トランスクリプトーム データの解析戦略とその周辺

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）
http://www.iu.a.u-tokyo.ac.jp/~kadota/
kadota@iu.a.u-tokyo.ac.jp

オーム（Ome）研究

ome：総体

DNA

遺伝子 (Gene) + ome → Genome

ゲノム研究：ヒトのもつ遺伝子情報の総体を研究

RNA

転写 (Transcription) + ome → Transcriptome

転写：遺伝子DNAの情報をRNAに写すこと

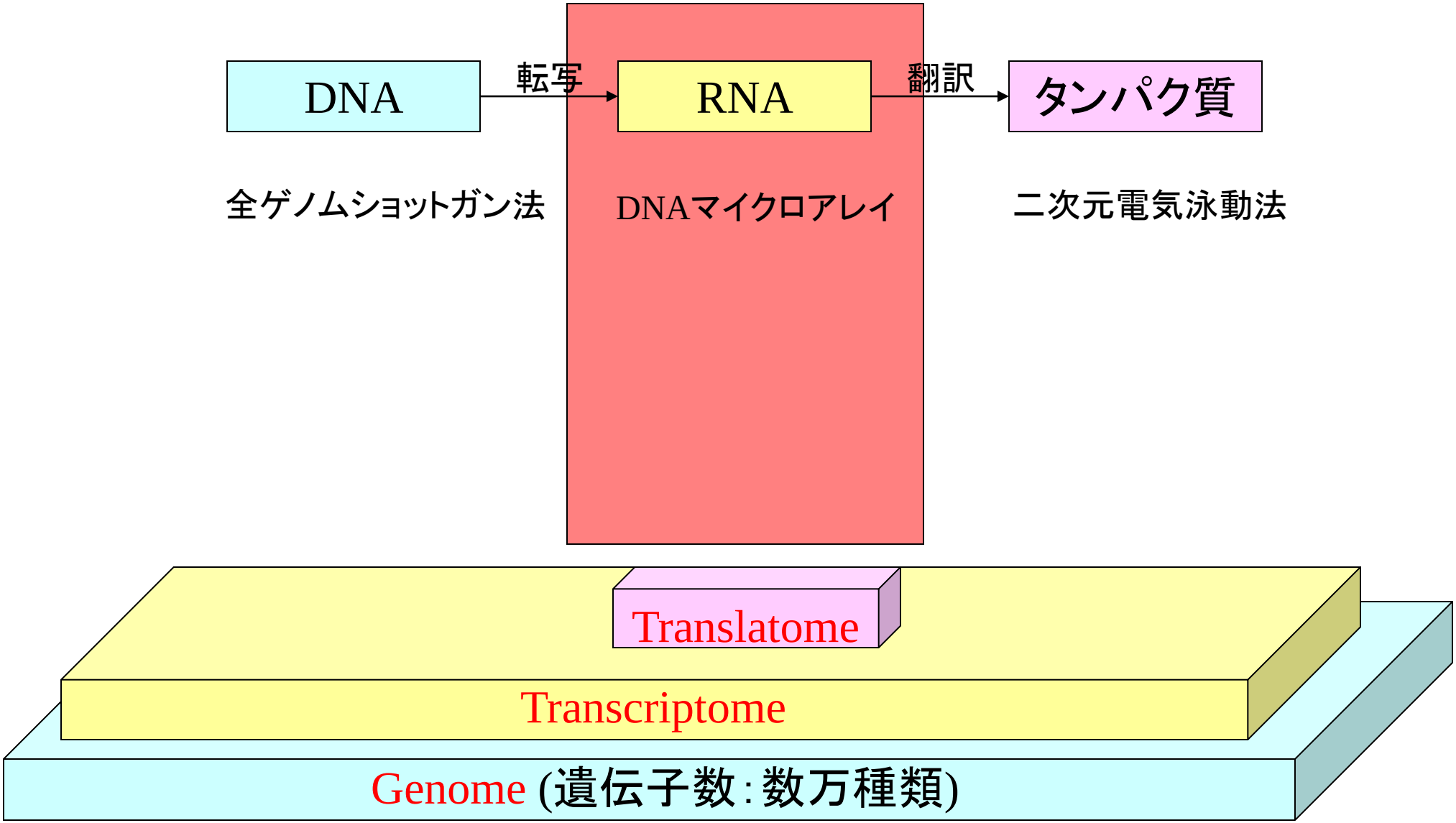
タンパク質

翻訳 (Translation) + ome → Translatome

翻訳：転写されたRNA情報をもとにタンパクを作ること

（タンパク質 (Protein) + ome → Proteome）

転写レベルの情報量は豊富



トランスクリプトームとは

- ある特定の状態の組織や細胞中に存在する全mRNA（ないしは一次転写産物、transcripts）の総体
- 様々なトランスクリプトーム解析技術
  - マイクロアレイ
    - cDNAマイクロアレイ、Affymetrix GeneChipなど
  - 配列決定に基づく方法
    - EST、SAGEなど
  - 電気泳動に基づく方法
    - Differential Display、AFLPなど

調べたい組織でどの遺伝子がどの程度発現しているのかを一度に観察

内容

- 様々なトランスクリプトーム解析技術
  - 概要、特徴、長所短所
  - 全て共通の“遺伝子発現行列”形式で取り扱いが可能
- “遺伝子発現行列”データ解析戦略
  - 発現変動遺伝子の同定
  - Gene Set Enrichment Analysis
  - クラスタリング
  - 分類
  - ネットワーク推定

トランスクリプトーム解析技術1

- マイクロアレイ
  - 配列既知遺伝子を搭載した“チップ”上に、調べたいサンプルから抽出・合成した蛍光標識済みcDNAをハイブリダイゼーションさせることによって、得られる蛍光シグナル強度をmRNAの発現量として観測
  - 比較する条件間で発現の異なる遺伝子の同定などの目的に利用される
  - ゲノム配列決定済みの生物種を対象

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
…	…

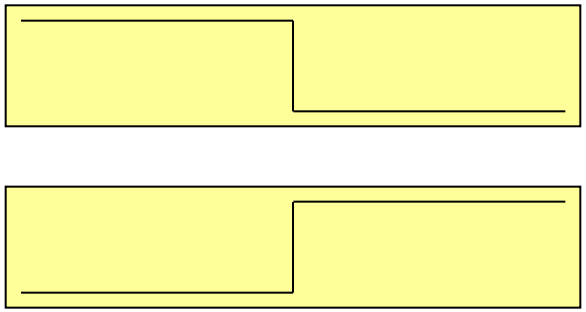
目的試料中の遺伝子4の発現レベルは対照試料に比べて2^{-2}倍高い

目的試料中で遺伝子3は沢山発現している

遺伝子発現行列

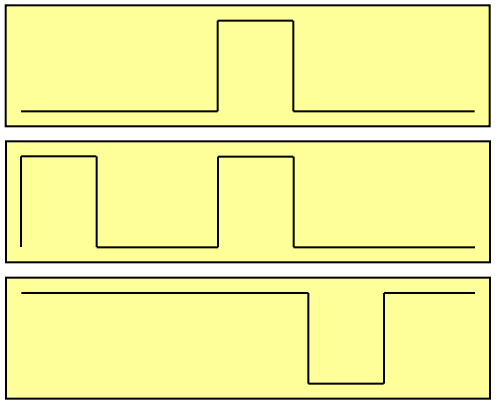
二群間比較

	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*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



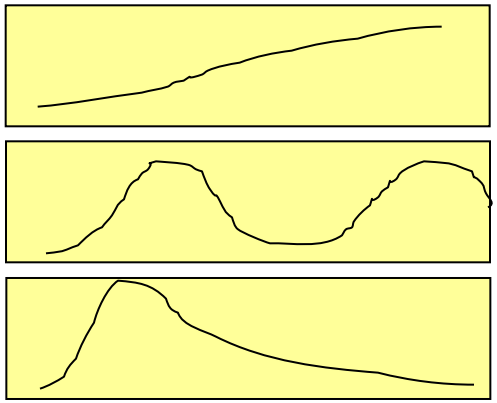
様々な組織（条件）

	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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



- 発現変動遺伝子の同定
- クラスタリング
- Gene Ontology解析
- パスウェイ解析

トランスクリプトーム解析技術2

- 配列決定に基づく方法
  - 調べたい目的サンプルから抽出・合成したcDNAの一部をsequencerで読みまくる
  - その配列をもつ転写物が沢山発現しているほど、その配列が多数読まれることを利用
    - EST（Expressed Sequence Tag）
      - 3’ or 5’側から数百塩基程度の配列を読んだもの。
    - SAGE（Serial Analysis of Gene Expression）
      - 特定の位置から数十塩基の配列（SAGEタグ）を分離し、他の転写物由来のタグをsequencerで読める程度まで連結して配列決定
  - ゲノム配列未知のサンプルを対象
    - 新規遺伝子の発見が原理的に可能

得られる発現データのイメージ

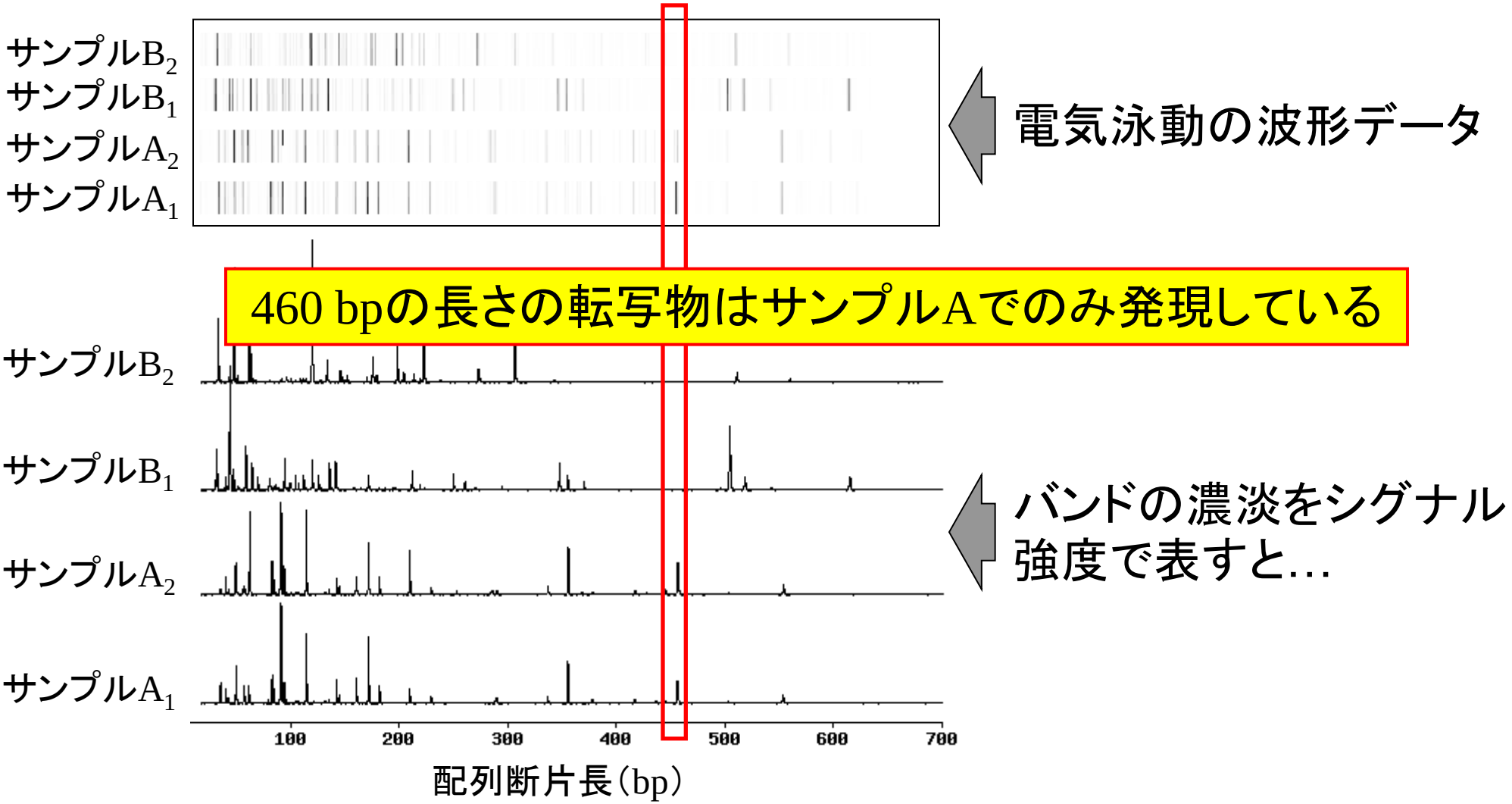
```
#TAG = tag sequence
#COUNT = count
TAG    COUNT
AAAAAAAAAAAAAAAAA   35
AAAAAAAAAAAAAAAAC   4
AAAAAAAAAAAAAAAAT   2
AAAAAAAAAAAAGACTTG  1
AAAAAAAAAAGGGTCAAA  1
AAAAAAAAATGGGTTCA   3
AAAAAAAAATGGGTTAAT  1
AAAAAAAAATGGGTTCAG  1
AAAAAAACTTCTTTCTA   1
AAAAAAAGAAGAAGAAG   1
AAAAAAATAAAAATCCC   3
AAAAAAATAGTCAATAA   1
AAAAAAATTTTGTAAAC   1
AAAAAACGAAGAAGAAG   1
AAAAAACGTTTCTTCCT   1
AAAAAAGATTTATTTTG   1
AAAAAAGCTGTAGAGAA   1
AAAAAAGGCCGTTTTCC   1
AAAAAAGGCGTTTTTGT   1
AAAAAAGTAAAGGGCCA   1
AAAAAATATCAGTCAAG   1
AAAAAATATCGGTCAAG   5
```

