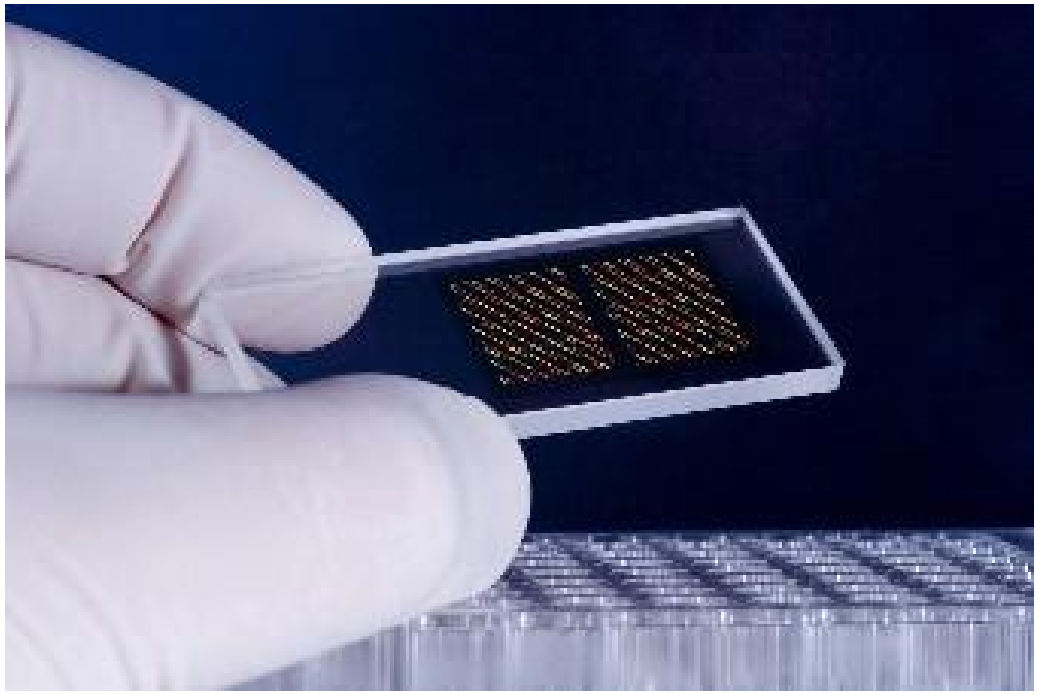
Βιοπληροφορική ΙΙ

Παντελής Μπάγκος
Αναπληρωτής Καθηγητής

Πανεπιστήμιο Θεσσαλίας
Λαμία, 2015

Μικροσυστοιχίες

Γυάλινο πλακίδιο που αποτελείται από συγκεκριμένες αλληλουχίες οι οποίες είναι ειδικές για συγκεκριμένα γονίδια, τους ανιχνευτές (probes), οι οποίοι είναι ακινητοποιημένοι σε μία κουκκίδα (spot) της γυάλινης επιφάνειας του πλακιδίου.



