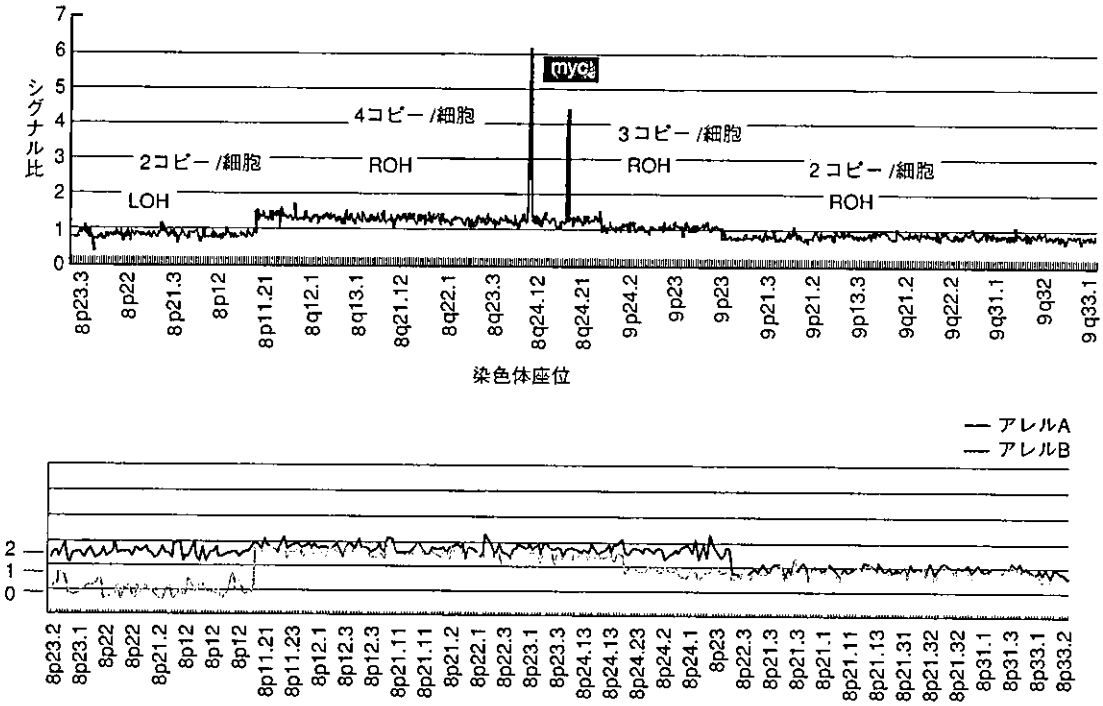

図1　DNAチップによるアレル別のゲノムコピー数解析
1番染色体から性染色体まで1万箇所に対するプローブが合成されたDNAチップを用いて癌細胞DNAのゲノムのコピー数を解析した．*myc*遺伝子を含む増幅部位の存在が検出されるほか，アレル別にコピー数を測れることからLOHの状態も判定できる（石川，油谷，未発表データ）

1 マイクロアレイによるゲノム多型解析

1）多型解析

SNP解析への応用によって1回のアレイ解析により5万箇所のSNPをタイピングできるようになり，ゲノム全体から10万SNP，およそ20 kbに1 SNPのタイピングが可能である．解析コストの面でもほかの測定法に対して競争力を有するようになりつつある．Affymetrix社製のタイピングアレイは公共データベースおよびPerlegen社が同定したSNP情報に基づいて設計されている．ゲノムワイドに有用な多型マーカーから関連解析により疾患関連遺伝子を探索している現状では多数のマーカーを高密度にタイピングする必要があるが，ハプロタイプが決定され将来的に解析すべきマーカー数が絞られた場合にはアレイによるタイピングもスクリーニングの有力手段であると思われる．家族集積例などの家系解析の場合は，同時にゲノムワイドに1万のマーカーをタイピングできれば，200 kbの解像度で解析できるので連鎖解析に十分有効であるし[1]，症例ごとにタイピングできるので実際の解析の上でも実用的である．

2）染色体異常解析

一方，癌細胞では癌化に先んじてゲノムの不安定化が生じていることが知られている．塩基レベルあるいは染色体レベルでのゲノム複製の確実性が失われることにより，変異が蓄積し癌化に至る．個人の全ゲノム配列を決定することも将来的には可能になると期待されるが，現時点ではコスト的に非現実的である．コピー数に関してはBAC（大腸菌人工染色体）クローンを配列したBACマイクロアレイを用いたアレイCGHやDNAチップを用いてかなり正確な遺伝子のコピー数の解析が可能となりつつあり[2)3)]，SNP測定と合わせることによりアレルごとにゲノムワイドな染色体変異の解明が期待される（図1）．さらに高密度オリゴヌクレオチドアレイの場合は高解像度であることが長所であり，小さな欠失領域や重複した配列の同定が