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

トランスクリプトーム解析技術3

- 電気泳動に基づく方法
  - 目的サンプルから得られた転写物由来DNA配列（断片）をPCR増幅して電気泳動にかける
  - サンプルの状態によって得られる電気泳動パターンが異なる（フィンガープリント）。得られるバンドの濃さ（シグナル強度の高さ）が転写物の発現レベルに（大まかに）対応。
  - ゲノム配列未知のサンプルを対象
    - 新規遺伝子の発見が原理的に可能

得られる発現データのイメージ



長所・短所

- 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、アレイ上に搭載されていない（未知）遺伝子の発現は観測不可能

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

長所・短所

- アノテーション情報
  - 配列決定（△）
    - 目的の配列情報をもとにBlast検索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	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

長所・短所

- アノテーション情報
  - 電気泳動（×）
    - 目的遺伝子の塩基配列情報を得る作業が（配列決定に基づく方法に比べて）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Blast検索

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

長所・短所

- データ解析の簡便さ
  - 配列決定（△）
    - 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

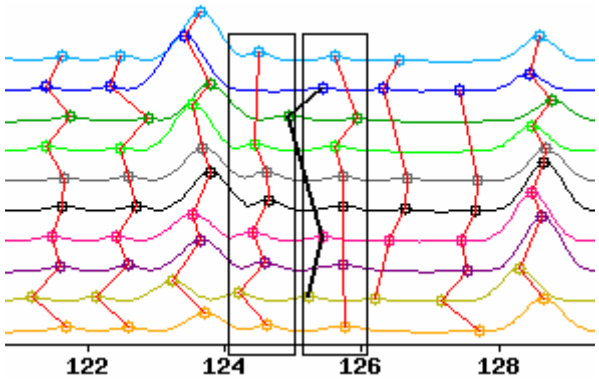
      →統計的なデータ解析が難しい

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

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

長所・短所

- データ解析の簡便さ
  - 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



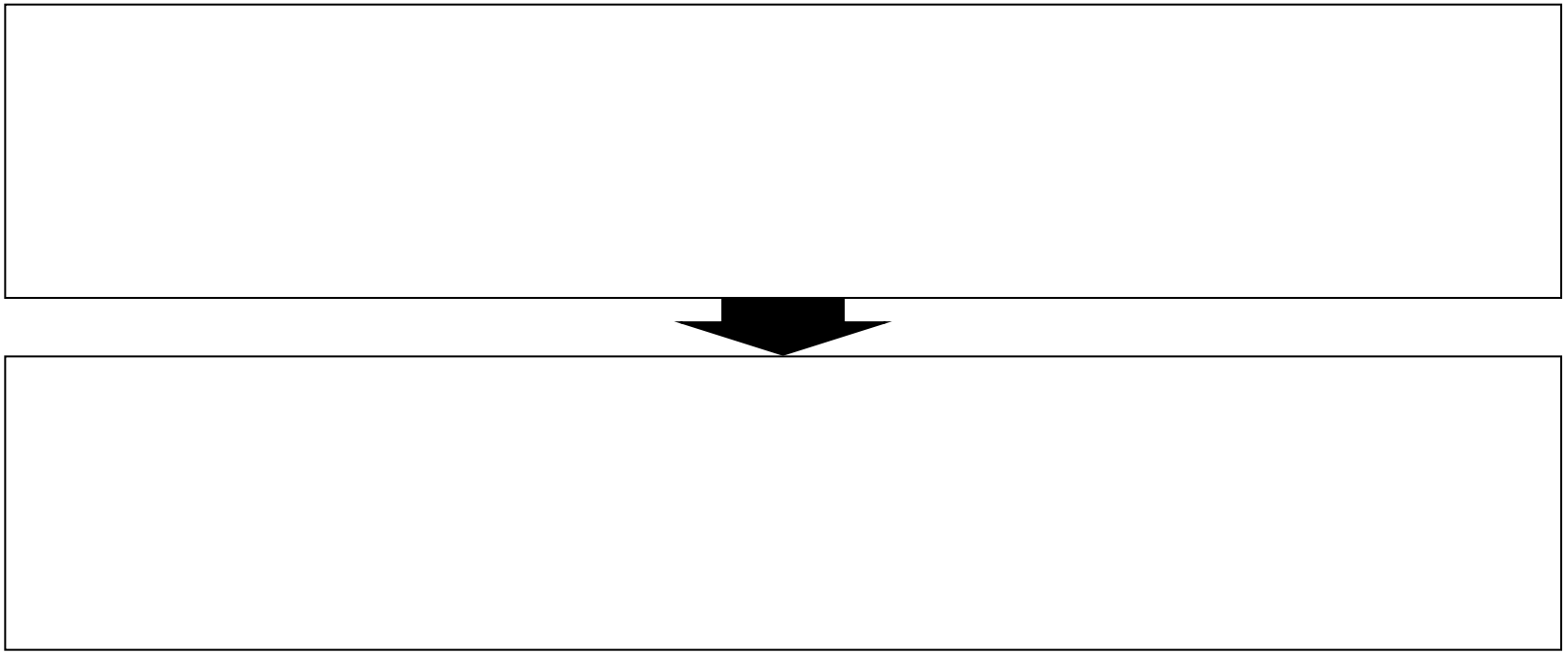
	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

他のトランスクリプトーム解析技術

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

- 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、アレイ上に搭載されていない（未知）遺伝子の発現は観測不可能

→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



「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

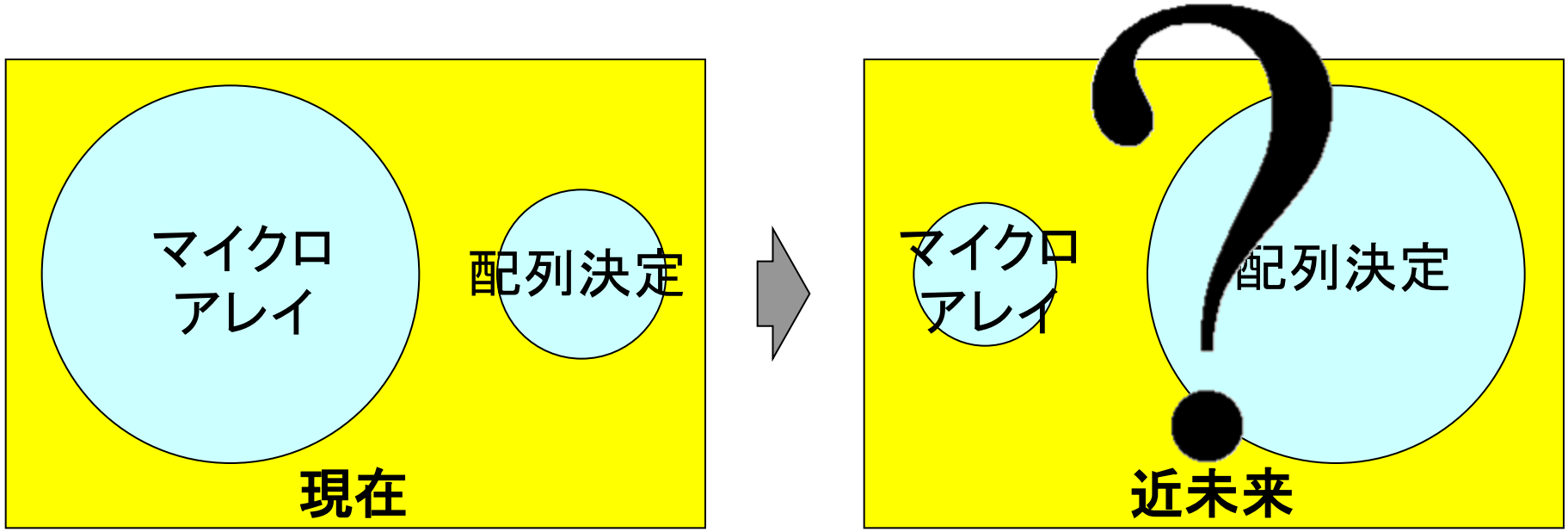
- タイリングアレイによる具体的な成果
  - ヒト21,22番染色体の解析により、従来よりはるかに多くの転写物が存在することを確認（Kapranov *et al.*, *Science*, 2002）
  - シロイヌナズナの解析により、既知の約27,000遺伝子領域以外に約5,200の領域で発現している新たな遺伝子構造を発見（Toyoda *et al.*, *Plant J.*, 2005）
  - 次期ヒトゲノム計画（ENCODE計画）でも採用され、ゲノム中の大部分の塩基が、タンパク質をコードしない転写産物や重複転写産物を含む、一次転写産物になることが示唆（The ENCODE　Project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

  →新型（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	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△

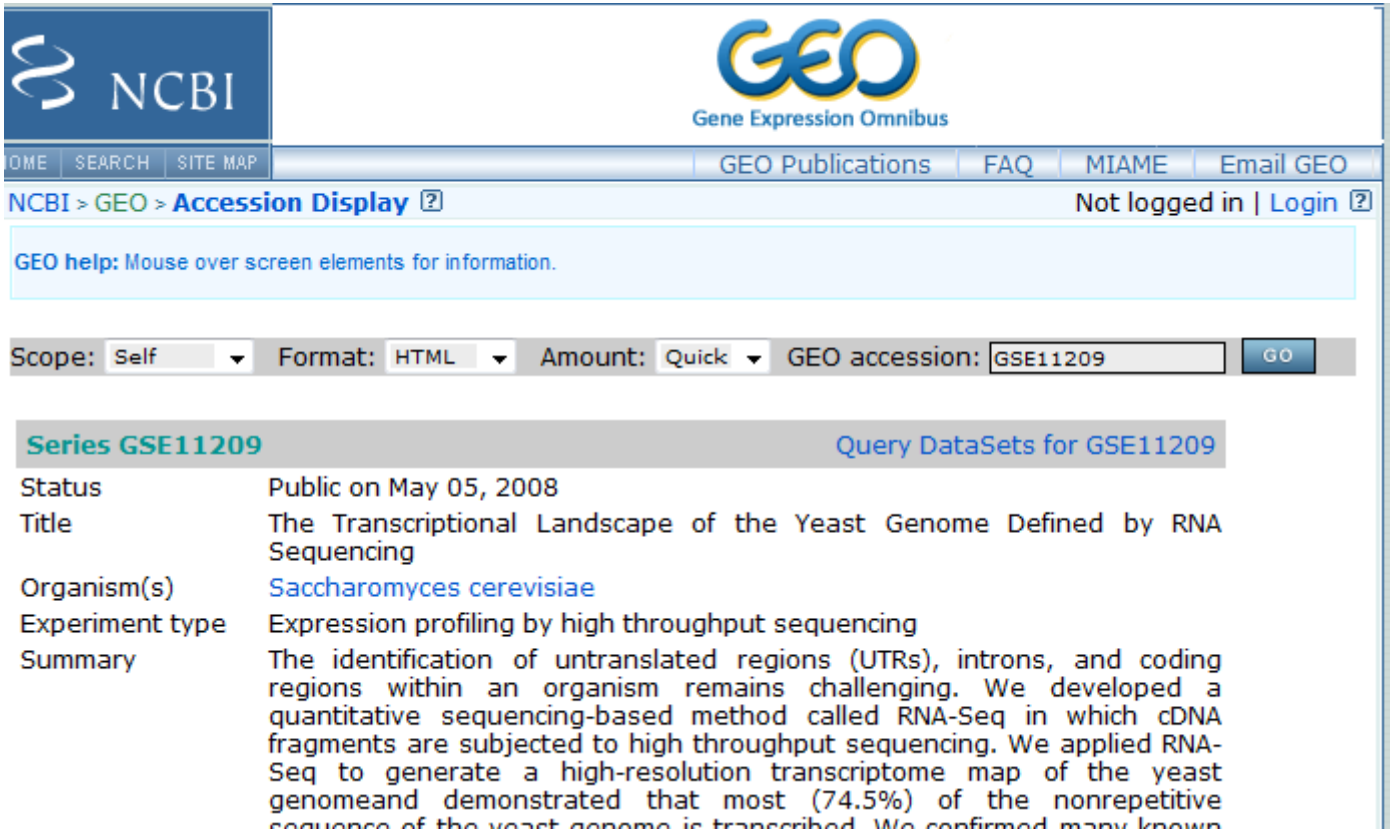


新型（次世代）シーケンサー

- 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
- 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

トランスクリプトーム解析例

- 出芽酵母のトランスクリプトームの全体像
  - Nagalakshmi *et al*., *Science*, **320**, 1344-1349, 2008.
  - polyA RNAのトランスクリプトームデータ（RNA-seq）
  - Illumina社の平均35bpの塩基配列
  - 公共遺伝子発現データベース（GEO）に登録済
    - GSE11209



トランスクリプトームデータ解析戦略

- ゲノム配列へのマッピング

ゲノム配列

- 新規転写物の同定
- Untranslated region (UTR)の同定
- 予測されていたイントロンの確認
- 選択的開始コドンの同定

etc...

