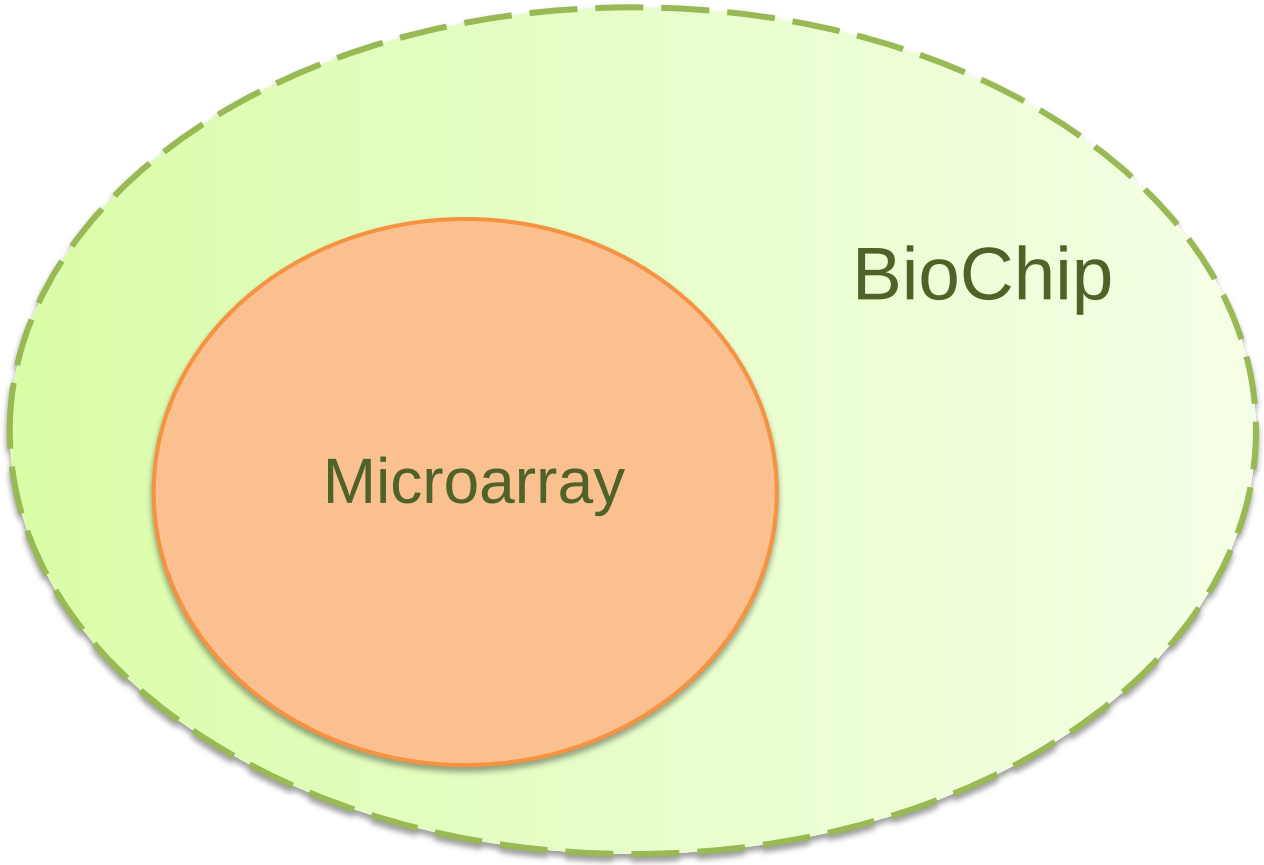
Microarray資料分析

黃德安　　張桂銘　　林哲宇

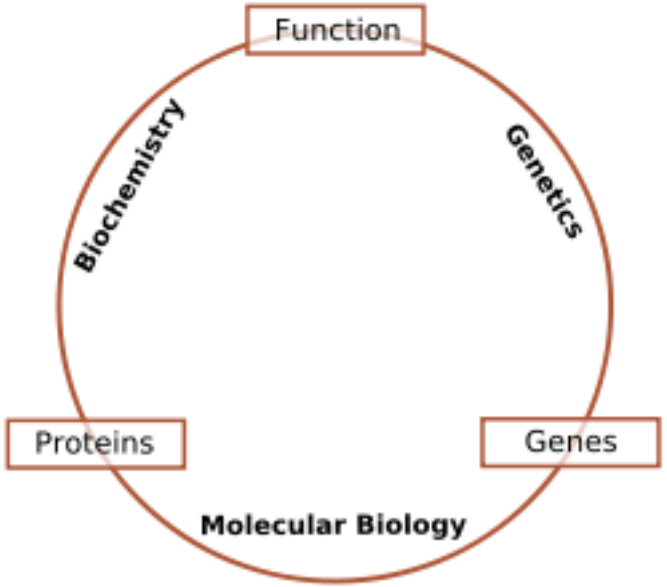
微陣列(Microarray)技術

微小化的生物技術平台

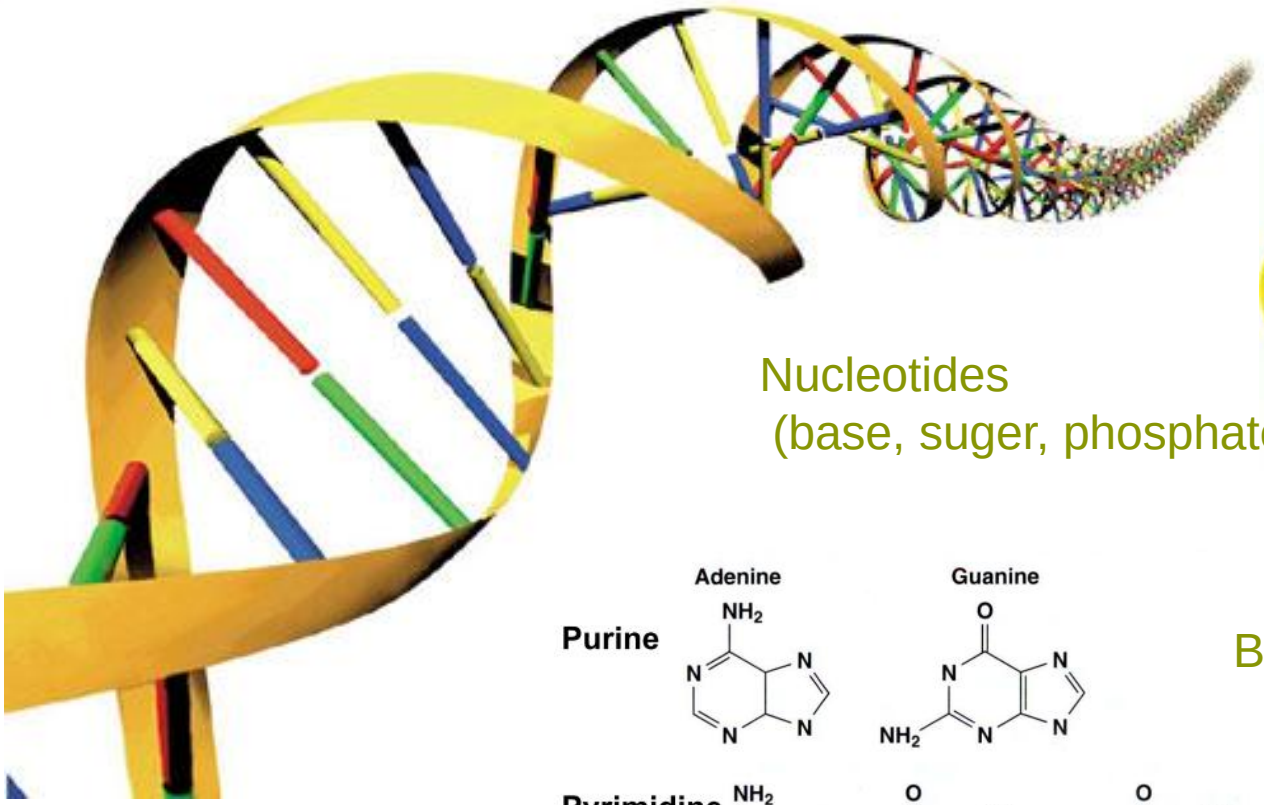


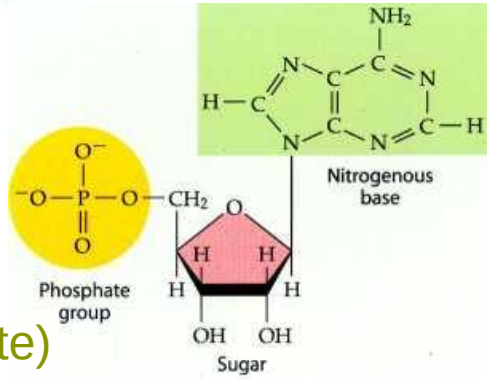
Molecular biology

- Concern with the interactions between the various systems of a cell.
- DNA, RNA and protein biosynthesis.



