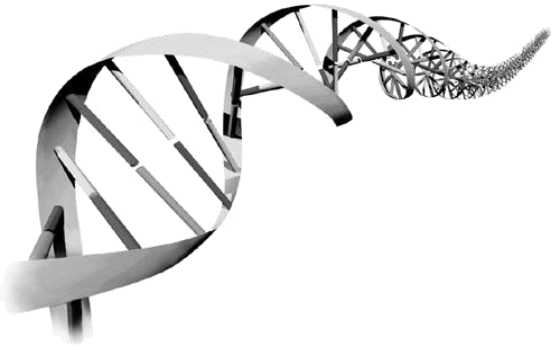


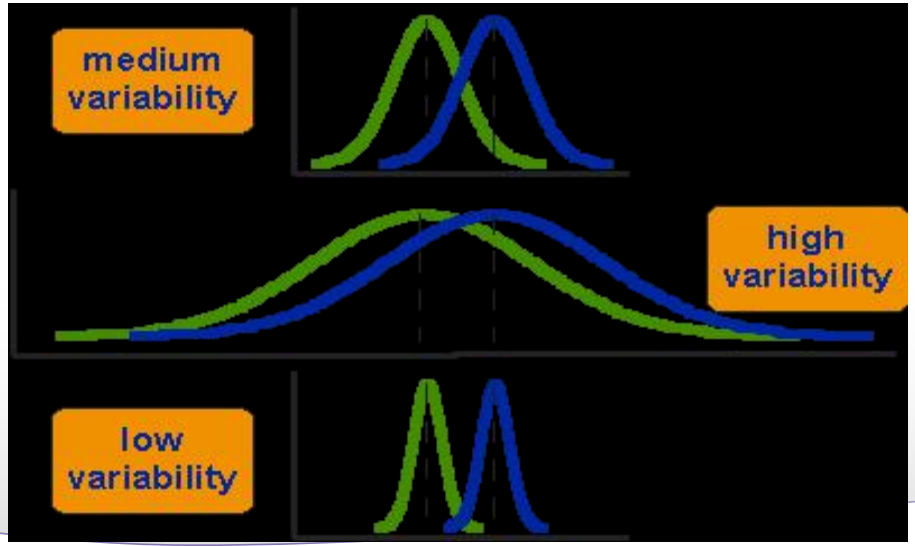
转录数据分析

3. 转录数据分析

- (1) 差异表达基因的分析
- (2) 基因共表达分析
- (3) 基因表达数据的聚类
- (4) 基因表达数据的分类
- (5) 与GO数据库关联分析
- (6) 基因调控网络

(1) 差异表达基因的分析

- 1.差异表达基因的分析: 寻找处理前后表达上调或者下调的基因
- 2. 使用标准的统计学方法检验 (t-test)，发现统计显著性差异表达的基因，
- 3. 如果处理本身并不显著，则结果无意义

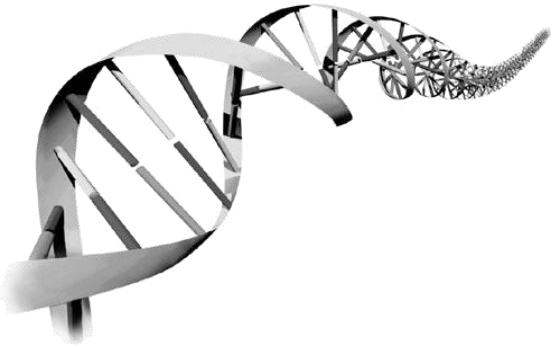


统计学分析

- 1. Fold change, 一般2-fold increase or decrease (平行实验的样本较少)
- 2. *P*-value (平行实验的样本较多)

Gene Name	T1	T2	T3	T4	T5	T6
Gene 1#	**20**	**22**	**15**	**19**	**22**	**12**
Gene 2#	**10**	**11**	**14**	**11**	**10**	**16**
Fold:	**2.0**	**2.0**	**1.1**	**1.7**	**2.2**	**0.8**

Paired sample t test: ***P*-value=0.03**



多重假设检验

How many differential genes to report?

