったPercellome自動換算・データ品質管理（QC）ソフトウエア，

図❸ Percellome Projectデータ解析の基本概念



水面を遺伝子カスケード，そこへ落とし込む石を投与化学物質と考え，石による波紋が投与実験から得られる遺伝子発現の初期応答であると考える。そして，三次元surfaceの類似する遺伝子をクラスターとして網羅的に抽出すると，1つのクラスターは同心円状の波の1つに当たることになる。このような波のクラスターを異なる「石」（別の化合物）のクラスターと比較検討することで，生命体が利用するカスケードの共通部分を特定してゆく。これに遺伝子ノックアウトマウスなどからの客観情報を組み合わせ，カスケードの描出範囲を広げてゆく。
（グラビア頁参照）

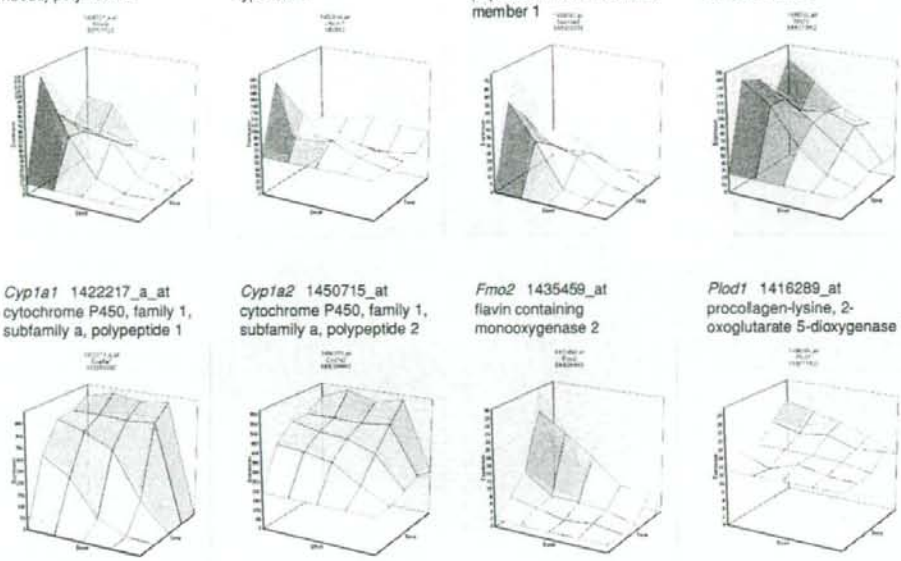
三次元多層データ（Millefeuille surface data, MFデータあるいはSurfaceデータと呼んでいる）のパターン類似性を元にした候補遺伝子検索ソフトウエア群を中心とした解析システム（MFソフトウエアシリーズ，開発：相﨑健一），およびその概念を発展させた教師なしクラスタリング[4]を独自に実用化し，現在も改良・開発を継続中である。

データの三次元可視化は，コンピュータが選び出した遺伝子クラスターの中身を確認する際，ことにmRNAの合成分解のスピードなどの常識から考えて生物学的にありえないパターン（用量軸の方向にも時間軸の方向にもジグザグな変化など）をノイズとして排除する際に威力を発揮している。さらに，いくつかの化合物の全遺伝子リストを比較し，化合物に共通して同期して発現する遺伝子群を自動抽出するシステムも開発ずみである。ここで得られた遺伝子群はシグナルカスケードの構成単位である可能性がある〔5TB規模のデータベース部分および，大量計算アルゴリズム実装，クラスタリングアルゴリズムなどはNTTコムウェアおよび日本NCR/Teradata（松本伸哉氏）との共同開発による〕。

Ⅵ. Percellome Projectの概要とデータ例

現在までに，単回強制経口投与による肝の初期反応データを，既知情報のある90以上の化学物質（医薬品，一般化学物質，食品関連物質を含む）について採取し終えた。例として，アリル炭化水素受容体（AhR，ダイオキシン受容体）に結合することが知られる2, 3, 7, 8-四塩化ジベンゾダイオキシン（ダイオキシン）の結果の一部を示す（図❹）。比較的少数のAhR直下の遺伝子が2時間目に誘導され，4, 8, 24と時間が経過するにつれ数が増して