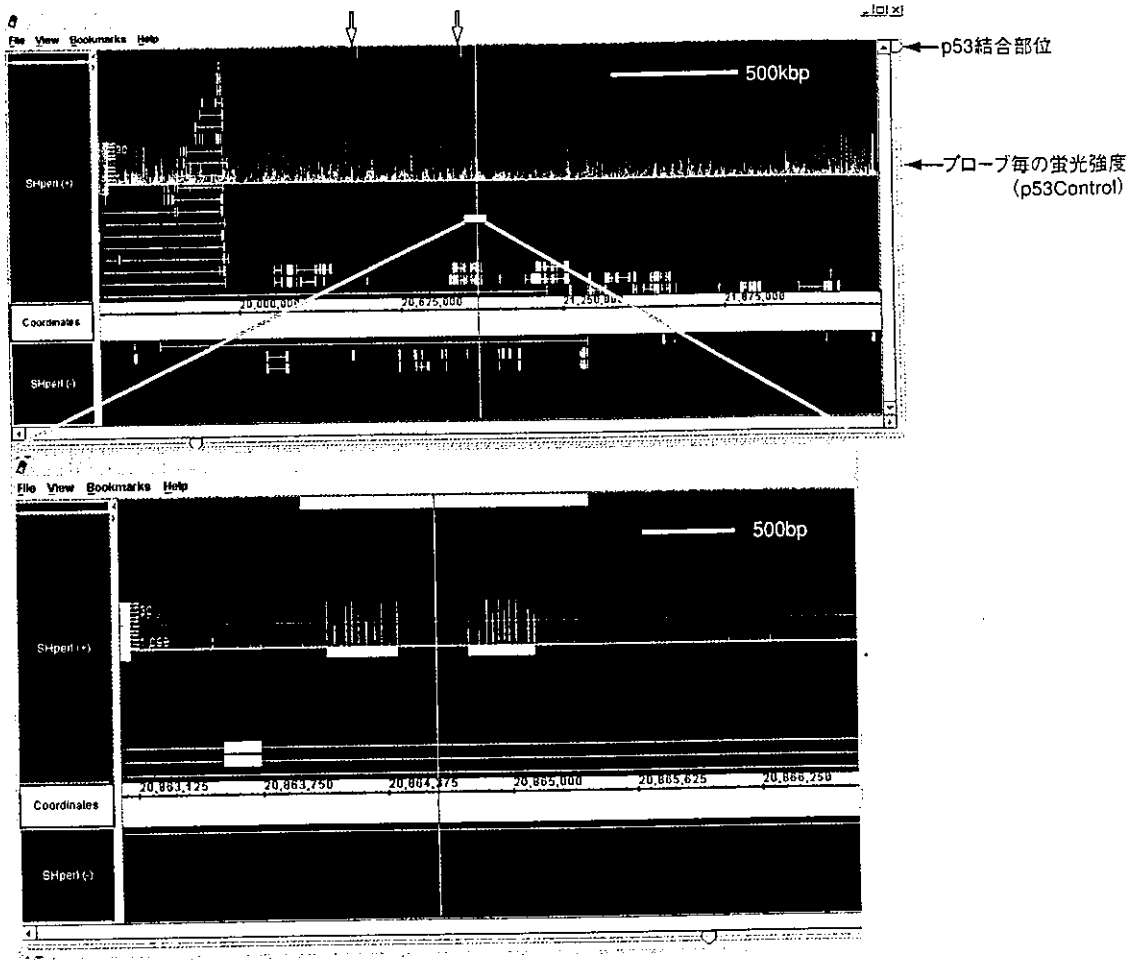

図2　ChIP_chip解析データのブラウザー表示（巻頭写真1－1参照）
p53による結合配列データをゲノムブラウザー内に表示したものである．連続する10前後のプローブのシグナル強度が有意に高いことから結合の有無を判定する

可能である．少量の検体から短時間でゲノム全体について増幅および欠失領域の判定を行うことができることから，現行のFISH法やBACマイクロアレイによる診断に替わる技術として期待される．現在Herceptin®などの分子標的医薬は標的分子の増幅を確認することが投与に際しての条件とされており，HER-2などの治療標的分子の増幅の判定への臨床応用が考えられる．

2 配列解析

1）シークエンシング

癌関連遺伝子の変異を解析するためにシークエンシングは有効な手段である．悪性黒色腫においてのBRAF (v-raf murine sarcoma viral oncogene homolog B1) 遺伝子の高頻度な変異の検出についても英国の癌ゲノムプロジェクトの成果であり[4]，網羅的な解析の有効性を物語る例である．すでにp53遺伝子変異[5]やp450遺伝子多型については配列決定用にマイクロアレイが用いられている．最近ではIressa®[R]の有効症例にはEGFR (epidermal growth factor receptor) 遺伝子に変異を生じている例が多いことが報告された[6][7]．発癌関連遺伝子について配列変異の有無を調べることが重要であることを物語る．

2）タイリングアレイ

アレイの高密度化により，ゲノム配列を網羅的に配

置することが現実的になりつつある．応用範囲はきわめて広く，転写領域の同定，スプライシングバリアントの検出[8]に用いることができるほか，転写因子結合部位（ChIP_chip解析*1）[9]や複製開始点[10]の解析にも用いることが試みられている．bleomycin刺激後にp53の結合配列の解析から，第21，22染色体上に48カ所の結合部位が同定された．われわれも，adriamycinなどのほかの薬剤による刺激でもそれらの結合部位がChIP後に濃縮されることを定量PCRおよびチップ解析で再現性よく確認することができた（図2，未発表データ）．p53のゲノムへの結合と近傍に位置する遺伝子の転写制御との関係について今後解析を進めていく予定である．

3 発現解析

マイクロアレイには医療における応用も期待されており，腫瘍組織や疾病に罹患した組織の遺伝子発現プロファイルすなわち遺伝子転写の状態の全貌を俯瞰することにより，分子レベルで新規な癌の分類[11]・診断が行われ（図3）[12]，治療への応答性および予後に関してより正確な予測に基づいた治療法の選択，すなわち「個別化医療」の実現が期待される．欧米では乳癌に代表されるように癌の治療効果や予後判定のために数千人レベルでの解析も計画されている[13]．薬理ゲノミクス（1章-3参照）の立場から化学療法剤の開発において癌細胞株パネルを用いたプロファイリングにより感受性予測への応用[14]が始められているほか，創薬のプロセスの初期に必要である新規創薬標的分子の同定あるいは検証に用いられている．進行性乳癌や前立腺癌では遺伝子増幅などによりHER-2遺伝子産物が過剰産生され，癌の増悪因子となっていることが知られており，抗HER-2ヒト化モノクローナル抗体（Herceptin®）はタキソールなどの化学療法剤との併用により生存期間，病勢進行までの期間を延長し，奏効率が向上したことが報告されている[15]．特にHER-2過剰発現レベルの著明に高い例では，約45％の生存期間の延長が認められている．