多重假设检验

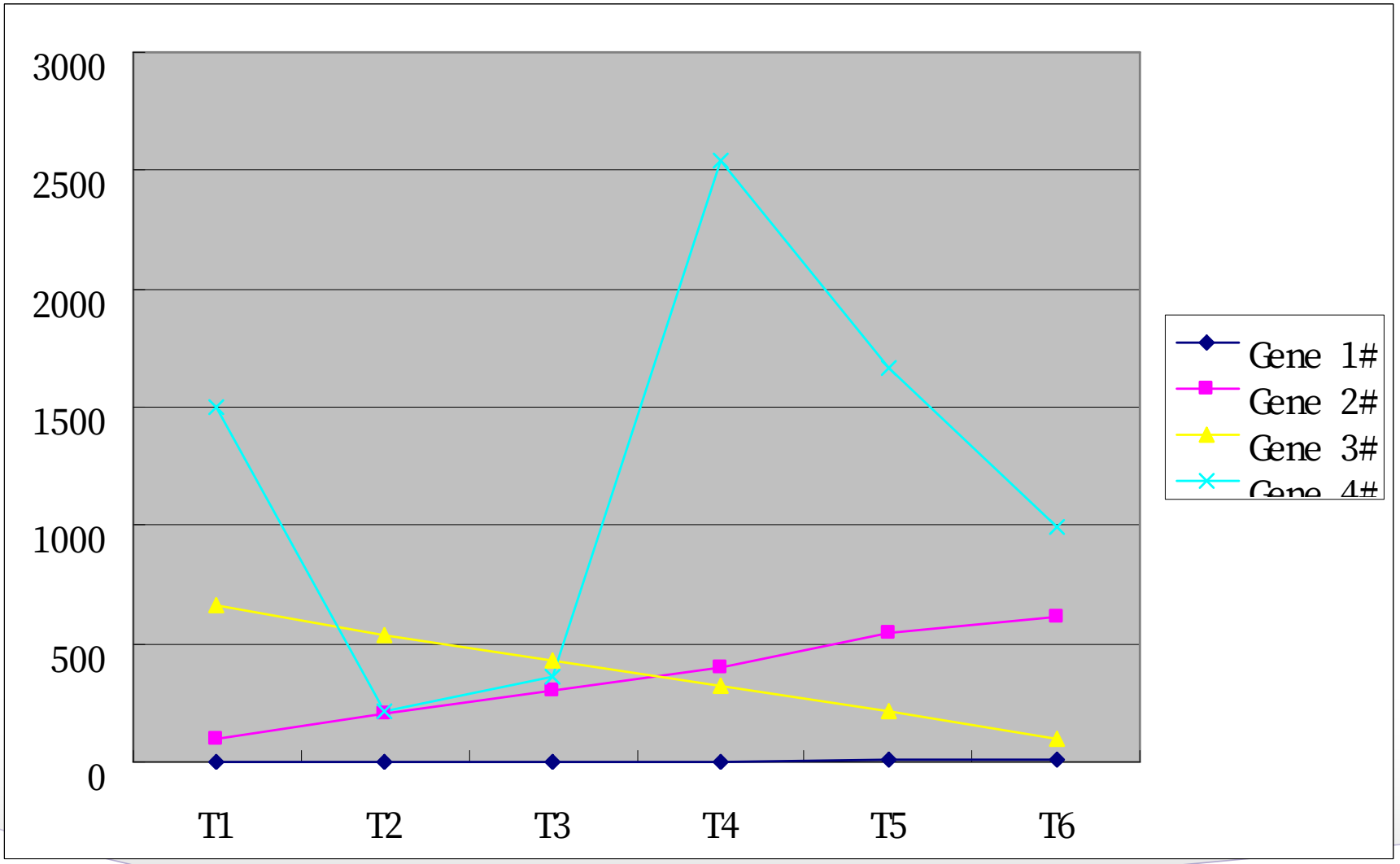
- 每个基因的差异表达 p-value假设为 0.01
- 假如两万个差异基因，将有0.01 x 20K = 200 基因的判断是错的
- Bonferroni 校正，p-value/N ；
- FDR（FalseDiscovery Rate）校正。

(2) 基因共表达分析

- 1. 在N个不同的条件下 (时间序列的芯片数据)，考察基因X和Y的表达是否相似
- 2. Gene 1#是否与Gene 2#、Gene 3#和Gene 4#共表达？
- 3. 共表达：
  - 正相关：相似的表达谱，可能存在正关联
  - 负相关：相反的表达谱，可能存在负调控

Gene Name	T1	T2	T3	T4	T5	T6
Gene 1#	**1**	**2**	**3**	**4**	**5**	**6**
Gene 2#	100	200	300	400	550	610
Gene 3#	660	540	430	320	210	101
Gene 4#	1504	215	357	2545	1670	998

没有相关性？


3000
2500
2000
1500
1000
500
0
T1
T2
T3
T4
T5
T6
Gene 1#
Gene 2#
Gene 3#
Gene 4#

基因相关性分析

- 1. Spearman 顺序相关系数
- 2. Pearson 积差相关系数
  - 衡量两个数据集合是否在一条线上面

$$\rho(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i$$

$$\bar{y} = \frac{1}{n}\sum_{i}^{n} y_i$$

Eisen MB, *et al.*, (1998) PNAS 95:14863-14868

Pearson相关系数

- **1. r ~ [-1, 1]**
  - **r~ 1，正相关**
  - **r~ -1，负相关**

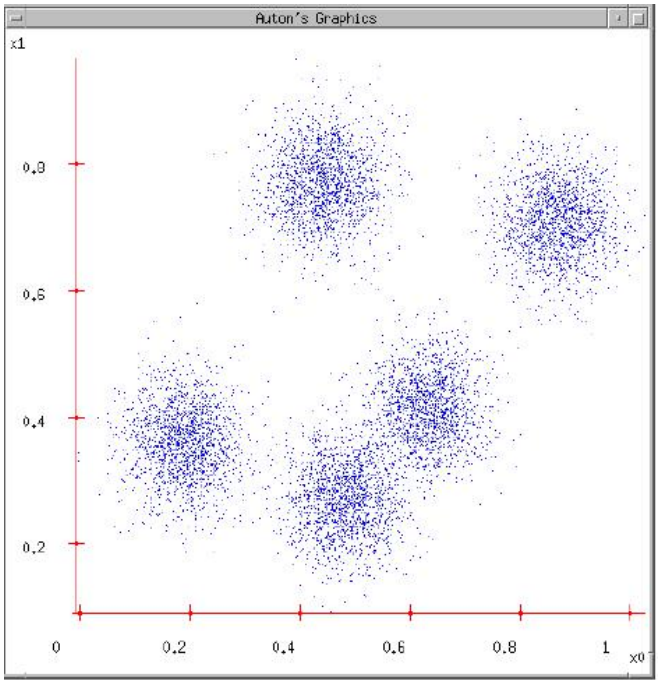
它是一种度量两个变量间相关程度的方法。
它是一个介于 1 和 -1 之间的值，其中，1 表示变量完全正相关， 0 表示无关，-1 表示完全负相关。

	Gene 1#	Gene 2#	Gene 3#
Gene 1#			
Gene 2#	0.996368		
Gene 3#	-0.99988	-0.99611	
Gene 4#	0.245292	0.254855	-0.2395