図❹　ダイオキシンによる肝遺伝子発現応答

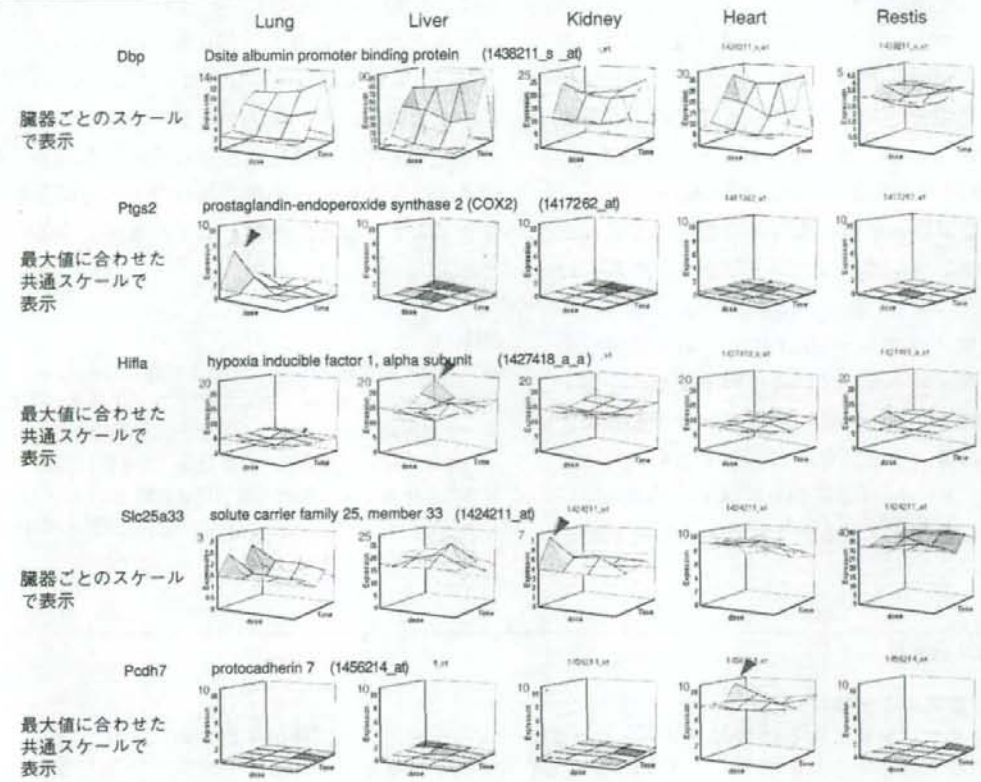


ダイオキシンの単回経口投与後の応答例を示す。2，3，7，8-四塩化ジベンゾダイオキシンの0，1，3，10，30 μg/kgを単回投与した後，2，4，8および24時間後の肝をサンプリングしたものである。*Tiparp*，*Zfp36l1*，*Serpine1*，および*Nfe2l2*（*Nrf2*）は2時間目に用量相関性をもって一過性に発現した。低用量では発現のピークが4～8時間にずれ込む傾向がみられる。*Tiparp*と*Nfe2l2*では高用量で24時間目に再誘導がみられる。*CYP1a1*および*CYP1a2*は2時間目より24時間目に向けて発現上昇が持続した。*Fmo2*は8時間目に一過性に誘導がかかり，*Plod1*は24時間目に向かって発現上昇を示した。2時間目に一過性に応答する初期応答遺伝子は10個程度，その後，時間とともに誘導される遺伝子数は急激に増加した。　（グラビア頁参照）

いった。一過性のパターンをとるもの，持続的に増加するもの，二峰性のパターンをとるものなどが観測された[5]。これに引き続くプロジェクトとして，食品・食品添加物関連，シックハウス症候群を考慮した低用量域での吸入毒性トキシコゲノミクス，1匹のマウスから多臓器を採取しそれらの連関を解析する多臓器トキシコゲノミクスを開始したほか，胎児発生，行動神経に関わるプロジェクトへの展開を図っている。

ここで，多臓器連関解析の一部を紹介する。四塩化炭素を0，0.7，2および7mg/kgの比較的低用量で強制経口投与し，その2，4，8および24時間後に，肝，腎，肺，心，精巣の5臓器を採取，Percellomeデータを得た（図❺）。まず，投与とは無関係に変動する概日遺伝子は精巣以外のすべての臓器において明瞭なパターンを示し，最高値が最低値の数倍を超えるものが多数みられた。この比較的低用量では上記のダイオキシンに比較すると，ごく限られた少数の遺伝子にのみ変動が認められた。2時間目に誘導のかかる遺伝子は肺に最も多く観測された。肺の初期応答は*heat shock protein*や*cycloxygenase 2*など，ストレス応答反応を示唆するものが主体であった。肺での変動遺伝子数は8時間目にピークを示した後，24時間目には減少傾向を示した。これに対し肝では2～4時間目にはごく少数の代謝酵素が軽度誘導されるのみであったが，8～24時間目に多くの遺伝子発現が観測され，それらには*HIF1a*，*Thyoredoxin*，*Gultathione S-transferase*，*SAA1*などが含まれていた。また少数ながら，腎，心に誘発される遺伝子も確認された。四塩化炭素

図❺ 四塩化炭素による多臓器影響



肺，肝，腎，心および精巣の四塩化炭素による遺伝子発現影響を示す。*Dbp*は概日遺伝子であり，精巣を除くすべてに投与に依存しない概日変動を認めた。最低値と最高値には数倍の差が多くの臓器で認められた。比較的低用量の四塩化炭素では，比較的限られた少数の遺伝子に変動が認められた。その中で2時間目に最も多くの遺伝子発現がみられたのは肺で，ここに示す*Ptgs2*（COX2）のほか，*heat shock protein*などのいわゆるストレス応答を示唆する遺伝子が変動した。肝では，2, 4, 8時間にはほとんど遺伝子発現はみられず，24時間目にここに示す*HIF1a*のほか，*Thyoredoxin, Gultathione S-transferase, SAA1*など比較的多数誘導された。腎においては2時間目（*Slc25a33*）と24時間目に少数の発現があり，心では8時間目（*Pcdh7*）を中心にいくつかの遺伝子の発現を認めた。（グラビア頁参照）

の単回経口投与時の最初の標的は肺であり，いわゆる酸化的ストレスが惹起された可能性が指摘される。その背景には肺の好気的環境が考慮され，*HIF*が肝では誘発されたのに対して，肺で誘導されないこともこれを支持している可能性が考えられた。

まとめ

「マイクロアレイデータには再現性がない」と言われることがあるが，その原因としては，その定量性の良し悪しのほかに，実験計画の不十分さが挙げられる場合がある。動物には概日リズムがあり，肝では数千の遺伝子に，他のほとんどの臓器でも明らかな概日変動がみられる。サンプリングのタイミングを一定にしないと，この影響を大きく受けてしまいデータが再現しないことが経験される。また，実験前の馴化期間中の動物舎の照明スケジュールが厳密に管理されていないと，タイミングを守っていても実験間誤差が大きくなる。その他，食べさせる餌の組成，飼育時のストレス（断水など）にも反応する。これに似た影響は，*in vitro*サンプルにも経験される。例えば，培養細胞

の培地交換はもとより，倒立顕微鏡での観察の有無（培地の攪拌などの影響），インキュベータ内の配置，96穴プレートでの辺縁効果（最外周ウェルの培養条件が中心と異なる）などが，ときに明瞭に遺伝子発現データに反映される。このような点に配慮することで，遺伝子発現データが安定し，再現性が向上することを経験している。

ノーザンブロットのような半定量的な手法において実験サンプルにだけバンドがあり，対照にはないという結果も，Percellome法では10コピーに対して実験サンプルが20コピーである場合がある。最近の医学・生物学には多因子疾患・多因子形質発現制御の概念が導入され，「21世紀初頭までは，患者の遺伝子多型を調べずして治療を行っていた時代」として，「血液型を調べずに輸血していた時代」と並び称されるようになるかもしれない勢いである。このような多因子概念の世界においては，生体反応の既述は「有（100％）」「無（0％）」の組み合わせではなく，「70％」「50％」「90％」といった半端な数の組み合わせによることになる。その際の網羅的トランスクリプトームデータにとって，定量化と標準化は必須の要件と思える。その確保は，これから実現されるであろう網羅的プロテオミクスなどの基盤としても重要となると考える。