DNA & RNA

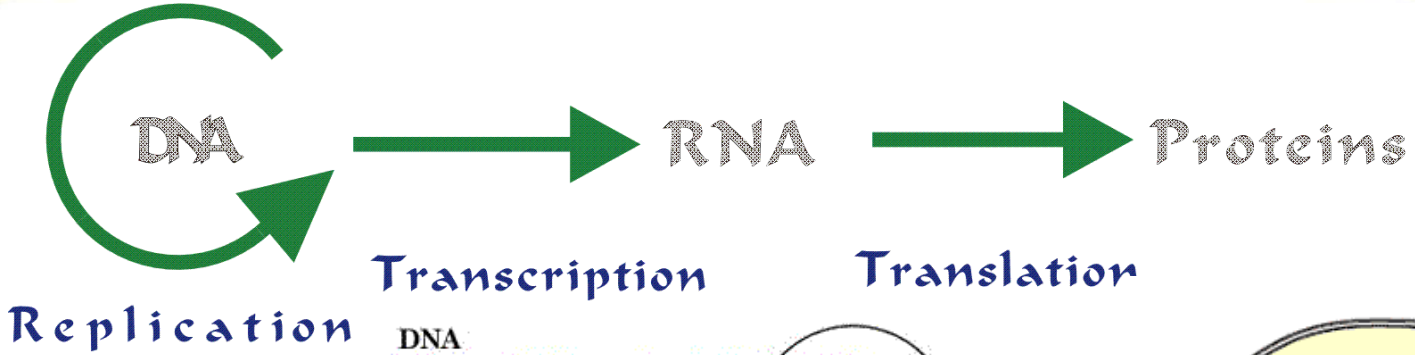


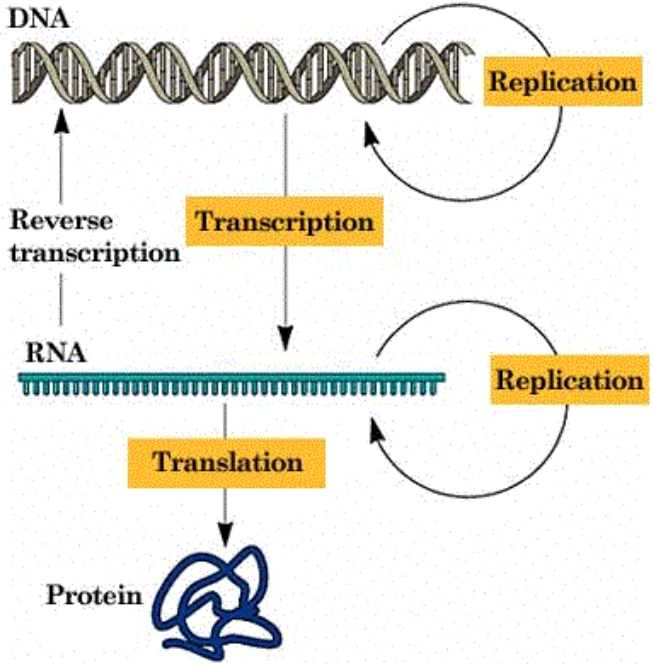


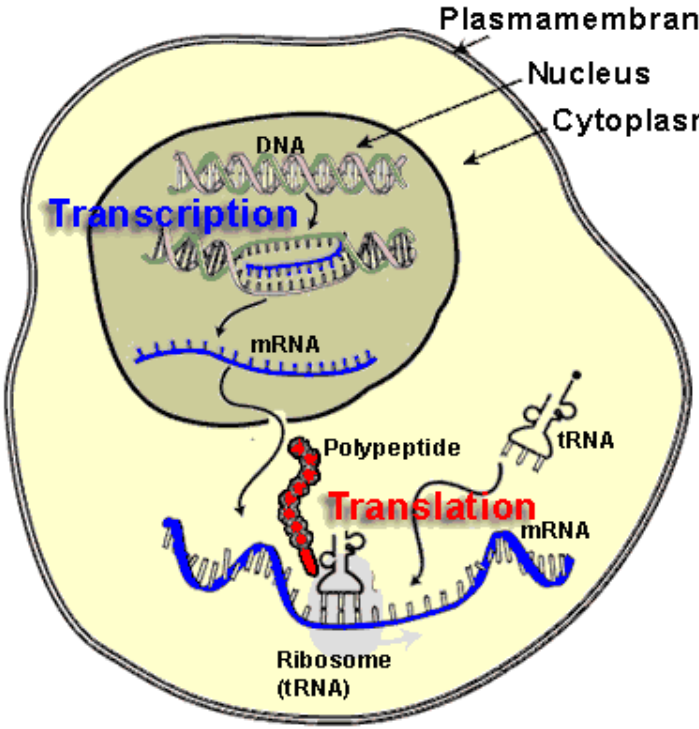
Nucleotides
(base, suger, phosphate)

Base – DNA: A-T, C-G
RNA: A-U, C-G

Central dogma







Protein synthesis

- Codon (genetic code)

  Three-nucleotide sequences used in mRNA to specify one of the 20 amino acid used for protein synthesis

TABLE 2-3 The Genetic Code

first position	second position: U	second position: C	second position: A	second position: G	third position
U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys	U
	UUC Phe	UCC Ser	UAC Tyr	UGC Cys	C
	UUA Leu	UCA Ser	UAA stop	UGA stop	A
	UUG Leu	UCG Ser	UAG stop	UGG Trp	G
C	CUU Leu	CCU Pro	CAU His	CGU Arg	U
	CUC Leu	CCC Pro	CAC His	CGC Arg	C
	CUA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CUG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	AUU Ile	ACU Thr	AAU Asn	AGU Ser	U
	AUC Ile	ACC Thr	AAC Asn	AGC Ser	C
	AUA Ile	ACA Thr	AAA Lys	AGA Arg	A
	AUG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GUU Val	GCU Ala	GAU Asp	GGU Gly	U
	GUC Val	GCC Ala	GAC Asp	GGC Gly	C
	GUA Val	GCA Ala	GAA Glu	GAA Gly	A
	GUG Val	GCG Ala	GAG Glu	GGG Gly	G

基因定序

- 人類基因體研究計劃
  (Human Genome Project，HGP)
  - 1990年開始，為期15年
  - 榮陽團隊
  - 第四號染色體(2000年解碼完畢)
    - 起因：國人常見的肝癌
    - 癌症是細胞基因體發生突變產生的結果
- 黑猩猩、靈芝、水稻的基因解碼