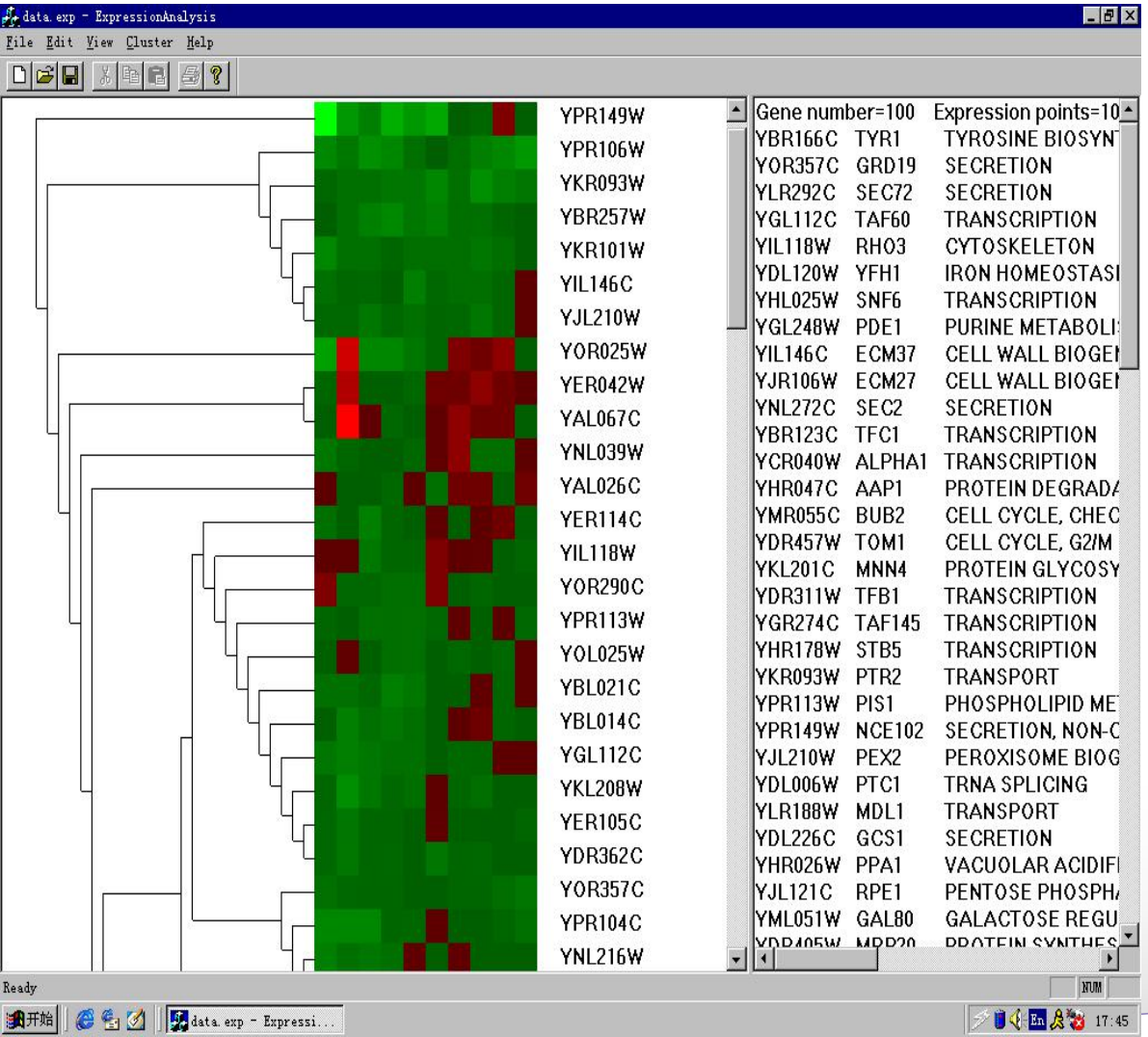
- 结论：**Gene 1#**与Gene 2#表达正相关，与Gene 3#表达负相关，与Gene 4#无关联