GEO ID: GSM282598

```
...
+SRR002059.1740 :7:1:446:160 length=33
IIIIIIIIIIIIAIBI6IIAII?=$66%5.)%/
@SRR002059.1741 :7:1:883:724 length=33
ATTAAACAAAAATATTATAATTAGGAAATATTT
+SRR002059.1741 :7:1:883:724 length=33
IIIIIIIIIIIIIIIIIIIIIII'@IIIIII
@SRR002059.1742 :7:1:568:594 length=33
TCGGAAGAGCTCGTATGCCGTCTTCTGCTTTCA
+SRR002059.1742 :7:1:568:594 length=33
IIIIIIIIIIIIIIIIIIIIII@IEIII,"8
@SRR002059.1743 :7:1:845:772 length=33
ATTTTTATATGAATGAAACGCCTATGGATATAT
+SRR002059.1743 :7:1:845:772 length=33
IIIIIIIIIIIIIIIII?GIIIIII<IIBICI
@SRR002059.1744 :7:1:303:168 length=33
TACTTGCCAAACTACGATGACATGAGACACTAT
...
```

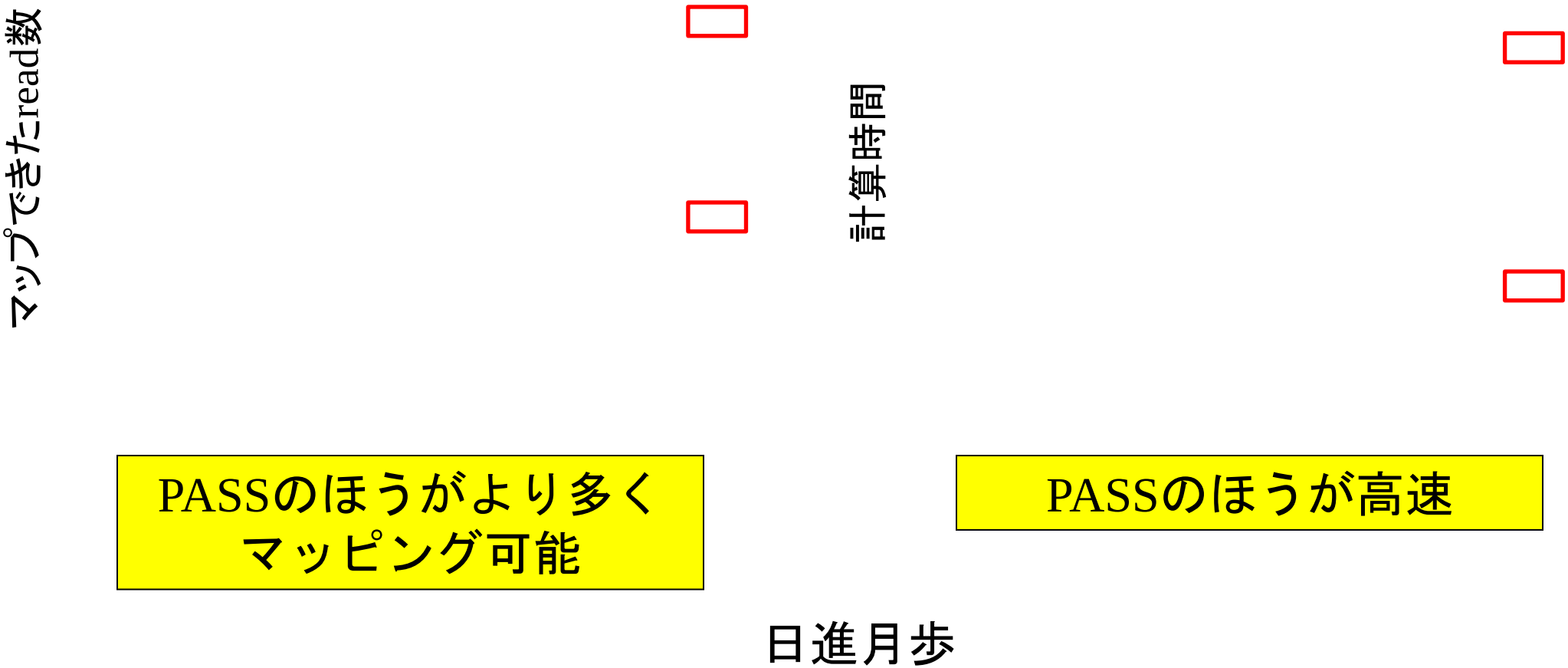
大量の短い配列（short read）をいかに正しく高速にゲノム配列にマップするか？

トランスクリプトームデータ解析戦略

- 「“大量の短い配列”を“一つのゲノム配列”」にマップするための専用のアルゴリズム開発の必要性
  - BLAST(Altschul et al., 1997)などは非現実的
    「“単一のクエリ配列”を“多数の配列データ”」に問い合わせることを想定
  - BLAT(Kent 2002)なども非現実的
    「“大量のそこそこ長い配列”を“一つのゲノム配列”」にマップすることを想定
- 新型シーケンサーデータ解析専用アルゴリズム
  - PatMaN (Prufer *et al.*, *Bioinformatics*, 2008)
  - RMAP (Smith *et al.*, B*MC Bioinformatics*, 2008)
  - MAQ (Li *et al.*, *Genome Res.*, 2008)
  - SeqMap (Jiang and Wong, *Bioinformatics*, 2008)
  - SOAP (Li *et al.*, *Bioinformatics*, 2008)
  - PASS (Campagna *et al.*, *Bioinformatics*, 2009)
  - SOAP2 (Li *et al.*, *Bioinformatics*, 2009)

どのアルゴリズムを採用するか？

- PASS(P) vs. SOAP(S)



解析技術（実験側）も日進月歩

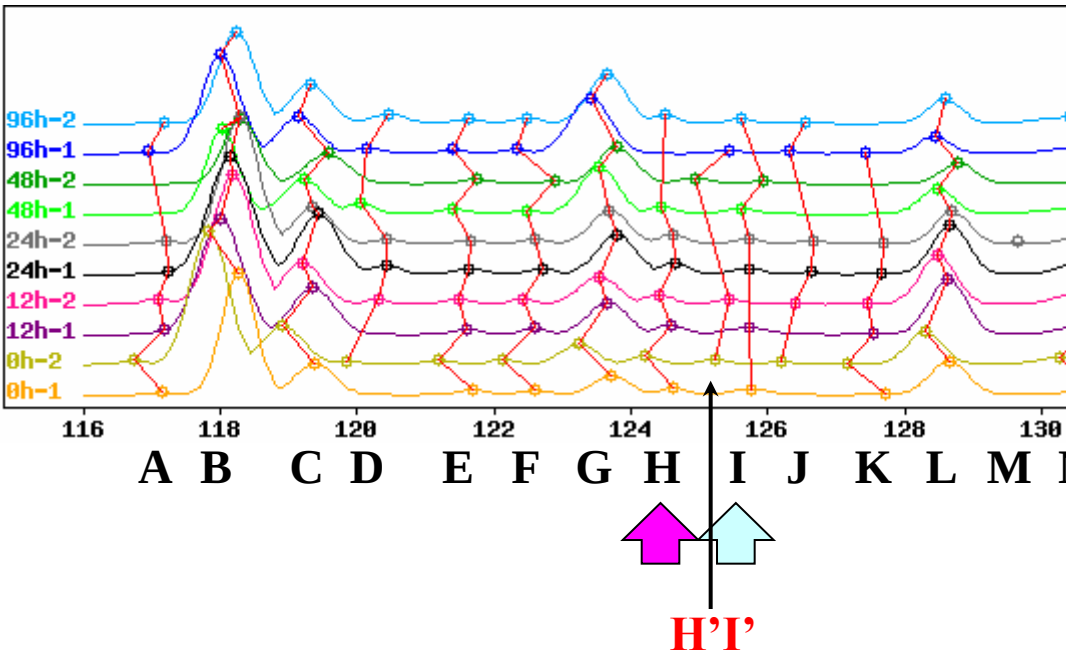
- 現在のマイクロアレイや市販の新型シーケンサーは、逆転写酵素を用いたcDNA合成などいくつかのステップを経る必要があるためバイアスが入り込む恐れがあった。

- RNAを直接配列決定する方法の開発（2009年10月）
  - Ozsolak *et al.*, *Nature*, **461**, 814-818, 2009
  - RNA分子1個の塩基配列を（cDNA合成などのステップを挟まずに）直接決定
  - 今後のスケールアップにより、バイアスのないハイスループットトランスクリプトーム解析法になりうると期待

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：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

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△



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**			**43**
H'I'		**24**		**25**				**27**	**18**	
I	**28**		**34**		**30**	**27**	**25**	**14**		**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

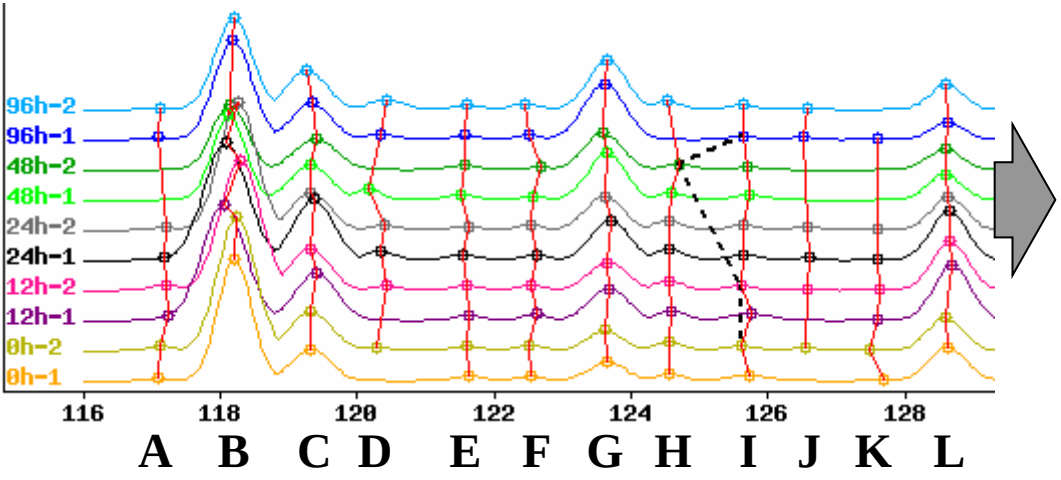
実験技術の開発も重要だがバイオインフォマティクス（解析手法の開発）も重要

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

様々なトランスクリプトーム解析技術

- バイオインフォマティクス技術の適用によりアラインメント精度の大幅な向上を達成

	マイクロアレイ	配列決定	電気泳動
解析対象の広さ	△	○	○
アノテーション情報	○	△	×
データ解析の簡便さ	○	△	△



遺伝子発現行列

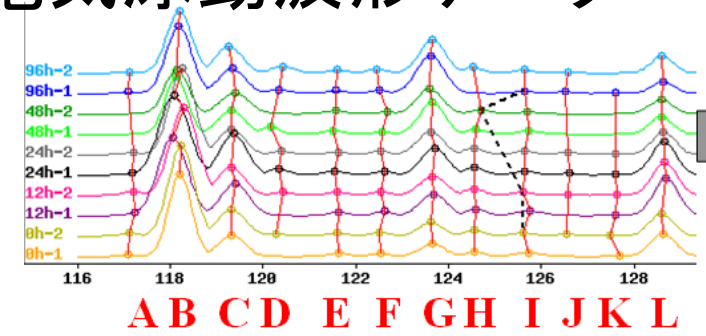
	0h-1	0h-2	12h-1	12h-2	24h-1	24h-2	48h-1	48h-2	96h-1	96h-2
A	11	23	21	29	13	15			10	9
B	607	664	576	649	582	634	421	326	491	456
C	156	191	233	209	301	186	172	151	181	195
D		19		24	44	25	53		27	48
E	23	21	28	25	28	19	24	22	20	21
F	21	25	30	26	28	26	19	15	24	29
G	93	100	160	139	196	166	234	184	276	245
H	**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