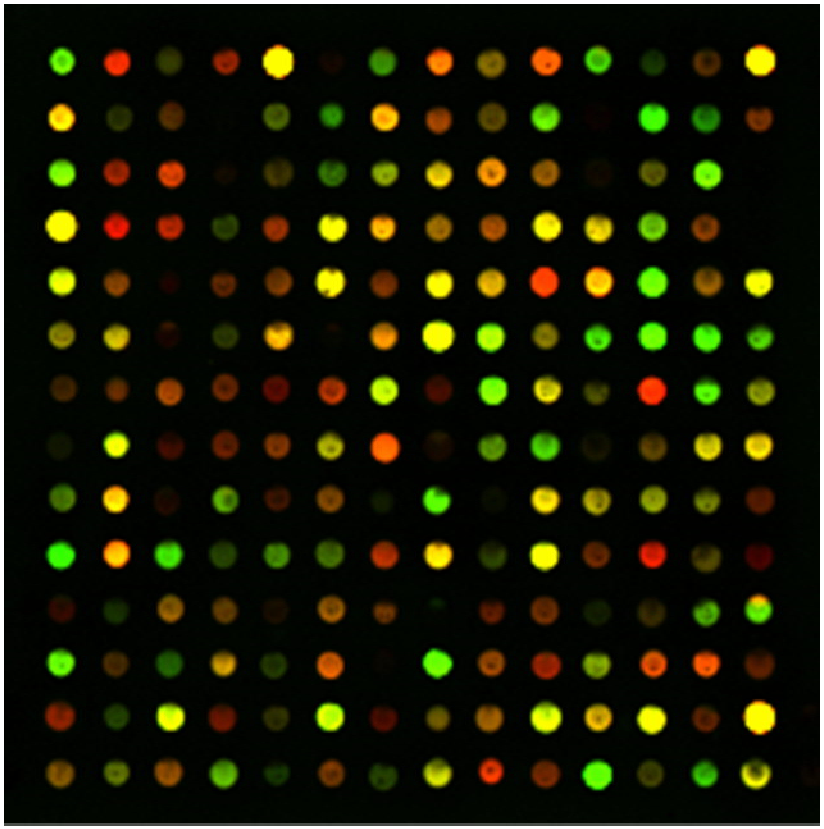
Μικροσυστοιχίες

- Ταυτόχρονη ανάλυση του τρόπου έκφρασης χιλιάδων γονιδίων σε διαφορετικά δείγματα ή σε διαφορετικά στάδια ανάπτυξης
- Σύγκριση έκφρασης σε φυσιολογικές και παθολογικές καταστάσεις
- Ανταπόκριση σε φαρμακευτικές ουσίες ή θεραπείες
- Παρέχουν χρήσιμες πληροφορίες για τη βιολογική λειτουργία ενός οργανισμού, βρίσκοντας ποια γονίδια ενεργοποιούνται ή καταστέλλονται σε διάφορα στάδια ανάπτυξης ή σε απόκριση σε ερεθίσματα του περιβάλλοντος, όπως η απόκριση σε ορμόνες ή σε υψηλή θερμοκρασία

Βασικά βήματα για ένα πείραμα μικροσυστοιχιών

- Διατύπωση του βιολογικού ερωτήματος
- Επιλογή του κατάλληλου τύπου μικροσυστοιχίας (τυπωμένες μικροσυστοιχίες cDNA, τυπωμένες μικροσυστοιχίες ολιγονουκλεοτιδίων, μικροσυστοιχίες που κατασκευάστηκαν με *in situ* σύνθεση ολιγονουκλεοτιδίων)
- Απομόνωση του RNA από τα δείγματα
- Σήμανση των δειγμάτων με φθορίζουσες ουσίες
- Υβριδισμός στην επιφάνεια της μικροσυστοιχίας
- Σάρωση μικροσυστοιχίας στα μήκη κύματος των φθορίζουσων ουσιών και μετρώντας τον αντίστοιχο φθορισμό της κάθε ουσίας
- Χρήση κατάλληλων προγραμμάτων για τη δημιουργία της τελικής εικόνας των μικροσυστοιχιών.