1）新たな疾患単位の発見

従来の分類では異質な症例が混在する集団において，発現プロファイル解析を用いることは分子レベルでの分類や遺伝子変異の推定に有効な場合があるほか，転移をきたす腫瘍の予測にも用いられている[16]．クラスター分析や主成分分析などの教師なし特徴抽出*2の手法が用いられ，データ全体を俯瞰するのに有用である．図3Aは小児白血病検体のマイクロアレイデータを主成分分析したものである．MLL転座*3を有する白血病は他の転座を有する白血病とは識別されることが認められるものの，予想に反して融合遺伝子による識別は明らかではなかった．そこでMLL転座白血病群のみのクラスター分析を行ったところ2群に分かれることが認められた（図3B）．さらに臨床情報との相関を検討したところ，無再発期間あるいは生存期間と相関することが認められ，予後不良群に対しては早期より強力な治療法を選択することが求められる．

2）アレイデータによる疾患の分類，識別遺伝子の抽出

有効群と無効群，罹患群と健常群というような2群間の比較においてはさまざまな「教師つき学習」のアルゴリズムが提唱され，それなりによい成績を収めており，その詳細は既報に詳しい[17][18]．一方，3群以上の比較あるいはあらかじめ分類すべきクラス数が不明の場合はまだ適当なアルゴリズムがない．また，教師つき特徴抽出を行う場合にはいずれかのクラスに強制的に分類してしまうことから「疑陽性」の問題があ

※1 ChIP_chip解析

クロマチン免疫沈降法（chromatin immunoprecipitation：ChIP）は個々のタンパクとゲノムDNAとの結合を検出する方法として用いられてきた．DNAチップ技術の高密度化により，全ゲノム配列を網羅的に合成，配置したタイリングアレイが使用可能となりつつある．DNA結合配列の同定やクロマチン構造の変換を解析するために用いられる．

※2 教師つき（なし）特徴抽出

マイクロアレイデータのような高次元のデータ空間をマイニングする機械学習の手法として，教師つき学習（supervised learning）と教師なし学習（unsupervised learning）に分けられる．前者はサンプルに分類名を与えることにより，識別するための変数の組み合せを知ろうとするのに対し，後者は全てのサンプルにわたり共通の制御を受ける遺伝子群の発見を目的とする．

※3 MLL転座

Mixed-Lineage Leukemia（MLL）遺伝子はTrithorax-related chromatin-modifying protooncogeneをコードしており，ポリコーム群（PcG）遺伝子と協調して高次クロマチン構造の制御に働き，Hox遺伝子群の活動を制御している．MLL遺伝子が位置する11q23は造血器腫瘍，とりわけ乳児白血病における染色体転座の主な切断点の1つである．

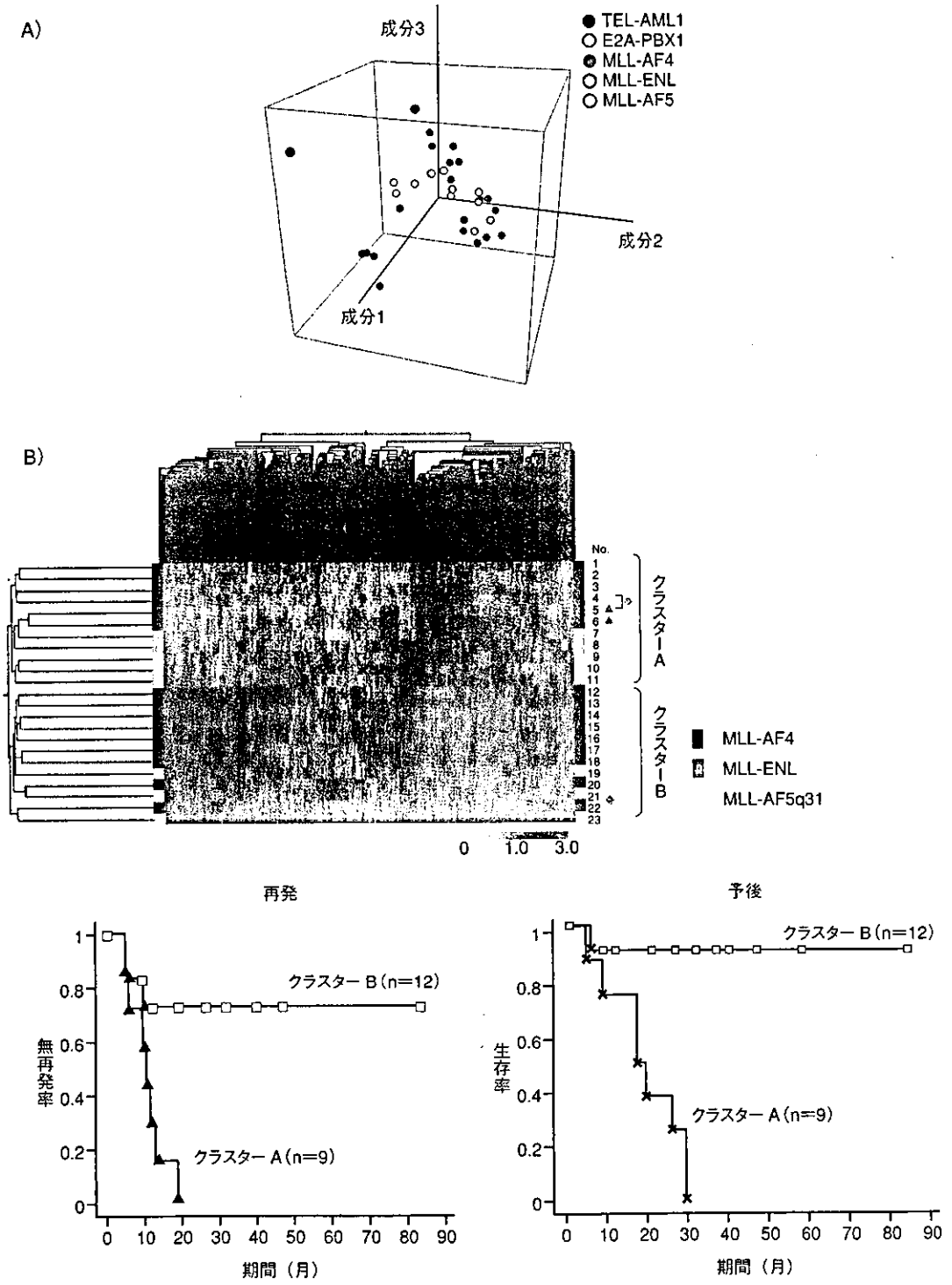

図3　アレイ解析による小児白血病における新たな病態の同定（巻頭写真1－2参照）
A）主成分分析による小児急性リンパ性白血病の分類，B）MLL転座を有する急性リンパ球性白血病患者より採取した骨髄について発現プロファイルをクラスタ解析したところ，2群に分かれ，2群間で無再発期間（P = 0.01）および予後（P = 0.0005）に有意な違いが認められた（文献21より引用改変）