Clustering-based peak alignment計算例

全てのトランスクリプトームデータは

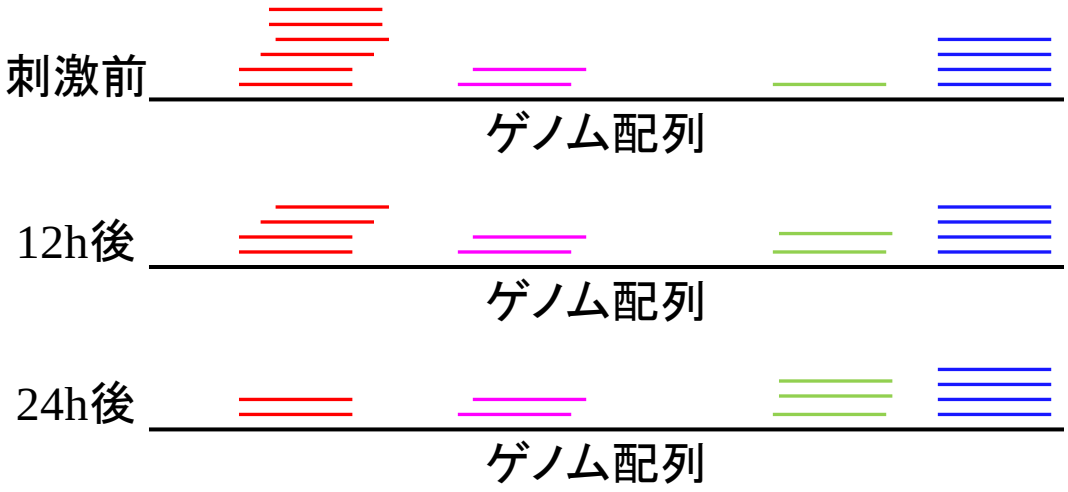
- “遺伝子発現行列”の形式に変換可能
  - 電気泳動波形データ



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

  - 新型シーケンサーデータ



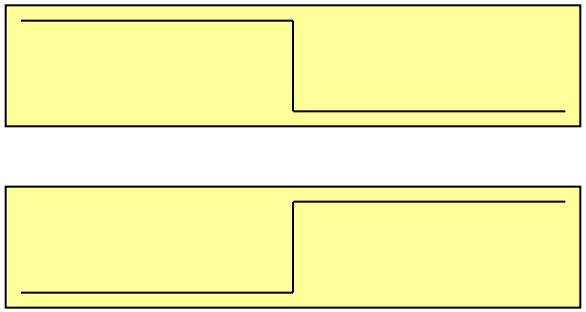
遺伝子発現行列

	刺激前	12h後	24h後	…
の配列をもつ転写物	6	4	2	
の配列をもつ転写物	2	2	2	
の配列をもつ転写物	1	2	3	
の配列をもつ転写物	4	4	4	
…				

様々な遺伝子発現行列

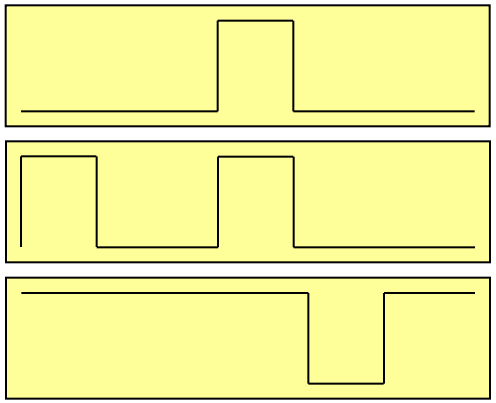
二群間比較

	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*	$x^A_{n,1}$	$x^A_{n,2}$		$x^B_{n,2}$	$x^B_{n,2}$	



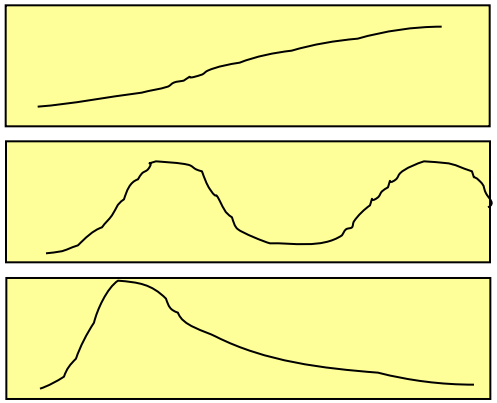
様々な組織（条件）

	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*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…



- 発現変動遺伝子の同定
- クラスタリング
- Gene Ontology解析
- パスウェイ解析

二群間比較解析

- 例）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）

白血病のタイプで発現の
異なる遺伝子群を同定

二群間比較解析戦略

- 「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

二群間比較解析戦略

- *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

参考資料

二群間比較解析戦略

- 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

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

統計量の絶対値が大きい → 候補発現変動遺伝子

二群間比較解析（様々な検出法）

- 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)

個々の遺伝子の発現変動の度合いを調べる研究

二群間比較解析戦略

- 発現変動遺伝子（マーカー遺伝子）の同定
  - 個々の遺伝子について統計量を算出し、ランキング
  - 手法選択のガイドライン（Kadota *et al.*, *AMB*, 2009）
    - 感度・特異度重視の場合
    - 再現性重視の場合
- Gene Set Enrichment Analysis (GSEA)
  - アノテーション情報が豊富な生物種用の解析手段
    - 同じセットに属する遺伝子をひとまとめにして解析
      - 例１：酸化的リン酸化に関係する遺伝子セット（KEGG: hsa00190）
      - 例２：脂肪酸β酸化に関係する遺伝子セット（ GO:0006635）
    - 比較する二群間でその遺伝子セットが動いたかどうかを評価
      - 帰無仮説：動いてない
      - 対立仮説：動いた
    - 沢山の遺伝子セットについて解析を行い、動いた遺伝子セットを列挙
      - positional gene sets
      - pathway gene sets
      - motif gene sets
      - GO gene sets

      etc...

      } 様々な視点での解析が可能

様々な遺伝子セットはMSigDBからゲット

- 例：KEGG Pathway遺伝子セット

	A	B	C	D	E	F	G	H	I	
1	HSA00010_GLYCOLYSIS_AND_GLUCONEOGENESIS	LDHC	LDHB	LDHA	ADH1C	PGAM1	ADH1B	PGAM2	ADH1A	A
2	HSA00020_CITRATE_CYCLE	OGDHL	OGDH	CLYBL	IDH3G	LOC28339	IDH2	IDH1	SUCLA2	F
3	HSA00030_PENTOSE_PHOSPHATE_PATHWAY	ALDOA	TALDO1	ALDOC	ALDOB	PGD	TKTL2	TKTL1	DERA	F
4	HSA00031_INOSITOL_METABOLISM	ALDH6A1	TPI1							
5	HSA00040_PENTOSE_AND_GLUCURONATE_INTERCONVERSION	UGDH	UGT1A7	UGT1A6	UGT1A9	UGT1A8	UGT1A3	UGT1A5	UGT1A4	U
6	HSA00051_FRUCTOSE_AND_MANNOSE_METABOLISM	ALDOA	SORD	PFKFB4	HSD3B7	PFKFB3	ALDOC	PFKFB2	ALDOB	F
7	HSA00052_GALACTOSE_METABOLISM	LALBA	HSD3B7	HK2	HK1	G6PC2	GLB1	GALK2	GALK1	H
8	HSA00053_ASCORBATE_AND_ALDARATE_METABOLISM	ALDH7A1	ALDH1B1	ALDH1A3	MIOX	UGDH	ALDH2	ALDH3A2	ALDH9A1	A
9	HSA00061_FATTY_ACID_BIOSYNTHESIS	OLAH	MCAT	ACACA	FASN	ACACB	OXSM			
10	HSA00062_FATTY_ACID_ELONGATION_IN_MITOCHONDRIA	HSD17B1C	ACAA2	PPT2	ECHS1	PPT1	HSD17B4	HADH	MECR	H
11	HSA00071_FATTY_ACID_METABOLISM	ACOX1	HSD17B1C	ACADSB	CPT2	ADHFE1	EHHADH	ADH5	ADH1C	A
12	HSA00072_SYNTHESIS_AND_DEGRADATION_OF_KETONE_BODI	HMGCS2	OXCT1	HMGCS1	OXCT2	BDH2	ACAT2	BDH1	ACAT1	H
13	HSA00100_BIOSYNTHESIS_OF_STEROIDS	TM7SF2	GGCX	EBP	MVD	CYP51A1	HMGCR	FDPS	LSS	F
14	HSA00120_BILE_ACID_BIOSYNTHESIS	ADHFE1	HSD3B7	ADH5	ADH1C	ADH6	ADH1B	ADH7	ADH1A	A
15	HSA00130_UBIQUINONE_BIOSYNTHESIS	ND1	NDUFB11	ND4	ND5	ND2	ND3	NDUFA13	COQ7	C
16	HSA00140_C21_STEROID_HORMONE_METABOLISM	HSD3B2	CYP17A1	HSD3B1	AKR1C4	CYP11A1	CYP21A2	CYP11B1	CYP11B2	H
17	HSA00150_ANDROGEN_AND_ESTROGEN_METABOLISM	ARSD	ARSE	CYP11B1	CYP11B2	SULT2B1	PRMT3	AKR1C4	PRMT2	F
18	HSA00190_OXIDATIVE_PHOSPHORYLATION	ATP6AP1	NDUFAB1	COX5A	COX5B	ATP8	ATP6	UQCR	COX6C	A
19	HSA00220_UREA_CYCLE_AND_METABOLISM_OF_AMINO_GROUI	SAT1	ALDH18A1	SRM	NAGS	ASS1	SAT2	AGMAT	ASL	A
20	HSA00230_PURINE_METABOLISM	ADCY3	FHIT	ADCY4	GDA	ADCY1	ADCY2	GMPR2	ADCY7	A
21	HSA00232_CAFFEINE_METABOLISM	XDH	CYP2A13	NAT1	NAT2	CYP2A6	CYP2A7	CYP1A2		
22	HSA00240_PYRIMIDINE_METABOLISM	CTPS	DTYMK	CAD	CANT1	PRIM1	NT5M	NT5C3	PRIM2	N
23	HSA00251_GLUTAMATE_METABOLISM	GCLC	GLUD2	GLUD1	GNPNAT1	CAD	QARS	NAGK	GCLM	F

Pathway ID　　Name　　Gene symbols

1行につき1セット

様々なGSEA系の解析手法

- GSEA (Subramanian *et al*., *PNAS*, 2005)
- PAGE (Kim and Volsky, *BMC Bioinformatics*, 2005)
- Hotelling's T^2-test (Kong *et al*., *Bioinformatics*, 2006)
- 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)
- GlobalANCOVA (Hummell *et al*., *Bioinformatics*, 2008)
- …

PAGE法

- 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　$AD^i = \overline{A^i} - \overline{B^i}, \qquad (i = 1,2,...,a)$
  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

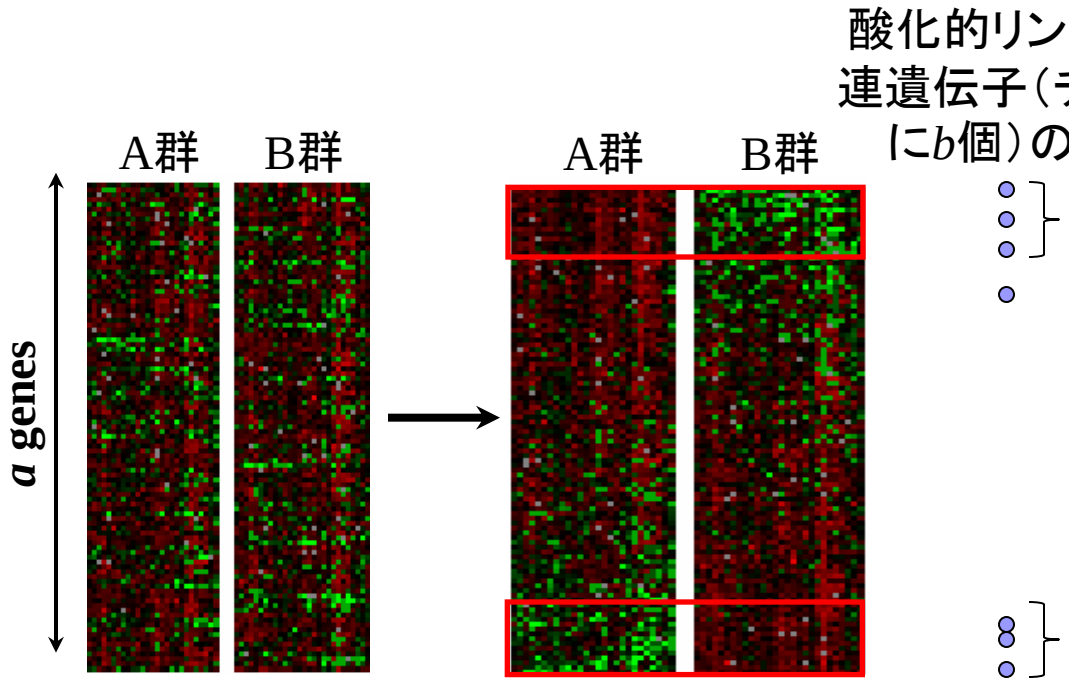
$$S_m = (AD^5 + AD^{89} + AD^{684} + AD^{2543} + ...)/m$$