Biochip

- 廣義上，指將與生物有關的分子利用微面積、高密度的技術，精確地點製在玻璃、矽片、塑膠等材質上。

- 依植入晶片物質的不同：
  - 基因晶片
  - 蛋白質晶片

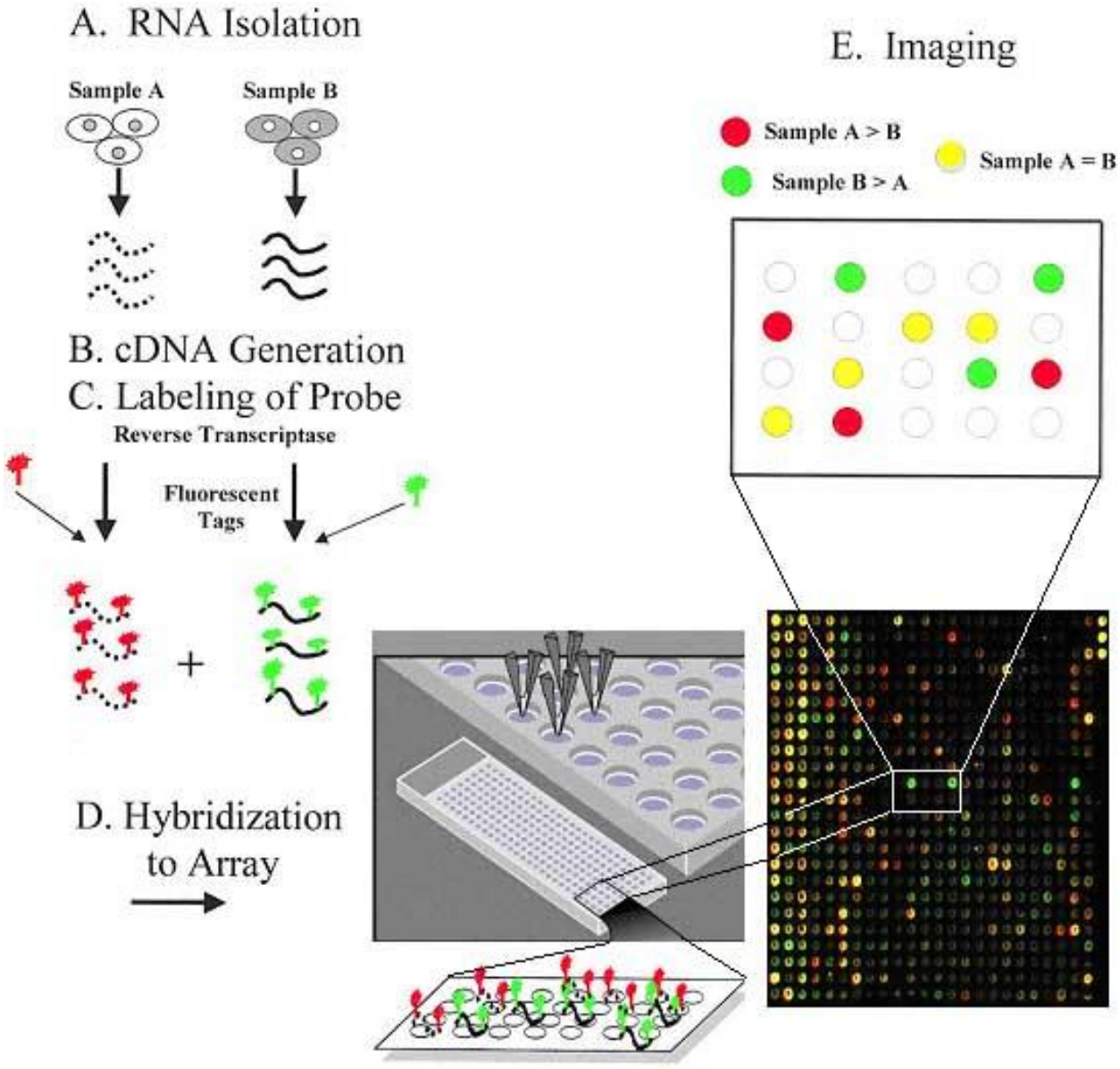
基因晶片

- Gene chip
- 帶有DNA微陣列的特殊玻璃片或矽晶元片

DNA Microarray

- 基因組學和遺傳學研究的工具
- 單股DNA的核酸探針與相匹配的標的 進行雜交反應(Hybridization)
- 藉由螢光、化學發光等方式標記
- 在**數平方公分**面積上佈放**成千上萬**個探針組成

多個願望一次滿足！


A. RNA Isolation
Sample A
Sample B
B. cDNA Generation