謝辞

本システムの開発とプロジェクトの遂行にあたっては，当毒性部の全メンバー，特に松田菜恵，辻昌貴，安東朋子，安部麻紀，森山紀子，森田紘一，近藤優子，古川佑介，青柳千百合，渡辺忍，相原妃佐子の各氏に深謝する。本研究は厚生労働科学研究費補助金H13-生活-012，H14-トキシコ指定-001，H15-化学-002，H17-化学-003H18-化学-一般-001などによる。

Technical Tips

マイクロアレイ実験

マイクロアレイによる網羅的トランスクリプトーム解析は，電子顕微鏡写真が世に現れた時の状況になぞらえることができる。すなわち，光学顕微鏡では見えない「もの」が新たに見えるようになったわけである。今までに見たことのないものを解析するには，十分な定量性のある方法でなるべく正確なデータを得て，汎用性のある標準化手法を用いて横断的なデータ解析を積み重ねる必要がある。その中で，はっきり見えるようになったものには，実験動物の概日リズムや培養細胞の培地攪拌による影響が含まれる。信頼できる良質のデータを得るには，実験条件設定に今までとは違った次元での細心の注意を払う必要がある所以である。また，動物にしろ細胞にしろ肉眼的あるいは光学顕微鏡的に確認される器質的変化の多くは，タンパク発現の後のものである。例えて言えば，全焼後の火事場の現場検証であり，出火原因の特定は経験的なものになる。ところが，トランスクリプトーム解析は，実際に出火したところを見るのが目的である。すなわち，タンパクのずっと前のmRNAの発現する段階を見るのである。現場検証の証拠所見を追認するだけの実験にしないように，マイクロアレイ実験のプロトコルはこの点を十分に考慮して設定されたい。

参考文献

1）Kanno J, Aisaki K, et al : BMC Genomics 7, 64, 2006.
2）菅野　純，相﨑健一，他：細胞工学23, 685-693, 2004.
3）菅野　純，北嶋　聡，他：細胞工学26, 71-77, 2007.
4）Matsumoto S, Aisaki K, et al : Genome Informatics 16, 183-194, 2005.
5）http://toxicomics.nihs.go.jp/db/

菅野　純
1981年　東京医科歯科大学医学部医学科卒業
1985年　同大学院医学研究科博士課程修了（病理学専攻，医学博士）
国立衛生試験所病理部リサーチレジデント（(財）がん研究振興財団）
1986年　東京医科歯科大学医学部病理学第二講座助手
1991年　米国国立衛生研究所NIH（NIEHS）実験病理部・客員研究員
1993年　東京医科歯科大学医学部感染免疫病理学講座助手（旧第二病理）
1995年　同講師
1997年　国立医薬品食品衛生研究所毒性部室長
2002年　同部長（現在に至る）

化学物質安全対策，食品安全，医薬品などに関連する毒性評価業務および受容体原生毒性（内分泌撹乱化学物質問題）などの分子毒性学的研究，発癌，トキシコゲノミクス（Percellome）プロジェクト，ナノマテリアル，食品関連物質などの安全性研究を推進している。

3

Principles of Data Mining in Toxicogenomics

Yoko Hirabayashi and Tohru Inoue

3.1 Introduction

When animals are exposed to ionizing radiation, the radioactivity induces a probabilistic quantum effect based on the uncertainty principle followed by sequential early events in physicochemical processes, which is in the realm of subatomic particle physics (Czapski and Peled, 1973, 1975; Hunt, 1976). Because the initial radiation hit may be probabilistic and random with respect to radiation beams, radiation-induced damage may be random and stochastic from one molecule to another, from one cell to another, from one tissue to another and, furthermore, from one individual to another. Consequently, data should not be statistically focused but must reveal a limited probabilistic divergence, including divergence from one cluster to another. Recent developments of DNA chips, photolithographic microchip analysis and integrated microarray methods (Brown and Botstein, 1999) have enabled the analysis of the global expressions of more than 34 000 genes as ultimate biological responses (Lovett, 2000; Hamadeh *et al.*, 2001; Storck *et al.*, 2002). The method is effective in not only providing deterministic genetic information, but also nondeterministic epigenetic information. Nondeterministic epigenetic information, together with computational toxicology, can be used to elucidate the mechanism underlying the above-mentioned background and to identify genes including responsible gene ontologies (GOs).

It is also of interest to compare gene expression profiles between radiation-induced myelogenous leukemias and spontaneous myelogenous leukemias as a sample analytical model for computational toxicology; principal component analysis (PCA) primarily differentiates the expression profiles between radiation-induced and spontaneous myelogenous leukemias. Although radiation-induced myelogenous leukemias are expected

Toxicogenomics: A Powerful Tool for Toxicity Assessment Edited by S. C. Sahu

to show a convergent gene expression pattern because of random but limited damage induced by radiation exposure linked to, for example, radiation-fragile sites (Hastie and Allshire, 1989; Yunis *et al.*, 1987), spontaneous myelogenous leukemias are associated with more highly stochastic diversities in gene expression patterns. The ordered list generated by PCA provides the contributing genes that differentiate the expressions associated with radiation-induced myelogenous leukemias from those associated with spontaneous myelogenous leukemias. These genes are assumed to provide useful biomarkers for differentiating these myelogenous leukemias that have different characteristics, although additional computational treatment is required to elucidate the radiation-specific gene expression profiles (see Section 3.8).