Μικροσυστοιχίες

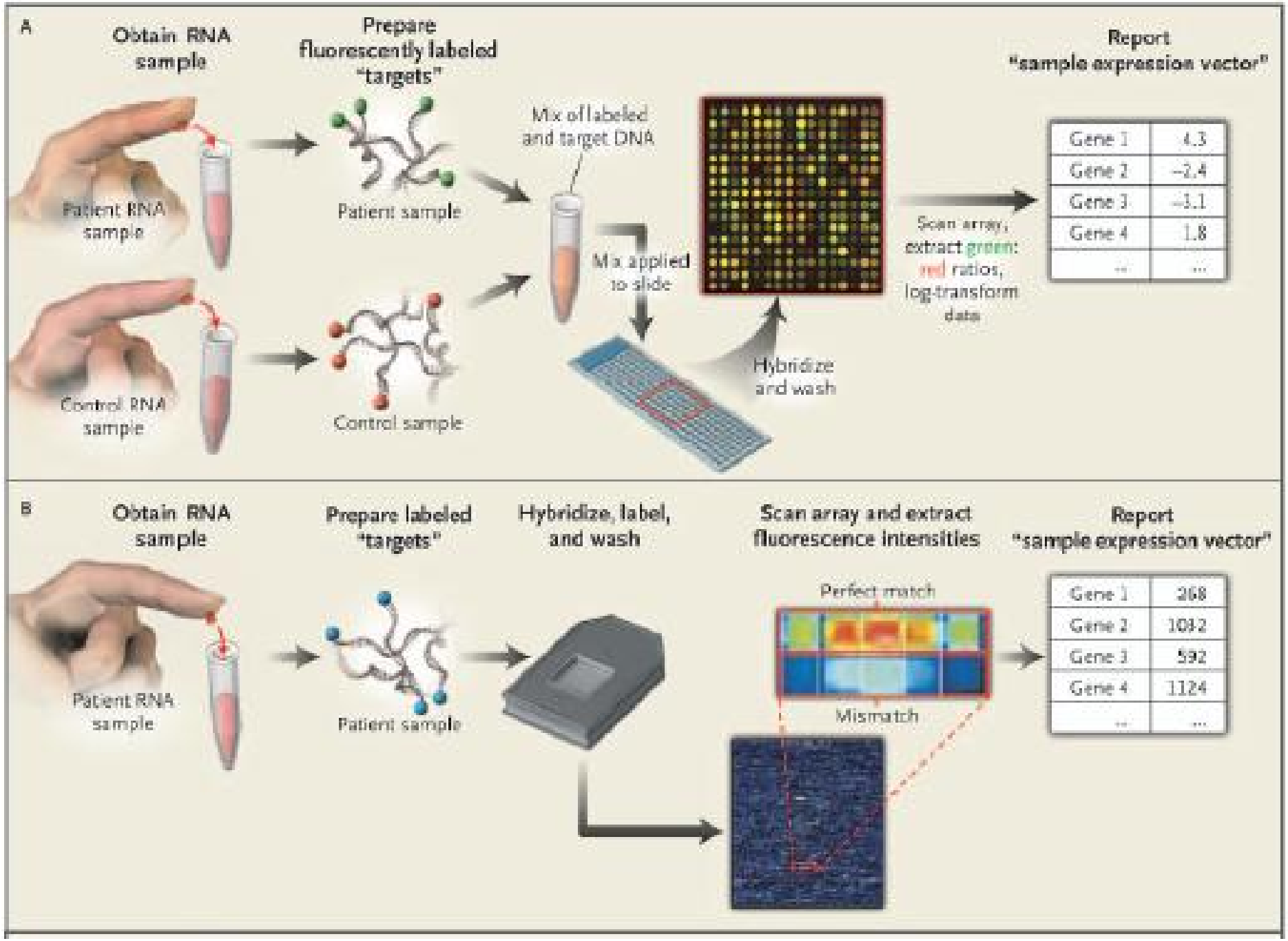


Η συνδυασμένη εικόνα της μικροσυστοιχίας παρέχει ένα βολικό τρόπο ώστε να βρεθούν τα γονίδια τα οποία βρίσκονται σε μεγαλύτερη έκφραση στο δείγμα ελέγχου σε σύγκριση με το δείγμα αναφοράς

Μικροσυστοιχίες

- Μονοχρωματικές μικροσυστοιχίες (Affymetrix): Κάθε δείγμα RNA σημαίνεται με μια χρωστική και τοποθετείται για υβριδισμό σε ένα τσιπ μικροσυστοιχιών.

- Διχρωματικές μικροσυστοιχίες: Δύο δείγματα RNA (ελέγχου – αναφοράς) σημαίνονται με 2 διαφορετικές φθορίζουσες ουσίες και το τοποθετούνται για υβριδισμό στο ίδιο τσιπ μικροσυστοιχιών.


A
Obtain RNA sample
Prepare fluorescently labeled "targets"
Report "sample expression vector"
Patient RNA sample
Patient sample
Mix of labeled and target DNA
Mix applied to slide
Scan array, extract green: red ratios, log-transform data
Gene 1 4.3
Gene 2 −2.4
Gene 3 −3.1
Gene 4 1.8
Control RNA sample
Control sample
Hybridize and wash
B
Obtain RNA sample
Prepare labeled "targets"
Hybridize, label, and wash
Scan array and extract fluorescence intensities
Report "sample expression vector"
Patient RNA sample
Patient sample
Perfect match
Mismatch
Gene 1 268
Gene 2 1082
Gene 3 592
Gene 4 1124

Μικροσυστοιχίες

•Με κόκκινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος ελέγχου είναι μεγαλύτερο
•Με πράσινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν η ποσότητα του δείγματος αναφοράς είναι μεγαλύτερο
•Με κίτρινο χρώμα εμφανίζεται μια κουκκίδα, αν σε αυτήν οι ποσότητες του δείγματος ελέγχου και του δείγματος αναφοράς είναι ίσες
•Με μαύρο χρώμα εμφανίζεται μία κουκκίδα αν κανένα δείγμα δεν έχει υβριδοποιηθεί
•Οι υπόλοιπες αποχρώσεις εμφανίζονται για αντίστοιχες ποσότητες των δύο δειγμάτων

Ποσοτικοποίηση δεδομένων

•Η ένταση του φθορισμού μετατρέπεται σε αριθμητικά δεδομένα και δίνει πληροφορίες σχετικά με την έκφραση των γονιδίων της μικροσυστοιχίας.

•Το σχετικό επίπεδο έκφρασης για κάθε γονίδιο αντιστοιχεί με την ποσότητα του κόκκινου ή του πράσινου φωτός που εκπέμπεται μετά από διέγερση.

•Για να συσχετίσουμε αυτές τις ποσότητες και να εξάγουμε το σχετικό επίπεδο έκφρασης κάθε γονιδίου χρησιμοποιούμε το λόγο έκφρασης

$$T_i = \frac{R_i}{G_i} \qquad T_i' = \log_2(T_i)$$

Σφάλματα στα πειράματα μικροσυστοιχιών

Τυχαία και συστηματικά σφάλματα συμβαίνουν σε ένα πείραμα μικροσυστοιχιών:

- Χρήση διαφορετικών φθορίζουσων ουσιών
- Χρήση διαφορετικών πλατφορμών
- Διαφορετικές πειραματικές συνθήκες
- Εισαγωγή θορύβου στα δεδομένα από το σαρωτή

Κανονικοποίηση

Τρόπος ελαχιστοποίησης των σφαλμάτων στα επίπεδα έκφρασης

- Κανονικοποίηση ολικής έντασης (total intensity normalization)

- Lowess (locally weighted linear regression) κανονικοποίηση

Βάσεις δεδομένων μικροσυστοιχιών

- **GeneExpression Omnibus (GEO):** Βάση δεδομένων του NCBI που παρέχει δεδομένα γονιδιακής έκφρασης
  http://www.ncbi.nlm.nih.gov/geo/

- **Array Express:** Δημόσια βάση δεδομένων μικροσυστοιχιών η οποία διατηρείται στο Ευρωπαϊκό Ινστιτούτο Βιοπληροφορικής EBI **http://www.ebi.ac.uk/arrayexpress/**