C. Labeling of Probe
Reverse Transcriptase
Fluorescent Tags
D. Hybridization to Array
E. Imaging
Sample A > B
Sample B > A
Sample A = B

蛋白質晶片

- 以蛋白質、抗原或是抗體當作生物探針，固定在晶片上。
- 藉著抗原、抗體間的專一性質，來檢測特定蛋白質。

功能

基因表現

癌症分類

新藥開發

疾病檢驗

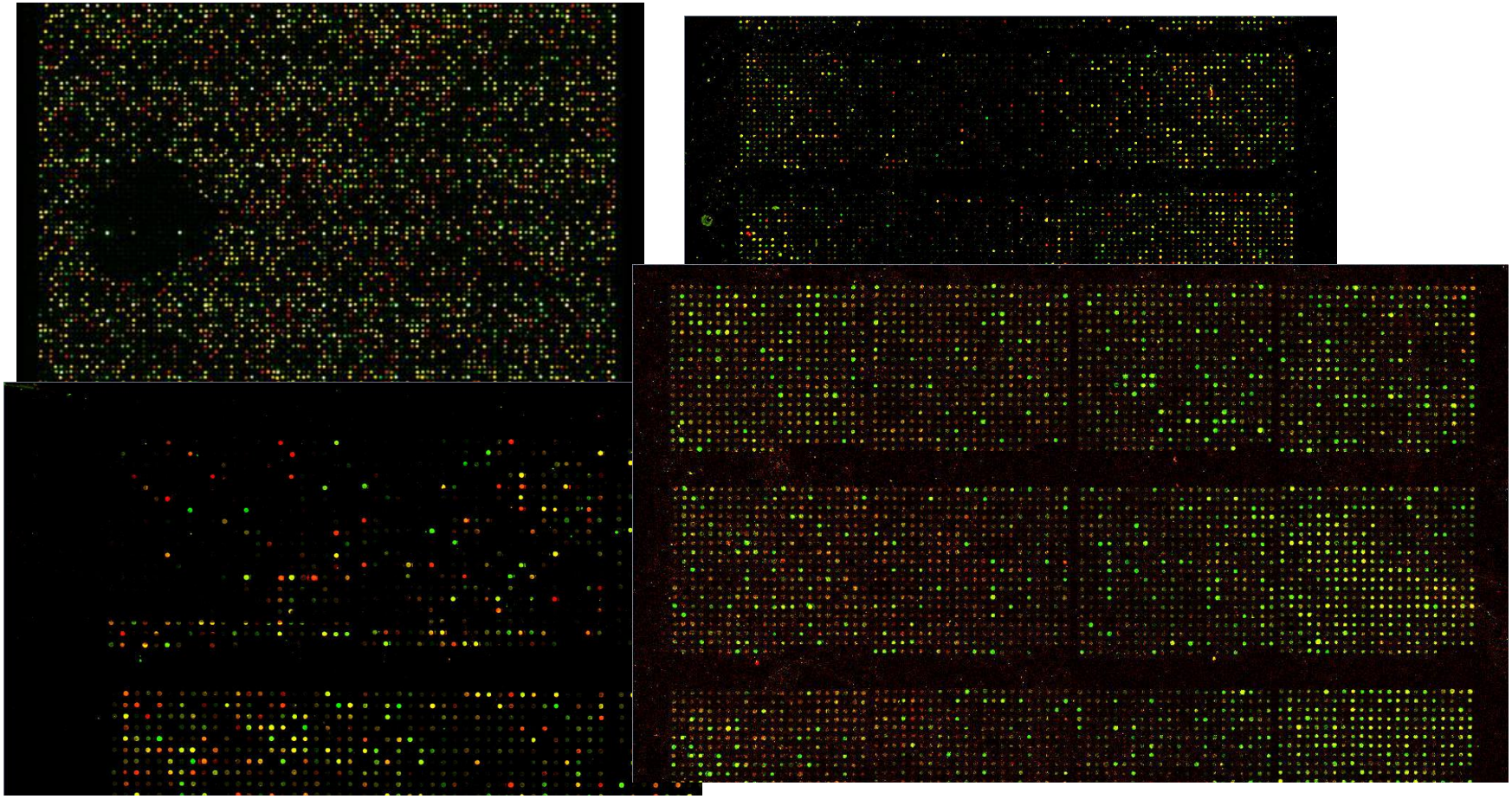
食品科技

outline

- Data pre-processing
- Differential Expressed Genes
- Correlation
- Classification