3.2 From gene Expression to Toxicological Application

Genetic information is carried by chromosomes, the major composition of which is nucleic acid, i.e. DNA, consisting of four different bases, namely guanine, adenine, thymine, and cytosine (Watson and Crick, 1953). Such genetic information is transcribed into messenger RNA (Brenner *et al.*, 1961), triplets of which, 'codons,' are translated into the 20 different amino acids (Crick *et al.*, 1961). On the basis of the translated information on the amino acid sequence, the protein, a source of life, is synthesized (Crick, 1958). A full set of chromosomes is called the genome, and its composition (DNA sequence) is replicated semiconservatively as genetic information (a master plan), which is transferred to descendant DNA in the case of germ cells or contributes to the expression of various biological activities, including cellular proliferation, apoptosis or metabolic activities, in the case of somatic cells (Singer and Berg, 1991; Bloomfield *et al.*, 2000; Lewin, 2004; Lodish *et al.*, 2004; Watson *et al.*, 2007). From the above-described background, the use of 'toxicogenomics' is aimed at collecting such information, which is assumed to reflect all the information about life, including on the one hand intrinsic biological activities and on the other hand xenobiotic responses (Inoue, 2003).

The sequencing of the genomes of *C. elegans* (*C. elegans* Sequencing Consortium, 1998), *Drosophila* (Adams *et al.*, 2000), mouse (Waterston *et al.*, 2002) and human (Lander *et al.*, 2001) was completed around the year 2000. This development makes it possible to establish a method of detecting global gene expression as macroscientific information, including the determination of xenobiotic hazardous effects (Inoue, 2003). Such information can be utilized for many different purposes, and its toxicological application, particularly for predicting toxicological gene expression, is called toxicogenomics (Borlak, 2005).

3.3 Toxicogenomics

Toxicogenomics can be assumed as an 'expression gene microscope' because the method focuses on patterns of global gene expression as a whole, rather than on individual gene expressions (Figure 3.1). Original data obtained and examined by gene chip or microarray technologies provide mathematical values based on expression intensities. These values have essentially no biological meaning; rather, they reveal only a pattern. The patterned expression intensities obtained and those of the corresponding reference findings are

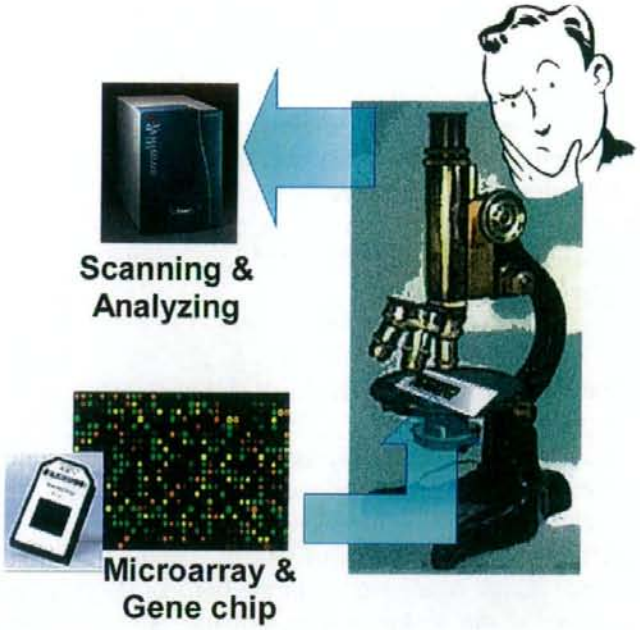


Figure 3.1 *Toxicogenomics methodology can be assumed as an 'expression gene microscope' because the method focuses on patterns of global gene expression as a whole rather than on individual gene expressions. See text. Two different methods are utilized: one using DNA microarray invented by Brown and Botstein (1999) and one using photolithographic gene chips invented by Fodor et al. (1993). The patterns of gene expression of targeted groups and the control are analyzed computationally using various pieces of external information (supervised analysis) or solely with internal data (unsupervised analysis).*

mathematically analyzed, which is considered 'transcriptomics analysis.' One can find similarities in the histopathological examination; there are different findings between untreated and treated histological specimens, and the gaps in the findings can be identified and considered as a histological alteration induced by the treatment. Toxicogenomics also requires an established database as well as an analytical computational methodology; this is similar to histopathology, which was established using the histopathological disease entities established during a hundred years of pathological history (Henke and Lubarsch, 1924–1952).

3.4 Mining of Information Provided by Toxicogenomics

Categorically, two different types of information are provided by gene expression profiles: 'genetic information' and 'epigenetic information.' The former is based on the genetically defined deterministic information. However, the latter is based on the developmental tissue-specific gene expressions, for example, or based on the consequences of xenobiotic responses with modified gene expressions by methylation, acetylation, or phosphorylation

of the genome; thus, the alteration would be probabilistic and plastic. The former, for example, includes single-individual gene-specific nucleotide polymorphism (SNP), which may be within a category of a new type of genomic bioassay. On the other hand, the latter seems to be analogous to conventional microscopy, which is considered to provide analytical information related to xenobiotic responses when one compares altered gene expressions after xenobiotic responses with an appropriate control database. Either type of information on gene expressions obtained from the former and the latter materials is useful in the field of toxicology; both are managed similarly from the technological viewpoint; thus, the latter is mainly discussed in the text.

Gene expression profiles from mice after whole-body irradiation were used as experimental models. In this experiment, mice were exposed to a single dose of radiation at 0.6, 1, and 3 Gy and the gene expression profile in the bone marrow of the exposed mice compared with that of the nonirradiated control. From the results, one can find gene expression profiles with sequential changes in intensity with increasing radiation dose (*dose-dependent profiling*). On the other hand, one can also find a dose-independent expression pattern among the groups (*dose-specific profiling*). The former profiling is considered to be a useful tool for comparing dose–response relationships from the outcome of various testing protocols. The latter, on the other hand, is not only dose specific, but, interestingly, also possibly valid for a wide range of interspecies extrapolation; thus, the gene expression profile of the latter can be another potentially useful biomarker (data not shown).

3.5 Points to Consider in Data Mining Using Gene Chip and Microarray Technologies

Gene chip and microarray technologies provide a large amount of genetic information distributed in the over 34 000-dimensional Euclidian universe, because such data are based on the independently regulated expression of 34 000 genes (GeneChip® Mouse Genome 430 2.0 Array; Affymetrix, Santa Clara, CA). Various computational aspects of trials have been made available to reduce such a large number of dimensions (Zhang and Shmulevich, 2006; Albanese *et al.*, 2007).

