マイクロアレイを用いた遺伝子発現解析

東京大学・大学院農学生命科学研究科
アグリバイオインフォマティクス教育研究ユニット
門田幸二（かどた　こうじ）

自己紹介

- 2002年3月
  - 東京大学・大学院農学生命科学研究科　博士課程修了
  - 学位論文：「cDNAマイクロアレイを用いた遺伝子発現解析手法の開発」（指導教官：清水謙多郎教授）
- 2002/4/1~
  - 産総研・生命情報科学研究センター　産総研特別研究員
- 2003/11/1~
  - 放医研・先端遺伝子発現研究センター　研究員
- 2005/2/16~
  - 東京大学・大学院農学生命科学研究科　特任助手
- 2007/4/1~現在
  - 東京大学・大学院農学生命科学研究科　特任助教

アグリバイオインフォマティクスプログラム

講義内容

- マイクロアレイ解析の流れ（一色法と二色法）
- アレイデータの正規化（前処理）
- 発現変動遺伝子（DEG）の同定
  - 二群間比較
    - 評価基準、評価法、および（Affymetrixチップの）ガイドライン
  - 多サンプル間比較
    - 組織特異的遺伝子
  - 時系列データ
    - 概日リズム関連遺伝子
    - 薬剤応答遺伝子

講義内容

- 機能解析（GSEA解析）
- クラスタリング
- 分類（or 診断）
- 遺伝子ネットワーク解析
- トランスクリプトームデータベース
- 他のトランスクリプトーム解析技術

様々なDNAマイクロアレイ（DNAチップ）

- スポット型（Stanford大学）
  - 搭載DNA：cDNA（またはoligonucleotide）
  - 解析法：2色法（比較したい2サンプルを同時に分析）
- プリント型（Agilent社）
  - 搭載DNA：oligonucleotide（60mer）
  - 解析法： 2色法または1色法
- 合成オリゴ型（Affymetrix社）
  - 搭載DNA：oligonucleotide（25mer）
  - 解析法： 1色法（調べたい1サンプルを分析）

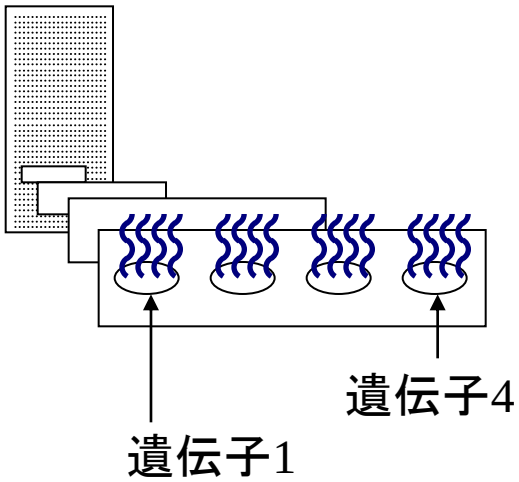
Stanford型

Affymetrix型

マイクロアレイ解析の流れ1

- 目的の生物種（ヒト、マウスなど）のマイクロアレイを入手

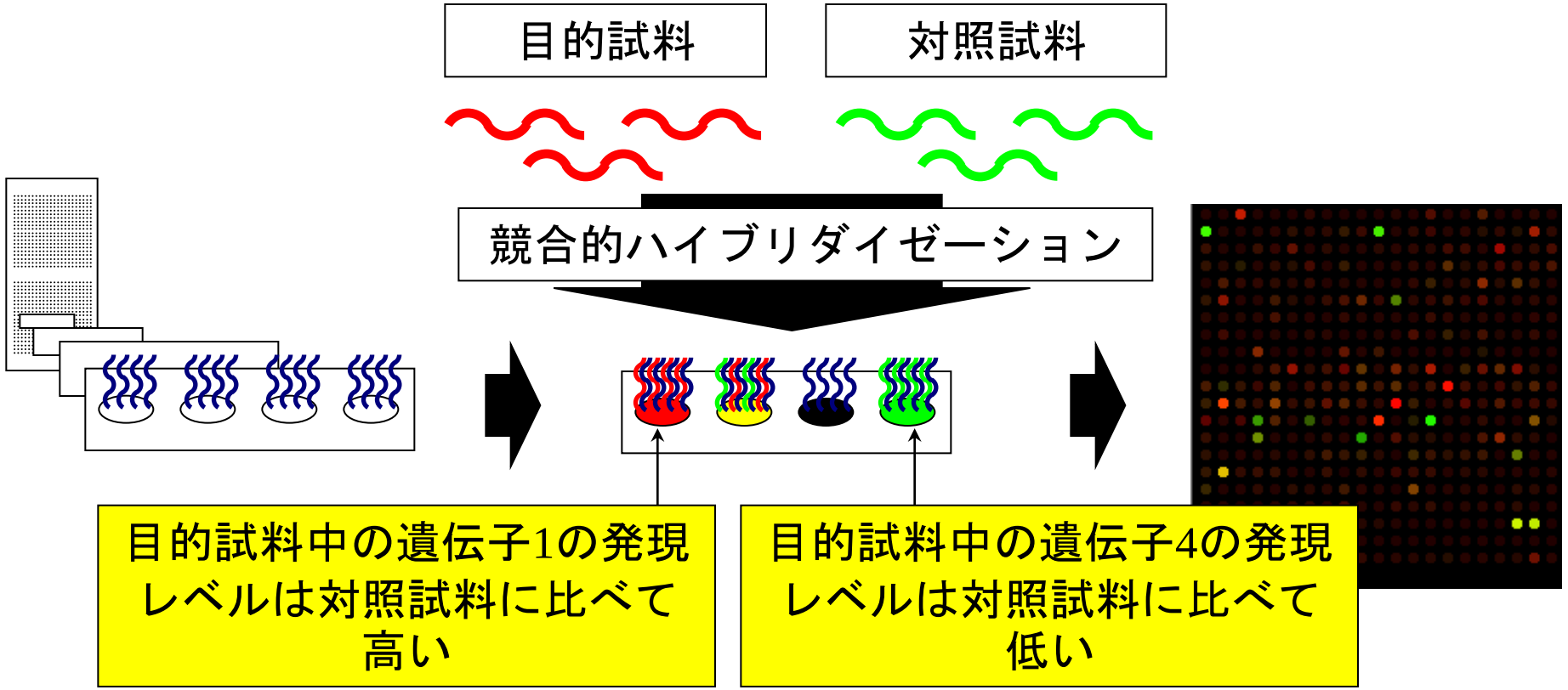
Affymetrix社がGeneChip®
という製品名で販売



- （基本的に）ゲノム配列が決定されている生物種のみ解析可能
- 同じ生物種（例えばヒト）でも、製品のバージョンによって、搭載されている遺伝子数（や種類）が異なる
- 搭載されていない遺伝子の発現量は不明（解析不可能）

マイクロアレイ解析の流れ2(二色法)

- 目的試料中の遺伝子発現レベルを対照試料に対する比として得る



マイクロアレイ解析の流れ1(一色法)

- 目的試料の遺伝子発現レベルをシグナル強度として得る

得られる遺伝子発現データのイメージ

- 二色法の場合

	目的試料	対照試料	目的/対照	$\log_2$(比)
遺伝子1	100	100	1	0
遺伝子2	4000	1000	4	2
遺伝子3	7000	7000	1	0
遺伝子4	2000	8000	0.25	−2
…	…	…	…	…

- 一色法の場合

	目的試料
遺伝子1	100
遺伝子2	4000
遺伝子3	7000
遺伝子4	2000
…	…

目的試料中の遺伝子4の発現レベルは対照試料に比べて2^{-2}倍高い

目的試料中で遺伝子3は沢山発現している

Affymetrix製チップ解析戦略

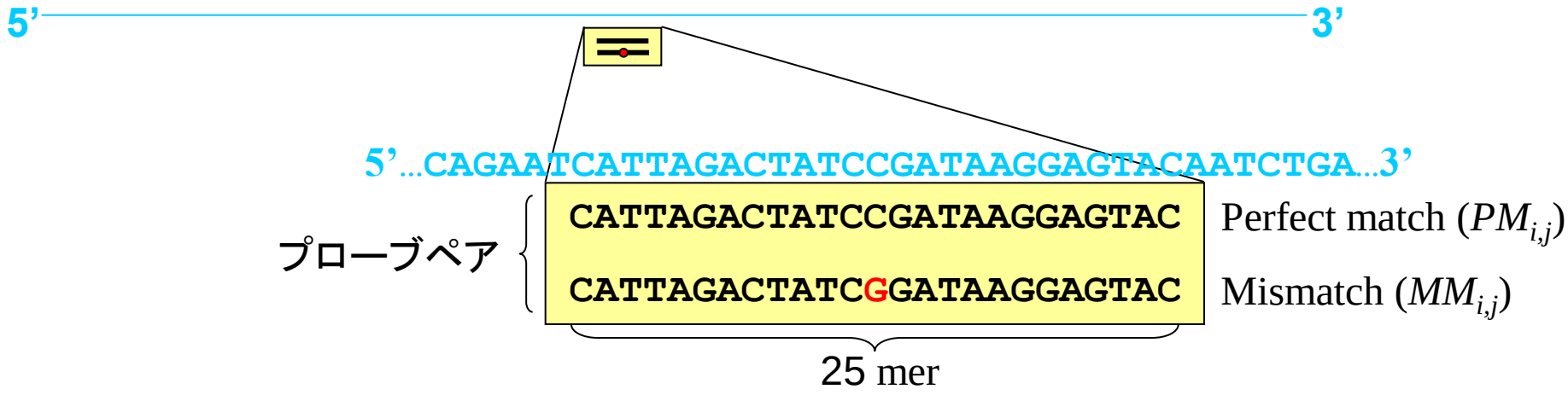
- 25-mer程度では
  - 本当に目的遺伝子の発現を調べられているのか？

    3Gbp(=3×10^9 bp) vs. 4^25 (=1×10^15 bp)

  - 発現量を正確に定量できるのか？

Affymetrix製チップ解析戦略

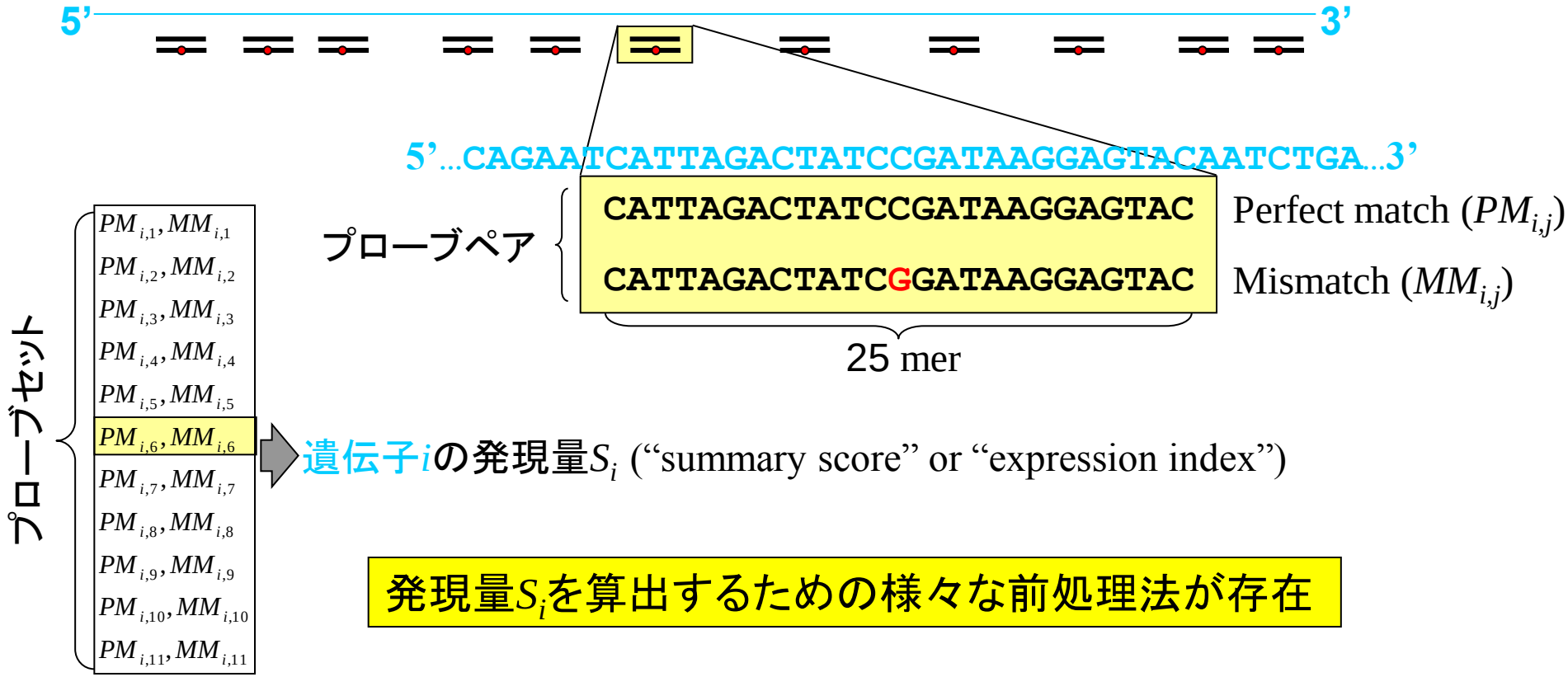
- 遺伝子iの発現量S_iを正確に知るために
  - PM/MMプローブ戦略（ユニークな配列選択と最適T_m）



特異的なハイブリダイゼーションと非特異的なハイブリダイゼーションを区別すべく、目的遺伝子配列に対してPMと一塩基MMがペアになっているのが特徴的

Affymetrix製チップ解析戦略

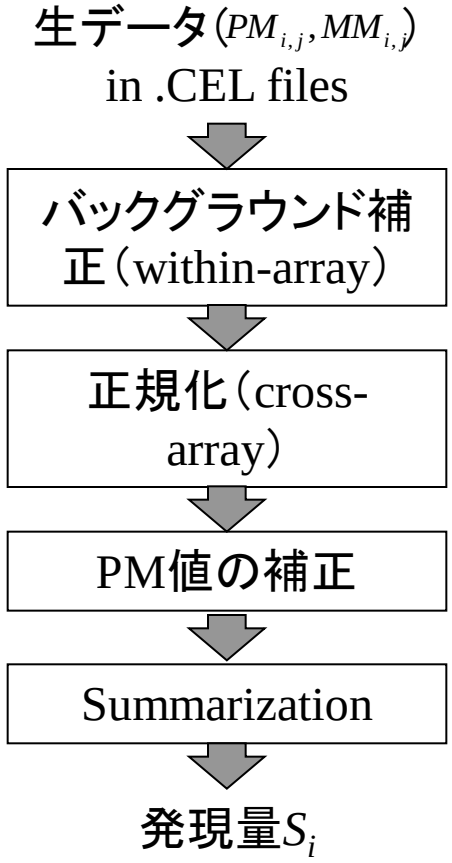
- 遺伝子iの発現量S_iをn_i(n_i=11~20)種類のプローブペアのシグナル強度をもとに計算



発現量S_iを算出するための様々な前処理法が存在

Affymetrix製チップ解析戦略（様々な前処理法）

- MBEI (Li and Wong, *PNAS*, **98**, 31-36, 2001)
- MAS5 (Hubbell *et al.*, *Bioinformatics*, **18**, 1585-92, 2002)
- RMA (Irizarry *et al.*, *Biostatistics*, **4**, 249-64, 2003)
- GCRMA (Wu *et al.*, *Tech. Rep.*, *John Hopkins Univ.*, 2003)
- PDNN (Zhang *et al.*, *Nat. Biotechnol.*, **21**, 818-21, 2003)
- PLIER (Affymetrix, 2004)
- SuperNorm (Konishi, T., *BMC Bioinformatics*, **5**, 5, 2004)
- multi-mgMOS (Liu *et al.*, *Bioinformatics*, **21**, 3637-3644, 2005)
- GLA (Zhou and Rocke, *Bioinformatics*, **21**, 3983-3989, 2005)
- FARMS (Hochreiter *et al.*, *Bioinformatics*, **22**, 943-949, 2006)
- DFW (Chen *et al.*, *Bioinformatics*, **23**, 321-327, 2007)
- Hook (Binder *et al.*, *AMB*, **3**, 11, 2008)



Availability: The R code for DFW is available upon request.

(Rで)マイクロアレイデータ解析 Microarray data analysis using R (last modified 2009.8.12)

What's new?
・全体的に必要に応じて変更すべき箇所を前半部分に移動させて、エラーがより起こりにくくするなどしています。(2009/7/10-)
・いろいろ追加しました。(2009/5/25-6/22)
・リンク切れの修正や、記述の統一などを行いました(2009/5/20-22)
・Bioconductor 2.4が2009/4/21にリリースされてたので、古いRパッケージのリンクを全て2.4のものに変更しました(2009/5/20)
・Affymetrix GeneChipデータ解析を行う上での推奨ガイドラインを掲載しました(2009/4/24)
・二群間比較用の私の最新手法を掲載しました(2008/6/26)

- はじめに (last modified 2009/8/7) NEW
- Rのインストールと起動(last modified 2009/8/10) NEW
- Rの昔のバージョンのインストール (last modified 2009/8/12) NEW
- 使用例(初心者向け)(last modified 2009/7/9)
- サンプルマイクロアレイデータ (last modified 2009/8/4) NEW
- 発現データ取得 | Affymetrix data全体 | Celsius (Day 2007) (last modified 2007/11/13)
- 発現データ取得 | Gene Expression Omnibus (GEO)から | GEOquery (Davis 2007) (last modified 2009/8/5) NEW
- 発現データ取得 | ArrayExpressから | ArrayExpress (last modified 2009/5/28)
- アノテーション情報取得 | Rのパッケージから (last modified 2009/8/5) NEW
- アノテーション情報取得 | GEOから (last modified 2009/8/5) NEW
- 正規化(cDNA or two-color or 二色法) について(last modified 2008/3/31)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: limma)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ (package: marray)
- 正規化 | Stanford型 (or cDNA)マイクロアレイ GPA (Xiong 2008) (last modified 2008/3/10)
- 正規化(Affymetrix) について(last modified 2009/7/9)
- 正規化 | Affymetrix GeneChip | RMA++, Extrapolation Averaging (Harbron 2007) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | RMA+, Extrapolation Strategy, refRMA (Harbron 2007) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | DFW (Chen 2007) (last modified 2009/8/12) NEW
- 正規化 | Affymetrix GeneChip | FARMS (Hochreiter 2006) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | multi-mgMOS (Liu 2005) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | GLA (Zhou 2005) (last modified 2007/4/20)
- 正規化 | Affymetrix GeneChip | GCRMA (Wu 2004) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | PLIER (Affymetrix 2004) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | PDNN (Zhang 2003) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | VSN (Huber 2002) (last modified 2009/8/6) NEW
- 正規化 | Affymetrix GeneChip | MAS, MBEI, RMA (package: affy) (last modified 2009/8/6) NEW

アレイデータの正規化（前処理）

- 実験によって得られた生のシグナル強度をそのまま利用することは普通はやりません
  - 二色法：蛍光色素（Cy3 and Cy5）の取り込み効率補正
  - 一色法：シグナルゲイン?!の補正

「こうであるべき！」という仮定を置いて、それを満たすような正規化を行った後のデータを利用する

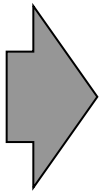
グローバル正規化

- 仮定：各サンプルから測定されたmRNAの全体量は一定

  チップ上の遺伝子数が少ない場合は非現実的だが、数千～数万種類の遺伝子が搭載されているので妥当（だろう）

	sample1	sample2	Ratio (sample1/sample2)	$\log_2$(Ratio)
gene1	10.5	12.4	0.84	-0.243
gene2	6.4	7.1	0.91	-0.141
gene3	8.0	8.5	0.94	-0.086
gene4	10.8	11.4	0.95	-0.075
gene5	5.6	6.7	0.83	-0.262
gene6	8.4	8.9	0.94	-0.090
gene7	6.2	7.0	0.90	-0.159
gene8	6.1	6.8	0.90	-0.145
gene9	6.6	6.5	1.01	0.010
gene10	5.1	5.8	0.89	-0.165
				-0.136

nomalization



$\log_2$(Ratio)
-0.107
-0.005
0.049
0.061
-0.126
0.045
-0.023
-0.010
0.145
-0.030
0.000

Quantile正規化

- 仮定：順位が同じならシグナル強度も同じ

正規化前

sample1	sample2
10.5	12.4
6.4	7.1
8.0	8.5
10.8	11.4
5.6	6.7
8.4	8.9
6.2	7.0
6.1	6.8
6.6	6.5
5.1	5.8

列ごとにソート →

sample1	sample2
5.1	5.8
5.6	6.5
6.1	6.7
6.2	6.8
6.4	7.0
6.6	7.1
8.0	8.5
8.4	8.9
10.5	11.4
10.8	12.4

行ごとの平均を算出 →

Average
5.4
6.1
6.4
6.5
6.7
6.8
8.3
8.6
10.9
11.6

対応する行の要素の元の位置に平均値を代入 →

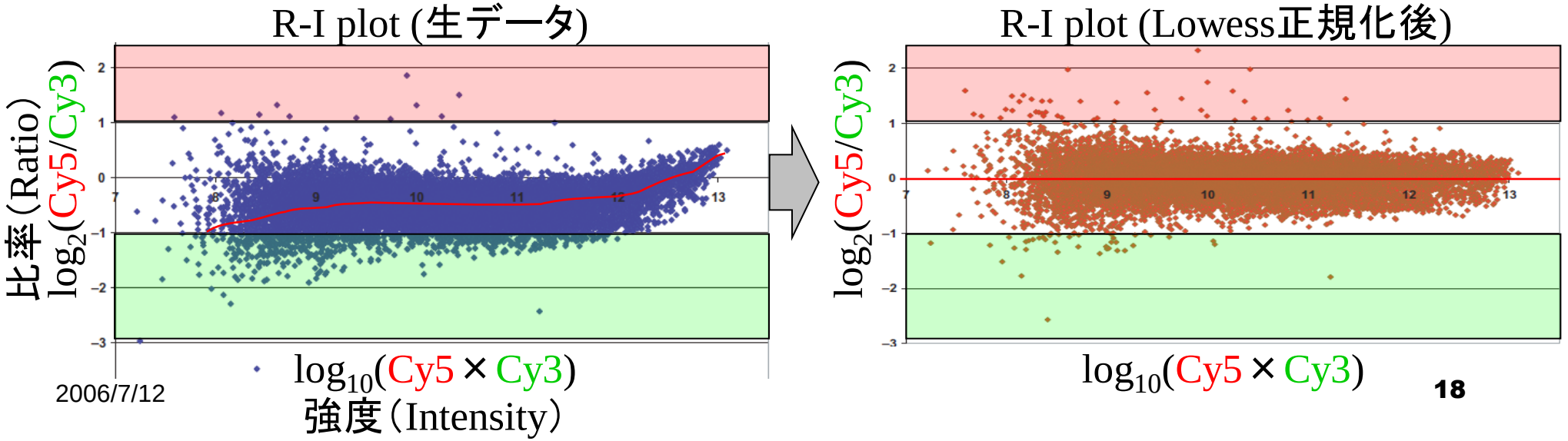
正規化後

sample1	sample2
10.9	11.6
6.7	6.8
8.3	8.3
11.6	10.9
6.1	6.4
8.6	8.6
6.5	6.7
6.4	6.5
6.8	6.1
5.4	5.4

データセット中のサンプル数が変わると結果が変わる

Lowess(**Lo**cally **wei**ghted **s**catterplot smoothing)正規化

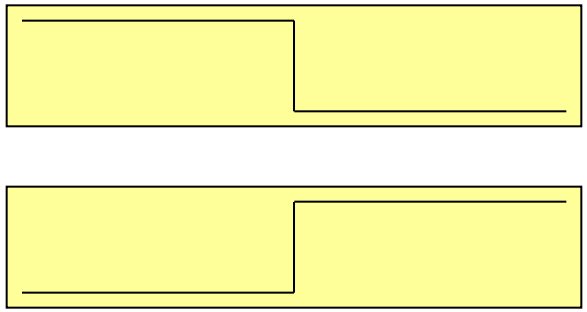
- 仮定：log比の分布はシグナル強度非依存である



正規化 → 遺伝子発現行列

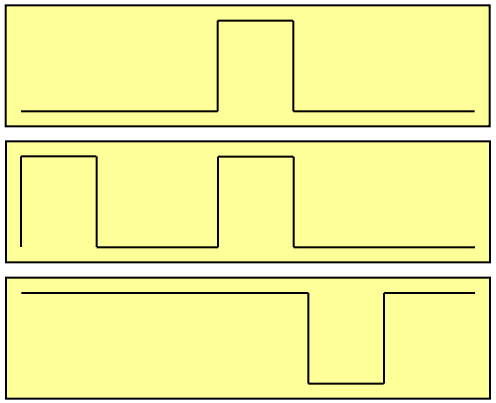
二群間比較

	A群 A1	A群 A2	…	B群 B1	B群 B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



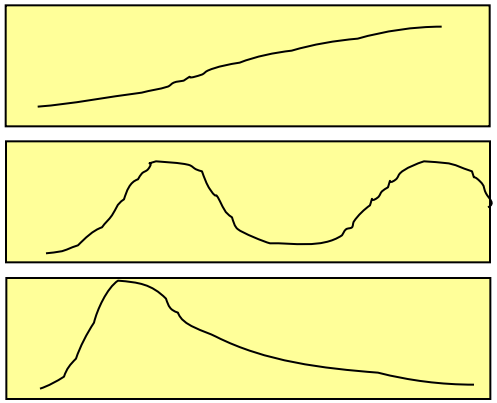
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



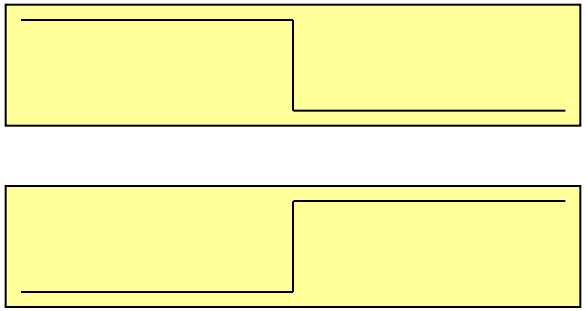
発現変動遺伝子の同定が可能な状態

二群間比較

- 例1）
  - A群：癌サンプル
  - B群：正常サンプル

  →癌と正常で発現の異なる遺伝子

	A群			B群		
	A1	A2	…	B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



二群間比較

- 例2）急性白血病
  - A群：リンパ性（27 サンプル）
  - B群：骨髄性（11サンプル）

二群間比較（解析手法）

- 倍率変化（Fold change; FC）に基づくランキング法
  - **2-fold, 3-fold (FC)**
  - The limit fold change model (Mutch *et al.*, *BMC Bioinformatics*, 2002)
  - **Rank product** (RP; Breitling *et al.*, *FEBS Lett.*, 2004)
  - **WAD** (Kadota *et al.*, *Algorithm. Mol. Biol.*, 2008)
  - …
- *t*-統計量に基づくランキング法
  - a signal-to-noise statistic (Golub *et al.*, Science, 1999)
  - **Student's (or Welch) *t*-test**
  - **SAM** (**samT**; Tusher *et al.*, *PNAS*, 2001)
  - **Samroc** (Broberg, P., *Genome Biol.*, 2003)
  - **a moderated *t* statistic** (Smyth, GK., *Stat. Appl. Genet. Mol. Biol.*, 2004)
  - **Intensity-based moderated *t* statistic** (IBMT; Sartor *et al.*, *BMC Bioinformatics*, 2006)
  - **Shrinkage *t* statistic** (Opgen-Rhein and Strimmer, *Stat. Appl. Genet. Mol. Biol.*, 2007)
  - …
- その他
  - Probability of Positive LogRatio (PPLR; Liu *et al.*, *Bioinformatics*, 2006)
  - FCPC (Qin *et al.*, *Bioinformatics*, 2008)

二群間比較（t-統計量に基づくランキング法）

- 「二群間の平均の差が大きく」、「群内のばらつきが小さい」遺伝子iを抽出
- a signal-to-noise(S2N)統計量

$$R(i) = \frac{\overline{A^i} - \overline{B^i}}{U_{A^i} + U_{B^i}}$$

← 二群間の平均の差

U_{A^i}：A群内のばらつき　U_{B^i}：B群内のばらつき

標本平均　$\overline{A^i} = \frac{1}{n_A}\sum_{j=1}^{n_A} A_j^i$

標本分散　$S_{A^i}^2 = \frac{1}{n_A}\sum_{j=1}^{n_A}(A_j^i - \overline{A^i})^2$

不偏分散　$U_{A^i}^2 = \frac{1}{n_A - 1}\sum_{j=1}^{n_A}(A_j^i - \overline{A^i})^2$

$n_A = 6, n_B = 5, n = n_A + n_B$

対数変換（log2変換）後のデータ

		A群						B群				
i		A_1^i	A_2^i	A_3^i	A_4^i	A_5^i	A_6^i	B_1^i	B_2^i	B_3^i	B_4^i	B_5^i
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = \frac{6.42 - 4.00}{0.08 + 0.35} = \frac{2.41}{0.43} = 5.64$$

$$R(2) = \frac{6.34 - 3.38}{0.54 + 1.65} = \frac{2.96}{2.20} = 1.35$$

$$R(3) = \frac{4.51 - 5.61}{0.81 + 0.07} = \frac{-1.11}{0.88} = -1.26$$

統計量の絶対値が大きい → 候補発現変動遺伝子

二群間比較（t-統計量に基づくランキング法）

- t検定（等分散を仮定）の統計量

検定統計量t^iは、自由度n_A+n_B-2のt分布に従う

$$R(i) = t^i = \frac{\overline{A^i} - \overline{B^i}}{\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}\sqrt{\frac{(n_A-1)U_{A^i}^2 + (n_B-1)U_{B^i}^2}{n_A+n_B-2}}}$$