Data pre-processing

- **Microarray data is highly noisy**



Data pre-processing

- Housekeeping Genes
- Background and noise subtraction
- Normalization
- GCRMA, MAS5, RMA

Gene1	100	50
Gene2	250	130
Gene3	50	25
…		
Gene5000	500	250

Differential Expressed Genes

- statistical significant difference

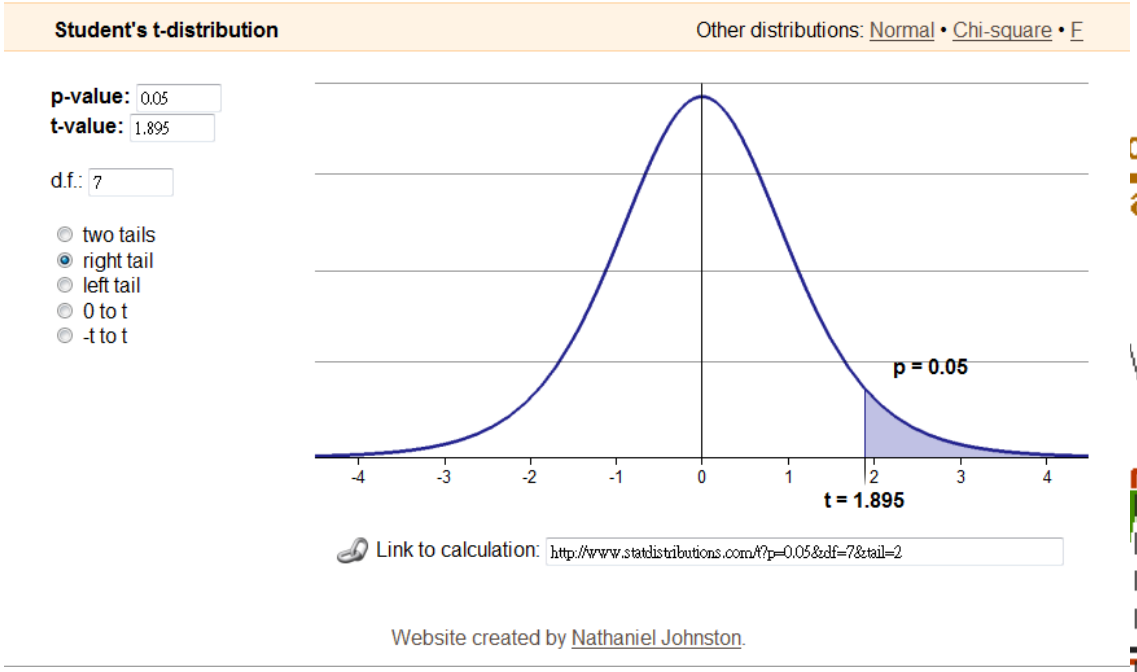
	CancerA 1	CancerA 2	CancerB 1	CancerB 2	
RAD21	9.49	10.19	7.76	7.33	YES
AP2B1	6.08	7.8	4.54	5.16	YES
CAP1	11.97	11.81	11.08	10.87	NO
CALM3	7.84	8.02	7.14	6.89	?
EEF1G	12.6	12.59	12.24	11.53	?

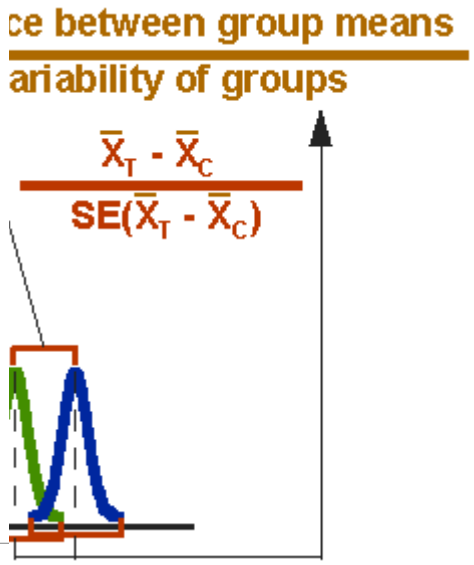
Statistical hypothesis testing

- t-test
- p<0.05

$$H_0: \mu = \mu_0$$

$$H_1: \mu > \mu_0$$





Gene-Gene Correlation

- Expression similarity ➔ functionally related
- Pearson's correlation coefficient