One piece of software provides a single-dimensional gene matrix based on each gene expression intensity, which is clusterized by Euclidean distance on the basis of gene expression intensity, and links them to each other according to their vector values (GeneSpring GX 7.3.1; Agilent Technologies Inc., Santa Clara, CA). A sample dendrogram is shown in Figure 3.2. These expression data can be clusterized and linked on the basis of two factors, namely gene expression intensity and groups of conditional trees. Two-dimensional and higher multiphasic dendrographic analyses can be carried out. These processes are carried out unsupervised, namely statistically autogenerated.

In this dendrographic analytical method, the arrangement of each gene is linked solely on the basis of expression intensity; thus, the relationships of genes are solely based on a single-dimensional diagram. Specifically, in this dendrographic analytical method, the statistical structure allows deterministic alterations, such as mutations induced directly by a genotoxic carcinogenic compound, the mechanism of which needs to be elucidated more clearly. Unpublished data, as an example, provided by Professor T. Iguchi, Okazaki

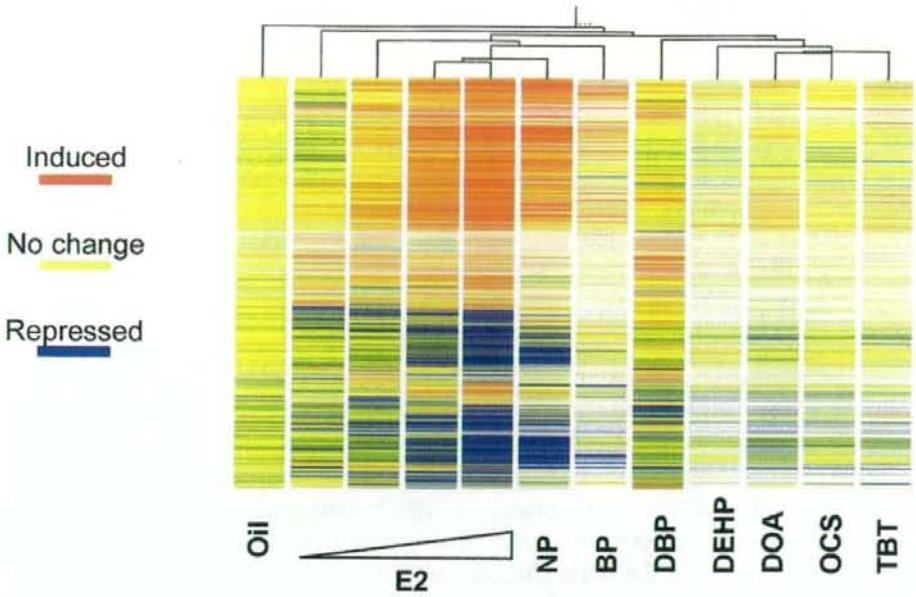


Figure 3.2 *Sample data for dendrographic analysis. Four columns, next to oil on the left, are from mice treated with a graded increase in dose of 17-beta-estradiol (E2). NP: nonylphenol; BP: benzophenone; DBP: dibutylphthalate; DEHP: diethylhexyl phthalate; DOA: dioctyl adipate; OCS: octachlorostyrene; TBT: tributyltin. (Unpublished data provided by Professor T. Iguchi, Okazaki National Biology Research Institute.)*

National Biology Research Institute, show a characteristic analytical function of the dendrographic analysis introduced above (Figure 3.2). In this figure, the gene expression profiles associated with estradiols at increasing reference doses and those associated with several endocrine-disrupting chemicals are shown. From the comparative gene expressions, the expression profiles associated with the reference estradiols, i.e. nonylphenol (NP) and benzophenone (BP), are clustered on the left half of the figure. On the other hand, the unsupervised autogenerated dendrogram does not show any representative profiles related to the role of endocrine-disrupting chemicals in the expression profiles associated with other known endocrine-disrupting chemicals, such as dibutylphthalate (DBP), diethylhexyl phthalate (DEHP) and dioctyl adipate (DOA). Specifically, dendrographic analysis is solely based on the expression intensity of each gene, and these autogenerated vectors clustered by approximation along the single-dimensional analysis cannot be compared with other dimensional data. Thus, multiphasic data profiling and multidimensional data analysis are based on different functions. That is, when one plots the expression data of each gene along a fixed dimension, it may be difficult to compare these data with the gene expression data along a different multidimensional axis having a pleiotropic interrelationship.

Furthermore, not all gene functions are important when genes are strongly expressed. Similarly, not all gene functions are of less importance when genes are weakly expressed, although the functions of weakly expressed genes are sometimes difficult to analyze. It is more difficult to analyze the interrelationship of genes on the basis of transcriptional activation and to elucidate the consequent gene expression pathway using this methodology.

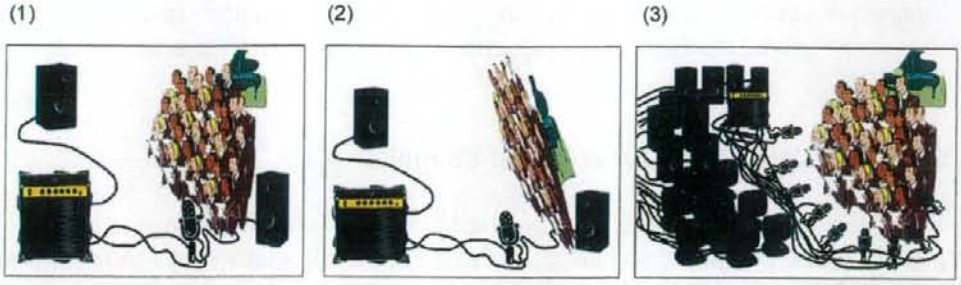


Figure 3.3 *Information from gene chip and/or microarray analyses is multidimensional. A microarray consisted of 34 000 genes behaving in the 30 000-dimensional Euclidian universe; thus, proper computational analysis requires appropriate multidimensional analytical power. See text.*

Figure 3.3a shows a cartoon where 30 000 persons are expressing (dispatching) their individual information. Depending on how a receiver receives the information, such information might be distorted because of the low-dimensional accepting system; the information might be distorted solely by expression intensities, as shown in Figure 3.3b for example. The possible reason why the intended information is not obtained from the microarray data is not always because of the poorly normalized raw data or the lack of significant data. One may have to realize that the reason may be the insufficient computational incorporation of multidimensional aspects of microarray data. Consequently, it may be appropriate when a possible multidimensional analysis of microarray data can be carried out to recover the data multidimensionally, as shown in Figure 3.3c.