← 二群間の平均の差

← ばらつき

対数変換（log2変換）後のデータ

		A群						B群				
i		$A_1{}^i$	$A_2{}^i$	$A_3{}^i$	$A_4{}^i$	$A_5{}^i$	$A_6{}^i$	$B_1{}^i$	$B_2{}^i$	$B_3{}^i$	$B_4{}^i$	$B_5{}^i$
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = t^1 = \frac{2.41}{0.15} = 16.64$$

$$R(2) = t^2 = \frac{2.96}{0.71} = 4.16$$

$$R(3) = t^3 = \frac{-1.11}{0.37} = -3.00$$

統計量の絶対値が大きい → 候補発現変動遺伝子

二群間比較（*t*-統計量に基づくランキング法）

- *t*検定（不等分散を仮定）の統計量

$$R(i) = t^i = \frac{\overline{A^i} - \overline{B^i}}{\sqrt{\frac{U_{A^i}^{\ 2}}{n_A} + \frac{U_{B^i}^{\ 2}}{n_B}}}$$

←二群間の平均の差

←ばらつき

検定統計量t^iは、自由度ν（にゅー）の*t*分布に従う

$$\nu = \frac{\left(U_{A^i}^2 / n_A + U_{B^i}^2 / n_B \right)^2}{\left\{ (U_{A^i}^2 / n_A)^2 / (n_A - 1) + (U_{B^i}^2 / n_B)^2 / (n_B - 1) \right\}}$$

対数変換（log2変換）後のデータ

		A群						B群				
i		A_1^i	A_2^i	A_3^i	A_4^i	A_5^i	A_6^i	B_1^i	B_2^i	B_3^i	B_4^i	B_5^i
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = t^1 = \frac{6.42 - 4.00}{\sqrt{0.08^2/6 + 0.35^2/5}} = 15.17$$

$$R(2) = t^2 = \frac{6.34 - 3.38}{\sqrt{0.54^2/6 + 1.65^2/5}} = 3.83$$

$$R(3) = t^3 = \frac{4.51 - 5.61}{\sqrt{0.81^2/6 + 0.07^2/5}} = -3.32$$

統計量の絶対値が大きい → 候補発現変動遺伝子

多重検定問題

		真実	
		H_0:差がない	H_1:差がある
検定結果	H_0	正しく判断 (1－α)	Type-II error (β)
	H_1	Type-I error (α)	正しく判断 (1－β)

- 「ある一つの遺伝子の発現データについて差があるかどうかを検定する」という作業を全遺伝子について行う

  帰無仮説H_0:**差がない**、　対立仮説H_1:**差がある**

- 有意水準（危険率；error rate）αを予め設定
  - Type-I error（本当は発現に差がないのに差があるとしてしまう誤り）を制御

これをN回（N個の遺伝子について）繰り返すと…

下手な鉄砲も数打ちゃ当たる

	普通の餌(Reference)					トクホ含有餌(Target)				
	R1	R2	R3	R4	R5	T1	T2	T3	T4	T5
gene1										
gene2										
…										
gene99										
gene100										

- N=100($\alpha = 0.05$)としてみると
  - 一連の検定(計100回)のどこかで第一種の誤り(Type-I error)をおかす確率(family-wise error rate; FWER)

$$= 1 - \text{間違わない確率}(1-\alpha)\text{が}N\text{回続けて起こる確率}$$

$$= 1-(1-\alpha)^N = 1-(1-0.05)^{100} = 0.994$$

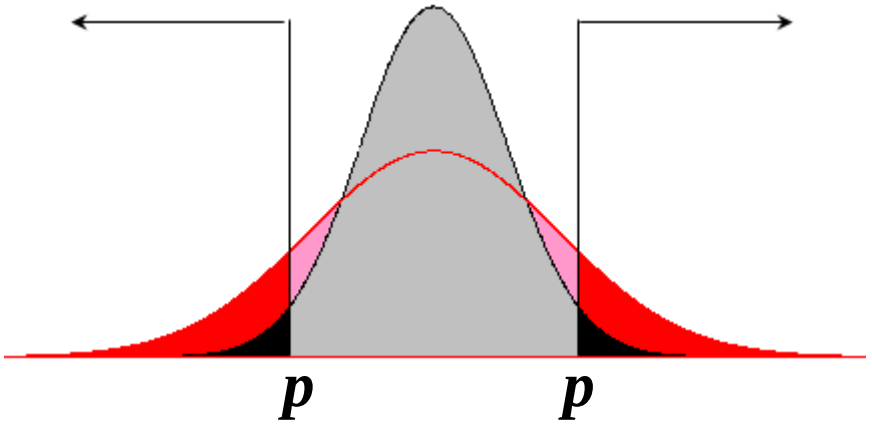
一連の検定のどこかで間違って帰無仮説を棄却してしまう確率(本当は「差がない」のに「差がある」としてしまう確率)はかなり大きい

→コントロールすべきはαではなくFWER

N	FWER	
	α=0.05	α=0.01
5	0.2262	0.0490
10	0.4013	0.0956
20	0.6415	0.1821
50	0.9231	0.3950
100	0.9941	0.6340
200	1.0000	0.8660
500	1.0000	0.9934
1000	1.0000	1.0000

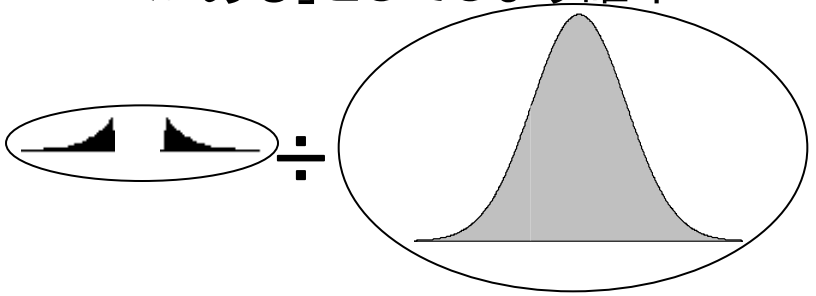
False Discovery Rate (FDR)を制御

- 検定によって帰無仮説が棄却された結果の数に占めるType-I errorの割合（FDR; *q*-value）を制御する、という考え方



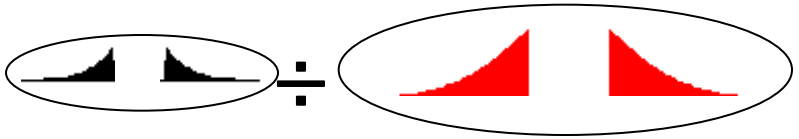
p-value (FPR)

本当は発現に「差がない」にもかかわらず「差がある」としてしまう確率



q-value (FDR)

発現に差が「ある」とされたもののうち、本当は発現に「差がない」ものの割合



FDR計算イメージ

$$R(1) = t^1 = \frac{85.50 - 16.40}{\sqrt{4.68^2/6 + 3.85^2/5}} = 26.88$$

$$R(2) = t^2 = \frac{85.50 - 16.40}{\sqrt{29.20^2/6 + 16.50^2/5}} = 4.93$$

$$R(3) = t^3 = \frac{25.50 - 49.00}{\sqrt{11.73^2/6 + 2.24^2/5}} = -4.81$$

1. 統計量を計算

例）*t*統計量（不等分散性を仮定；Welch検定）

$$R(i) = t^i = \frac{\overline{A^i} - \overline{B^i}}{\sqrt{\frac{U_{A^i}^{\ 2}}{n_A} + \frac{U_{B^i}^{\ 2}}{n_B}}}$$

← 二群間の平均の差

← ばらつき

	A群						B群					
	A_1	A_2	A_3	A_4	A_5	A_6	B_1	B_2	B_3	B_4	B_5	統計量
gene1	87	79	91	82	90	84	12	21	19	13	17	26.88
gene2	56	122	106	47	84	98	7	44	2	11	18	4.93
gene3	15	28	33	9	27	41	48	46	52	50	49	-4.81
gene4	46	28	33	20	27	41	48	46	52	27	49	-2.00
gene5	30	60	81	69	42	39	58	41	27	92	73	-0.34
gene6	46	28	33	20	27	41	48	47	52	26	49	-1.95

|統計量|	Observed
0.3	6
0.4	5
1.0	5
2.0	4
3.0	3
4.0	3
5.0	1

\|統計量\| ≧1.0を満たす遺伝子を「差がある」とすると5個ある、という意味

FDR計算イメージ

	A群						B群					
	A_1	A_2	A_3	A_4	A_5	A_6	B_1	B_2	B_3	B_4	B_5	統計量
gene1	87	79	91	82	90	84	12	21	19	13	17	26.88
gene2	56	122	106	47	84	98	7	44	2	11	18	4.93
gene3	15	28	33	9	27	41	48	46	52	50	49	-4.81
gene4	46	28	33	20	27	41	48	46	52	27	49	-2.00
gene5	30	60	81	69	42	39	58	41	27	92	73	-0.34
gene6	46	28	33	20	27	41	48	47	52	26	49	-1.95

2. 並べ替え検定（random permutation test）の実行

「偶然差があるとされる遺伝子数」を見積もる

1回目

	A群						B群					
	B_5	A_2	B_4	A_1	A_5	B_3	B_2	B_1	A_3	A_4	A_6	統計量
gene1	17	79	13	87	90	19	21	12	91	82	84	-0.31
gene2	18	122	11	56	84	2	44	7	106	47	98	-0.43
gene3	49	28	50	15	27	52	46	48	33	9	41	0.15
gene4	49	28	27	46	27	52	46	48	33	20	41	0.08
gene5	73	60	92	30	42	27	41	58	81	69	39	-0.27
gene6	49	28	26	46	27	52	47	48	33	20	41	0.03

2回目

	A群						B群					
	A_3	A_1	B_3	B_4	A_2	B_1	A_6	A_5	B_2	B_5	A_4	統計量
gene1	91	87	19	13	79	12	84	90	21	17	82	-0.38
gene2	106	56	2	11	122	7	98	84	44	18	47	-0.29
gene3	33	15	52	50	28	48	41	27	46	49	9	0.34
gene4	33	46	52	27	28	48	41	27	46	49	20	0.33
gene5	81	30	27	92	60	58	39	42	41	73	69	0.40
gene6	33	46	52	26	28	48	41	27	47	49	20	0.28

3回目

…

|統計量|	Observed	Randomized				
		1回目	2回目	…	mean	FDR
0.3	6	2	4		3	**50.0%**
0.4	5	1	0		0.5	**10.0%**
1.0	5	0	0		0	**0.0%**
2.0	4	0	0		0	**0.0%**
3.0	3	0	0		0	**0.0%**
4.0	3	0	0		0	**0.0%**
5.0	1	0	0		0	**0.0%**

二群間比較(倍率変化に基づくランキング法)

- log比:(対数変換後のデータなので)*t*検定系の数式の分子のみに相当

$$R(i) = \log(FC) = \overline{A^i} - \overline{B^i}$$ ←二群間の平均の差

対数変換(log2変換)後のデータ

		A群						B群				
i		$A_1{}^i$	$A_2{}^i$	$A_3{}^i$	$A_4{}^i$	$A_5{}^i$	$A_6{}^i$	$B_1{}^i$	$B_2{}^i$	$B_3{}^i$	$B_4{}^i$	$B_5{}^i$
1	gene1	6.4	6.3	6.5	6.4	6.5	6.4	3.6	4.4	4.2	3.7	4.1
2	gene2	5.8	6.9	6.7	5.6	6.4	6.6	2.8	5.5	1	3.5	4.2
3	gene3	3.9	4.8	5	3.2	4.8	5.4	5.6	5.5	5.7	5.6	5.6

$$R(1) = 6.42 - 4.00 = 2.41$$
$$R(2) = 6.34 - 3.38 = 2.96$$
$$R(3) = 4.51 - 5.61 = -1.11$$

統計量の絶対値が大きい → 候補発現変動遺伝子

参考資料

二群間比較（倍率変化に基づくランキング法）

- WAD：log比を基本としつつ、全体的にシグナル強度の高い遺伝子が上位にくるように重みをかけた統計量

unlogged data

Gene	A1	A2	A3	B1	B2
gene1	128	64	128	128	64
gene2	1024	1024	1024	1024	1024
gene3	512	1024	1024	2048	246
gene4	1024	1024	2048	256	256
gene5	2	2	2	32	32
gene6	2	4	4	64	128
gene7	16	8	32	64	8

$\log_2$-transformed data

Gene	A1	A2	A3	B1	B2
gene1	7	6	7	7	6
gene2	10	10	10	10	10
gene3	9	10	10	11	8
gene4	10	10	11	8	8
gene5	1	1	1	5	5
gene6	1	2	2	6	7
gene7	4	3	5	6	3

Average Difference（AD）統計量

$AD_i = \left|\overline{B_i} - \overline{A_i}\right|$

AD	rank
0.17	6
0.00	7
0.20	5
2.33	3
4.00	2
4.83	1
0.50	4

平均シグナル強度

$x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$

xを(0~1)の範囲に規格化

$w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$

WAD統計量

$WAD_i = AD_i \times w_i$

x	w	WAD	rank
6.58	0.51	0.09	5
10.00	1.00	0.00	6
9.57	0.94	0.18	3
9.17	0.88	2.06	1
3.00	0.00	0.00	6
4.08	0.15	0.75	2
4.25	0.18	0.09	4

$AD_i = \left|\overline{B_i} - \overline{A_i}\right|$ より

$AD_{gene6} = \left|(6+7)/2 - (1+2+2)/3\right|$

$= 4.83$

$x_i = \left(\overline{B_i} + \overline{A_i}\right)/2$ より

$x_{gene6} = \left((6+7)/2 + (1+2+2)/3\right)/2$

$= 4.08$

$w_i = \dfrac{x_i - \min(x)}{\max(x) - \min(x)}$ より

$w_{gene6} = \dfrac{4.08 - 3.00}{10.00 - 3.00} = 0.15$

WADの一位：gene4, ADの一位：gene6

二群間比較（倍率変化に基づくランキング法）

- Rank products (RP)：A群 vs. B群の総当たりの比を計算し、その順位の相乗平均を統計量とする

入力データ

	A1	A2	A3	B1	B2	B3
gene1	a11	a12	a13	b11	b12	b13
…	…	…	…	…	…	…
genei	ai1	ai2	ai3	bi1	bi2	bi3
…	…	…	…	…	…	…
genen	an1	an2	an3	bn1	bn2	bn3

$n_A = 3$　$n_B = 3$

総当りの発現比を計算

$(n_A \times n_B) = 9$通り

A1/B1	A1/B2	A1/B3	A2/B1	A2/B2	A2/B3	A3/B1	A3/B2	A3/B3
a11/b11	a11/b12	a11/b13	a12/b11	a12/b12	a12/b13	a13/b11	a13/b12	a13/b13
ai1/bi1	ai1/bi2	ai1/bi3	ai2/bi1	ai2/bi2	ai2/bi3	ai3/bi1	ai3/bi2	ai3/bi3
an1/bn1	an1/bn2	an1/bn3	an2/bn1	an2/bn2	an2/bn3	an3/bn1	an3/bn2	an3/bn3

列ごとにRankを計算した後、各行に対して相乗平均値（RPs）を計算

	RP
gene1	RP1
…	
genei	RPi
…	
genen	RPn

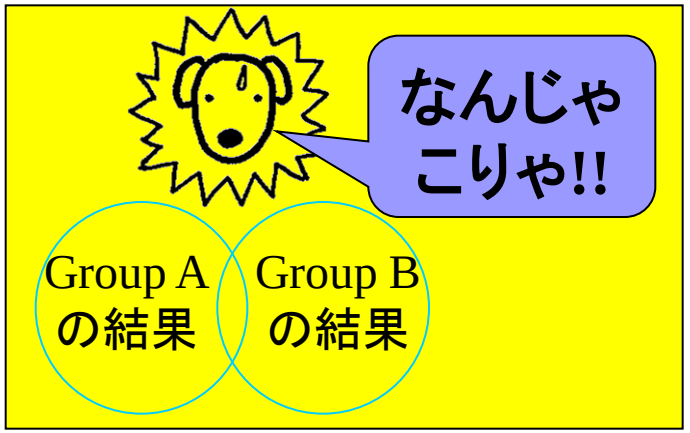
実用化にむけた取り組み

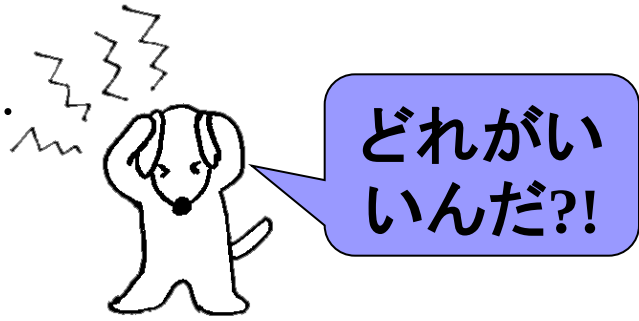
- 国外
  - MicroArray Quality Control (MAQC)プロジェクト (2005/2-2006/9)
  - External RNA Control (ERC) Consortium
  - MAQC-II (2006/9-2009/3)
- 国内
  - バイオチップコンソーシアム(JMAC)
    - 2007年10月に設立
    - バイオ産業分野の業界団体

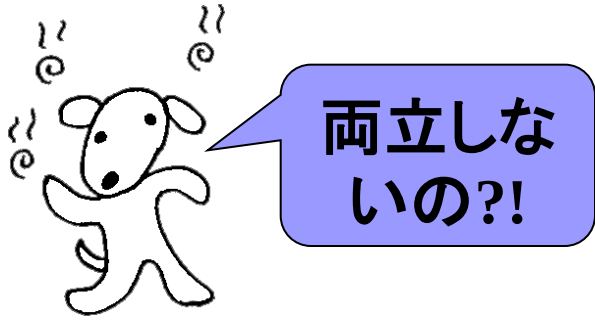
解決すべき課題

- 再現性は本当にあるのか？
  - プラットフォーム間（メーカーの違い）
  - プラットフォーム内（実験場所の違い）
- どの解析手法がいいか？
  - 前処理（正規化）法：MAS5, RMA, MBEI, …
  - 発現変動遺伝子検出法
    - 組織特異的遺伝子：Dixon test, ROKU, …
    - 二群間比較（癌 vs. 正常）：*t*-test, SAM, …
- 重視すべき評価基準は？
  - 「感度・特異度」重視派
  - 「再現性（MAQCプロジェクト提唱）」重視派

「感度・特異度」と「再現性」は両立しない？！





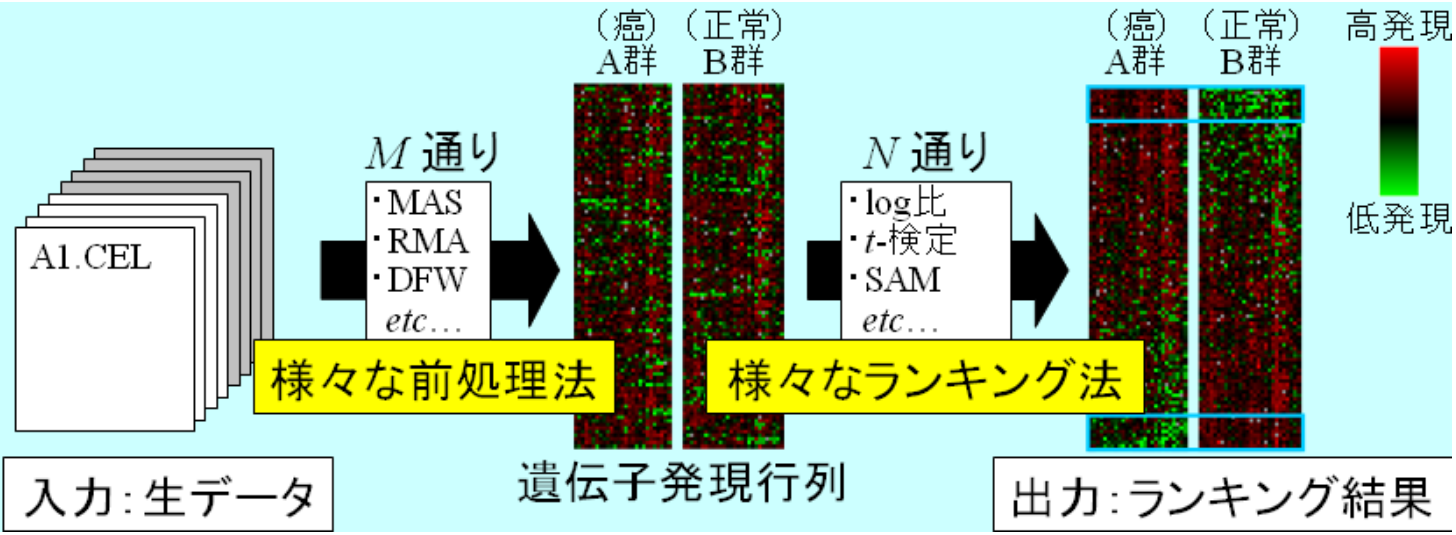


これまでの流れ

- 「マイクロアレイ再現性が低いぞ、やべー」
- 「これだけ再現性が低かったら臨床応用とかできるの？」
- MicroArray Quality Control (MAQC) プロジェクト(2005/2-)
  - 2006年秋ごろの*Nature Biotechnology*誌に一連の研究成果を発表
  - 「再現性が低いのは*t*-統計量系の方法(*p*値を出すやつ)を使っていたから。しかもかなりキツメの*p*値だったから。」
  - 「*t*-統計量系の方法は感度・特異度は高いかもしれんが、再現性がいまいちだな。**倍率変化に基づく方法は再現性が非常に高いことが分かったよ**。」
  - どのメーカーのアレイを使っても、発現変動遺伝子を検出するという観点では実用に耐えうる。
  - 「*t*-統計量系と倍率変化系の方法は感度・特異度と再現性の点においてトレードオフの関係にあるね。よって、実際の利用として、**緩めの*p*値でカットオフしつつ倍率変化でのランキング**すると再現性高く発現変動遺伝子を得られるのでは。」

評価の実際

- 例：Affymetrixの二群間比較（←最もよく研究されている）



- Gene Ontology解析
- （未知サンプルの）分類
- モチーフ解析
- パスウェイ解析

- 感度・特異度

  既知の発現変動遺伝子をどれだけ上位にランキング可能か？

- 再現性

  同じサンプルの比較結果（発現変動遺伝子リスト）が場所間でどれだけ一致しているか？

「感度・特異度」をAUC値で評価

- どの**前処理法**がいい？（比較例：**MAS5** vs. **RMA**）

  既知の発現変動遺伝子をどれだけ上位にランキング可能か？（AUC値の高さ）

MAS5 の遺伝子発現行列

Gene	sample C1	sample C2	sample D1	sample D2
gene 1				
gene 2				
gene 3				
gene 4				
gene 5				

|log比|を計算

$\lvert\log_2(C/D)\rvert$
0.4
3.0
0.2
2.0
0.7

|log比|でランキング

$\lvert\log_2(C/D)\rvert$	Gene
3.0	gene 2
2.0	gene 4
0.7	gene 5
0.4	gene 1
0.2	gene 3

AUC値=100%

RMA の遺伝子発現行列

Gene	sample C1	sample C2	sample D1	sample D2
gene 1				
gene 2				
gene 3				
gene 4				
gene 5				

$\lvert\log_2(C/D)\rvert$
0.8
1.9
0.5
1.3
1.4

$\lvert\log_2(C/D)\rvert$	Gene
1.9	gene 2
1.4	gene 5
1.3	gene 4
0.8	gene 1
0.5	gene 3

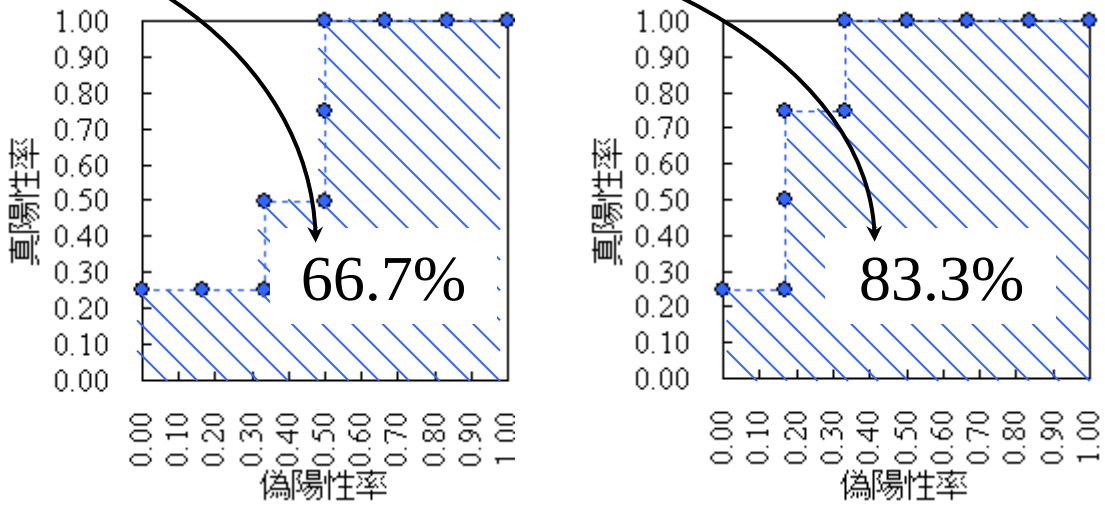
AUC値=83.3%

「感度・特異度」をAUC値で評価

- どのランキング法がいい？（比較例：*t*-検定 vs. 倍率変化）

  既知の発現変動遺伝子をどれだけ上位にランキング可能か？（AUC値の高さ）

ランキング法				
rank	*t*-検定		倍率変化	
1	gene8	真	gene8	真
2	gene5	偽	gene5	偽
3	gene4	偽	gene3	真
4	gene3	真	gene2	真
5	gene7	偽	gene7	偽
6	gene1	真	gene1	真
7	gene2	真	gene4	偽
8	gene9	偽	gene9	偽
9	gene10	偽	gene10	偽
10	gene6	偽	gene6	偽



Area **U**nder the ROC **C**urve (ROC曲線の下部面積：**AUC**)

ROC曲線が左上にあるほどよい方法

ROC曲線の求め方

予測 \ 真実	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

	真実	予測	上位k個を発現変動遺伝子(Positive)とすると					偽陽性率 (1−特異度)	真陽性率 (感度)
Rank	Gene	Gene	k	TP	TN	FP	FN	FP/(FP+TN)	TP/(TP+FN)
1	gene8	gene8	1	1	6	0	3	0.000	0.250
2	gene5	gene5							
3	gene3	gene3							
4	gene2	gene2							
5	gene7	gene7							
6	gene1	gene1							
7	gene4	gene4							
8	gene9	gene9							
9	gene10	gene10							
10	gene6	gene6							

ROC曲線の求め方

予測 \ 真実	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

	真実	予測	上位k個を発現変動遺伝子(Positive)とすると				偽陽性率 (1−特異度)	真陽性率 (感度)
Rank	Gene	Gene	k	TP	TN	FP	FN	FP/(FP+TN)	TP/(TP+FN)
1	gene8	gene8	1	1	6	0	3	0.000	0.250
2	gene5	gene5	2	1	5	1	3	0.167	0.250
3	gene3	gene3	3	2	5	1	2	0.167	0.500
4	gene2	gene2							
5	gene7	gene7							
6	gene1	gene1							
7	gene4	gene4							
8	gene9	gene9							
9	gene10	gene10							
10	gene6	gene6							

ROC曲線の求め方

予測 \ 真実	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

	真実	予測	上位k個を発現変動遺伝子(Positive)とすると					偽陽性率 (1−特異度)	真陽性率 (感度)
Rank	Gene	Gene	k	TP	TN	FP	FN	FP/(FP+TN)	TP/(TP+FN)
1	gene8	gene8	1	1	6	0	3	0.000	0.250
2	gene5	gene5	2	1	5	1	3	0.167	0.250
3	gene3	gene3	3	2	5	1	2	0.167	0.500
4	gene2	gene2	4	3	5	1	1	0.167	0.750
5	gene7	gene7	5	3	4	2	1	0.333	0.750
6	gene1	gene1	6	4	4	2	0	0.333	1.000
7	gene4	gene4	7	4	3	3	0	0.500	1.000
8	gene9	gene9	8	4	2	4	0	0.667	1.000
9	gene10	gene10	9	4	1	5	0	0.833	1.000
10	gene6	gene6	10	4	0	6	0	1.000	1.000