- **ONCOMINE:** Βάση δεδομένων που περιέχει πειράματα μικροσυστοιχιών που αφορούν διαφόρους τύπους καρκίνου. Επίσης παρέχει στο χρήστη εργαλεία διαχείρισης των δεδομένων για την αποδοτικότερη εύρεση των επιθυμητών πειραμάτων και γονιδίων **http://www.oncomine.org/**

Δεδομένα μικροσυστοιχιών

"ID_REF"	"GSM800742"	"GSM800743"	"GSM800744"	"GSM800745"	"GSM800746"	"GSM800747"	"GSM800748"
"1007_s_at"	10.98412	11.11354	11.00474	11.1896	10.91828	10.97255	10.73529
"1053_at"	5.426875	5.940344	6.649304	6.436227	6.820375	6.378402	6.597894
"117_at"	5.265565	5.564025	5.817803	5.631599	5.231561	5.960334	5.546186
"121_at"	8.424749	8.27857	8.670591	8.590728	8.854136	8.911889	8.939678
"1255_g_at"	3.666319	3.737454	3.615889	3.642348	3.449736	3.715736	3.63812
"1294_at"	8.747962	8.904438	8.151748	9.039548	8.316022	8.413626	8.490143
"1316_at"	6.944241	6.310405	6.499703	6.195474	6.757213	6.446918	6.374754
"1320_at"	4.873748	5.029888	4.914144	4.770439	4.737259	4.976627	4.697868
"1405_i_at"	8.374513	6.55758	7.46423	6.75511	7.09383	8.290524	6.733742	7.4756	6.58236
"1431_at"	3.578451	3.570486	3.590849	3.486636	3.484305	3.599472	3.61397
"1438_at"	6.547136	6.172181	6.52345	6.491056	6.636283	6.492878	6.566638
"1487_at"	9.487066	10.11535	9.950716	9.501676	9.721743	9.94243	9.398685
"1494_f_at"	6.015643	6.029665	6.276587	6.275242	5.83434	6.03582	5.735351	5.76565
"1552256_a_at"	9.44986	8.520991	8.410604	8.527215	8.535135	8.668772	8.416965
"1552257_a_at"	9.731705	10.22188	10.48709	10.08306	10.27443	9.989593	10.148
"1552258_at"	4.568571	4.462218	4.542326	4.312526	4.259055	4.366878	4.25006
"1552261_at"	4.717711	4.547108	4.450623	4.606178	4.521616	4.6143	4.295897
"1552263_at"	5.712723	6.616152	7.193986	7.012848	7.118447	7.355439	7.75506
"1552264_a_at"	6.29749	6.153967	6.855052	6.428255	6.409831	7.231188	6.9378	7.3
"1552266_at"	4.444225	4.353084	4.128518	4.606124	4.323753	4.322809	4.18074
"1552269_at"	3.687429	3.475274	3.84762	3.686121	3.528507	3.418546	3.640182

Ανάλυση Μικροσυστοιχιών

1) Στατιστική ανάλυση για εύρεση γονιδίων που υπέρ ή υποεκφράζονται

2) Ομαδοποίηση (Clustering)

3) Πρόγνωση (Prediction)

Στατιστική Ανάλυση Μικροσυστοιχιών

- t-test

$$t_k = d_k / \sigma_k \quad \text{με} \quad d_k = \bar{x}_{1k} - \bar{x}_{2k} \quad \text{και} \quad \sigma_k = \sqrt{\sigma^2{}_{k1} / n1_{k1} + \sigma^2{}_{k2} / n2_{k2}}$$

- Παραλλαγές του t-test

1) $t_k = d_k / \left(s_0 + \sigma_k\right)$ με $\sigma_k = \sqrt{\sigma^2{}_{k1} / n1_{k1} + \sigma^2{}_{k2} / n2_{k2}}$

2) $t_k = d_k / \delta_k$ με $d_k = \bar{x}_{1k} - \bar{x}_{2k}$ και $\delta_k = \sqrt{A_k + \sigma^2{}_{1k} / m_{1k} + \sigma^2{}_{2k} / m_{2k}}$

με το A_κ να παίρνει τιμές: $A_k = \begin{cases} 1\left(d_k > \sigma_k < 1\right) \\ 0\left(otherwise\right) \end{cases}$

Στατιστική Ανάλυση Μικροσυστοιχιών

3) $t_p = \dfrac{\overline{X_1} - \overline{X_2}}{a + \sqrt{\dfrac{{s_1}^2}{n_1} + \dfrac{{s_2}^2}{n_2}}}$

με ***α*** ίσο με το 90% των σφαλμάτων όλων των γονιδίων

4) $t = \dfrac{\overline{x}_{2k}}{\overline{x}_{1k}}$

Μια άλλη επιλογή είναι να χρησιμοποιήσουμε τον λόγο των μέσων τιμών για τις δύο καταστάσεις.