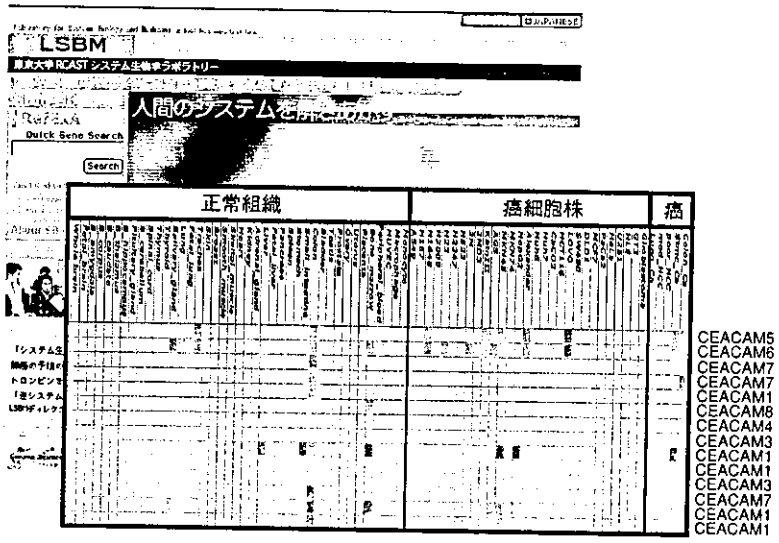

図4　遺伝子発現データベースSMB-DB
GeneChipあるいはCode Linkアレイにより測定したヒト正常組織，初代培養細胞，癌細胞株について2万遺伝子の発現データに，システム生物医学ラボラトリーのホームページ（http://www.lsbm.org/）からアクセスできる．遺伝子機能アノテーションについてはAffymetrix社により提供されるもの以外にもNCBIなどへのリンクがはられている

る．1施設において解析に十分な症例数を収集することは臨床的に容易ではなく，解析コストの制約から症例数が限られることもある．同様な症例においての解析結果をまとめて統計的手法で再解析する（メタ解析）必要がある[19]．

3）転写ネットワークの推定

クラスター解析により発現パターンが共通な遺伝子おのおのの上流に共通の転写因子結合部位が抽出できれば，遺伝子制御と機能との関係がわかり，遺伝子間相互作用あるいは転写制御ネットワークを推定することが可能である．酵母ではプロモーター配列が転写開始点の近傍に存在すること，酵母自体が転写で制御される単細胞生物であることなどから，このようなアプローチがきわめて有効であった．ヒトを含む高等動物での問題点は，プロモーター配列どころか遺伝子の転写開始点の同定が不十分であり，スプライスバリアントの存在もアレイ解析のデータのノイズとなる．さまざまな細胞種が混在する生体組織の解析も当然ながらデータのばらつきが増えるので，マイクロダイセクションやフローサイトメトリーなどにより特定の細胞種を選別することも必要である．時系列解析や*in vitro*の培養細胞系において薬剤による介入実験をうまくデザインすることが成功の鍵になる．

4）発現プロファイルデータベース

得られた発現プロファイル解析データのデータベース化については数年来議論されているが，解析プラットフォームの相違やプローブ配列の違いといった本質的な面のみならず，実験プロトコール，検体採取法，スキャニング条件などの技術的な違いによっても解析データが異なることがあるため，複数の施設間での解析データを相互比較することは容易ではない．特に2色の蛍光を用いる方式の場合には対照検体の選択も重要な因子となる．学術論文にアレイデータを用いる場合にはMIAME（minimum information about a microarray experiment）[20]に準拠して公開することが求められるようになり，NCBI，EBIなどで公共データベースGEO（Gene Expression Omnibus），ArrayExpressが提案されている．われわれの研究室でもヒト正常組織，初代培養細胞，癌細胞株など80検体について2万個の遺伝子を解析した結果をデータベース化して公開を行っているのでご利用いただけたら幸いである（システム生物医学データベース：http://www.lsbm.org/）（図4）．

1. 遺伝子発現解析とデータ解析

星田有人，油谷浩幸

マイクロアレイによるトランスクリプトームデータ解析はすでに日常的に行われるようになり，さまざまな生物学的な研究課題に特化した問題点や課題が徐々に明らかになっている．それに伴い新たな解析手法の開発や，プローブ設計を含むプラットフォーム自体の改良が急速に行われており，当初の目標の1つであった臨床応用も全く非現実的なものではなくなりつつある．また，トランスクリプトームデータは網羅性が高く，次々と生成される多様なゲノミクスデータを統合する際の核の1つとしてもさらに重要になっている．

はじめに

現在，mRNA，miRNA，タンパク質，遺伝子多型，転写因子結合部位，染色体の構造変化などの多様なゲノミクスデータを測定するテクノロジーが次々と登場している．これらの嚆矢（こうし）となったDNAマイクロアレイによるトランスクリプトーム解析は，再現性，コスト，コンテンツであるプローブのデザインや遺伝子のアノーテーションなどの点で十分日常的に使用できるレベルに達しつつある．またそのデータ解析は新たな研究領域を創出し解析に関するコンセンサスも形成されつつある[1)~3)]．本稿ではそのプラットフォーム，データ解析の手法などに関する最近の動向を概説する．