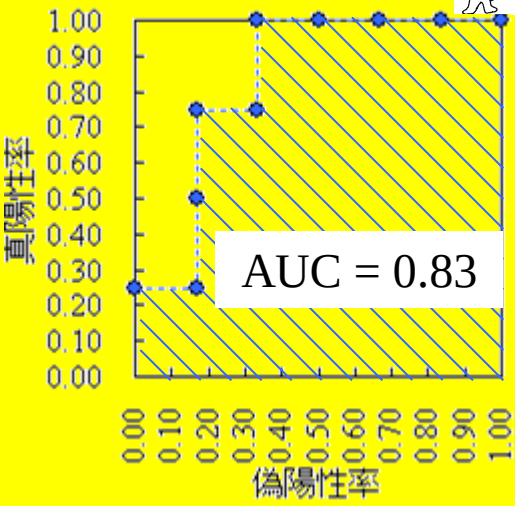
ROC曲線の求め方

予測 \ 真実	Positive	Negative
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

	真実	予測	上位k個を発現変動遺伝子(Positive)とすると				
Rank	Gene	Gene	k	TP	TN	FP	FN
1	gene8	gene8	1	1	6	0	3
2	gene5	gene5	2	1	5	1	3
3	gene3	gene3	3	2	5	1	2
4	gene2	gene2	4	3	5	1	1
5	gene7	gene7	5	3	4	2	1
6	gene1	gene1	6	4	4	2	0
7	gene4	gene4	7	4	3	3	0
8	gene9	gene9	8	4	2	4	0
9	gene10	gene10	9	4	1	5	0
10	gene6	gene6	10	4	0	6	0

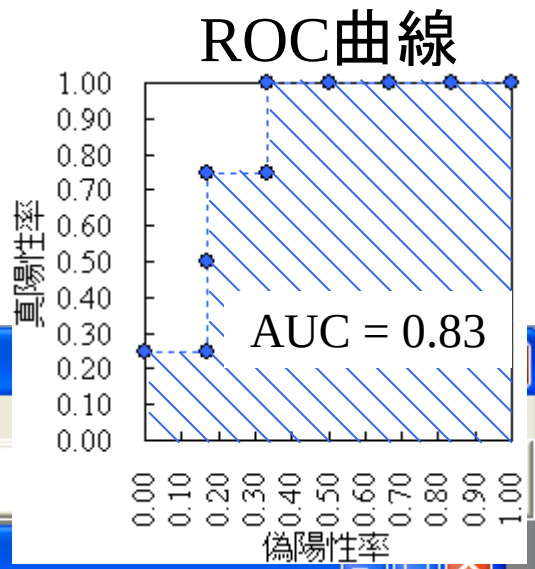
偽陽性率 (1−特異度)	真陽性率 (感度)
FP/(FP+TN)	TP/(TP+FN)
0.000	0.250
0.167	0.250
0.167	0.500
0.167	0.750
0.333	0.750
0.333	1.000
0.500	1.000
0.667	1.000
0.833	1.000
1.000	1.000

ROC曲線



AUC値はRで簡単に計算できます

	真実
Rank	Gene
1	gene8
2	gene5
3	gene3
4	gene2
5	gene7
6	gene1
7	gene4
8	gene9
9	gene10
10	gene6



RGui

ファイル 編集 閲覧 その他 パッケージ ウインドウ ヘルプ Vignettes

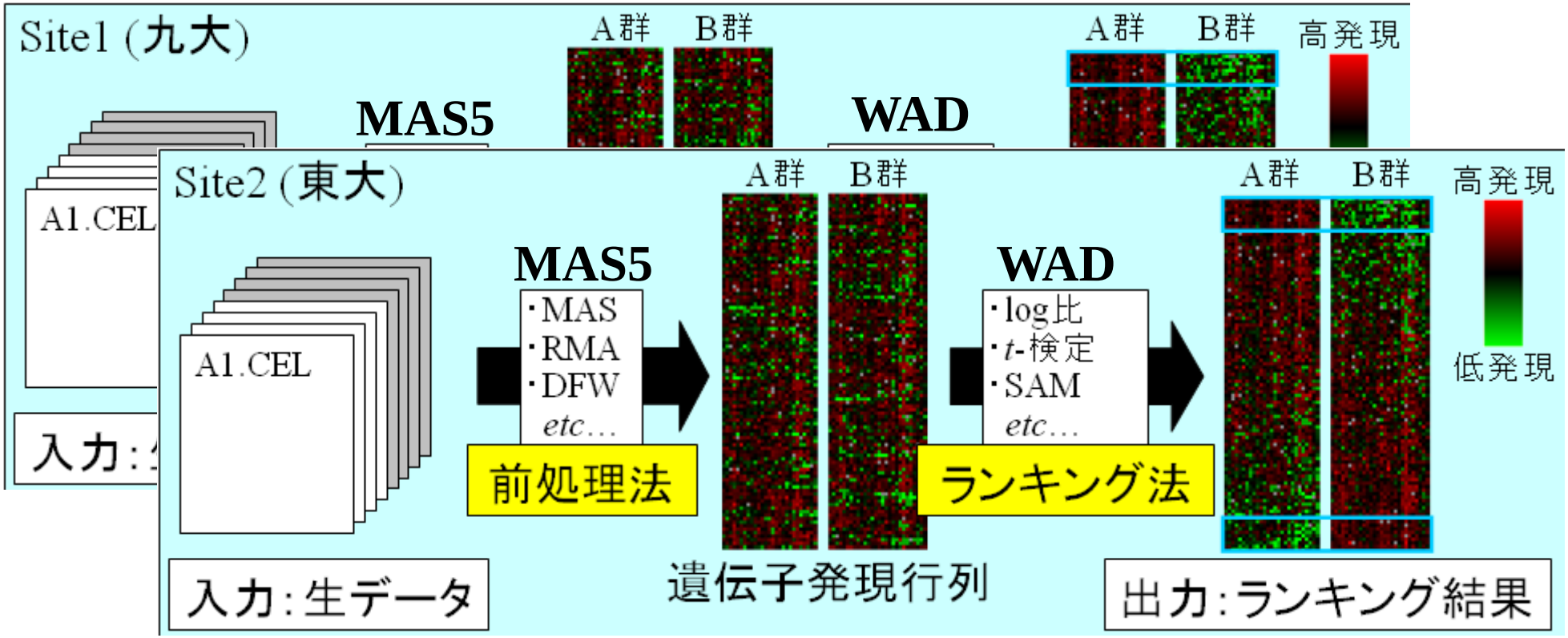
R Console

```
> library(ROC)
> x <- c(1,0,1,1,0,1,0,0,0,0)
> rank_x <- c(1,2,3,4,5,6,7,8,9,10)
> AUC(rocdemo.sca(truth = as.vector(x), data = as.vector(-rank_x), rule = dxrule.sca))
[1] 0.8333333
>
```

R version 2.6.2 (2008-02-08)

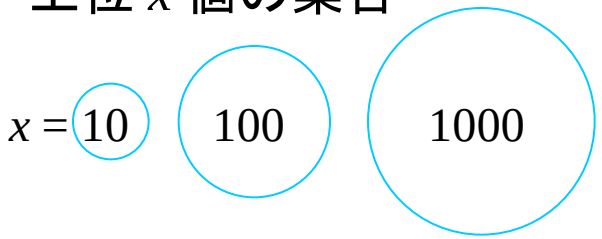
「再現性」を一致度で評価

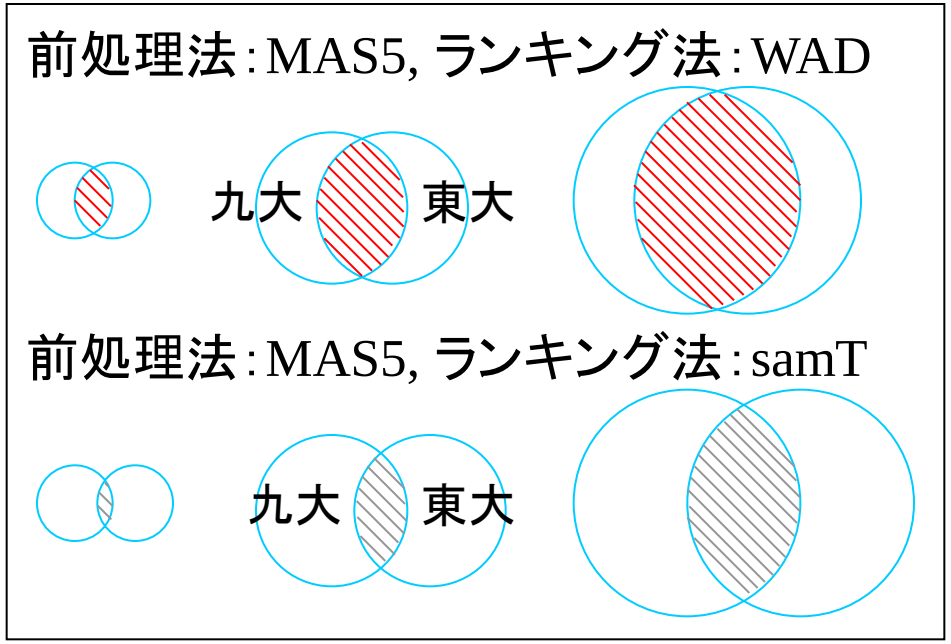
- MicroArray Quality Control (MAQC) プロジェクトで提唱（0≦POG≦100%）
- POG値が高い → ランキング結果の頑健性（再現性）が高い方法

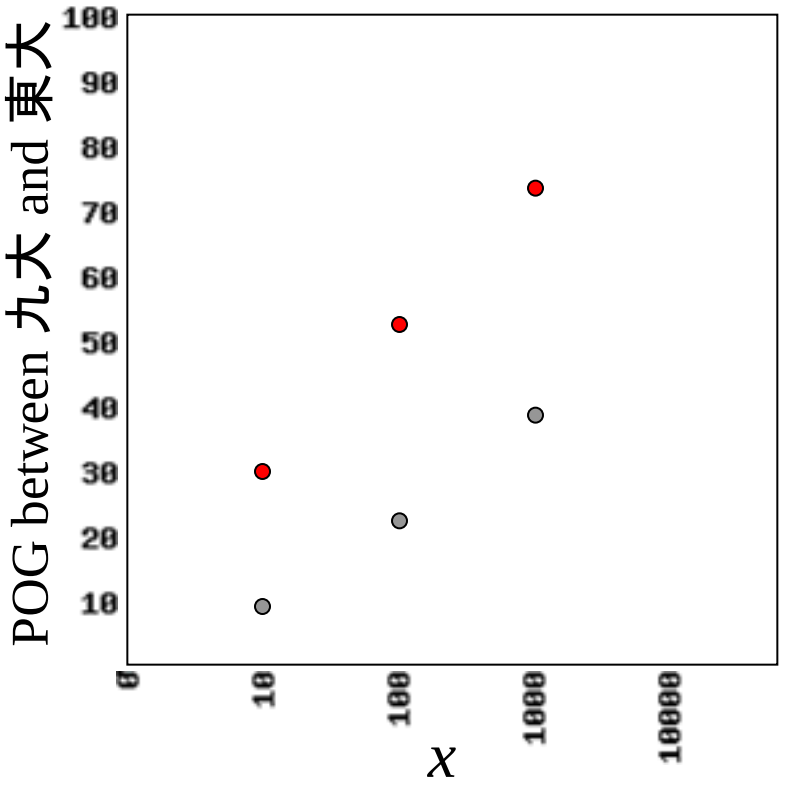


「再現性」を一致度で評価

- MicroArray Quality Control (MAQC) プロジェクトで提唱（0≦POG≦100%）
- POG値が高い → ランキング結果の頑健性（再現性）が高い方法
- 上位 *x* 個の集合



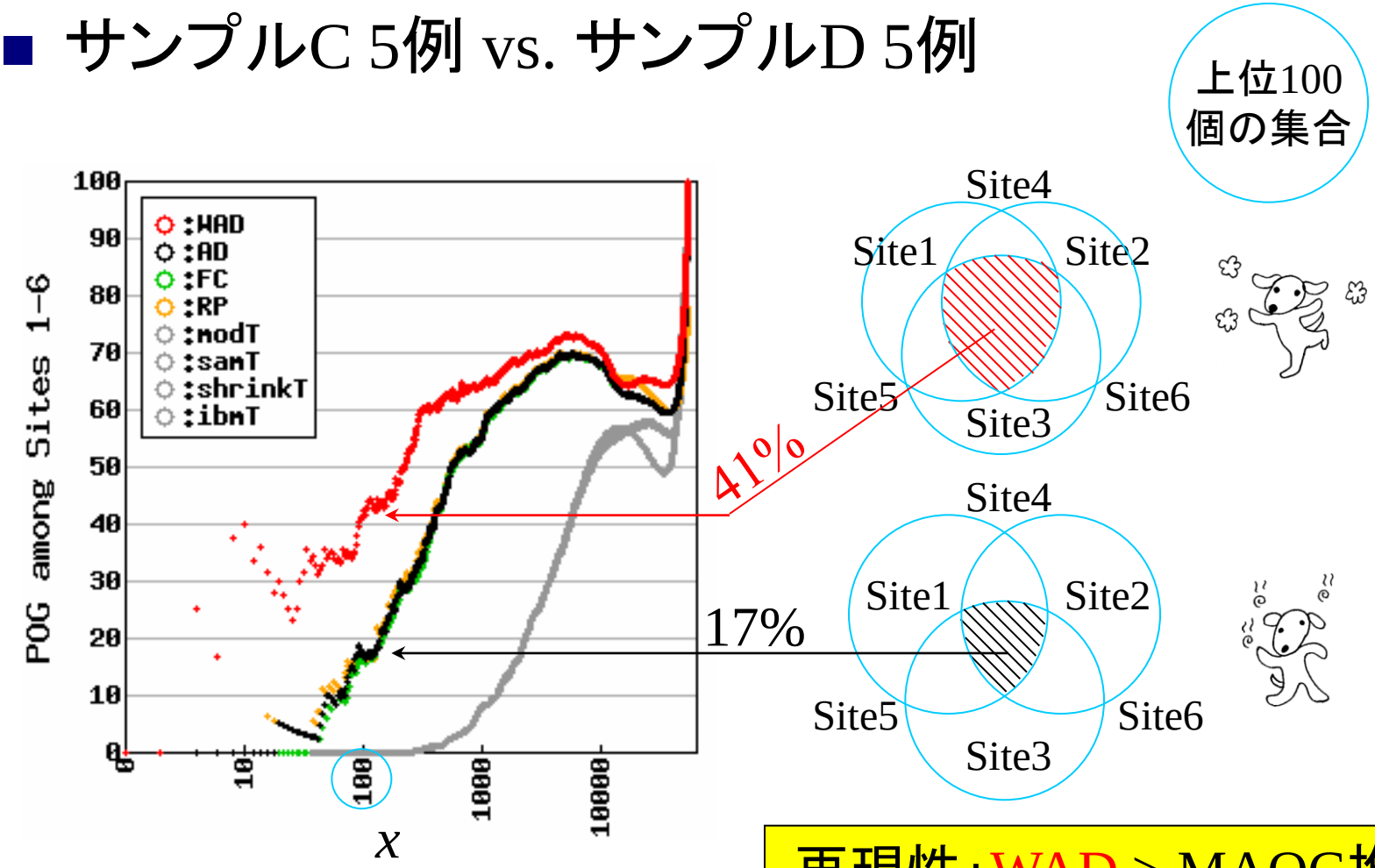




再現性：WAD > samT

「再現性」解析結果（前処理法：FARMS）

- サンプルC 5例 vs. サンプルD 5例



再現性：WAD > MAQC推奨法（AD）

結論（Affymetrixデータ; 二群間比較）

- 「感度・特異度」が高い方法（**組合せが重要**である！）

前処理法	MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW
ランキング法	WAD	WAD	RP	RP	RP	RP	RP	RP	RP

WAD: Weighted Average Difference
RP: Rank Products

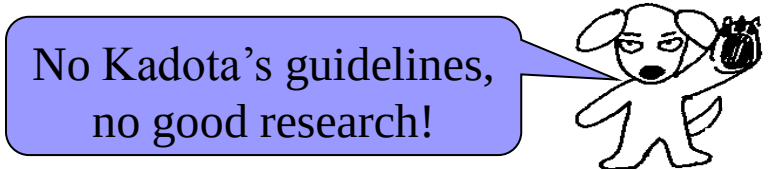
Fold Changeに基づく方法

↕

従来：*t*-統計量に基づく方法

- （発現変動遺伝子リストの）「再現性」が高い方法
  - （前処理法によらず）WAD ↔ 従来：Average Difference (AD)法

MAQC Consortium, *Nat. Biotechnol.*, **24**:1151-1161, 2006



推奨ガイドラインの比較

- 「感度・特異度」の高いランキング法
  - *t*-検定系の方法（*P*値） ⟷ FC系の方法（WAD or RP）
- 「再現性」の高いランキング法
  - Fold Change（FC）系の方法（AD法） ⟷ FC系の方法（WAD）

MAQC
- MAQC Consortium, *Nat. Biotechnol.*, 2006
- Shi *et al.*, *BMC Bioinformatics.*, 2008

門田ら
- Kadota *et al.*, *AMB.*, 2008
- Kadota *et al.*, *AMB.*, 2009

「感度・特異度」の高いランキング法

- *t*-検定系の方法（MAQC推奨）⇔ FC系の方法（門田推奨）

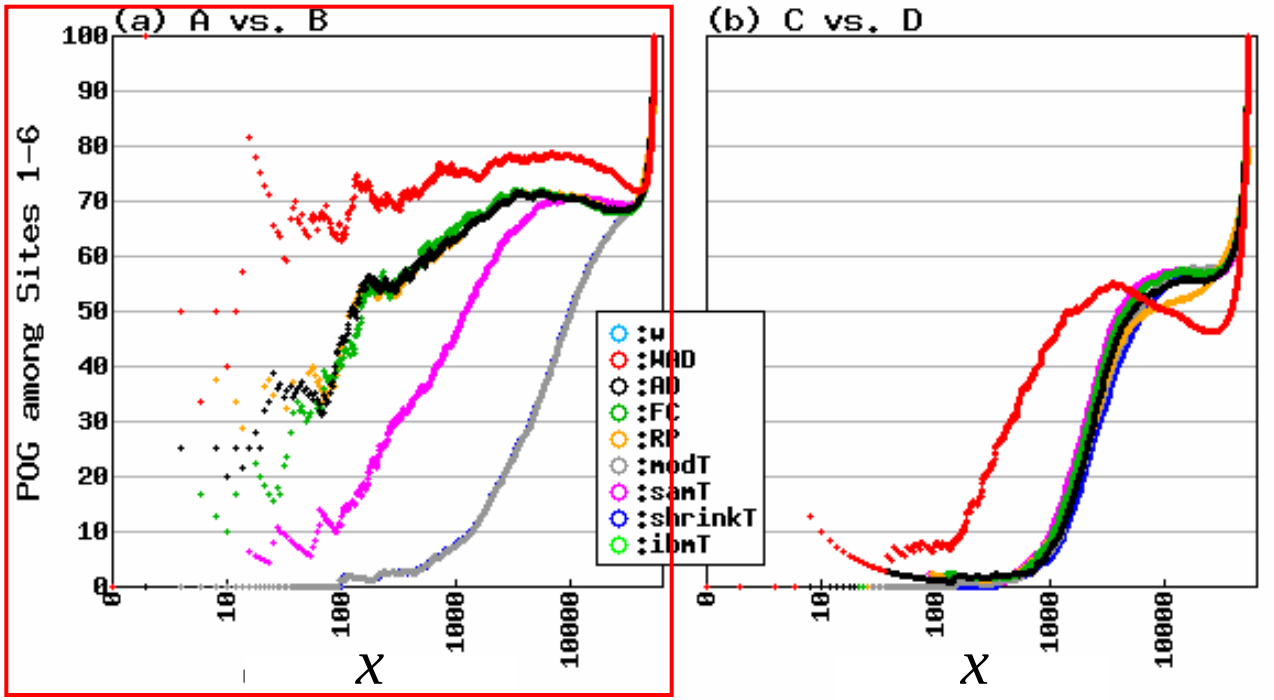
前処理法

ランキング法		MAS5	multi-mgMOS	RMA	VSN	GCRMA	MBEI	PLIER	FARMS	DFW	
	Datasets 3-26: **MAS5**, FC系が4, *t* 検定系が8, 両方同時が12										
	WAD	**96.74**	**94.84**	91.37	90.97	92.67	88.19	89.15	91.58	91.41	Fold Change (FC)系
	AD	93.76	93.13	93.10	**92.96**	**93.42**	89.14	87.32	92.47	92.24	
	FC	93.63	92.82	**93.12**	92.92	93.16	89.71	86.20	92.49	92.24	
	RP	91.51	92.20	92.54	92.48	93.23	**90.07**	**92.01**	**93.06**	**92.53**	
	modT	95.67	91.91	91.38	90.70	91.19	86.79	85.43	89.23	90.11	*t*検定系
	samT	95.95	92.99	91.23	90.94	91.07	87.09	85.25	89.40	89.96	
	shrinkT	95.73	92.56	91.32	90.62	92.12	86.89	84.65	-	91.45	
	ibmT	96.34	93.11	91.77	91.02	91.06	87.10	86.27	90.04	90.25	
	Datasets 27-38: **RMA**, FC系が2, *t* 検定系が7, 両方同時が3										
	WAD	**92.42**	**92.36**	96.73	95.41	95.75	93.39	91.30	93.55	94.09	Fold Change (FC)系
	AD	87.41	86.99	**96.77**	96.22	96.18	93.11	89.01	93.81	94.22	
	FC	88.23	85.92	96.73	96.20	96.06	92.94	88.82	93.81	94.22	
	RP	84.55	86.94	96.53	**96.53**	**96.25**	**93.53**	**91.57**	**94.76**	**94.67**	
	modT	90.90	89.61	95.28	94.53	93.33	90.50	89.28	91.62	92.36	*t*検定系
	samT	90.31									
	shrinkT	90.97									
	ibmT	91.92	90.60	95.49	94.80	93.67	90.66	89.89	89.95	92.43	

赤枠の中だけで評価すると*t*-検定系がよい

「再現性」の高いランキング法は“FC系”で一致

- AD（MAQC推奨）⇔ WAD（門田推奨）



MAQCの解析は:
- 用いた前処理法がPLIERのみ
- 比較したランキング法がAD, samT, …のみ
- C vs. Dの比較結果にsamTが含まれてない

門田らの解析は:
- 用いた前処理法は9種類
- 比較したランキング法は8種類

その他のメーカーではどの方法がいい？

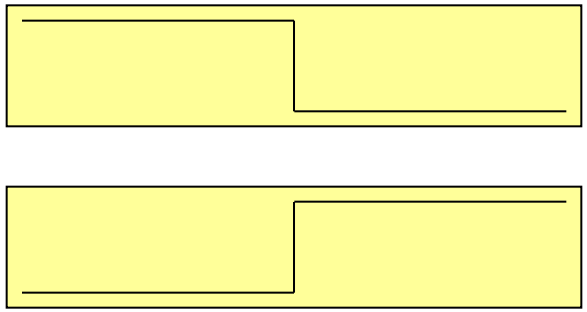
- そもそも前処理法はAffymetrix以外はほとんど開発されていない

  →メーカーのデフォルト（or 推奨）の前処理法をやる以外にない
- ではランキング法はどれがいい？
  - 一色法の場合：（手前味噌ながら）WAD
  - 二色法の場合：わかりません
  - WADの根拠は？
    - （おそらく）Affymetrix以外のメーカーはチップごとの正規化法しかない。
    - Affymetrixのチップごとの正規化法はMAS5だけで、MAS5と最も相性がよかったのはWADだから…。

遺伝子発現行列

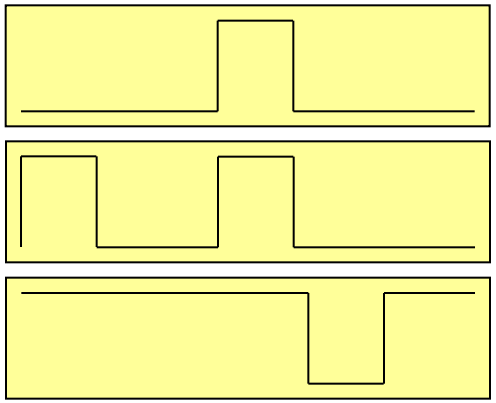
二群間比較

	A群 A1	A群 A2	…	B群 B1	B群 B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



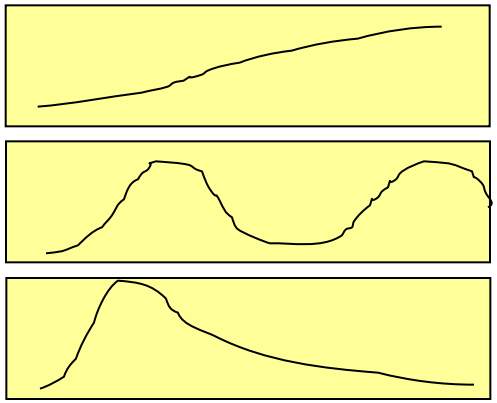
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



組織特異的遺伝子検出法

- ランキングに基づく方法
  - Dixon test（Greller and Tobin, *Genome Res.*, **9**, 282-296, 1999）
  - Pattern matching（Pavlidis and Noble, *Genome Biol.*, **2**, research0042, 2001）
  - Entropy（Schug *et al.*, *Genome Biol.*, **6**, R33, 2005）
  - Tissue specificity Index（Yanai *et al.*, *Bioinformatics*, **21**, 650-659, 2005）
- 外れ値検出に基づく方法
  - Akaike's Information Criterion (AIC)（Kadota *et al.*, *Physiol. Genomics*, **12**, 251-259, 2003）
  - Sprent's non-parametric method（Ge *et al.*, *Genomics*, **86**, 127-141, 2005）
- その他
  - Tukey-Kramer's Honest Significance Difference (HSD) test（Liang *et al.*, *Physiol. Genomics*, **26**, 158-162, 2006）
  - ROKU (Kadota *et al.*, *BMC Bioinformatics*, **7**, 294, 2006)

組織特異的遺伝子検出法

	方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
①	Dixon test	×	×	?	○	-
	Pattern matching	○	○	×	○	-
	Tissue specificity index	○	×	-	○	-
②	Entropy (H)	○	×	×	○	-
	Tukey-Kramer's HSD test	○	×	?	○	-
④	AIC	○	○	○	×	○
	Sprent's method	○	○	○	×	×
③	ROKU (AIC+a modified H)	○	○	○	○	○

結論：おすすめはROKU

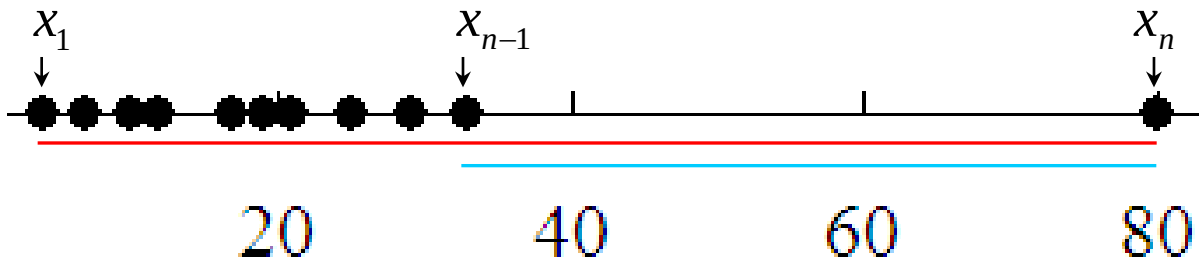
Dixon WJ, *Biometrics*, 1953

組織特異的遺伝子検出法

- ①Dixon test（$0 \leqq D \leqq 1$）
  - 一組織のみで高発現（低発現）しているパターンを検出

x	
組織	発現量
肺	4
骨	7
脳	10
皮膚	17
延髄	19
心臓	21
胃	25
小腸	29
膵臓	33
肝臓	80

一般化	
組織	発現量
Tissue 1	x_1
Tissue 2	x_2
…	…
Tissue i	x_i
…	…
Tissue n-1	x_{n-1}
Tissue n	x_n