- SAM (Significance Analysis of Microarrays) – Στατιστική Ανάλυση Μικροσυστοιχιών

$$S = \frac{\overline{x_{k1}} - \overline{x_{k2}}}{\delta + \sqrt{{\sigma^2}_{k1} / n1_{k1} + {\sigma^2}_{k2} / n2_{k2}}}$$

Computationally Intensive methods

- Bootstrap
- Permutation
- Bayesian t-test

Στατιστική Ανάλυση Μικροσυστοιχιών

- Παράδειγμα: Ας υποθέσουμε ότι εξετάζονται 10000 γονίδια τότε με p-value<0.05, 500 γονίδια αναμένεται να βρεθούν στατιστικά σημαντικά κατά τύχη (by chance)

- Ανάγκη χρησιμοποίησης των μεθόδων διόρθωσης για πολλαπλές συγκρίσεις
  - Bonferroni: $p_{cor(i)} = p_{(i)} * n$
  - Sidak: $p_{cor(i)} = 1 - (1 - p_{(i)})^{\frac{1}{n}}$
  - Holm: $p_{cor(i)} = (n - i) * p_{(i)}$
  - Holland: $p_{cor(i)} = (n - i + 1) * p_{(i)}$
  - FDR: $p_{cor(i)} = \frac{n}{n - i} * p_{(i)}$

Εφαρμογή t-test σε ζευγαρωτές παρατηρήσεις

TABLE 7.1: Data for ACAT2 from Data Set 7A

Patient	Before Treatment	After Treatment	Log Ratio	Fold Difference
7	−0.86	−2.17	−1.30	−2.47
10	−1.97	−1.93	0.04	+1.03
12	−2.07	−1.28	0.79	+1.73
14	−1.91	−2.32	−0.41	−1.33
15	−0.94	−2.00	−1.06	−2.09
18	−1.29	−1.74	−0.45	−1.37
26	−1.09	−1.54	−0.44	−1.36
27	−0.65	−0.60	0.06	+1.04
39	−1.69	−2.06	−0.37	−1.30
41	−0.79	−1.22	−0.43	−1.35
47	−1.19	−2.11	−0.91	−1.88
48	−1.36	−1.40	−0.04	−1.03
53	−1.11	−1.59	−0.48	−1.40
61	−1.82	−1.72	0.10	+1.07
100	−2.22	−2.13	0.10	+1.07
101	−1.76	−1.94	−0.18	−1.14
102	−1.51	−2.37	−0.86	−1.81
104	−1.65	−1.98	−0.33	−1.25
109	−0.78	−1.49	−0.71	−1.63
112	−1.80	−1.82	−0.03	−1.02
Average	−1.42	−1.77	−0.35	−1.21
Sample SD	0.48	0.43	0.48	

Note: In this experiment, the samples from before and after treatment have been hybridised to two separate arrays, with a common reference sample in the second channel. The measurements before and after treatment are the log ratios of the experimental sample to the reference sample. The log ratio is the difference between these two values; the logs are taken to base 2, so a value of 1 represents a 2-fold up-regulation, and −1 represents a 2-fold down-regulation. The sample standard deviations have been calculated with a denominator of $n - 1 = 19$ to ensure that they are unbiased estimators of the population standard deviation.

Εφαρμογή t-test σε παρατηρήσεις δυο δειγμάτων

TABLE 7.2: Data for Metallothionein IB from Data Set 7B

Patient	ALL Log	Patient	AML Log
1	8.60	28	8.42
2	7.85	29	8.35
3	8.85	30	9.58
4	8.20	31	9.18
5	7.60	32	9.41
6	8.21	33	8.96
7	8.47	34	8.81
8	8.51	35	9.55
9	8.75	36	8.18
10	6.75	37	8.71
11	7.93	38	9.46
12	7.71		
13	7.88		
14	7.55		
15	6.61		
16	8.75		
17	9.32		
18	8.40		
19	7.16		
20	8.41		
21	4.75		
22	7.92		
23	7.82		
24	8.42		
25	7.08		
26	7.38		
27	9.29		
Average	7.93		8.97
Sample s.d.	0.94		0.51
Fold Ratio	−1.84		+1.84