  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

GSEA以前の解析手段

- 例：酸化的リン酸化関連遺伝子セット
  1. Average Differenceのような統計量を各遺伝子について算出
  2. 上位x個を抽出し、**酸化的リン酸化関連遺伝子群**のバックグラウンド（b/a）に対する濃縮度合い（c/x）を評価

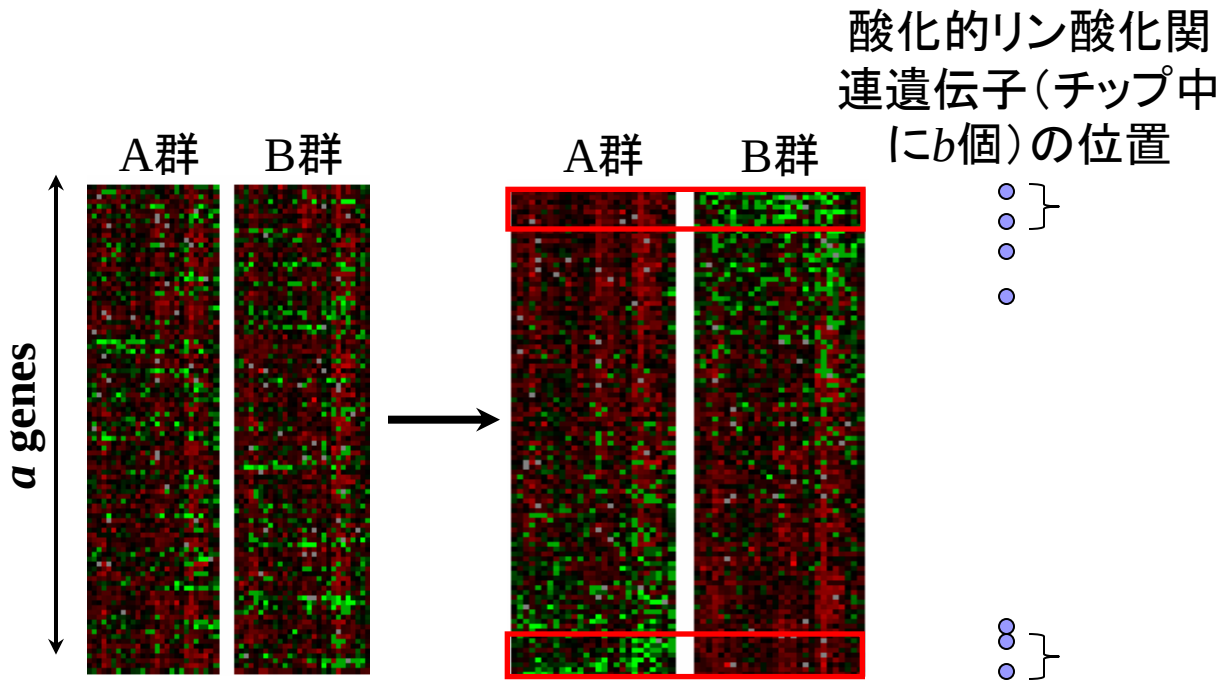


酸化的リン酸化関連遺伝子（チップ中にb個）の位置

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

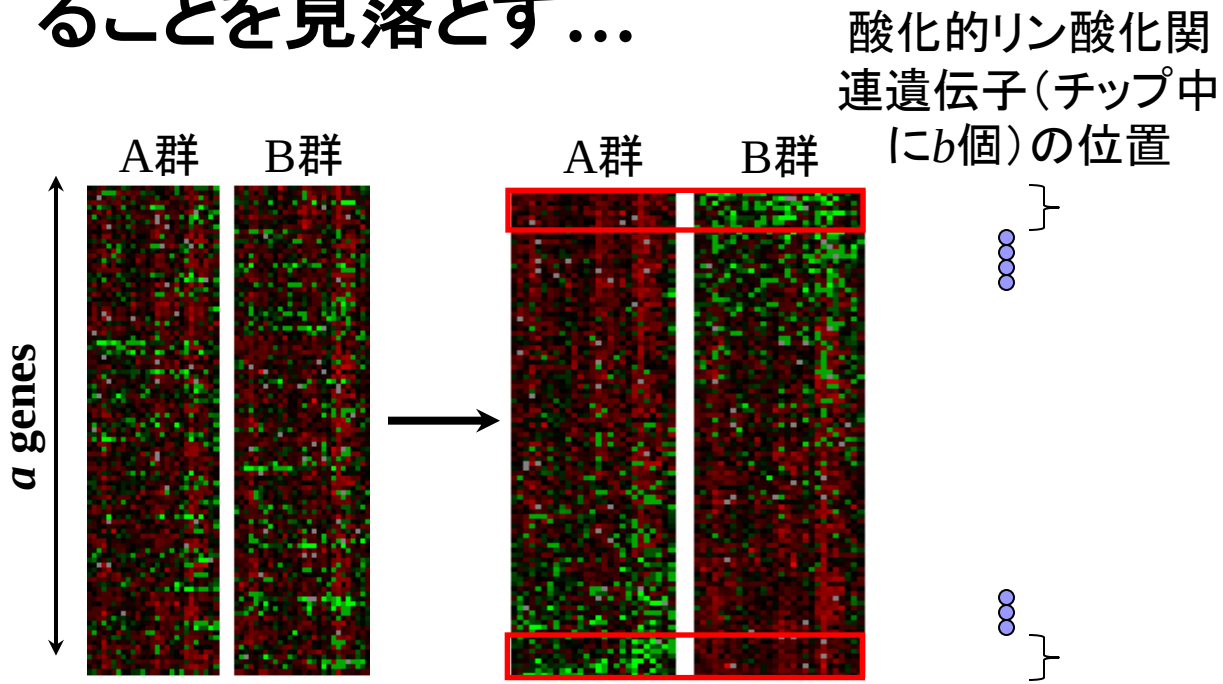
GSEA以前の解析手段の問題点1

- 上位x個のx次第で結果が変わる



GSEA以前の解析手段の問題点2

- 下図のように、全体としては**酸化的リン酸化関連遺伝子セット**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**酸化的リン酸化関連遺伝子セットが動いていることを見落とす…**



様々なGSEA系手法

- 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（gene-level statistics）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
    - その他：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）を見積もる統計量（gene set statistics）はいくらでもある
    - PAGE：Z検定
    - GSEA：Enrichment Score
    - その他：平均%順位, AUC, medianなど
  - Ans.3：有意性を評価する手段もいくつか考えられる
    - sample label permutation
    - gene resampling

極論：論文になっていない組合せを“新規手法だ！”とすることも可能…

手法選択のガイドラインはない（に等しい）

- どの遺伝子セットが動いている・いないという正解情報（“地上の真実”）を知るすべがない
  - 論文でありがちなプレゼンテーション
    - 既知の遺伝子セットはちゃんと上位にあった。我々はさらに他に動いている遺伝子セットを見つけた。（感度の高さをアピール）
    - “感度の高さ”という点については正しいのかもしれないが、“特異度”は低いのかも…。（本当は動いていない遺伝子セットまで動いていると判断してしまうこと）
  - シミュレーションで本当は動いていないデータセットを作成することはできるが、その結果と現実の結果には相当のギャップがある

GSEA系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→GSEA系手法を適用可能
- 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

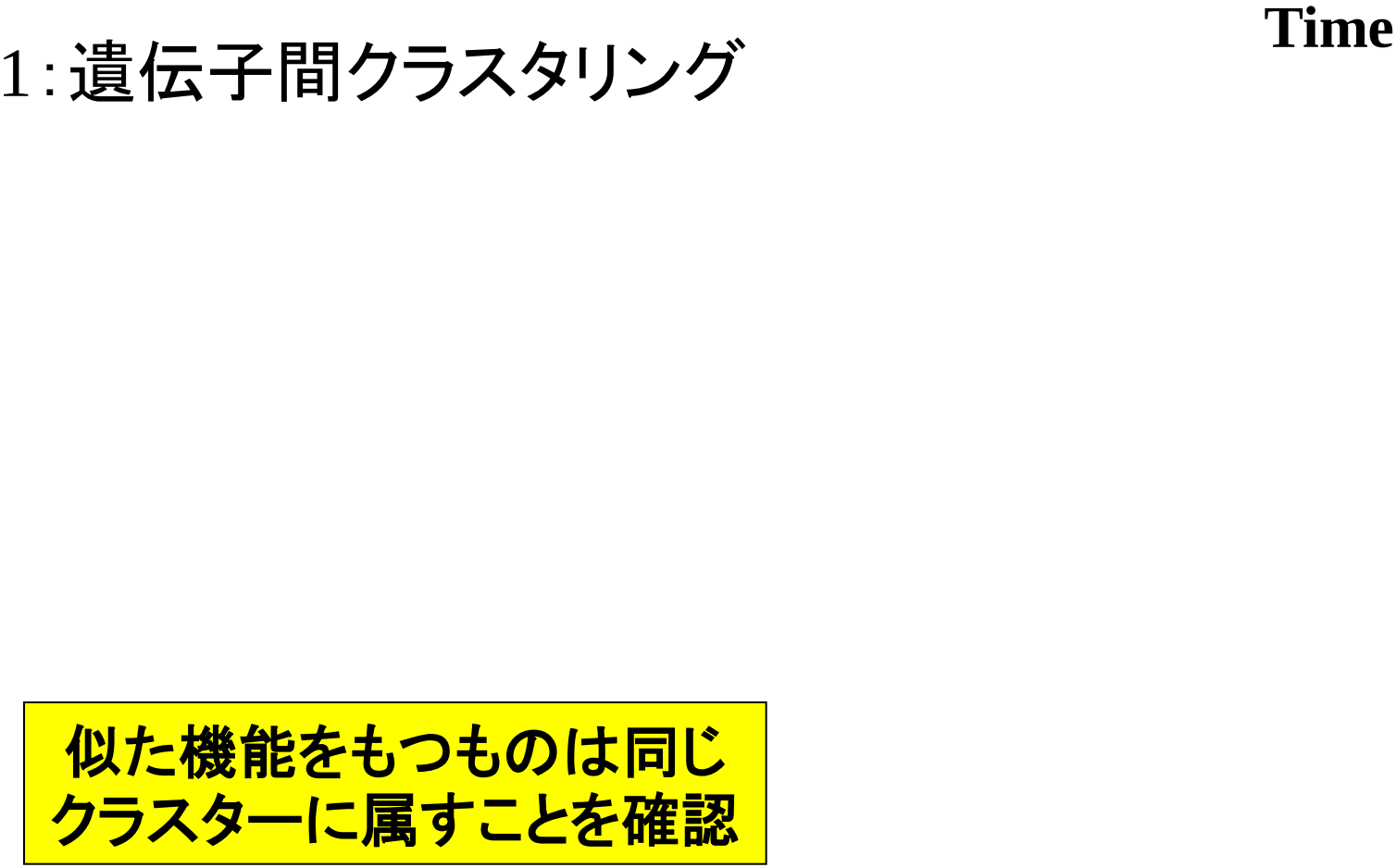
	S1	S2	S3	S4	…
gene *1*	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$	…
gene *2*	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$	…
…	…	…	…	…	…
gene *i*	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$	
…	…	…	…	…	…
gene *n*	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$	…