3.6 Reproducibility of Microarray Data and Comparison among the Data Obtained from Different Platforms

Reproducibility, homogeneity and stability during data sampling are critical so that global gene expressions can be compared, because RNAs sampled from tissues are labile in general. Thus, the proper robotic administration of test compounds and the standardization of test materials are of special importance, the technical advancement of which has contributed markedly to the present microarray technologies (Schadt *et al.*, 2001; Churchill, 2002; Kroll and Wolfl, 2002; Astrand, 2003; Stoyanova *et al.*, 2004). Furthermore, spotting and spike gene incorporations also contributed to the development of a hardware system that supports qualitative data evaluation and semiquantitative gene expression profiles (Hill *et al.*, 2001). Consequently, newly introduced journals in the field of computational sciences publish a number of informational studies. Data from different platforms are also actively compared and reported because of minimum requirements in the field of microarray study (Brazma *et al.*, 2001). Actual sample evaluation data are not presented in this paper; however, the similarities of those obtained data with those obtained from different platforms are often confirmed by PCA. This is not only because of equivalences at each technical level of data processing, but also because it is more likely that those gene expression data are sequentially linked together. Current gene expression data obtained by

microarray and gene chip technologies are largely comparable, unless specific spike genes or other additional modifiers are incorporated in the system (Petersen *et al.*, 2005).

3.7 Pathological and Toxicological Endpoints

In this section, let us consider the general data mining of observed outcomes obtained by gene chip and microarray technologies. From Figure 3.4, which shows different routes from the bottom to the mountain top, one can assume that the top where the routes merge is a pathological endpoint where common genes are mostly expressed, whereas the other enclosed area at the foot of the mountain can be considered a toxicological endpoint, which is assumed to be based on the different routes to the summit represented by stochastic and probabilistic gene expression clusters. The former is considered to include a group of specific and deterministic gene expression profiles, i.e. an 'essential leukemogenic gene' profile, which could be used as possible biomarker genes for diagnosis, whereas the latter clusters represent various probabilistic uncertainties whose profiles are considered to be different from one cluster (route) to another, i.e. 'stochastically necessary genes.' In this regard, toxicological endpoints do not seem to provide definitive information of gene expression; however, toxicological endpoints represented by 'stochastically necessary genes'

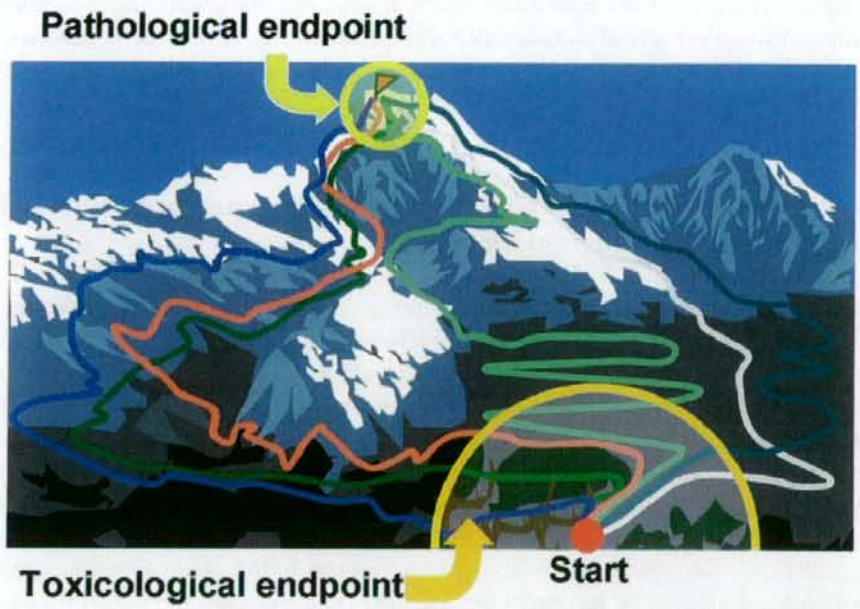


Figure 3.4 *'Pathological endpoint' and 'toxicological endpoint'. Each route from the start to the summit represents an individual probabilistic variety of different gene expression profiles. Depending on the characteristics of toxicological impacts, responders may show different toxicological endpoints in different clusters, even if one uses an identical and homogeneous experimental protocol with a highly purified inbred strain (a probabilistic quantum effect based on the uncertainty principle; see text).*

can provide probabilistic but cluster-specific predictability among various independent clusters. The latter clusters may not be statistically determined unless hundreds of a relatively large number of quantum cases are examined. Both an 'essential leukemogenic gene' profile and a 'stochastically necessary gene' profile are required for the early prediction of radiation-induced myelogenous leukemogenesis.

3.7.1 Pathological Endpoints

The strain C3H/He mouse develops a similar myelogenous leukemia both spontaneously and upon irradiation. The average incidence of the latter is 35% after 3 Gy irradiation, whereas that of the former is 1% (Seki *et al.*, 1991; Yoshida *et al.*, 1997). The line configuration in Figure 3.5 shows six cases of spontaneous leukemias on the left and six cases of radiation-induced myelogenous leukemias on the right. In this figure, the line configuration

