Lecture 10: Systems Biology

Torgeir R. Hvidsten

Professor
Norwegian University of Life Sciences

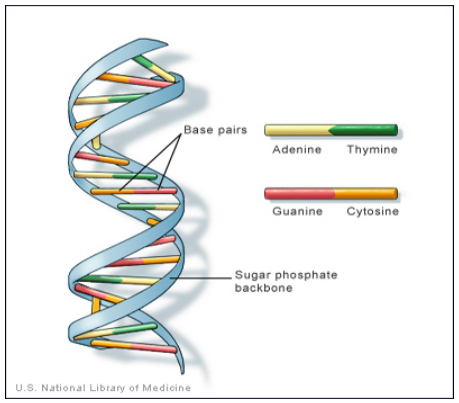
Guest lecturer
Umeå Plant Science Centre
Computational Life Science Cluster (CLiC)

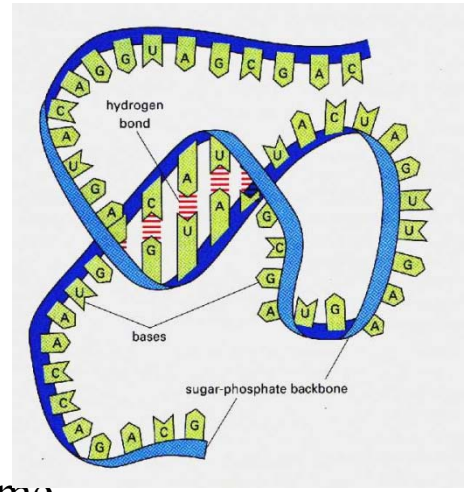
Systems Biology

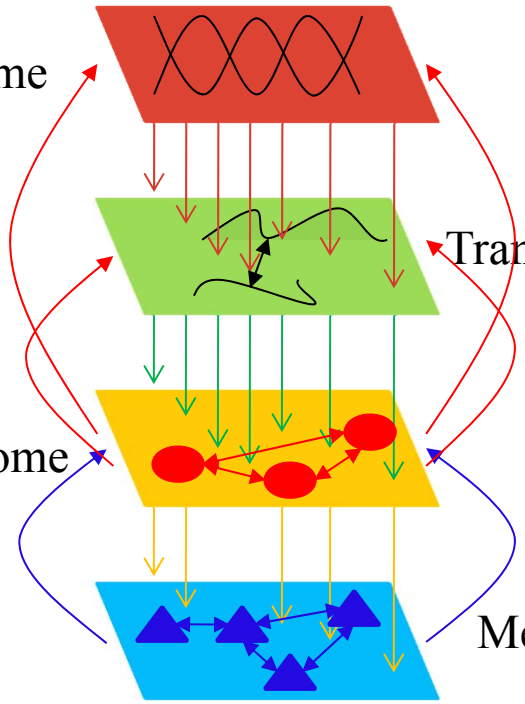
- Emergent properties
  - From reductionistic models (e.g. single genes)
  - to models describing interactions (e.g. gene networks)

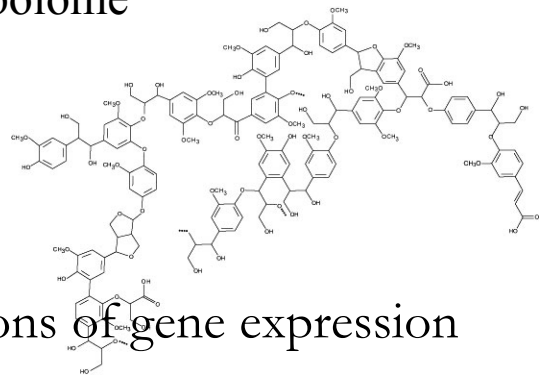
- Data integration
  - From one experimental platform (e.g. transcriptomics)
  - to integration of many (e.g. transcriptomics and proteomics)

'omics data





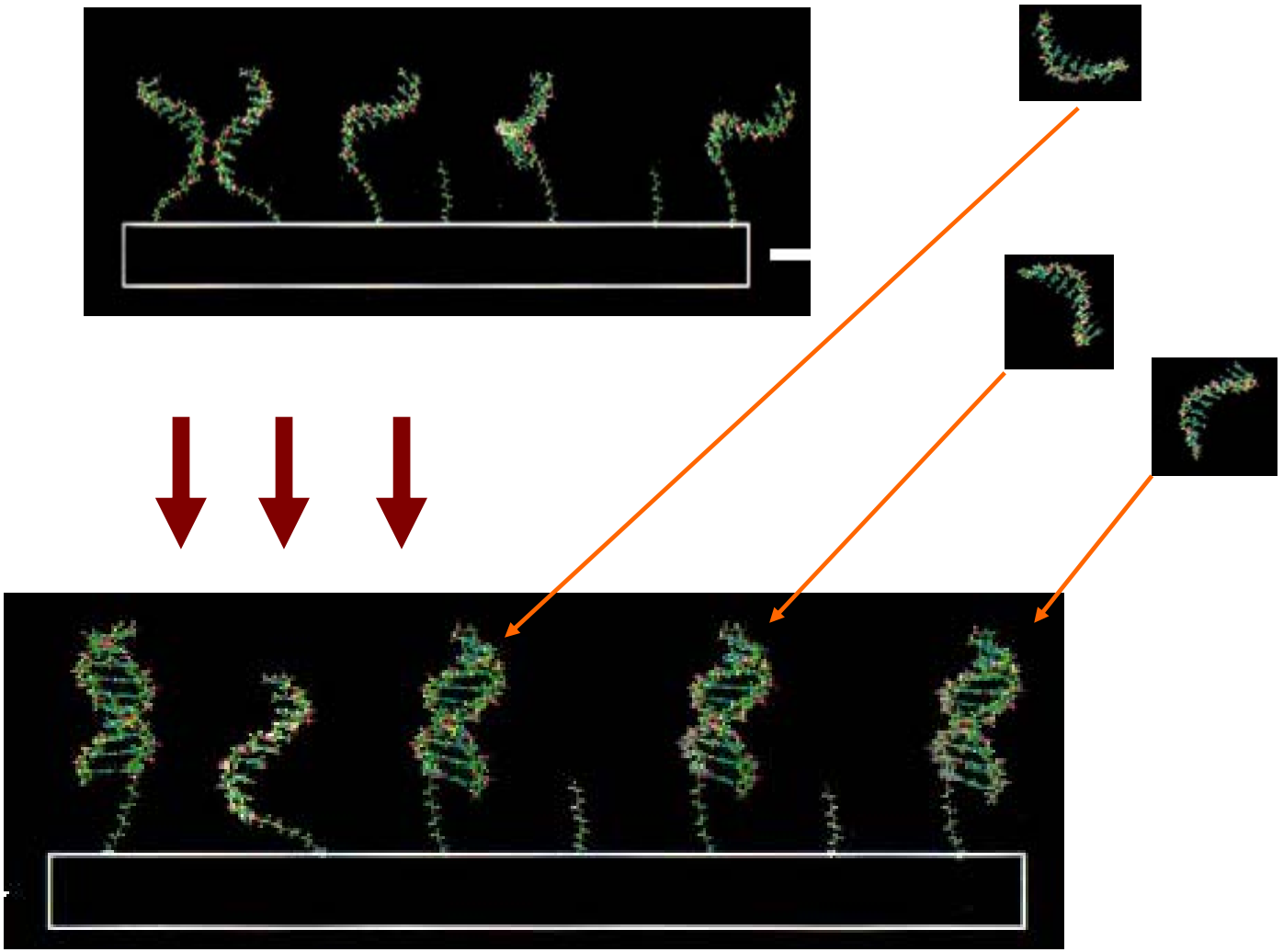




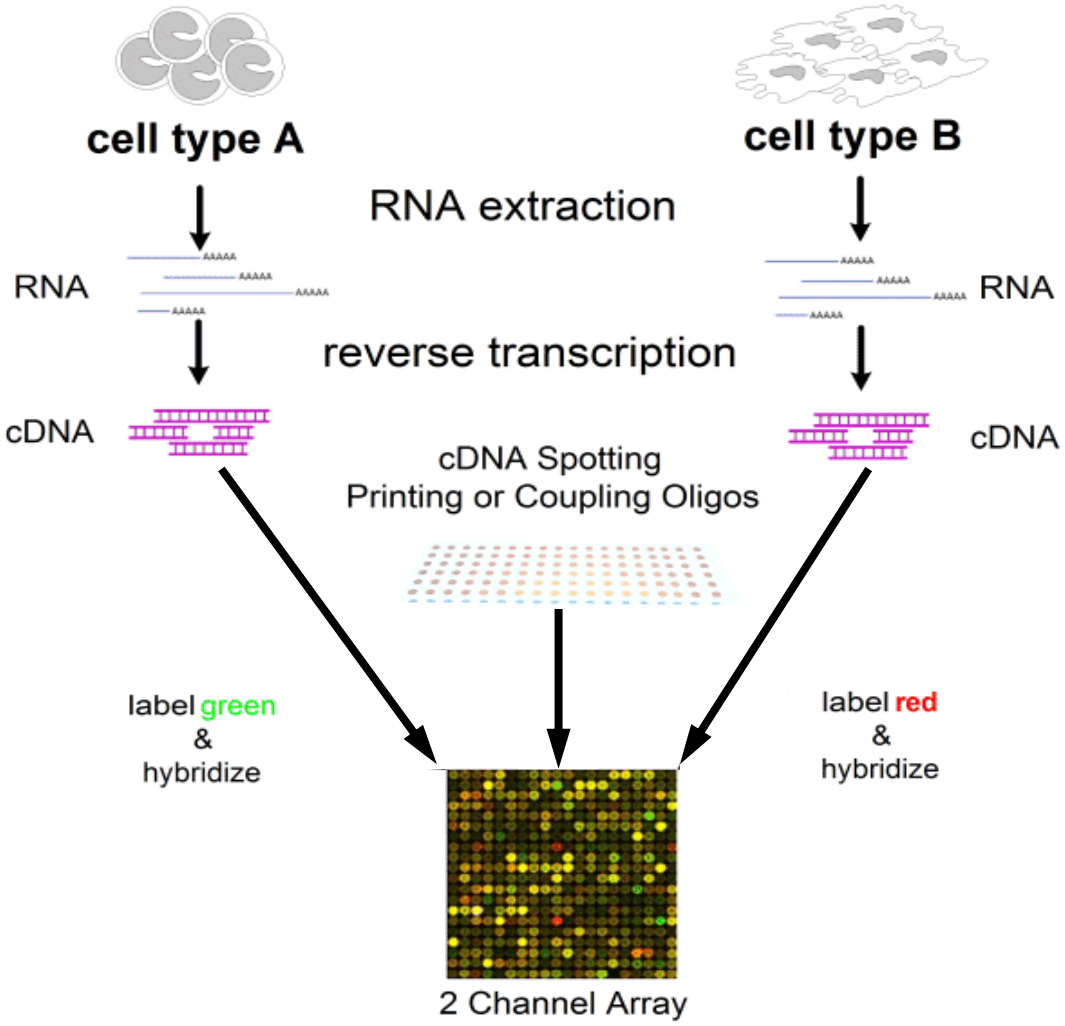
- Transcriptomics - quantifications of gene expression
- Proteomics - quantifications of proteins (peptides)
- Metabolomics - quantifications of metabolites

Transcriptomics

hybridization



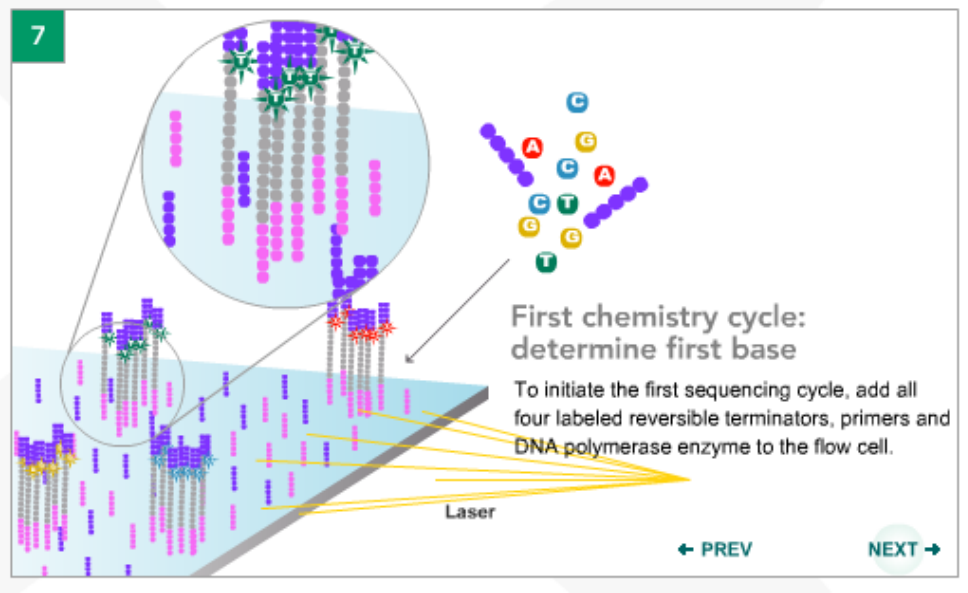
(Two-channel) microarrays



RNA-Seq

- Illumina HiSeq2000
  - Read length: 100bp
  - Paired-end reads: 2·100 bp
  - 150-300 Gbp per run
  - 10 lanes per run (flow cell)
  - 75-150 M reads per lane

- Multiplexing (bar-coding): many samples per lane



Illumina (HiSeq 2000)
Sequencing by synthesis

Illumina

Produce a “fasta file” of reads!

