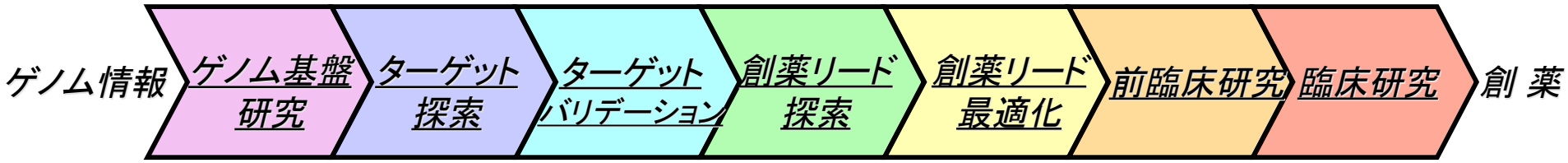
応用バイオインフォマティクス講義

# 遺伝子発現解析とデータマイニング

統合薬学教育開発分野
奥野　恭史
okuno@pharm.kyoto-u.ac.jp

# ゲノム情報から創薬へ




- 分子生物学技術
- 細胞生物学技術
- ゲノム解析

- 遺伝子発現解析
- プロテオーム解析
- 疾患モデル(KO)解析

- コンビナトリアルケミストリー
- ハイスループットスクリーニング
- 情報化学・計算化学

- ケミカルゲノミクス
- 医薬品合成化学
- 構造生物学

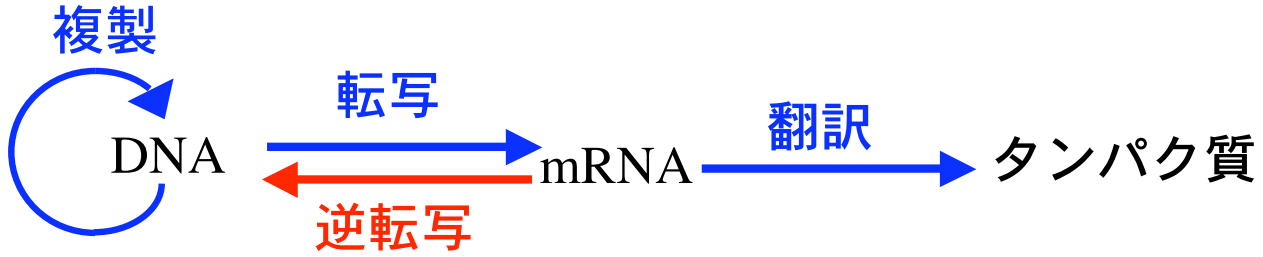
- 薬理学
- 薬物代謝学
- 薬物動態学

- バイオスタティスティクス
