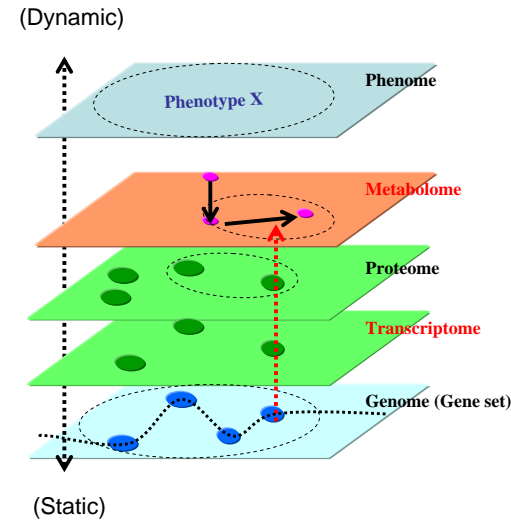


# マイクロアレイデータおよび生物学データの統計解析

金谷 重彦

情報科学研究科・比較ゲノム学

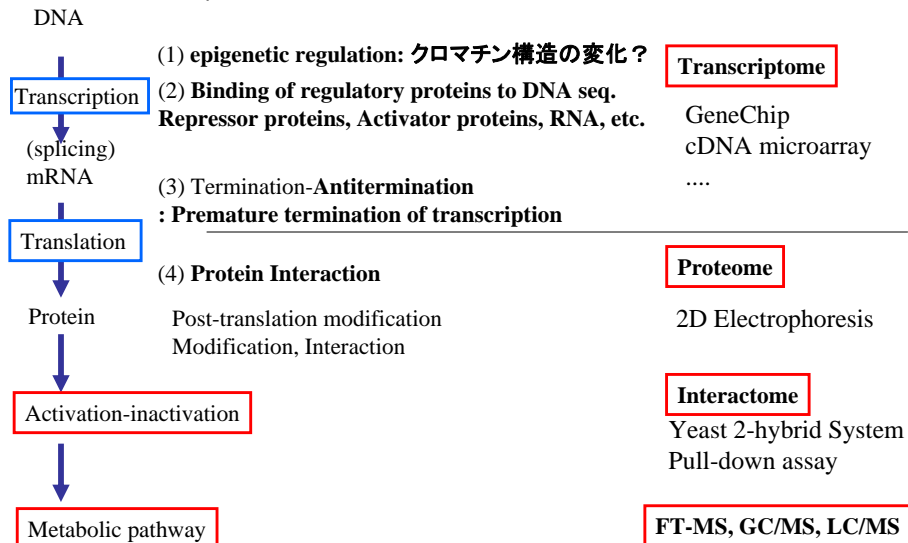
## From Genome to Phenome



(Dynamic)  
(Static)  
Progressing genome projects, many kinds of “-omics” works have been progressed such as transcriptome, .... These are dynamic information reflecting to Phenome.  
Of them, metabolites are fundamentally important as molecular phenotype

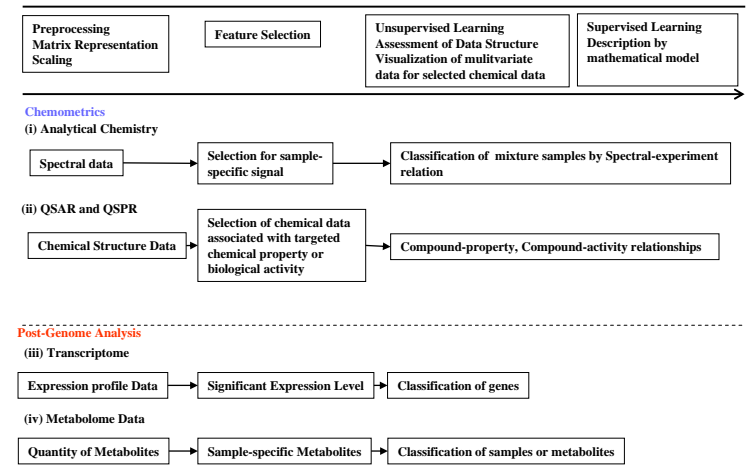
## Regulation of Gene Expression (遺伝子の発現メカニズム)

Comprehensive Analysis



## Mathematical Modeling

P.1



## 2.統計学

「統計学は、観測に基づくデータを対象とする数学であり、(i)集団の研究、(ii)変動の研究、(iii)データの簡約方法に関する研究である」(統計学者フィッシャー)

トランスクリプトーム解析やメタボローム解析

遺伝子の転写量あるいはメタボライトの量を繰り返し測定における再現性、区間推定

## 2.基本統計量と用語

$$\Sigma$$

P.2

$$\sum_{i=1}^n x_i = x_1 + x_2 + \dots + x_n$$

### 問題 1

以下の式を $\Sigma$ を用いて表せ。

(1)  $x_1y_1 + x_2y_2 + \dots + x_ny_n =$

(2)  $x_1^2 + x_2^2 + \dots + x_n^2 =$

(3)  $(x_1x_1 + x_1x_2 + \dots + x_1x_n) + (x_2x_1 + x_2x_2 + \dots + x_2x_n) + \dots + (x_nx_1 + x_nx_2 + \dots + x_nx_n) =$

(4)  $(x_1x_1 + x_1x_2 + \dots + x_1x_n) + (x_2x_1 + x_2x_2 + \dots + x_2x_n) + \dots + (x_nx_1 + x_nx_2 + \dots + x_nx_n) =$

p.2

### 問題 2

$x_1 = 1, x_2 = 2, x_3 = 3, \dots, x_{10} = 10$  のとき以下の値をもとめよ。

(1)  $x_1 + x_2 + \dots + x_{10} =$

(2)  $x_1^2 + x_2^2 + \dots + x_{10}^2 =$

(3)  $(x_1x_1 + x_1x_2 + \dots + x_1x_{10}) + (x_2x_1 + x_2x_2 + \dots + x_2x_{10}) + \dots + (x_{10}x_1 + x_{10}x_2 + \dots + x_{10}x_{10}) =$

## 確率変数 出現する確率が決まっている変数

### 期待値(平均値)

$$E(x) = x_1P_1 + x_2P_2 + \dots + x_nP_n = \sum_{i=1}^n x_iP_i$$

$$P_1 = P_2 = \dots = P_i = \dots = P_n = \frac{1}{n} \quad \text{のとき}$$

$$E(x) = \frac{\sum_{i=1}^n x_i}{n}$$

## p.3 (問題3)

### 問題 3

(1) 1,2,3,4,5,6の目の出現確率がそれぞれ 1/6 のさいころがある。このさいころを一回ふったときに出る目の期待値を求めなさい。

(2) 1,2,3,4,5,6の目の出現確率がそれぞれ 1/12, 1/12, 4/12, 4/12, 1/12, 1/12のさいころがある。このさいころを一回ふったときに出る目の期待値を求めなさい。

$$E(x) = x_1P_1 + x_2P_2 + \dots + x_nP_n = \sum_{i=1}^n x_iP_i$$

## 分散、標準偏差

$$V(x) = E\left[\{x - E(x)\}^2\right]$$

$$P_1 = P_2 = \dots = P_i = \dots = P_n = \frac{1}{n} \quad \text{のとき}$$

$$\begin{aligned} V(x) &= E\left[\{x - E(x)\}^2\right] \\ &= E\left[\{x_1 - E(x)\}^2 + \{x_2 - E(x)\}^2 + \dots + \{x_n - E(x)\}^2\right] \\ &= \{x_1 - E(x)\}^2 P_1 + \{x_2 - E(x)\}^2 P_2 + \dots + \{x_n - E(x)\}^2 P_n \\ &= \{x_1 - E(x)\}^2 \frac{1}{n} + \{x_2 - E(x)\}^2 \frac{1}{n} + \dots + \{x_n - E(x)\}^2 \frac{1}{n} \end{aligned}$$

$$= \frac{\sum_{i=1}^n \{x_i - E(x)\}^2}{n}$$

## p.3 (問題4)

### 問題 4

(1) 1,2,3,4,5,6の目の出現確率がそれぞれ 1/6 のさいころがある。このさいころを一回ふったとき出る目の分散を求めなさい。

(2) 1,2,3,4,5,6の目の出現確率がそれぞれ 1/12, 1/12, 4/12, 4/12, 1/12, 1/12のさいころがある。このさいころを一回ふったとき出る目の分散を求めなさい。

$$D(x) = \sqrt{V(x)}$$

## 平均値、分散の和

$x_1, x_2, \dots, x_i, \dots, x_k$  互いに独立な  $k$  個の確率変数

$a_1, a_2, \dots, a_i, \dots, a_k$  定数

$z = a_1x_1 + a_2x_2 + \dots + a_kx_k$  の期待値および分散は

$$E(z) = a_1E(x_1) + a_2E(x_2) + \dots + a_kE(x_k)$$

$$V(z) = a_1^2V(x_1) + a_2^2V(x_2) + \dots + a_k^2V(x_k)$$

分散は二乗の係数であることに注意！

**問題 5** 互いに独立な二つの確率変数  $x, y$  について、 $z = x - y$  の期待値  $E(x - y)$  と分散  $V(x - y)$  を  $E(x)$ ,  $E(y)$ ,  $V(x)$ ,  $V(y)$  を用いて表してみよう。

次式  $z = a_1x_1 + a_2x_2 + \dots + a_kx_k$  の期待値および分散は

$$E(z) = a_1E(x_1) + a_2E(x_2) + \dots + a_kE(x_k) \quad (2.7)$$

$$V(z) = a_1^2V(x_1) + a_2^2V(x_2) + \dots + a_k^2V(x_k) \quad (2.8)$$

である。

$$E(x - y) = E(x) - E(y)$$

$$V(z) = V(x) + V(y)$$

**問題 6** 互いに独立な二つの確率変数  $x, y$  のいずれも 0 と 1 の値をとる。また、それぞれの値が出現する確率が、

$$P(x=0) = \frac{1}{2}, P(x=1) = \frac{1}{2}, P(y=0) = \frac{1}{2}, P(y=1) = \frac{1}{2}$$

で与えられるとき、 $z = x - y$  の期待値  $E(x - y)$  と分散  $V(x - y)$  を求めてみよう。

**問題 6** 互いに独立な二つの確率変数  $x, y$  のいずれも 0 と 1 の値をとる。また、それぞれの値が出現する確率が、

$$P(x=0) = \frac{1}{2}, P(x=1) = \frac{1}{2}, P(y=0) = \frac{1}{2}, P(y=1) = \frac{1}{2}$$

で与えられるとき、 $z = x - y$  の期待値  $E(x - y)$  と分散  $V(x - y)$  を求めてみよう。

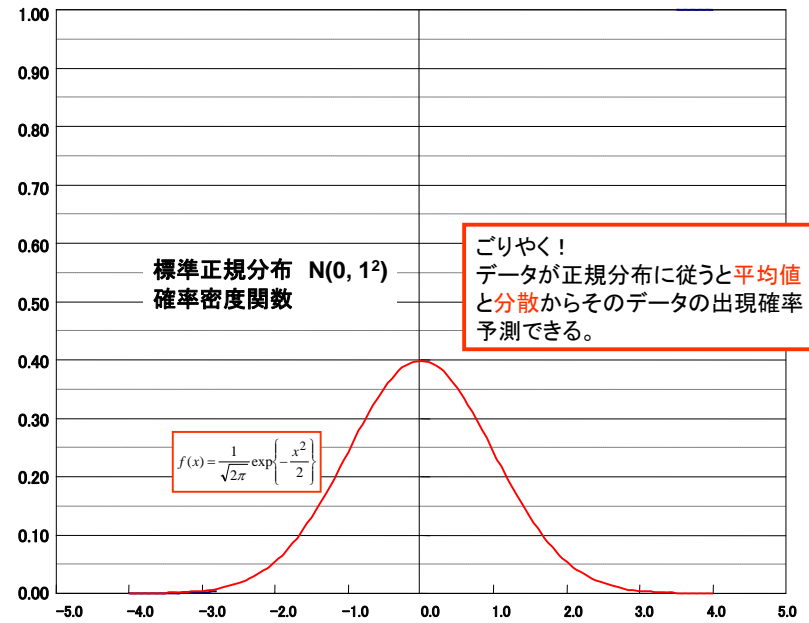
	$y$		
	0	1	
$x$	0	1	$E(x-y) = 0 \times 1/4 + 1 \times 1/4 + (-1) \times 1/4 + 0 \times 1/4 = 0$
	0(1/4)	-1(1/4)	
	1	0(1/4)	

$$V(x-y) = (0-0)^2 \times 1/4 + (1-0)^2 \times 1/4 + (-1-0)^2 \times 1/4 + (0-0)^2 \times 1/4 = 0 + 1/4 + 1/4 + 0 = 1/2 \quad (> V(x) = 1/4)$$

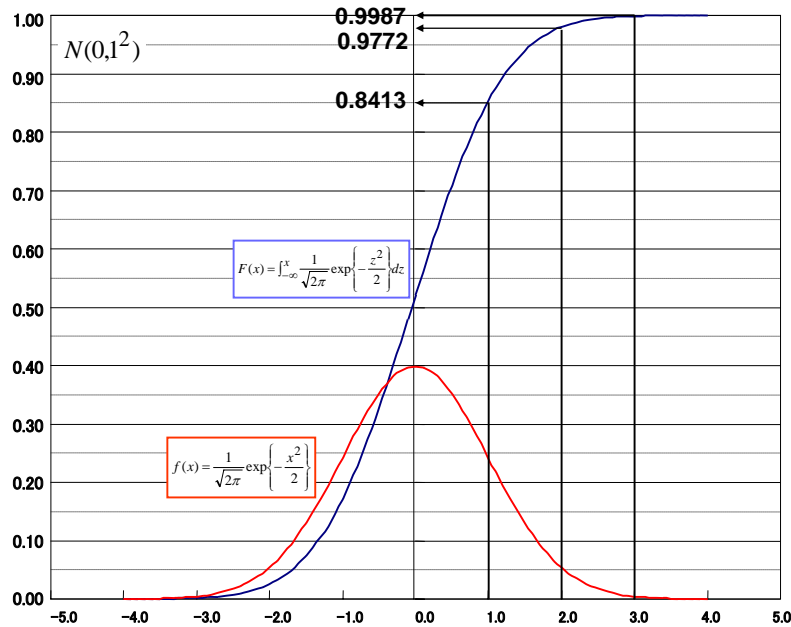
$$V(x) = (0-0.5)^2 \times 1/2 + (1-0.5)^2 \times 1/2 = 1/2 \times 1/2 \times 1/2 + 1/2 \times 1/2 \times 1/2 = 1/4$$