高発現の場合：$D(x) = \dfrac{x_n - x_{n-1}}{x_n - x_1} = \dfrac{80-33}{80-4} = 0.618$

（低発現の場合：$D(x) = \dfrac{x_2 - x_1}{x_n - x_1}$）

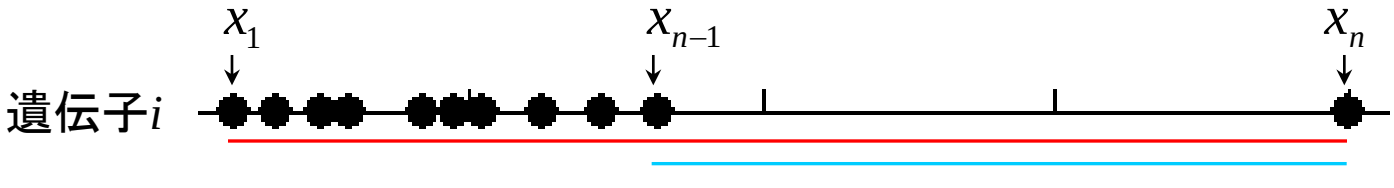
統計量Dの大きい遺伝子を抽出

Dixon WJ, *Biometrics*, 1953

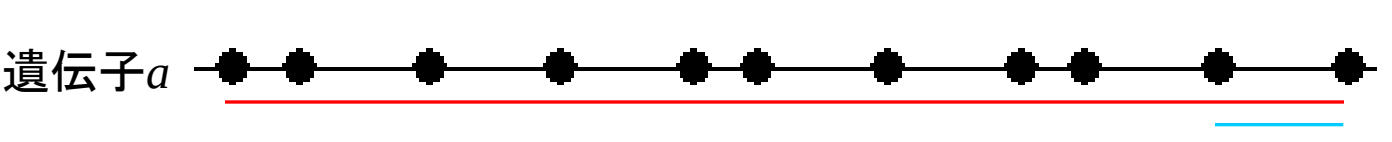
組織特異的遺伝子検出法

①Dixon testの欠点（$0 \leqq D \leqq 1$）

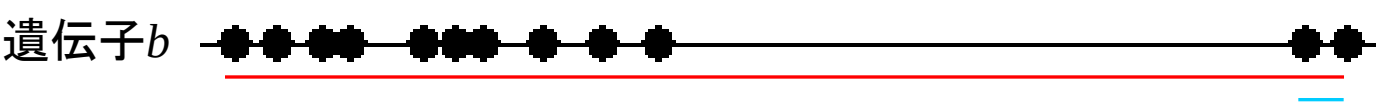
- 複数の外れ値が互いに外れ値をかばいあう効果（**マスク効果**）の影響を受ける



$$D(x^i) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.6$$



$$D(x^a) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.1$$



$$D(x^b) = \frac{x_n - x_{n-1}}{x_n - x_1} \approx 0.03$$

Dixon統計量によるランキングでは複数外れ値に対応不可

組織特異的遺伝子

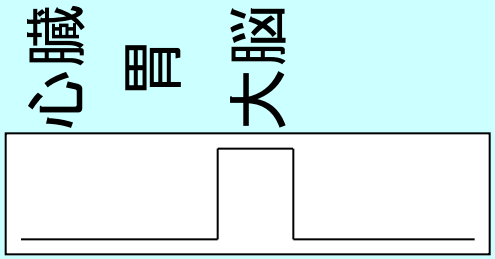
- やりたいこと1

方法	複数外れ値への対応	様々な特異的発現パターンへの対応	目的組織特異性	ランキング	頑健性
① Dixon test	×	×	?	○	-
Pattern matching	○	○	×	○	-
Tissue specificity index	○	×	-	○	-
② Entropy (H)	○	×	×	○	-
Tukey-Kramer's HSD test	○	×	?	○	-
④ AIC	○	○	○	×	○
Sprent's method	○	○	○	×	×
③ ROKU (AIC+a modified H)	○	○	○	○	○

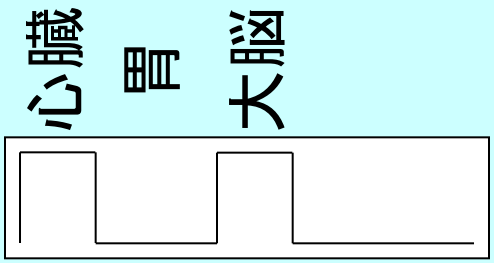
	心臓	肝	大脳	
gene 1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$ …
gene 2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$ …
…	…	…	…	… …
gene i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
…	…	…	…	… …
gene p	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$ …

入力：遺伝子発現行列

「大脳」特異的高発現遺伝子



「心臓と大脳」特異的高発現遺伝子



出力：任意の組織特異的遺伝子

様々な特異的発現パターンを組織特異性の度合いで統一的にランキングしたい

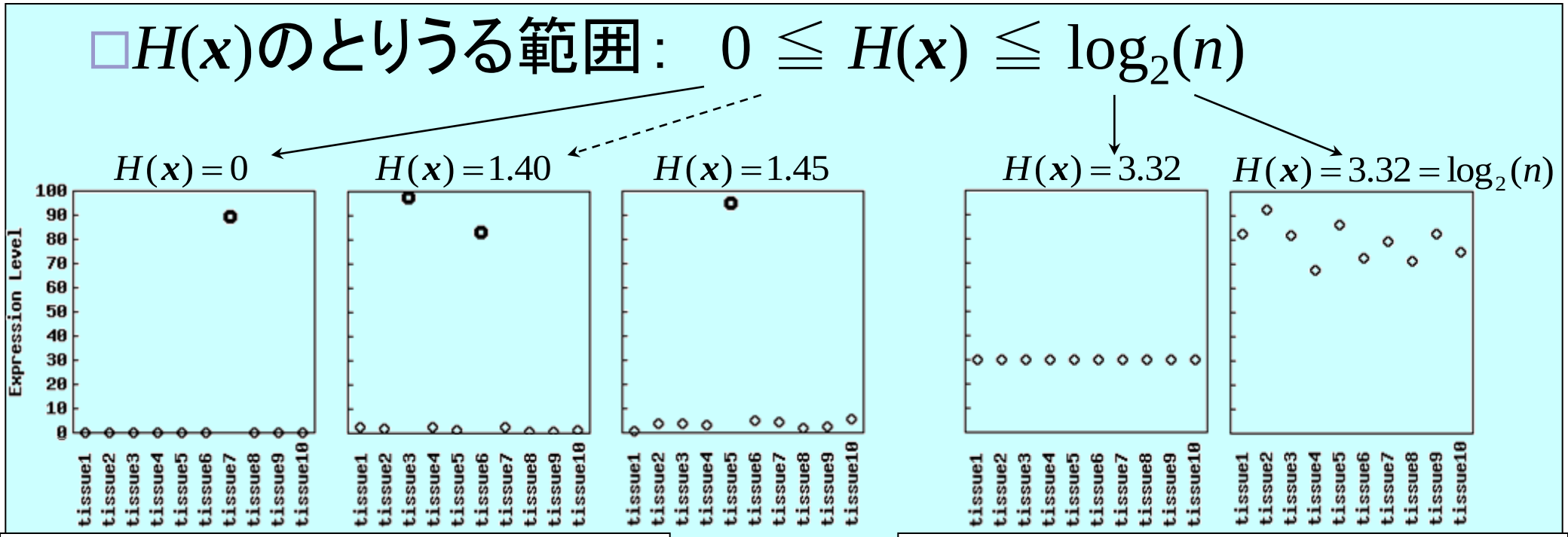
組織特異的遺伝子検出法

②エントロピーによるランキング

- 遺伝子$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$のエントロピー $H(\boldsymbol{x})$

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textbf{where } p_i = x_i / \sum x_i$$

- $H(\boldsymbol{x})$のとりうる範囲: $0 \leqq H(\boldsymbol{x}) \leqq \log_2(n)$



エントロピーが低い → 組織特異性が高い

エントロピーが高い → 組織特異性が低い

エントロピーでランキングすることにより複数外れ値に対応可能

②エントロピー計算例

- 遺伝子iのエントロピー $H(x_i)$

$$H(x_i) = -\sum_{j=1}^{N} p_{ij} \log_2(p_{ij}) \qquad p_{ij} = x_{ij} / \sum_{j=1}^{N} x_{ij}$$

$$0 \leqq H \leqq \log_2 N$$

gene	Tissue_1	…	Tissue_j	…	Tissue_N	$H(x)$
gene_1	x_{11}		x_{1j}		x_{1N}	$H(x_1)$
…						
gene_i	x_{i1}		x_{ij}		x_{iN}	$H(x_i)$
…						
gene_m	x_{m1}		x_{mj}		x_{mN}	$H(x_m)$

$0 \leqq H \leqq 2.32$

	x_i	p_i	$-p_{ij}\log_2(p_{ij})$
組織1→	$x_{i1} = 0.1$	$p_{i1} = 0.01$	$-p_{i1}\log_2(p_{i1}) = 0.06$
組織2→	$x_{i2} = 10.1$	$p_{i2} = 0.96$	$-p_{i2}\log_2(p_{i2}) = 0.05$
組織3→	$x_{i3} = 0.1$	$p_{i3} = 0.01$	$-p_{i3}\log_2(p_{i3}) = 0.06$
組織4→	$x_{i4} = 0.1$	$p_{i4} = 0.01$	$-p_{i4}\log_2(p_{i4}) = 0.06$
組織5→	$x_{i5} = 0.1$	$p_{i5} = 0.01$	$-p_{i5}\log_2(p_{i5}) = 0.06$
	sum = 10.5	sum = 1	$H(x_i) = 0.31$

特異的発現パターン

→低いエントロピー

	x_i	p_i	$-p_{ij}\log_2(p_{ij})$
組織1→	x_{i1} 40.1	$p_{i1} = 0.17$	$-p_{i1}\log_2(p_{i1}) = 0.44$
組織2→	x_{i2} 35.2	$p_{i2} = 0.15$	$-p_{i2}\log_2(p_{i2}) = 0.41$
組織3→	x_{i3} 60.8	$p_{i3} = 0.26$	$-p_{i3}\log_2(p_{i3}) = 0.50$
組織4→	x_{i4} 50.4	$p_{i4} = 0.21$	$-p_{i4}\log_2(p_{i4}) = 0.48$
組織5→	x_{i5} 48.7	$p_{i5} = 0.21$	$-p_{i5}\log_2(p_{i5}) = 0.47$
	sum = 235	sum = 1	$H(x_i) = 2.30$

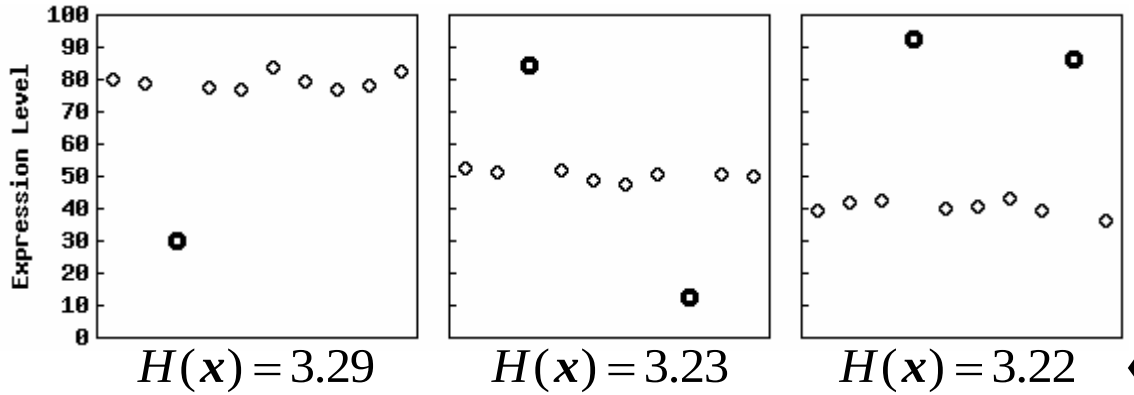
そうでないパターン

→高いエントロピー

組織特異的遺伝子検出法

②エントロピーの短所

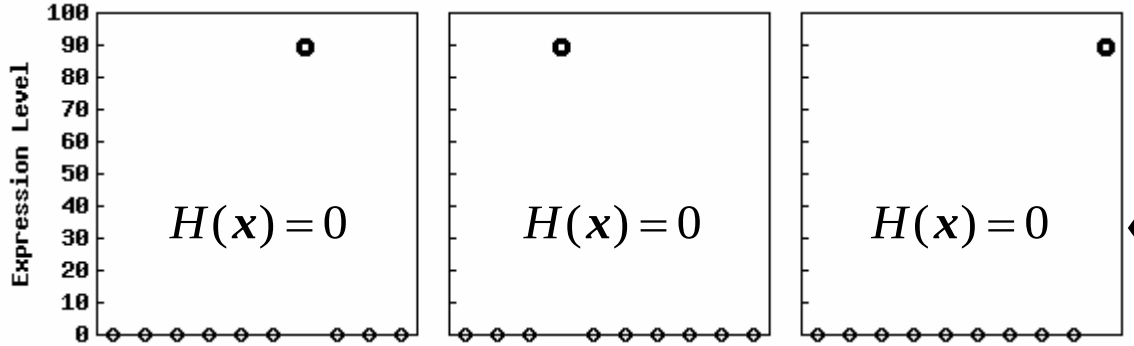
1. 組織特異的低発現パターンなどの検出が不可能



$$0 \leqq H(x) \leqq \underbrace{\log_2(n)}_{\| \\ 3.32}$$

上位にランキングされない

2. 特異的組織の同定が不可能

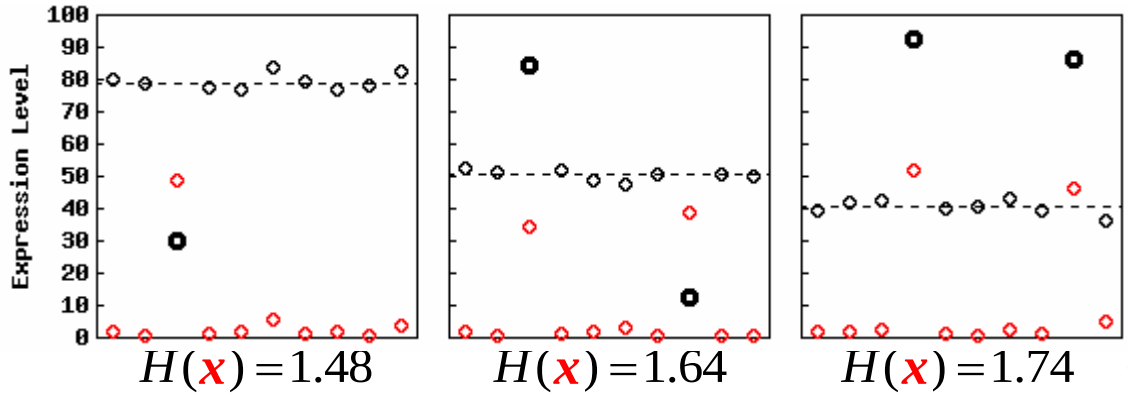


どの組織で特異的なのか分からない

組織特異的遺伝子検出法

③ROKU

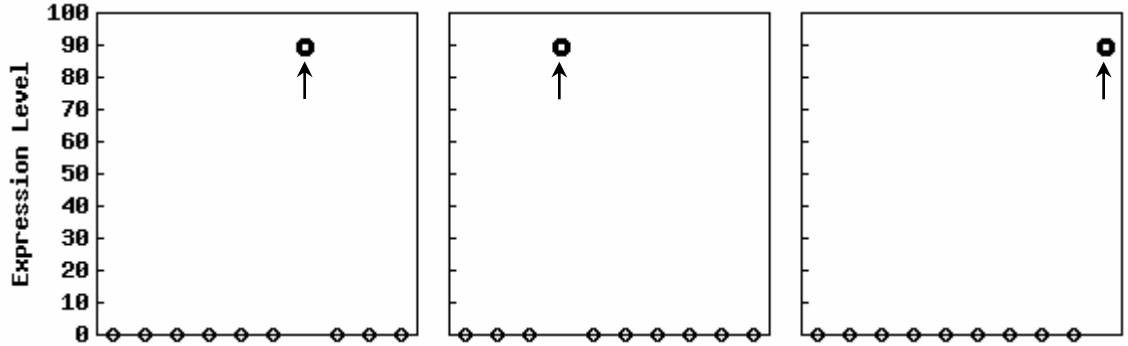
1. 遺伝子発現ベクトルxを変換: $x \rightarrow \boldsymbol{x}$ by $x_i = |x_i - T_{bw}|$



$$0 \leqq H(\boldsymbol{x}) \leqq \underbrace{\log_2(n)}_{\| \\ 3.32}$$

上位にランキングされる

2. AICに基づく外れ値検出法を採用



どの組織で特異的なのか分かる

上田太一郎, *応用統計学*, 1996
Kadota *et al.*, *Physiol. Genomics*, 2003

組織特異的遺伝子検出法

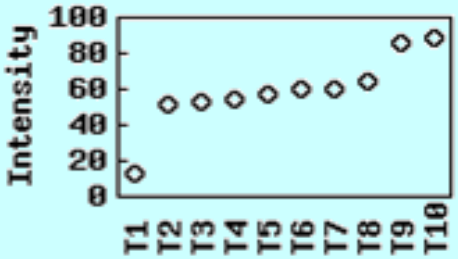
④AICに基づく外れ値検出法

- Akaike's Information Criterion (AIC)
- 様々な外れ値の組み合わせモデルから*AIC*が最小の組み合わせ(MAICE)を探索

$$AIC = n_n \log \hat{\sigma} + \sqrt{2} \times n_o \times \frac{\log n_n!}{n_n}$$

$n_n + n_o (= n)$: サンプル数
n_o : *Outlier*（外れ値）の数
n_n : *Non-outlier* の数
$\hat{\sigma}$: 標準偏差

計算例:



入力

組織	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
発現量	12	51	52	54	57	59	60	63	85	88

↓

出力

出力結果	-1	0	0	0	0	0	0	0	1	1

低発現側の外れ値 : -1, 高発現の～ : 1, それ以外 : 0

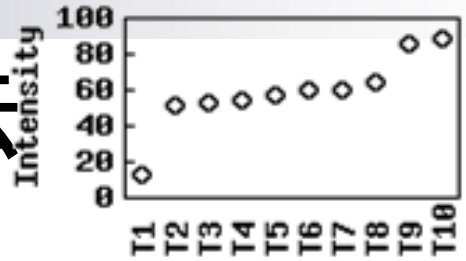
組織特異的遺伝子検出法

④AICに基づく外れ値検出法

- 様々な外れ値の組み合わせモデルから*AIC*が最小の組み合わせ(MAICE)を探索
- 様々な外れ値の組み合わせモデル最大探索範囲*Nmax* = *n*/2 = 5

$$AIC = n_n \log \hat{\sigma} + \sqrt{2} \times n_o \times \frac{\log n_n!}{n_n}$$

$n_n + n_o (= n)$: サンプル数
n_o : *Outlier*（外れ値）の数
n_n : *Non-outlier* の数
$\hat{\sigma}$: 標準偏差



(i)Mean-SD scaling

(ii)Calculate AIC

Nmax = 2　　Nmax = 5

outliers(high)

outliers(low)	none	T10	T9-10	T8-10	T7-10	T6-10
none	-0.53	0.68	1.27	3.14	4.67	5.67
T1	-2.22	-1.97	-6.19	-4.66	-2.91	
T1-2	-0.01	0.19	-4.18	-2.91		
T1-3	1.82	1.94	-2.91			
T1-4	3.27	3.24				
T1-5	4.31					

(iii)Detect outliers

Expression Data

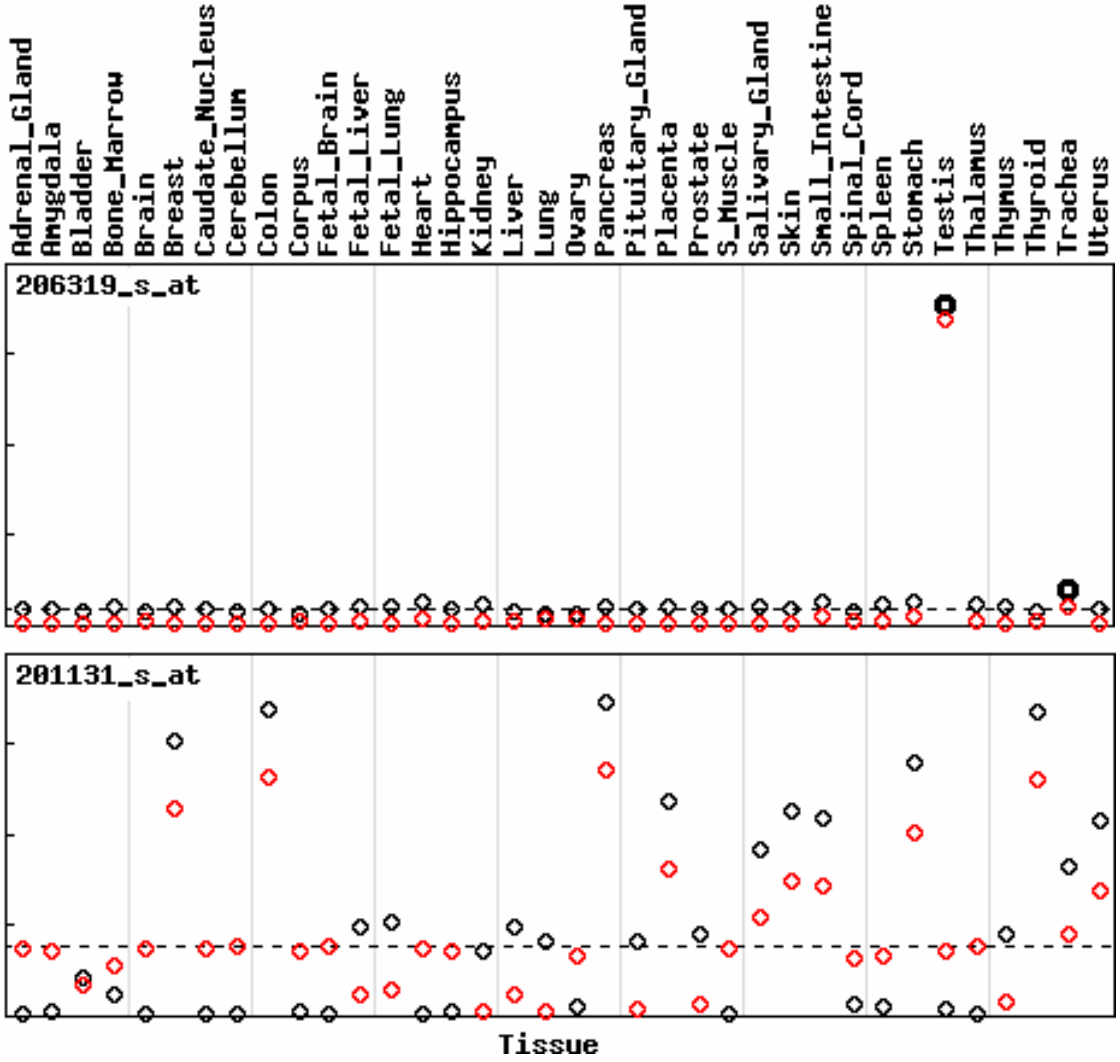
	12	51	52	54	57	59	60	63	85	88
5	-1	0	0	0	0	0	0	0	1	1

1:High-side outlier
0:Non-outlier
-1:Low-side outlier

実データで比較

22,263遺伝子×36組織のデータ(Ge *et al.*, *Genomics*, 2005)

- 全体的な組織特異性の度合いで正しくランキングできるのは？



② Schug et al., *Genome Biol.*, 2005	③ Kadota et al., *BMC Bioinfo.*, 2006
$H(\boldsymbol{x}) = 4.235$	$H(\boldsymbol{x}) = 1.950$
$H(\boldsymbol{x}) = 4.228$	$H(\boldsymbol{x}) = 4.729$

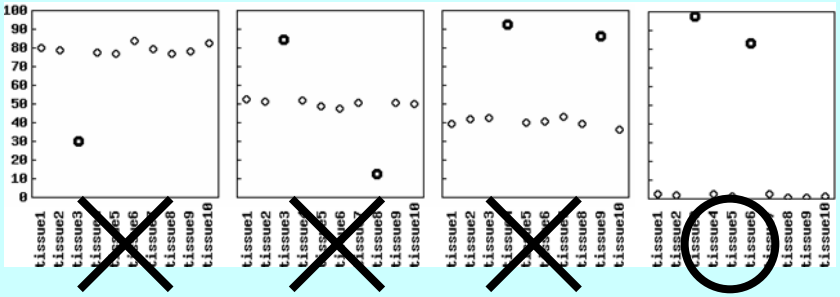
③のほうが正しくランキング可能

目的組織特異性が高いのは？

② Schug et al., *Genome Biology*, 2005

1) 遺伝子 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$の全体的な組織特異性度合いを表す統計量

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textbf{where } p_i = x_i / \sum x_i$$



2) 組織tにおける特異性度合いを表す統計量

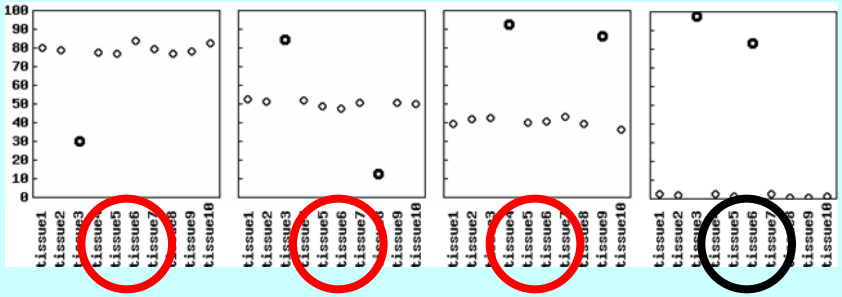
$$Q_t(\boldsymbol{x}) = H(\boldsymbol{x}) - \log_2(p_t)$$

全遺伝子について統計量を計算し、最低の統計量をもつものが最もt組織特異的高発現遺伝子

③ Kadota et al., *BMC Bioinformatics*, 2006

1)遺伝子$\boldsymbol{x}$を変換($x_i = |x_i - T_{bw}|$)し、変換後のベクトル$\boldsymbol{x}$のエントロピーを利用

$$H(\boldsymbol{x}) = -\sum_{i=1}^{n} p_i \log_2(p_i), \textbf{where } p_i = x_i / \sum x_i$$