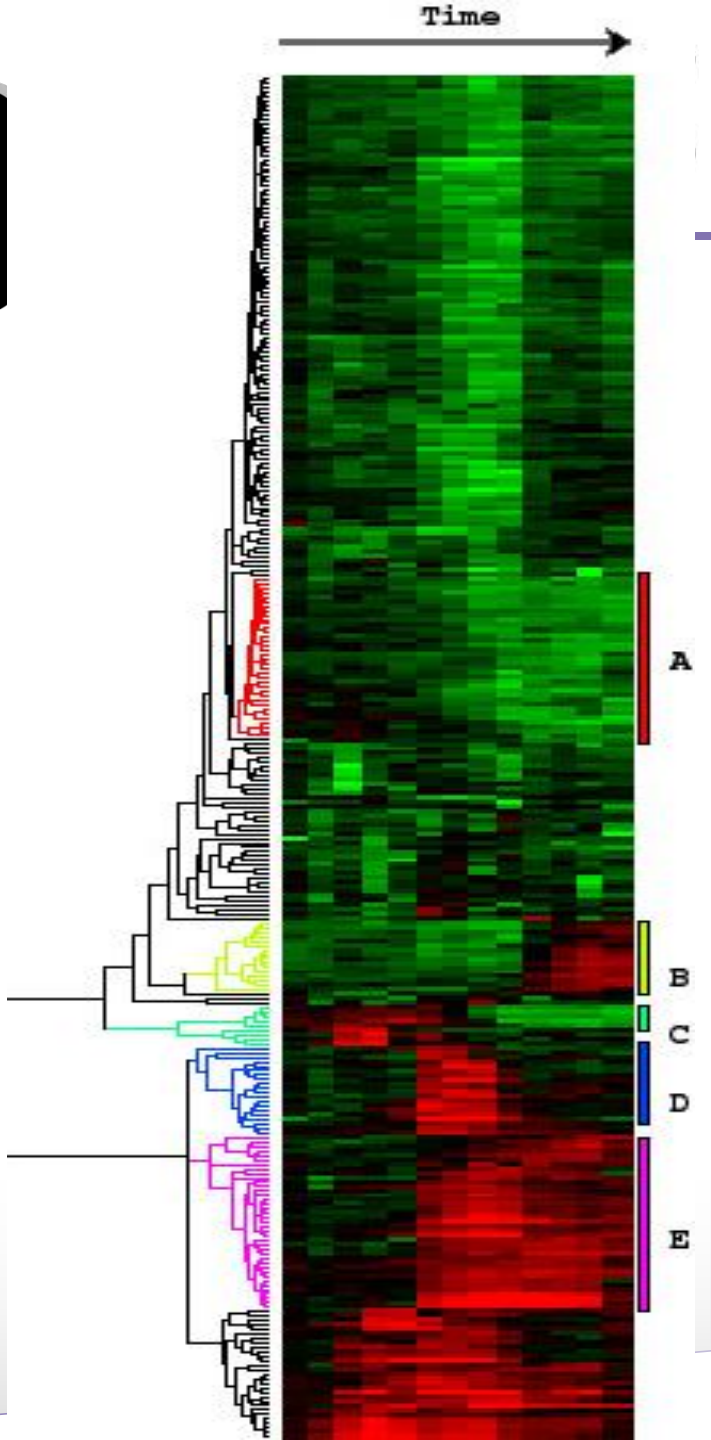
(3) 基因表达数据的聚类

- 将表达谱相似的基因聚类在一起
- 聚类方法：
  - A. *K*-means clustering
  - B. Hierarchical clustering
- 聚类的目的
  - 可诱导基因是共表达的
  - 可以揭示细胞的生理状态
  - 可以帮助研究未知基因的功能

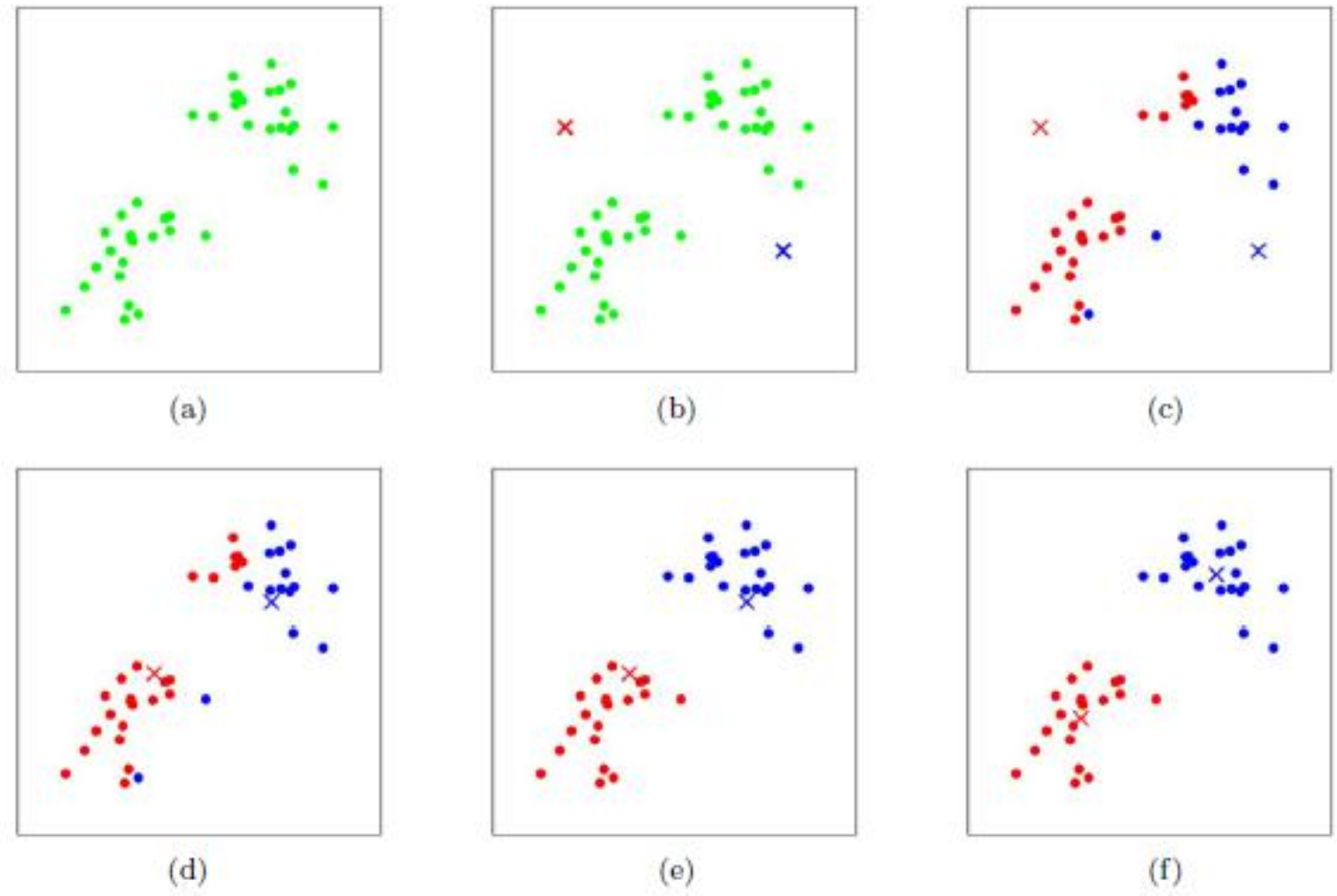


聚类结果显示:
Cluster, TreeView





K-means clustering



k-means 算法

k-means 算法基本步骤

- （1） 从 n个数据对象任意选择 k 个对象作为初始聚类中心；
- （2） 根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离；并根据最小距离重新对相应对象进行划分；
- （3） 重新计算每个（有变化）聚类的均值（中心对象）；
- （4） 计算标准测度函数，当满足一定条件，如函数收敛时，则算法终止；如果条件不满足则回到步骤（2）。