引き算をしても分散は増える！

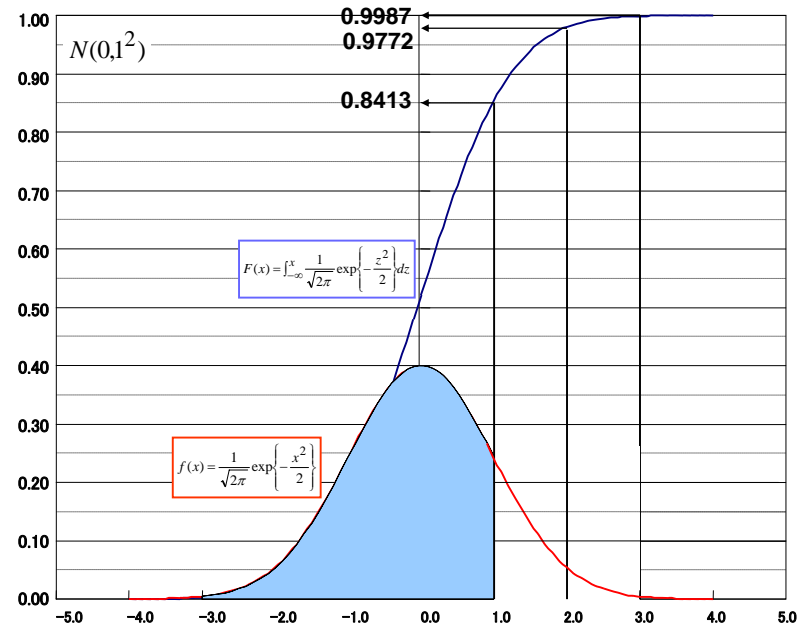
2.1 正規分布(p.5)

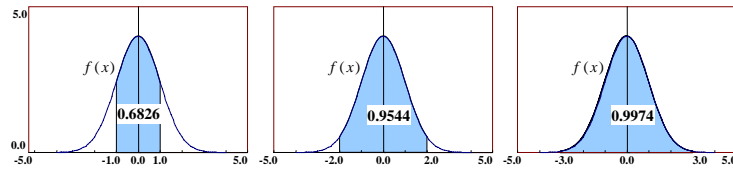


2.1 正規分布



2.1 正規分布



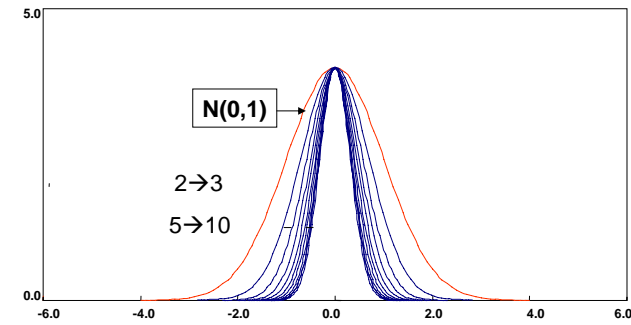


N(0,1)の累積密度関数からP(-1<x<1), P(-2<x<2), P(-3<x<3)を求めてみよう。

$$F(1) = \int_{-\infty}^1 f(z)dz = 1 - 0.1587 = 0.8413$$

$$(0.8413 - 0.5) \times 2 = 0.6826$$

2.2 平均値の分布



$$D(\bar{x}) = \frac{\sigma}{\sqrt{n}}$$

← 母標準偏差

← 実験回数 N(0,1)からn個の標本をとった。

2.2.1 母平均  $\mu$  の検定、区間推定

帰無仮説  $\bar{x} = \mu$

$\sigma$  が既知の場合は正規分布

$\sigma$  が未知の場合はt分布

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$t = \frac{\bar{x} - \mu}{\frac{\hat{\sigma}}{\sqrt{n}}} = \frac{\bar{x} - \mu}{\frac{\sqrt{V}}{\sqrt{n}}}$$

$$V = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

検定 (帰無仮説が棄却される)

$$|u_o| \geq \text{crit}(p)$$

$$|t_o| \geq t(n-1, p)$$

区間推定

$$\bar{x} - \text{crit}(p) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \text{crit}(p) \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - t(n-1, p) \sqrt{\frac{V}{n}} \leq \mu \leq \bar{x} + t(n-1, p) \sqrt{\frac{V}{n}}$$

問題7

問題7 マススペクトルのある標準サンプルに対するピークのm/z値の平均値104.840、標準偏差 $\sigma=0.028$ (N=100)であった。元旦後、新たに同一の標準サンプルのピークのm/z値を10回測定したところ、

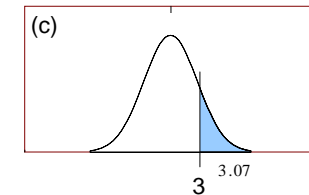
104.902, 104.840, 104.839, 104.835, 104.872,  
104.870, 104.877, 104.832, 104.900, 104.905.

個の値を得た。これらの10回の測定結果は以前のm/z値と同一とみなすことができるだろうか。

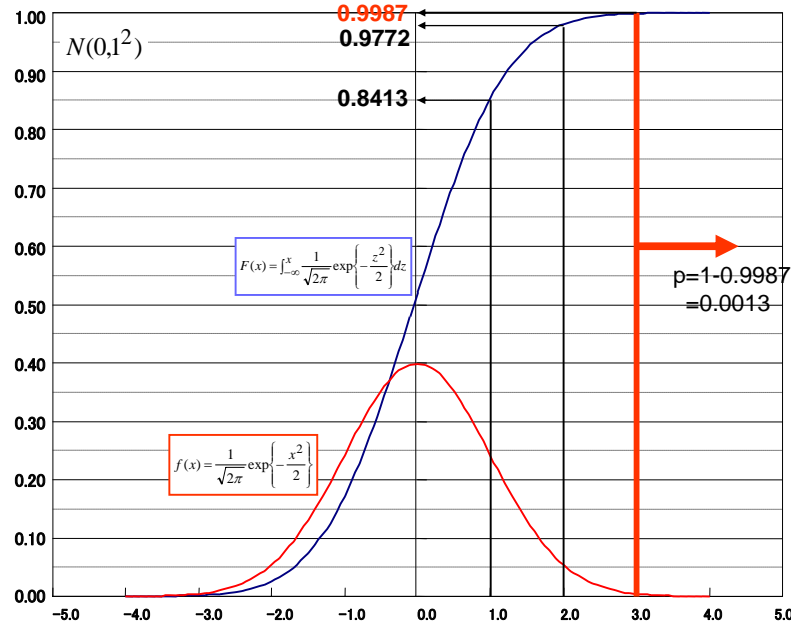
母標準偏差 $\sigma=0.0280$ がわかっているので正規分布が使える。

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

←  $\mu=104.840$   
←  $\sigma=0.0280$   
←  $n=10$



$$u_0 = \frac{104.8672 - 104.840}{\frac{0.028}{\sqrt{10}}} = 3.07193$$



問題7 マススペクトルのある標準サンプルに対するピークのm/z値の平均値104.840、標準偏差σ=0.028(N=100)であった。元旦後、新たに同一の標準サンプルのピークのm/z値を10回測定したところ、

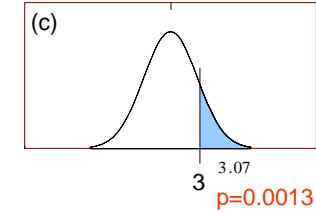
104.902, 104.840, 104.839, 104.835, 104.872,  
104.870, 104.877, 104.832, 104.900, 104.905.

個の値を得た。これらの10回の測定結果は以前のm/z値と同一とみなすことができるだろうか。

母標準偏差σ=0.0280がわかっているため正規分布が使える。

$$u = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

← μ=104.840  
← σ=0.0280  
← n=10



$$u_0 = \frac{104.8672 - 104.840}{\frac{0.028}{\sqrt{10}}} = 3.07193$$

p=0.0013で同一であるとみなせる!

通常p=0.05あるいは、p=0.01で検定を行う。このときH<sub>0</sub>: μ=μ<sub>0</sub>は棄却される。すなわち、元旦前後で装置の挙動が変化した。

平均値の検定、区間推定

σが既知の場合は正規分布

σが未知の場合はt分布

区間推定

$$\bar{x} - crit(p) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + crit(p) \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} - t(n-1, p) \sqrt{\frac{V}{n}} \leq \mu \leq \bar{x} + t(n-1, p) \sqrt{\frac{V}{n}}$$

どのくらい違うの!

問題8 (p.12)

問題8

[A] マススペクトルのある標準サンプルに対するピークのm/z値の平均値と不偏分散値は実験回数が2-10回において全て $\bar{x} = 205.002$ 、および $V = 0.010$ であった。それぞれの実験回数について平均値のt検定5%信頼限界における区間を推定せよ。

[B] マススペクトルのある標準サンプルに対するピークのm/z値の母分散は既知で $\sigma^2 = 0.010$ である。また、平均値は実験回数が2-10回において全て $\bar{x} = 205.002$ であった。それぞれの実験回数について平均値の正規分布5%信頼限界における区間を推定せよ。

自由度	1	2	3	4	5	6	7	8	9
t値	12.706	4.303	3.182	2.776	2.571	2.447	2.365	2.306	2.262

正規分布において

$$F(1.645) = \int_{-\infty}^{1.96} f(z) dz = 0.975$$

である。

問題8

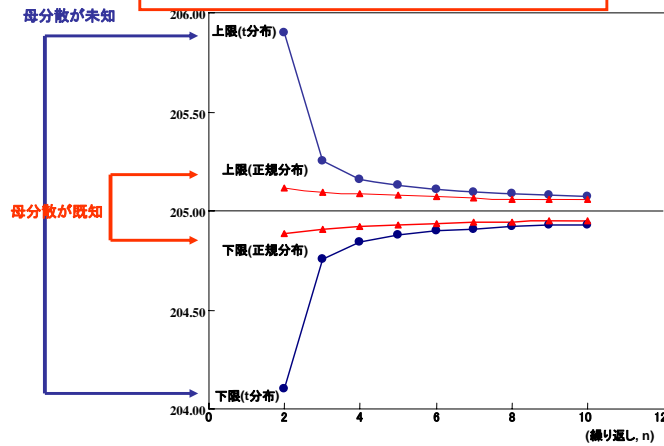
p.12

$$\bar{x} - t(n-1, p) \sqrt{\frac{v}{n}} \leq \mu \leq \bar{x} + t(n-1, p) \sqrt{\frac{v}{n}}$$

t-分布を仮定

$$\bar{x} - \text{crit}(p) \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + \text{crit}(p) \frac{\sigma}{\sqrt{n}}$$

正規分布を仮定



t-分布  
繰り返し実験回数  
に対する区間の  
変動に注目!

検定まとめ(p.16)

		仮説検定の結果	
		H <sub>0</sub> を採択する。	H <sub>0</sub> を棄却する。
事実	H <sub>0</sub> は真である。 (陽性)	true positive 1-α True positive (TP)	false positive 第一種の誤り (type I error) 有意水準=α
	H <sub>0</sub> は偽である。 (陰性)	false negative 第二種の誤り (type II error) = β False negative (FN)	true negative 検定力=1-β True negative (TN)

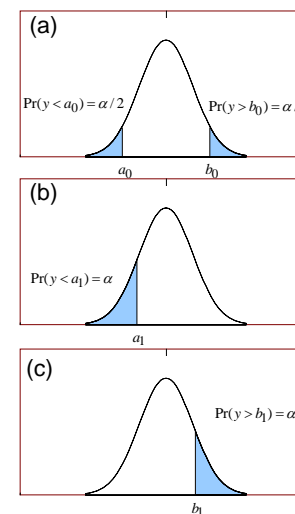
問題9 (p.16)

問題9 イヌの妊娠検査薬について統計検定をおこなった。仮説 H<sub>0</sub>: 「イヌが妊娠している」とした場合、

- (a) 検査薬が妊娠していると判定し、イヌが妊娠していない場合、
- (b) 検査薬が妊娠していると判定し、イヌが妊娠している場合
- (c) 検査薬が妊娠していないと判定し、イヌが妊娠していない場合
- (d) 検査薬が妊娠していないと判定し、イヌが妊娠している場合

のそれぞれは、 false positive, false negative, true positive, true negative のいずれか

両側検定と片側検定 (p.17)



N個の遺伝子について

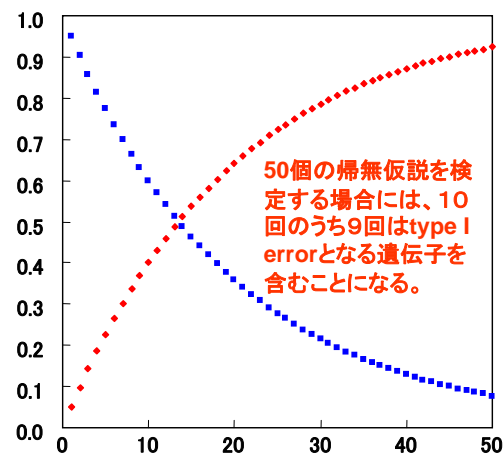
帰無仮説  $H_0$ : 二つの実験間における遺伝子発現量に有意な差がない。

p.18

対立仮説  $H_1$ : 二つの実験間における遺伝子発現量に有意な差がある。

ここで、有意水準  $\alpha=0.05$  とする。

として t 検定を行う。



$$1 - (1 - \alpha)^N$$

少なくとも一つの遺伝子の検定において Type I error を起こす確率

二つの実験においてN個の遺伝子について発現量に有意な差がないにもかかわらず、危険率  $\alpha=0.05$  において少なくとも一つの遺伝子について有意な差があると判定される確率

$$(1 - \alpha)^N$$

全ての遺伝子について帰無仮説が成り立つ確率

二つの実験においてN個の遺伝子について発現量に有意な差がなく、かつ危険率  $\alpha=0.05$  において全ての遺伝子が有意な差がないと判定される確率

### Dunn-Sidak補正とBonferroni補正

Dunn-Sidak補正

$$\alpha' = 1 - (1 - \alpha)^{1/N}$$

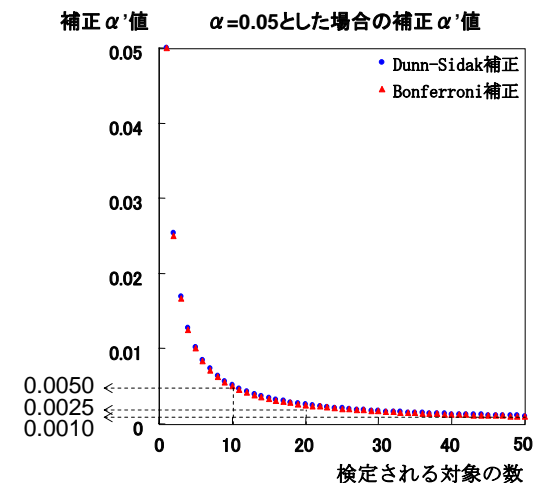
Bonferroni補正

$$\alpha' = \alpha / N$$



閾値を  $\alpha'$  により求める。

$$t(n-1, \alpha')$$



### Dunn-Sidak補正とBonferroni補正

p.19

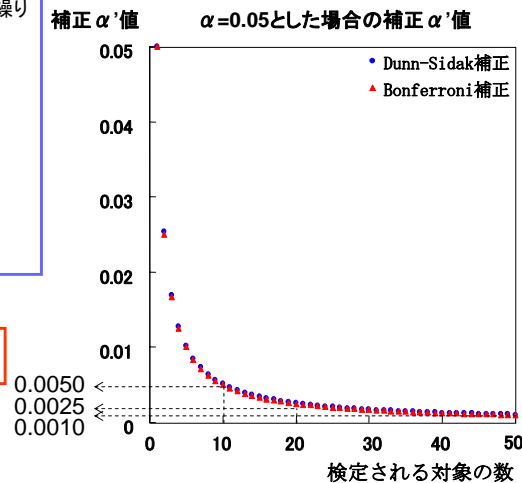
それぞれの遺伝子について  $n=10$  回の繰り返し実験がなされたとすると

N=1の場合  
 $t(n-1=9, \alpha=0.05)=1.833$

N=10の場合  
 $t(n-1=9, \alpha'=0.0050)=3.250$

N=100の場合  
 $t(n-1=9, \alpha'=0.0005)=4.781$

$|$ 実験データから得られる  $t$  値  $> t(n-1, \alpha')$  の時、帰無仮説が棄却される。



### 問題10(p.19)

問題 10 N=100, 1000 において、有意水準  $\alpha=0.05$  に対する Dunn-Sidak 補正ならびに Bonferroni 補正値を求めよ。