FASTQ Format:

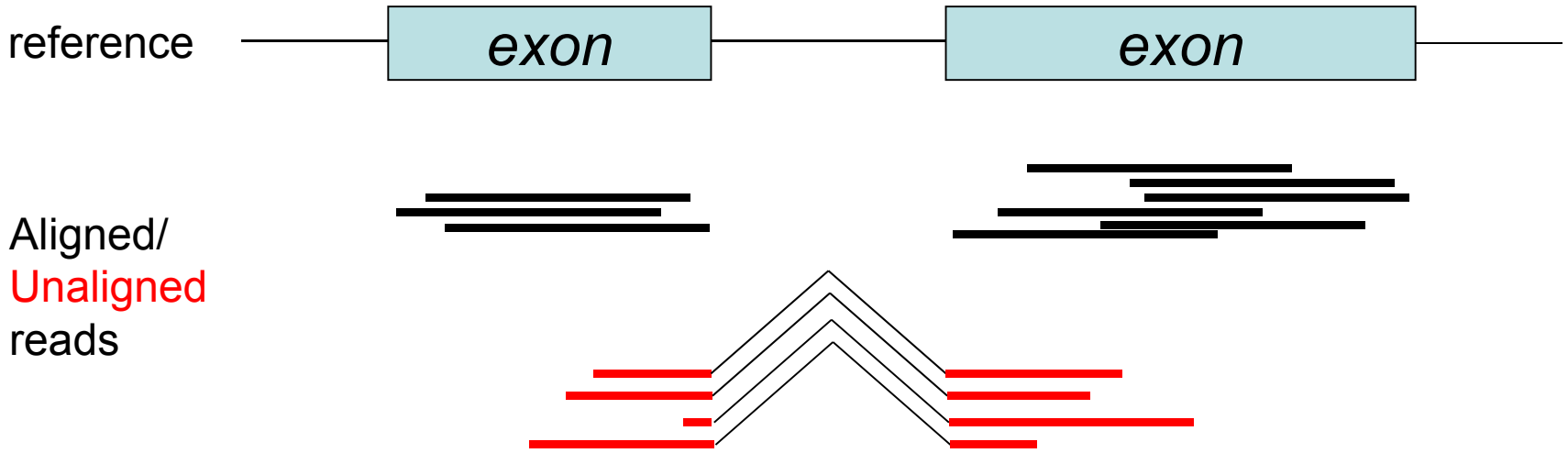
```
@HWI_ST139:1:1:1184:1942#CTTGTA/1 1
NTTGATCACAGAAACACAAGAATTTCTTCGAGATTTCACACTGTTTTCAGCTACAAATAAGCAAGGAATATCCAATTCCATGTATCAGGAACTCCTGGCAA
+HWI_ST139:1:1:1184:1942#CTTGTA/1 1
BMMKPUWTWW[[[[XQRUVR[YYZX^^VV^^^^^^^^^^^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI_ST139:1:1:1167:1955#CTTGTA/1 1
NAAACTTCCACATGATGTCCCATAACCAAGTCGAGTGCAGTCAGCTAGTCCGCAGGCATAGCTTACACTTAGTGGCACTTGTGGGTCATCAAGCGTGGCCA
+HWI_ST139:1:1:1167:1955#CTTGTA/1 1
BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI_ST139:1:1:1200:1991#CTTGTA/1 1
ATTTTTTTTATGATAGATTTATAGATAGTCATTGAGGTAAATTCTGTTTACCTGTTTGACTTTTATGTATCGGTAGCTAGAAAGGGGACGTCATGGCAATA
+HWI_ST139:1:1:1200:1991#CTTGTA/1 1
eeeeeeeecc\acT_V_Sab]H]U]TZWS]eecabc[Pcccc`ccb`b_cabP]`\[`cadbb__YYS^T`P^[^BBBBBBBBBBBBBBBBBBBBBBB
@HWI_ST139:1:1:1310:1934#CTTGTA/1 1
NTTTAAGAGTTTATCTGTGTGGCTTTTGGTAAGATGATTTTGCTGCAATAGAGTTTGGACTTGGAACATCTTGGGAGGTAAAGATGATGCTACTTAAGGAA
+HWI_ST139:1:1:1310:1934#CTTGTA/1 1
BUUMRSUTPU^[^^^[[]][[[[][XXZXX[XYYY[[V[[[[][[^^^^^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI_ST139:1:1:1328:1929#CTTGTA/1 1
NTTACATTTACCCACCCAGATTACTCAACCAAAAAAAAGCAATCAACGCTACAAAAGGTTGCAGTTTGAGATACCAAGTGACACTGGATCAGCTATGGTGCA
+HWI_ST139:1:1:1328:1929#CTTGTA/1 1
BUUUPSRRQTcc_cXVYUccc_cccc\ccc^^^^^UYScc]S\Z[^^^^^^V^BBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI_ST139:1:1:1350:1964#CTTGTA/1 1
GCAGATGTTTGGCTTCTTTAGCATCCTCACCCTCCATGCCCGATGGATTTTCACTGACTCTTGAATTTAAAGTGCCTGGTTTAATCCTCTGTACCTCCTCT
+HWI_ST139:1:1:1350:1964#CTTGTA/1 1
ggggggfggfgegegfdgggeggefgagggggdgggegggfffddcdbaaeeZebeeebdZZ[^c`dceebY[\]b]`]T`abbSY[O[R]````bbeXcB
@HWI_ST139:1:1:1629:1926#CTTGTA/1 1
NGATCATCTATAATTTTGATGAGATAAAAATCAATGTTTTGCAATTGAAACCATTCAGAAAAGACAGGGTTGCAGGAATGAAAATCCTGAAACCGGGAAAA
```

Mapping reads to reference genome

Raw data:

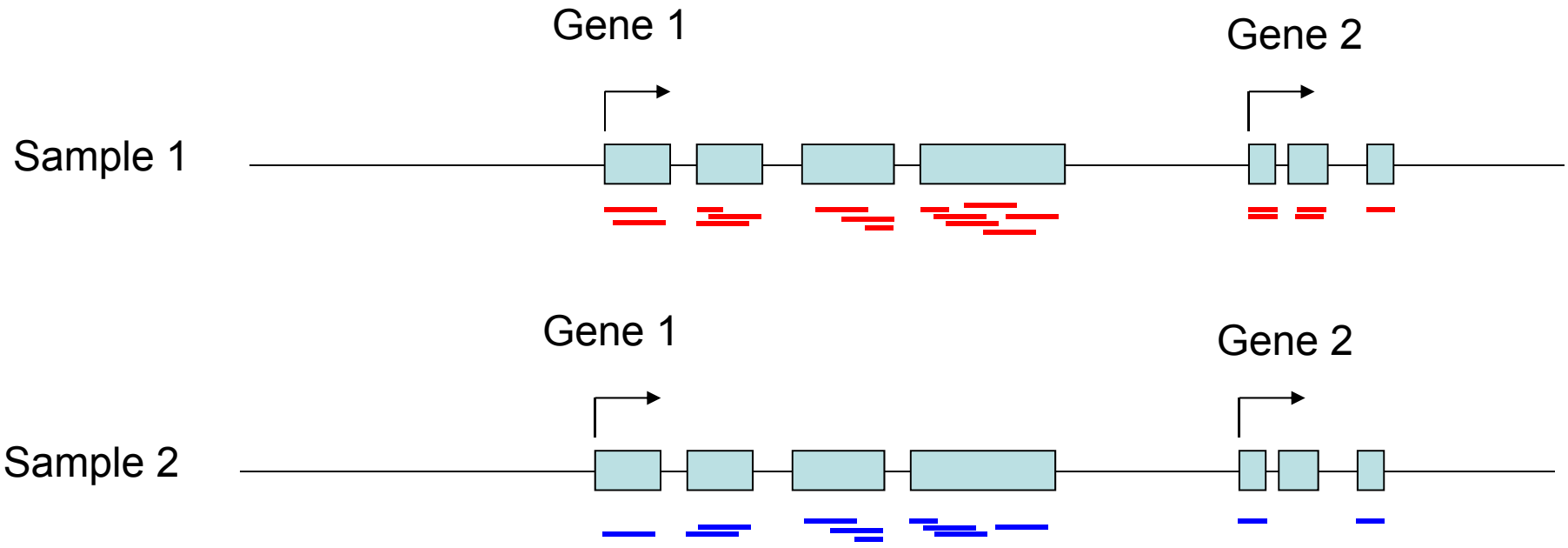
```
@HWI-EAS293:1:1:4:447#0/1
TTAAAGCGATCCAATGGTCGGATCTATATTTATGGACCTTTTGAGCTGGTACTCTAGTAGTGTGGGTGGAAT
@HWI-EAS293:1:1:7:1410#0/1
ATGGGGAAAGAAATTTGAGAATAAGTGACAGTAGAGAATTTCATGGGGACAGTGATGAAATGGTGAGTGAAAATGA
....
```

One file per sample or lane: 100 M reads, 20GB file



Quantifying expression

Count the number of reads mapped to each gene



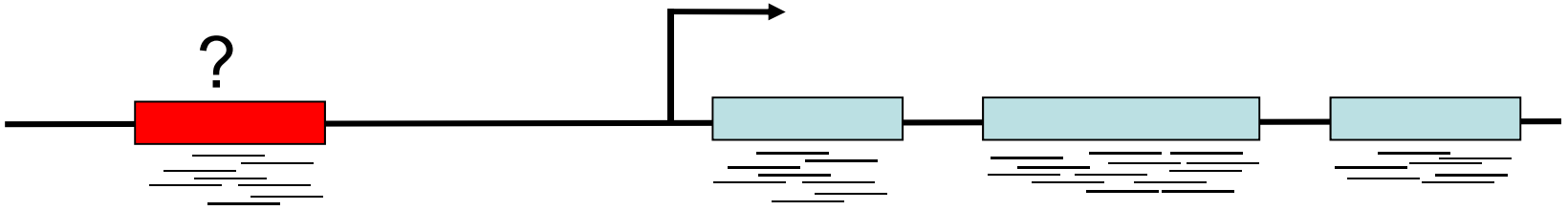
RPKM = Reads Per Kilobase of exon model per Million mapped reads

	Gene 1	Gene 2
Sample 1	**14 reads**	**5 reads**
Sample 2	**10 reads**	**2 reads**

	Gene 1	Gene 2
Sample 1	**0.18 RPKM**	**0.25 RPKM**
Sample 2	**0.25 RPKM**	**0.2 RPKM**

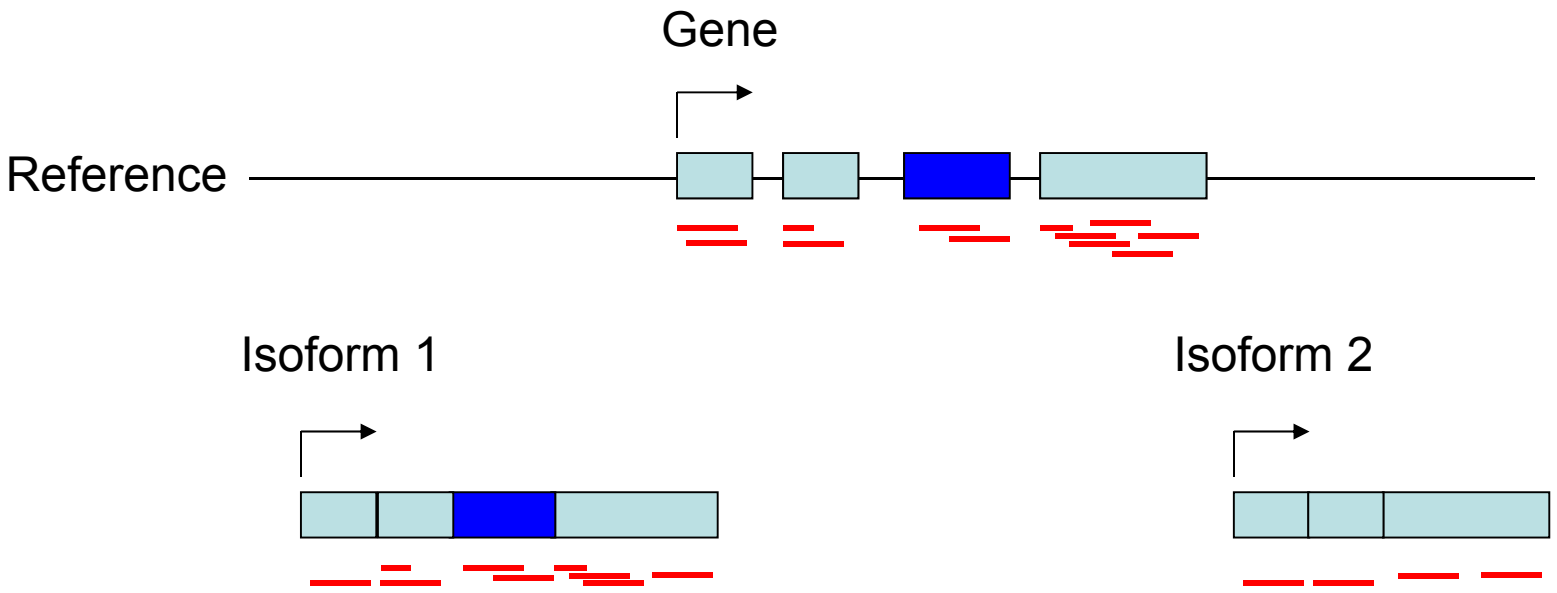
Novel transcribed regions

Detect regions outside known gene models



Isoform detection (splicing variants)

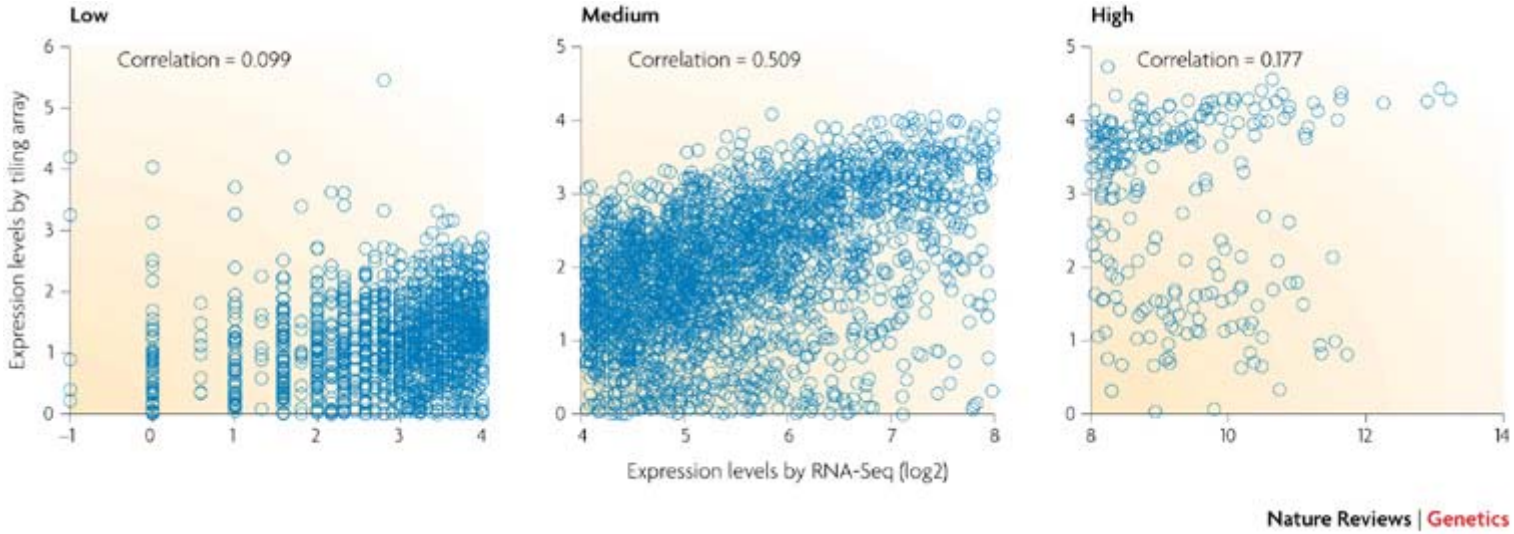
Detected by bethods that reconstruct entire transcripts