[キーワード&略語]
機能ゲノミクス，トランスクリプトーム，マイクロアレイ
FDR：false discovery rate
FWER：family-wise error rate
GIM：genome imbalance map

1 トランスクリプトームデータ解析

トランスクリプトーム解析の普及の理由として，RNA塩基の生化学的な多様性がタンパク質のように大きくない，増幅が容易，RNA安定化技術の向上，豊富なアノーテーションなどのみならず，分子生物学的プロセスのプロファイリングや興味のある表現型に関連する遺伝子（群）のスクリーニングにある程度成功を収めてきたことがあげられるだろう．mRNAの発現量は，生物学的プロセスにおいて中心的な役割をもつタンパク質の発現量とは必ずしも相関しておらず，翻訳後修飾やリン酸化による制御など転写産物レベルで直接確認できない過程が多数存在する．しかしながら，主要な生物学的プロセスの状態の変化は，直接にせよ間接にせよ転写産物のプロファイル（シグネチャー）に生じる影響として捉えることが可能であると考えられる．

また，「同様な発現パターンを示す遺伝子群は機能的な関連を有している」という仮定がある程度有効であることもあげられる（「遺伝子クラスタリングドグマ」ともよばれている[1)]）．実際に，リボソームタン

Transcriptome data analysis
Yujin Hoshida/Hiroyuki Aburatani：Genome Science Division, Research Center for Advanced Science and Technology, University of Tokyo（東京大学先端科学技術研究センターゲノムサイエンス部門）

表　大規模トランスクリプトーム解析の主要なプラットフォーム

プラットフォーム		プローブ（/転写産物）	転写産物数	ラベル	URL
1）スポット型cDNAアレイ	Stanford型アレイ	任意	任意	2色	http://cmgm.stanford.edu/pbrown/
2）オンチップ合成オリゴアレイ	Affymetrix社	25mer×11	47K	単色	http://www.affymetrix.com
	NimbleGen社	60mer×5	38K	単色	http://www.nimblegen.com
	Agilent Technologies社	60mer	41K	2色	http://jp.home.agilent.com
3）スポット型オリゴアレイ	Amersham Biosciences社	30mer	57K	単色	http://www.jp.amershambiosciences.com
4）ビーズアレイ	Illumina社	50mer×2	48K	単色	http://www.illuminakk.co.jp
	Lynx社	17～20mer	*	**	http://www.lynxgen.com

*　ビーズ上にクローニングした任意のcDNAライブラリの配列決定を同時に行う（Massively Parallel Signature Sequencing［MPSS］）
**　SAGEのようにクローンのカウント数を遺伝子発現の絶対量とする．競合ハイブリにより発現の異なる遺伝子も検討可能

パク質や転写開始複合体などのタイトな生物学的プロセスに関与する遺伝子群はどのような測定プラットフォームや解析手法においてもほぼ同様の強固なクラスターを形成する．

一方，プラットフォーム依存的な実験手技上のアーチファクトや生物学的なばらつきによるノイズが相当量存在することも事実である．機能ゲノミクス研究におけるデータ解析には，現時点において特性が完全には明らかになっていないこれらのノイズの中から生物学的，臨床的に意味のある知見を抽出するフィルターの役割が期待されている．

2 遺伝子発現解析のプラットフォーム

遺伝子発現測定のためのDNAマイクロアレイは，スライドグラスやビーズなどの支持体上に固定したDNAプローブとのハイブリダイゼーションにより，数千～万に及ぶRNA種を同時に定量する技術である．主要なアレイプラットフォームの概要を表に示す．これらのアレイの多くについてはカスタムアレイ作製サービスも行われている．

アレイ実験の結果は，プラットフォーム間，同一プラットフォームの世代間，実験者や実験施設間の差に影響されることが知られており，とりわけアレイ間の誤差を解消するためには工業化を含めた品質管理の重要性が指摘されている[4]．

3 データ解析

アレイデータは，従来の分子生物学データとも臨床疫学データとも異質な特性をもつため，その解析手法の開発自体が新たな研究分野となっており，情報科学，統計学などの幅広い領域の研究者が参入している．これらは既存の解析手法の単なる適用のみならず，データに対してどのような仮定が適切かを明らかにすることによりトランスクリプトームの生物学的な挙動の解明に寄与する研究であるとも言えるだろう．AlizadehやGolubらのアレイデータは，統計学におけるFisherのアイリスデータのように新たな解析手法のベンチマーキングに使われている[1]．

1）データの前処理，実験デザイン

アレイデータの正規化はプローブのハイブリダイゼーションなどに関する実験データや仮定に基づいており，解析結果に大きな影響を及ぼすことから，アレイ実験のデザインとともに多くの研究がなされている[1]．また遺伝子のフィルタリングを含む前処理の手法の選択は，マーカー探索やクラス発見などの解析の目的により異なる．またゲノミクスデータに特有の問題として，遺伝子アノーテーションの頻繁な更新に対するアップデートも重要である[5]．

2）データマイニング※1

大きな分類として，データセットに内在する構造を探索するi）～iii）を教師なし（unsupervised）法，興味のある外部情報に関連するデータを抽出するiv）～vi）を教師付き（supervised）法とする場合が多い．

i）次元削減

アレイデータはきわめて高次元であるため，解釈が容易になるように次元削減がしばしば行われる．デー

※1　データマイニング
データセットに含まれる要素間の関係や内在するパターンなどを探索すること．

タセット内の変動の大きい成分のみを抽出する主成分分析（PCA：principle component analysis）や要素間の距離に基づく多次元尺度（MDS：multidimensional scaling）などはデータの視覚化に用いられる．特異値分解（SVD：singular value decomposition）や非負値行列分解（NMF：non-negative matrix factorization）などの行列分解による次元削減は多数の遺伝子を少数の「メタジーン※2」にまとめることによる識別や機能クラスタリングに使われている[1) 6) 7)]．