False Discovery Rate (FDR) (p.19)

図9上段(ランダムデータ)

N=20,000遺伝子について、帰無仮説「二つの条件において遺伝子の発現量に有意な差がない」のもとで検定を行うことを考える。

0.01の刻み幅でグラフに表すと遺伝子の出現個数は常に20,000 x 0.01=200となる。

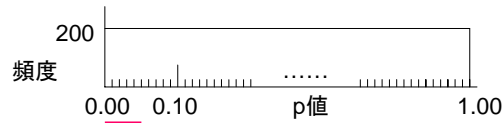


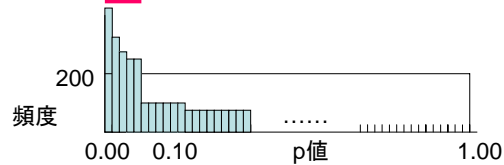
図9中段(測定データ)

N=20,000遺伝子におけるp値の分布は、データがランダムでなく有意に変化したものが含まれるとすると、p値最小(最大)の領域において、**真中殿**の頻度を上回ることになる。

p値において最小あるいは最大の領域で

$$FDR = N_{\text{Rand}} / N_{\text{exp}} < 1$$

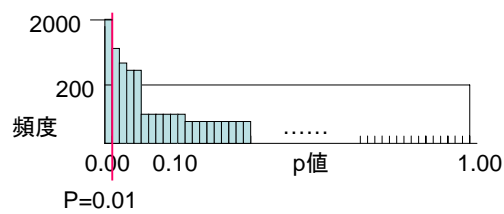
が成り立つ間、実験間の発現量の差が有意である遺伝子が含まれている。



$$FDR = 0.01$$

$N_{\text{Rand}}=200$ についてFDR= 0.01  
であるとすると、 $N_{\text{exp}}=2000$

真に発現量に変化があった遺伝子の個数=2000-200=1800



### 3. マイクロアレイデータ解析

## [2] Transcriptome

(1) Methods

1A GeneChip

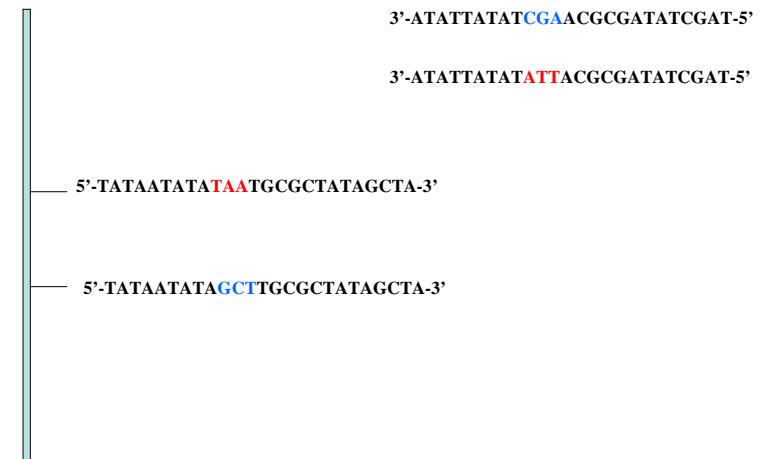
1B cDNA microarray

(2) Informatics for cDNA microarray

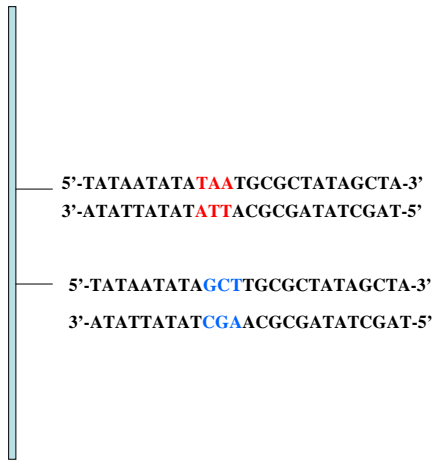
2A Signal processing

2B Data mining

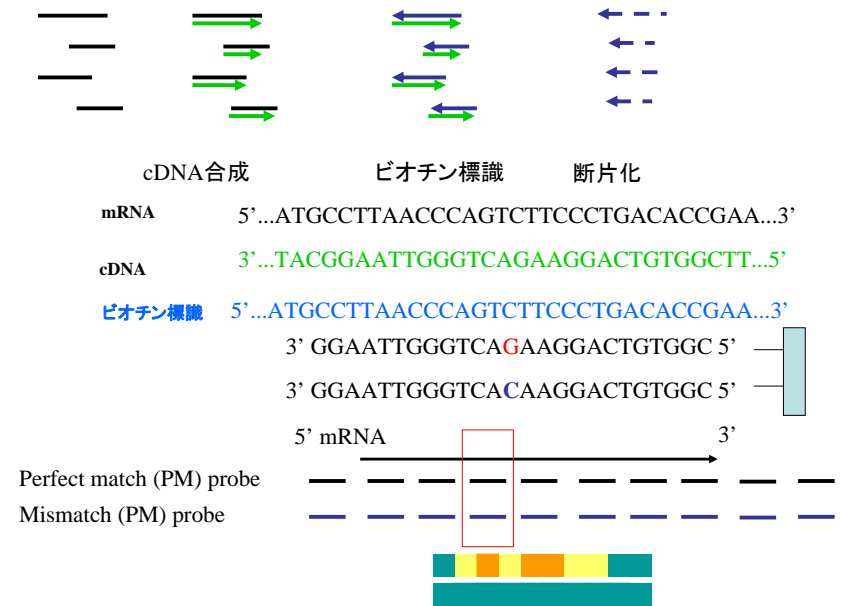
DNA as information-transfer molecules: (DNA hybridization )



### DNA as information-transfer molecules: (DNA hybridization )



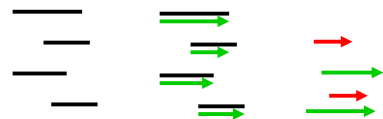
### 1A GeneChip



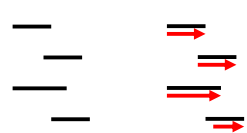
### 1B cDNA microarray (p.22)

#### (i) mRNAの調製

条件1: **Cy3**  
(コントロール)

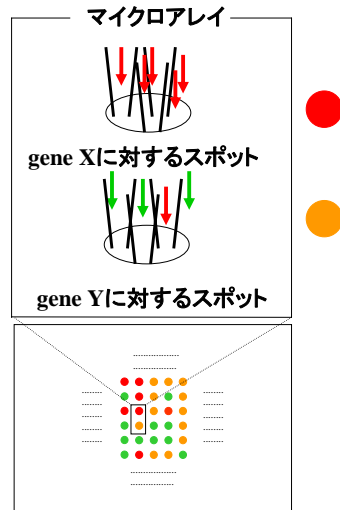


条件2: **Cy5**  
(比較したい実験)

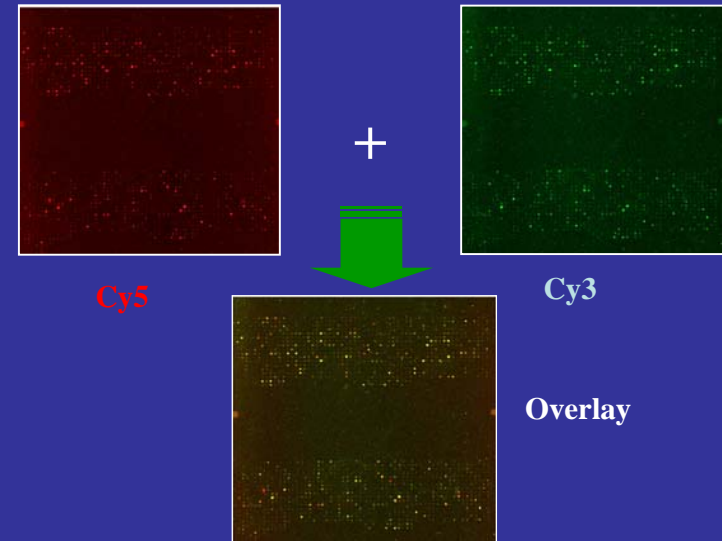


#### (ii) 蛍光色素でラベルした cDNAの作製

#### (iii) 競合ハイブリダイゼーション

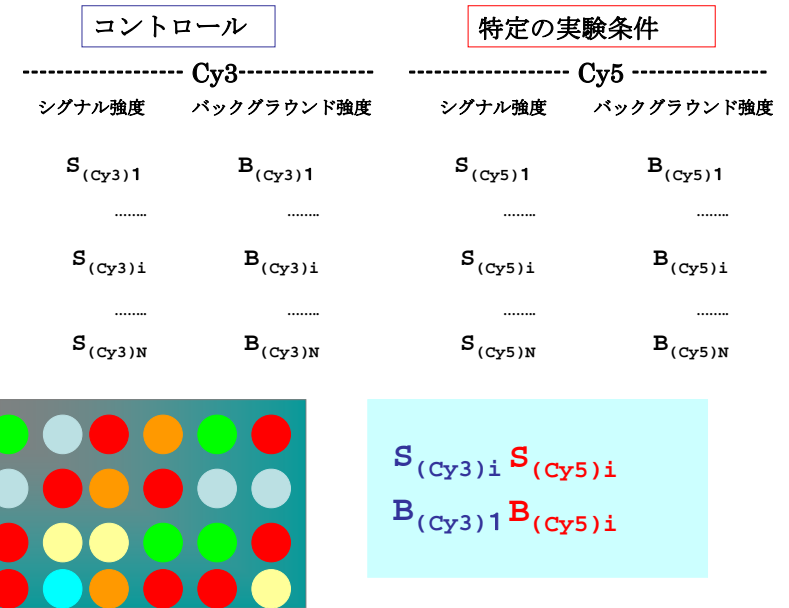


### cDNA microarray



シグナルプロセッシング→3.2 シグナル評価 (p.22)  
 データの体系化→4 多変量解析 (p.28)

3.2 シグナル評価

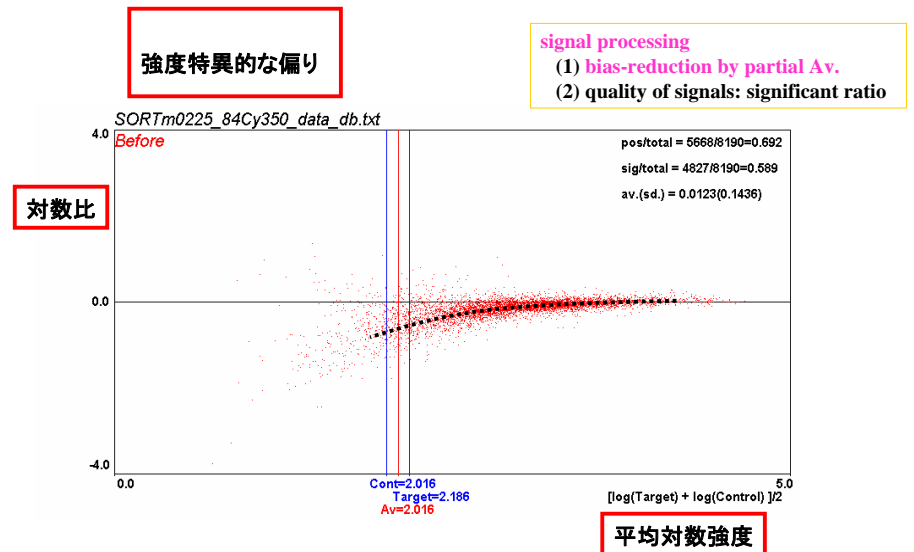


p.22

強度に依存した部分統計量を用いる方法

Quackenbush J., Nature genetics supplement, 32, 496-501, 2001

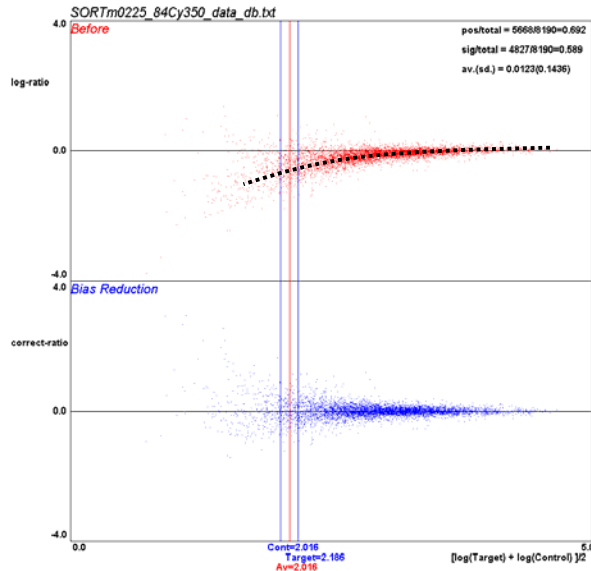
P.22



In many case intensity-specific bias is observed.