- 臨床インフォマティクス
- ゲノム多型解析

バイオインフォマティクス

# 機能ゲノミクス

・セントラルドグマ（ゲノム情報から機能発現までのプロセス）

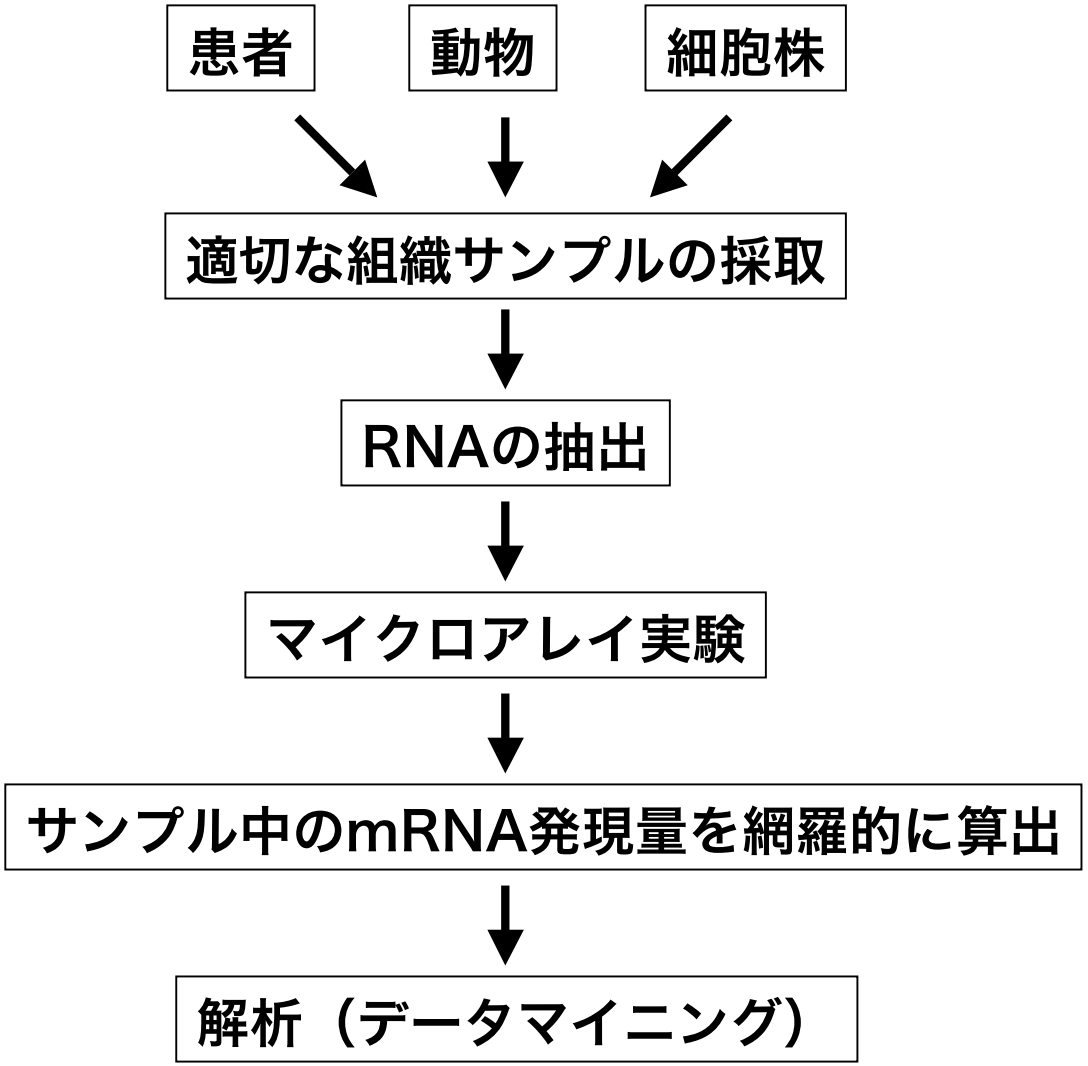



・生命の機能発現の大半はタンパク質が担う。
・mRNAはタンパク質の機能発現を示す指標の一つである。

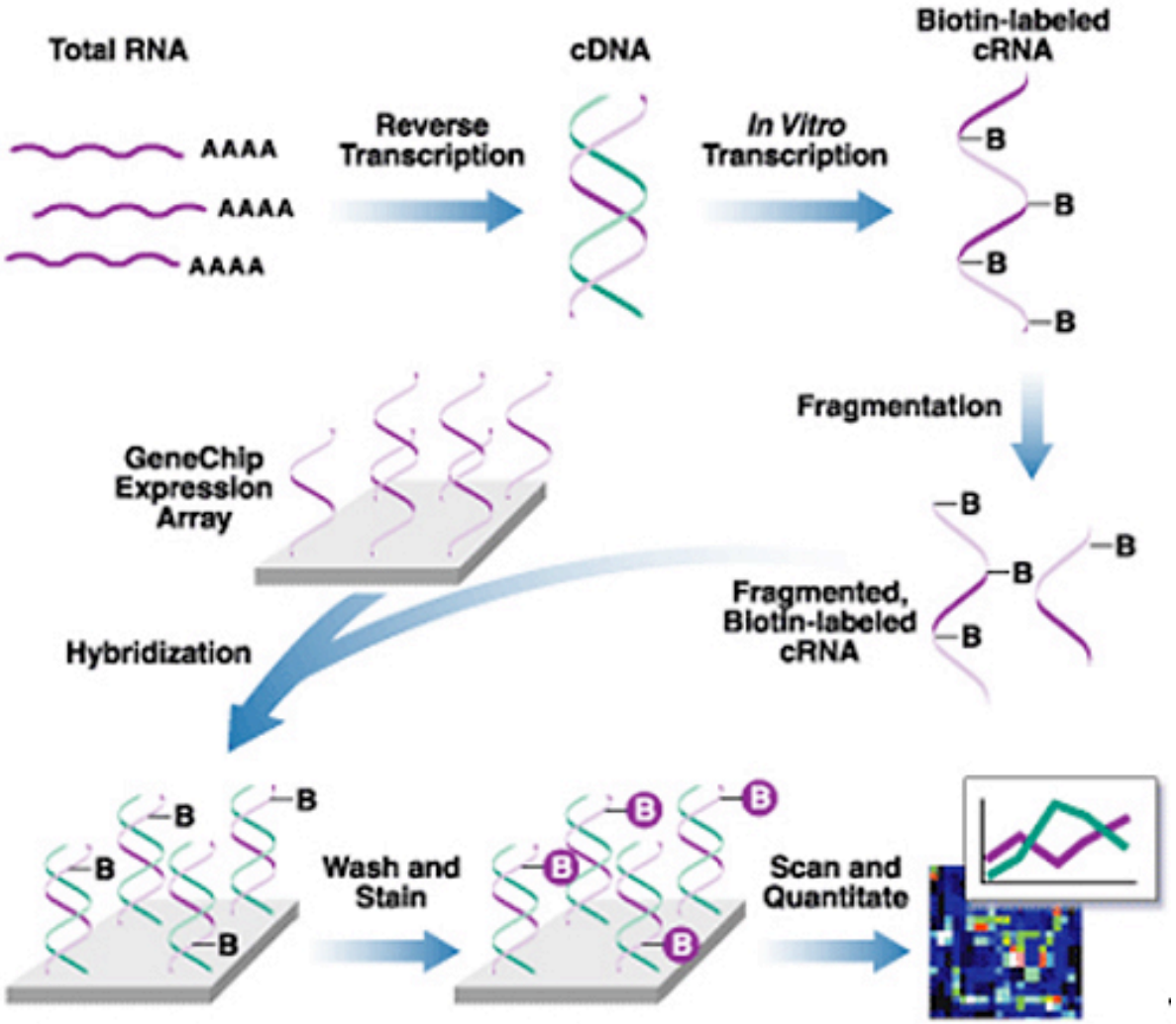
mRNAの発現量（遺伝子発現情報）を検出すれば
機能解析につながる

# マイクロアレイ

マイクロアレイは
mRNAの発現量
（遺伝子発現情報）
を網羅的に検出する



# マイクロアレイ解析フロー (GeneChip)




1.cDNA / cRNA合成

2.ハイブリダイゼーション