Microarrays versus RNA-Seq

RNA-Seq advantages:

- Can detect ”all” transcribed regions (including small RNA)
- Can be applied to all organisms (no reference genome needed)
- Broader dynamic range: higher sensitivity and specificity
- Can do much more than just quantifying gene expression (SNP detection, Isoform detection, etc)



Expression data

M < 100

Gene/Expr	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	...	EM
G1	0,72	0,10	0,57	1,08	0,66	0,39	0,49	0,28	0,50	0,66	...	0,52
G2	1,58	1,05	1,15	1,22	0,54	0,73	0,82	0,82	0,90	0,73	...	0,75
G3	1,10	0,97	1,00	0,90	0,67	0,81	0,88	0,77	0,71	0,57	...	0,46
G4	0,97	1,00	0,85	0,84	0,72	0,66	0,68	0,47	0,61	0,59	...	0,65
G5	1,21	1,29	1,08	0,89	0,88	0,66	0,85	0,67	0,58	0,82	...	0,60
G6	1,45	1,44	1,12	1,10	1,15	0,79	0,77	0,78	0,71	0,67	...	0,36
G7	1,15	1,10	1,00	1,08	0,79	0,98	1,03	0,59	0,57	0,46	...	0,39
G8	1,32	1,35	1,13	1,00	0,91	1,22	1,05	0,58	0,57	0,53	...	0,43
G9	1,01	1,38	1,21	0,79	0,85	0,78	0,73	0,64	0,58	0,43	...	0,47
...	...	...	...	...	...	...	...	...	...	...	...	...
GN	0,85	1,03	1,00	0,81	0,82	0,73	0,51	0,24	0,54	0,43	...	0,51

N ≈ 10000

Two-channel experiments: ratio-based intensities (“Red/Green”)

One-channel experiments: ”absolut” intensities

RNASeq: ”number” of transcripts expressed

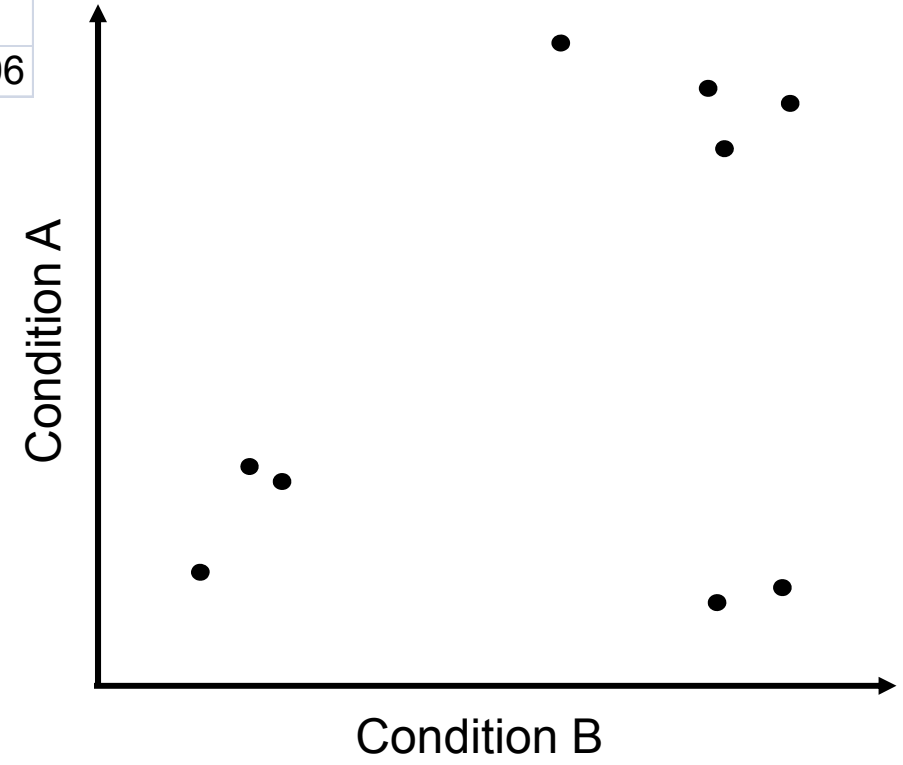
Multidimensional data

Conditions/tissues/time

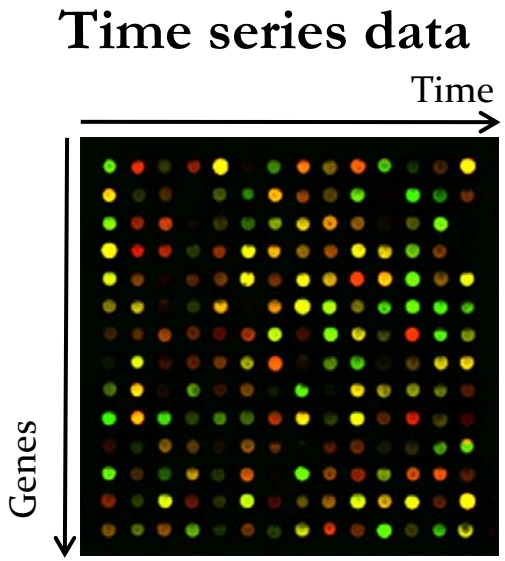
Genes/metabolites/proteins

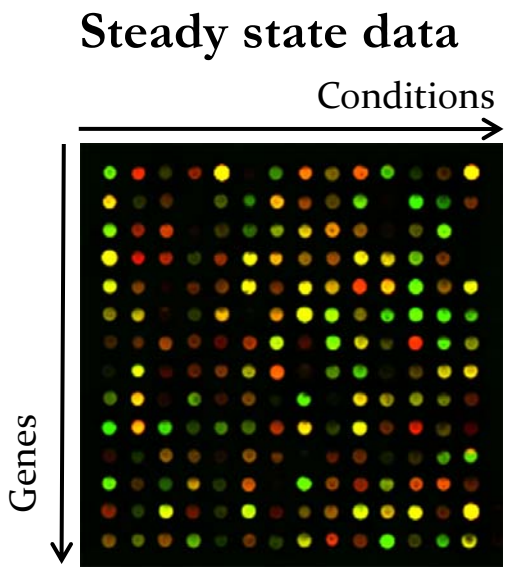
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06

Time series versus
Feature space



Data dimentionality:
How many samples do I need?

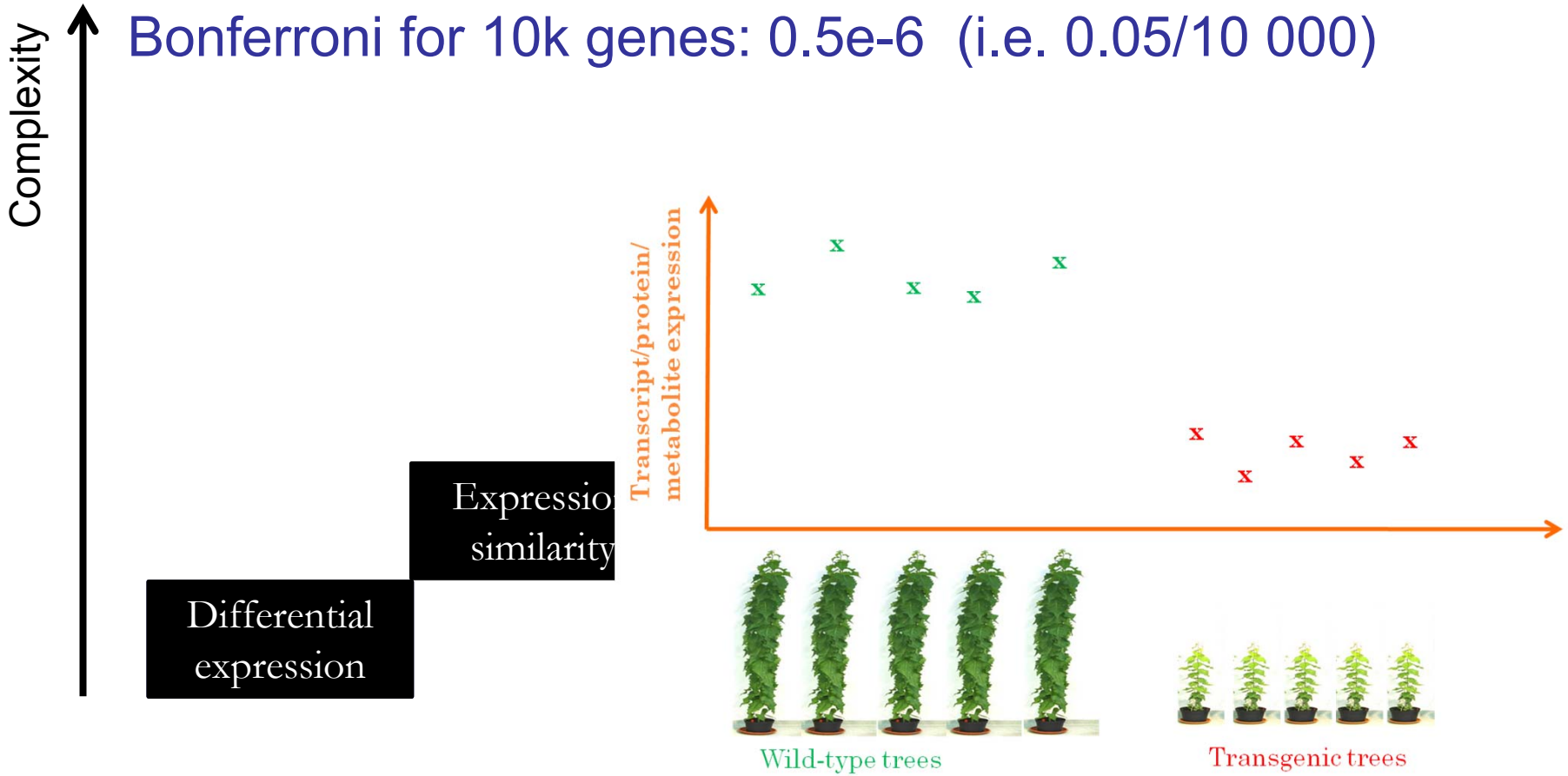




Complexity of data analysis

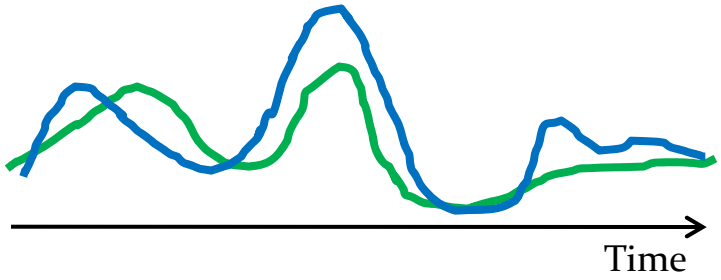
To do e.g. a t-test you need at least three biological replicates from each class

Bonferroni for 10k genes: 0.5e-6 (i.e. 0.05/10 000)



Complexity of data analysis

Class discovery/Clustering
Co-expression network
…

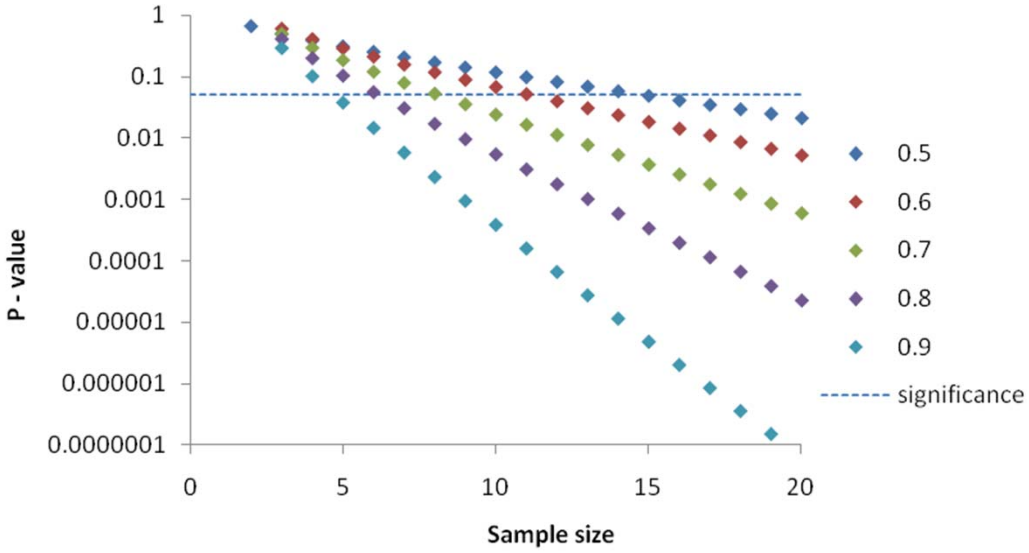


Complexity

Similarity can happen by chance
(e.g. Pearson correlation).