(1) Bias reduction by partial averages

signal processing  
 (1) bias-reduction by partial Av.  
 (2) quality of signals: significant ratio



$$M_i = \log(T_i / R_i)$$

$$A_i = \{\log(T_i) + \log(R_i)\} / 2$$

$$M_i = f(A_i) + \varepsilon_i$$

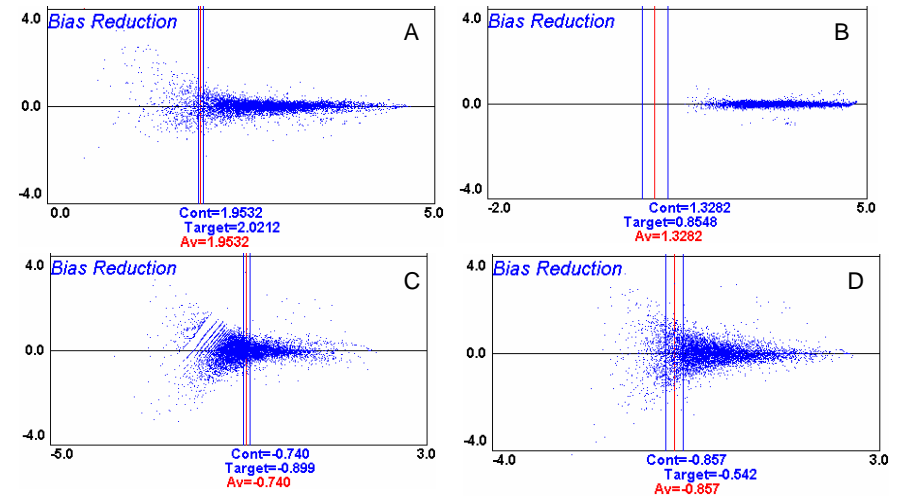
ベースライン

$$\hat{M}_i = f(A_i)$$

実測値とベースラインの差

$$\tilde{M}_i = M_i - \hat{M}_i$$

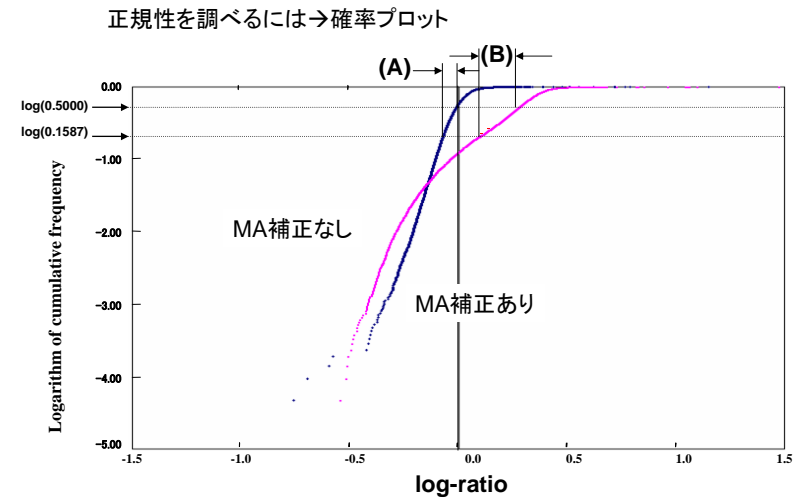
A,B,C,Dは二つの実験における遺伝子の発現量のMAプロットである。測定のクオリティが最も高いものはどれか。

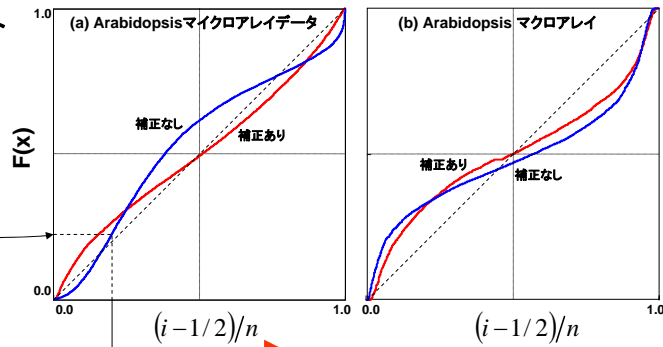


p.24

3.3 観測データの分布が正規分布に従うか否かの判定法

p.24

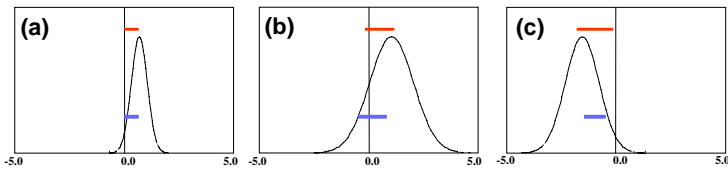




$$F(z_i) = \int_{-\infty}^{z_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\} dz$$

3.4 2色蛍光型マイクロアレイデータの発現量変化における有効スポットの検出法 (p.26)

0とみなせる確率



1. サンプル平均  $\bar{x}$  と不偏分散  $v = \frac{\sum_{i=1}^n (y_i - \bar{x})^2}{n-1}$  をもとに  $t_0 = \frac{\bar{x} - 0}{\sqrt{v/n}}$  を計算する。
2. 理論分布  $t(n-1, p)$  をもとに  $t_0 = t(n-1, p)$  となる p 値を求める。この p 値により再現性の程度を評価する。

1-p(0とみなせる確率): 符合が有効か否かの判定  
多重性の問題

表1 2回の cDNA マイクロアレイ測定結果

	1st	2nd	Av	SD	p
[1] At2g38270	0.32832	0.32831	0.32832	0.00001	0.00001
[2] At2g34620	0.35022	0.35028	0.35025	0.00003	0.00003
[3] At2g18710	0.26919	0.26927	0.26923	0.00004	0.00004
[4] At2g22500	0.42060	0.42043	0.42052	0.00008	0.00006
[5] At4g37870	-0.68499	-0.68533	-0.68516	0.00017	0.00008
[6] At1g69080	-1.29664	-1.29767	-1.29716	0.00052	0.00013
[7] At3g05290	-0.44634	-0.44598	-0.44616	0.00018	0.00013
[8] At2g40940	-0.21412	-0.21395	-0.21404	0.00009	0.00013
---					
[9] At5g13740	0.03155	-0.03124	0.00015	0.03140	0.49843
[10] At4g32620	-0.09012	0.08925	-0.00043	0.08968	0.49846
[11] At4g09510	0.13364	-0.13245	0.00060	0.13304	0.49857
[12] At3g45620	-0.12305	0.12198	-0.00053	0.12252	0.49861
[13] At4g01560	-0.19832	0.19982	0.00075	0.19907	0.49880
[14] At5g35570	-0.07307	0.07351	0.00022	0.07329	0.49905
[15] At2g47710	-0.05806	0.05785	-0.00011	0.05795	0.49940
[16] At2g27450	-0.06773	0.06796	0.00011	0.06785	0.49947

### 4. 多変量解析

#### 1. 前処理

解析目的に適切な縮尺にデータをスケールリングする。

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1M} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2M} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{iM} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{NM} \end{pmatrix}$$

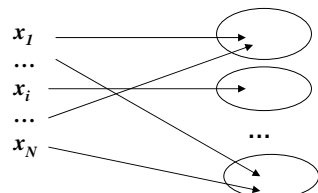
#### 2. 特徴抽出

識別に有効な変量を選択する。同一の情報をもつ変量を除去する。

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1M'} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2M'} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{iM'} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{N1} & x_{N2} & \dots & x_{Nj} & \dots & x_{NM'} \end{pmatrix}$$

#### 3. 教師なし学習

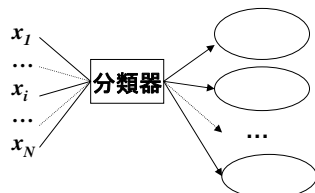
多変量データの類似性に基づいて対象の分類する。



類似性に基づいてグループ

#### 3. 教師あり学習

多変量データに基づいて対象をユーザー指定のクラスに分類をする。



ユーザが指定したクラス

## Multivariate Analysis

### (a) Similarity

#### 1. Unsupervised learning

配列間比較

系統樹

クラスター分析

#### 2. Supervised learning

kNN法

### (b) Multivariate Vectors

#### 1. Unsupervised learning

主成分分析

マルコフモデル

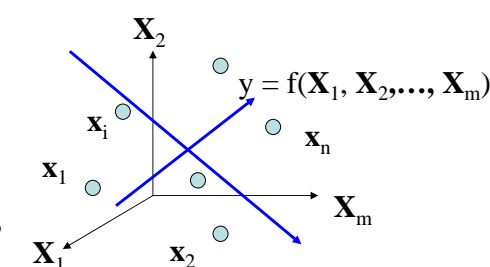
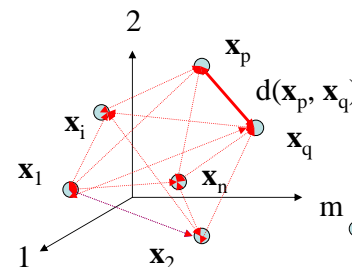
自己組織化法

#### 2. Supervised Learning

ベイズ法

線形識別関数法

種々の人工ニューラルネット



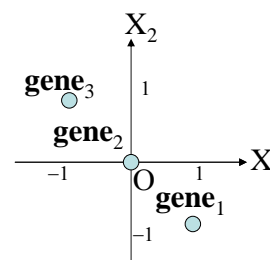
## 4.2 特徴抽出

### ピアソン相関係数とコサイン係数

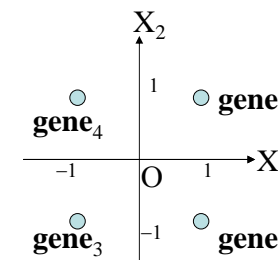
$$r_{ij} = \frac{\sum_{j=1}^M (x_{ij} - \bar{x}_i)(x_{ij} - \bar{x}_j)}{\sqrt{\sum_{j=1}^M (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^M (x_{ij} - \bar{x}_j)^2}}$$

$$\cos_{ij} = \frac{\sum_{j=1}^M x_{ij} x_{ij}'}{\sqrt{\sum_{j=1}^M x_{ij}^2 \sum_{j=1}^M x_{ij}'^2}}$$

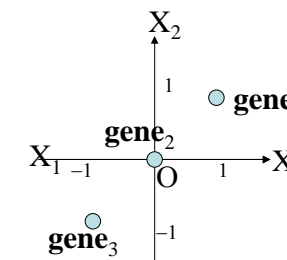
## 相関係数を求めてみよう！



(a)	X <sub>1</sub>	X <sub>2</sub>
gene <sub>1</sub>	1	-1
gene <sub>2</sub>	0	0
gene <sub>3</sub>	-1	1

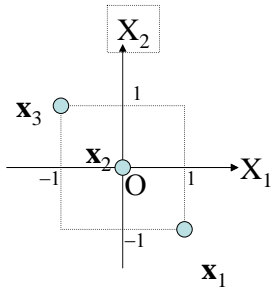


(b)	X <sub>1</sub>	X <sub>2</sub>
gene <sub>1</sub>	1	1
gene <sub>2</sub>	1	-1
gene <sub>3</sub>	-1	-1
gene <sub>4</sub>	-1	1



(3)	X <sub>1</sub>	X <sub>2</sub>
gene <sub>1</sub>	1	1
gene <sub>2</sub>	0	0
gene <sub>3</sub>	-1	-1

(Ex. 3 . Calc.)



$$\begin{aligned} X_1 & X_2 \\ x_1 &= (-1, 1) \\ x_2 &= (0, 0) \\ x_3 &= (1, -1) \end{aligned}$$

$$r_{ii'} = \frac{\sum_{j=1}^M (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^M (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^M (x_{i'j} - \bar{x}_{i'})^2}}$$

$$\begin{aligned} s_{12} &= -1.0 \\ s_{11} &= 1.0 \\ s_{22} &= 1.0 \\ r_{11} &= 1.0 \\ r_{12} &= -1.0 \end{aligned}$$

問題 11 異なった培養条件 1-4 について二色蛍光法マイクロアレイ測定をおこない変異株と野生株の間に以下の対数比を得た。遺伝子間のピアソン相関係数およびコサイン係数を求めよ。

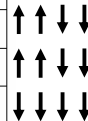
遺伝子 ID	実験 1	実験 2	実験 3	実験 4
g1	0.30	0.10	-0.10	-0.30
g2	0.60	0.20	-0.20	-0.60
g3	-0.30	-0.50	-0.70	-0.90

問題 11 異なった培養条件 1-4 について二色蛍光法マイクロアレイ測定をおこない変異株と野生株の間に以下の対数比を得た。遺伝子間のピアソン相関係数およびコサイン係数を求めよ。

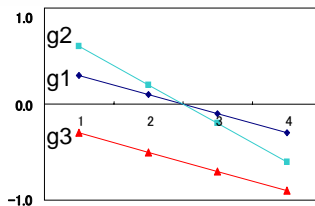
遺伝子 ID	実験 1	実験 2	実験 3	実験 4
g1	0.30	0.10	-0.10	-0.30
g2	0.60	0.20	-0.20	-0.60
g3	-0.30	-0.50	-0.70	-0.90

p.31

	ピアソン相関係数			コサイン係数		
	g1	g2	g3	g1	g2	g3
g1	1.0	1.00	1.00	1.00	1.00	0.35
g2		1.00	1.00		1.00	0.35
g3			1.00			1.00

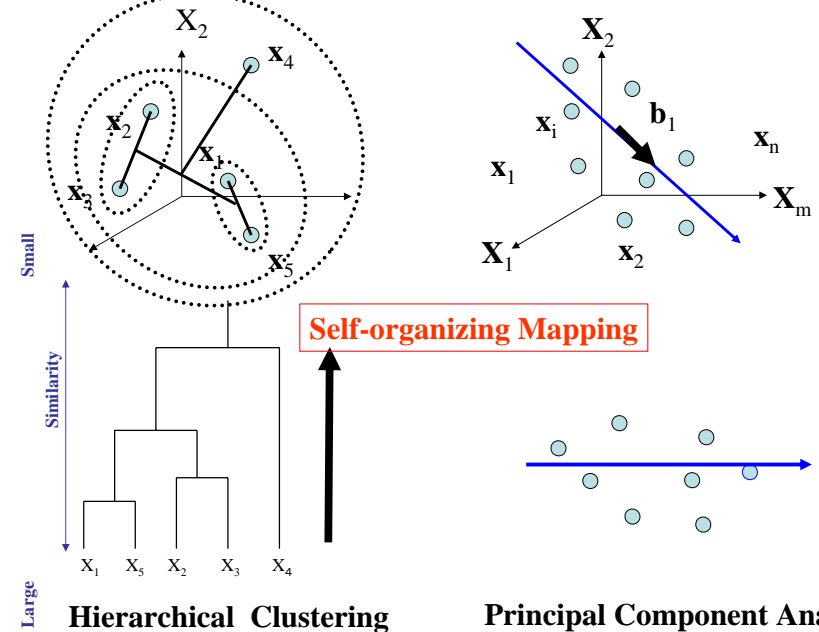


$$r_{ii'} = \frac{\sum_{j=1}^M (x_{ij} - \bar{x}_i)(x_{i'j} - \bar{x}_{i'})}{\sqrt{\sum_{j=1}^M (x_{ij} - \bar{x}_i)^2 \sum_{j=1}^M (x_{i'j} - \bar{x}_{i'})^2}}$$



$$\cos_{ii'} = \frac{\sum_{j=1}^M x_{ij} x_{i'j}}{\sqrt{\sum_{j=1}^M x_{ij}^2 \sum_{j=1}^M x_{i'j}^2}}$$