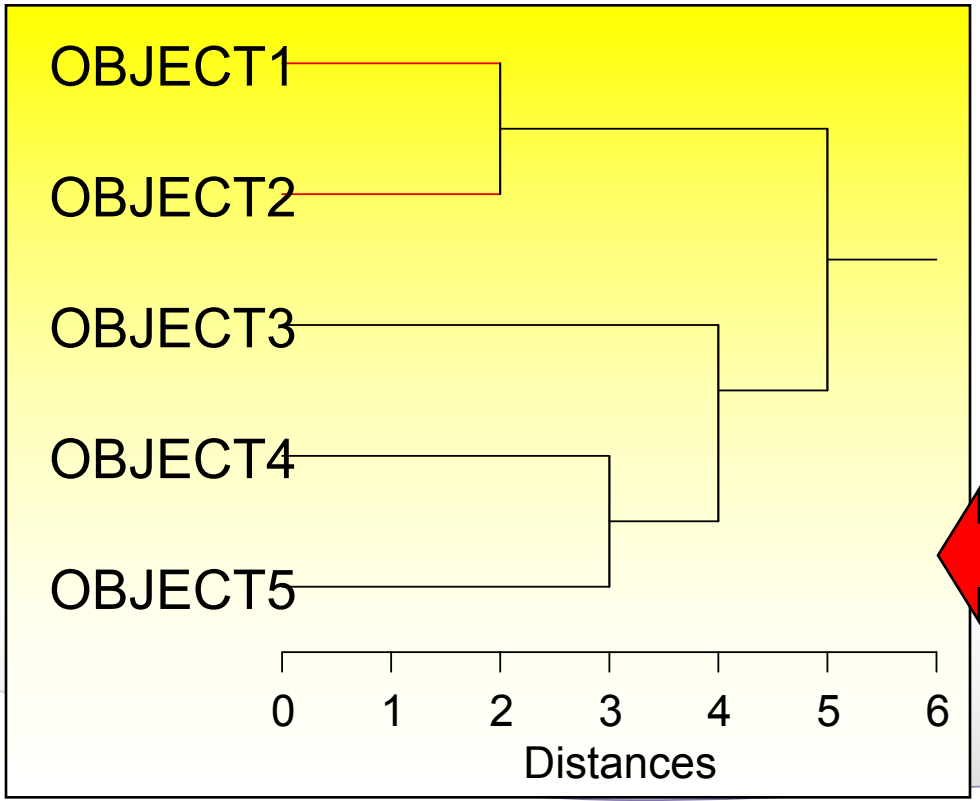
Hierarchical clustering

- 1. 用树状结构来表征基因表达之间的相似性/相关性
- 2. 优点：不需要指定结果有多少类

Object	1	2	3	4	5
1					
2	2				
3	6	5			
4	10	9	4		
5	9	8	5	3	

Distance matrix

Distance	Cluster
0	1,2,3,4,5
2	(1, 2), 3, 4, 5
3	(1, 2), 3, (4, 5)
4	(1, 2), (3, 4, 5)
5	(1, 2, 3, 4, 5)

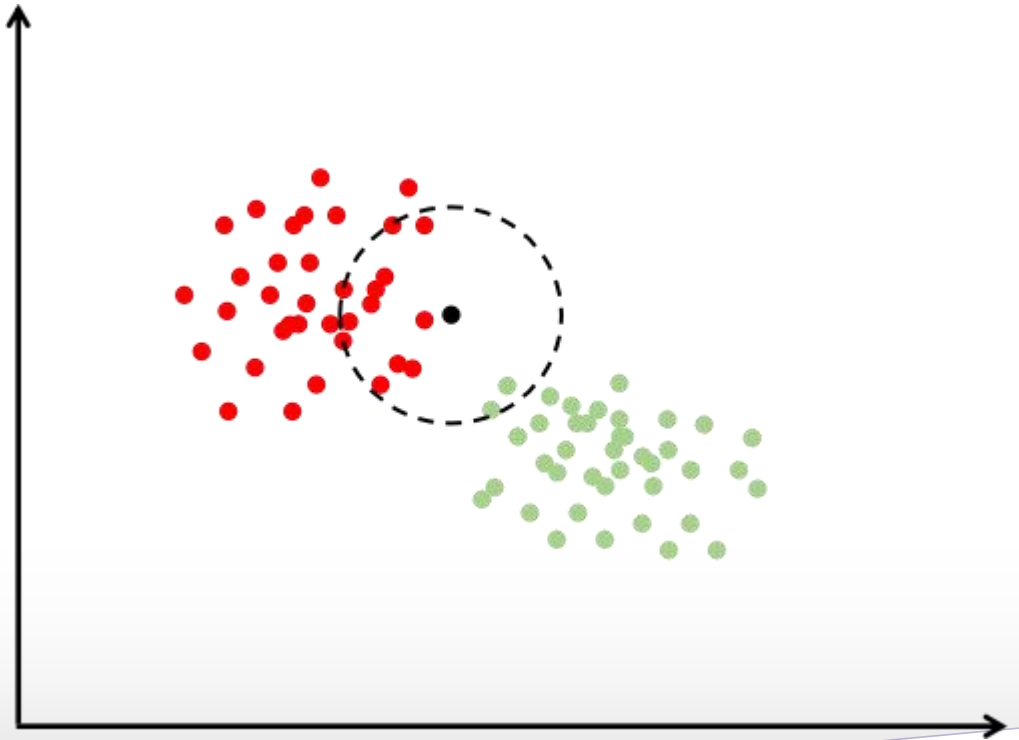


(4) 基因表达数据的分类

- 1. 根据基因表达的数据将样本分成两类或多类；
- 2. 根据发现的pattern进行预测
- 3. 应用：
  - 癌症 vs. 正常组织
  - 癌症的亚型、不同阶段 (良性的 vs. 恶性的)
  - 对药物的敏感性 (tamoxifen for breast cancer)