Expression similarity

Differential expression



Class discovery/Clustering

Need to define;

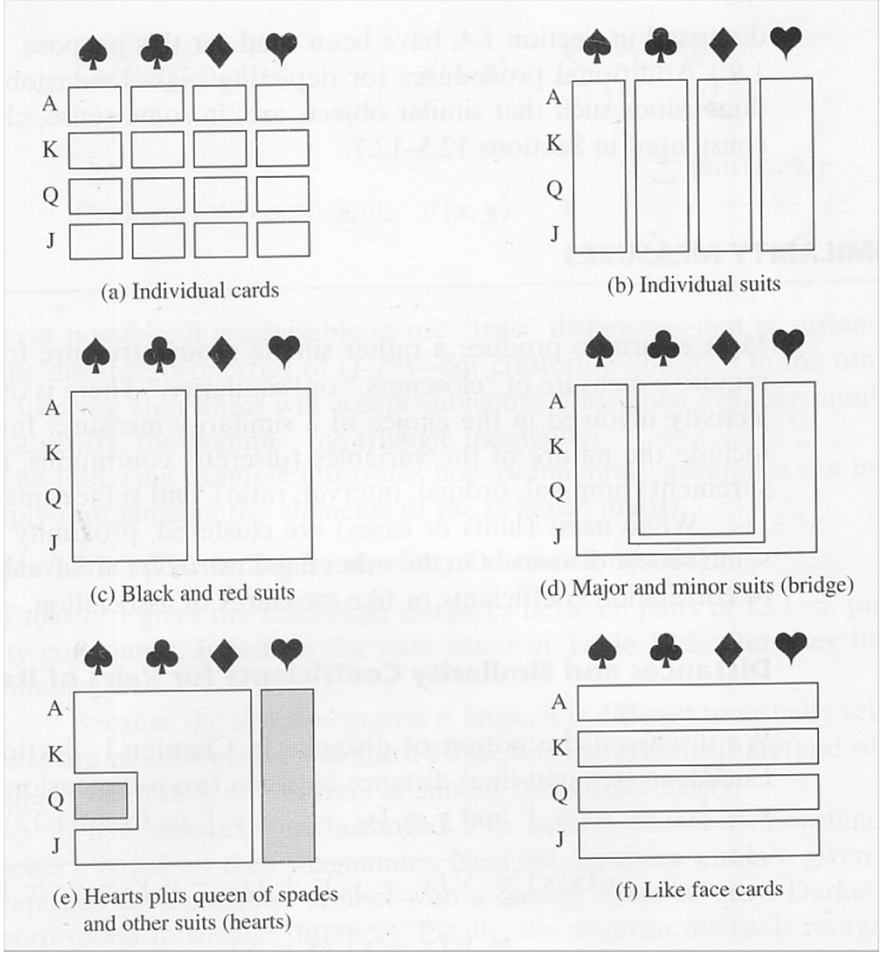
- measure of similarity
- algorithm for using the measure of similarity to discover natural groups in the data

The number of ways to divide n items into k clusters: $k^n/k!$

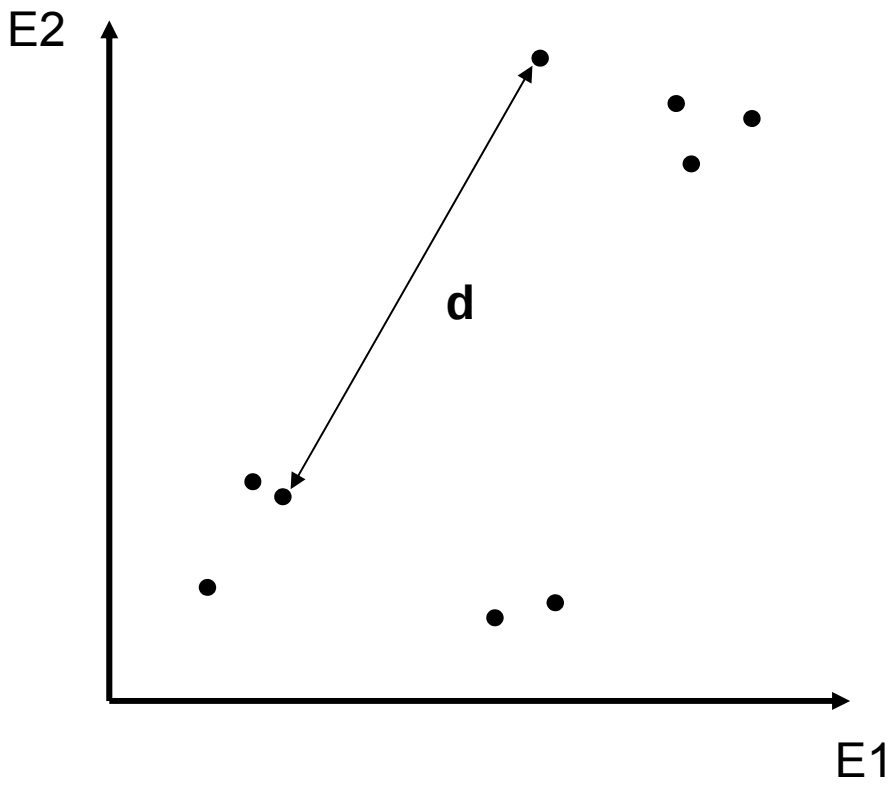
Example: $10^{500}/10! = 2.756 \times 10^{493}$

Measure of similarity

What is similar?

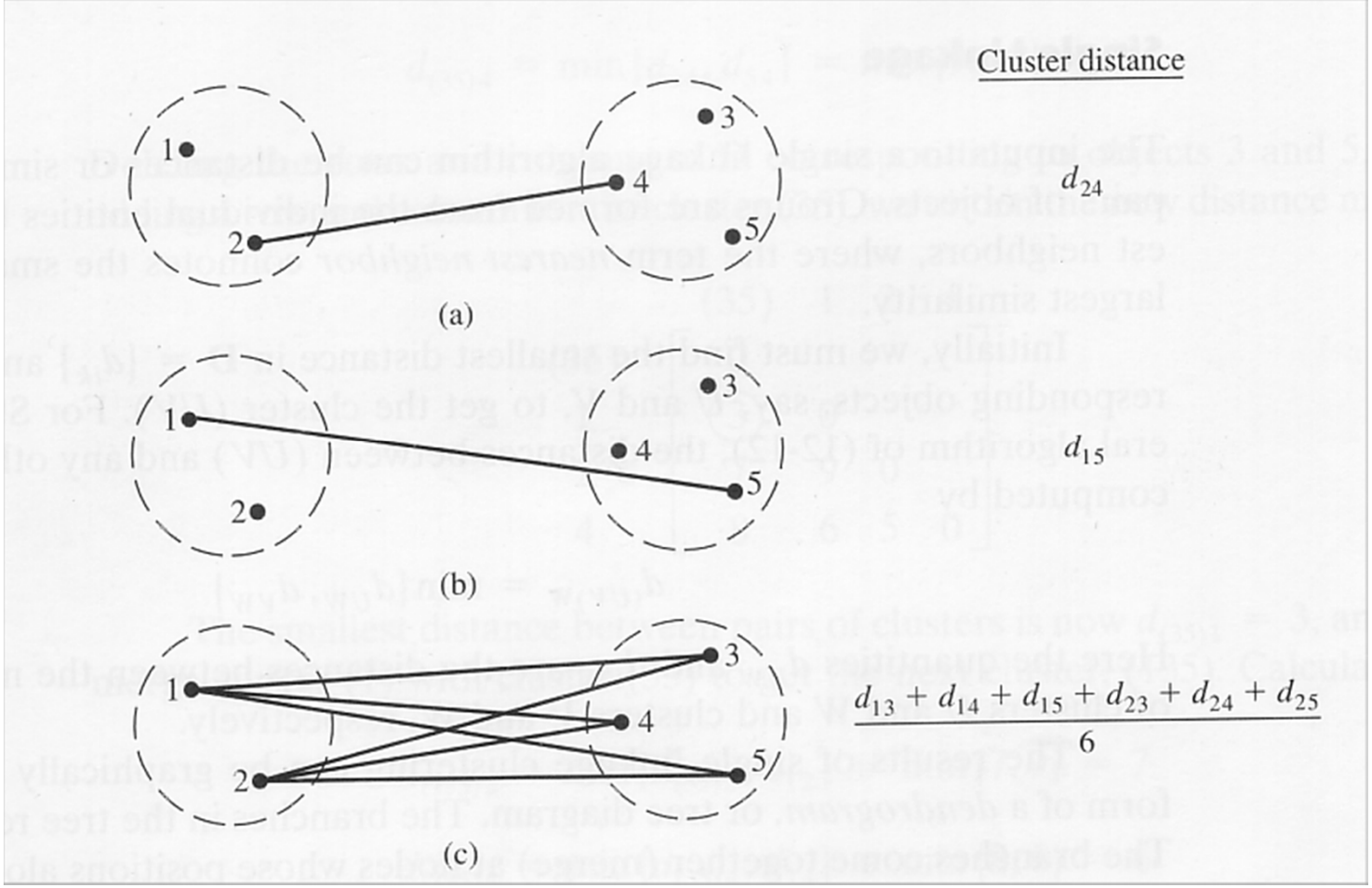


Euclidean distance



Hierarchical clustering

Inter-cluster similarity measures: (a) single linkage, (b) complete linkage and (c) average linkage



Example of hierarchical clustering: languages of Europe

TABLE 12.3 NUMERALS IN 11 LANGUAGES

English (E)	Norwegian (N)	Danish (Da)	Dutch (Du)	German (G)	French (Fr)	Spanish (Sp)	Italian (I)	Polish (P)	Hungarian (H)	Finnish (Fi)
one	en	en	een	eins	un	uno	uno	jeden	egy	yksi
two	to	to	twee	zwei	deux	dos	due	dwa	ketto	kaksi
three	tre	tre	drie	drei	trois	tres	tre	trzy	harom	kolme
four	fire	fire	vier	vier	quatre	cuatro	quattro	cztery	negy	neua
five	fem	fem	vijf	funf	cinq	cinco	cinque	piec	ot	viisi
six	seks	seks	zes	sechs	six	seis	sei	szesc	hat	kuusi
seven	sju	syv	zeven	sieben	sept	siete	sette	siedem	het	seitseman
eight	atte	otte	acht	acht	huit	ocho	otto	osiem	nyolc	kahdeksan
nine	ni	ni	negen	neun	neuf	nueve	nove	dziewiec	kilenc	yhdeksan
ten	ti	ti	tien	zehn	dix	diez	dieci	dziesiec	tiz	kymmenen

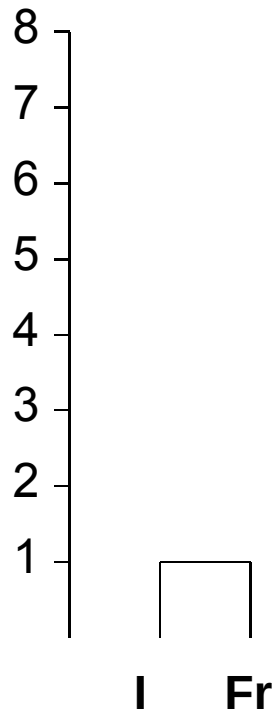
Distance: Frequency of numbers with different first letter e.g.

$d_{EN} = 2 \quad d_{EDu} = 7 \quad d_{SpI} = 1$

Inter-cluster strategy: SINGEL LINKAGE

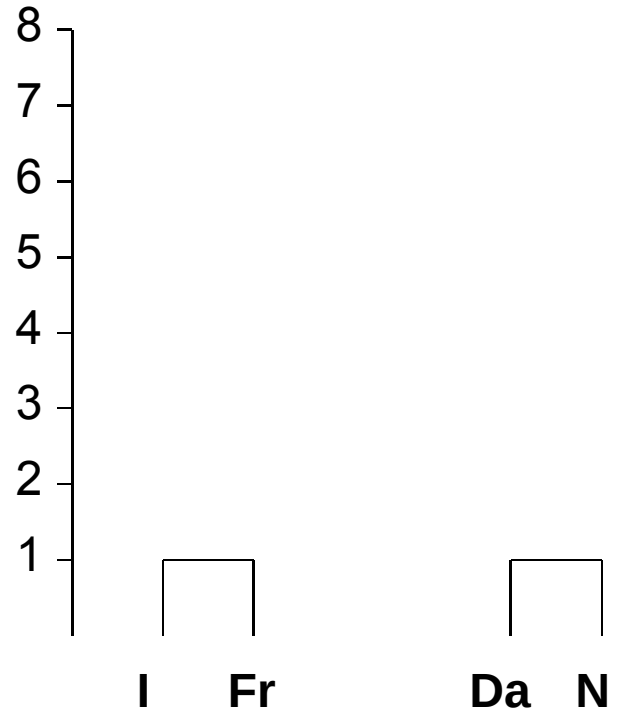
Iteration 1

	E	N	Da	Du	G	Fr	Sp	I	P	H	Fi
E	0										
N	2	0									
Da	2	1	0								
Du	7	5	6	0							
G	6	4	5	5	0						
Fr	6	6	6	9	7	0					
Sp	6	6	5	9	7	2	0				
I	6	6	5	9	7	1	1	0			
P	7	7	6	10	8	5	3	4	0		
H	9	8	8	8	9	10	10	10	10	0	
Fi	9	9	9	9	9	9	9	9	9	8	0



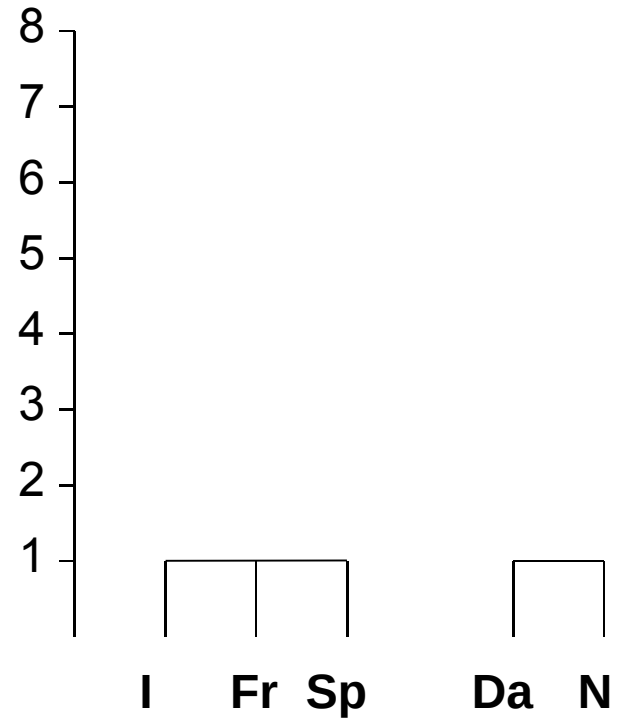
Iteration 2

	I Fr	E	N	Da	Du	G	Sp	P	H	Fi
I Fr	0									
E	6	0								
N	6	2	0							
Da	5	2	1	0						
Du	9	7	5	6	0					
G	7	6	4	5	5	0				
Sp	1	6	6	5	9	7	0			
P	4	7	7	6	10	8	3	0		
H	10	9	8	8	8	9	10	10	0	
Fi	9	9	9	9	9	9	9	9	8	0