ii）クラスタリング

標準的に用いられている階層クラスタリングの欠点を克服すべくさまざまな研究がなされている．ノイズの中から意味のある小クラスターを抽出するSuper-Paramagnetic Clustering[8)]や，乱数に基づく多数の初期条件を用いてクラスターの再現性を評価するConsensus Clustering[9)]などは有用である．

iii）遺伝子ネットワークの推定

ベイジアンネットワーク※3を応用した酵母における遺伝子間の機能ネットワーク推定が報告されている[26)]．遺伝子発現データ間の相関係数に基づく共発現ネットワーク（coexpression networks）による遺伝子の機能クラスタリングが試みられている[10)]．また非線形の関連を捉えるために相互情報量を用いたパッケージも公開されている[27)]．

iv）表現型の代用

遺伝子発現プロファイルを細胞の分化を示す表現型の代わりに用いることにより，薬剤のスクリーニングが行われている[11)]．

v）特異的に発現する遺伝子

アレイ解析の主要な応用の1つであり，倍率変化のほか，平均と標準偏差で標準化した後に任意の割合の遺伝子を用いる手法（Zスコア），群間比較の仮説検定，ある測定データが与えられたときに発現が変化していると考えられる確率を計算するベイズ法，統計モデルなどが一般的に用いられている[1)]．

vi）臨床マーカーの構築

表現型などの外部情報をガイドとして新たなサンプルの分類や識別を行う．パターン識別や臨床診断学の手法がほぼそのまま適用されている領域であり，4に述べるパイプライン化のよい対象であろう．分子マーカーの構築は魅力的なテーマであるが，ゲノミクス研究に特異的な問題点に留意する必要がある[12)]．

vii）Semi-supervised法

完全に理想的な表現型情報が揃ったデータを得ることが難しい生物医学研究において，限られた情報から統計学的に有意かつ生物医学的に意味のある知見を最大限に引き出すため，前述のi）～vi）の教師なし，教師付き法の組み合わせが試みられている[13)]．例えば教師なし法で発見したクラスを教師として識別器を構築するなどである．

3）有意性の評価

データ解析の結果を「有意」と判定する閾値を決定するための重要なステップであり，活発な研究がなされている[1)]．注意しなければならないのは，たとえ有意と判定されたとしてもそのデータセットのみに過剰適合した一般化できない結果である場合が大部分であり，4）に述べる検証の過程が必須になるということである[12)]．

i）並べ替え検定

データマイニングの過程で用いる統計量やスコアの分布が既知の確率分布（ガウス分布，二項分布など）に従う保証はどこにもない．並べ替え検定はこのような状況に対処するきわめて現実的な手法である．もとのデータセットに含まれるデータをランダムに並べ替えて計算された統計量の分布を帰無分布とし，もとの統計量より有意な並べ替えの数の割合をp値（nominal p-value）として用いる．どの並べ替えデータのセットを用いるかは，データセットに対する仮定に依存する（遺伝子間の相互作用の有無に対する仮定など）[1) 2)]．

ii）多重性の補正

数千～万という遺伝子の数だけ仮説検定を繰り返す場合，多重性を補正する必要が生じる．1つのデータセット全体のどこかでタイプⅠエラー（偽陽性）が起きる確率（FWER：family-wise error rate）をコントロールするBonferroni補正は保守的で多くの偽陰性の原因となるため，FDR（false discovery rate）な

※2 メタジーン

複数の同様なパターンを示す遺伝子をまとめた少数の仮想的な遺伝子．「メタジーン」をクラスタリングなどに用いている報告が散見される[6)]．

※3 ベイジアンネットワーク

事前確率と事後確率を尤度により結びつけるベイズの定理に基づき遺伝子間の相互作用の確率的なネットワークを構築する手法．

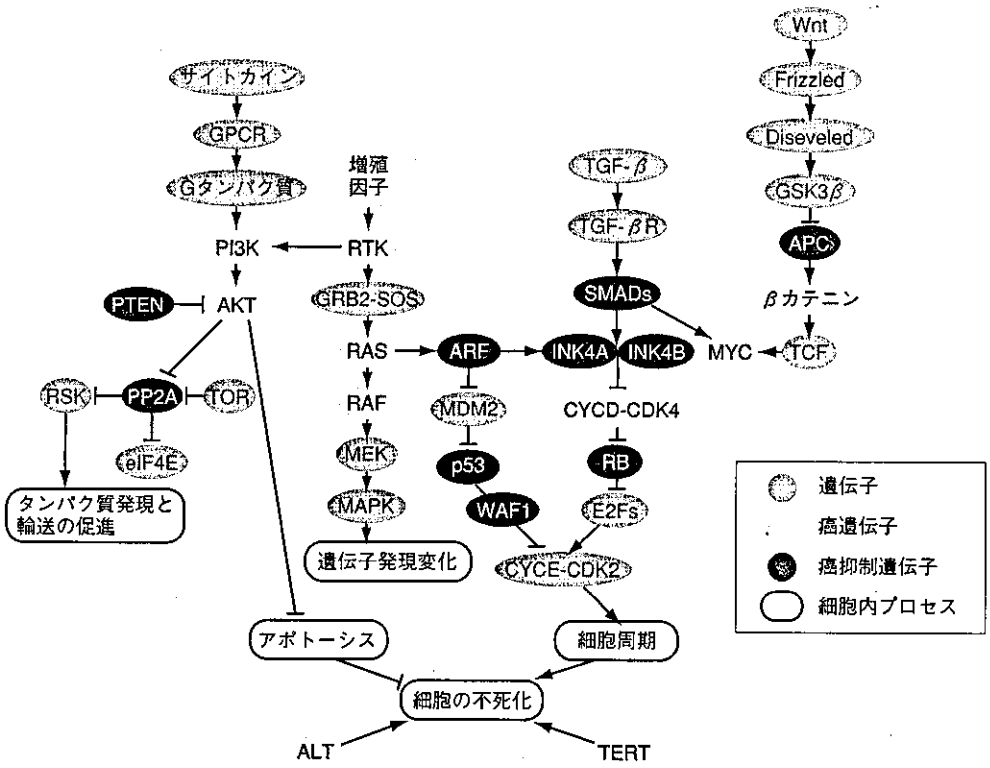

図1　文献情報から抽出した遺伝子間の相互作用ネットワーク
PubMedおよび主要生物医学雑誌の論文からテキストマイニングにより抽出した遺伝子群を用いて再構築した遺伝子間のネットワーク．Wnt/βカテニン，Smadシグナル伝達経路，MAPKカスケードなどの相互作用を示す