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

- Positive or Negative correlation

Correlation is not enough

Genetic network of cell cycle control in Caulobacter

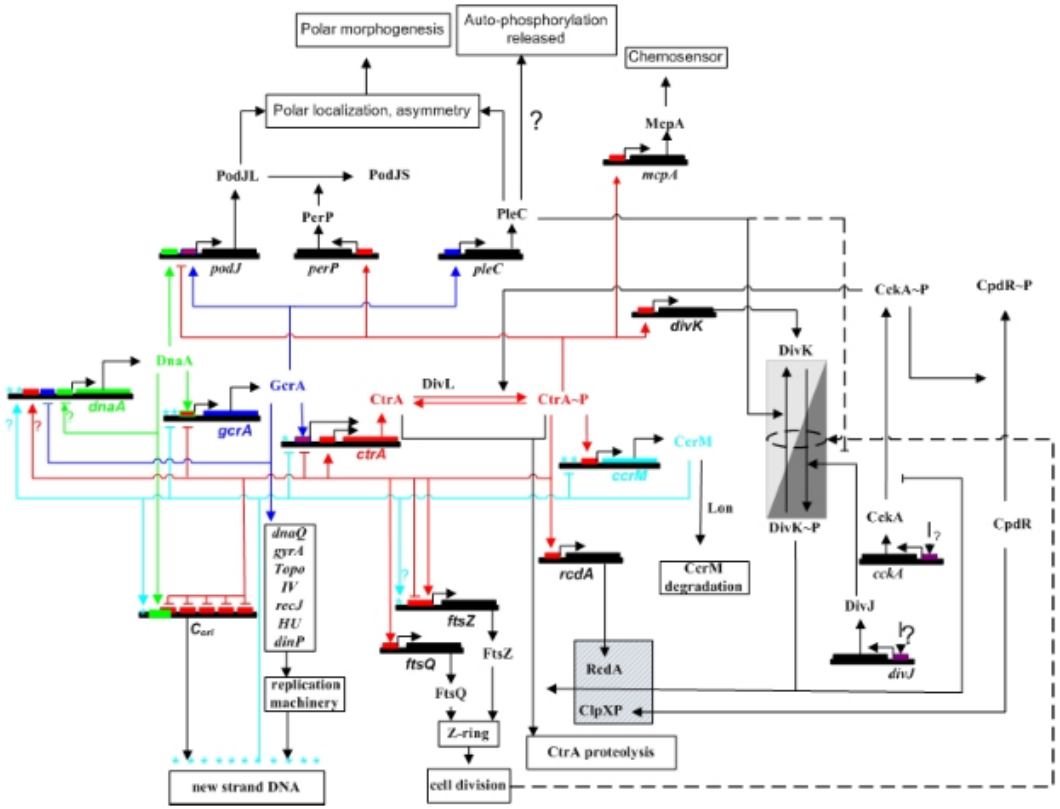
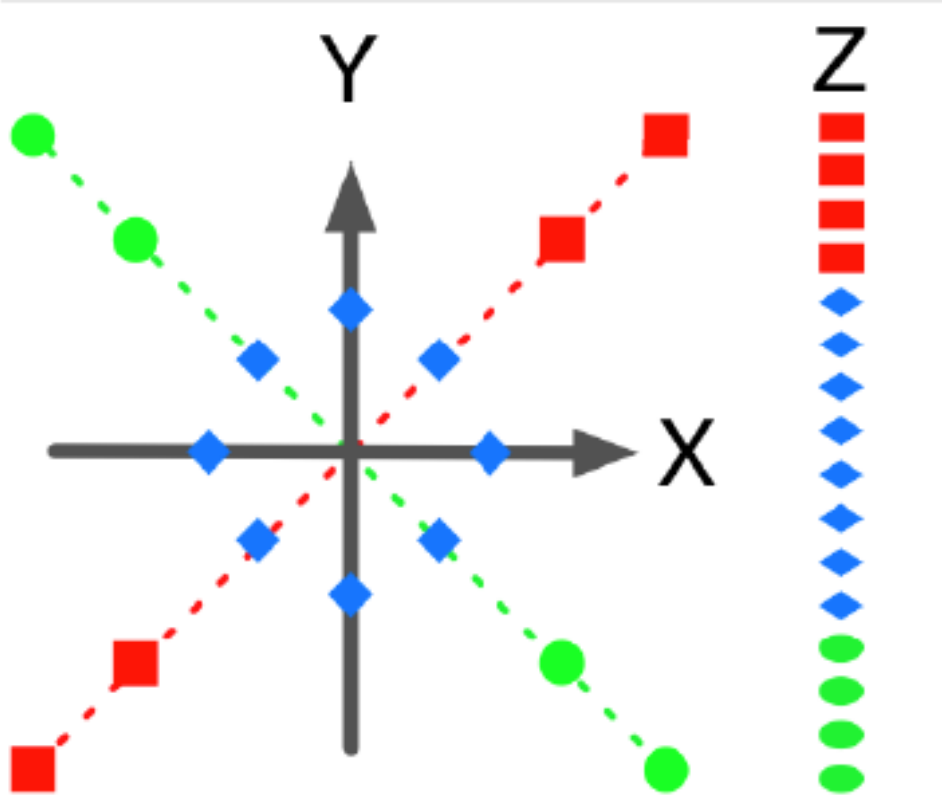


Figure 3. Known cell cycle genes in *Caulobacter crescentus* (adapted from (McAdams and Shapiro, 2003)). Regulation of genes by CtrA is shown in red, by GcrA in blue, by DnaA in green, and by CcrM in cyan. The cyan stars indicate those genes whose transcription is regulated by DNA methylation The CtrA-driven upregulation of the dnaA gene (red line with ? mark) is suggested by microarray data (Laub et al., 2002; McAdams, 2005). DnaA self-regulation (blue line with ? mark) is proposed from the fact that the dnaA promoter has DnaA boxes (Zweiger and Shapiro, 1994).