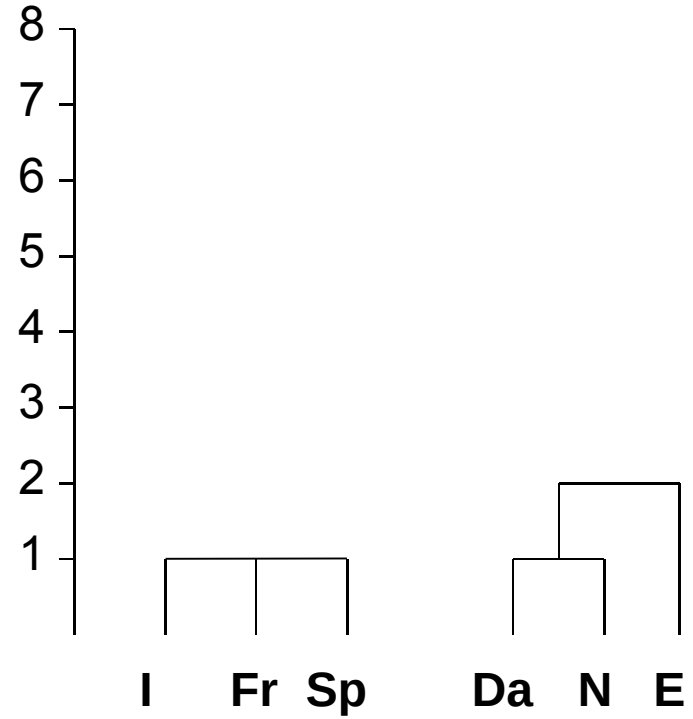
Iteration 3

	Da N	I Fr	E	Du	G	Sp	P	H	Fi
Da N	0								
I Fr	5	0							
E	2	6	0						
Du	5	9	7	0					
G	4	7	6	5	0				
Sp	5	1	6	9	7	0			
P	6	4	7	10	8	3	0		
H	8	10	9	8	9	10	10	0	
Fi	9	9	9	9	9	9	9	8	0



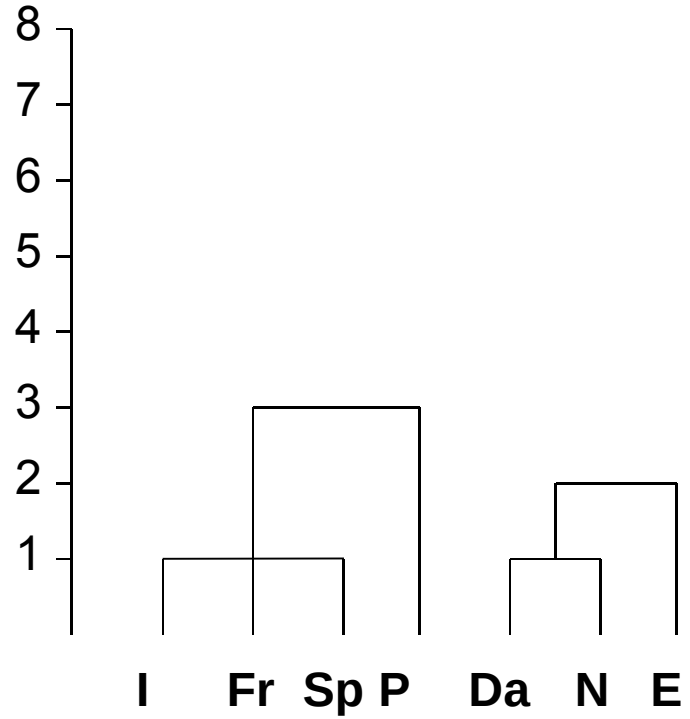
Iteration 4

	Sp I Fr	Da N	E	Du	G	P	H	Fi
Sp I Fr	0							
Da N	5	0						
E	6	2	0					
Du	9	5	7	0				
G	7	4	6	5	0			
P	3	6	7	10	8	0		
H	10	8	9	8	9	10	0	
Fi	9	9	9	9	9	9	8	0



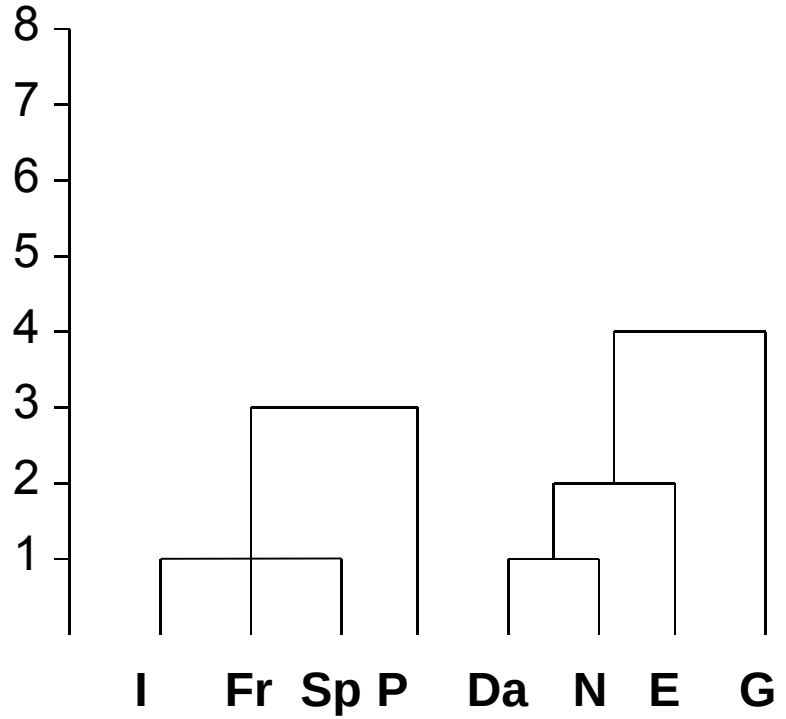
Iteration 5

	E Da N	Sp I Fr	Du	G	P	H	Fi
E Da N	0						
Sp I Fr	5	0					
Du	5	9	0				
G	4	7	5	0			
P	6	3	10	8	0		
H	8	10	8	9	10	0	
Fi	9	9	9	9	9	8	0



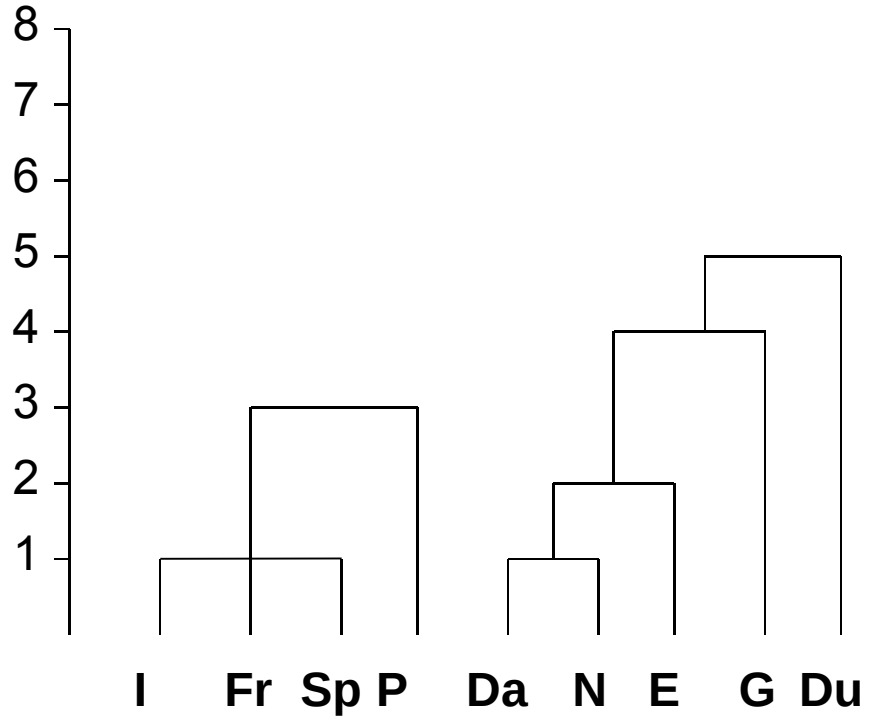
Iteration 6

	P Sp I Fr	E Da N	Du	G	H	Fi
P Sp I Fr	0					
E Da N	5	0				
Du	9	5	0			
G	7	4	5	0		
H	10	8	8	9	0	
Fi	9	9	9	9	8	0



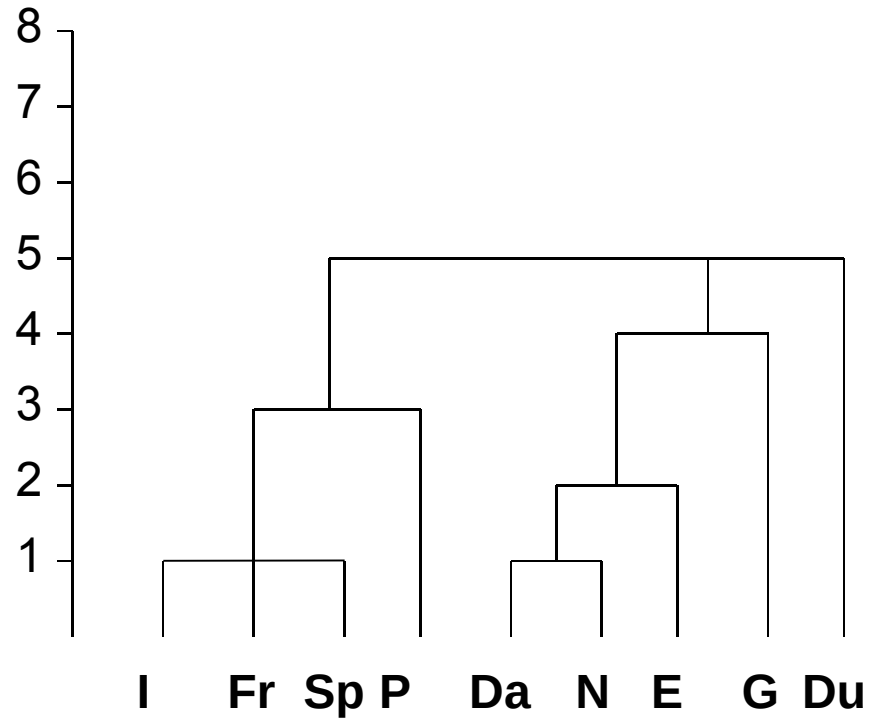
Iteration 7

	G E Da N	P Sp I Fr	Du	H	Fi
G E Da N	0				
P Sp I Fr	5	0			
Du	5	9	0		
H	8	10	8	0	
Fi	9	9	9	8	0