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）

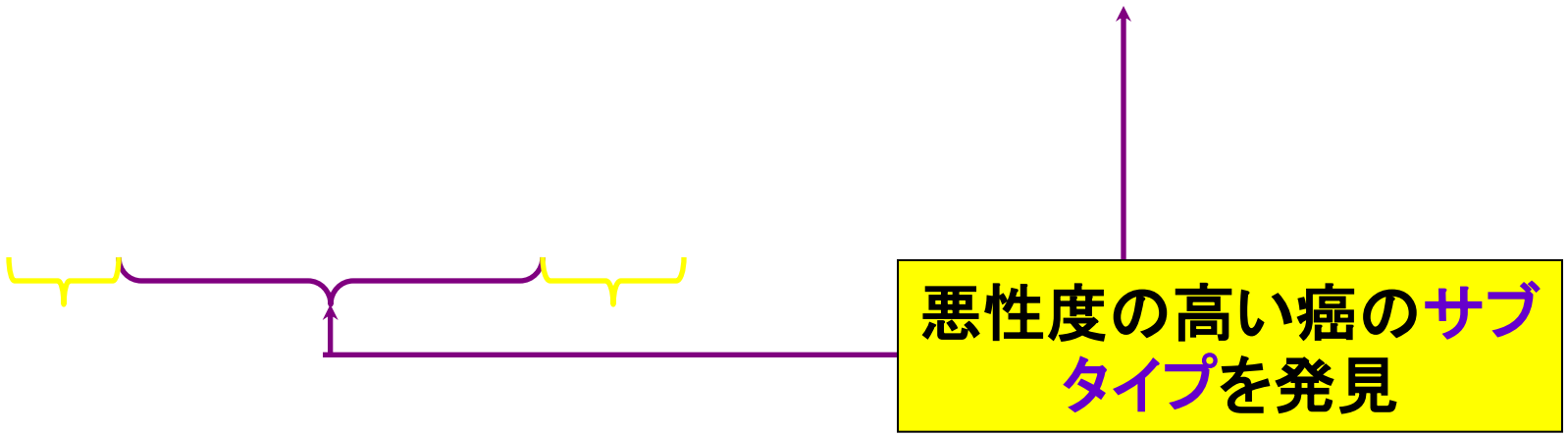
- 例1：遺伝子間クラスタリング



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



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

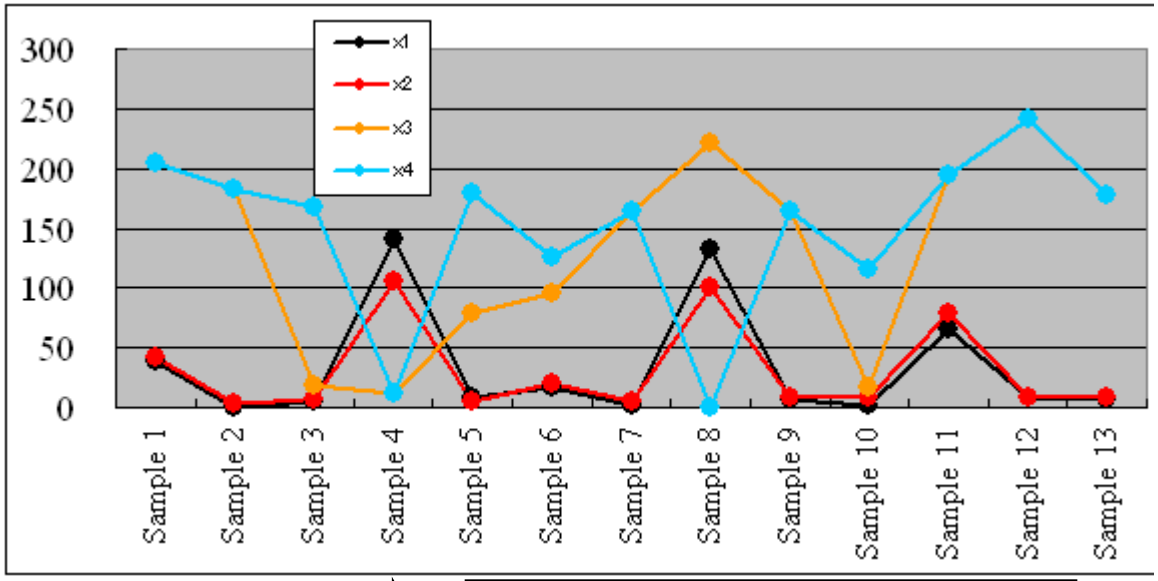
- xとyの発現パターンが酷似 $\rightarrow$ $r \approx 1$
- xとyの発現パターンがばらばら $\rightarrow$ $r \approx 0$
- xとyの発現パターンがほぼ正反対 $\rightarrow$ $r \approx -1$

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

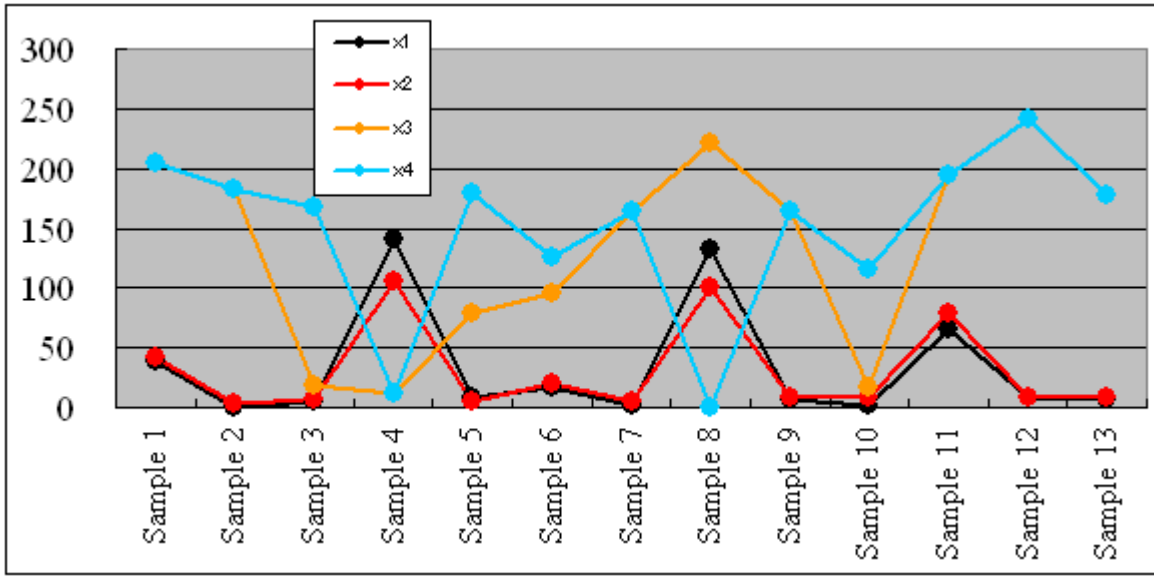
相関係数 $r_{1,3} = -0.01 \rightarrow$ 距離 $D_{1,3} = \dfrac{1-(-0.01)}{2} = 0.50$

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

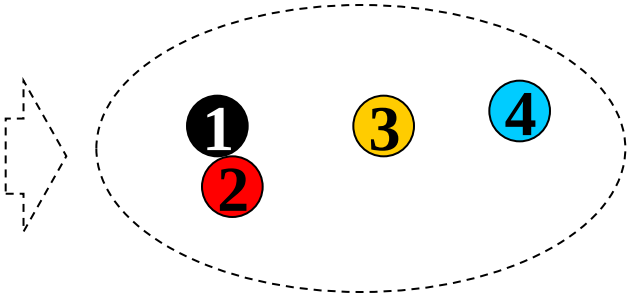
距離 $D_{1,3} = \frac{1-(-0.01)}{2} = 0.50$