分类算法

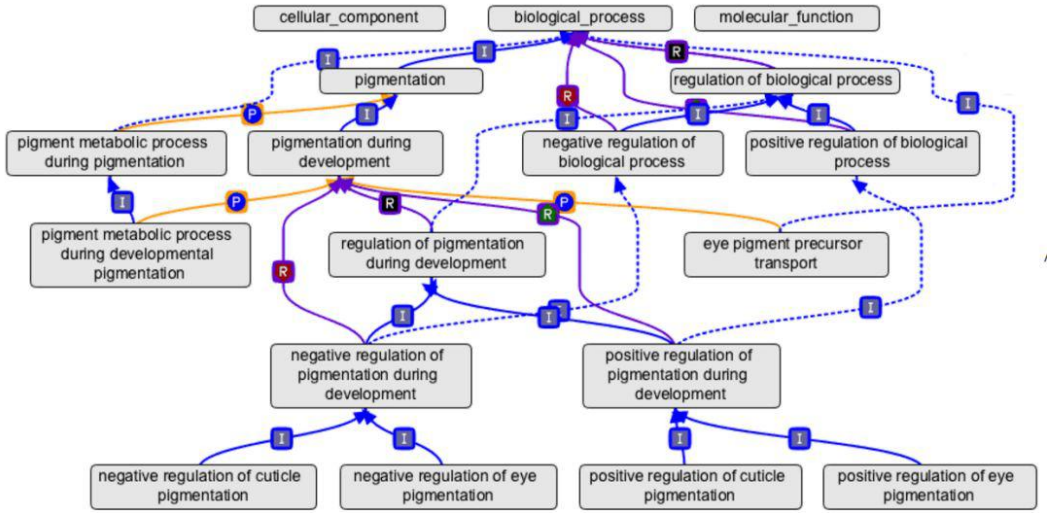
- 决策树
- 支持向量机SVM
- 逻辑回归
- 神经网络NN
- 随机森林
- K近邻