4.3 教師なし学習

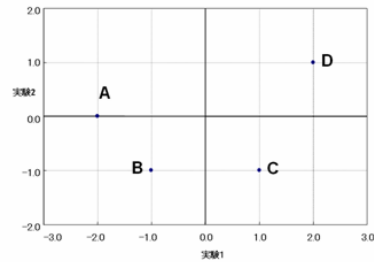


Self-organizing Mapping

Hierarchical Clustering

Principal Component Analysis

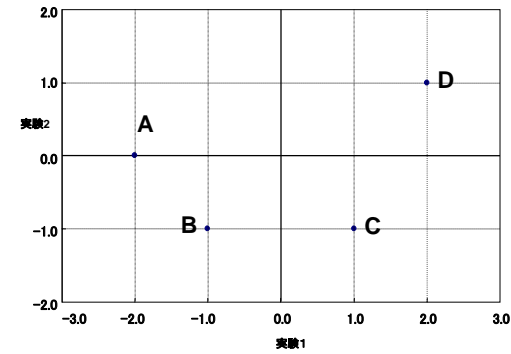
問題 12



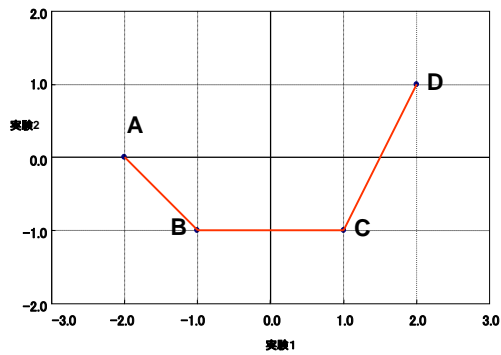
二つの異なる条件でマイクロレイ実験をおこなったところ4つの遺伝子 A,B,C,D に対して図 10 のような散布図を得た。お互いのユークリッド距離が小さいものから線でつなぎ、4つの遺伝子全てが3本の線でつながったところで終了せよ。M個の変量における遺伝子  $s_t$  と  $t$  の

ユークリッド距離は  $d(x_s, x_t) = \sqrt{\sum_{j=1}^M (x_{sj} - x_{tj})^2}$  である。

図 12 二つの実験における二次元プロットの例



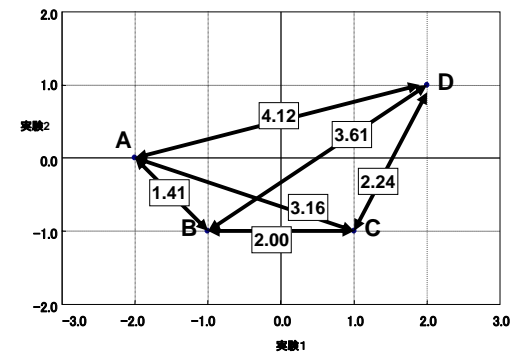
階層的クラスタリング



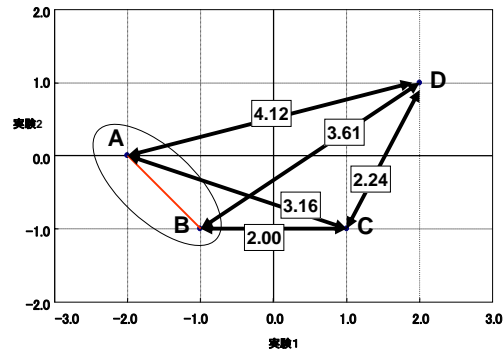
階層的クラスタリング:最短距離法 (p.34)

最小距離法

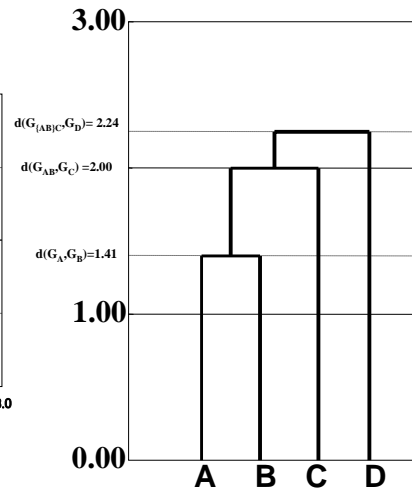
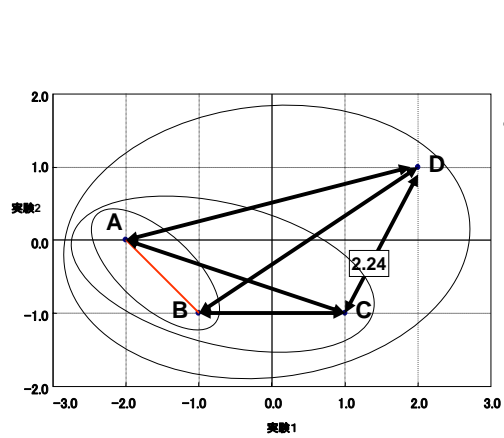
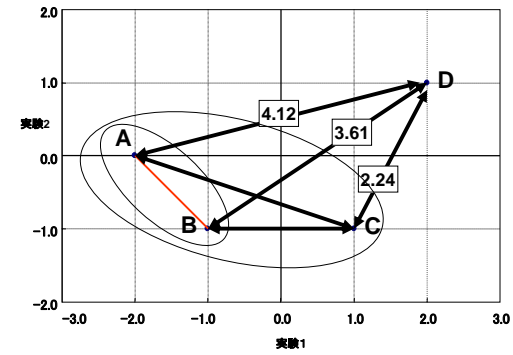
- (i) グループ間の代表距離をそれぞれのグループに属する要素間の最小距離と定義する。
- (ii) はじめに、全ての遺伝子は個別のグループに帰属するものとする。
- (iii) グループ間の距離が小さい順にグループを融合し、一つのグループになったら終了する。



階層的クラスタリング



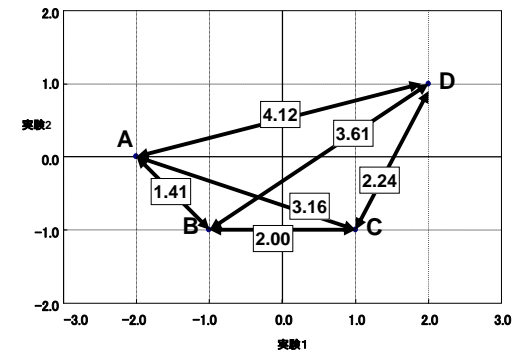
階層的クラスタリング



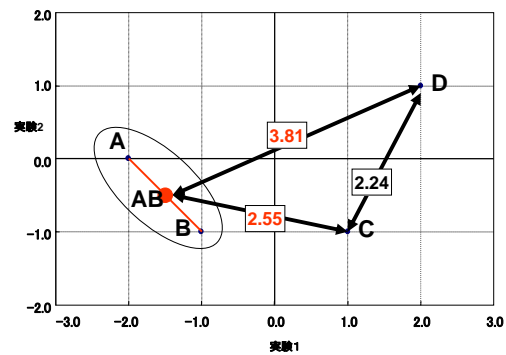
問題13 (p.35)

階層的クラスター分析法	グループ間距離
Nearest Neighbor(最近隣法)	それぞれのグループにおける要素間の距離の最小値
Furthest Neighbor(最遠隣法)	それぞれのグループにおける要素間の距離の最大値
Centroid(重心間距離法)	それぞれのグループにおける要素から求めた重心間の距離
Average(平均距離法)	それぞれのグループにおける要素間の距離の平均値

問題 13 問題 12 のデータを用いて重心距離法によりクラスター分析をせよ。グループ間の距離にはユークリッド距離を用いることとする。



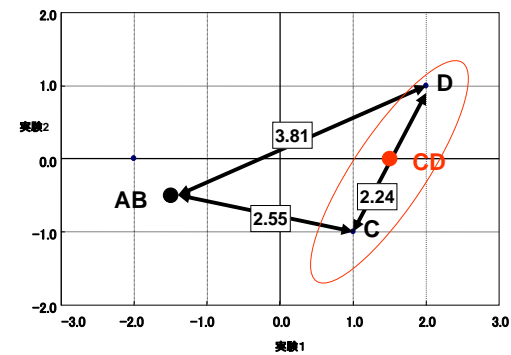
問題13: Step 1



	C	D
AB	2.55	3.81
C		2.24

AB (-1.5, -0.5)

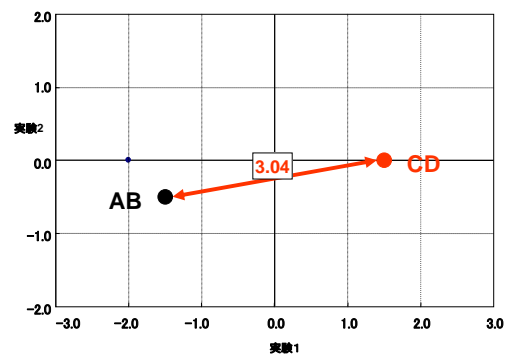
問題13: Step 2



	C	D
AB	2.55	3.81
C		2.24

CD (1.5, 0.0)

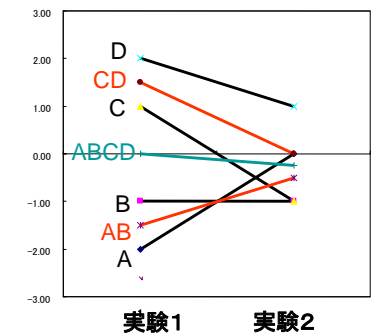
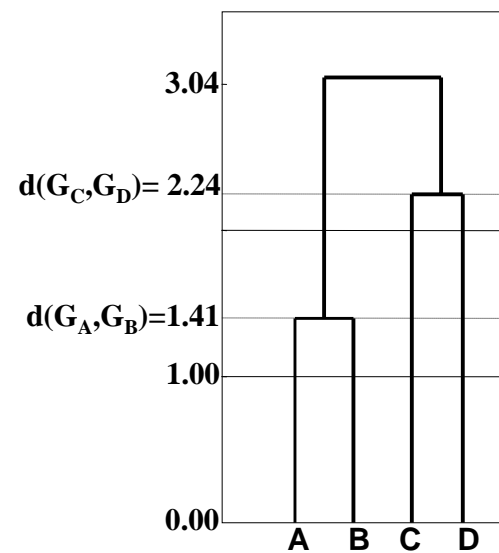
問題13: Step 3

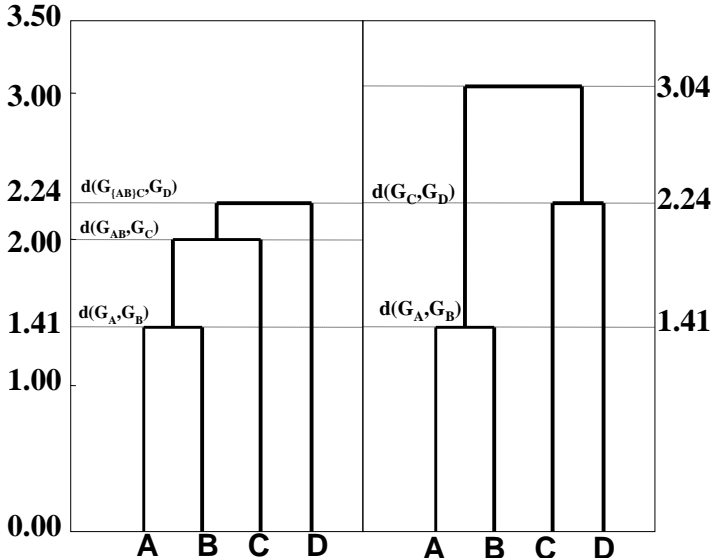


	CD
AB	3.04

ABCD (1.5, 0.0)

問題13





問題 14 二つの異なる条件でマイクロレイ実験をおこなったところ 5つの遺伝子 A,B,C,D,E に対して左図のような散布図を得た。最小距離法によりデンドログラムを作成せよ。なお距離にはユークリッド距離は  $d(x_s, x_t) = \sqrt{\sum_{j=1}^M (x_{sj} - x_{tj})^2}$  を用いよ。

4.5主成分分析法

Journal of Educational Psychology (1933)  
24, 417-441.

Hotelling (1933)

ANALYSIS OF A COMPLEX OF STATISTICAL VARIABLES INTO PRINCIPAL COMPONENTS<sup>1</sup>

HAROLD HOTELLING  
Columbia University

I. INTRODUCTION

Consider  $n$  variables attaching to each individual of a population. These statistical variables  $x_1, x_2, \dots, x_n$  might for example be scores made by school children in tests of speed and skill in solving arithmetical problems or in reading; or they might be various physical properties of telephone poles, or the rates of exchange among various currencies. The  $x$ 's will ordinarily be correlated. It is natural to ask whether some more fundamental set of independent variables exists, perhaps fewer in number than the  $x$ 's, which determine the values the  $x$ 's will take. If  $\gamma_1, \gamma_2, \dots$  are such variables, we shall then have a set of relations of the form

$$x_i = f_i(\gamma_1, \gamma_2, \dots) \quad (i = 1, 2, \dots, n) \quad (1)$$

Quantities such as the  $\gamma$ 's have been called mental factors in recent psychological literature. However in view of the prospect of application of these ideas outside of psychology, and the conflicting usage attaching to the word "factor" in mathematics, it will be better simply to call the  $\gamma$ 's components of the complex depicted by the tests.