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

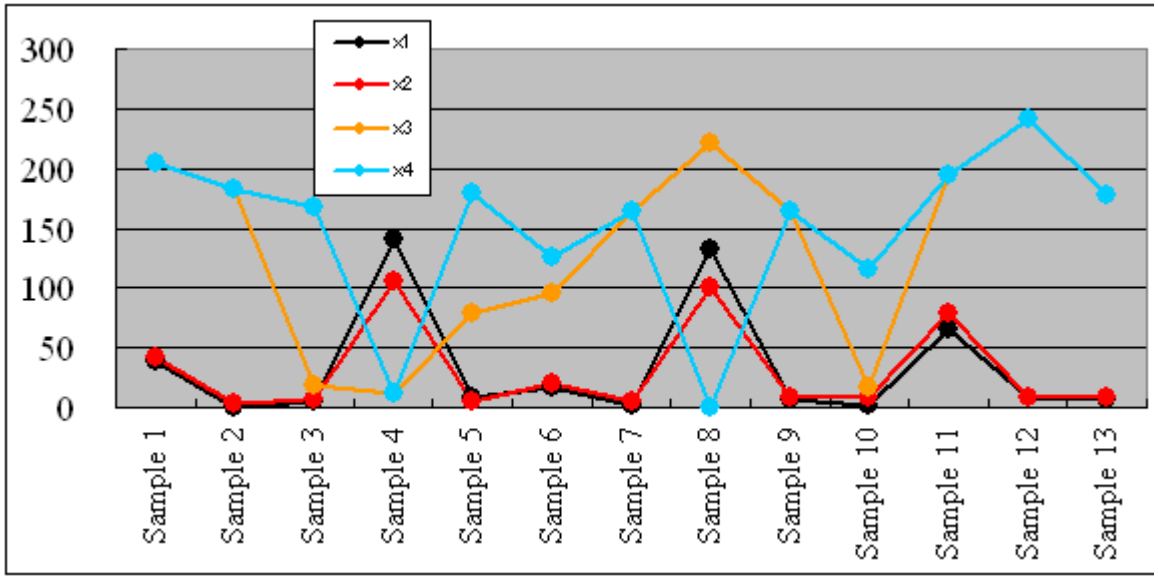
距離行列



イメージ

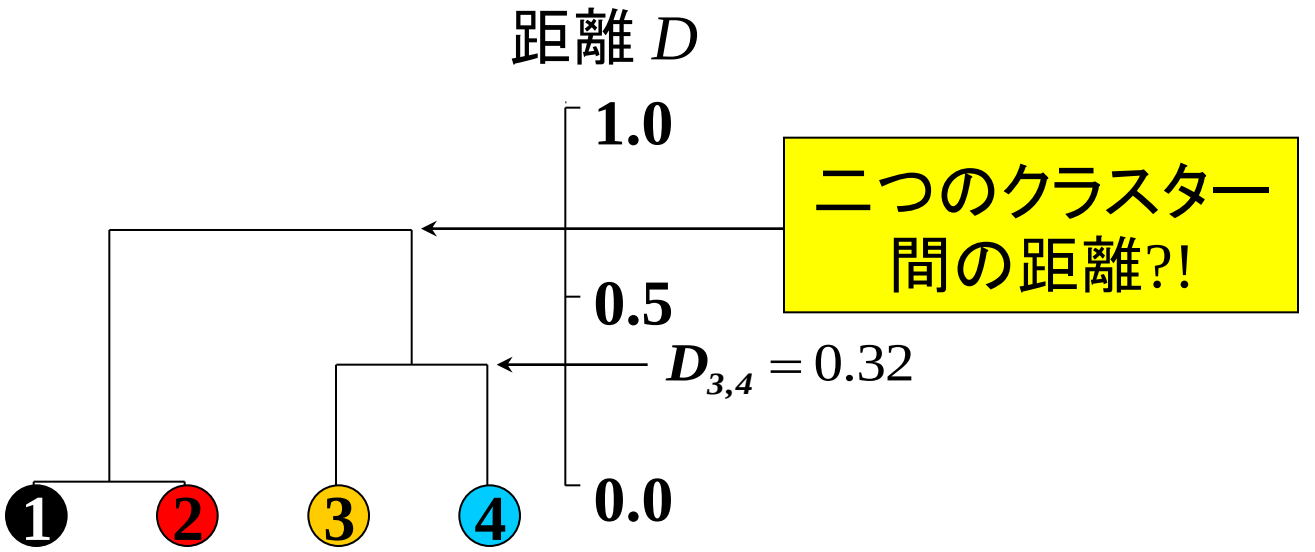
階層的クラスタリング

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

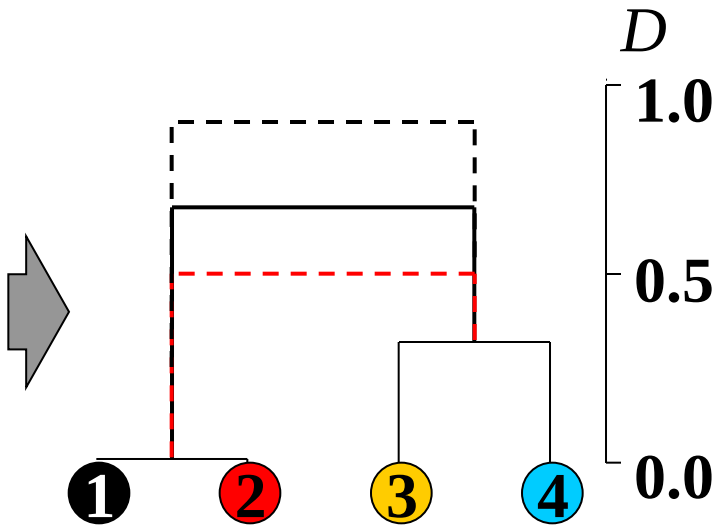
距離行列



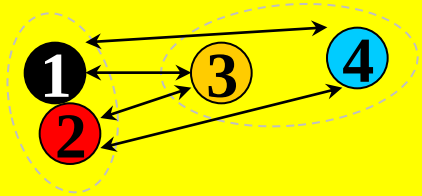
階層的クラスタリング

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

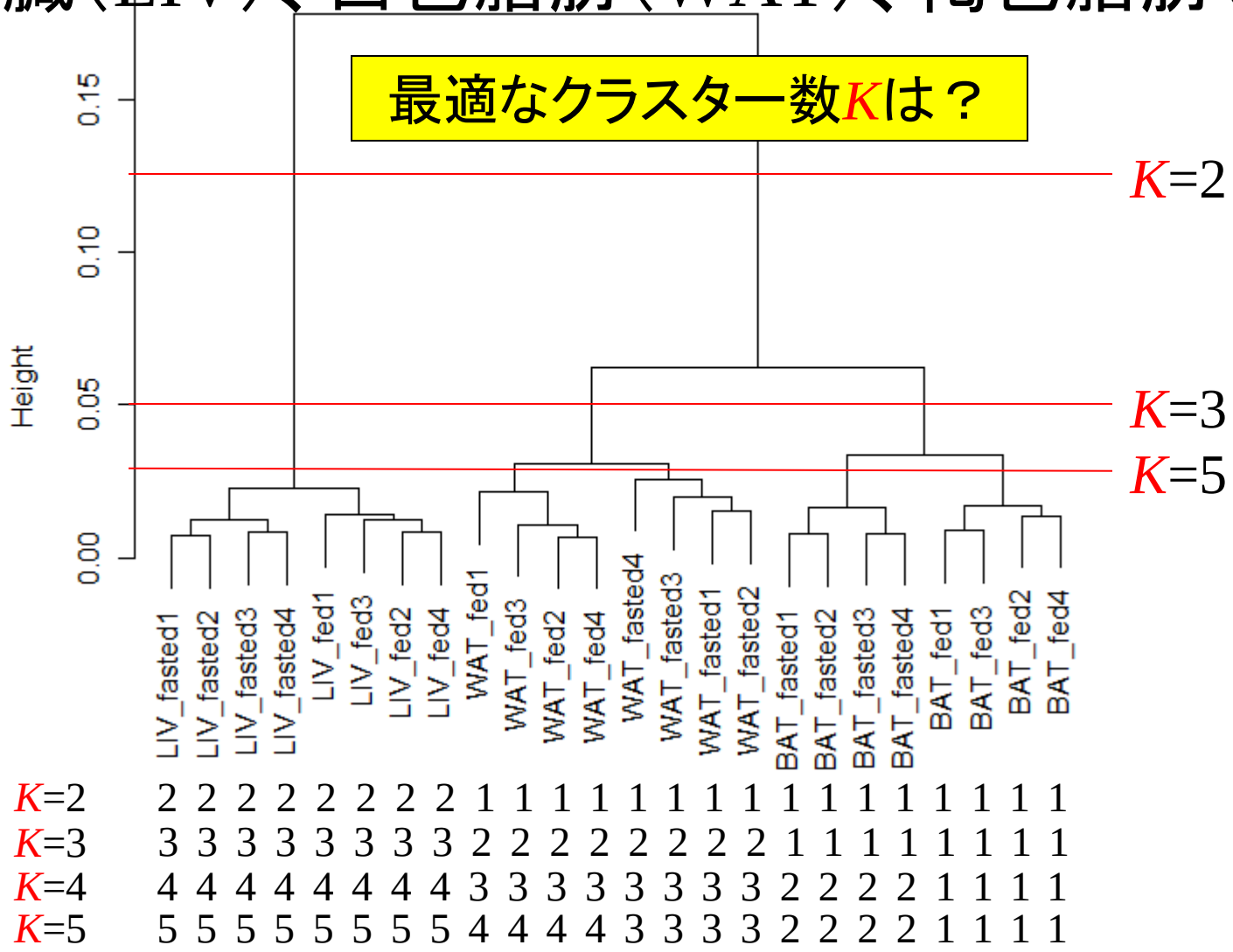
$$\min(D_{1,3},D_{1,4},D_{2,3},D_{2,4})$$
$$=0.47$$

完全連結法の場合

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

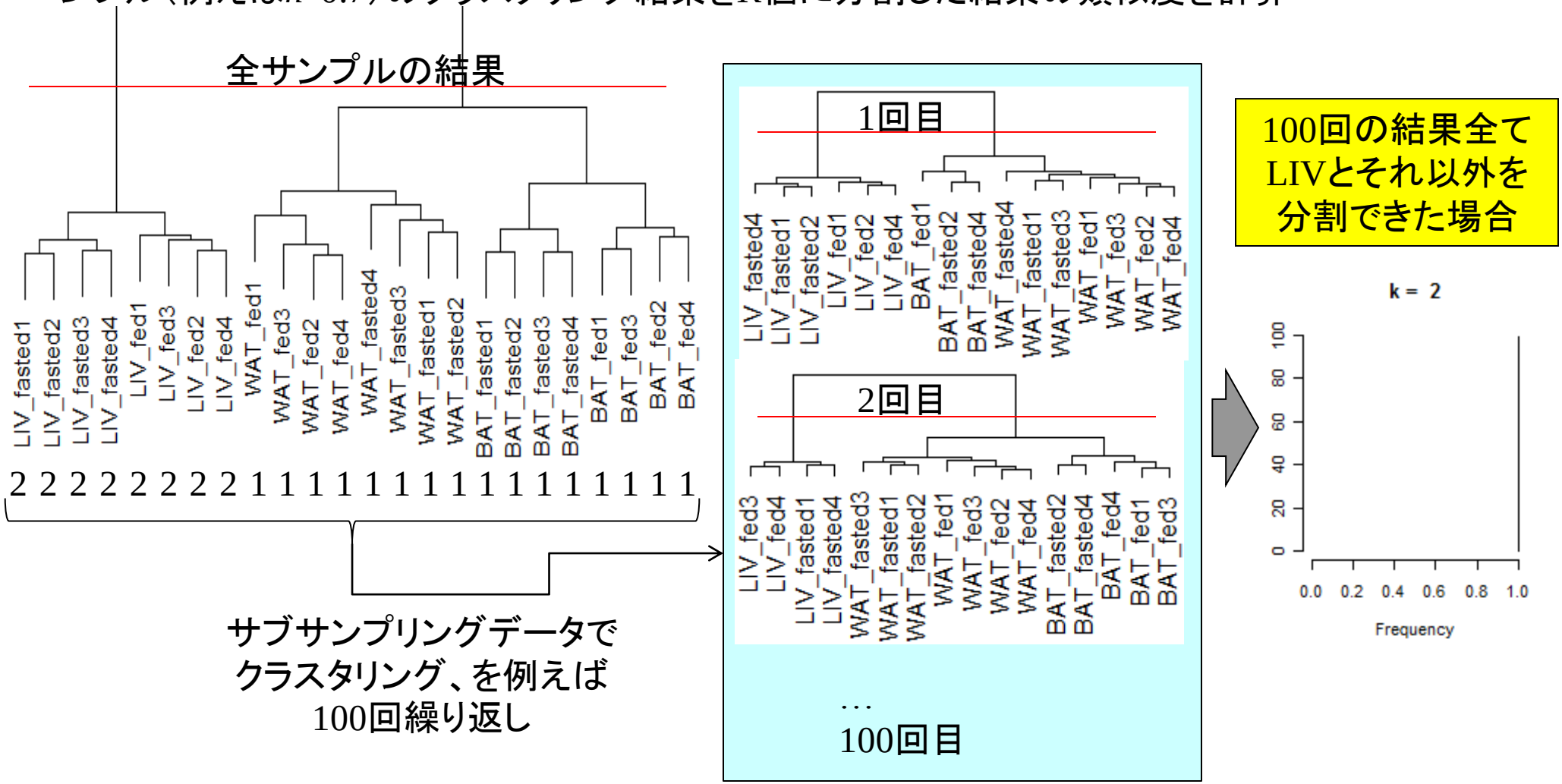
階層的クラスタリング例

- 肝臓(LIV)、白色脂肪(WAT)、褐色脂肪(BAT)