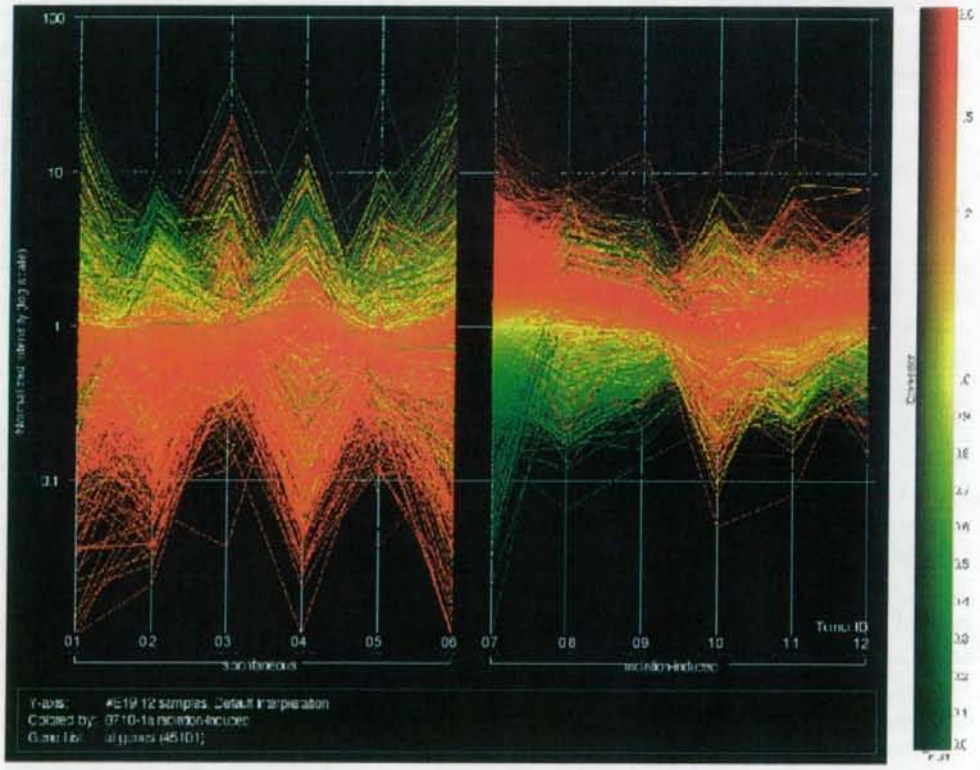


Figure 3.5 *Linear configurations of spontaneous and radiation-induced myelogenous leukemias. Six individual data on the left are from spontaneously developed myelogenous leukemias in C3H/He mice. The other six individual cases on the right are from radiation-induced myelogenous leukemias in C3H/He mice after 3 Gy X-ray exposure (Seki et al. 1991; Yoshida et al. 1997). Along the gene expression intensity from the highest (red) to the lowest (green) of group #07, the same gene in the other groups was connected and designated with the same color. Accordingly, overexpressed genes in the radiation-induced myelogenous leukemia groups are largely repressed in the spontaneous myelogenous leukemias. See text.*

associated with the spontaneous leukemias on the left shows very prominent and wide divergence in expression intensities from one individual mouse to another, and individual differences in gene ordering are also significant among each other. In contrast, each case of radiation-induced myelogenous leukemia on the right shows relatively homogeneous expression intensities compared with the former cases of spontaneous leukemias. When one compares the linear configurations of spontaneous and radiation-induced myelogenous leukemias, genes associated with radiation-induced myelogenous leukemias are not expressed in the spontaneous leukemias similarly but rather diversely. These findings are compatible with the observation by real time RT-PCR (Applied Biosystems 7900 Sequence Detection System, ABI, Foster, CA) shown in Figure 3.6, in which the former shows diverged *Sh2d1a* expressions from spontaneous myelogenous leukemias (*Sh2d1a* top) and the latter, relatively homogeneous ones from radiation-induced myelogenous leukemias (*Sh2d1a* bottom). Similarly, another gene, *Gngt1*, depicted from radiation-induced myelogenous leukemia shows relatively homogeneous expressions in most of radiation-induced myelogenous leukemias (*Gngt1* bottom), whereas the expressions of *Gngt1* were not detected in spontaneous myelogenous leukemias (*Gngt1* top). These expression profiles are further analyzed by PCA and the gene cluster associated with radiation-induced myelogenous leukemia can be discriminated from that associated with spontaneous leukemia (Figure 3.7). Representative genes that can be used to differentiate between the two types of myelogenous leukemia can be determined by PCA. The list of genes that differentiate both leukemias include *Met* (met proto-oncogene), *Fosl2* (fos-like antigen2), *Fancd2* (Fanconi anemia, complementation group D2), and *Fmr2* (fragile X mental retardation 2 homolog), among others. It is important that the expression intensities of these discriminant genes described above are not always high in all radiation-induced myelogenous leukemias, but sometimes low in a stochastic manner. Therefore, it is assumed that these genes cannot be the 'essential leukemogenic genes,' but 'stochastically necessary genes' for radiation-induced myelogenous leukemias. As described later, these genes, however, still seem to be less discriminant for radiation-induced myelogenous leukemias, although the function of each gene is linked to radiation injury. Why are radiation-induced leukemia-specific gene expression profiles not determined by the above computational analysis?

Regardless of the difference between spontaneous and radiation-induced myelogenous leukemias, it took nearly a lifetime for both myelogenous leukemias to develop fatally. Thus, the profile of each type of myelogenous leukemia may be associated with an age-related gene expression profile. Such an age-related gene expression profile may overlap to some extent with the gene expression profiles associated with spontaneous and radiation-induced myelogenous leukemias. The relationship of this factor with the above-mentioned differentiation will be considered later.

3.7.2 Toxicological Endpoints

Toxicological endpoints are different from pathological endpoints in terms of the nature of probabilistic clusters, as discussed above. One such frequent epigenetic modification, e.g. DNA methylation, constitutes a post-replicative modification, in which a methyl group is added covalently to a DNA residue. As it is known that cancer diagnosis before the cancer reaches the 'point of no return' is considered 'logically' difficult, toxicologists consider the possibility of identifying gene repertoires that are possibly associated with