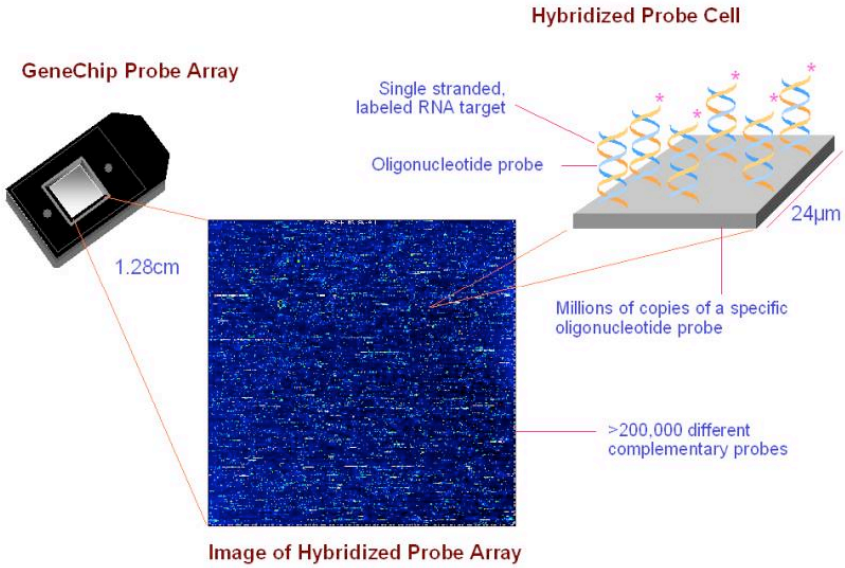
3.蛍光標識

4.スキャニング

5.データ解析

# DNAチップ

スライドガラス等の基盤上に数千～数万種類の遺伝子断片（プローブ）を固定させたもの

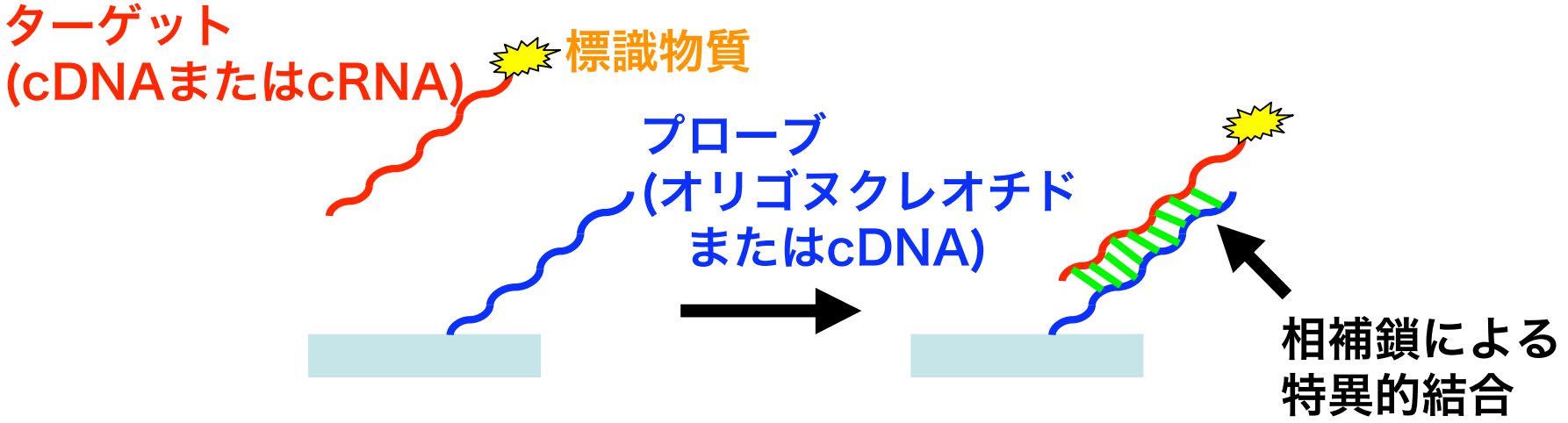



＜分類＞

・スポット型マイクロアレイ：Stanford方式
　　cDNAプローブ配列をガラススライドにスポットしたマイクロアレイ
　　Stanford大P.Brownが開発
・オリゴヌクレオチドマイクロアレイ：Affymetrix方式
　　２５塩基長のオリゴヌクレオチドをアレイ上に合成
　　Affymetrix社が開発、GeneChip
・スポット型オリゴヌクレオチドマイクロアレイ