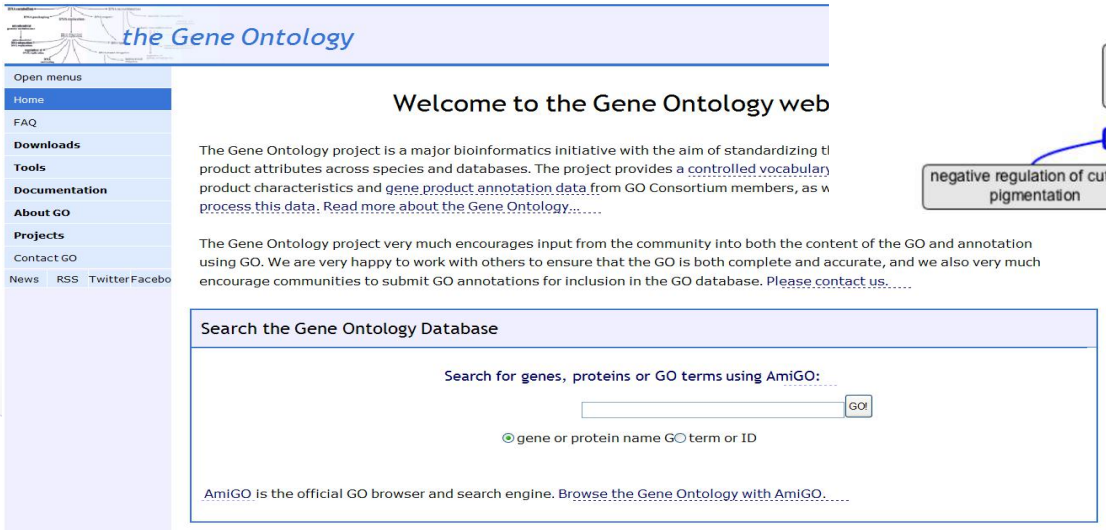


(5) 与GO数据库关联分析

GO:一个结构化的标准生物学模型，旨在建立基因及其产物知识的标准词汇体系，涵盖了基因的细胞组分、分子功能、生物学过程。。

- 1. 通过基因芯片，找到了一批“interesting” 的基因
- 2. 生物学功能上是否存在关联？
  - 某种功能是否显著？

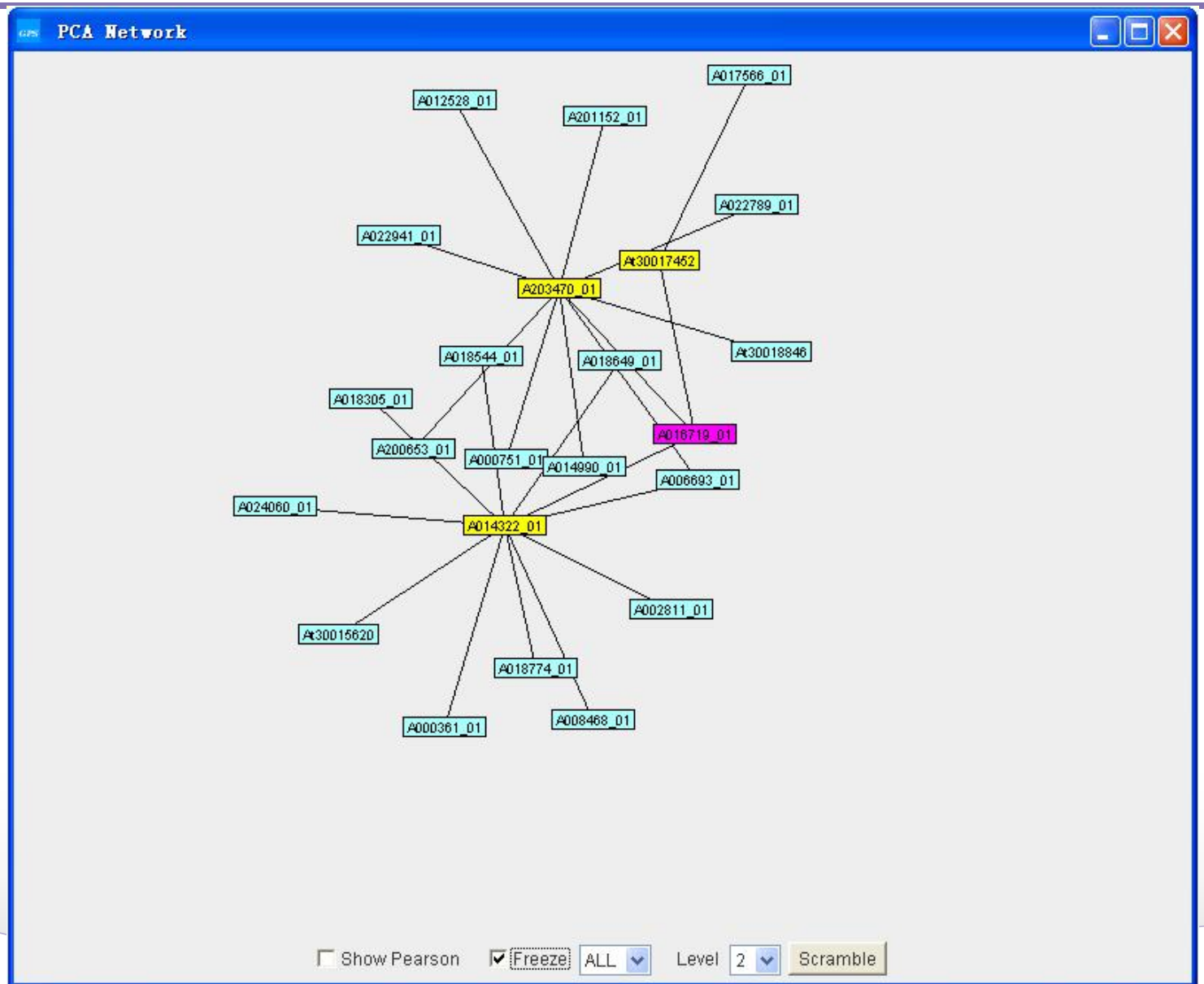




(6) 基因调控网络

- 1. 早期观点：表达谱相似的基因可能存在功能上的关联，可能相互作用 (直接作用)
- 2. 当前的观点：表达谱相似的基因可能具有共同的调控元件 (基因UTR区域存在共同的Promotor), 能够被同一个上游因子所调控

相关系数：基因共表达网络



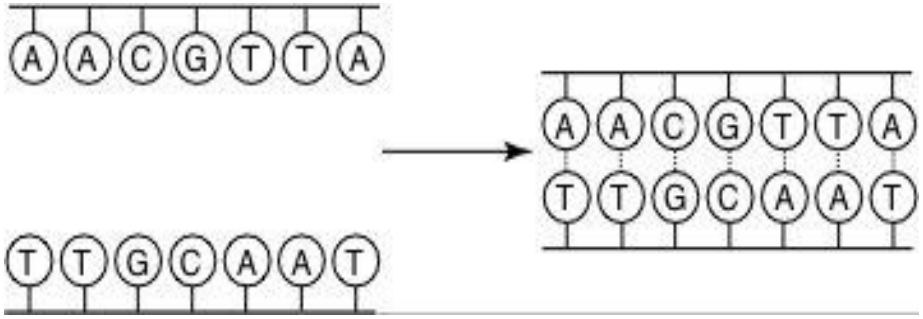
获取转录数据的方法

- **转录组测序**
- **生物芯片：**
  - 借助微加工和微电子技术，将大量已知序列的核酸或蛋白质片段有序地组合在一个微小基片表面,通过与标记的核酸或蛋白质分子进行反应，分析待检标本的相应成分。