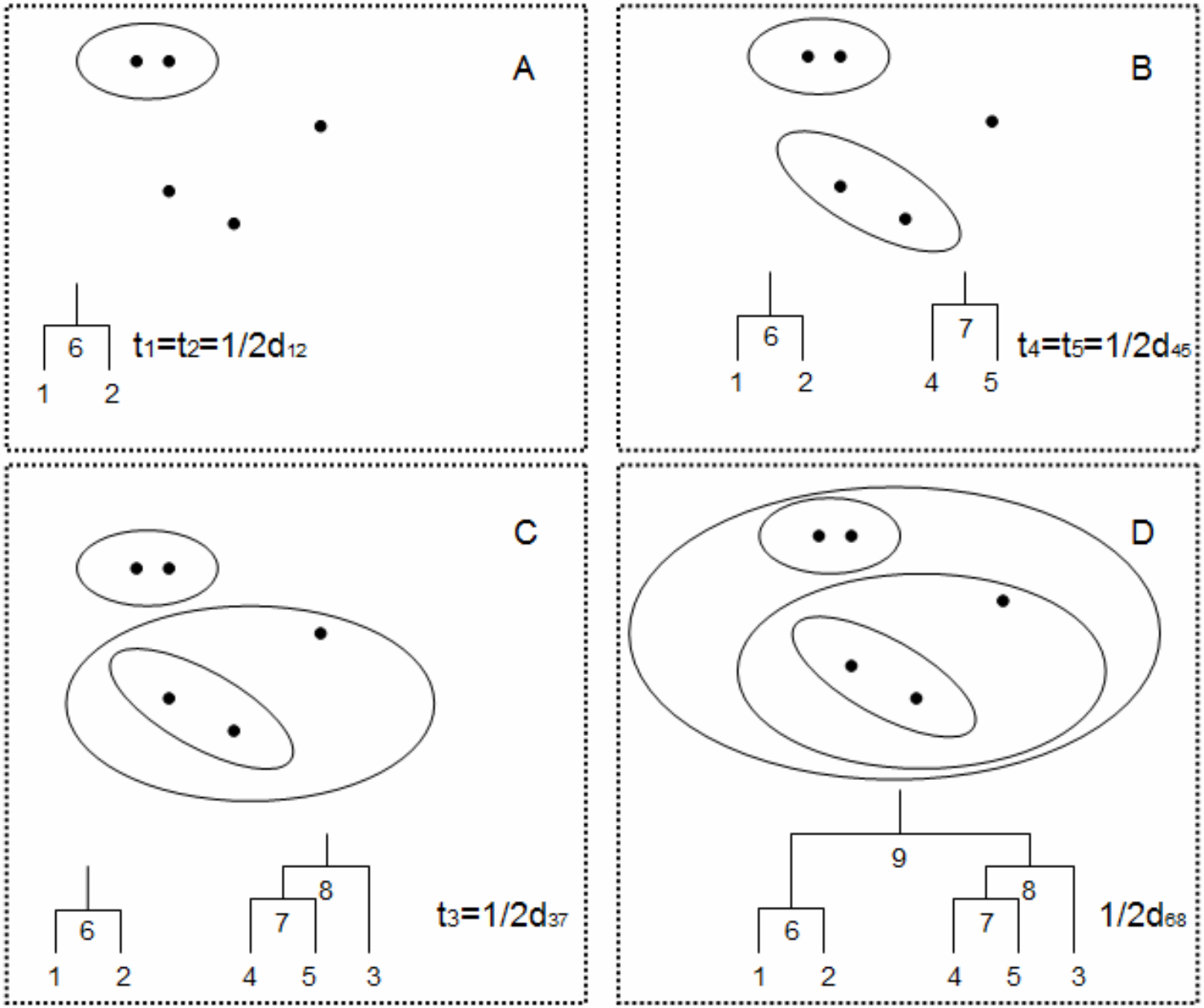
Note: This data came from Affymetrix arrays; the values have been logged (to base 2) to ensure that the data are normally distributed.

Ομαδοποίηση (Clustering)

- Ομαδοποιούνται μαζί γονίδια με βάση τα επίπεδα έκφρασης τους
- Αναπαράσταση των ομάδων αυτών με σκοπό την εύρεση πιθανών σχέσεων μεταξύ των γονιδίων
- Αλγόριθμοι ομαδοποίησης μπορούν να διαχωριστούν σε επιβλεπόμενους (supervised) και μη-επιβλεπόμενους (unsupervised)
- Η απόσταση (distance) μεταξύ δύο γονιδίων χρησιμοποιηται ως είσοδος στους αλγορίθμους ομαδοποίησης:
  - **Ευκλείδεια απόσταση** $$d_{AB} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$
  - **Απόσταση Manhattan** $$d_{AB} = \sum_{i=1}^{n} |x_i - y_i|$$
  - **Συντελεστής Συσχέτισης του Pearson** $$r = \frac{\sum_{i=1}^{n} (x_i - \bar{X})(y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{X})^2} \sqrt{\sum_{i=1}^{n} (y_i - \bar{Y})^2}}$$