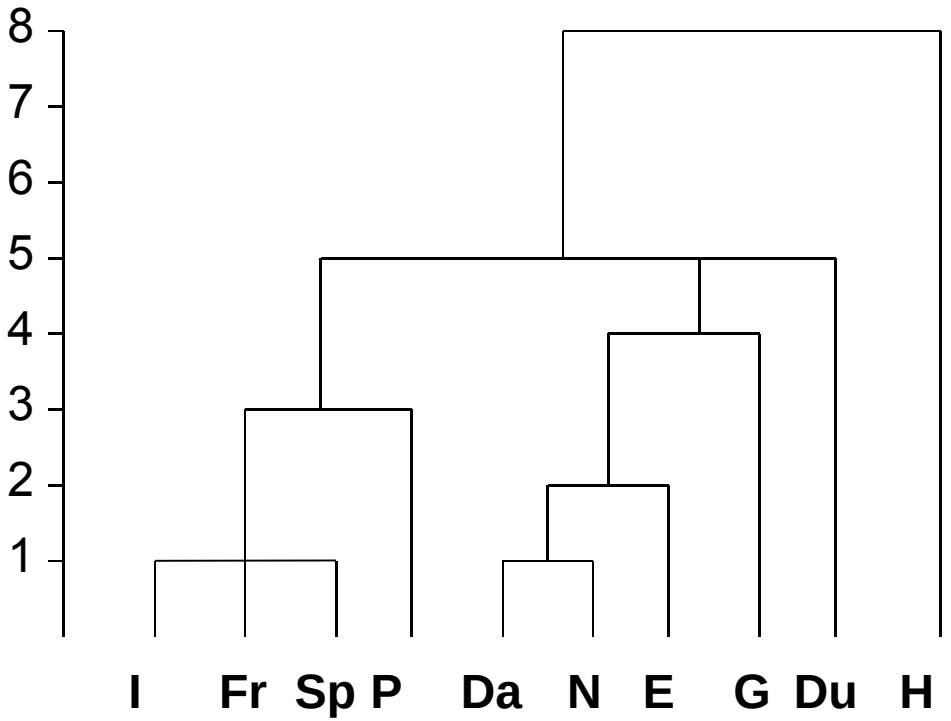
Iteration 8

	Du G E Da N	P Sp I Fr	H	Fi
Du G E Da N	0			
P Sp I Fr	5	0		
H	8	10	0	
Fi	9	9	8	0



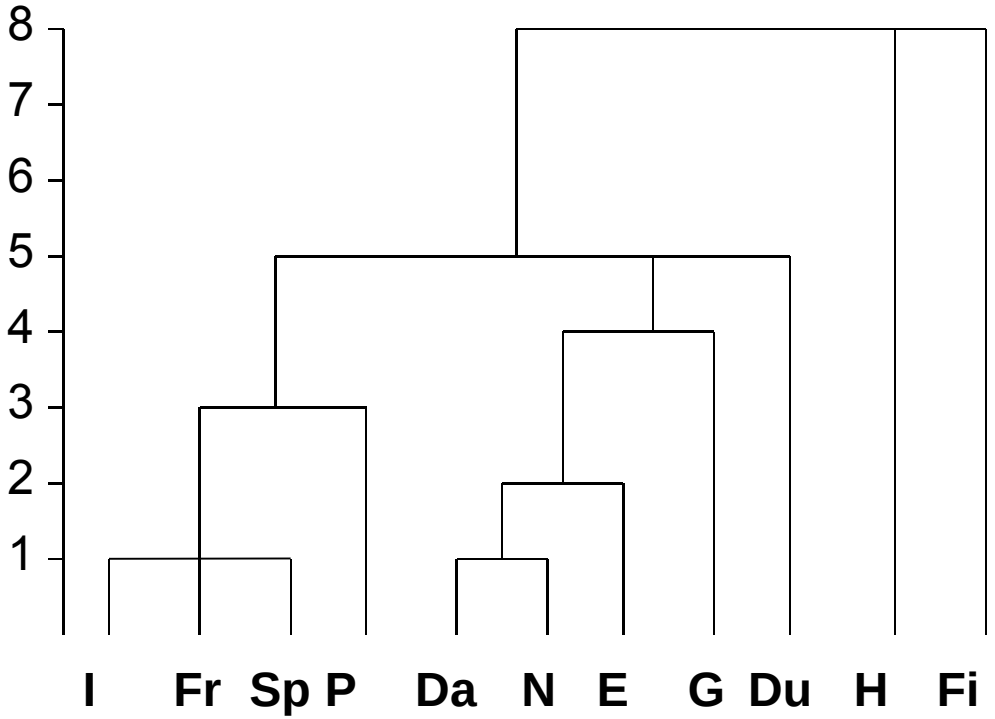
Iteration 9

	P Sp I Fr Du G E Da N	H	Fi
P Sp I Fr Du G E Da N	0		
H	8	0	
Fi	9	8	0



Iteration 10

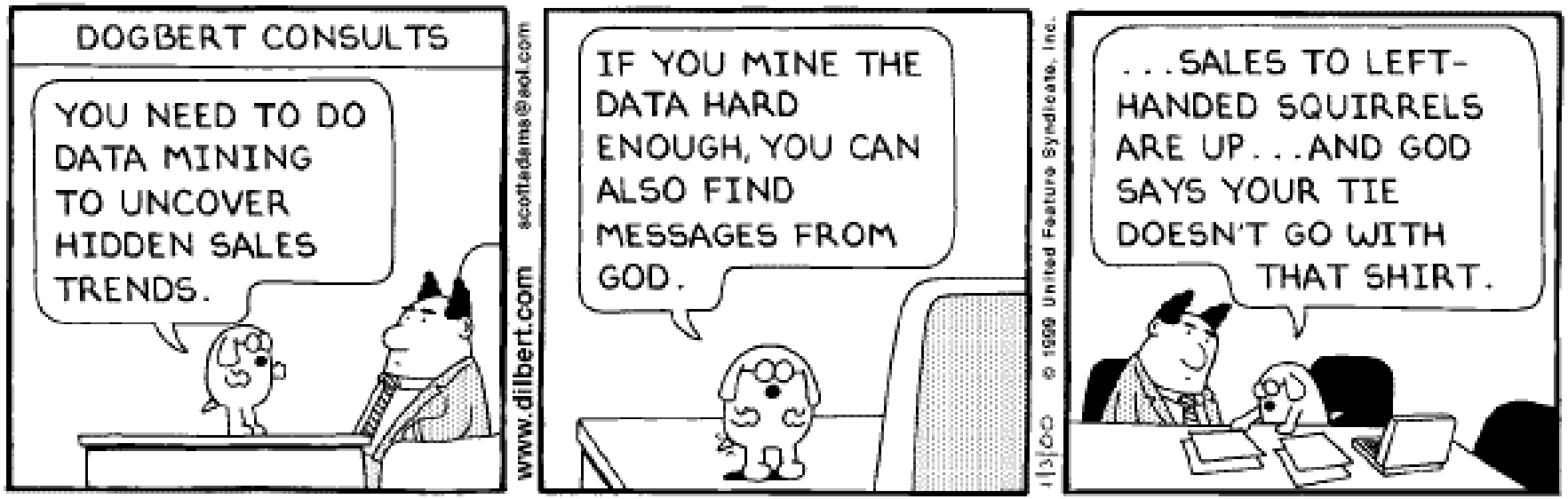
	Fi H	P Sp I Fr Du G E Da N
Fi H	0	
P Sp I Fr Du G E Da N	8	0



Hierarchical clustering: properties

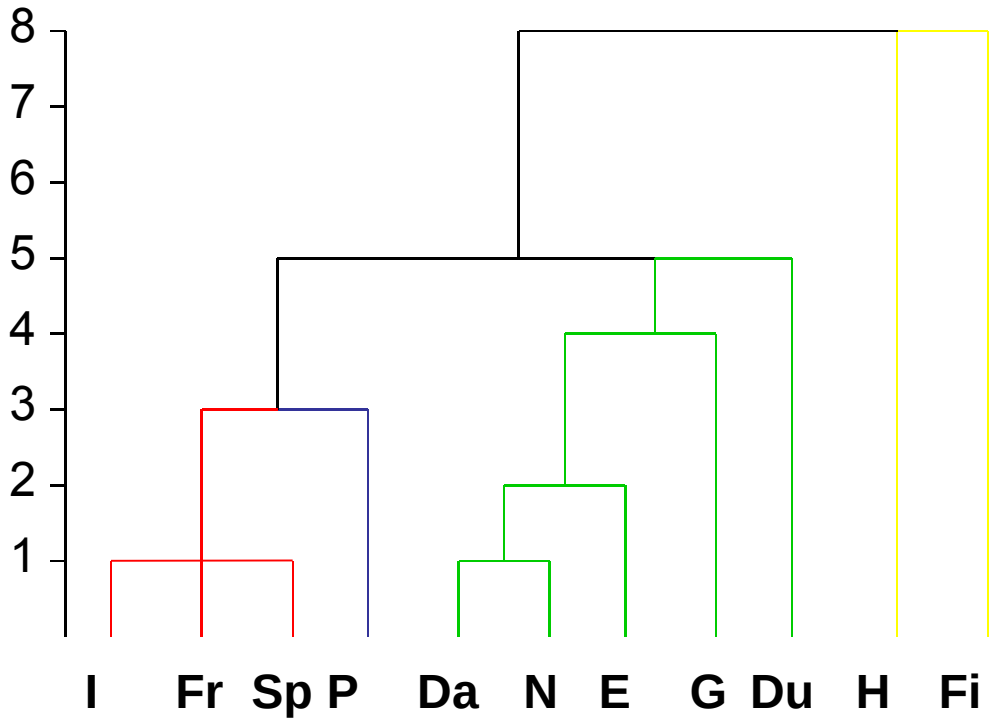
- Huge memory requirements: stores the $n \times n$ matrix
- Running time: $O(n^3)$
- Deterministic: produces the same clustering each time
- Nice visualization: dendrogram
- Number of clusters can be selected using the dendrogram

Any data mining result needs to be consistent
BOTH with the data and current knowledge!



Evaluation of clusters

Clusters may be evaluated according to how well they describe current knowledge



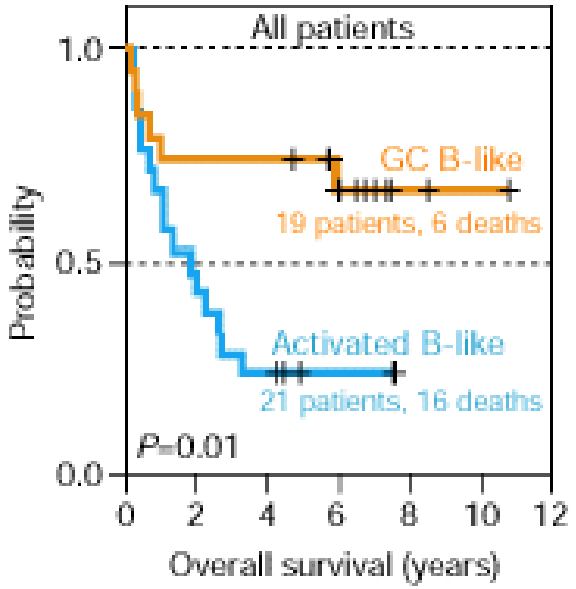
Roman
Slavic
Germanic
Ugro-Finnish

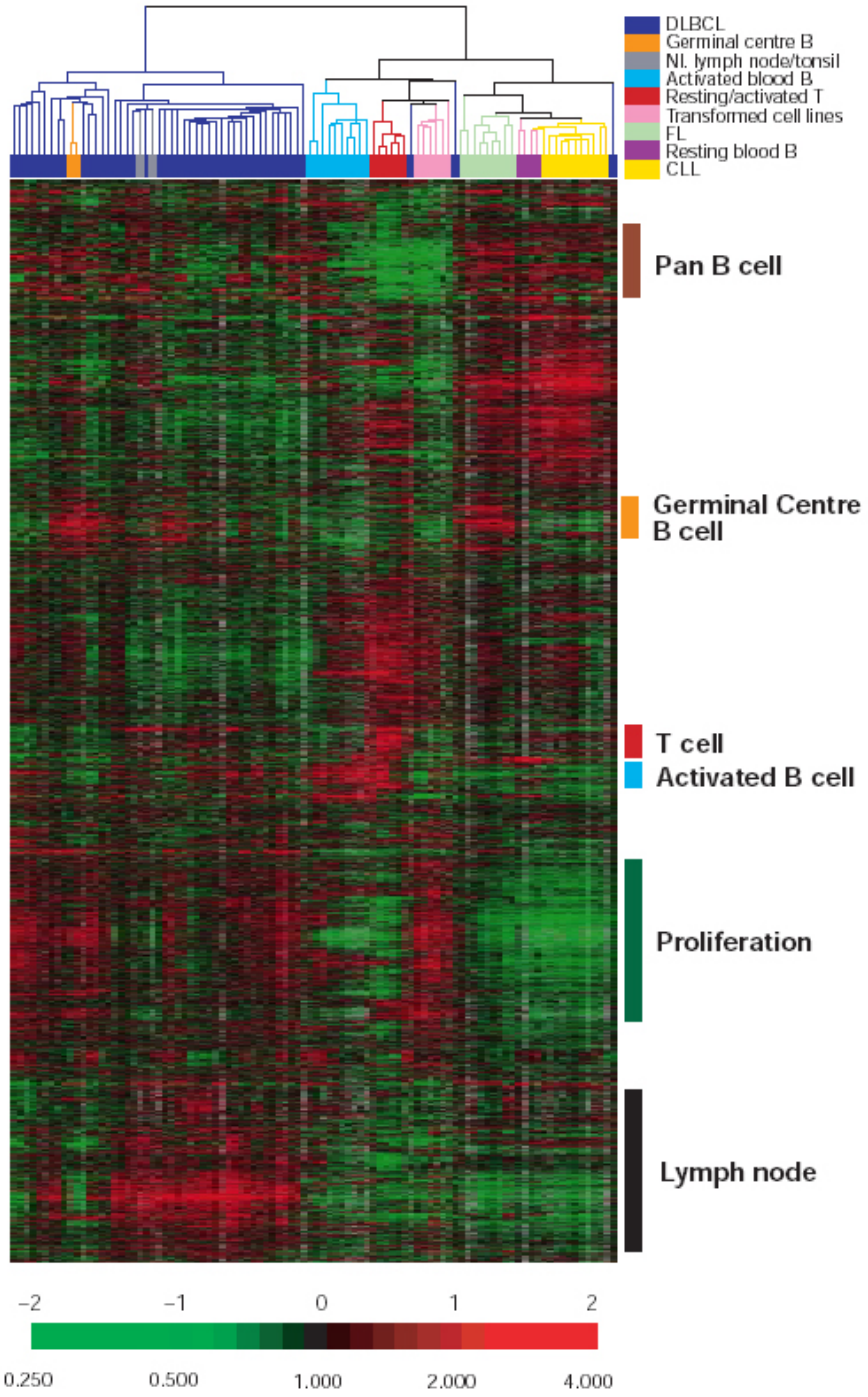
Example: Hierarchical clustering

96 normal and malignant lymphocyte samples

Almost 20 000 cDNA clones

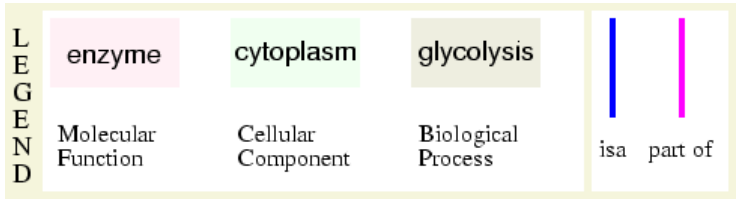
Two sub-clusters of DLBCL were shown to include patients with significantly different expected survival time!

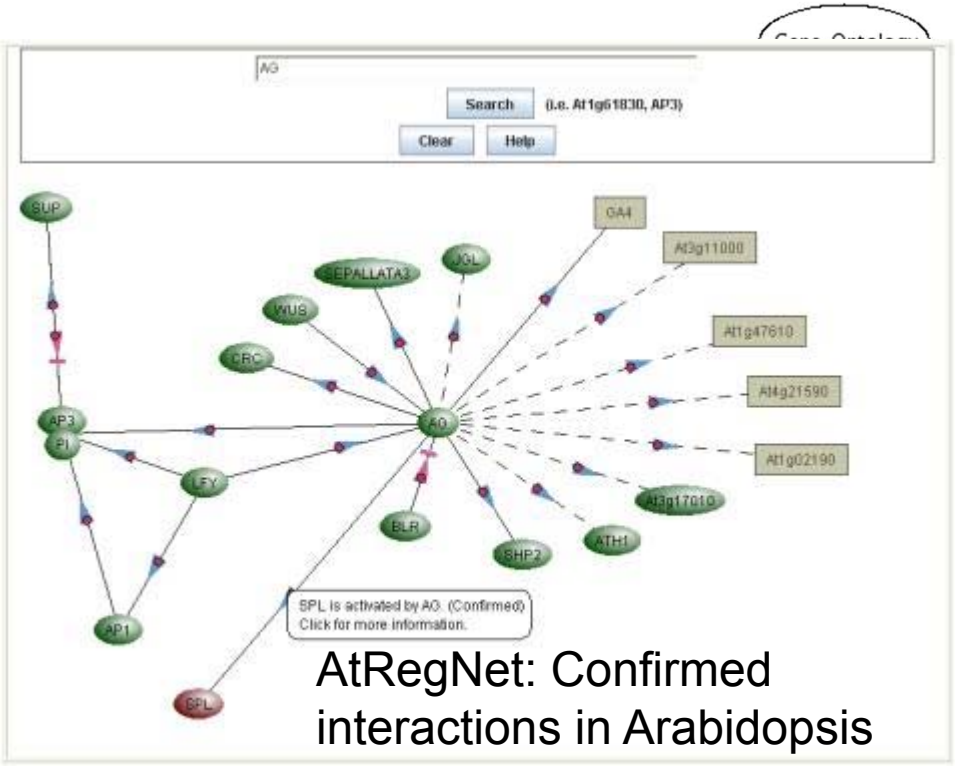




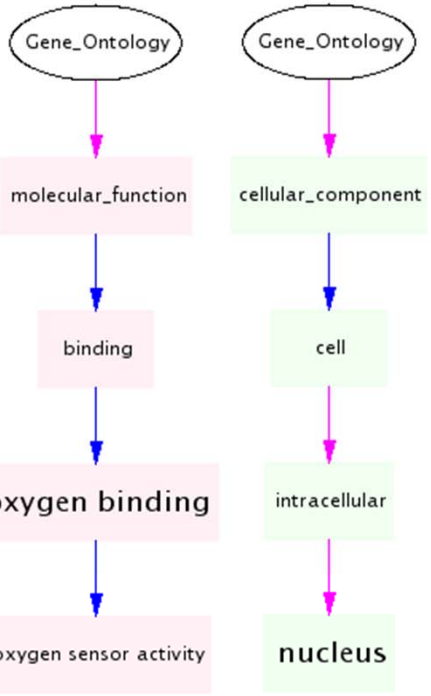
Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503-511, 2000.