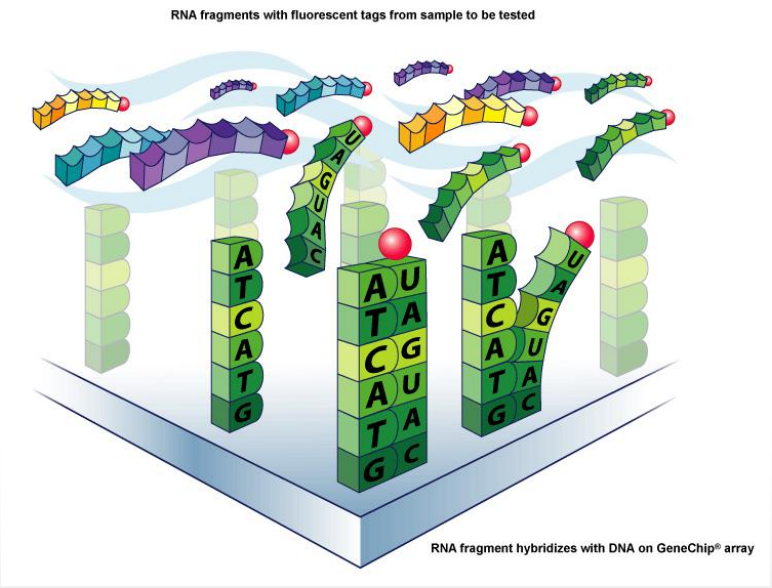
生物芯片的分类

- 基因芯片
- 蛋白芯片
- 组织芯片
- 细胞芯片

基因芯片的原理示意图

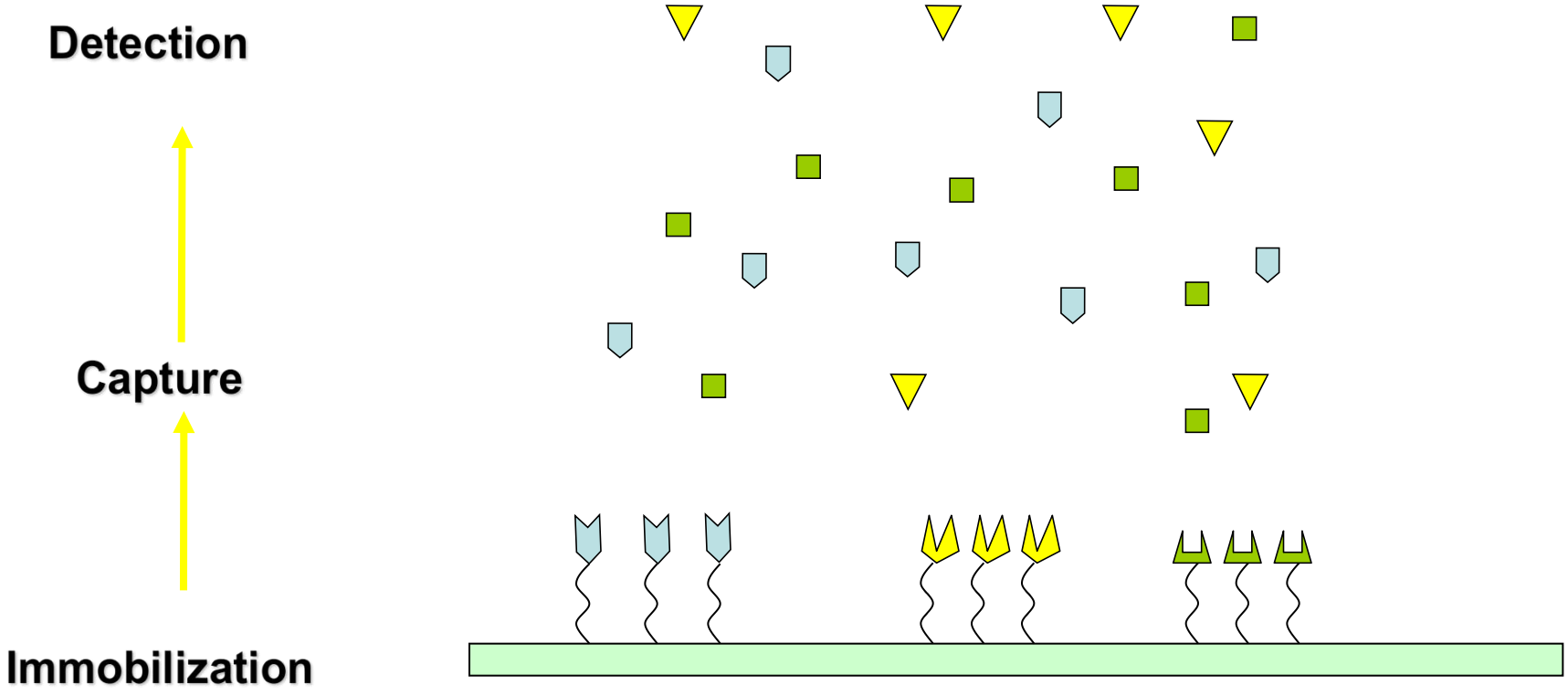


碱基互补



将样品中的DNA/RNA表上荧光标记，则可以定量检验基因的表达水平

蛋白质芯片原理示意图



基因芯片分析的优点

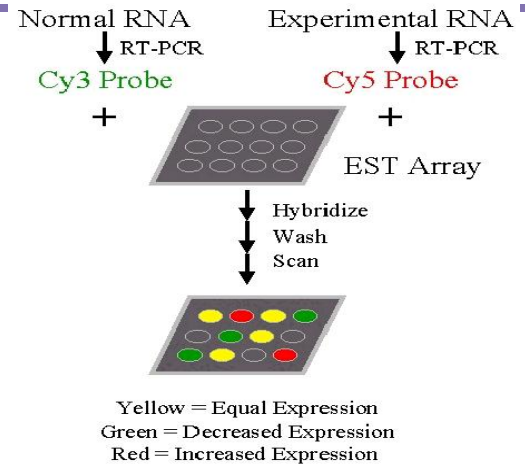
- 快速
- 高通量(10^4 - 10^6)
- 自动化
- 使用的试剂少
- 低成本

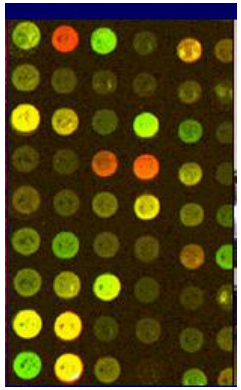
基因芯片的分类

- 芯片设计分类
  - 寡核苷酸芯片（20～80碱基，特异性好）
  - cDNA芯片（300～1500碱基，成本低）

- 芯片荧光物质分类
  - 双通道（两个样本用不同荧光标记后一起杂交到同一张芯片上，Cy5（红）和Cy3（绿）两种荧光物质，可靠性高）
  - 单通道（单一荧光标记，数据变异大，需要重复）

基因芯片的实验流程

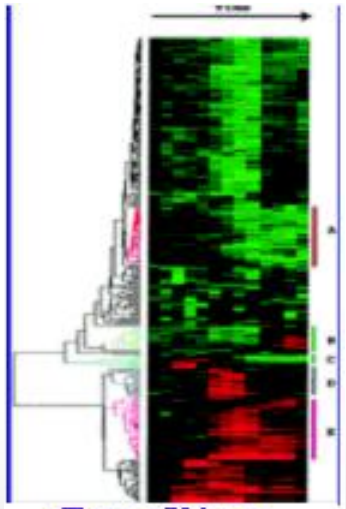





Microarray Scanner

	A	B	C	D	E	F	G
1	YORF	NAME	GWEIGHT	spo0	spo30	spo2	spo5
2	EWEIGHT			1	1	1	1
3	YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56
4	YAL004W	YAL004W	1	0.41	-0.38	-0.89	-1.06
5	YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29
6	YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25


Cluster


TreeView

探针设计

- 一致性:所有探针的杂交温度、杂交时间、杂交液成分、洗脱温度等尽量相同。
  - 控制探针的长度、Tm 值、GC 含量等。
- 灵敏性：与样本中靶标序列相互结合的能力
  - 探针不会形成稳定的二级结构和同源二聚体
- 特异性：不与样本中的非靶标序列结合。
  - 探针和靶标序列的连续匹配长度不能过大或过长，低复杂度的序列要尽可能的小。

基因表达的数据

mRNA samples

Genes	sample1	sample2	sample3	sample4	sample5	...
1	0.46	0.30	0.80	1.51	0.90	...
2	-0.10	0.49	0.24	0.06	0.46	...
3	0.15	0.74	0.04	0.10	0.20	...
4	-0.45	-1.03	-0.79	-0.56	-0.32	...
5	-0.06	1.06	1.35	1.09	-1.09	...

Gene expression level of gene *i* in mRNA sample *j*

GEO https://www.ncbi.nlm.nih.gov/geo