Πρόγνωση

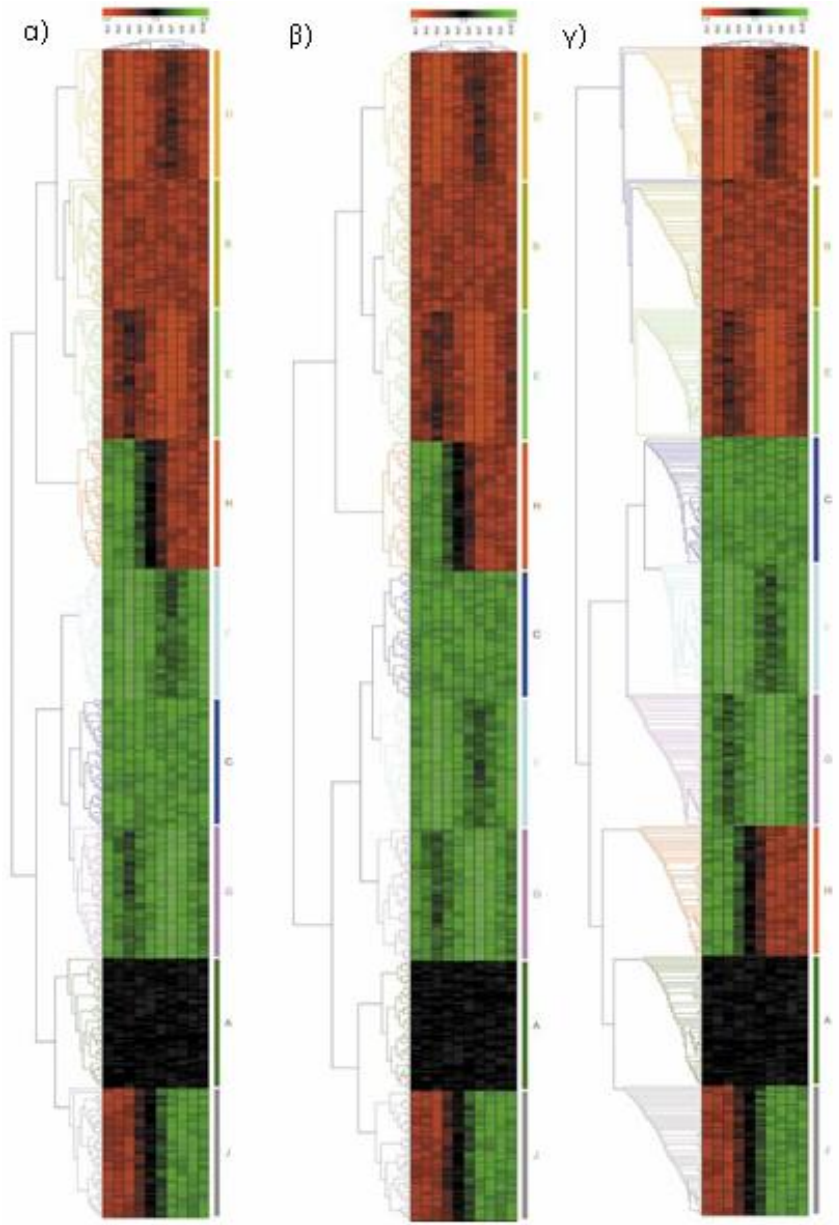
- Ενδιαφερόμαστε κυρίως για τη σωστή πρόγνωση (ταξινόμηση) των ασθενών.
- Έχει σημασία σε περιπτώσεις πρόβλεψης της ασθένειας, σαν διαγνωστική δοκιμασία
- Χρησιμοποιούνται οι συνηθισμένες μέθοδοι ταξινόμησης (Νευρωνικά Δίκτυα, SVM, κλπ)
- Πολλές φορές απαιτείται κάποια μέθοδος επιλογής των πιο σημαντικών γονιδίων


A
6
1
2
$t_1=t_2=1/2d_{12}$
B
6
7
1
2
4
5
$t_4=t_5=1/2d_{45}$
C
6
8
7
1
2
4
5
3
$t_3=1/2d_{37}$
D
9
6
8
7
1
2
4
5
3
$1/2d_{68}$

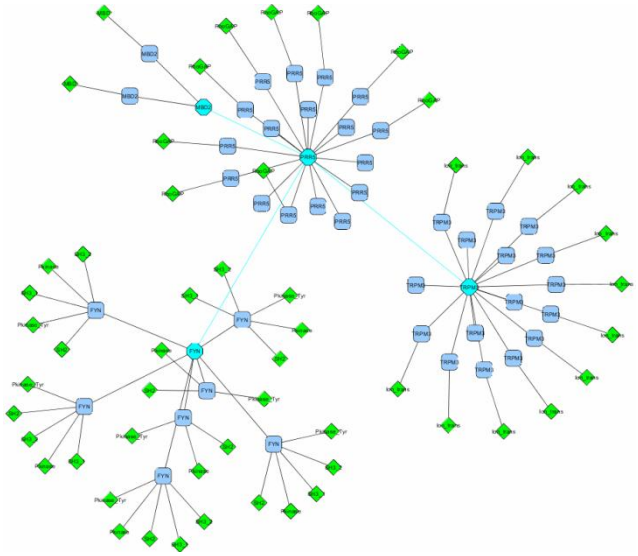
Αλγόριθμοι Ομαδοποίησης

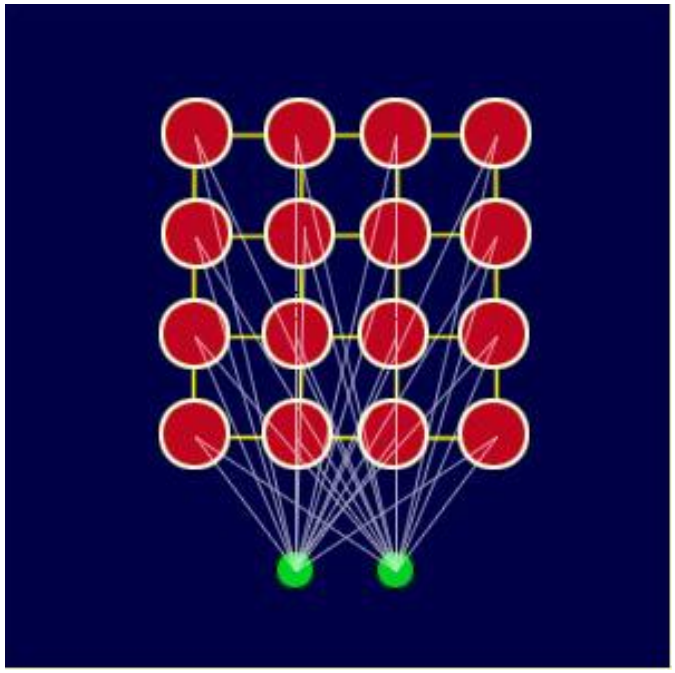
Ιεραρχική ταξινόμηση:

α)Single Linkage Clustering
β) Complete Linkage Clustering
γ) Average Linkage Clustering



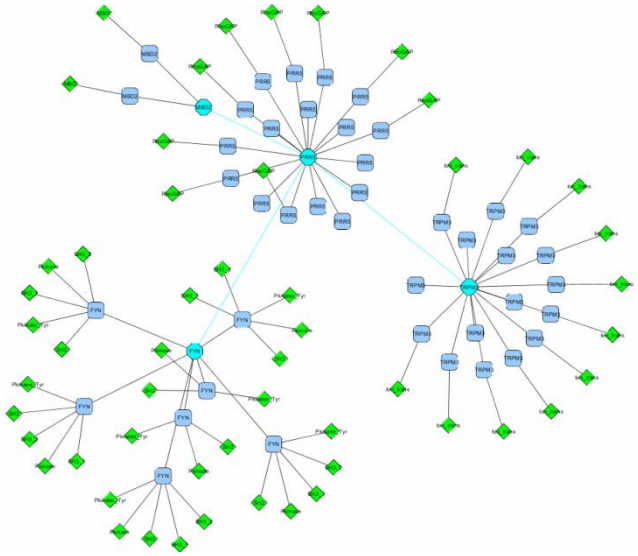
Αλγόριθμοι Ομαδοποίησης

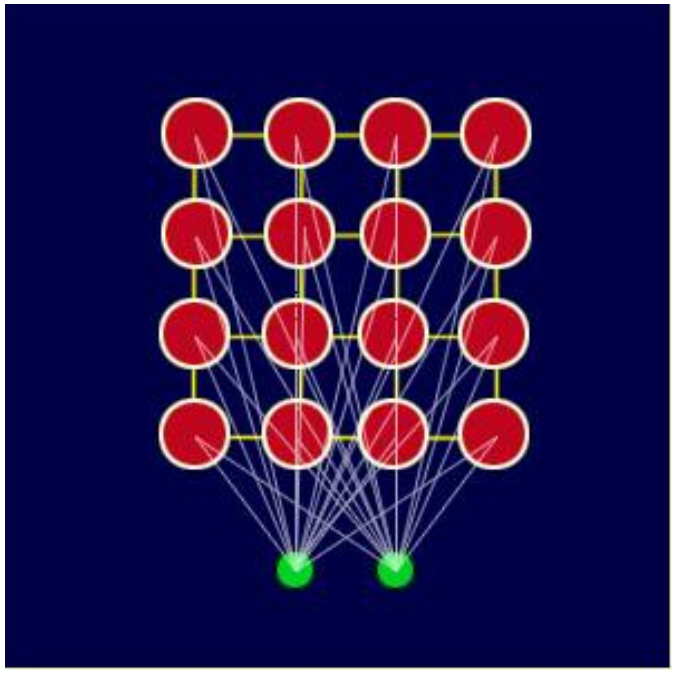




- K-means
- SOMs
- SVM
- PCA
- MCL

Αλγόριθμοι Ομαδοποίησης





- K-means
- SOMs
- SVM
- PCA
- MCL

Μετα-Ανάλυση

- Παρουσία θορύβου στα αποτελέσματα
- Μη επαναλήψιμα αποτελέσματα μεταξύ των πειραμάτων

- Στατιστικό εργαλείο που επεξεργάζεται τα δεδομένα και τα αποτελέσματα μελετών που ερευνούν το ίδιο ερώτημα

- Παρέχει ένα τελικό συμπέρασμα το οποίο προέρχεται από μια σύνθεση ανεξάρτητων συνόλων δεδομένων

Normand, S. L. (1999). "Meta-analysis: formulating, evaluating, combining, and reporting." Stat Med 18(3): 321-59

Μετα-ανάλυση Μικροσυστοιχιών

Μέθοδοι μετα-ανάλυσης:

- **t-test**

$$d_i = \frac{\bar{x}_{1i} - \bar{x}_{2i}}{sd_i} \qquad sd_i = \sqrt{\frac{(n_{1i}-1)sd_{1i}^2 + (n_{2i}-1)sd_{2i}^2}{n_{1i}+n_{2i}-2}}$$

- **Rank Product (Γινόμενο των βαθμών κατάταξης)**

$$RP_g = (\Pi_i \Pi_k r_{gik}) \frac{1}{k}$$

- **Συνδυασμός των p-values**

$$s_i = -2\sum_{k=1}^{K} \log(p_{ik})$$

Hong, F. and R. Breitling (2008). "A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments." Bioinformatics 24(3): 374-82.

Μετά το clustering και τη μετα-ανάλυση?

- Χρήση λογισμικών για εύρεσης κοινών χαρακτηριστικών μεταξύ ομάδων γονιδίων
- Δημιουργία γονιδιακών υπογραφών με σκοπό την πρόβλεψη ασθενειών