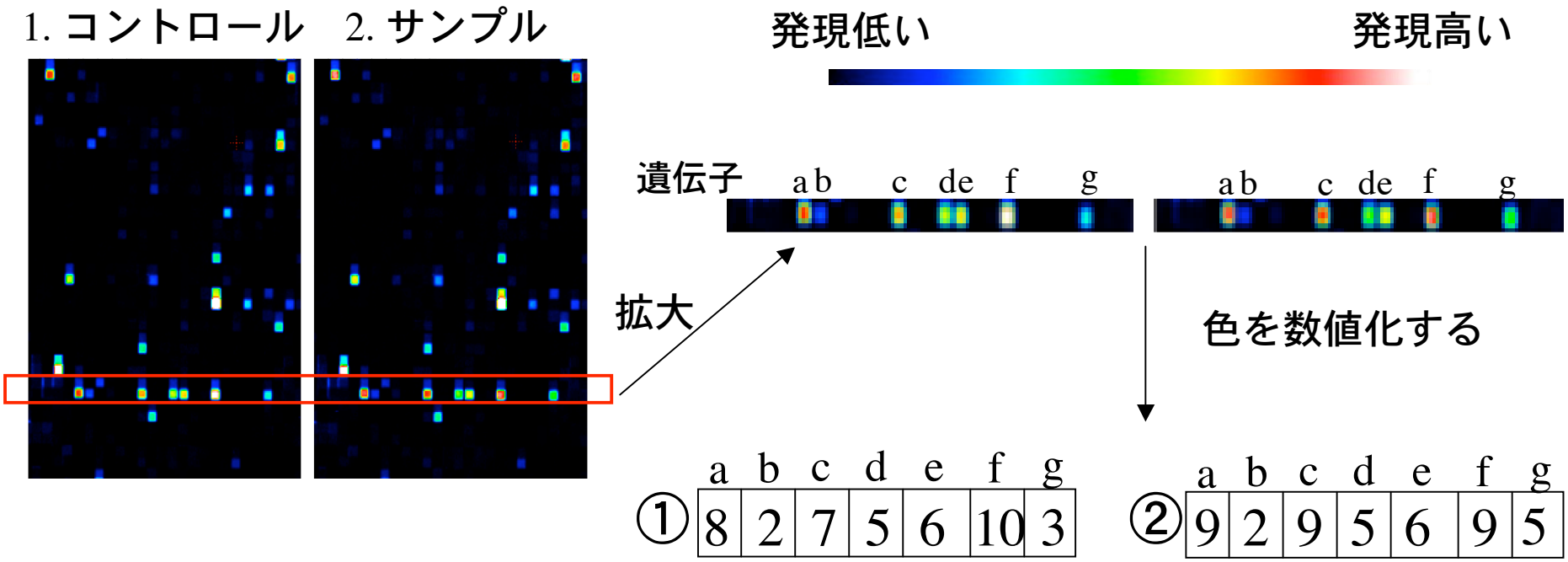
# ハイブリダイゼーション

- 物質がお互いを認識し、結合する現象
- マイクロアレイでは、相補的配列をもつ核酸同士であるプローブと、ターゲットが特異的に結合すること



# マイクロアレイから得られるデータ

## ＜GeneChipのスキャン結果＞

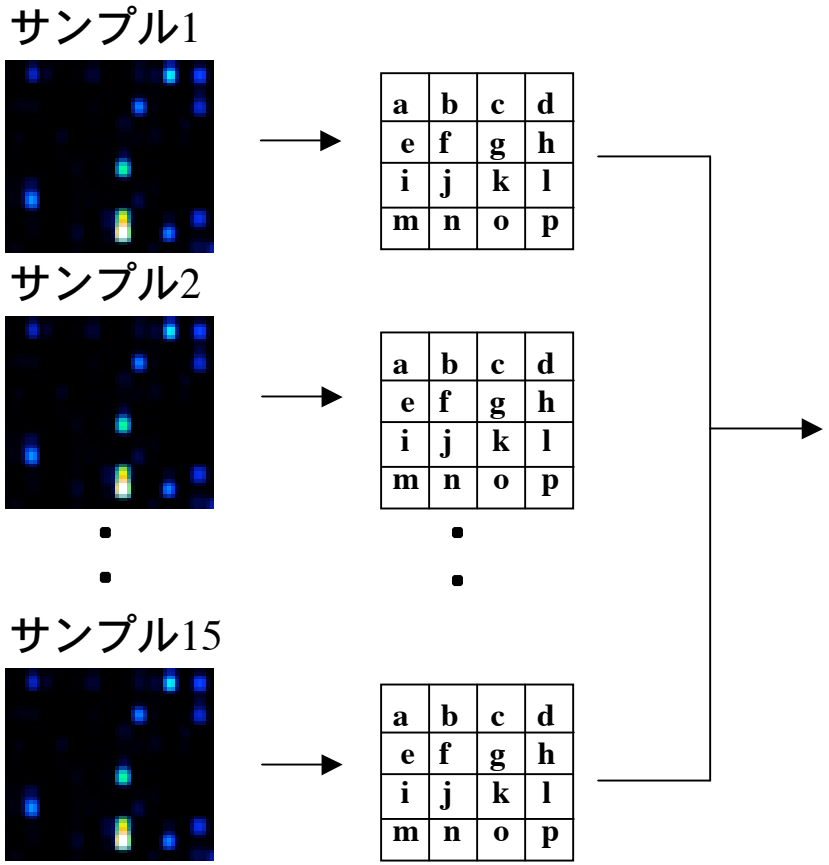


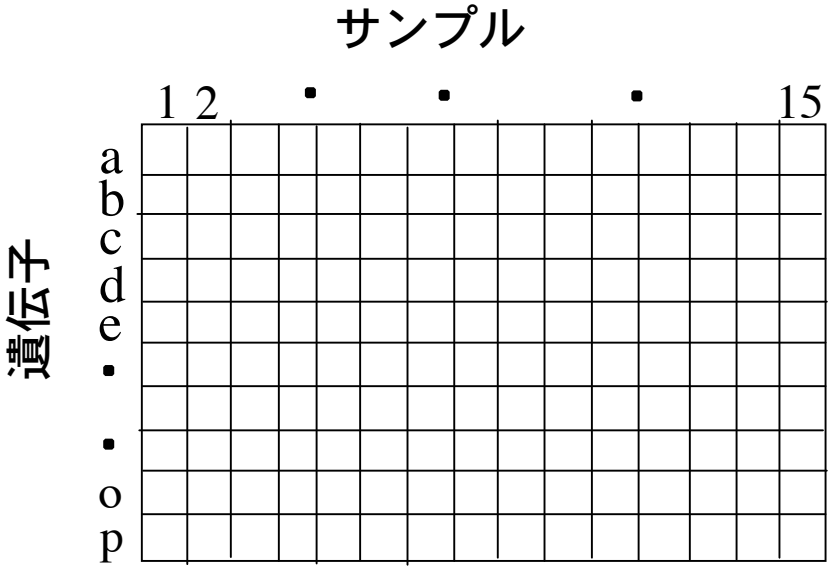
# マイクロアレイ解析（データ行列）

1. Scan data

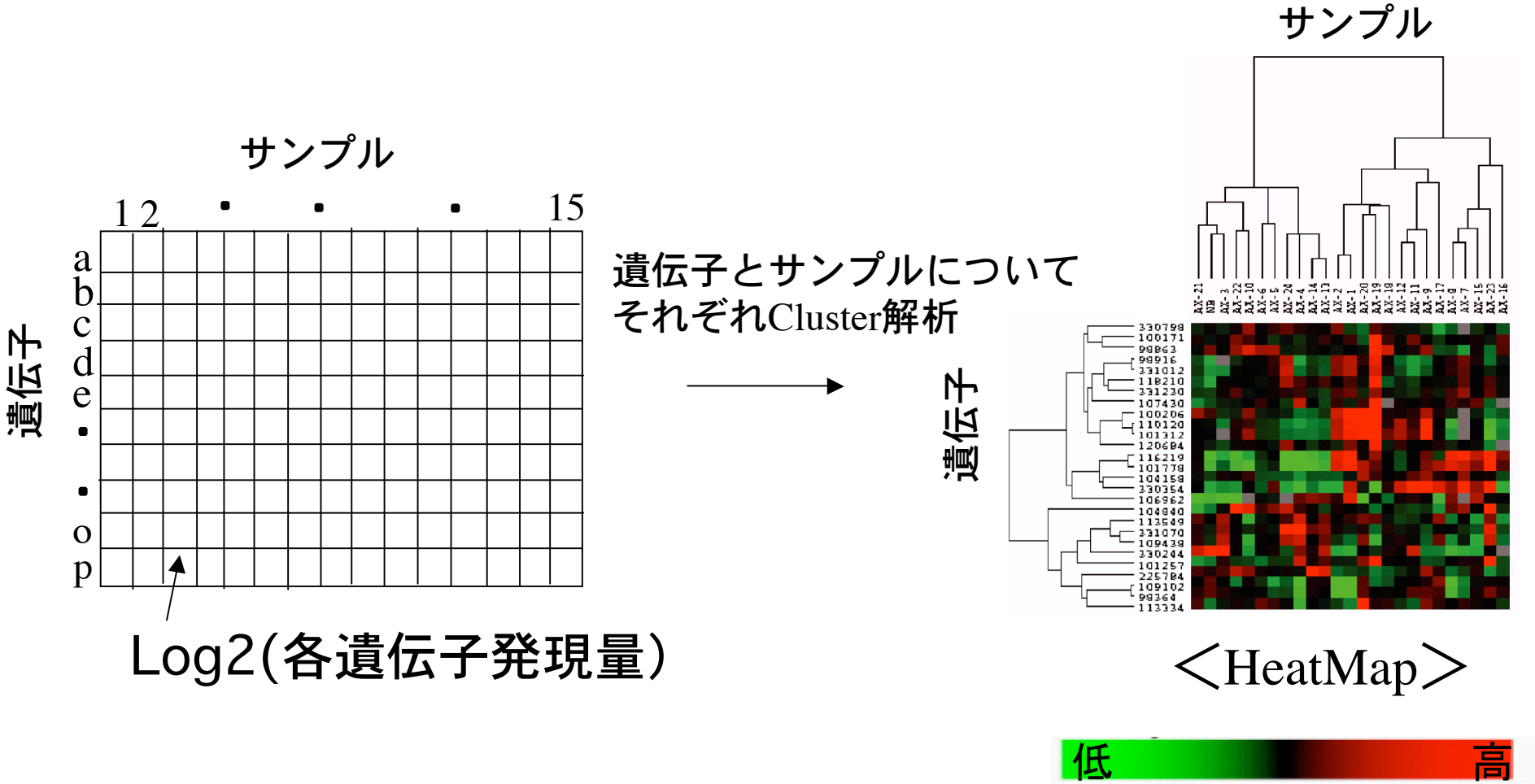
2. 遺伝子の発現量の数値化

3. サンプルの遺伝子の発現量の数値のマトリックス化



# マイクロアレイ解析（クラスター解析）



# クラスター解析復習

# クラスター解析

例えば、5科目のテスト結果から、能力別（理系、文系、優秀など）にクラス分けを行いたい場合、どうすれば良いのか？

| | 国語 | 社会 | 数学 | 理科 | 英語 |
|---|---|---|---|---|---|
| a | 29 | 33 | 55 | 79 | 74 |
| b | 71 | 68 | 72 | 64 | 97 |
| c | 74 | 91 | 79 | 76 | 100 |
| d | 52 | 56 | 58 | 60 | 85 |
| e | 77 | 92 | 96 | 88 | 98 |