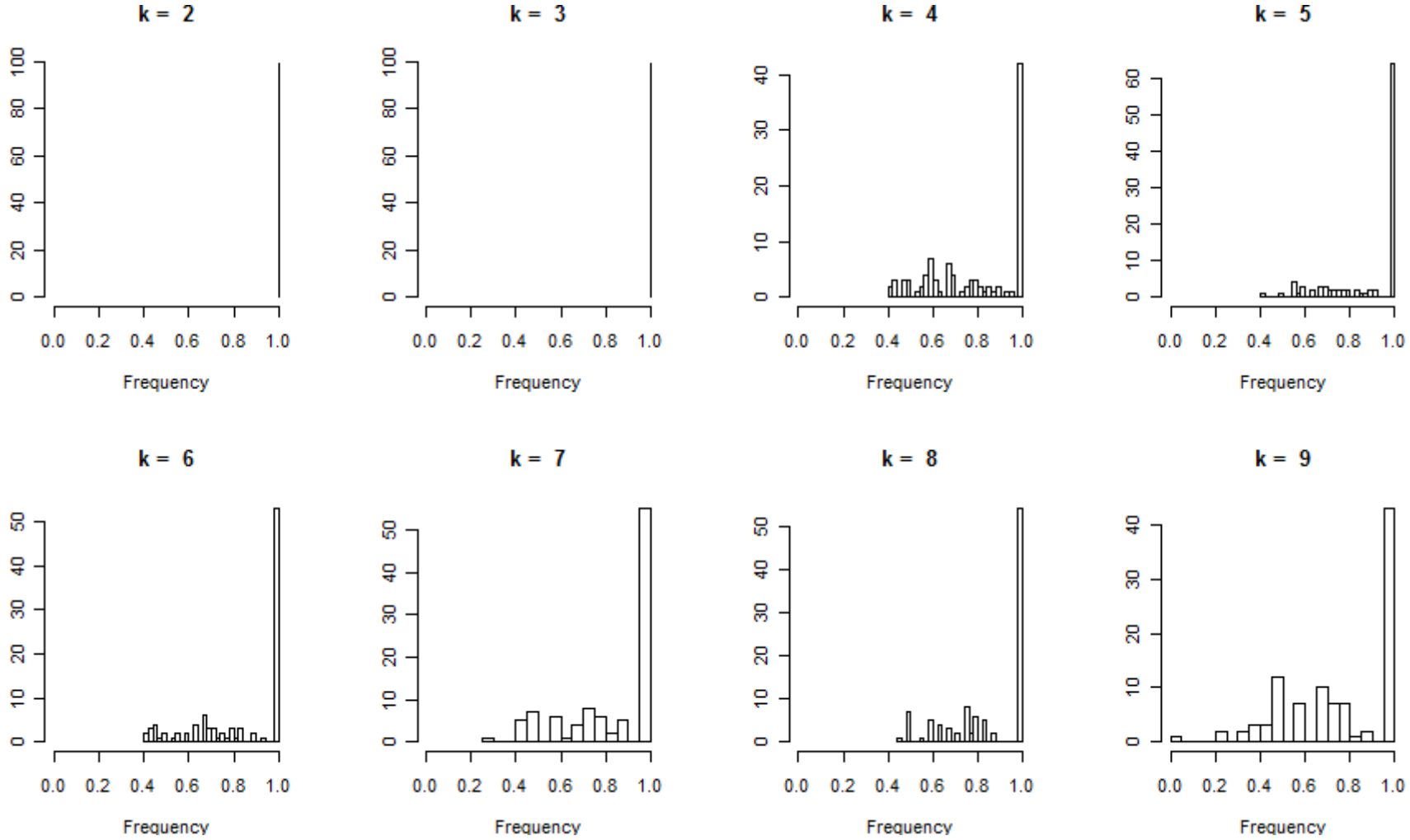
最適なクラスター数を見積もる方法

- 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）

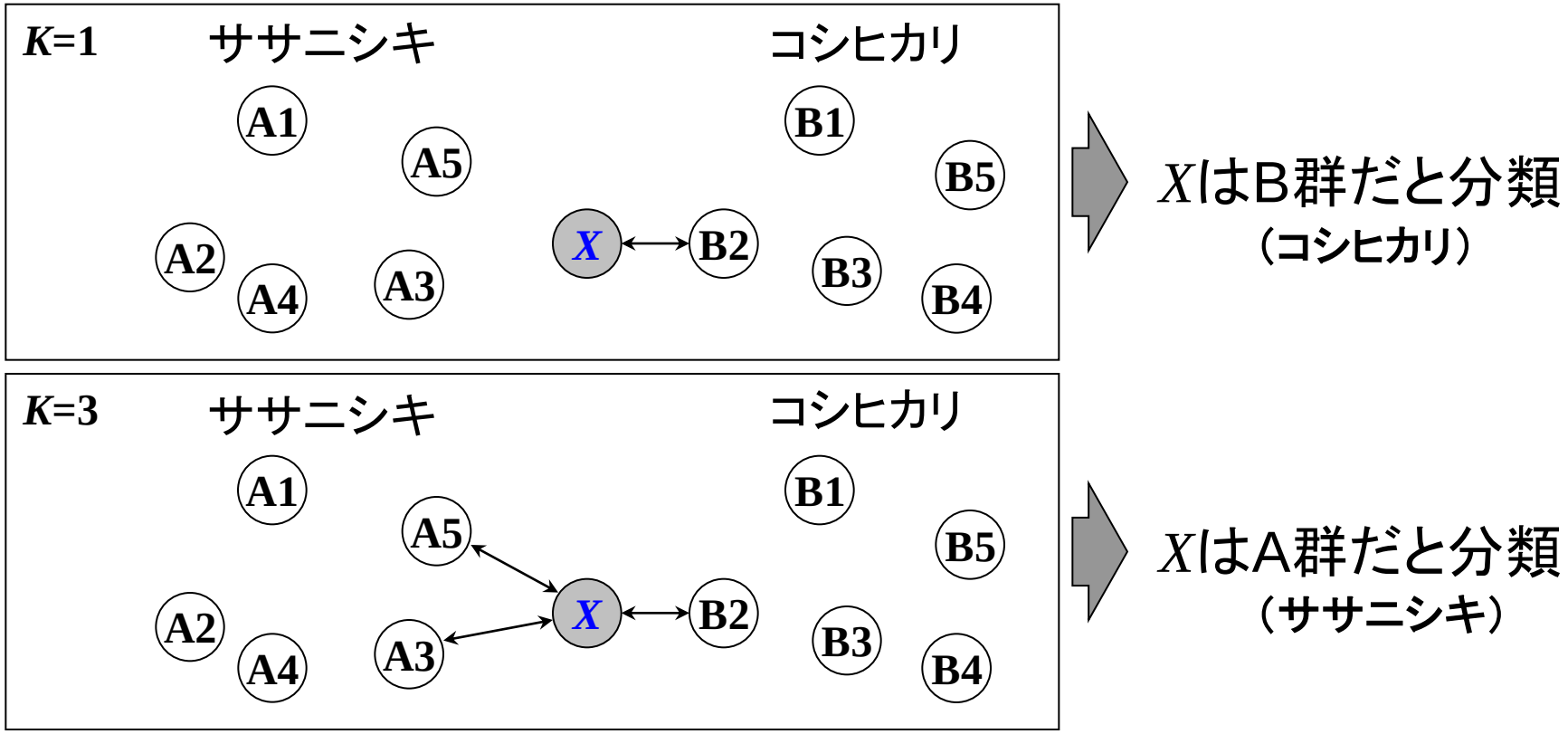
- 未知サンプルを分類するための様々な方法
  - **K-Nearest Neighbor (K-NN; K-最近傍法)**
  - Support Vector Machine (SVM)
  - Neural Network (NN)
  - Naïve Bayesian (NB)
  - Multi-Layer Perceptron (MLP; 多層パーセプトロン)
  - Weighted Voting (WV; 重みつき多数決法)
  - Decision Tree

  etc…

冬学期開講科目：
ゲノム知識情報処理論

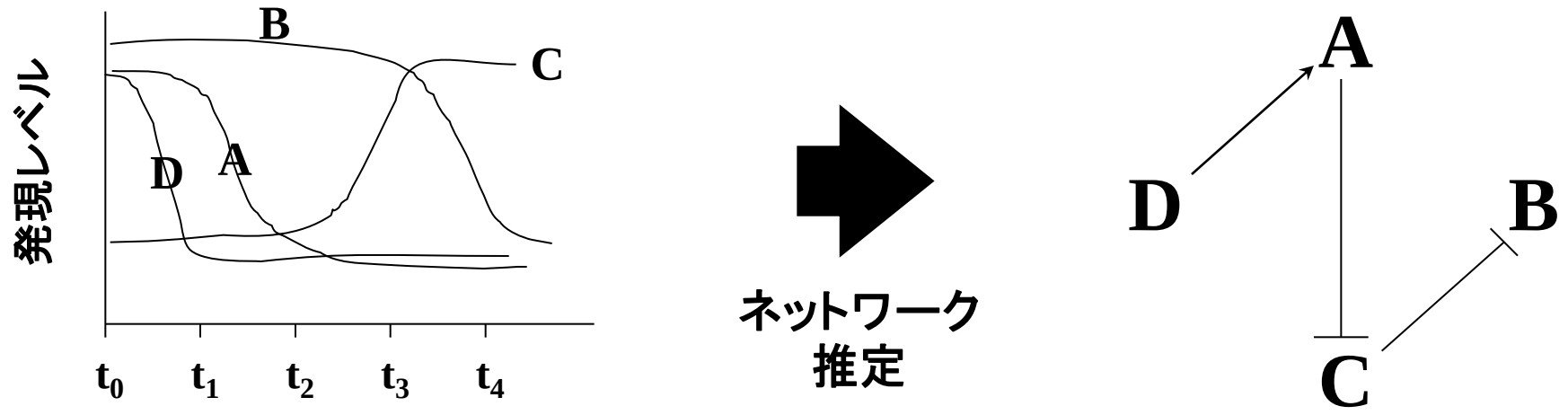
K-Nearest Neighbor (*K*-NN) 法

- 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



遺伝子の発現制御ネットワーク推定

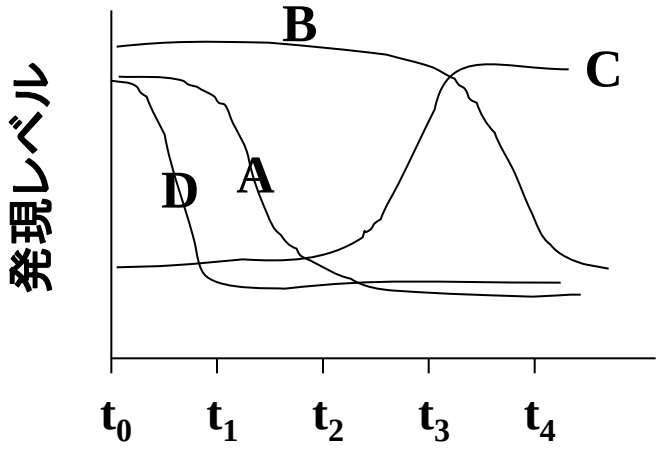
- 時系列データ
  - 遺伝子Dの発現を抑制し、他の遺伝子の挙動を観察



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

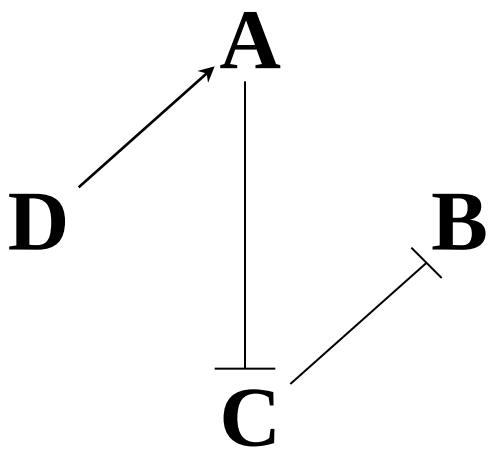
$$\begin{pmatrix} A^t \\ B^t \\ C^t \\ D^t \end{pmatrix} = \begin{pmatrix} w_{A,A} & w_{A,B} & w_{A,C} & w_{A,D} \\ w_{B,A} & w_{B,B} & w_{B,C} & w_{B,D} \\ w_{C,A} & w_{C,B} & w_{C,C} & w_{C,D} \\ w_{D,A} & w_{D,B} & w_{D,C} & w_{D,D} \end{pmatrix} \begin{pmatrix} A^{t-1} \\ B^{t-1} \\ C^{t-1} \\ D^{t-1}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

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

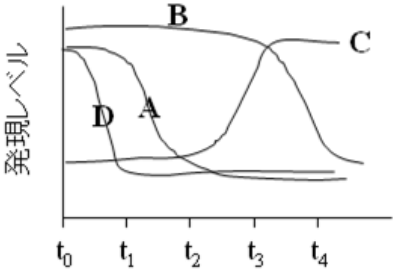
Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

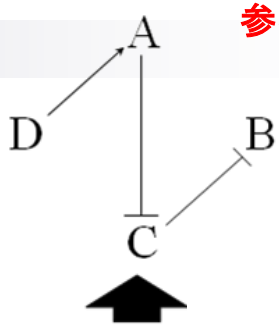
$$
\begin{cases}
-1 = w_{A,A}(-1) + w_{B,A}(0) + w_{C,A}(1) + w_{D,A}(-1) & \rightarrow w_{C,A} = 0 \\
-1 = w_{A,A}(-1) + w_{B,A}(0) + w_{C,A}(0) + w_{D,A}(-1) & \rightarrow w_{A,A} = 0 \\
-1 = w_{A,A}(0) + w_{B,A}(0) + w_{C,A}(0) + w_{D,A}(-1) & \rightarrow w_{D,A} = 1 \\
0 = w_{A,A}(0) + w_{B,A}(0)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

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！



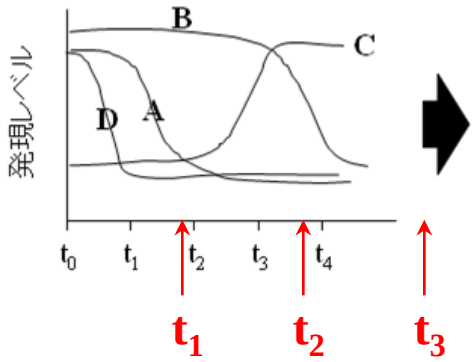


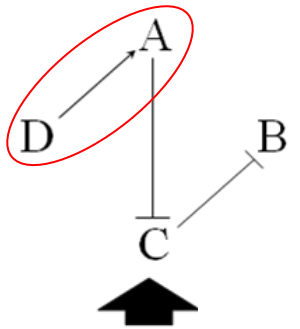
Gene	t_0	t_1	t_2	t_3	t_4
A	0	0	-1	-1	-1
B	0	0	0	0	-1
C	0	0	0	1	1
D	0	-1	-1	-1	-1

- 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

まとめ

- 様々なトランスクリプトーム解析技術
  - 概要、特徴、長所短所
  - 全て共通の“遺伝子発現行列”形式で取り扱いが可能
- “遺伝子発現行列”データ解析戦略
  - 発現変動遺伝子の同定（二群間）
  - Gene Set Enrichment Analysis
  - クラスタリング
  - 分類
  - ネットワーク推定

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プログラム
「基盤バイオインフォマティクス・フォーラム」主催

「マイクロアレイデータ解析講習会」

日時（第一回）：2009年11月20日（金）　13:00-17:00 ← **定員に達しました**
場所：東京大学農学部2号館1階第3講義室（化3）

日時（第二回）：2009年11月24日（火）　13:00-17:00 ← **定員に達しました**
場所：東京大学農学部2号館2階第1講義室（化1）

講師：門田幸二（東京大学大学院農学生命科学研究科）

マイクロアレイ解析に特化したセミナーは
11/20 or 11/24に開催予定

概要：
マイクロアレイはトランスクリプトーム解析のための基盤技術として広く普及しており、現在では比較的小規模な研究室でも受託解析などで大量の数値データを簡単に得られる状況となっている。しかしながら、簡単なデータ解析結果は得られるもののどのように解釈すればいいか分からない、自分でやってみたものの妥当なデータ解析ができているのか不安、などの理由からマイクロアレイ解析を普段から行っている専門家から実習を交えた講習会を受けてみたいという要望が多く寄せられている。そこで本講習会では、以下のような人々を**想定受講対象者**とする。

1. すでにマイクロアレイデータはあるが解析が手つかず
2. これからマイクロアレイを始めようと思っている
3. 一通りデータ解析をやってみたものの本当にそれでいいのか不安

講習会の**前半の講義**（13:00-14:20）では、バイオインフォマティクスを活用することによってそれぞれの研究を発展させるための基本的な考え方や一連の解析手法の概要を紹介する。**後半の実習**では、フリーソフトRを用いて以下に示すデータ解析の実習を行う予定であるが受講者のリクエストに応じて柔軟に対応する予定である。

1. サンプルクラスタリング
2. 発現変動遺伝子検出
3. 遺伝子セット偏り解析（いわゆるGSEA系の解析）
   (ア) Gene Ontology解析
   (イ) パスウェイ解析

アグリバイオインフォマティクス教育研究プログラムのフォーラム活動について

本プログラムでは、研究課題ごとにフォーラムを形成し、セミナー、シンポジウムの開催から、企業との共同研究、学位論文の指導などを行い、当該課題の研究・教育の活性化を図ります。フォーラムのメンバーは、本研究科の教員のほか、他大学、企業、試験研究機関の方々から構成されます。これらのメンバーから、「農学生命情報科学実習II」の受講を通して**学位論文の研究におけるバイオインフォマティクスに関係した研究の指導を受けることができます**。バイオインフォマティクスを利用した農学生命科学の研究、あるいは、バイオインフォマティクスそのものの研究を行って学位を取得した人には、「修了認定証」を発行します。修了の認定は、各専攻の学位審査とは別にフォーラムのメンバーが審査会を開いて行います。研究指導は、研究室の指導教員との合意に基づいて行いますので、**希望する人は、指導教員と相談の上、アグリバイオインフォマティクス教育研究プログラム事務局までご連絡下さい**。現在のところ、以下の4つのフォーラムが形成されています：

- **微生物インフォマティクス・フォーラム**
- **基盤バイオインフォマティクス・フォーラム**
- **アグリ／バイオ・センシングと空間情報学フォーラム**
- **食品インフォマティクス・フォーラム**