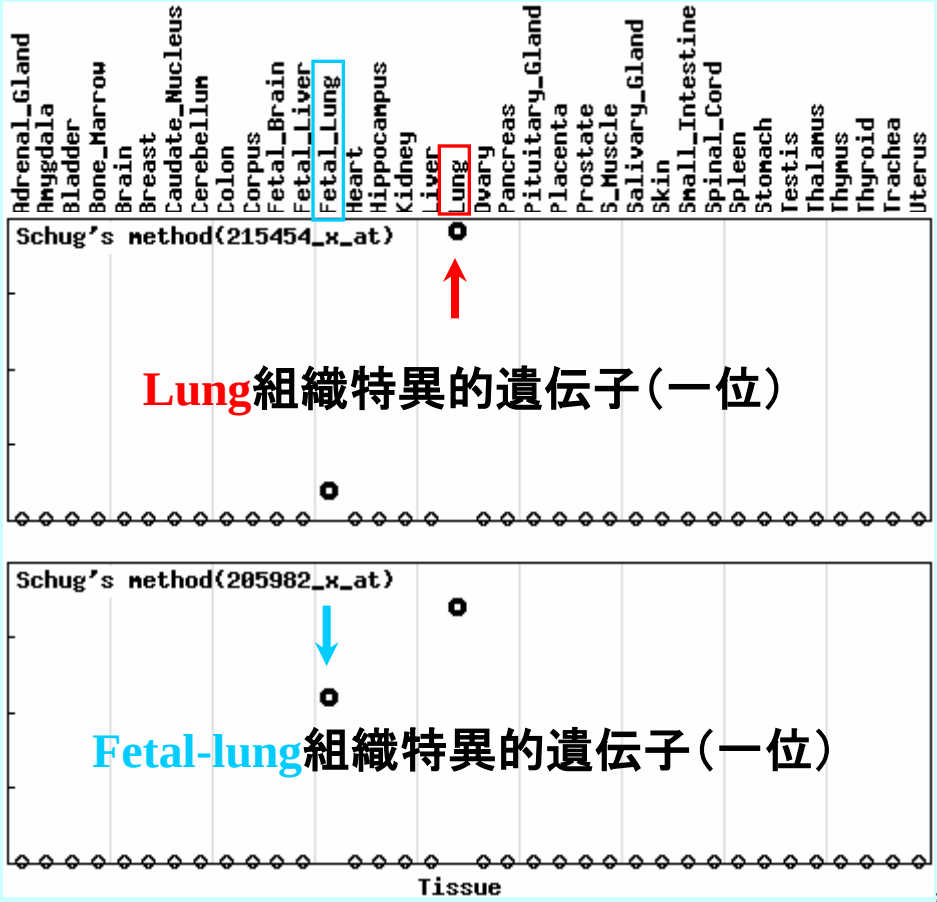
2) AICに基づく外れ値検出法の適用

入力	2.1	1.3	9.0	1.9	1.1	2.0	0.8	0.5	1.2
				t					
出力	0	0	1	0	0	0	0	0	0

組織tのみで1、それ以外で0の遺伝子群を抽出。その中で最低の$H(\boldsymbol{x})$をもつものが最もt組織特異的高発現遺伝子

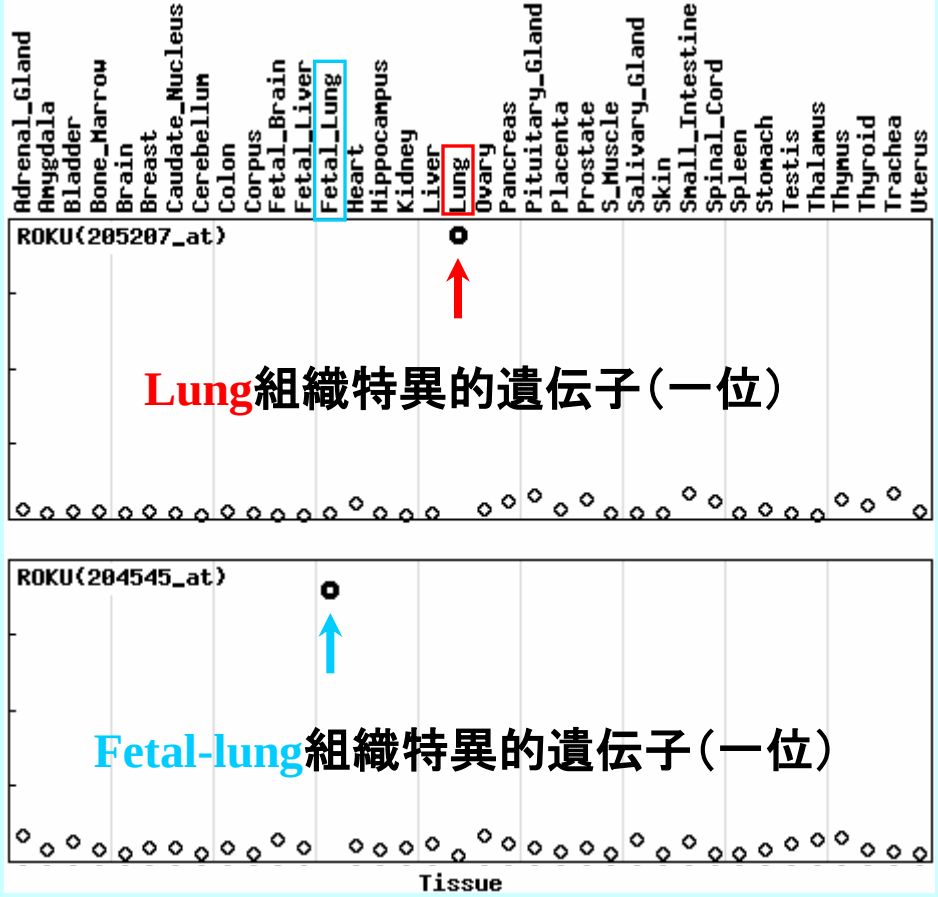
目的組織特異性が高いのは？

② Schug et al., *Genome Biology*, 2005



目的組織**以外でも**特異的：×

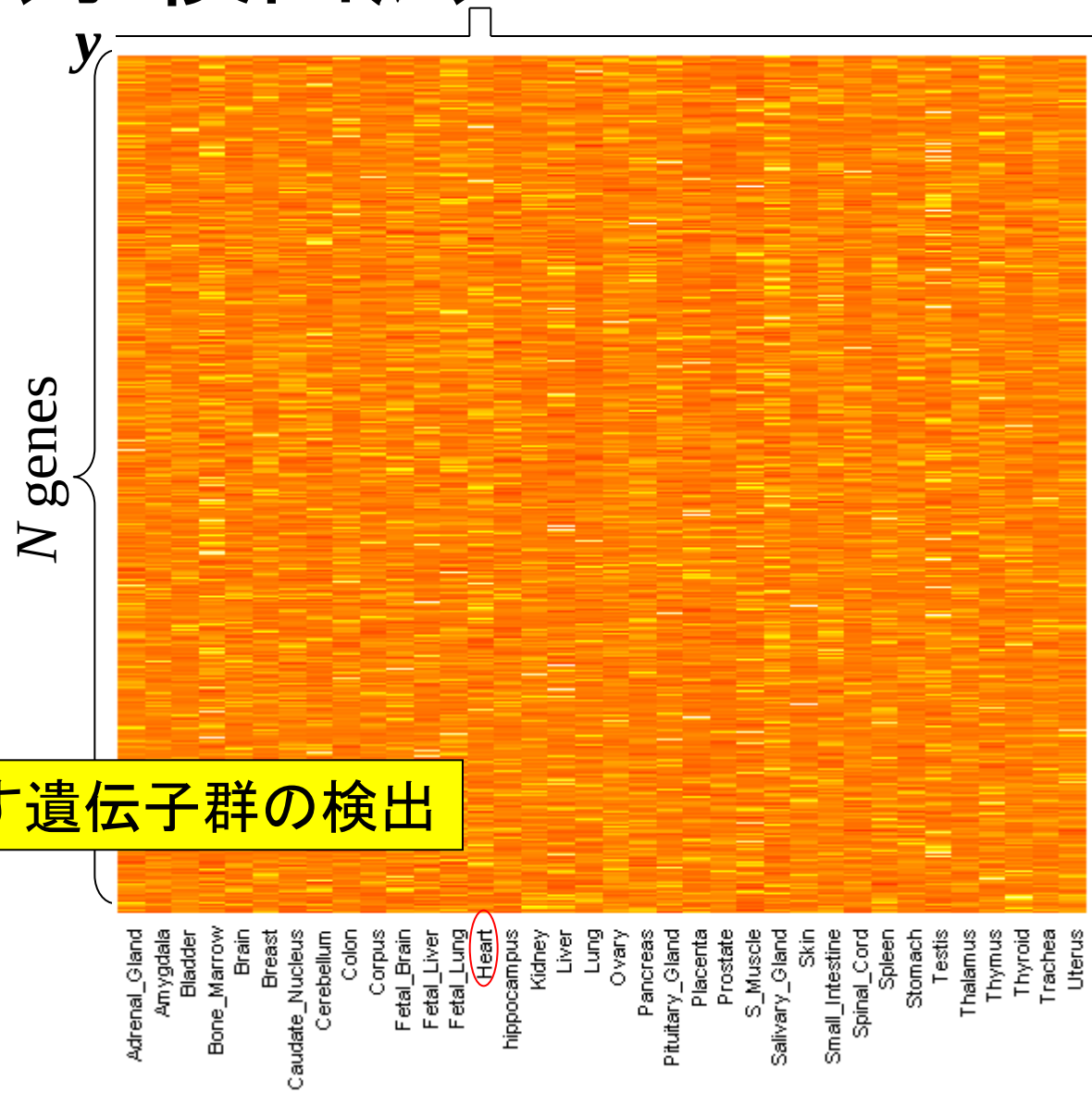
③ Kadota et al., *BMC Bioinformatics*, 2006



目的組織**のみで**特異的：○

組織特異的遺伝子検出法

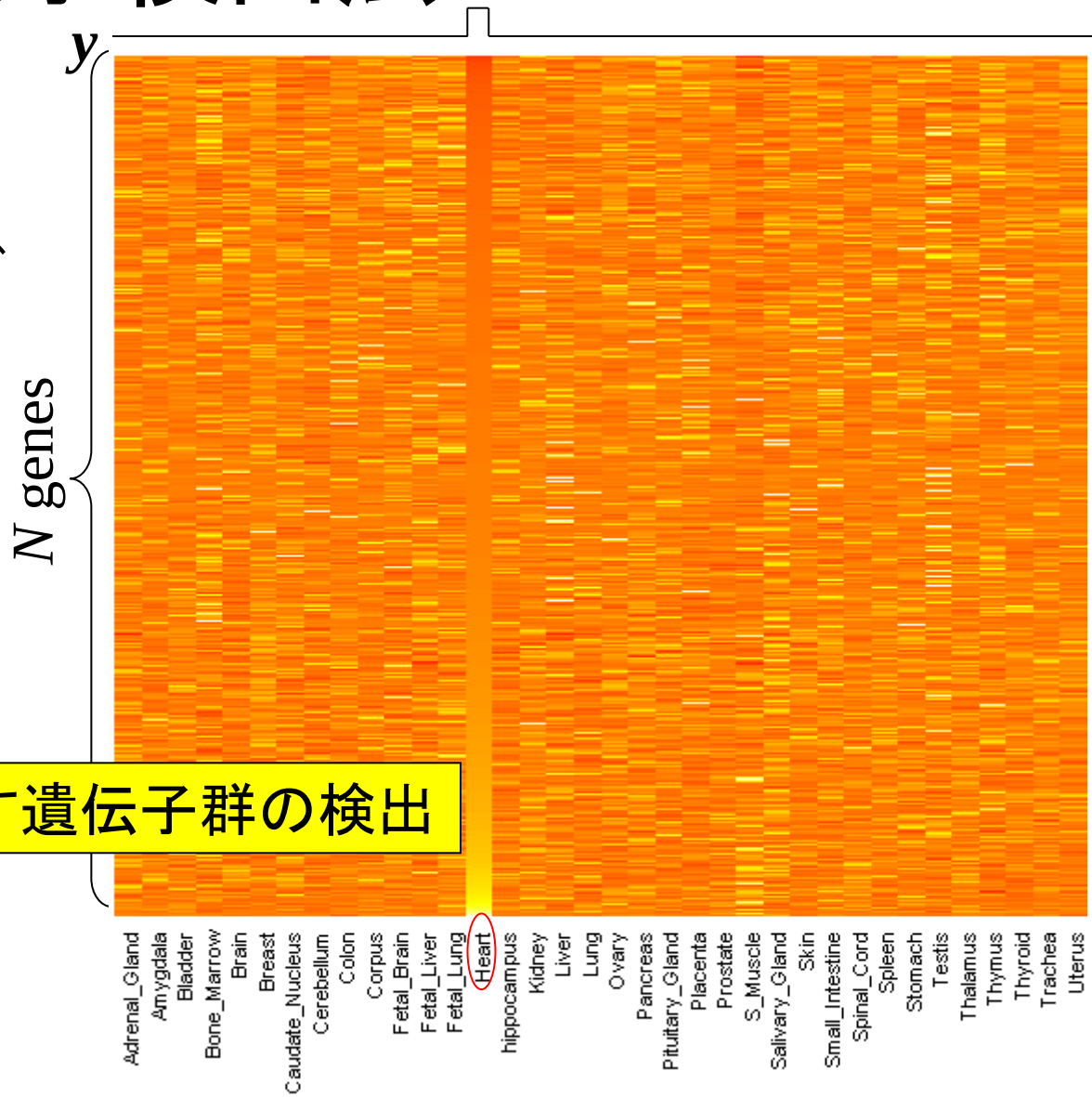
- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順にランキング



例：心臓特異的パターンを示す遺伝子群の検出

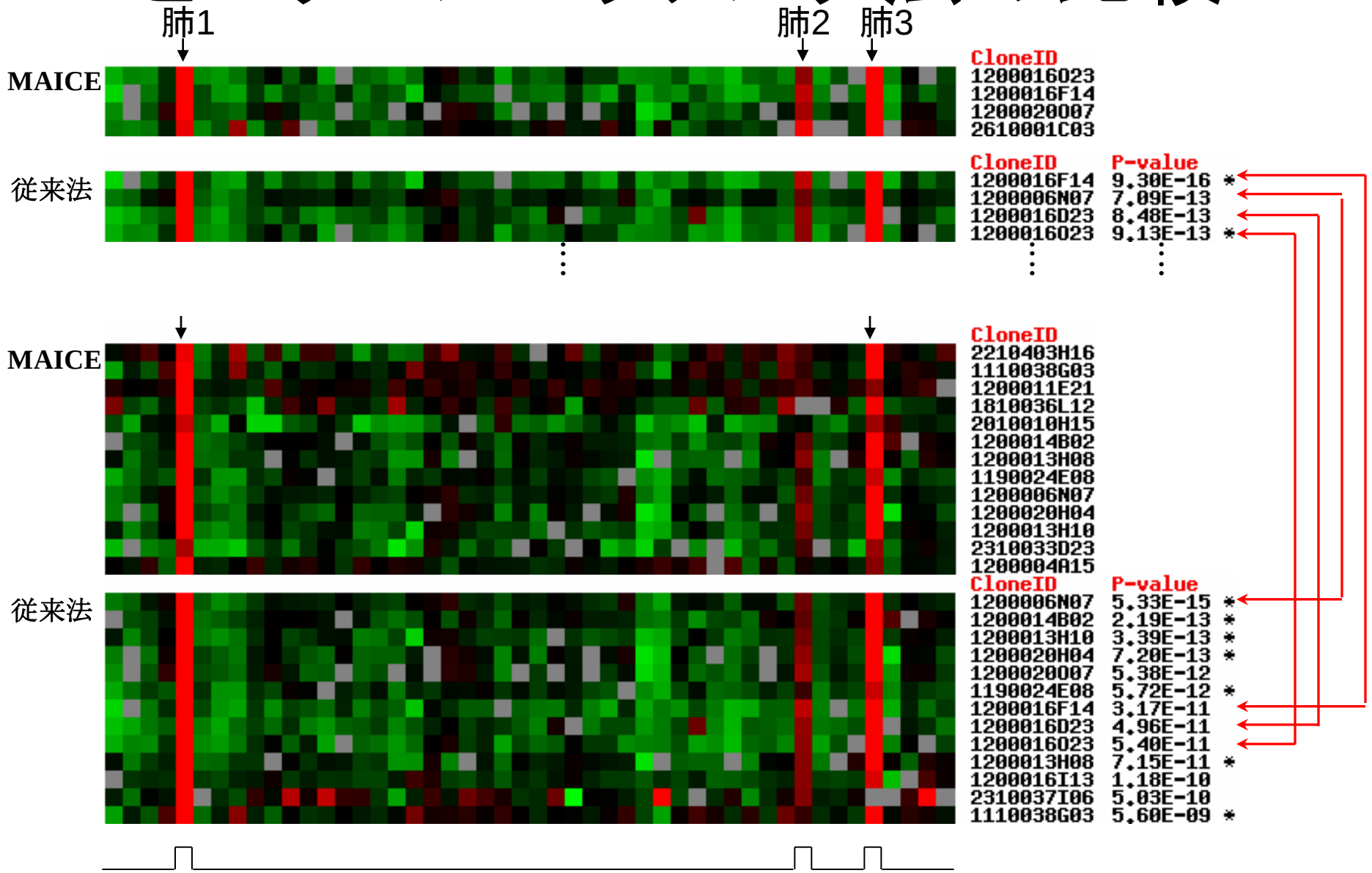
組織特異的遺伝子検出法

- パターンマッチング法
  - 理想的なパターンyとの類似度が高い順にランキング



例：心臓特異的パターンを示す遺伝子群の検出

AICとパターンマッチング法の比較



組織特異的遺伝子検出法

- Tissue specificity index τ
  - Yanai *et al.*, *Bioinformatics*, **21**, 650-659, 2005
  - 遺伝子発現行列 $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$に対し、

$$\tau = \frac{\sum_{i=1}^{n}(1-p_i)}{n-1}, \quad \textit{where} \quad p_i = x_i / \max(\boldsymbol{x})$$

  - 例 : $\boldsymbol{x}$ = (0, *8*, 0, 0, 0, 2, 0, 2, 0, 0, 0, 0)

    $\boldsymbol{p}$ = (0, 1, 0, 0, 0, 0.25, 0, 0.25, 0, 0, 0, 0)

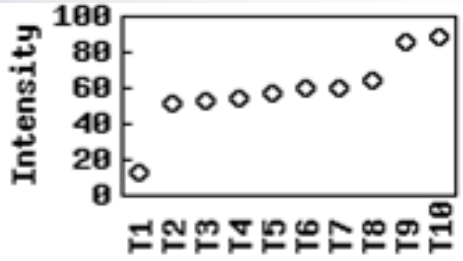
    τ= (1+0+1+1+1+0.75+1+0.75+1+1+1+1)/(12-1) = 0.95
  - τ($\boldsymbol{x}$)のとりうる範囲 : 0 ≦τ≦1

Housekeeping gene **Tissue-specific gene**

統計量τの大きい遺伝子を抽出

参考資料

組織特異的遺伝子検出法

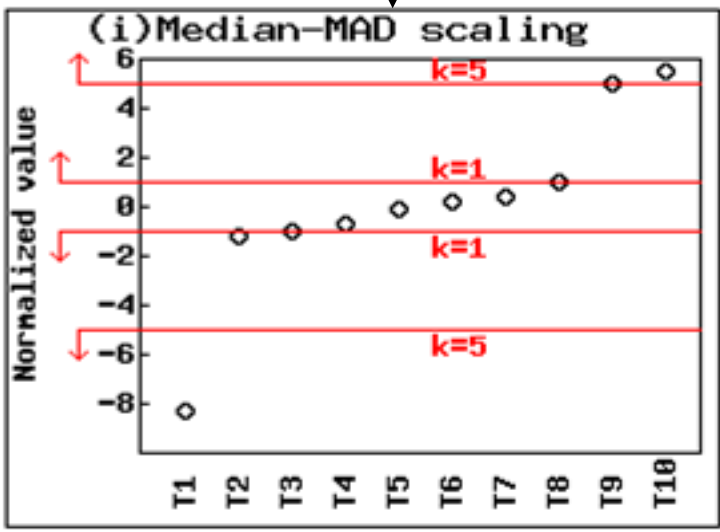


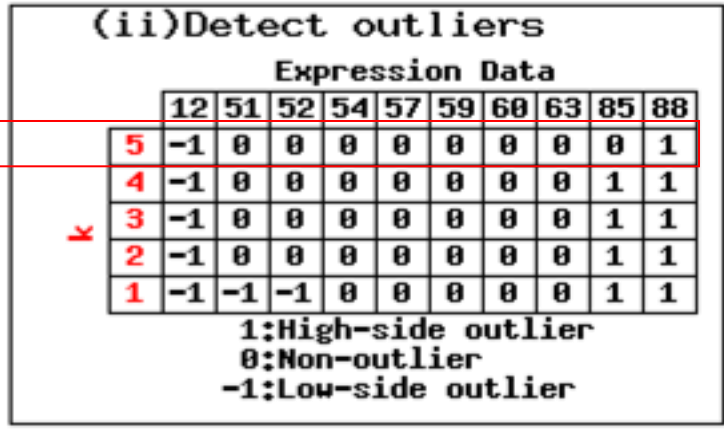
- Sprent's non-parametric method
  - 遺伝子発現ベクトル$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$に対して、

    $x_i < \text{median}(\boldsymbol{x}) - k \times \text{MAD}(\boldsymbol{x})$ and

    $x_i > \text{median}(\boldsymbol{x}) + k \times \text{MAD}(\boldsymbol{x})$

    を満たすx_iを外れ値とする
  - $k = 5$ (原著論文)



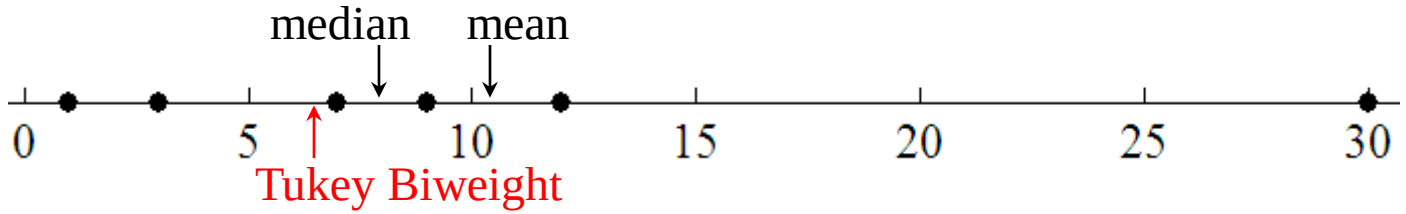


デフォルトの結果

kが変わると得られる結果が異なることには論文中では触れられていない

T_{bw} : Tukey's biweight algorithm

- $\boldsymbol{x} = (\overset{x_1}{1}, \overset{x_2}{3}, \overset{x_3}{7}, \overset{x_4}{9}, \overset{x_5}{12}, \overset{x_6}{30})$ の重みつき平均を求める
  - mean = (1+3+7+9+12+30)/6=10.3
  - median M= (7+9)/2=8
  - 外れ値の影響をなるべく受けないようにしたい
    - median近辺の数値（7や9）には1に近い重み
    - 遠く離れるほど重みを軽くしたい



T_{bw} : Tukey's biweight algorithm

- Median Absolute Deviation (*MAD*)の計算（→全体のバラツキを数値化）

$MAD(\boldsymbol{x})$ = median ($|x_1-M|$, $|x_2-M|$, $|x_3-M|$, $|x_4-M|$, $|x_5-M|$, $|x_6-M|$)

= median (|1-8|, |3-8|, |7-8|, |9-8|, |12-8|, |30-8|)

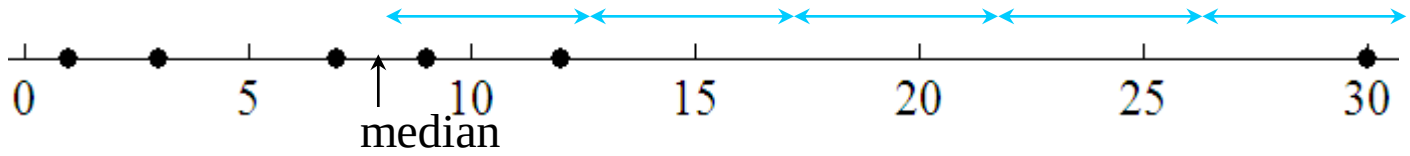
= median (7, 5, 1, 1, 4, 22)

= (4+5)/2= 4.5

- 標準化（≒*Z-score*化）

$$t_1 = \frac{x_1 - M}{c \times MAD + \varepsilon} = \frac{x_1 - M}{5 \times MAD + 0.0001} = \frac{x_1 - 8}{5 \times 4.5 + 0.0001} = -0.311$$

$$t_2 = -0.222,\ t_3 = -0.044,\ t_4 = 0.044,\ t_5 = 0.178,\ t_6 = 0.978$$



T_{bw} : Tukey's biweight algorithm

- 重み関数（bisquare weight function）

$$w(t_i) = \begin{cases} (1-t_i^2)^2, if\ |t_i| \le 1 \\ 0, \quad else \end{cases}$$

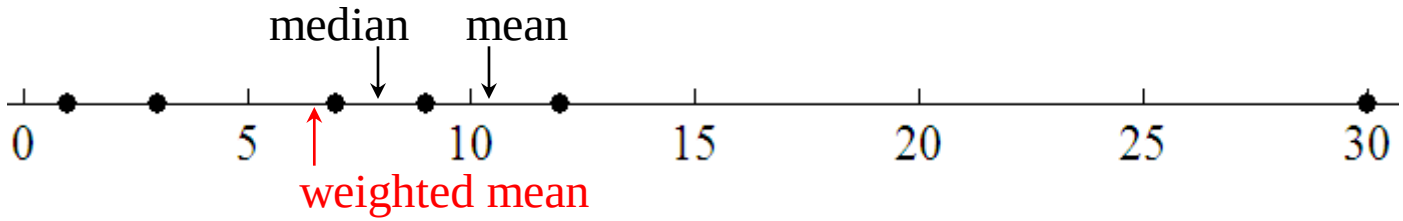
i	x_i	t_i	$w(t_i)$	
1	1	-0.311	0.816	
2	3	-0.222	0.904	
3	7	-0.044	0.996	Median（=8）に近いので重みが1に近い
4	9	0.044	0.996	
5	12	0.178	0.938	
6	30	0.978	0.002	←Median（=8）より非常に遠い（30）ので、重みが限りなく0に近い

- 重みつき平均

$$T_{bi}(x_1, x_2, ..., x_n) = \frac{\sum_{i=1}^{n} w(t_i) x_i}{\sum_{i=1}^{n} w(t_i)}$$

$$= \frac{0.816\times(-0.311) + 0.904\times(-0.222) + 0.996\times(-0.044) + 0.996\times 0.044 + 0.938\times 0.178 + 0.002\times 0.978}{0.816 + 0.904 + 0.996 + 0.996 + 0.938 + 0.002}$$

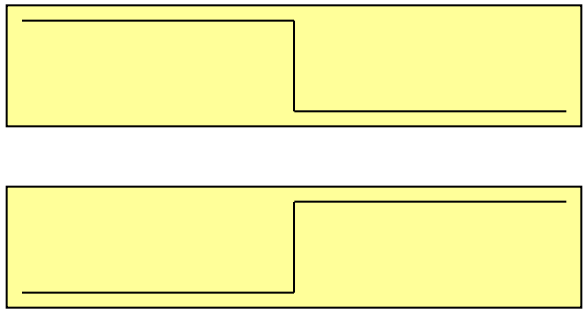
$$= 6.62$$



遺伝子発現行列

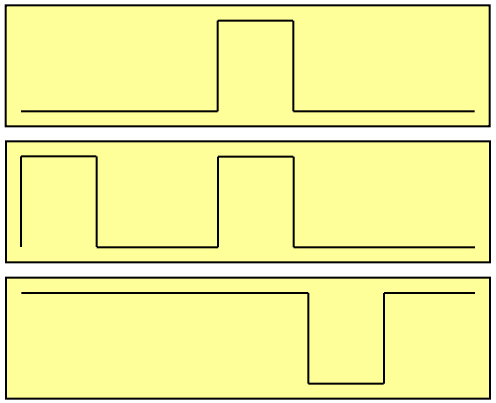
二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



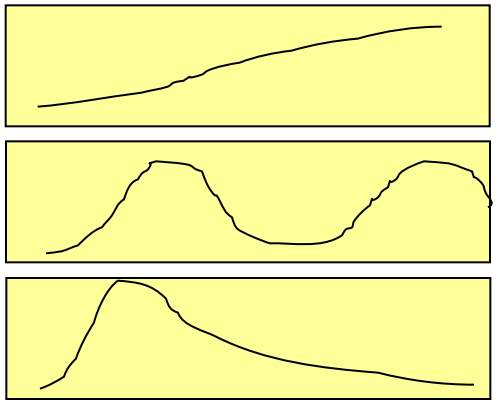
様々な組織（条件）

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



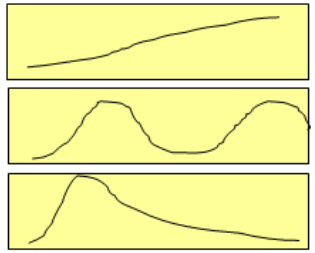
時系列データ

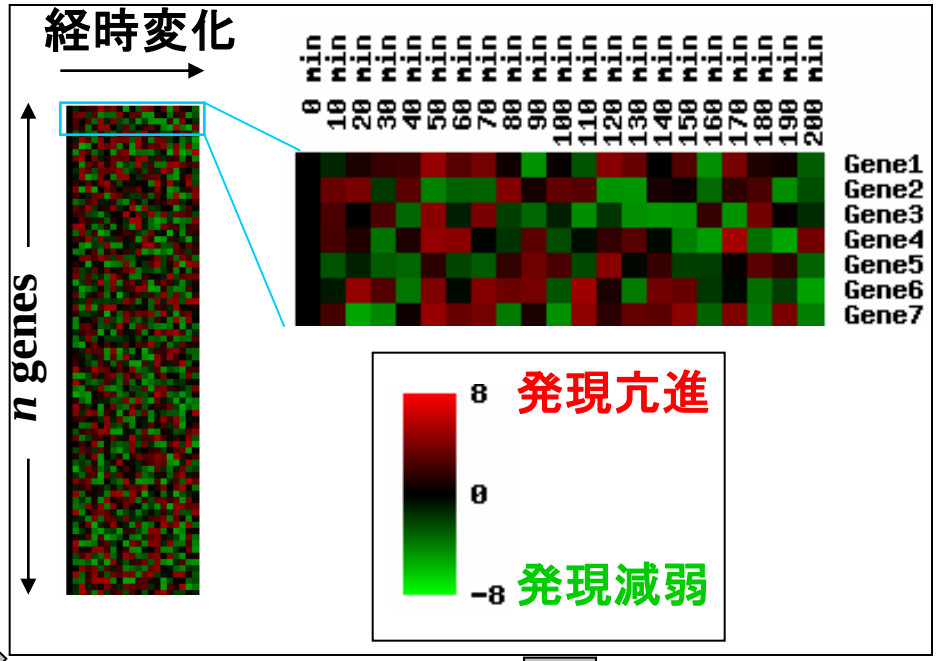
	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