人間的に

a～eさんの点数のパターンを眺める ⇒ パターンが似ている者どうしを同じグループにする

数学的に

a～eさんの変数をベクトル表現する ⇒ 似ているか似ていないかを距離という尺度で定義する

## ベクトル表現から類似度定義

a～eさんの変数をベクトル表現する

$\vec{V}a = (29,\ 33,\ 55,\ 79,\ 74)$

$\vec{V}b = (71,\ 68,\ 72,\ 64,\ 97)$

$\vec{V}c = (74,\ 91,\ 79,\ 76,\ 100)$

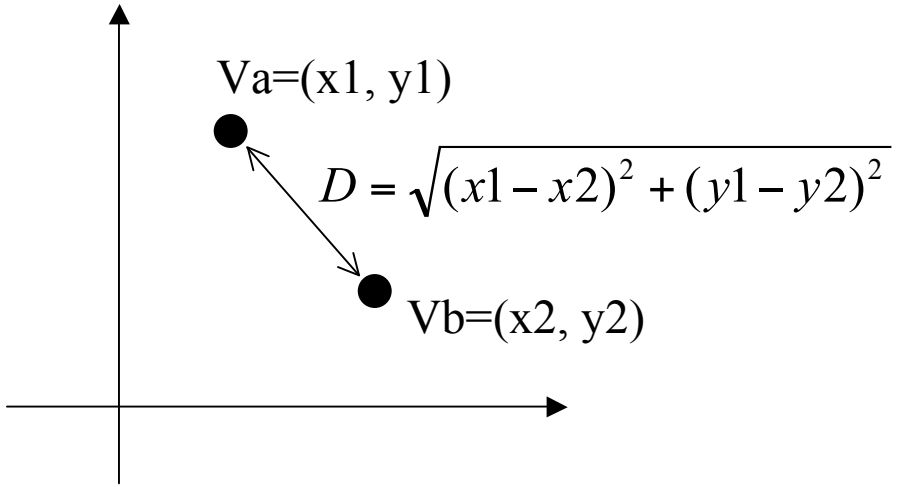
……

似ているか似ていないかを
距離という尺度で定義する

ユークリッド距離で表現する
（似ているものは距離が小さい）

$$D = \sqrt{(\vec{V}a - \vec{V}b)^2}$$

### 簡単のため、2次元の場合




### 今の場合、5次元になる

$$Dab = \sqrt{(29-71)^2 + (33-68)^2 + .... + (74-97)^2}$$