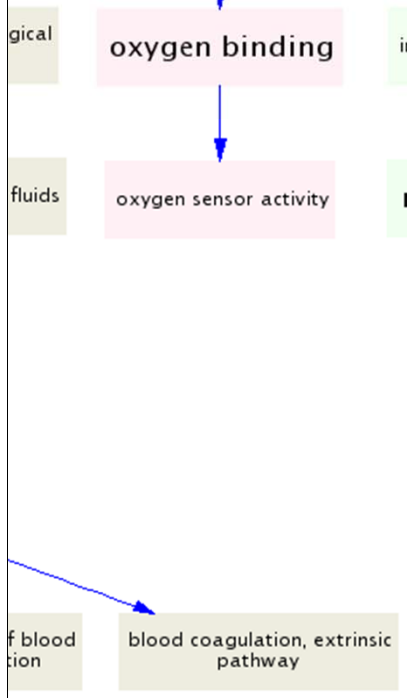
Existing knowledge





AtRegNet: Confirmed interactions in Arabidopsis





Randomization

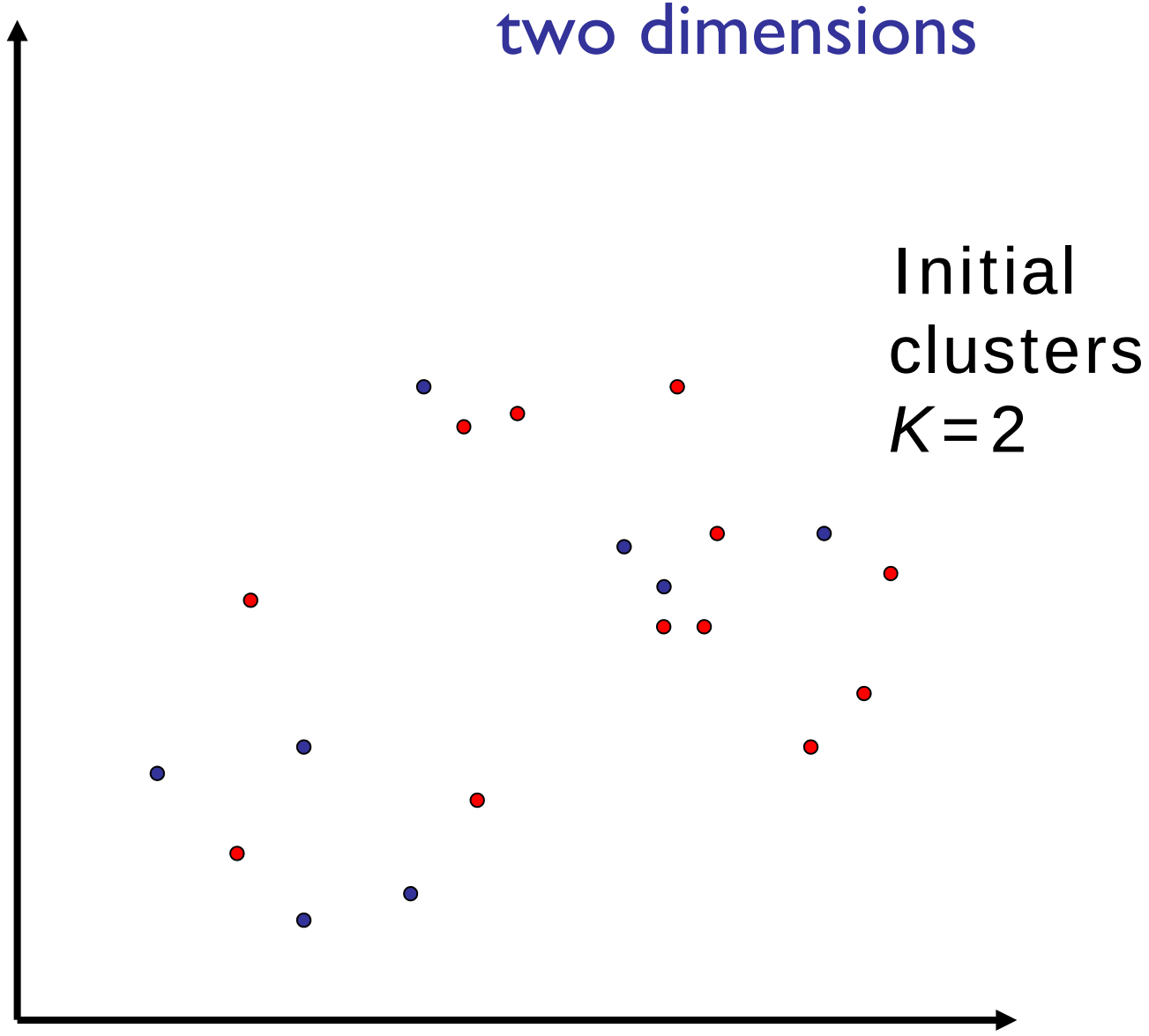
- Randomize the input data
- Test whether model properties from real data differs from those obtained using randomized data
- P-values: fraction of randomized datasets resulting in ”better” models than the real data

- Examples:
  - ✓ To what degree the model explains existing knowledge (e.g. gene function)
  - ✓ To what degree the model predicts observed data (cross validation)

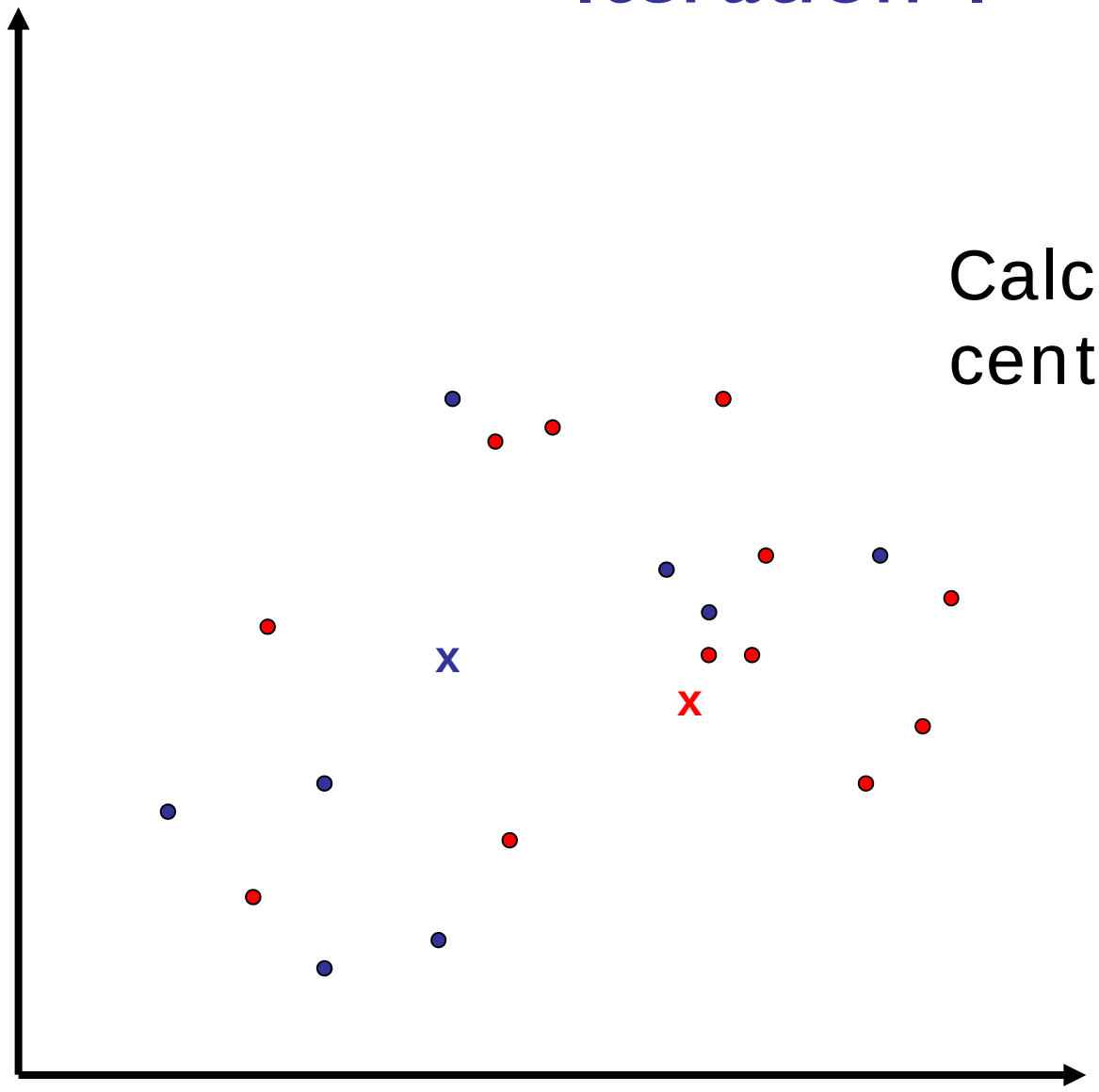
K-means clustering

- Split the data into *k* random clusters
- Repeat
  - calculate the centroid of each cluster
  - (re-)assign each gene/experiment to the closest centroid
  - stop if no new assignments are made