Example : Liquid Association

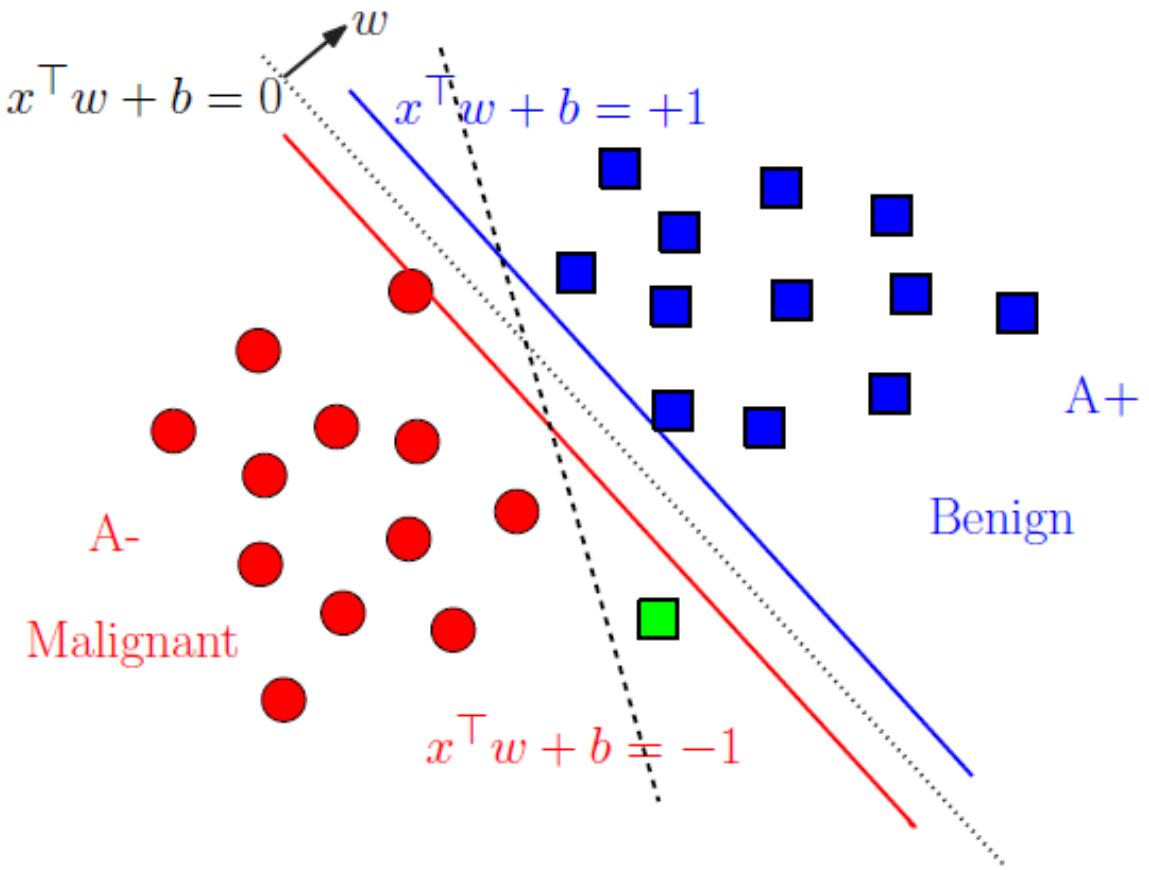


Classification

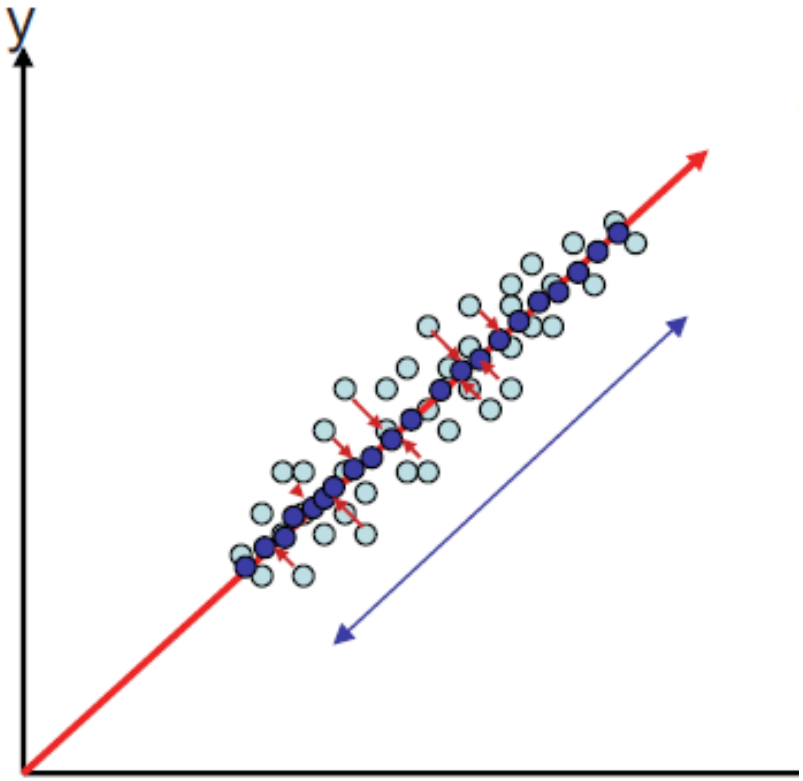
- Pattern Recognition
- Supervised/Unsupervised Learning
- Clustering
- SVM
- PCA
- LDA
- KNN

	CancerA 1	CancerA 2	CancerB 1	CancerB 2	Cancer?
RAD21	9.49	10.19	7.76	7.33	9.26
AP2B1	6.08	7.8	4.54	5.16	7.87
CAP1	11.97	11.81	11.08	10.87	10.59
CALM3	7.84	8.02	7.14	6.89	7.93
EEF1G	12.6	12.59	12.24	11.53	12.4
…	…	…	…	…	…