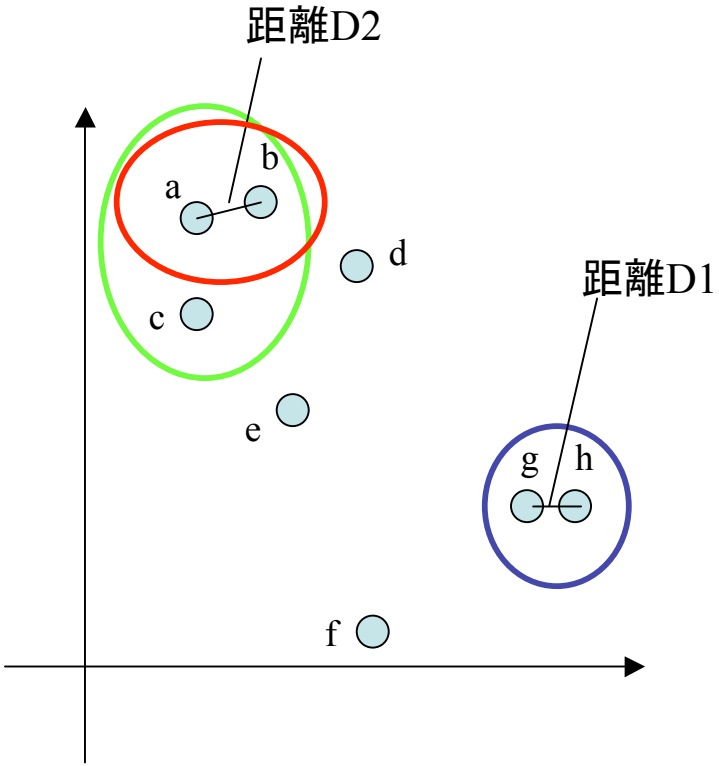
$Dac = \sqrt{\ }.....$

$Dbc = \sqrt{\ }.....$

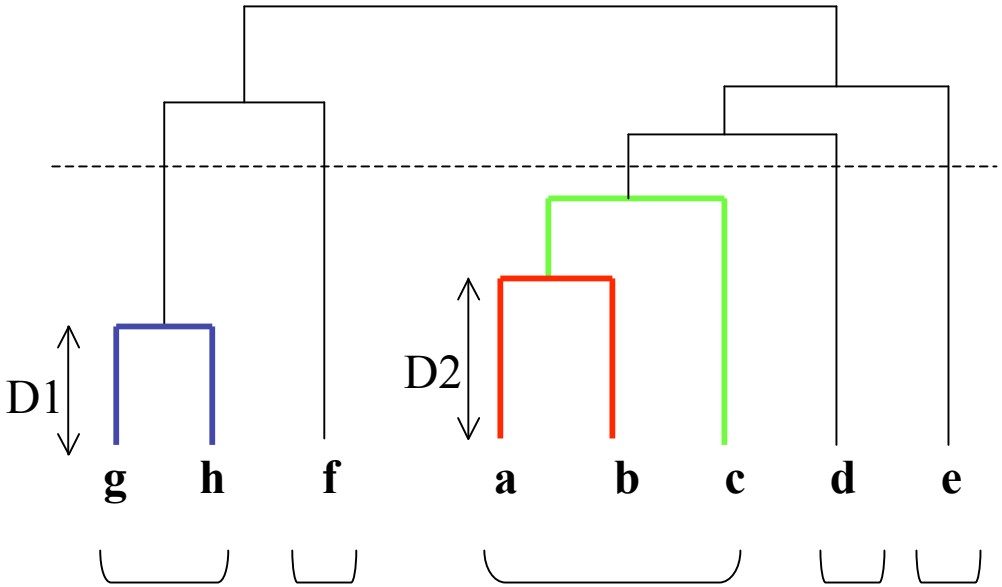
Rで計算してみる

# 階層型クラスタリング

距離の近いものから、グルーピングしていく。




## クラスター表記：デンドログラム（系統樹）




簡単にするため2次元で表現している

# データマイニング

データマイニングとは
蓄積したデータから、意味のある相関関係やパターンなどを発見する技術

対象がゲノム情報の場合、
大量のゲノム情報を有用な発見が潜む「鉱山」とみなし、そこから有用な情報を「採掘（mining)」するという意味

# 何がマイニングできるか？

＜遺伝子側クラスタリング＞

1. データセットの中に何種類の遺伝子発現パターンが含まれているか？
2. 遺伝子Xはどの機能カテゴリーに属するか？
3. 機能未知の遺伝子群の発現パターンの中に、すでによく知られた遺伝子の発現パターンと似たものはあるか？