We shall consider only normally distributed systems of components having zero correlations and unit variances. If we use the symbol  $E$  to denote the expectation, or mean value in the population, of the quantity following it, the condition that the means shall be zero is expressed by

$$E\gamma_i = 0.$$

The assumptions of unit variances and zero correlations may be combined in the statement

<sup>1</sup> A study made in part under the auspices of the Unitary Traits Committee and the Carnegie Corporation.  
The author is indebted to Professor Truman L. Kelley, who was responsible

## 4.5 主成分分析法(PCA) p.37

主成分スコア

$$\begin{aligned} Z_1 &= a_{11}X_1 + \dots + a_{1j}X_j + \dots + a_{1M}X_M \\ &\dots \\ Z_j &= a_{j1}X_1 + \dots + a_{jj}X_j + \dots + a_{jM}X_M \\ &\dots \\ Z_M &= a_{M1}X_1 + \dots + a_{Mj}X_j + \dots + a_{MM}X_M \end{aligned}$$

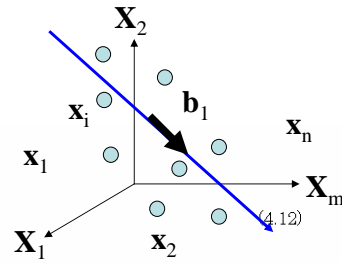
寄与率

オリジナルの多変量データの全分散に対する  $k$  番目の主成分の分散を **寄与率**(4.13)と呼ぶ。寄与率が高いほどオリジナルデータの分布が反映される。

$$\%Var[Z_k] = \frac{V[Z_k]}{\sum_{u=1}^M V[X_u]} \cdot 100 \quad (4.13)$$

因子負荷量(Factor Loadings)

$r(X_j, Z_k)$  : 第  $j$  番目の変量と第  $k$  主成分における相関係数

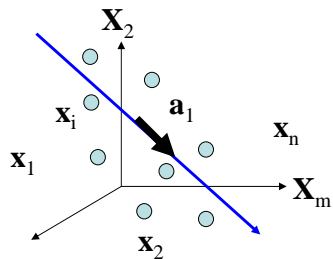


## 問題15 (p.40)

問題 15 4つの遺伝子について3種のマイクロアレイ測定を行ったところ以下のような値を得た。遺伝子を主成分第1軸および第2軸にプロットし、因子負荷量および寄与率を求めてこのデータを解釈せよ。

	実験1	実験2	実験3
遺伝子1	-1.0	-1.0	2.0
遺伝子2	1.0	-1.0	0.0
遺伝子3	-1.0	1.0	0.0
遺伝子4	1.0	1.0	-2.0
遺伝子5	0.0	0.0	0.0

### 問題15:主成分分析



$$\begin{aligned} Z_1 &= a_{11}X_1 + \dots + a_{1j}X_j + \dots + a_{1M}X_M \\ &\dots \\ Z_M &= a_{M1}X_1 + \dots + a_{Mj}X_j + \dots + a_{MM}X_M \end{aligned}$$

これを求める!

$$\begin{pmatrix} \text{cov}[X_1, X_1] & \dots & \text{cov}[X_1, X_j] & \dots & \text{cov}[X_1, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_j, X_1] & \dots & \text{cov}[X_j, X_j] & \dots & \text{cov}[X_j, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_M, X_1] & \dots & \text{cov}[X_M, X_j] & \dots & \text{cov}[X_M, X_M] \end{pmatrix} \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} = \lambda \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} \quad (3.18)$$

$$\text{cov}[X_s, X_t] = \frac{1}{N-1} \sum_{i=1}^N (x_{is} - \bar{x}_s)(x_{it} - \bar{x}_t)$$

### 問題15:

(p.40)

$$\begin{pmatrix} \text{cov}[X_1, X_1] & \dots & \text{cov}[X_1, X_j] & \dots & \text{cov}[X_1, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_j, X_1] & \dots & \text{cov}[X_j, X_j] & \dots & \text{cov}[X_j, X_M] \\ \dots & \dots & \dots & \dots & \dots \\ \text{cov}[X_M, X_1] & \dots & \text{cov}[X_M, X_j] & \dots & \text{cov}[X_M, X_M] \end{pmatrix} \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} = \lambda \begin{pmatrix} a_{11} \\ \dots \\ a_{1j} \\ \dots \\ a_{1M} \end{pmatrix} \quad (3.18)$$

	実験1	実験2	実験3
遺伝子1	-1.0	-1.0	2.0
遺伝子2	1.0	-1.0	0.0
遺伝子3	-1.0	1.0	0.0
遺伝子4	1.0	1.0	-2.0
遺伝子5	0.0	0.0	0.0

$$\begin{pmatrix} \frac{1}{4}((-1)^2 + 1^2 + (-1)^2 + 1^2 + 0^2) & \frac{1}{4}((-1)^2 + 1 \cdot (-1) + (-1) \cdot 1 + 1^2 + 0^2) & \frac{1}{4}((-1) \cdot 2 + 1 \cdot 0 + (-1) \cdot 0 + 1 \cdot (-2) + 0^2) \\ \frac{1}{4}((-1)^2 + 1 \cdot (-1) + (-1) \cdot 1 + 1^2 + 0^2) & \frac{1}{4}((-1)^2 + (-1)^2 + 1^2 + 1^2 + 0^2) & \frac{1}{4}((-1) \cdot 2 + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot (-2) + 0^2) \\ \frac{1}{4}((-1) \cdot 2 + 1 \cdot 0 + (-1) \cdot 0 + 1 \cdot (-2) + 0^2) & \frac{1}{4}((-1) \cdot 2 + (-1) \cdot 0 + 1 \cdot 0 + 1 \cdot (-2) + 0^2) & \frac{1}{4}(2^2 + 0^2 + 0^2 + (-2)^2 + 0^2) \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \lambda \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \quad (\text{p.40})$$

$$\begin{vmatrix} 1-\lambda & 0 & -1 \\ 0 & 1-\lambda & -1 \\ -1 & -1 & 2-\lambda \end{vmatrix} = (1-\lambda)(1-\lambda)(2-\lambda) - (1-\lambda) - (1-\lambda) = 0$$

$$\lambda = 3, 1, 0$$

$$Z_1 = -\frac{1}{2}\sqrt{\frac{2}{3}}X_1 - \frac{1}{2}\sqrt{\frac{2}{3}}X_2 + \sqrt{\frac{2}{3}}X_3$$

$$Z_2 = -\sqrt{\frac{1}{2}}X_1 + \sqrt{\frac{1}{2}}X_2$$

### 主成分スコア

(p.42)

$$Z_1 = -\frac{1}{2}\sqrt{\frac{2}{3}}X_1 - \frac{1}{2}\sqrt{\frac{2}{3}}X_2 + \sqrt{\frac{2}{3}}X_3$$

$$Z_2 = -\sqrt{\frac{1}{2}}X_1 + \sqrt{\frac{1}{2}}X_2$$

遺伝子	実験1	実験2	実験3	Z <sub>1</sub>	Z <sub>2</sub>
遺伝子1	-1.0	-1.0	2.0	$3\sqrt{\frac{2}{3}}$	0
遺伝子2	1.0	-1.0	0.0	0	$-2\sqrt{\frac{1}{2}}$
遺伝子3	-1.0	1.0	0.0	0	$2\sqrt{\frac{1}{2}}$
遺伝子4	1.0	1.0	-2.0	$-3\sqrt{\frac{2}{3}}$	0
遺伝子5	0.0	0.0	0.0	0	0
分散	1	1	2	8	1

### 因子負荷量 (Factor loading)

(p.43)

遺伝子	実験1	実験2	実験3	Z <sub>1</sub>	Z <sub>2</sub>
遺伝子1	-1.0	-1.0	2.0	$3\sqrt{\frac{2}{3}}$	0
遺伝子2	1.0	-1.0	0.0	0	$-2\sqrt{\frac{1}{2}}$
遺伝子3	-1.0	1.0	0.0	0	$2\sqrt{\frac{1}{2}}$
遺伝子4	1.0	1.0	-2.0	$-3\sqrt{\frac{2}{3}}$	0
遺伝子5	0.0	0.0	0.0	0	0
分散	1	1	2	8	1

相関

	Z <sub>1</sub>	Z <sub>2</sub>
実験1	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{2}$
実験2	$-\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{2}$
実験3	1.00	0

### 寄与率 λ = 3, 1, 0 より

解析結果のまとめ

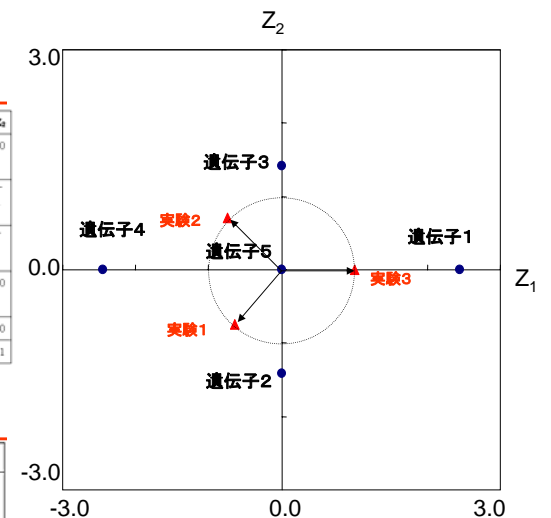
3/4 x 100 = 75% (第1主成分)  
1/4 x 100 = 25% (第2主成分)

### 主成分スコア

遺伝子	実験1	実験2	実験3	Z <sub>1</sub>	Z <sub>2</sub>
遺伝子1	-1.0	-1.0	2.0	$3\sqrt{\frac{2}{3}}$	0
遺伝子2	1.0	-1.0	0.0	0	$-2\sqrt{\frac{1}{2}}$
遺伝子3	-1.0	1.0	0.0	0	$2\sqrt{\frac{1}{2}}$
遺伝子4	1.0	1.0	-2.0	$-3\sqrt{\frac{2}{3}}$	0
遺伝子5	0.0	0.0	0.0	0	0
分散	1	1	2	8	1

### 因子負荷量 (Factor loading)

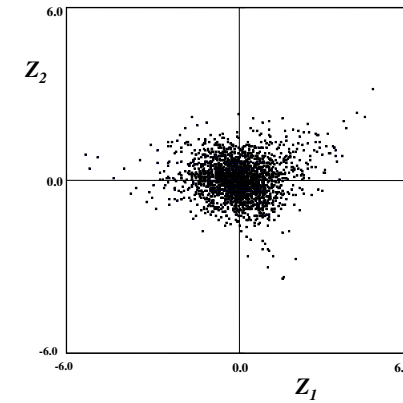
	Z <sub>1</sub>	Z <sub>2</sub>
実験1	$-\frac{1}{2}\sqrt{2}$	$-\frac{1}{2}\sqrt{2}$
実験2	$-\frac{1}{2}\sqrt{2}$	$\frac{1}{2}\sqrt{2}$
実験3	1.00	0



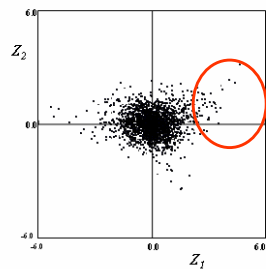
# PCA analysis of microarray data

Microarray Data for 6 mutants (stationary-phase)  
LRP, HNS, Hu, IHF, FIS, CRP

Scatter plot for genes between 1<sup>st</sup> and 2<sup>nd</sup> PCs

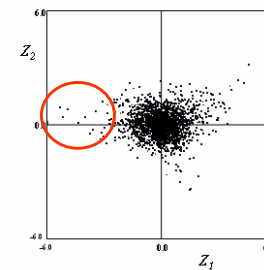


## Positive Z1



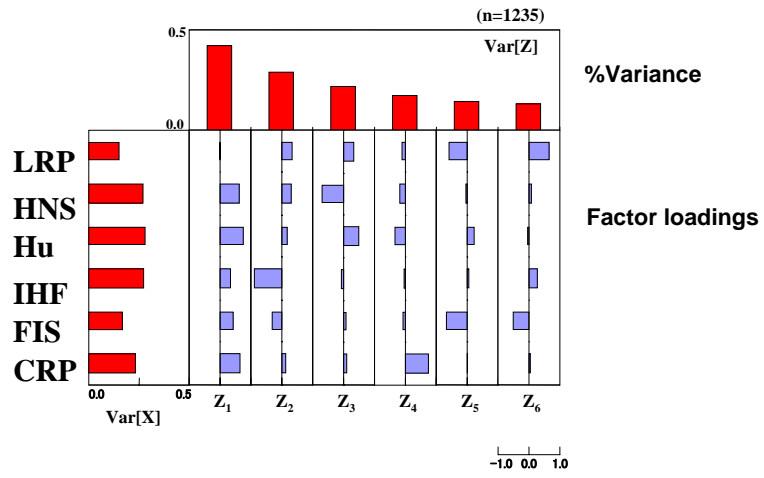
	$Z_1$	$Z_2$	$Z_3$
<i>hdeB</i> :	4.66	<i>hdeB</i> : 3.13	<i>261#9</i> : 1.89
<i>ycaC</i> :	4.40	<i>yedU</i> : 2.35	<i>yciG</i> ): 1.69
<i>yedU</i> :	4.12	<i>trpG</i> : 2.30	<i>narG</i> : 1.62
<i>yjgB</i> :	3.72	<i>yefE</i> : 2.16	<i>udk</i> : 1.61
<i>msyB</i> :	3.65	<i>ycaC</i> : 2.16	<i>frdA</i> : 1.60
<i>ynhC</i> :	3.60	<i>333#1</i> : 2.13	<i>yfiD</i> : 1.54
<i>aceB</i> :	3.50	<i>osmB</i> : 2.11	<i>purB</i> : 1.53
<i>msyB</i> :	3.43	<i>yhaH</i> : 2.05	<i>274#5</i> : 1.53
<i>katE</i> :	3.37	<i>msyB</i> : 2.02	<i>ydbH</i> : 1.52
<i>bfr</i> :	3.35	<i>otsB</i> : 1.97	<i>parE</i> : 1.49
<i>yeaG</i> :	3.30	<i>yccL</i> : 1.96	<i>himD</i> : 1.47
<i>amyA</i> :	3.30	<i>yhiX</i> : 1.96	<i>nirB</i> : 1.43
<i>yeaG</i> :	3.25	<i>427#1</i> : 1.90	<i>pgi</i> : 1.41
<i>ynhE</i> :	3.16	<i>gcvT</i> : 1.80	<i>nirB</i> : 1.37
<i>yzzM</i> :	3.13	<i>ompT</i> : 1.79	<i>oppC</i> : 1.37
<i>yciG</i> :	3.09	<i>yjgB</i> : 1.79	<i>ugpA</i> : 1.34
<i>ygaM</i> :	3.03	<i>334#5.1</i> : 1.74	<i>yojI</i> : 1.33
<i>gttF</i> :	2.98	<i>arp</i> : 1.72	<i>dicA</i> : 1.33
<i>yhjG</i> :	2.98	<i>glgS</i> : 1.72	<i>yajQ</i> : 1.31
<i>aceA</i> :	2.96	<i>334#5.1</i> : 1.70	<i>ydcY</i> : 1.30