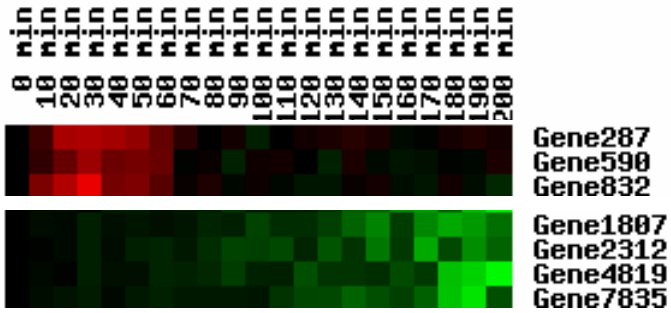
時系列データ

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

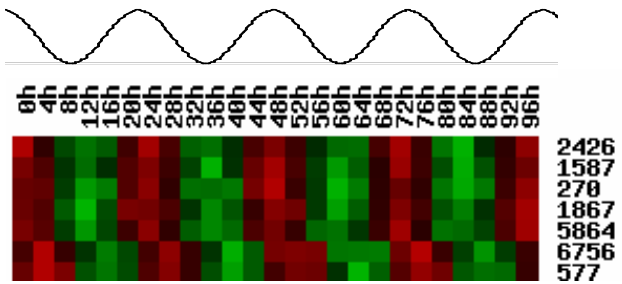




薬物投与後の発現変化モニタリング



概日リズム関連遺伝子探索



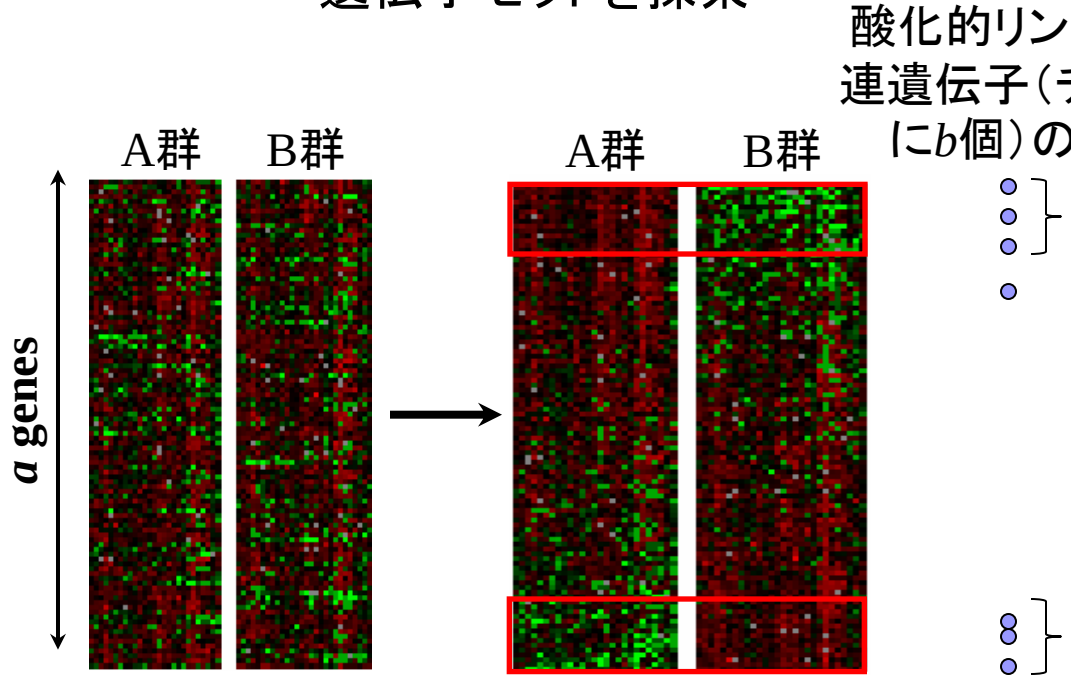
（機能性食品の量・濃度）

様々な時系列データ解析手法

- 周期性解析（概日リズム、細胞周期）
  - **Lomb-Scargle method (Glynn *et al.*, *Bioinformatics*, 22, 310-316, 2006)**
  - C&G procedure (Chen J., *BMC Bioinformatics*, **6**, 286, 2005)
  - A model-based method (Luan and Li, *Bioinformatics*, **20**, 332-339, 2004)
  - **GeneTS (Wichert *et al.*, *Bioinformatics*, 20, 5-20, 2004)**
- その他
  - **Di Camillo *et al.*, *BMC Bioinformatics*, 8 (Suppl 1), S10, 2007.**
  - Ahnert *et al.*, *Bioinformatics*, **22**, 1471-1476, 2006.
  - ICA (Frigyesi *et al.*, *BMC Bioinformatics*, **7**, 290, 2006.)
  - **maSigPro (Conesa *et al.*, *Bioinformatics*, 22, 1096-1102, 2006.)**
  - dynamic model-based clustering (Wu et al., *J. Bioinform. Comput. Biol.*, **3**, 821-836, 2005.)
  - Step-down quadratic regression (Liu *et al.*, *BMC Bioinformatics*, **6**, 106, 2005)

機能解析(GSEA解析)

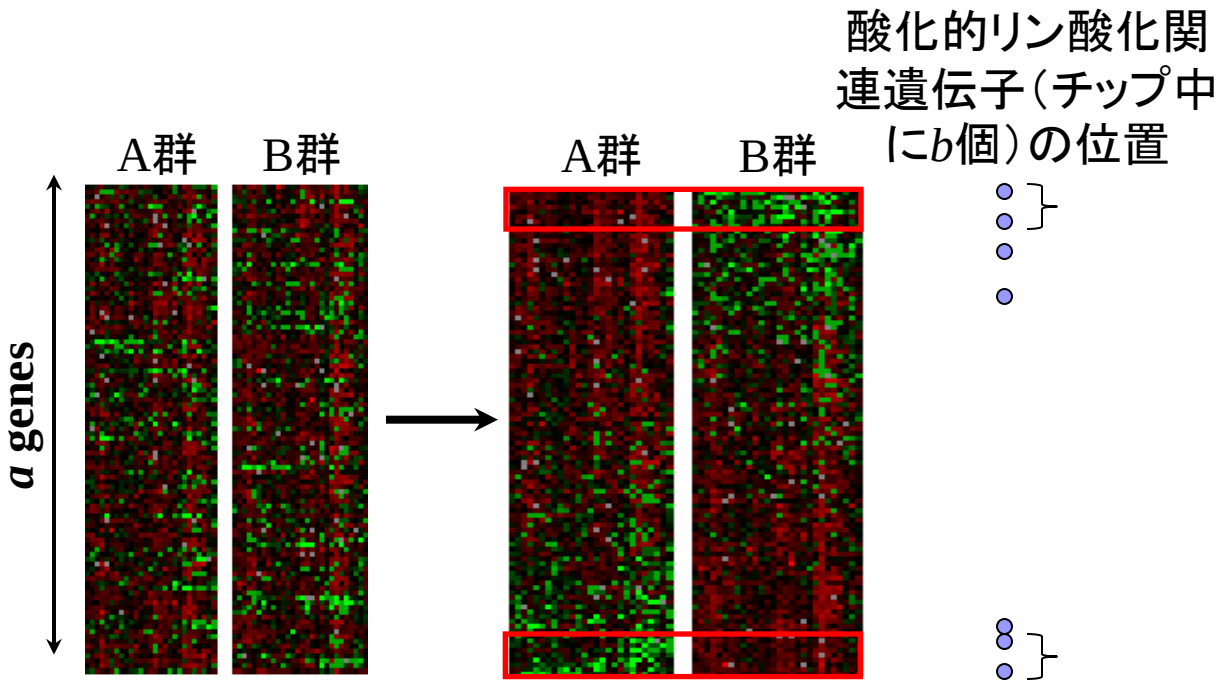
- この種の解析法の論文が出る前のメジャーな機能解析手段
  - 例:二群間比較
    1. 何らかの手段で発現変動の度合いでランキング
    2. 上位*x*個を抽出し、**XXX(例:酸化的リン酸化)関連遺伝子群**(**G**ene **S**et:遺伝子セット)がどれだけ濃縮(**E**nrichment)されているのかを解析(**A**nalysis)
    3. 遺伝子セット(**XXX**に相当)をいろいろ変えて、二群間で発現変動している遺伝子セットを探索



帰無仮説:
「チップ中の全遺伝子数(*a*)に対する酸化的リン酸化関連遺伝子数(*b*)の割合(*b*/*a*)」
と
「酸化的リン酸化関連遺伝子数(*b*)に対する上位*x*個の中に占める酸化的リン酸化関連遺伝子数(*c*)の割合(*c*/*x*)」
は等しい

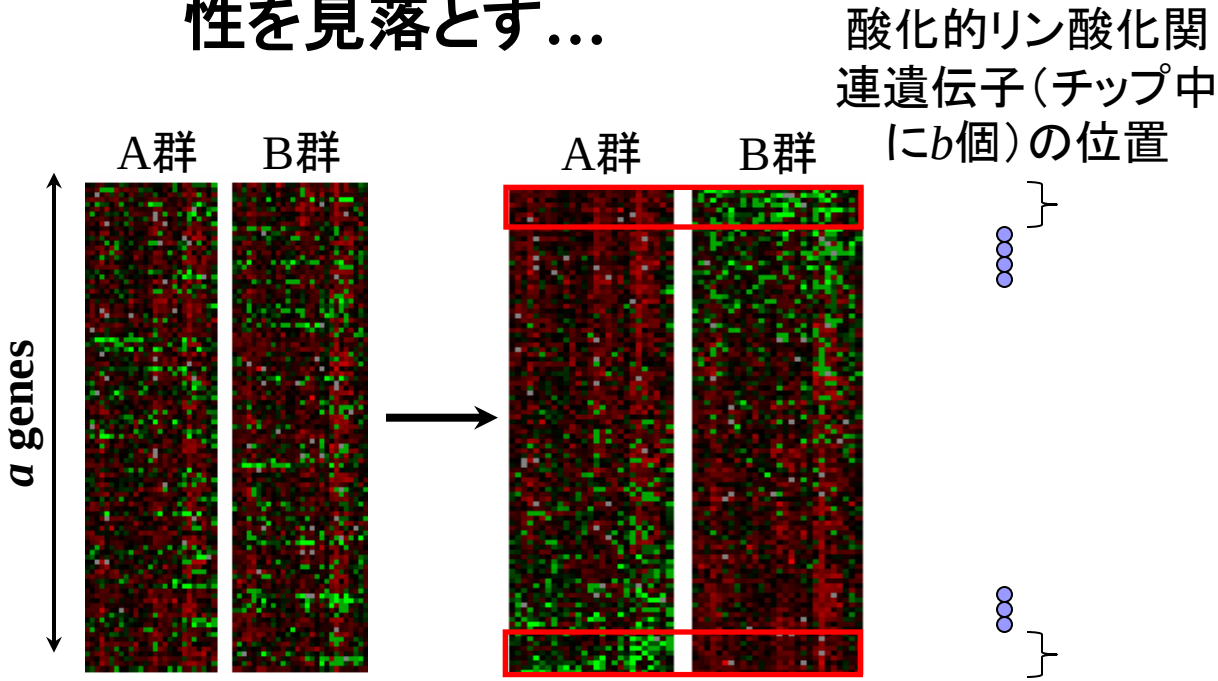
機能解析（GSEA解析）

- この種の解析法の論文が出る前のメジャーな機能解析手段の問題点1
  - 上位x個のx次第で結果が変わる



機能解析（GSEA解析）

- この種の解析法の論文が出る前のメジャーな機能解析手段の問題点2
  - 下図のように、全体としては**XXX(例：酸化的リン酸化)関連遺伝子群**が有意差があるといえるような場合でも、上位*x*個の中に一つも含まれないので有意差があるといえなくなる…。
  - 現実の解析では**XXX(例：酸化的リン酸化)関連遺伝子群の重要性を見落とす…**



様々な機能解析手法

- GSEA (Subramanian *et al*., *PNAS*, 2005)
- PAGE (Kim and Volsky, *BMC Bioinformatics*, 2005)
- GSA (Efron and Tibshirani, *Ann. Appl. Stat*., 2007)
- GeneTrail (Backes *et al*., *NAR*, 2007)
- SAM-GS (Dinu *et al*., *BMC Bioinformatics*, 2007)
- GSEA-P (Subramanian *et al*., *Bioinformatics*, 2007)
- …

PAGE法の概略

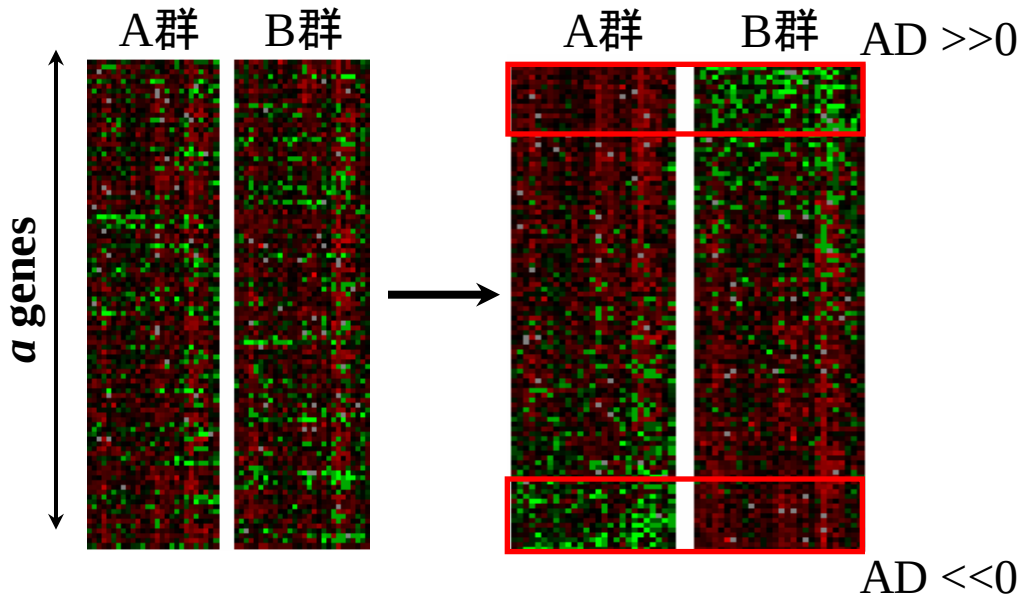
- Parametric Analysis of Gene set Enrichmentの略
  1. 各遺伝子iについて対数変換後のデータのAverage Difference (AD^i)を計算　$AD^i = \overline{A^i} - \overline{B^i}, \qquad (i = 1,2,...,n)$
  2. AD^iの平均μと標準偏差σを計算
  3. 興味ある遺伝子セット（例：i=5,89, 684, 2543, …に相当する計m個の遺伝子）のADの平均S_mを計算

$$S_m = (AD^5 + AD^{89} + AD^{684} + AD^{2543} + ...)/m$$

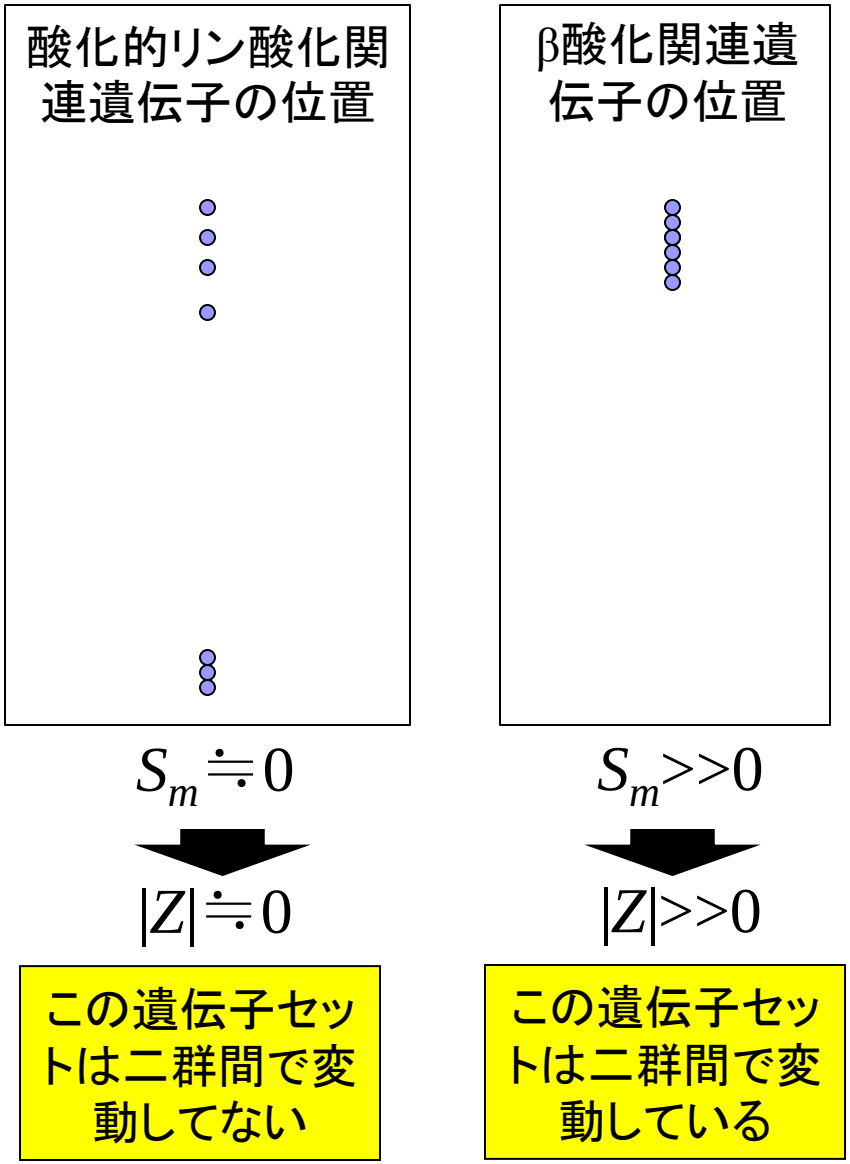
  4. Zスコアを計算 $Z = (S_m - \mu) \times \sqrt{m} / \sigma$

Zスコアの絶対値が大きい遺伝子セットほど二群間でより発現変動している、と解釈

「(Rで)マイクロアレイ」のPAGE(現状)



Zスコアの絶対値が大きい遺伝子セットほど二群間でより発現変動している、と解釈



様々な機能解析手法

- なぜ次々と提案されるのか？
  - Ans.1：発現変動遺伝子のランキング法はいくらでもある
    - PAGE：Average Difference (*AD*) ← 倍率変化そのもの
    - GSEA：S2N統計量など
    - Rank products, WAD, SAMなど
  - Ans.2：興味ある遺伝子セットの偏り度合い（濃縮度）を見積もる方法はいくらでもある
    - PAGE：Z検定
    - GSEA：Kolmogorov-Smirnov統計量の改良版
    - 平均%順位, AUC, t検定など

機能解析手法を使えるのはごく一部の生物種

- アノテーション情報が豊富な生物種はGene Ontologyやパスウェイの情報が豊富
  →多くの遺伝子セットを用意できる→機能解析手法を適用可能
- それ以外の生物種は、まずは様々な発現変動遺伝子をひたすら同定しまくるなどして地道にアノテーション情報を増やしていく以外にない（のではないだろうか）

クラスタリング（教師なし学習）

- サンプルの属性情報（癌 or 正常など）を使わずに、発現情報のみを用いて発現パターンの類似した遺伝子（またはサンプル）をクラスター（群）にしていく手法（Unsupervised learning

二群間比較

	A群 A1	A2	…	B群 B1	B2	…
gene *1*	$x^A_{1,1}$	$x^A_{1,2}$		$x^B_{1,2}$	$x^B_{1,2}$	
gene *2*	$x^A_{2,1}$	$x^A_{2,2}$		$x^B_{2,2}$	$x^B_{2,2}$	
…	…	…	…	…	…	…
gene *i*	$x^A_{i,1}$	$x^A_{i,2}$		$x^B_{i,2}$	$x^B_{i,2}$	
…	…	…	…	…	…	…
gene *n*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	

多サンプル

	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

時系列解析

	T1	T2	T3	T4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

クラスタリング（教師なし学習）

- 例1：遺伝子間クラスタリング

似た機能をもつものは同じ
クラスターに属すことを確認

クラスタリング（教師なし学習）

- 例2：サンプル間クラスタリング

悪性度の高い癌のサブタイプを発見

クラスタリング（教師なし学習）

- 階層的クラスタリング
  - 発現パターンの類似した遺伝子を集めて系統樹を作成
- 非階層的クラスタリング
  - K-meansクラスタリング
    - 「K個のクラスターに分割（Kの数は主観的に決定）する」と予め指定し、各クラスター内の遺伝子（サンプル）間の距離の総和が最小になるようなK個のクラスターを作成
  - 自己組織化マップ（SOM）
  - 主成分分析（PCA）

距離(類似度)の定義

- 遺伝子(or サンプル)xとyの発現パターンの距離D

相関係数 $r_{xy} = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ $\quad (-1 \le r_{xy} \le 1)$

i	x	y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
4	x_4	y_4
5	x_5	y_5
…	…	…
n	x_n	y_n

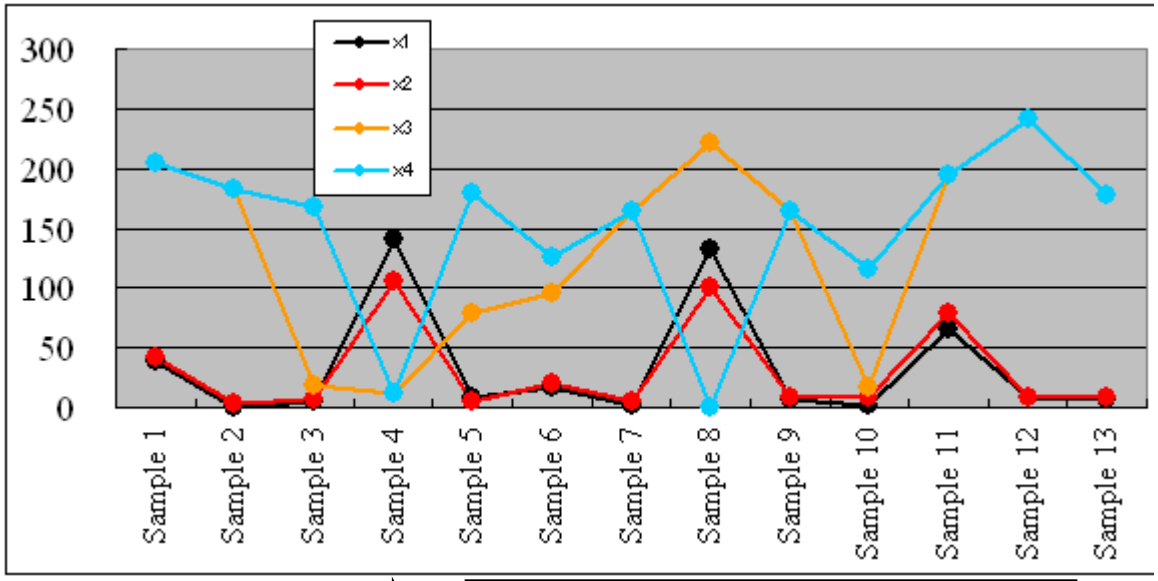
- xとyの発現パターンが酷似 $\rightarrow$ $r \approx 1$
- xとyの発現パターンがばらばら $\rightarrow$ $r \approx 0$
- xとyの発現パターンがほぼ正反対 $\rightarrow$ $r \approx -1$

距離 $D = 1 - r \quad (0 \le D \le 2)$

- $r = 1 \rightarrow D = 1 - 1 = 0$
- $r = 0 \rightarrow D = 1 - 0 = 1$
- $r = -1 \rightarrow D = 1 - (-1) = 2$

階層的クラスタリング

1. 遺伝子間距離を計算



例：4遺伝子の場合

	x^1	x^2	x^3	x^4
Sample 1	38	42	204	204
Sample 2	0	3	182	182
Sample 3	6	6	19	168
Sample 4	141	106	11	11
Sample 5	8	5	79	179
Sample 6	16	20	96	126
Sample 7	2	5	164	164
Sample 8	132	101	222	0
Sample 9	7	9	164	164
Sample 10	2	8	16	116
Sample 11	66	79	195	195
Sample 12	8	9	241	241
Sample 13	6	8	177	177

距離 $D = 1 - r \quad (0 \le D \le 2)$ ⇒ 距離 $D = \dfrac{1-r}{2} \quad (0 \le D \le 1)$

相関係数 $r_{1,2} = 0.98 \rightarrow$ 距離 $D_{1,2} = \dfrac{1-0.98}{2} = 0.01$

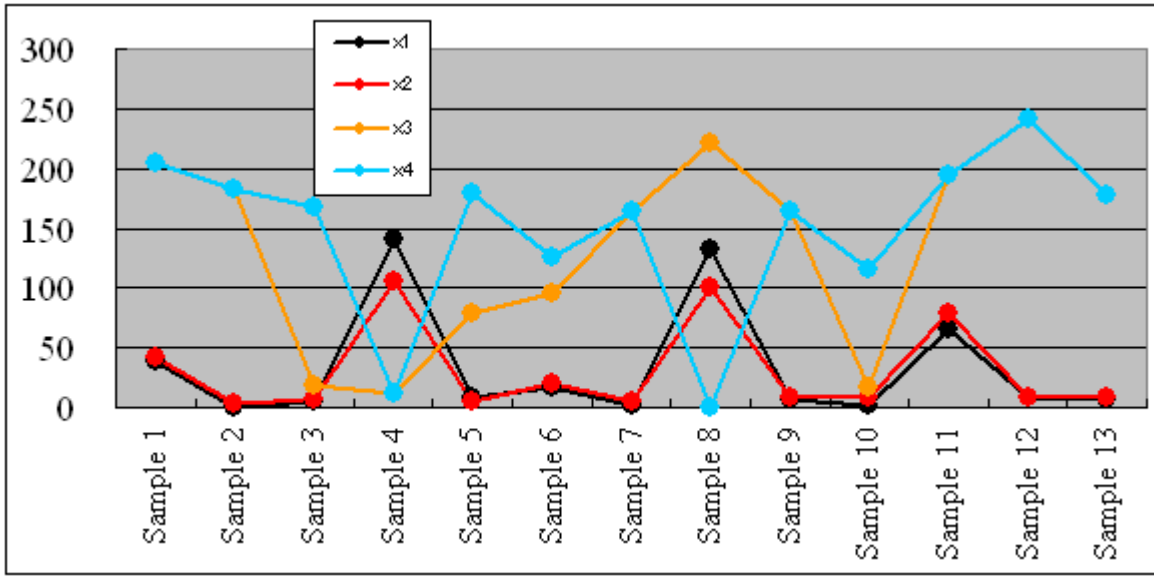
相関係数 $r_{1,3} = -0.01 \rightarrow$ 距離 $D_{1,3} = \dfrac{1-(-0.01)}{2} = 0.50$

相関係数 $r_{1,4} = -0.78 \rightarrow$ 距離 $D_{1,4} = \dfrac{1-(-0.78)}{2} = 0.89$

…

階層的クラスタリング

2. 距離行列を作成



距離 $D_{1,2} = \frac{1-0.98}{2} = 0.01$

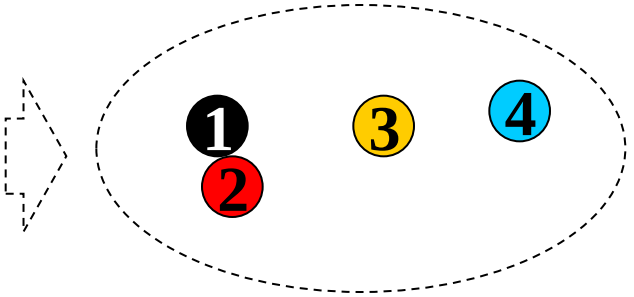
距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