どのステップワイズな補正が頻用される[1) 2)]．

4）解析結果の検証，生物学的な意味づけ

i）計算機上の検証

a．交差検証（cross validation）

分類や識別の評価には，新たなデータによる検証や，サンプル数が限られている場合には元データを用いた交差検証やブートストラップ法※4などが標準的に用いられる[1) 2)]．独立したデータセットを用いた検証が行われているマイクロアレイ研究はわずか10％しかないと言われている．また十分なサイズの検証セットを用いていない研究も多くみられる．これらの検証は解析結果が一般化できるか否かに直接かかわっている[12)]．

b．パスウェイへのマップ

解析の結果得られるのは通常，何らかの統計量やスコアの順にソートされたプローブIDのリストである．これらから生物学的な意味を汲み取るための試みの1つとして既知の生物学的機能に属する遺伝子セット（ここではパスウェイとよぶ）が用いられている[14)]．

選び出した遺伝子の発現レベルを単純にパスウェイ上にマップし視覚的なチェックを行う方法から，選び出した遺伝子リスト中に有意に多く含まれるパスウェイの探索，ソートされた遺伝子リストにおいて有意な偏りを示すパスウェイの探索を行う手法などが用いられている[15) 16)]．またわれわれの研究室においては遺伝子リストを文献情報に基づくタンパク質間の相互作用ネットワークにマップするシステムを構築している（図1）．

※4　ブートストラップ法

標本数が限られているため用いる統計量の分布が不明な場合に，オリジナル標本から反復を許しサンプリングしたランダム標本における統計量を得ることにより推定する手法（通常数千回のサンプリングを行う）．推定した分布をもとにp値や統計量の信頼区間が計算できる．

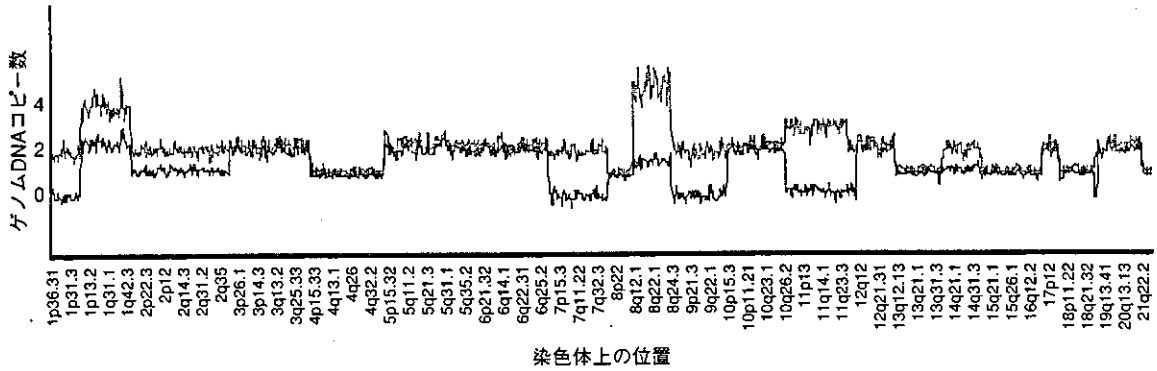

図2　SNPアレイによるgenome imbalance map（巻頭写真1参照）
Affymetrix社のGeneChip®，Mapping arrayのシグナルデータに対して独自の正規化を行い描出したゲノムのコピー数変化．Mapping arrayはもともと，一塩基多型（SNP）のタイピングのために設計されたアレイであるため，従来のCGH解析と異なり，多型の情報を利用してアレル別のコピー数変化を描出できる．また1～10万座位の高密度データから高解像度のゲノムの構造情報を得ることができる．グレーはコピー数が多いアレル，黒はコピー数が少ないアレルを表す

ii）生物学的な検証

標準的に行われている定量PCRのほかにも，組織上で発現遺伝子を定量的に評価する手法が使われはじめている[17]．

5）メタ解析

アレイデータの集積に伴い[18]，プラットフォーム間の再現性やサンプルサイズの問題を克服すべくメタ解析が報告されはじめている[19]．群間比較[20]，共発現ネットワークのメタ解析[11) 21)]が報告されている．

6）他のゲノミクスデータとの統合

ORFomeやpromoteromeやreactomeなど留まるところを知らず多様化するゲノミクスデータは，トランスクリプトームデータがカバーできない領域を補完するものと期待される．われわれの研究室ではSNPアレイのデータからゲノムのアレル別のコピー数変化を描出するGIM（genome imbalance map）を開発している（図2）．このようなゲノム構造の変化は遺伝子の発現レベルにも影響を及ぼしており，発現データと組合わせることにより効率的に疾患関連遺伝子の発現変化を捉えることが可能になると思われる．

アレイデータ解析には，既存の統計学的なツールを巧みに組み合わせて「尤もらしい有意性」をつくり出そうとするトリッキーな側面があるのは否めない．劣決定性を有するデータセットからの過剰な解釈は容易に起こりうる（数千～万もの遺伝子に対して，わずか数十～百のサンプルしかなければ答えはいくらでも存在しうるかもしれない）．以上に紹介したような手法が前提としている仮定のどれが妥当であるかを判断するには，トランスクリプトームの挙動の詳細がより明らかになる必要があるだろう[22]．