## Negative Z1

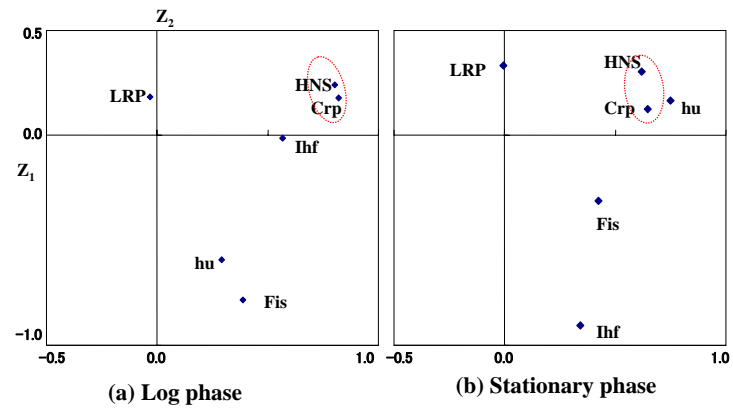


	$Z_1$	$Z_2$	$Z_3$
<i>frdA</i> :	-2.62	<i>hoxK</i> : -1.60	<i>mppA</i> : -1.68
<i>malM</i> :	-2.65	<i>ispA</i> : -1.61	<i>cspB</i> : -1.68
<i>treC</i> :	-2.74	<i>glyQ</i> : -1.64	<i>sucD</i> : -1.69
<i>nirB</i> :	-2.78	<i>cytR</i> : -1.78	<i>malE</i> : -1.71
<i>ygiN</i> :	-2.84	<i>nuoB</i> : -1.81	<i>aldA</i> : -1.74
<i>yfiA</i> :	-2.86	<i>xerB</i> : -1.81	<i>cobU</i> : -1.76
<i>crp</i> :	-2.88	<i>ybdE</i> : -1.86	<i>416#3</i> : -1.88
<i>ompF</i> :	-3.01	<i>cysS</i> : -1.98	<i>mgIB</i> : -1.88
<i>glpQ</i> :	-3.03	<i>cysS</i> : -2.10	<i>ynhE</i> : -1.89
<i>yfiA</i> :	-3.10	<i>ybdE</i> : -2.11	<i>dadX/alr2</i> : -1.90
<i>malF</i> :	-3.12	<i>yhdG</i> : -2.22	<i>hisC</i> : -1.90
<i>malF</i> :	-3.23	<i>yccC</i> : -2.40	<i>aldA</i> : -2.00
<i>nmpC</i> :	-3.45	<i>prpR</i> : -2.41	<i>yigM</i> : -2.01
<i>lamB</i> :	-3.60	<i>prpR</i> : -2.45	<i>dppA</i> : -2.02
<i>ansB</i> :	-3.78	<i>yhdG</i> : -2.65	<i>crp</i> : -2.06
<i>lamB</i> :	-3.99	<i>yidD</i> : -2.68	<i>ptkB</i> : -2.41
<i>malK</i> :	-4.37	<i>yagQ</i> : -2.74	<i>ydcW</i> : -2.52
<i>malK</i> :	-4.94	<i>sbmA</i> : -3.05	<i>fdnG</i> : -3.17
<i>malE</i> :	-5.20	<i>sbmA</i> : -3.38	<i>tnaA</i> : -3.43
<i>malE</i> :	-5.33	<i>paaX</i> : -3.45	<i>tnaA</i> : -3.55

# Microarray Data for 6 mutants (stationary-phase)

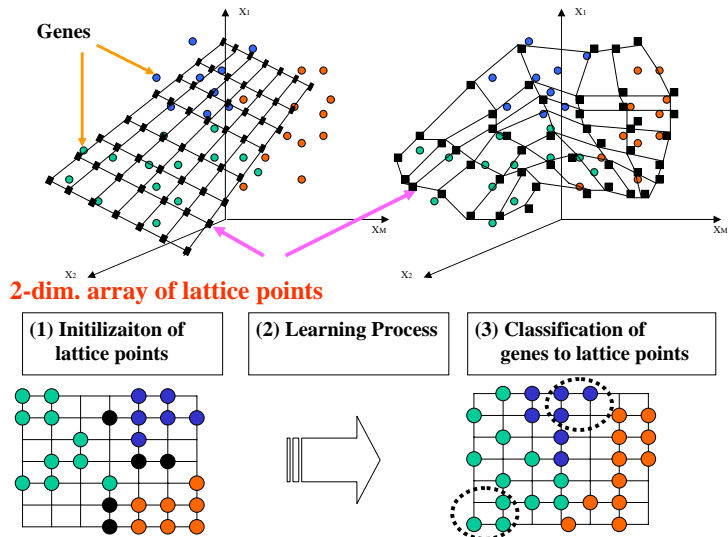


The first two component



## 4.7 Self-Organizing Mapping (Concept)

p.47



Visualization of gene distribution in multidimensional expression profile.

### BLSOM (Batch-Learning SOM)

**Parameters in learning process**

$$W_{ij}^{(new)} = W_{ij}^{(old)} + \alpha(t) \left( \frac{\sum_k x_k}{N_{ij}} - W_{ij}^{(old)} \right)$$

$$S_{ij} = \{W_{ij} \mid |i - \beta(t) \cdot \langle i \rangle| < i + \beta(t), |j - \beta(t) \cdot \langle j \rangle| < j + \beta(t)\}$$

$$\alpha(t) = \max \left\{ 0.01, \alpha_{init} \left( 1 - \frac{t}{T} \right) \right\}$$

$$\beta(t) = \max \{ 1, \beta_{init} - t \}$$

**Learning process of lattice points by input vectors (Iteration (t = 1, 2, ..., T))**

```

graph LR
    A[Initializing Weight Vectors (PCA)  
:reflecting data distribution] --> B[Classification of Input Vectors]
    B --> C[Updating Weight Vectors  
corresponding to lattice points]
    C --> D[Mapping of Genes]
    C --> E[Feature Map]
    
```

Batch-Learning: independent of order of input vectors

**Comparison of stages**  
Lattice points corresponding to gene expression levels in individual experimental points

Kanaya, S. et al. Gene (1999)  
Abe, T. et al. Genome Res. (2003)  
Hirai et al, Proc. Natl. Acad. Sci. USA, (2004).

In algorithm, we revised original SOM by two points  
(1) Applying PCA : reflecting data distribution  
(2) independent of order of input vectors

**PCA**

$$W_{ij}^0 = x_{mv} + 5\sigma_i \left\{ b_1 \left( \frac{i-1/2}{I} \right) + b_2 \left( \frac{j-1/2}{J} \right) \right\}$$

$x_k = (x_{k1}, x_{k2}, \dots, x_{kT})$

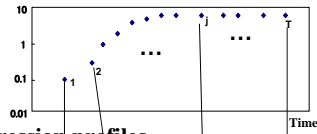
$W_{00}, W_{10}, W_{i0}, W_{0j}, W_{ij}, W_{1-j}, W_{0-j-1}, W_{i-1-1}, W_{i-1-j-1}$

Similar vectors

Similar vectors

## 遺伝子発現プロファイルの時系列解析

Growth curve



Expression profiles

Gene <sub>1</sub>	x <sub>11</sub>	x <sub>12</sub>	...	x <sub>1j</sub>	...	x <sub>1T</sub>	≡	Expression similarity
Gene <sub>2</sub>	x <sub>21</sub>	x <sub>22</sub>	...	x <sub>2j</sub>	...	x <sub>2T</sub>		
...	...	...	...	...	...	...		
Gene <sub>i</sub>	x <sub>i1</sub>	x <sub>i2</sub>	...	x <sub>ij</sub>	...	x <sub>iT</sub>		
Gene <sub>D</sub>	x <sub>D1</sub>	x <sub>D2</sub>	...	x <sub>Dj</sub>	...	x <sub>DT</sub>		

Expression similarity

Stage 1 2 ... j ... T

T, # of time-series microarray experiments  
D, # of genes in a microarray

Stage similarity

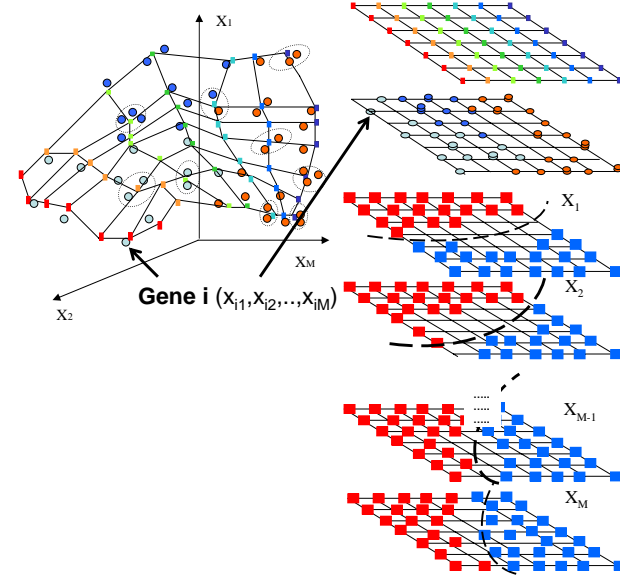
STATES

→ State-Transition

SOM : expression similarity of genes and stage similarity simultaneously.

When we measure time-series microarray, gene expression profile is represented by a matrix. SOM makes it possible to examine gene similarity and stage similarity simultaneously.

## Self-organizing Mapping (Summary)



Arrangement of lattice points in multi-dimensional expression space  
Lattice points are optimized for reflecting data distribution

Gene Classification  
Genes with similar expression profiles are clustered to identical or near lattice points

Feature Mapping  
In the i-th condition, lattice points containing highly (low) expressed genes are colored by red (blue).

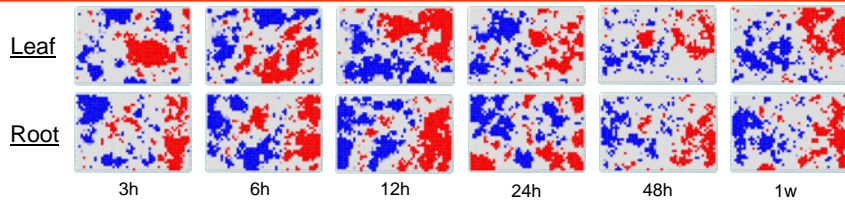
(ex.)  
■  $X_k > Th.(k)$   
■  $X_k < -Th.(k)$

Visually comparing among each stage of time-series data

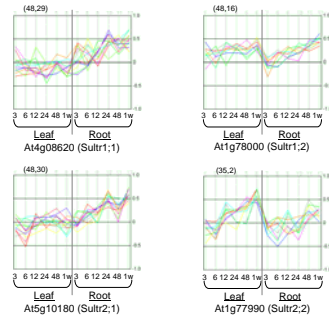
Non-linear projection of multi-dimensional expression profiles of genes. Original dimension is conserved in individual lattice points.

Several types of information is stored in SOM

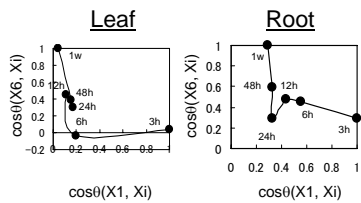
## Time series analysis of expression profile of *A. thaliana* by SOM (Sulfur starvation conditions after 3 w cultivation)



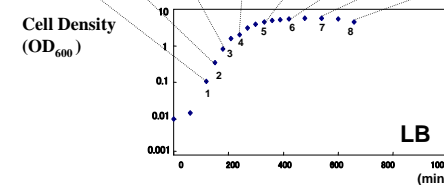
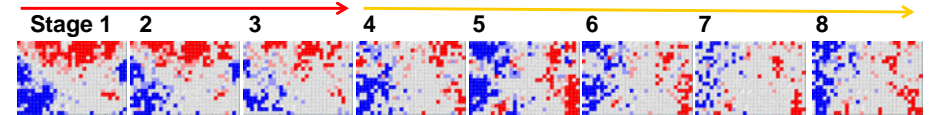
Expression profiles of genes encoding sulfate transporters



Correlation coefficients



## Time-series Expression Profile in *Bacillus subtilis* (cultivated in LB medium) (Data: Kazuo Kobayashi, Naotake Ogasawara (NAIST))



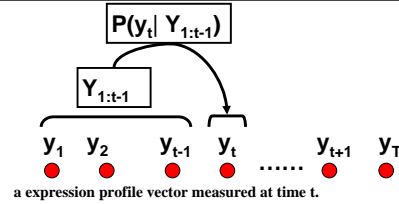
SOM for time-series expression profile

Validation:

“How do we estimate state-transition points based on statistics?”

Statistical estimation of state transitions (proposed by Ryoko Morioka Med.Eng.(in press) (Minato Lab.)

Time-series Data measured ( $Y_{1:T}$ )



Idea:

$P(y_t | Y_{1:t-1})$ : Prob. of  $y_t$  on condition of  $Y_{1:t-1}$   
 If  $P(y_t | Y_{1:t-1})$  is very small,  $y_t$  is independent of  $Y_{1:t-1}$ .  
 → Transition occurs between  $t-1$  and  $t$ .

Latent State transition model : assessment of the Prob.

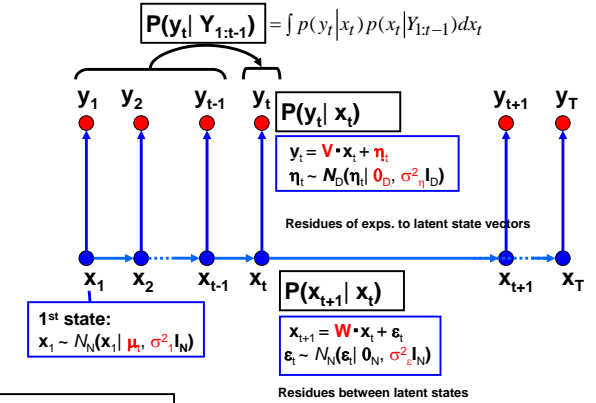
Model Structure for Statistical estimation of state transitions

Time-series Data measured ( $Y_{1:T}$ )

Latent states  $X_{1:T}$

Parameter estimation by EM-algorithm

$$\theta = \{\mu_1, \sigma_1, W, \sigma_\varepsilon, V, \sigma_\eta\}$$



Model Structure for statistical estimation of state transitions  
 Time-series data. Individual time-series measurements are linked linearly via state vectors  $x$  in latent model.