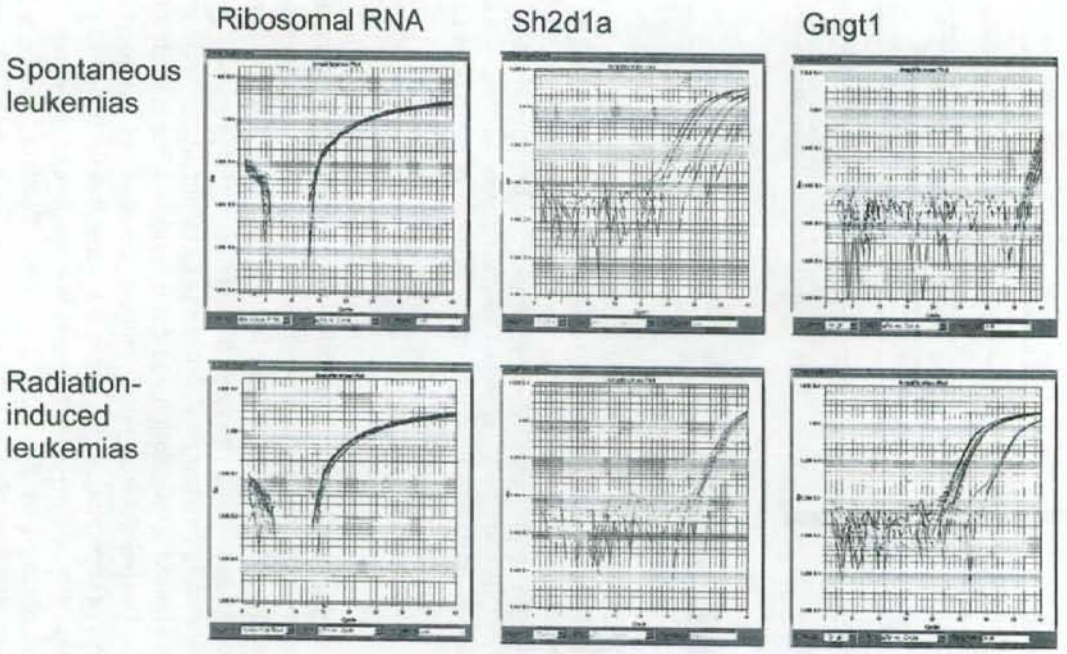


Figure 3.6 *Specific messenger RNA products of genes detected by fluorescent probes (TaqMan™ probes, ABI). Ordinate axes represent relative fluorescence (VIC™, 6-FAM™) and horizontal axes the number of cycles amplified. Ribosomal RNAs as control for RT-PCR on the left; a sample gene from spontaneous leukemia, Sh2d1a, in the middle; and a sample gene from radiation-induced leukemia Gngt, on the right are evaluated in both spontaneous and radiation-induced myelogenous leukemias (six cases each, upper and lower row respectively). Sh2d1a expressions in spontaneous leukemias (Sh2d1a top) are diverged in expression intensities, whereas those from radiation-induced leukemias are relatively homogeneous (Sh2d1a bottom). Similarly, another gene, Gngt1, shows relatively homogeneous expressions in most of radiation-induced leukemias (Gngt1 bottom), whereas the expressions were not detected in spontaneous leukemias (Gngt1 top). Quantitative real-time PCR triplicates are shown in each figure.*

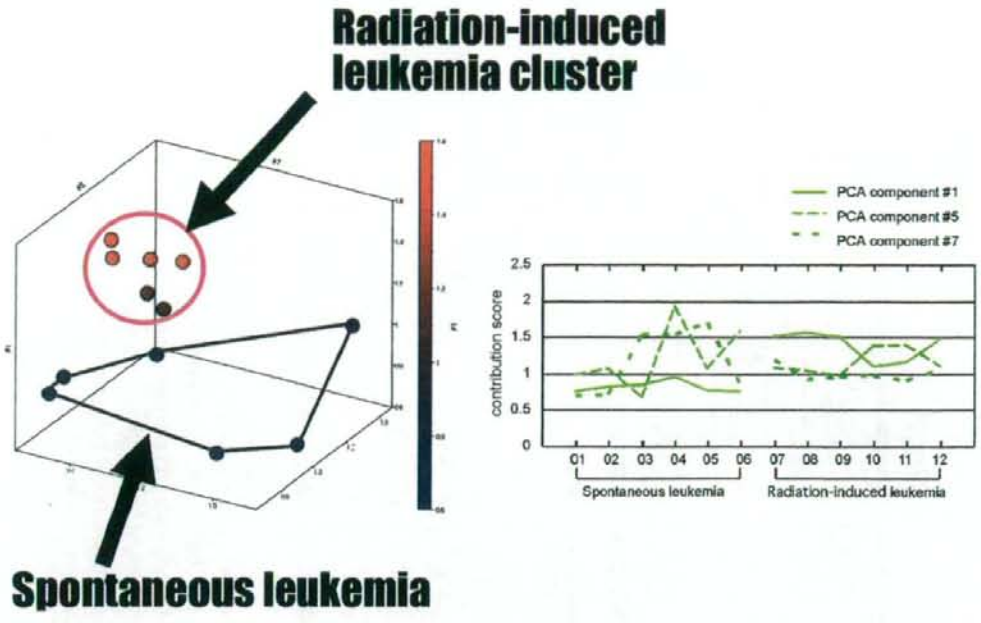


Figure 3.7 *PCA of six each of the spontaneous and radiation-induced myelogenous leukemias is shown in the three-dimensional contribution scores for components #1, #5 and #7, which discriminate the radiation-induced myelogenous leukemia cluster from the spontaneous myelogenous leukemia cluster. The line graph on the right shows actual contribution scores converted from each eigenvector value, which were used for the three-dimensional expression on the left. Note that the contribution scores of the spontaneous leukemias, except for component #1, are relatively divergent in comparison with those of radiation-induced leukemias.*

carcinogenesis at the early stage. There seems to be no definitive theoretical understanding of this issue. The most discouraging reason concerning this issue is that previous results obtained by the study of the International Life Science Institute (ILSI) consortium showed that the gene expression profiles associated with genophilic and nondirect genophilic pharmaceutical compounds were clearly differentiated, but both expression profiles from mice 4 h and 24 h after treatment showed up-regulation and down-regulation respectively (data not shown; Hu *et al.*, 2004). The possible diagnostic profiles, therefore, are considered to change with observation time or depending on the strain or treatment dose. Where can one find an appropriate discriminant axis? This is the question that needs to be answered.

The experimental results obtained by the ILSI consortium, however, can be interpreted differently. Those profiles may change with observation time, dose and strain; however, those profiles associated with genophilic compounds are always found to change in a direction opposite to that of the change of profiles associated with nondirect genophilic compounds; it is plausible that discriminant biomarkers for both genophilic and nondirect genophilic compounds would show this trend in both groups. In this regard, some sample trials carried out to determine these discriminant biomarker genes are discussed next.