距離 $D_{1,4} = \frac{1-(-0.78)}{2} = 0.89$

…

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

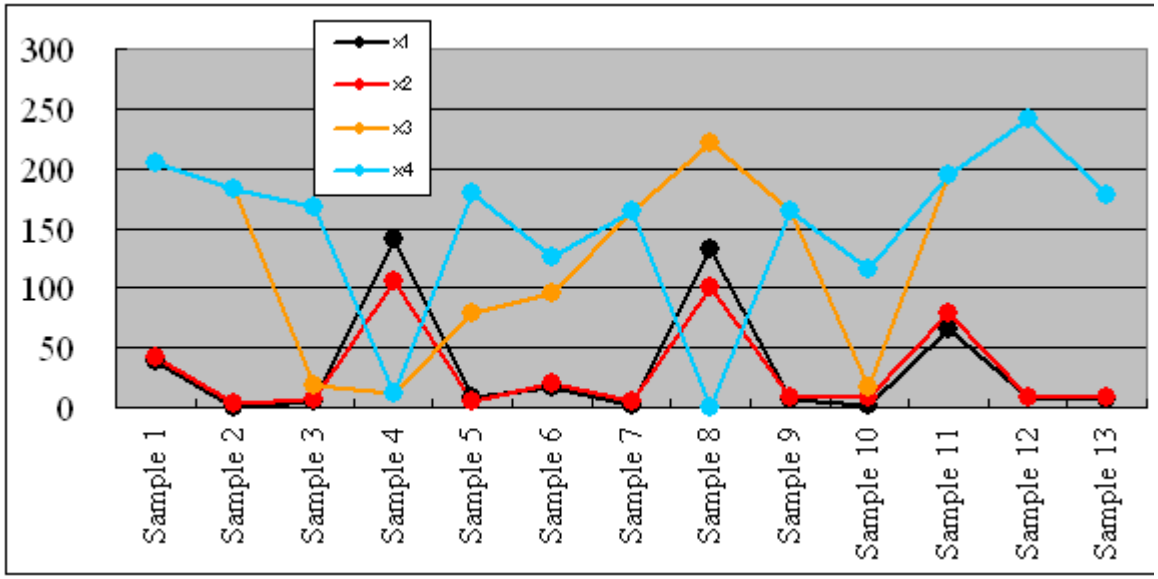
距離行列



イメージ

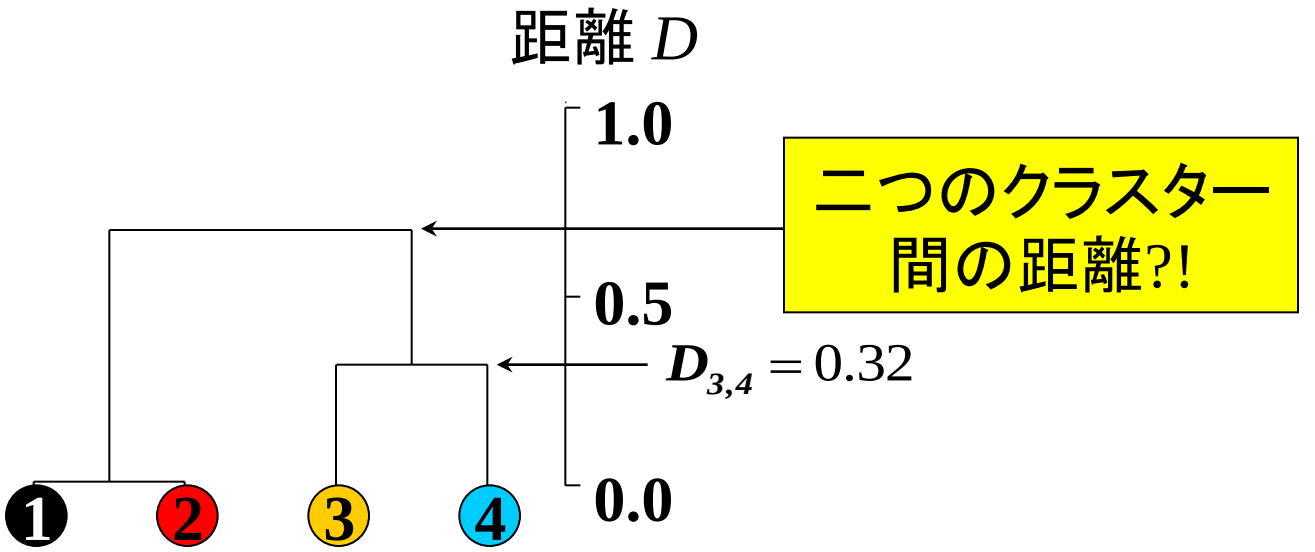
階層的クラスタリング

3. 樹形図を作成



	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32

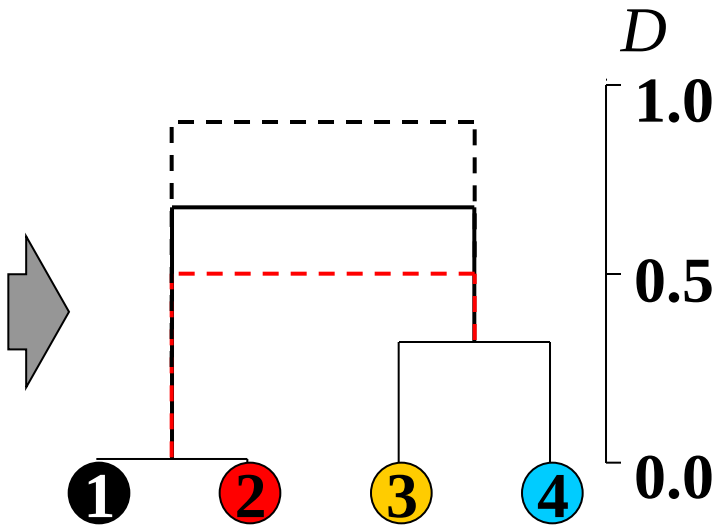
距離行列



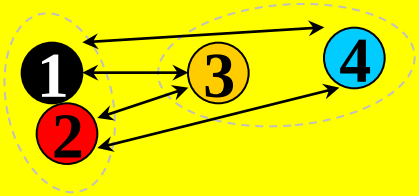
階層的クラスタリング

3. 樹形図を作成

	x^2	x^3	x^4
x^1	0.01	0.50	0.89
x^2		0.47	0.84
x^3			0.32



平均連結法の場合



$$(D_{1,3}+D_{1,4}+D_{2,3}+D_{2,4})/4$$
$$=(0.50+0.89+0.47+0.84)/4$$
$$=0.68$$

単連結法の場合

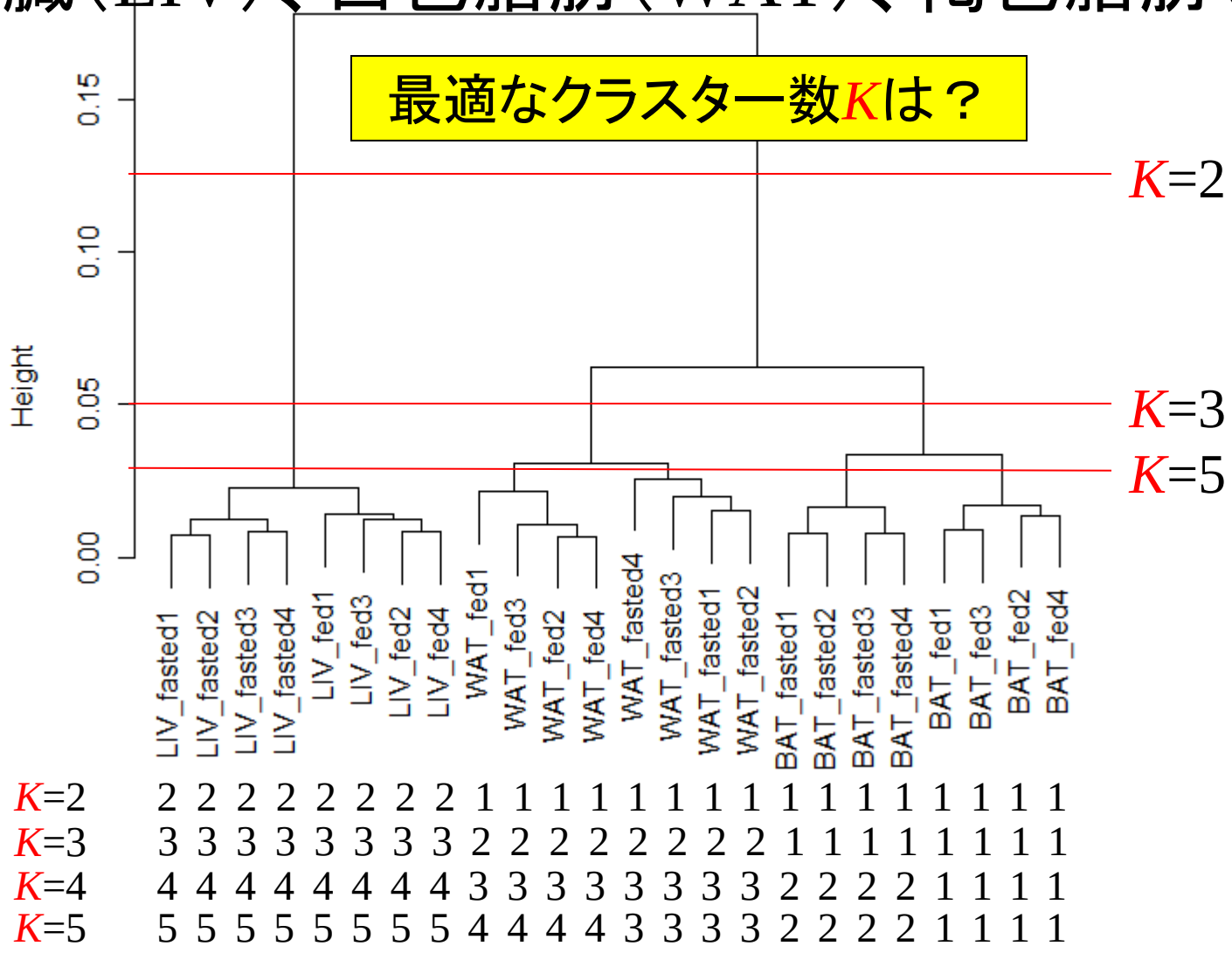
$$\min(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.47$$

完全連結法の場合

$$\max(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.89$$

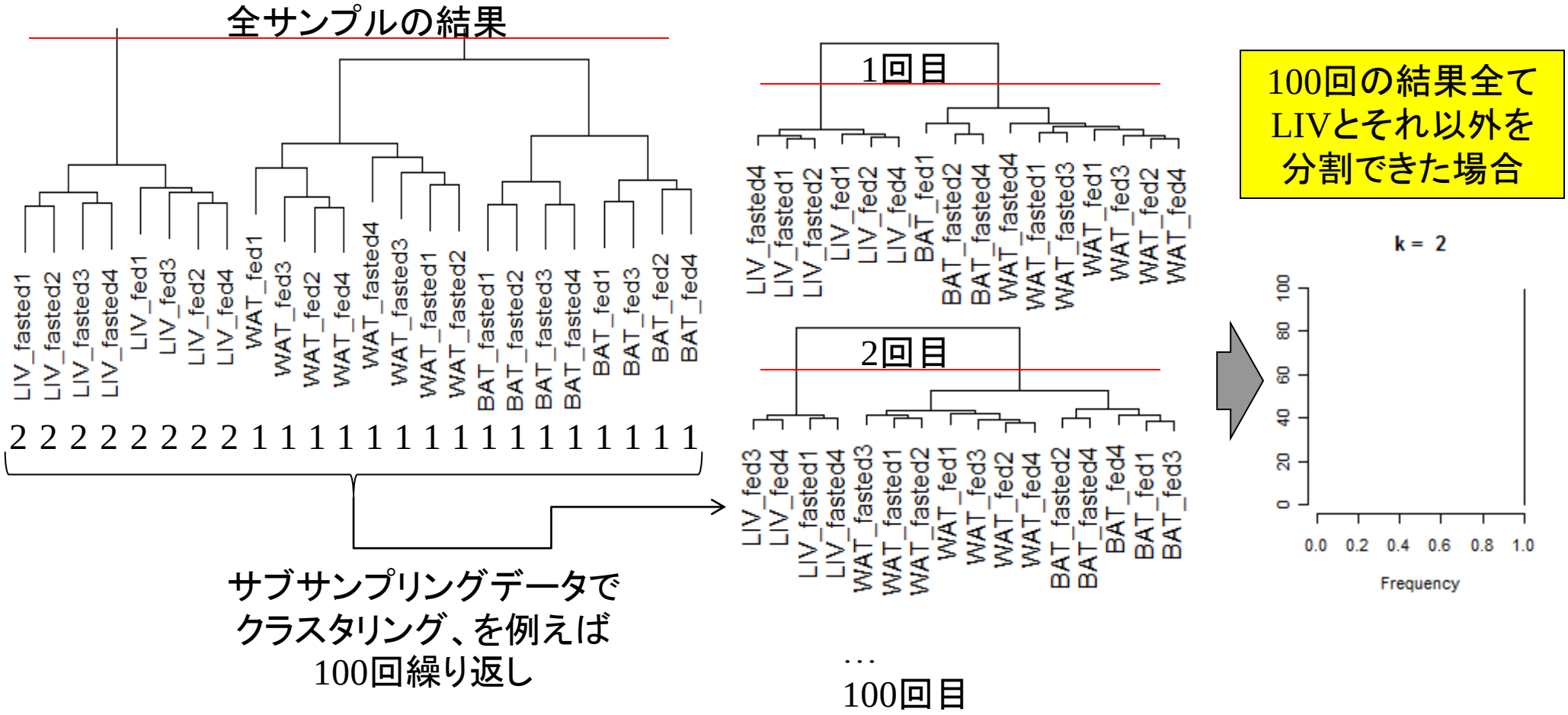
階層的クラスタリング例

- 肝臓(LIV)、白色脂肪(WAT)、褐色脂肪(BAT)



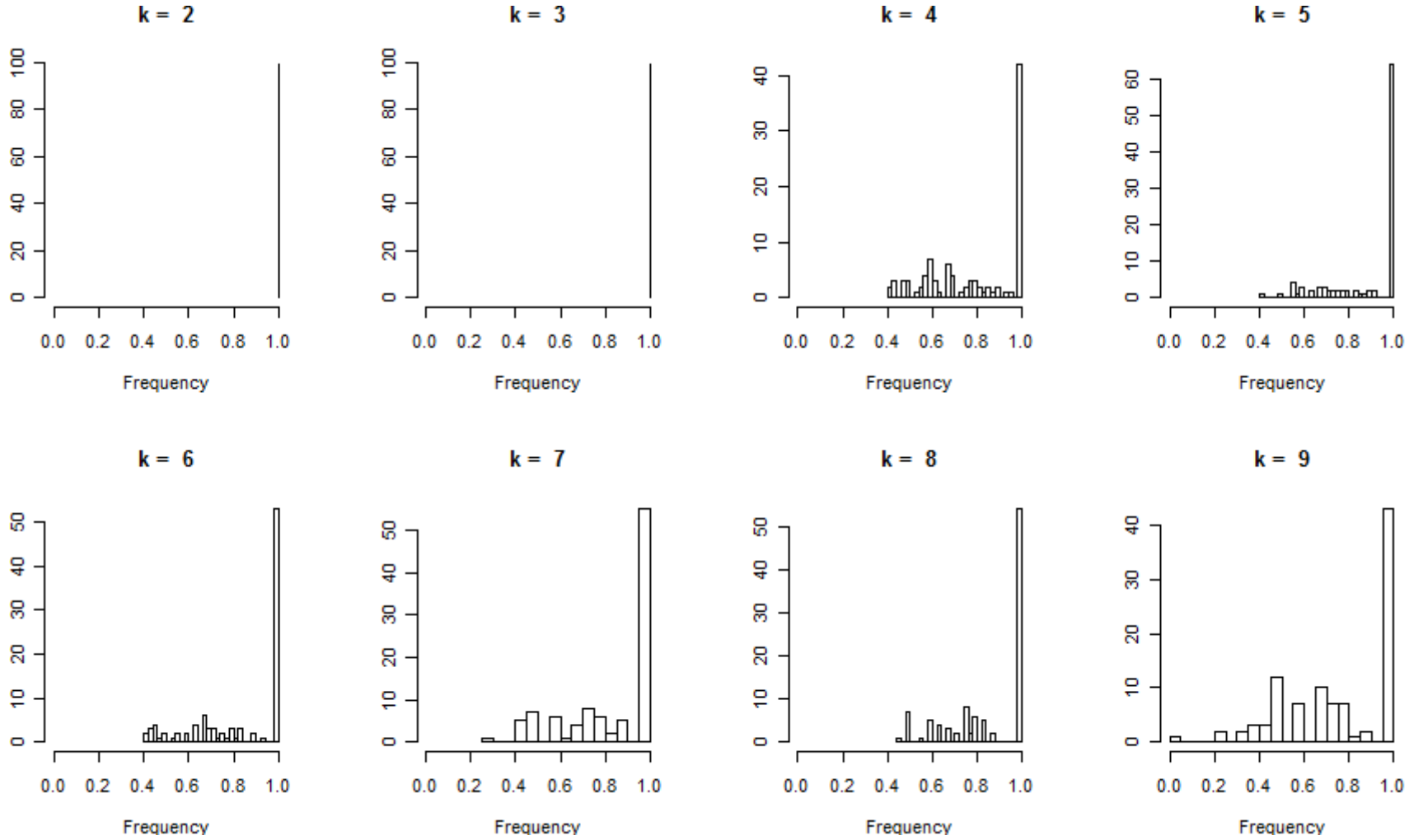
最適なクラスター数を見積もる方法

- 様々なKについて（例えばK=2）全サンプル（n）のクラスタリング結果をK個に分割した結果とサブサンプル（例えばn*0.7）のクラスタリング結果をK個に分割した結果の類似度を計算



最適なクラスター数を見積もる方法

- *K*の値をいくつか試して（例では2~9）、最適な*K*の値を同定



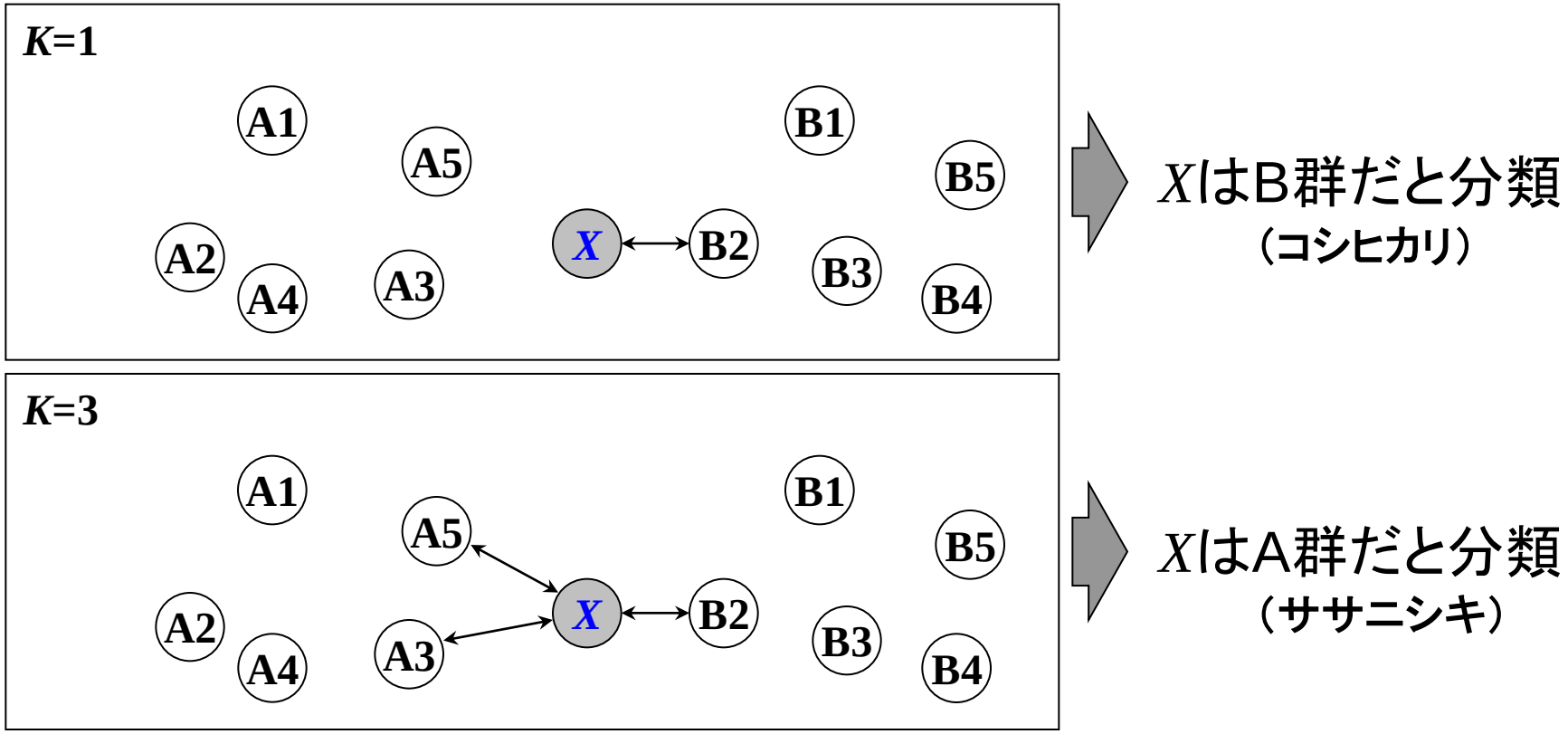
この場合は*K*=2, 3が最適なクラスター数

分類（教師あり学習）

- **未知サンプルを分類するための様々な方法**
  - **K-Nearest Neighbor (K-NN; K-最近傍法)**
  - Support Vector Machine (SVM)
  - Neural Network (NN)
  - Naïve Bayesian (NB)
  - Multi-Layer Perceptron (MLP; 多層パーセプトロン)
  - Weighted Voting (WV; 重みつき多数決法)
  - Decision Tree
  - etc…

K-Nearest Neighbor (*K*-NN) 法

- 未知サンプル*X*からの距離がもっとも近い*K*個のサンプルのうち、所属するクラスが最も多いクラスに分類

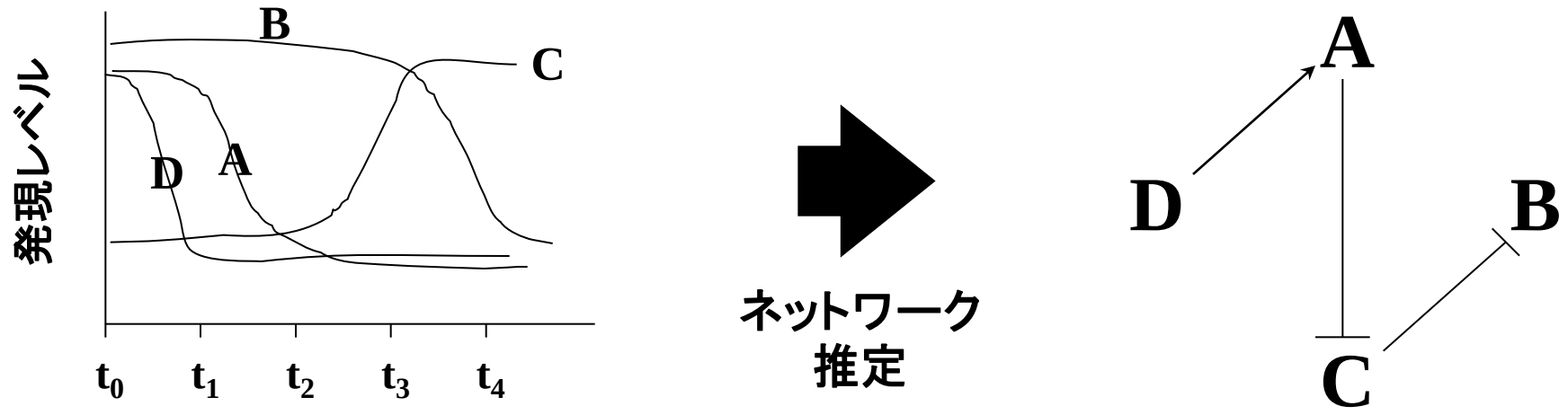


交差検証法（Cross-validation）

- 手持ちのデータセットを利用して、用いた分類器（K-NNなど）とその中で用いたパラメータ（Kの数）採用時の分類精度を評価する手段
  - Leave-one-out (take-one-sample-out or hold-one-out) approach
    1. 手持ちのデータセット（nサンプル）の中から一つを（本当はクラス既知だが）未知のテストサンプルとしてデータセットから除く
    2. 残りの(n-1)サンプルからなるデータセットから、分類に用いる遺伝子サブセット(predictor genes)を得る
    3. Predictor genesの発現プロファイルを用いて、テストサンプルを予測
    4. 全サンプルに対して1-3を繰り返し、予測精度を見積もる
  - Cross-fold approach
    1. 各クラスから一定（例えばA群から$n_A/2$個、B群から$n_B/2$個）サンプル数をテストセットとする
    2. 残り（$n/2$）のサンプルからなるデータセットから、分類に用いる遺伝子サブセット(predictor genes)を得る
    3. Predictor genesの発現プロファイルを用いて、テストサンプルを予測
    4. 指定回数1-3を繰り返し、予測精度を見積もる

遺伝子の発現制御ネットワーク推定

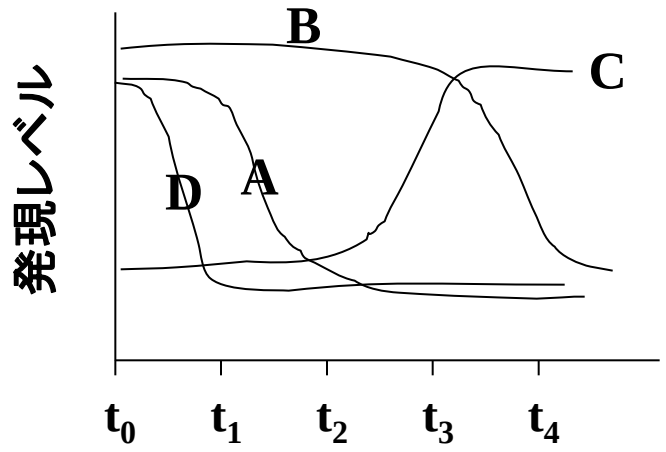
- 時系列データ
  - 遺伝子Dの発現を抑制し、他の遺伝子の挙動を観察



遺伝子の発現制御ネットワーク推定

- 時系列データ
  - 遺伝子発現行列の作成

    例）t_0に対するlog比などで表現



Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

遺伝子の発現制御ネットワーク推定

- 時系列データ
  - 「(基本的な)線形モデル法」で解いてみる

仮定:遺伝子x_kの時間tにおける発現レベル$x_k{}^t$は、時間t-1における他のすべての遺伝子発現レベルの線形結合で表される

$$x_k^t = \sum_{i=1}^{N} w_{i,k} x_i^{t-1}$$

$w_{i,k}$: x_iの発現レベルがx_kの発現レベルに及ぼす影響を示す重み係数

Gene	0	1	2	3	4
x_1	$x_1{}^0$	$x_1{}^1$	$x_1{}^2$	$x_1{}^3$	$x_1{}^4$
x_2	$x_2{}^0$	$x_2{}^1$	$x_2{}^2$	$x_2{}^3$	$x_2{}^4$
x_3	$x_3{}^0$	$x_3{}^1$	$x_3{}^2$	$x_3{}^3$	$x_3{}^4$
x_4	$x_3{}^0$	$x_3{}^1$	$x_3{}^2$	$x_3{}^3$	$x_3{}^4$

$\longleftrightarrow$

Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

「(基本的な)線形モデル法」で解く

- 行列で表すと以下のような感じになる

$$\begin{pmatrix} A^t \\ B^t \\ C^t \\ D^t \end{pmatrix} = \begin{pmatrix} w_{A,A} & w_{A,B} & w_{A,C} & w_{A,D} \\ w_{B,A} & w_{B,B} & w_{B,C} & w_{B,D} \\ w_{C,A} & w_{C,B} & w_{C,C} & w_{C,D} \\ w_{D,A} & w_{D,B} & w_{D,C} & w_{D,D} \end{pmatrix} \begin{pmatrix} A^{t-1} \\ B^{t-1} \\ C^{t-1} \\ D^{t-1} \end{pmatrix}$$

遺伝子発現行列(時系列データ)

Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

目的:4^2個の未知の$w_{i,k}$を決める

重み行列→相互作用行列

$$x_k^t = \sum_{i=1}^{N} w_{i,k} x_i^{t-1}$$

「(基本的な)線形モデル法」で解く

- 計算結果

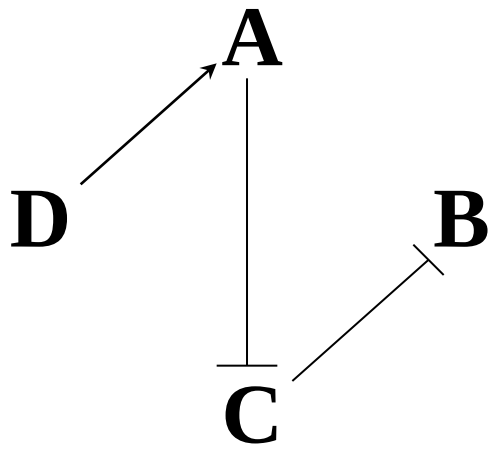
$$\begin{pmatrix} A^t \\ B^t \\ C^t \\ D^t \end{pmatrix} = \begin{pmatrix} w_{A,A} & w_{A,B} & w_{A,C} & w_{A,D} \\ w_{B,A} & w_{B,B} & w_{B,C} & w_{B,D} \\ w_{C,A} & w_{C,B} & w_{C,C} & w_{C,D} \\ w_{D,A} & w_{D,B} & w_{D,C} & w_{D,D} \end{pmatrix} \begin{pmatrix} A^{t-1} \\ B^{t-1} \\ C^{t-1} \\ D^{t-1} \end{pmatrix}$$

遺伝子発現行列(時系列データ)

Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

遺伝子間相互作用行列

Gene	A	B	C	D
A			-1	
B				
C		-1		
D	1			



「(基本的な)線形モデル法」で解く

- 目的:重み係数$w_{i,k}$を解として得る
  - 例)遺伝子Aの発現調節を支配している方程式を解く

$$x_k^t = \sum_{i=1}^{N} w_{i,k} x_i^{t-1}$$

$$\begin{cases} A^{t4} = w_{A,A}A^{t3} + w_{B,A}B^{t3} + w_{C,A}C^{t3} + w_{D,A}D^{t3} \\ A^{t3} = w_{A,A}A^{t2} + w_{B,A}B^{t2} + w_{C,A}C^{t2} + w_{D,A}D^{t2} \\ A^{t2} = w_{A,A}A^{t1} + w_{B,A}B^{t1} + w_{C,A}C^{t1} + w_{D,A}D^{t1} \\ A^{t1} = w_{A,A}A^{t0} + w_{B,A}B^{t0} + w_{C,A}C^{t0} + w_{D,A}D^{t0} \end{cases}$$

Gene	0	1	2	3	4
x_1	x_1^0	x_1^1	x_1^2	x_1^3	x_1^4
x_2	x_2^0	x_2^1	x_2^2	x_2^3	x_2^4
x_3	x_3^0	x_3^1	x_3^2	x_3^3	x_3^4
x_4	x_3^0	x_3^1	x_3^2	x_3^3	x_3^4

↕

Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

「(基本的な)線形モデル法」で解く

- 目的:重み係数$w_{i,k}$を解として得る
  - 例)遺伝子Aの発現調節を支配している方程式を解く

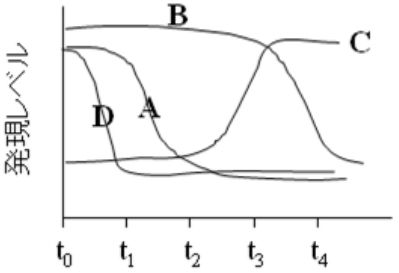
Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

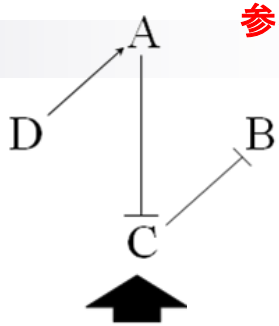
$$
\begin{cases}
-1 = w_{A,A}(-1) + w_{B,A}(0) + w_{C,A}(1) + w_{D,A}(-1) & \rightarrow w_{C,A} = 0 \\
-1 = w_{A,A}(-1) + w_{B,A}(0) + w_{C,A}(0) + w_{D,A}(-1) & \rightarrow w_{A,A} = 0 \\
-1 = w_{A,A}(0) + w_{B,A}(0) + w_{C,A}(0) + w_{D,A}(-1) & \rightarrow w_{D,A} = 1 \\
0 = w_{A,A}(0) + w_{B,A}(0) + w_{C,A}(0) + w_{D,A}(0)
\end{cases}
$$

DはAをプラスに制御

問題点

- 例題の時系列データ
  - 4遺伝子×5 time points
  - ネットワークが解けた！



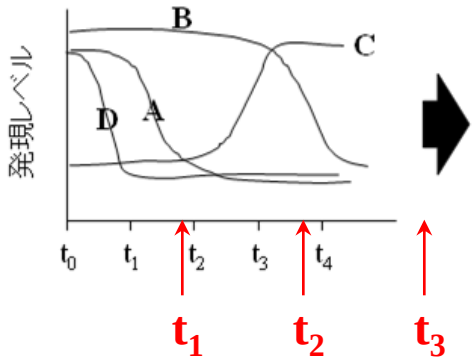


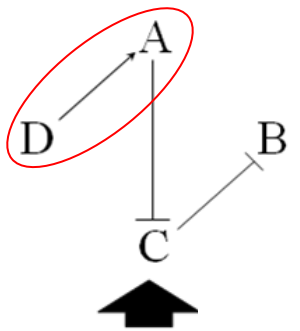
Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

- 一般論
  - N個の遺伝子間相互作用の可能性はN^2通り存在する→N^2個の未知のパラメータ（重み係数$w_{i,k}$）を一意に求めるためには、最低でもN^2個の線形独立な方程式が必要
  - （例題のように）時点数 > 遺伝子数であれば…

問題点

- 次元の問題(劣決定性の問題)
  - 理想:遺伝子数 ≦ 時点数
  - 現実:遺伝子数 >> 時点数
  - 例:「数万遺伝子×(せいぜい)数十時点」のデータ
    →N^2個あるパラメータを解くための方程式が足りない!
    (解が多数得られてしまう…)
- 時間解像度の問題
  - 相互作用イベントの起こる順番を**明確に分離できる時点間隔**となっているか?