Parameters estimated by EM-algorithm

$$\theta = \{\mu_1, \sigma_1, W, \sigma_\varepsilon, V, \sigma_\eta\}$$

$$F[q_x(X_{1:T}), \theta] \equiv \int q_x(X_{1:T}) \log p(X_{1:T}, Y_{1:T} | \theta) dX_{1:T}$$

Here,  $q_x(X_{1:T}) \equiv p(X_{1:T} | Y_{1:T}, \bar{\theta})$

Expectation occurring latent states in condition of experimental data

(in k-loop)

$$q_x^{(k)}(X_{1:T}) \equiv p(X_{1:T} | Y_{1:T}, \theta^{(k-1)})$$

E-step: Maximization of F in the fixation of  $\theta^{(k)}$

$$F[q_x^{(k)}(X_{1:T}), \theta] \equiv \int q_x^{(k)}(X_{1:T}) \log p(X_{1:T}, Y_{1:T} | \theta) dX_{1:T} \quad (\text{update of log p()})$$

M-step: Maximization of log-likelihood by  $\theta^{(k)}$

$$\theta^{(k)} = \max_{\theta} \{ \int q_x^{(k)}(X_{1:T}) \log p(X_{1:T}, Y_{1:T} | \theta) dX_{1:T} \} \quad (\text{update of parameter set})$$

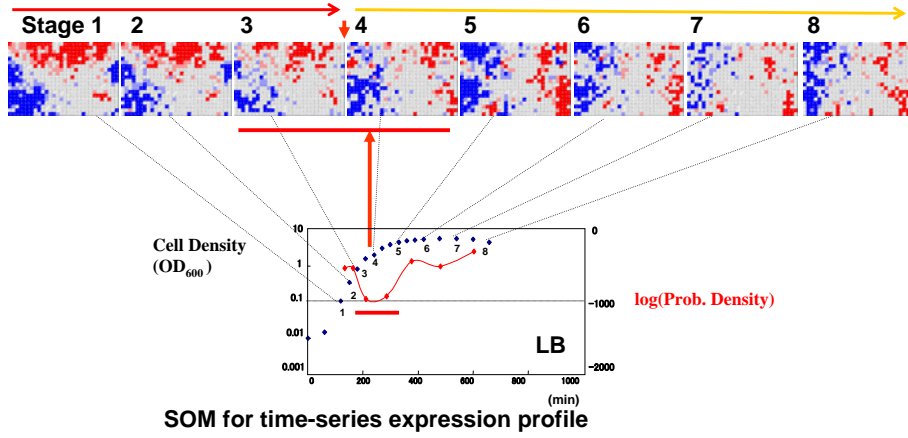
Expectation of latent states in condition of experimental data is maximized by EM algorithm. E step, M-step, iteration

Time series microarray data for 13 conditions (*B. subtilis*): state-transition points

Abb.	Growth Condition	# of experiments
LB	LB medium	8
HS	Heat shock	5
SS	NaCl shock	5
CS	cold shock	6
SOS	SOS stress	6
AG	anaerobic growth	6
MGM	MGM medium	8
GS	Glucose starvation	5
PS	Phosphate starvation	6
CM	competence medium	5
CSM	CS medium (Sporulation)	13
DSM	DSM medium (Sporulation)	8
DGG	DSM plus 2% Glucose, 0.1% Glutamine (inhibiting Sporulation)	6
		<b>Total 98</b>

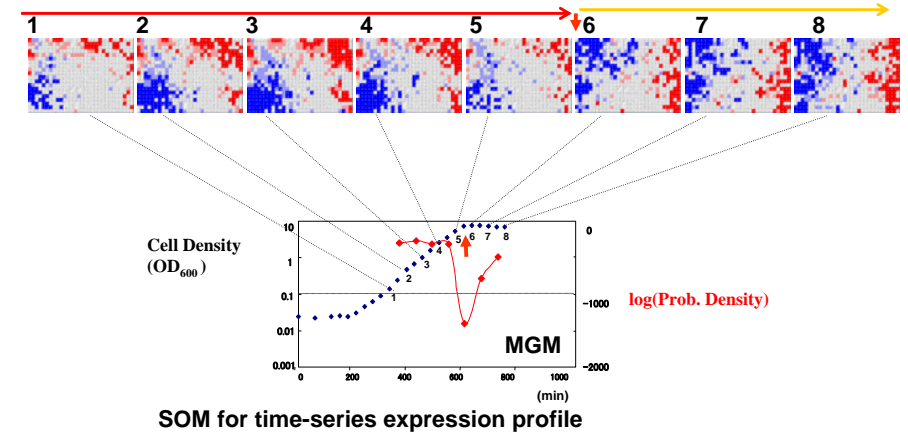
Data: Kobayashi, K. and Ogasawara, N (NAIST)

Expression Profile in *Bacillus subtilis* (LB medium)  
 (Data: Kazuo Kobayashi, Naotake Ogasawara (NAIST))

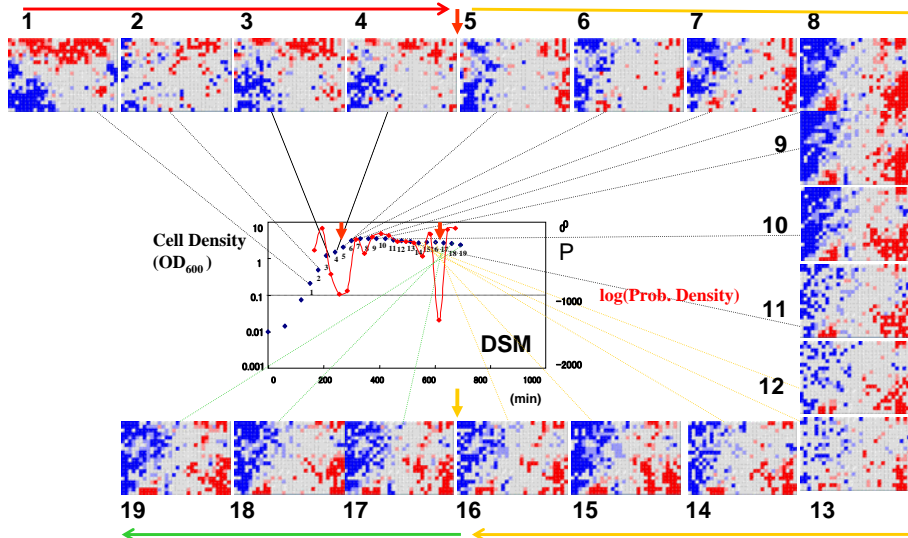


State transition point is observed between stages 3 and 4

Expression Profile in *Bacillus subtilis* (Minimal Glucose Medium)  
 (Data: Kazuo Kobayashi, Takeshi Ogasawara (NAIST))

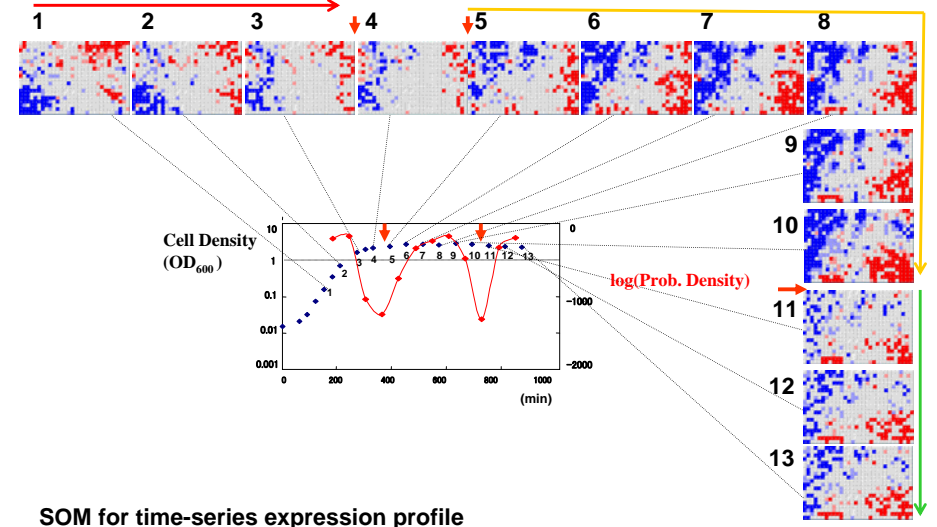


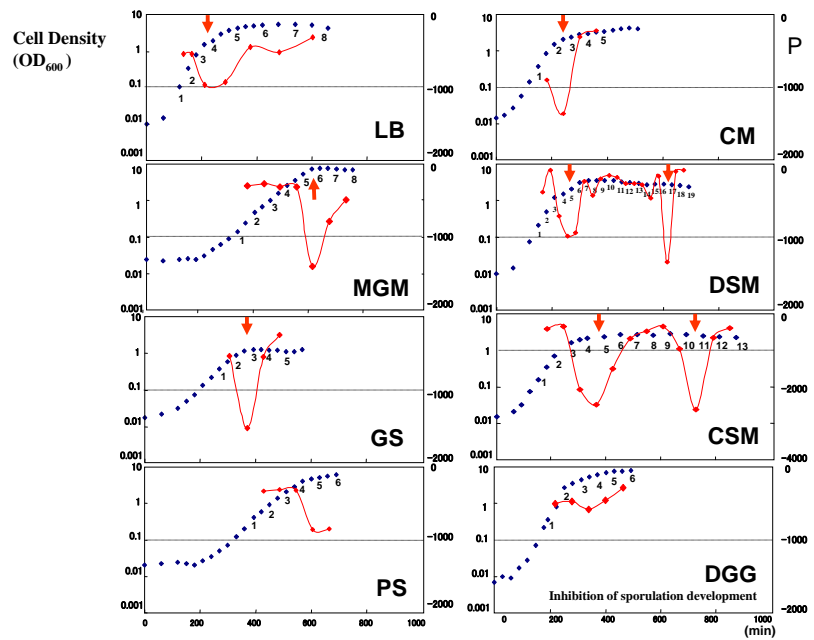
Expression Profile in *Bacillus subtilis* (DSM Sporulation Medium)  
 (Data: Kazuo Kobayashi, Takeshi Ogasawara (NAIST))



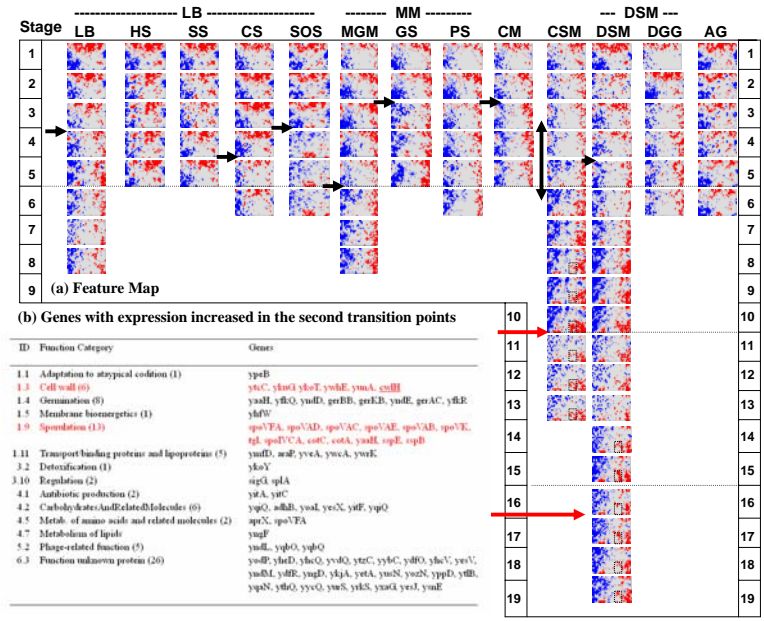
Relations of state transition point to growth curve and to expression profile

Expression Profile in *Bacillus subtilis* (Competence Sporulation Medium)  
 (Data: K. Kobayashi, N. Ogasawara (NAIST))

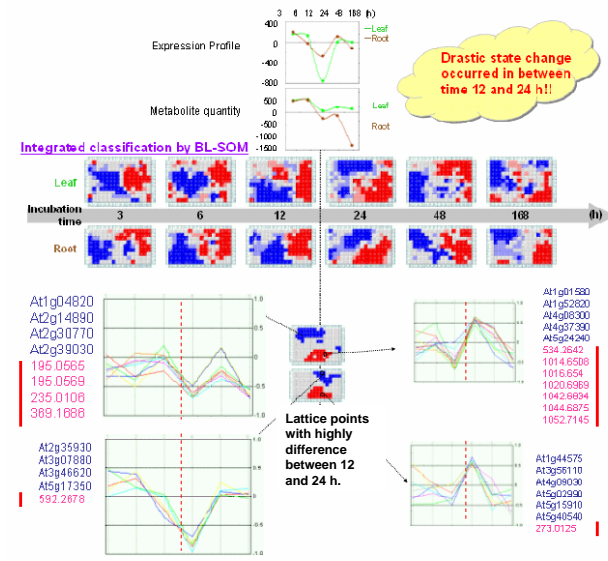
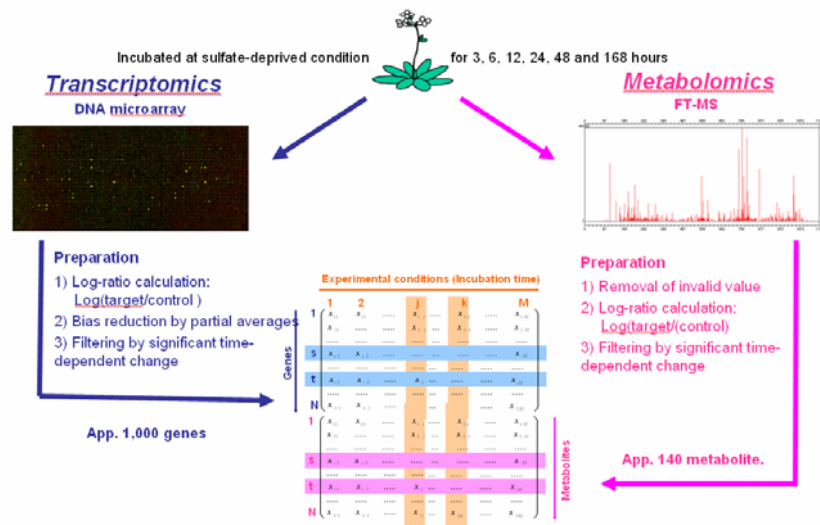




Summary of relations of transition points to growth curve. One significant transition point is obtained for LB, MGM, GS, CM. Two Transition points is obtained for DSM and CSM.



## [2] Integrated analysis of gene expression profile and metabolite quantity data of Arabidopsis thaliana (sulfur def./cont.)

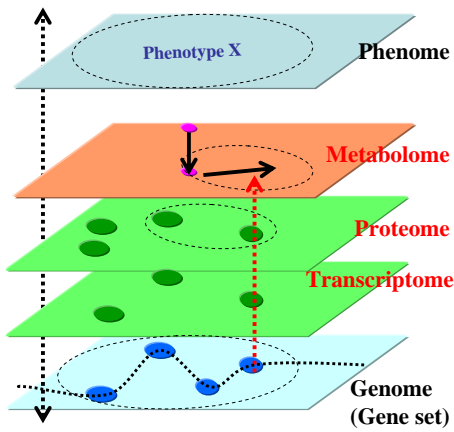


Accurate molecular weights and their quantities are measured by FT-MS!  
 → Candidate metabolites corresponding to accurate molecular weights  
 → **Metabolite Informatics**

# From Genome to Phenome

(Dynamic)

配列解析   多変量解析   ネットワーク解析



○

○

- プロファイル解析
- 2.プロファイル解析
- 1.コドン解析
- 4.メタボライトネットワーク
- 3.相互作用解析
- 転写制御
- ゲノム上の遺伝子編成

(Static)

Progressing genome projects, many kinds of “-omics” works have been progressed such as transcriptome, ....  
 These are dynamic information reflecting to Phenome.  
 Of them, metabolites are fundamentally important as molecular phenotype

# 問題14の図