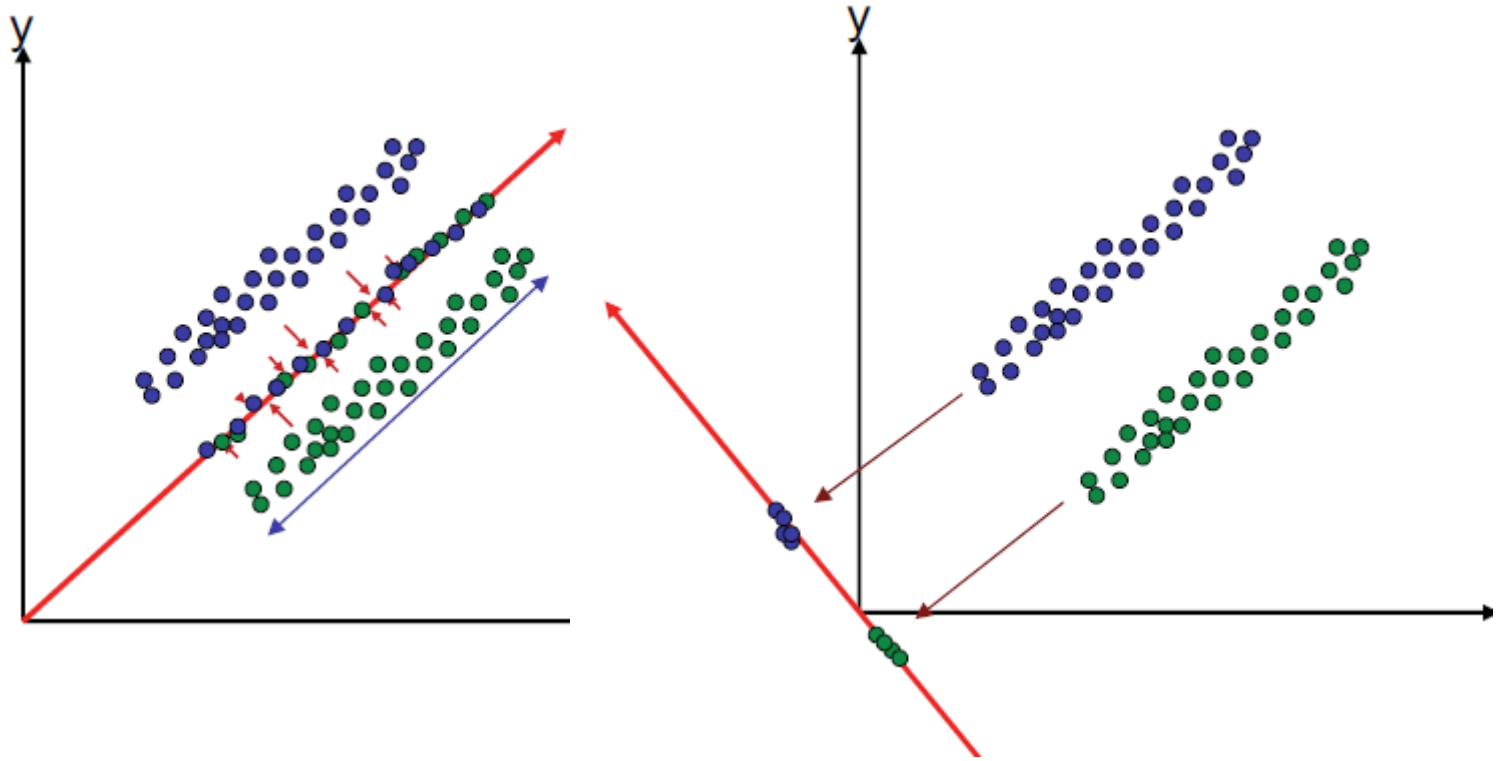
SVM



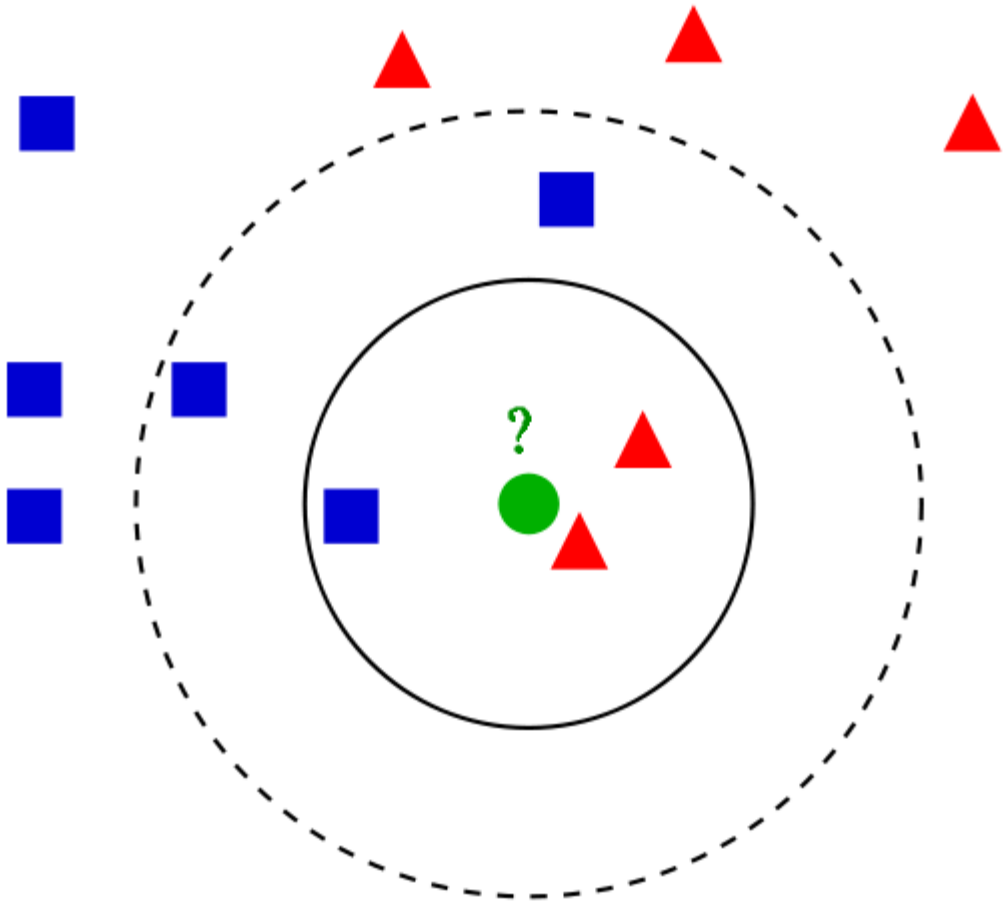
PCA



LDA



KNN



參考資料

http://en.wikipedia.org/wiki/

http://ultrasound.ee.ntu.edu.tw/belab/

http://en.wikipedia.org/wiki/KNN

http://www.socialresearchmethods.net/kb/stat_t.php

http://www.statdistributions.com/t/?p=0.1&df=12&tail=2

http://www.phalanxbiotech.com/tech_support/troubleshoot.html

http://www.bio.davidson.edu/projects/gcat/protocols/Troubleshooting_tiffs.html

http://discover.nci.nih.gov/cellminer/home.do

http://mpf.biol.vt.edu/research/caulobacter/pp/genetic_network.php

http://www.stat.sinica.edu.tw/kcli/