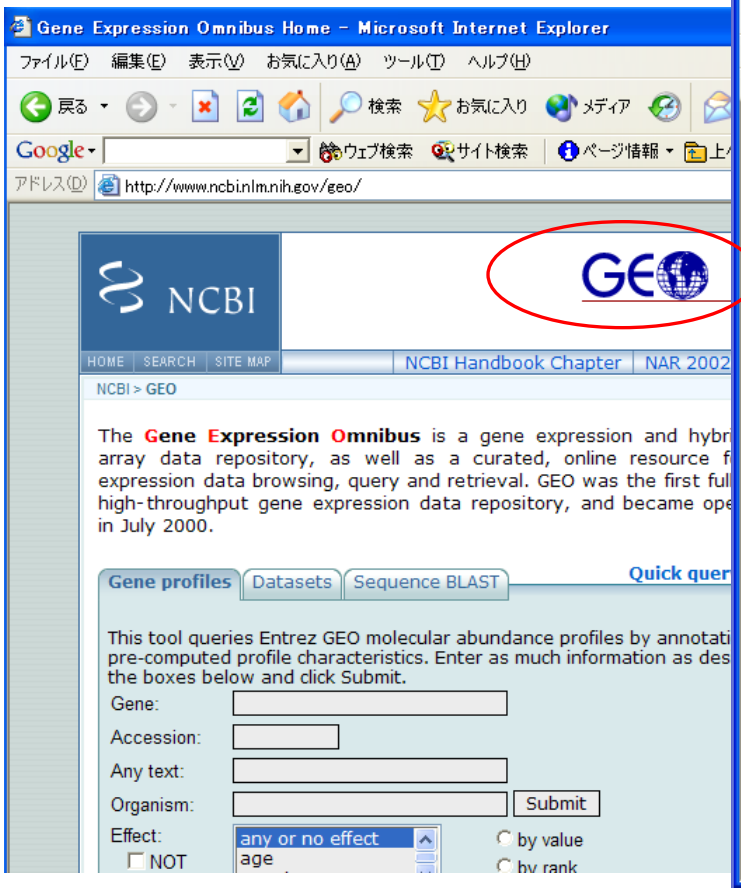
Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

遺伝子の発現制御ネットワーク推定

- 閾値検定モデル
  - 発現量の変化から2遺伝子間の制御関係を推定
- Bayesianモデル
  - 実験データから作り出す条件付き確率分布から 推定
  - Imoto *et al.*, *J. Bioinform. Comput. Biol.*, **1**, 231-252, 2003.
- S-systemモデル
  - 複数の遺伝子が関与する発現プロファイルの時系列データをもとに推定
  - Tominaga *et al.*, *J. Bioinform. Comput. Biol.*, **4**, 503-514, 2006.
- 多階層有向グラフモデル
  - 2遺伝子間の関係から遺伝子のグループ化と制御関係を推定

トランスクリプトームデータベース

- 多くの遺伝子発現データは公共データベースに格納されている



トランスクリプトームデータへ

- 用途
  - 検証
  - 異なる解析手法で再解析
  - 異なる視点で再解析
  - …
- DNAマイクロアレイ以外のデータも格納されている
  - 例）SAGEデータなど

GEO ID: GSM26978

#TAG = tag sequence
#COUNT = count

TAG	COUNT
AAAAAAAAAAAAAAAAA	35
AAAAAAAAAAAAAAAAC	4
AAAAAAAAAAAAAAAAT	2
AAAAAAAAAAAAGACTTG	1
AAAAAAAAAGGGTCAAA	1
AAAAAAAAATGGGTTCA	3
AAAAAAAATGGGTTAAT	1
AAAAAAAATGGGTTCAG	1
AAAAAAACTTCTTTCTA	1
AAAAAAAGAAGAAGAAG	1
AAAAAAATAAAAATCCC	3
AAAAAAATAGTCAATAA	1
AAAAAAATTTTGTAAAC	1
AAAAAACGAAGAAGAAG	1
AAAAAACGTTTCTTCCT	1
AAAAAAGATTTATTTTG	1
AAAAAAGCTGTAGAGAA	1
AAAAAAGGCCGTTTTCC	1
AAAAAAGGCGTTTTTGT	1
AAAAAAGTAAAGGGCCA	1
AAAAAATATCAGTCAAG	1
AAAAAATATCGGTCAAG	5
AAAAAATGTTGCCAGGA	1
AAAAAATTGAGGCACTC	5
AAAAAATTGAGGCATTC	1

“AAAAAATATCGGTCAAG”という配列が5回sequenceされた

他のトランスクリプトーム解析技術

- 配列断片タグのsequenceに基づく方法
  - Expressed Sequence Tags (ESTs)
  - Serial Analysis of Gene Expression (SAGE)
    - long SAGE
  - Massively Parallel Signature Sequencing (MPSS)
  - Cap Analysis Gene Expression (CAGE)
- PCR＋電気泳動に基づく方法
  - Amplified Fragment Length Polymorphism (AFLP)
    - Introduced AFLP (iAFLP)
    - High-coverage expression profiling (HiCEP)
  - Differential Display (DD)

他のトランスクリプトーム解析技術

- 配列断片タグのsequenceに基づく方法
  - 沢山発現している遺伝子の配列断片はより多くsequenceされる

    →sequenceされた回数がその遺伝子の発現レベルそのもの

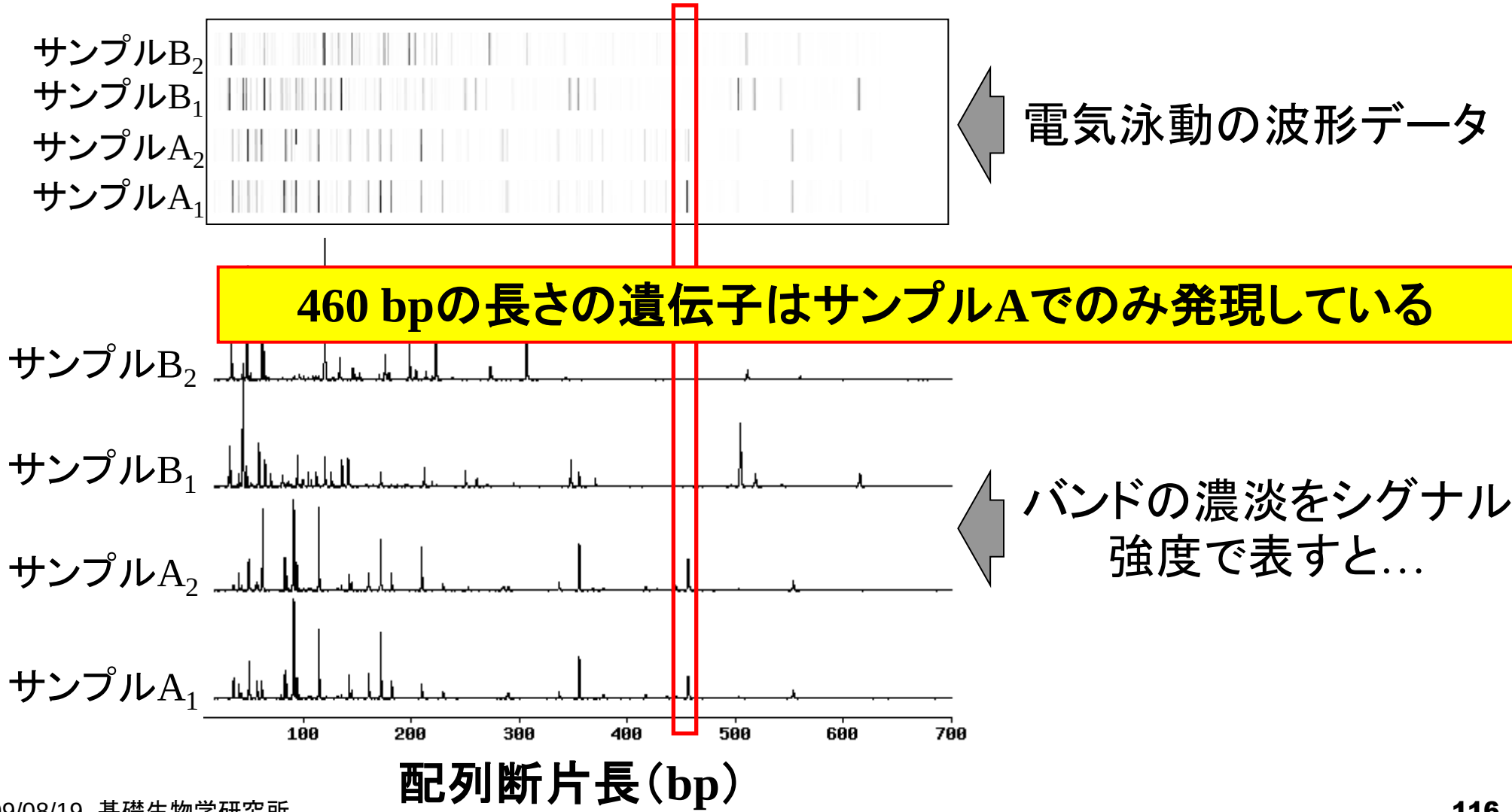
#TAG = tag sequence
#COUNT = count

TAG	COUNT
AAAAAAAAAAAAAAAAA	35
AAAAAAAAAAAAAAAAC	4
AAAAAAAAAAAAAAAAT	2
AAAAAAAAAAAGACTTG	1
AAAAAAAAAGGGTCAAA	1
AAAAAAAAATGGGTTCA	3
AAAAAAAATGGGTTAAT	1
AAAAAAAATGGGTTCAG	1
AAAAAAACTTCTTTCTA	1
AAAAAAAGAAGAAGAAG	1
AAAAAAATAAAAATCCC	3
AAAAAAATAGTCAATAA	1
AAAAAAATTTTGTAAAC	1
AAAAAACGAAGAAGAAG	1
AAAAAACGTTTCTTCCT	1
AAAAAAGATTTATTTTG	1
AAAAAAGCTGTAGAGAA	1
AAAAAAGGCCGTTTTCC	1
AAAAAAGGCGTTTTTGT	1
AAAAAAGTAAAGGGCCA	1
AAAAAATATCAGTCAAG	1
AAAAAATATCGGTCAAG	5
AAAAAATGTTGCCAGGA	1
AAAAAATTGAGGCACTC	5
AAAAAATTGAGGCATTC	1

“AAAAAAATATCGGTCAAG”という配列が5回sequenceされた

他のトランスクリプトーム解析技術

- **PCR＋電気泳動に基づく方法**



様々なトランスクリプトーム解析技術

- 特徴（解析対象の広さ）
  - 目的生物種のDNAマイクロアレイが用意されていないものは解析不可能

    例）バクテリア、柿、桃などのマイクロアレイはない
  - マイクロアレイがあったとしても、未知遺伝子の解析はできない（アレイ上に搭載されていないため）

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

- 特徴（アノテーション情報）
  - 配列断片タグ（△）
    - 目的の配列情報は分かるが、その遺伝子名などはBlastサーチなどを行う必要性あり
    - 配列長が短いため、候補遺伝子群の中からの特定が難しい

```
AAAAAATAGCCTAGAGA   1
AAAAAATAGTCAATAAA   1
AAAAAATATCAGTCAAG   1
AAAAAATATCGGTCAAG   5
AAAAAATGTTGCCAGGA   1
AAAAAATTGAGGCACTC   5
AAAAAATTGAGGCATTC   1
```

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

- 特徴（アノテーション情報）
  - PCR＋電気泳動（×）
    - 目的遺伝子の塩基配列情報を得る作業が（配列断片タグに比べて）余分に必要
      - バンドの切り出し
      - 抽出、PCR増幅
      - クローニング（塩基配列決定）
    - 得られた塩基配列をもとにBlastサーチ

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

- 特徴（データ解析の簡便さ）
  - 配列断片タグ（△）
    - Sequenceコストがかかるため、それほど多くのsequenceができるわけではない

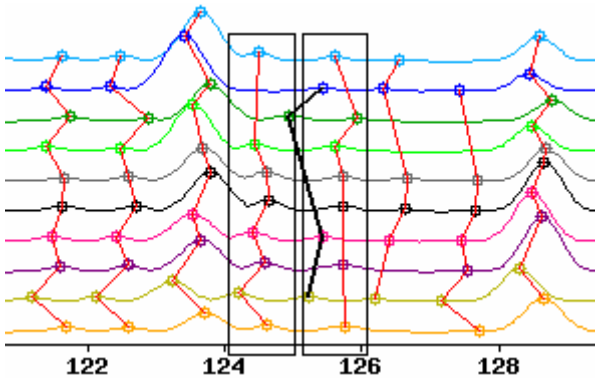
      →統計的なデータ解析が難しい

```
AAAAAATAGCCTAGAGA   1
AAAAAATAGTCAATAAA   1
AAAAAATATCAGTCAAG   1
AAAAAATATCGGTCAAG   5
AAAAAATGTTGCCAGGA   1
AAAAAATTGAGGCACTC   5
AAAAAATTGAGGCATTC   1
```

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

- 特徴（データ解析の簡便さ）
  - PCR＋電気泳動（△）
    - ピークアラインメント（同一遺伝子の認識）が難しい



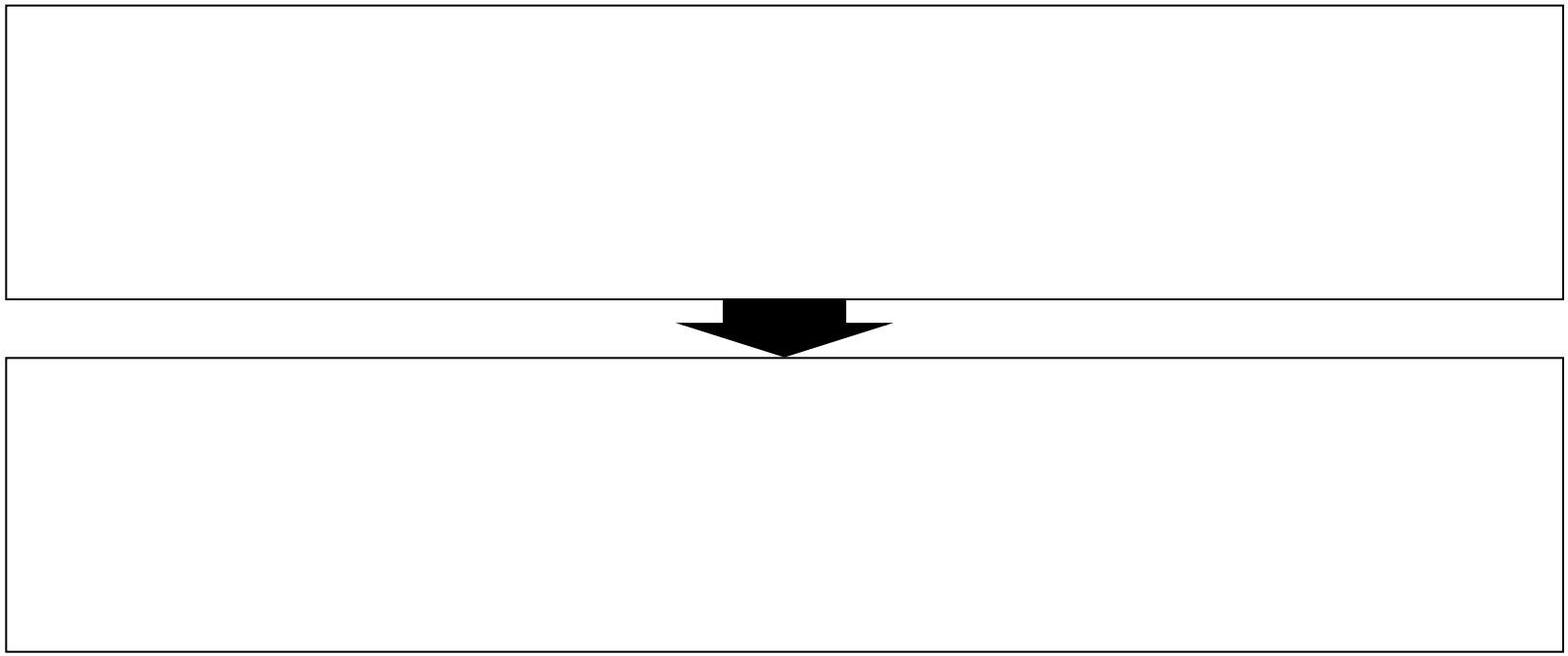
	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

- 改良に向けた取り組み：マイクロアレイ
  - 短所：マイクロアレイがあったとしても、未知遺伝子の解析はできない（アレイ上に搭載されていないため）

  →タイリングアレイの開発により、未知遺伝子の発現も検出可能に

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△



「タンパク質をコードする遺伝子」の解析から「ゲノム全体」の発現解析へ

様々なトランスクリプトーム解析技術

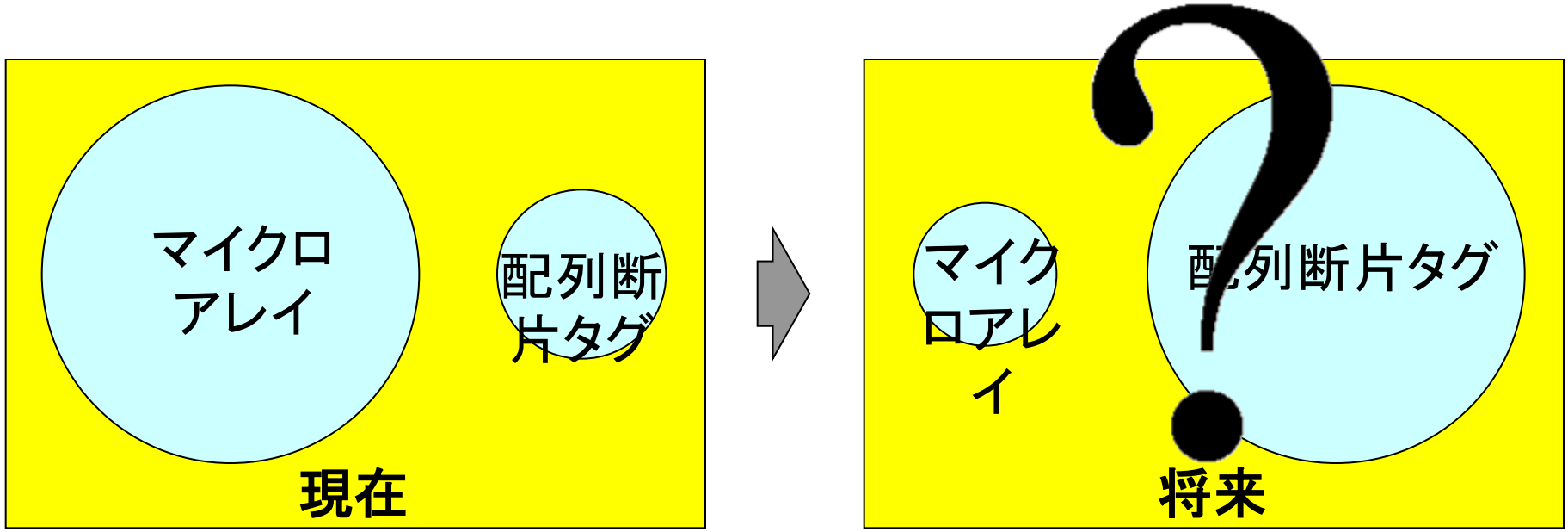
- タイリングアレイによる具体的な成果
  - ヒト21,22番染色体の解析により、従来よりはるかに多くの転写物が存在することを確認（Kapranov *et al.*, *Science*, 2002）
  - シロイヌナズナの解析により、既知の約27,000遺伝子領域以外に約5,200の領域で発現している新たな遺伝子構造を発見（Toyoda *et al.*, *Plant J.*, 2005）
  - 次期ヒトゲノム計画（ENCODE計画）でも採用され、ゲノム中の大部分の塩基が、タンパク質をコードしない転写産物や重複転写産物を含む、一次転写産物になることが示唆（The ENCODE　Project Consortium, *Nature*, 2007）

様々なトランスクリプトーム解析技術

- 改良に向けた取り組み：配列断片タグ
  - 短所： Sequenceコストがかかるため、それほど多くのsequenceができるわけではない。そのため、統計的なデータ解析が難しい

→次世代シーケンサーの開発によりコストを大幅に削減可能に

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△



次世代シーケンサー

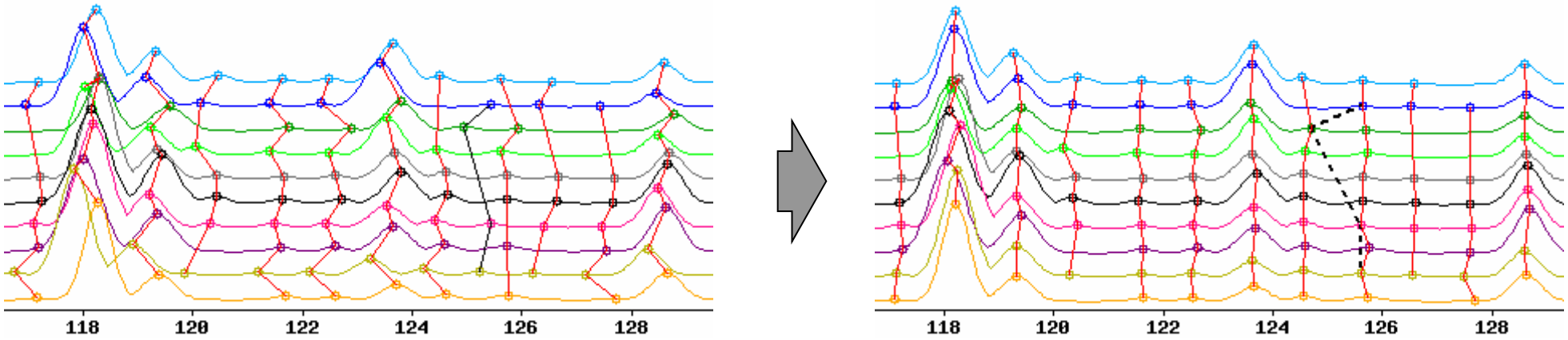
- パンダ（大熊猫）ゲノム解読（2008/10）
  - ヒトゲノム解読に10年 → 半年
  - 猫よりも犬・熊に近い動物
- アジア人（中国人）一個体の全ゲノム配列決定（2008/11/6, *Nature*）
  - 36倍のカバー率
  - 個人ゲノムとしてはJ.D. WatsonとJ.C. Venterに次いで3人目
- 2010年ごろ発売される予定のものは、ヒトゲノムを8分程度で解読できるらしい（Levene *et al*. *Science*, 2003）
- 国際プロジェクト
  - 1000人ゲノム計画（1人1人の遺伝情報の違いを詳細に調査）
  - 国際癌ゲノムプロジェクト
  - 感染症の同定

他のトランスクリプトーム解析技術

- 改良に向けた取り組み：PCR+電気泳動
  - 短所：ピークアラインメント（同一遺伝子の認識）が難しい

→バイオインフォマティクス手法の適用によるアラインメント精度の大幅な向上

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△



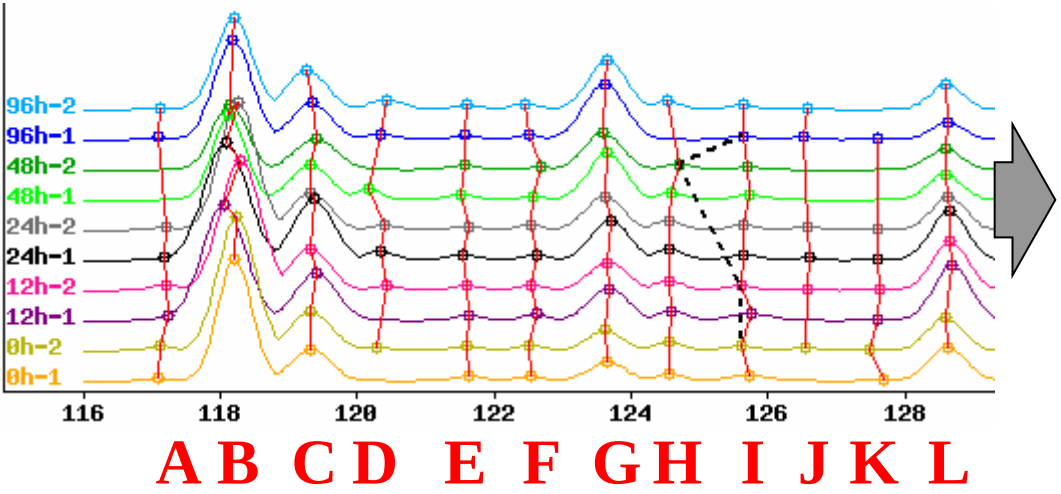
実験技術の開発も重要だがバイオインフォマティクス（解析手法の開発）も重要

Kadota *et al.*, *BMC Bioinformatics*, 2005
Kadota *et al.*, *AMB*, 2007

他のトランスクリプトーム解析技術

	マイクロアレイ	配列断片タグ	PCR+電気泳動
解析対象の広さ	△	○	○
遺伝子に関する情報の量（アノテーション情報）	○	△	×
データ解析の簡便さ	○	△	△

- アラインメント精度の大幅な向上により、正確な遺伝子発現行列を作成可能になった



遺伝子発現行列

	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**33**	**41**	**47**	**49**	**55**	**48**	**34**	**27**		**43**
I	**28**	**24**	**34**	**25**	**30**	**27**	**25**	**14**	**18**	**23**
J		16		8	13	17			14	7
K	7	9	8	9	8	9			4	
L	168	163	275	242	246	165	126	102	85	123
M						10				
N	30	34	54	49	68	52	25	11	33	31

マイクロアレイ解析用に開発された手法が電気泳動波形データ解析にも利用可能

まとめ

- 様々なマイクロアレイ解析手法を紹介

  二群、多群、時系列、クラスタリング、分類、ネットワーク
- マイクロアレイの位置づけ

  長所（解析が容易）、短所（搭載されてない遺伝子など）
- 他の解析技術によって得られたトランスクリプトームデータへの適用可能性

  遺伝子発現行列さえできれば次世代シーケンサーもOK
- 「**（Rで）マイクロアレイデータ解析**」でお幸せに

  よりよい手法を、よりお求めやすく