4 データ解析のパイプライン

以上のようにアレイデータ解析は多くのステップを経るが，最近の論文では詳細が記述されないこともままある．よく使われているBioConductor[23]のlimmaやaffyなどのパッケージなどを使ったとしても数値変換などの解析フローの詳細が不明であれば解析結果が再現できない場合もある．このような状況を避けるため，大量のデータ解析を日常的に行っている研究室では解析フローの自動化を行っているであろう．

最近ではそのようなデータ解析パイプラインや解析環境の構築をサポートするフリーのパッケージが登場している．GenePatternではPerl，Java，R，Matlabといった任意のコードのモジュールを組合わせて作製したパイプラインファイルをやり取りすることにより解析が完全に再現できる[24]（図3）．クラス識別と交差検証などの定番の解析はすでに実装されている．またサーバー/クライアント型のソフトなので，数日を要するような重い計算は別のサーバーに投げておくこともできる．米国NCIのcaArrayは統合データ解析環

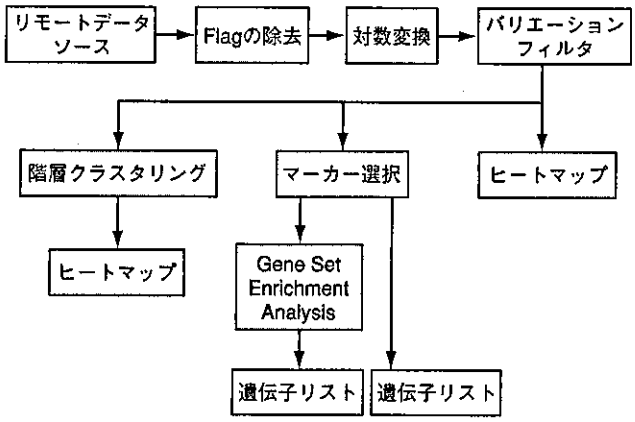

図3　GenePatternにおけるデータ解析パイプラインの例
データ解析ソフトウェアGenePatternでは，データの前処理，複数の並行する解析とその結果のリスト作成や視覚化の手順をパイプラインファイルとして構築，保存することができる．このファイルを実行することにより，全く同じデータ解析が容易に再現できる

境の構築をサポートするオープンソースプロジェクトである[25]．

おわりに

機能ゲノミクスが生物医学研究に貢献するためには，質の高い表現型データの取得が重要な問題になる．また今後，技術開発とコストダウンによりゲノミクスデータの入手はより安価で容易になり，十分な検出力が得られる症例数でデータ測定ができるようになるであろうことを考えると，サンプルサイズ計算法の整備も望まれる[1]．以上に述べたような情報解析のためのスタッフおよびインフラの整備はゲノミクス研究においてクリティカルな問題になると思われる．

文献

1）『統合ゲノミクスのためのマイクロアレイデータアナリシス』(Kohane, I. S. 他／著，星田有人／訳)，シュプリンガーフェアラーク東京，東京，2004
2）Geoffrey, J. et al.：Analyzing microarray gene expression data, Wiley, UK, 2004
3）Genes in action：http://www.sciencemag.org/sciext/genome2004/
4）Tumor Analysis Best Practices Working Group：Nature Rev. Genet., 5：229-237, 2004
5）Onto-Translate：http://vortex.cs.wayne.edu/projects.htm#Onto-Translate
GeneCruiser：http://www.broad.mit.edu/cancer/genecruiser/src/main.jsp
6）Brunet, J. P. et al.：Proc. Natl. Acad. Sci. USA, 101：4164-4169, 2004
7）Kim, P. M. & Tidor, B.：Genome Research, 13：1706-1718, 2003
8）Getz, G. et al.：Proc. Natl. Acad. Sci. USA, 97：12079-12084, 2000
9）Monti, S. et al.：Machine Learning Journal, 52：91-118, 2003
10）Segal, E. et al.：Nature Genet., 36：1090-1098, 2004
11）Stegmaier, K. et al.：Nature Genet., 36：257-263, 2004
12）Ransohoff, D. F.：Nature Review Cancer, 4：309-314, 2004
13）Bair, E. & Tibshirani, R.：PLoS Biol., 2：E108, 2004
14）GENE ONTOLOGY CONSORTIUM：http://www.geneontology.org/index.shtml
BioCarta：http://www.biocarta.com/
GenMapp：http://www.genmapp.org/
15）Onto-Express：http://vortex.cs.wayne.edu/Projects.html#Onto-Express
DAVID：http://apps1.niaid.nih.gov/david/
16）Mootha, M. K. et al.：Nature Genet., 34：267-273, 2003
17）Luminex：http://www.luminexcorp.com/
QuantumDot：http://www.qdots.com/live/index.asp
18）Gene Expression Omnibus：http://www.ncbi.nlm.nih.gov/geo/
ArrayExpress：http://www.ebi.ac.uk/arrayexpress/
19）Moreau, Y. et al.：Trends Genet., 19：570-577, 2004
20）Rhodes, D. R. et al.：Proc. Natl. Acad. Sci. USA, 101：9309-9314, 2004

21) Lee, H. K. et al.：Genome Research, 14：1085-1094, 2003
22) Ueda, H. R. et al.：Proc. Natl. Acad. Sci. USA, 101：3765-3769, 2004
23) Bioconductor：http://www.bioconductor.org/
24) GenePattern：http://www.broad.mit.edu/cancer/software/genepattern/index.html
25) caArray：http://caarray.nci.nih.gov/ caARRAY
26) Lee, I. et al.：Science, 306：1555-1558, 2004
27) ARACNE：http://amdecbioinfo.cu-genome.org/html/caWorkBench.htm

<筆頭著者プロフィール>
星田有人：筑波大学医学専門学群卒業，2004年より Cancer Genomics, Broad Institute, MIT and Harvard University（旧 Whitehead Institute, Center for Genome Research），Golub研究室に留学中．生物学的に新たな知見を抽出する解析アルゴリズムとテクノロジーを臨床情報と統合したトランスレーショナルな機能ゲノミクス研究をめざしています．
E-mail：hoshida@broad.mit.edu