＜細胞・組織（サンプル）側クラスタリング＞

4. 疾患Xのサブタイプを組織の遺伝子発現パターンで認識、発見することができるか？
5. 対象の組織サンプルはどの組織由来か？

# 何がマイニングできるか？

＜遺伝子相互作用ネットワーク＞

6. 対象の組織サンプルで観測される全ての遺伝子間の相互作用の違いは？
7. 発現パターンが類似した全ての遺伝子ペアを明らかにできるか？

＜遺伝子ハンティング＞

8. 正常と疾患など２群の組織サンプルを最もよく識別できる遺伝子群はどれか？
9. 薬物の影響を受けている遺伝子群は？
10. ある遺伝子Xの発現パターンは他の遺伝子群と比較してどれくらい特異か？
11. 医薬品のターゲットとなる遺伝子は？

# R復習

Rは統計計算とグラフィックスのためのフリーソフト

R参考ページ
http://www.R-project.org/
http://cse.naro.affrc.go.jp/takezawa/r-tips/r.html

＜R基本操作＞

Rの起動と終了
Working Directoryの変更

とにかく計算
```
> 256*14+5
[1] 3589
> 5^3
[1] 125
```

結果を変数（オブジェクト）に代入（変数は好きな名前を付けられる）
```
> anyname <- 5^3
> anyname
[1] 125
> x <- 5^3
> x
[1] 125
> anyname+100
[1] 225
```

関数を使う Function_Name(,,,,,)
```
> sin(2)
[1] 0.9092974
> sin(sin(2))
[1] 0.7890723

> y <- sin(2)
> sin(y)
[1] 0.7890723
```

関数を調べる ？Function_Name

```
> ?hclust
>
```

**Usage**

```
hclust(d, method = "complete", members=NULL)
```

**Arguments**

```
d   a dissimilarity structure as produced by dist.
method  the agglomeration method to …….
```

外部ファイルを読み込む

```
> x <- read.table("sample.txt")
> x
  kokugo syakai sugaku rika eigo
a     29     33     55   79   74
b     71     68     72   64   97
c     74     91     79   76  100
d     52     56     58   60   85
e     77     92     96   88   98
f     60     85     66   66   88
.....
```

行列の取り扱い

```
> x[2,3]
[1] 72
> x[4,]
  kokugo syakai sugaku rika eigo
d     52     56     58   60   85
> x[4,2:4]
  syakai sugaku rika
d     56     58   60
> x[4:6,2:4]
  syakai sugaku rika
d     56     58   60
e     92     96   88
f     85     66   66
> y <- x[2:6, 3:5]
> y
```

計算結果を外部ファイルに出力する

```
> write.table(y, "out.txt")
```

履歴の保存

```
> savehistory("rireki.txt")
```

＊ファイルの確認

＜クラスター解析＞

ファイルの中味を EXCEL で確認してください。

R の起動　（Change Directory の変更を忘れない）

手順　read.table 関数　→　dist 関数　→　hclust 関数　→　plot 関数

ファイルの読み込み

```
dat <- read.table("exp.txt")
```

遺伝子側クラスタリング

```
Dis <- dist(dat)
Cl <- hclust(Dis, “complete”)
```

＊single-linkage をしたいときは、complete→single にする

```
par(ps=8)                グラフの文字サイズ変更
plot(Cl)                 画像コピー、Power point へペーストしてみる
plot(Cl, hang=-1)        クラスタリングの表示を変えてみる
cutree(Cl, k=4)          グループを４つに分ける
```

細胞・組織サンプル側（変数側、属性）のクラスタリング

```
Tdat <- t(dat)           転置行列を作成
Tdis <- dist(Tdat)
Tcl <- hclust(Tdis, “complete”)
plot(Tcl, hang=-1)
```

Heatmap を作る

```
Dcl <- as.dendrogram(Cl)
DTcl <- as.dendrogram(Tcl)
Mdat <- as.matrix(dat)
heatmap(Mdat, Rowv=Dcl, Colv=DTcl, col=cm.colors(256))
```

色を変える

```
heatmap(Mdat, Rowv=Dcl, Colv=DTcl, col=heat.colors(256)[256:1])
```