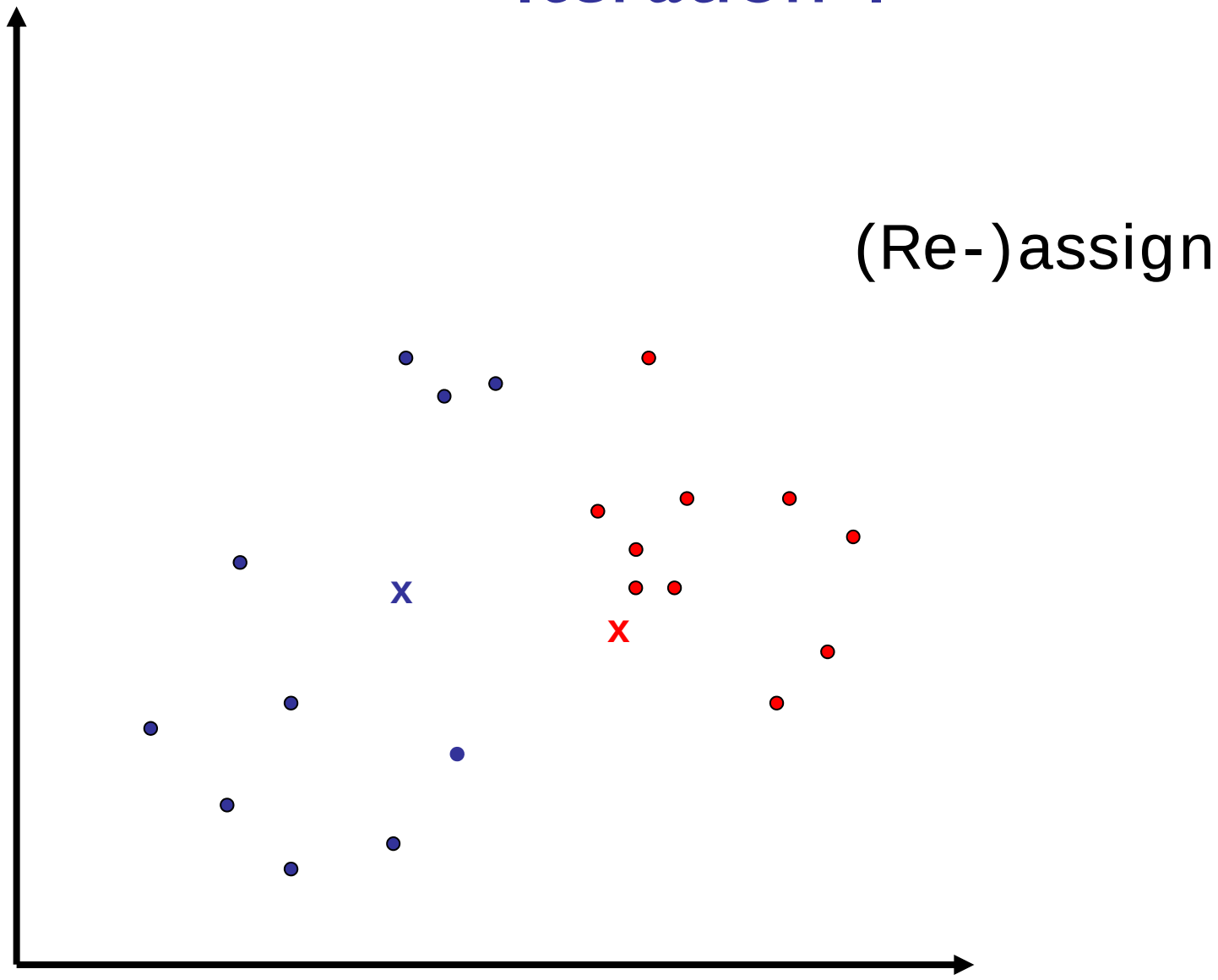
Example of K-means: two dimensions



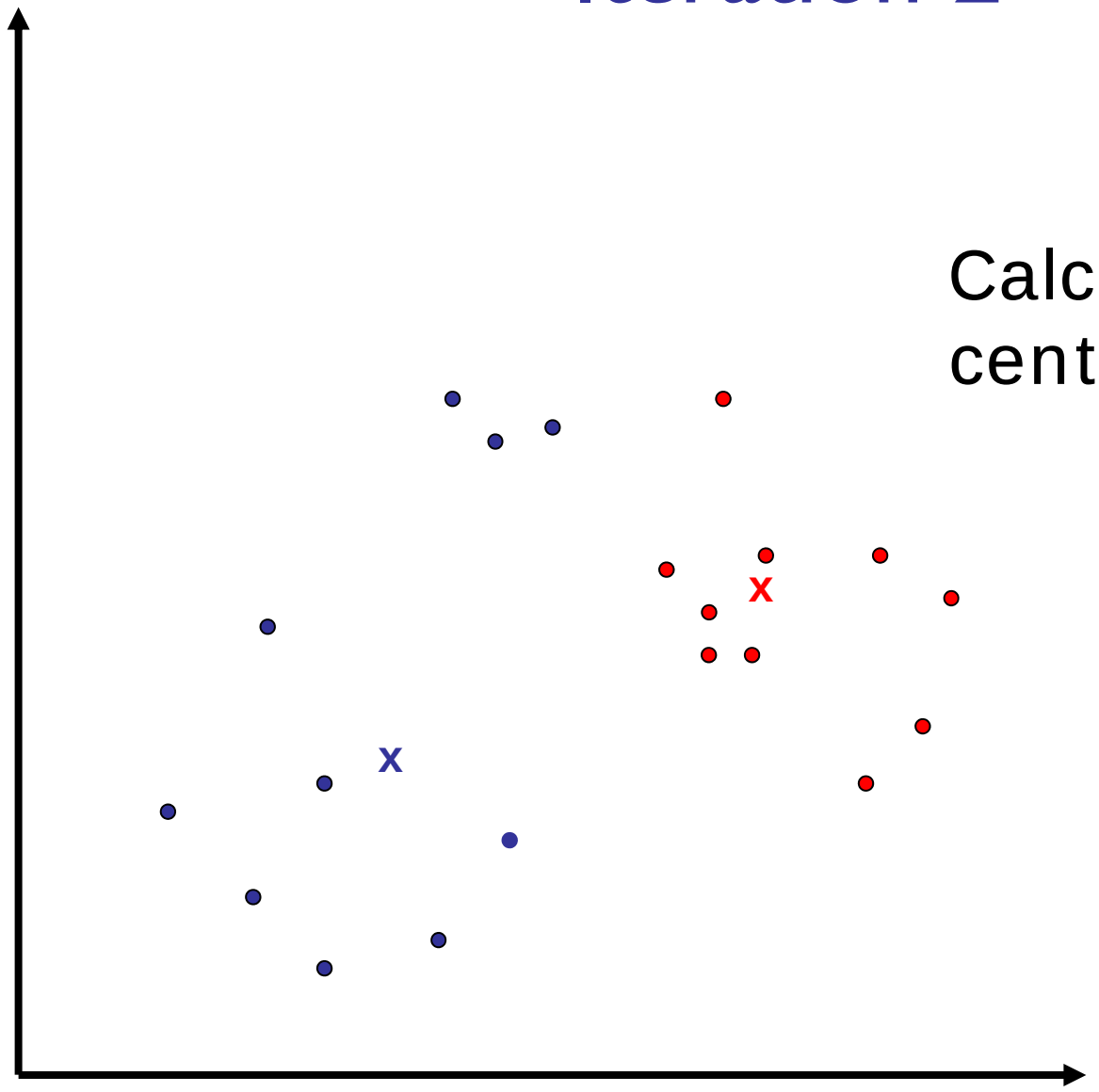
Iteration 1



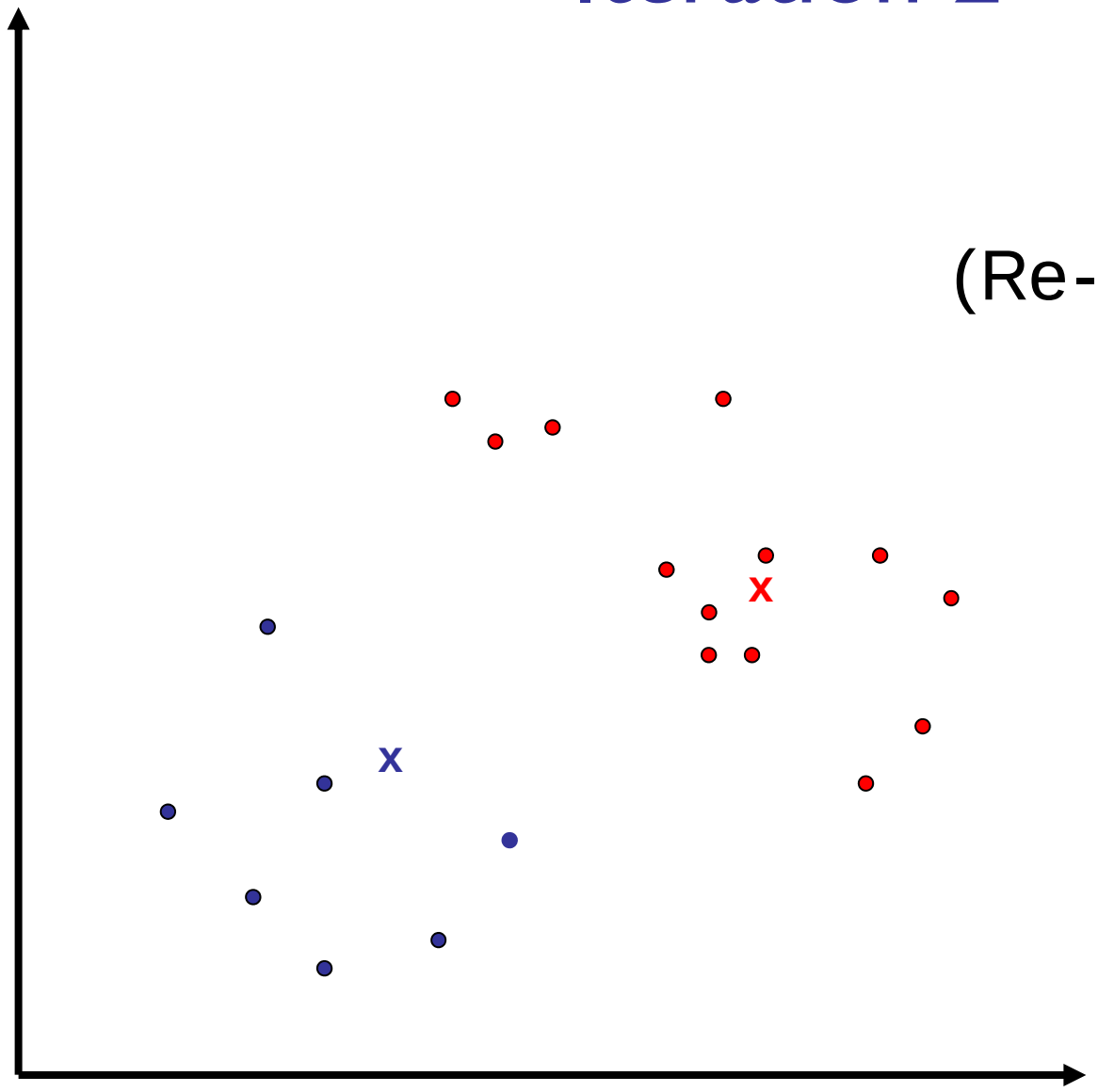
Iteration 1



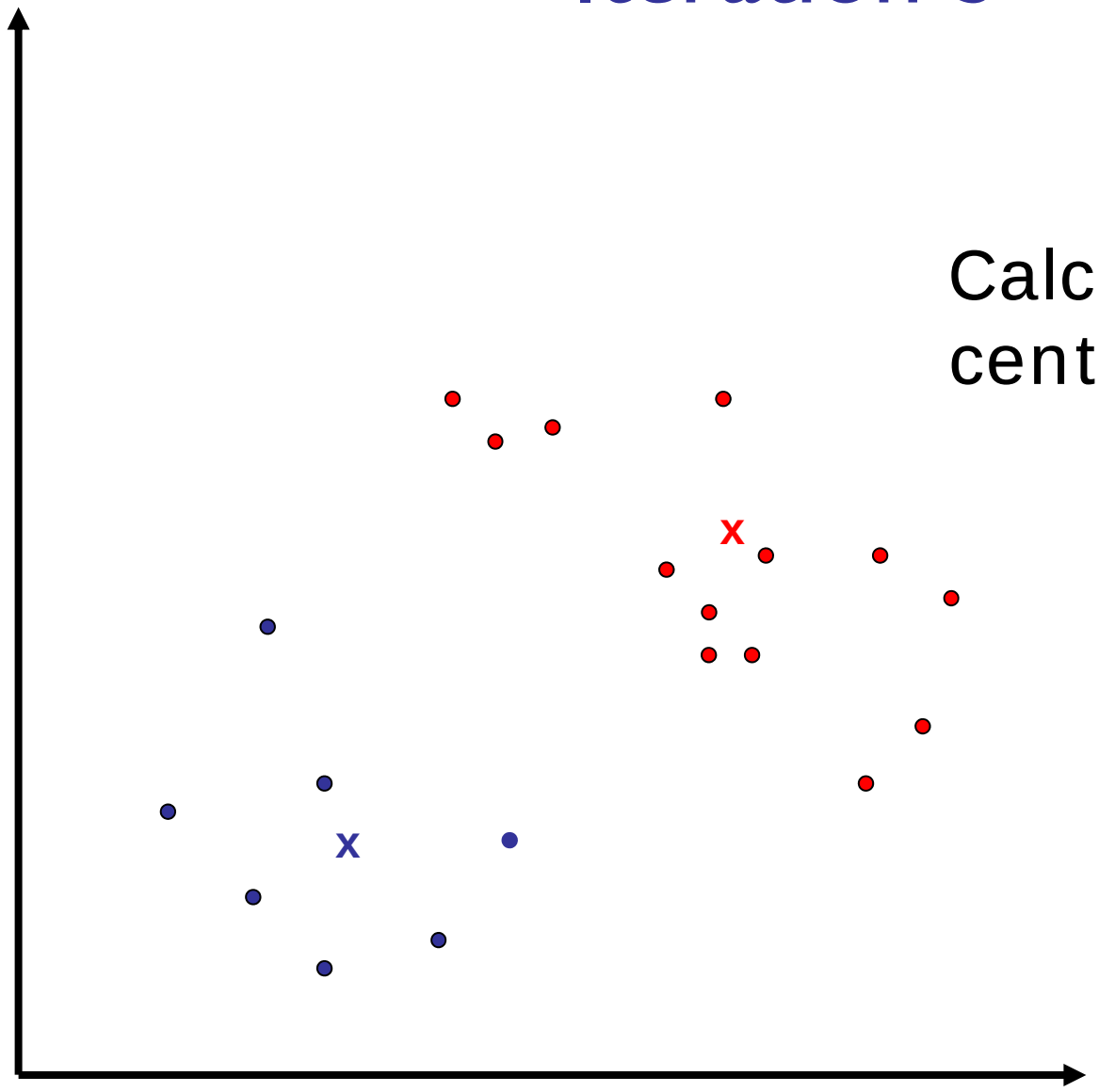
Iteration 2



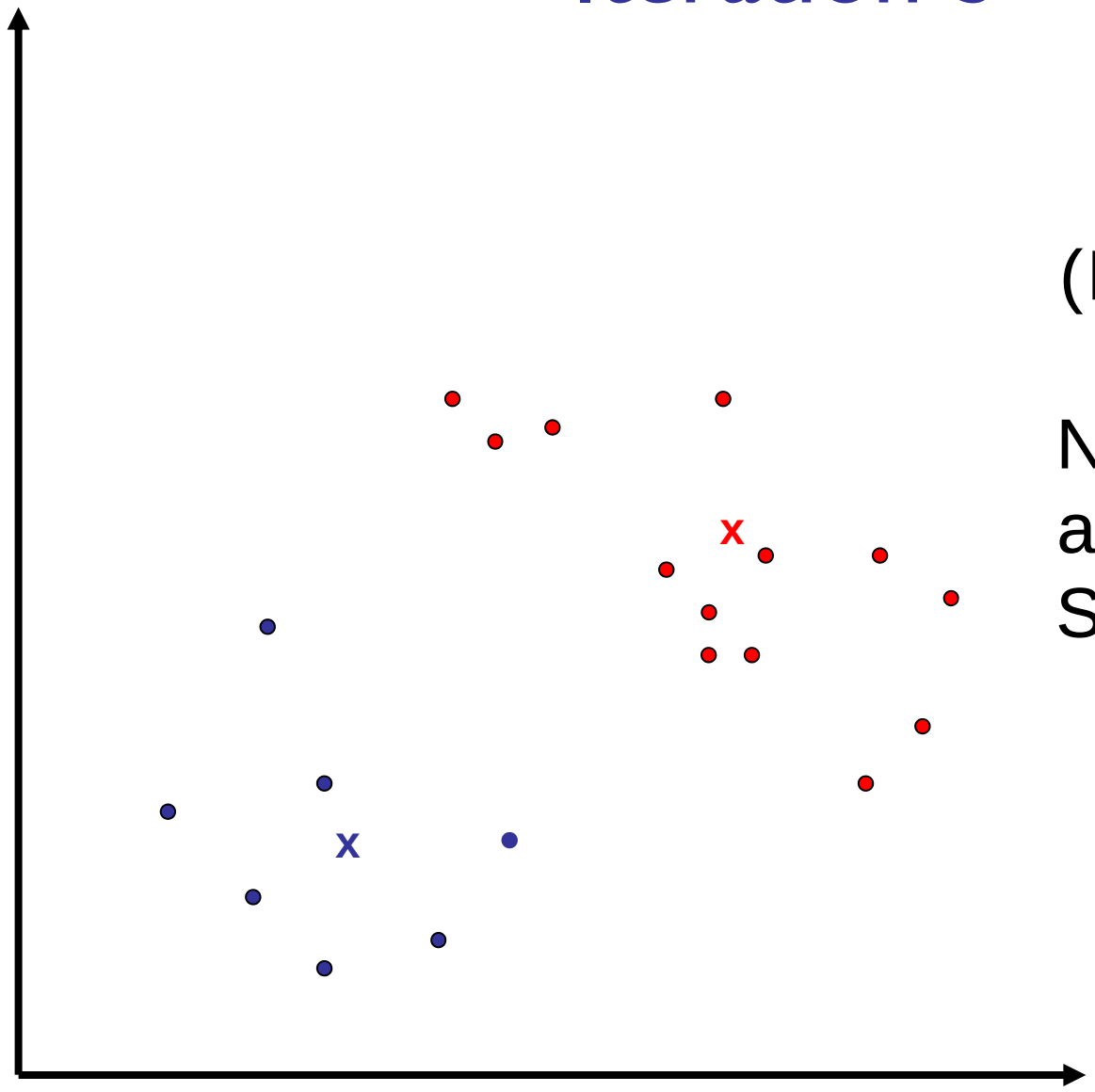
Iteration 2



Iteration 3



Iteration 3



(Re-)assign

No new assignments!
STOP

K-means: properties

- Low memory usage
- Running time: $O(kn)$, where k is the number of iterations
- Improves iteratively: not trapped in previous mistakes
- Non-deterministic: will in general produce different clusters with different initializations
- Number of clusters must be decided in advance

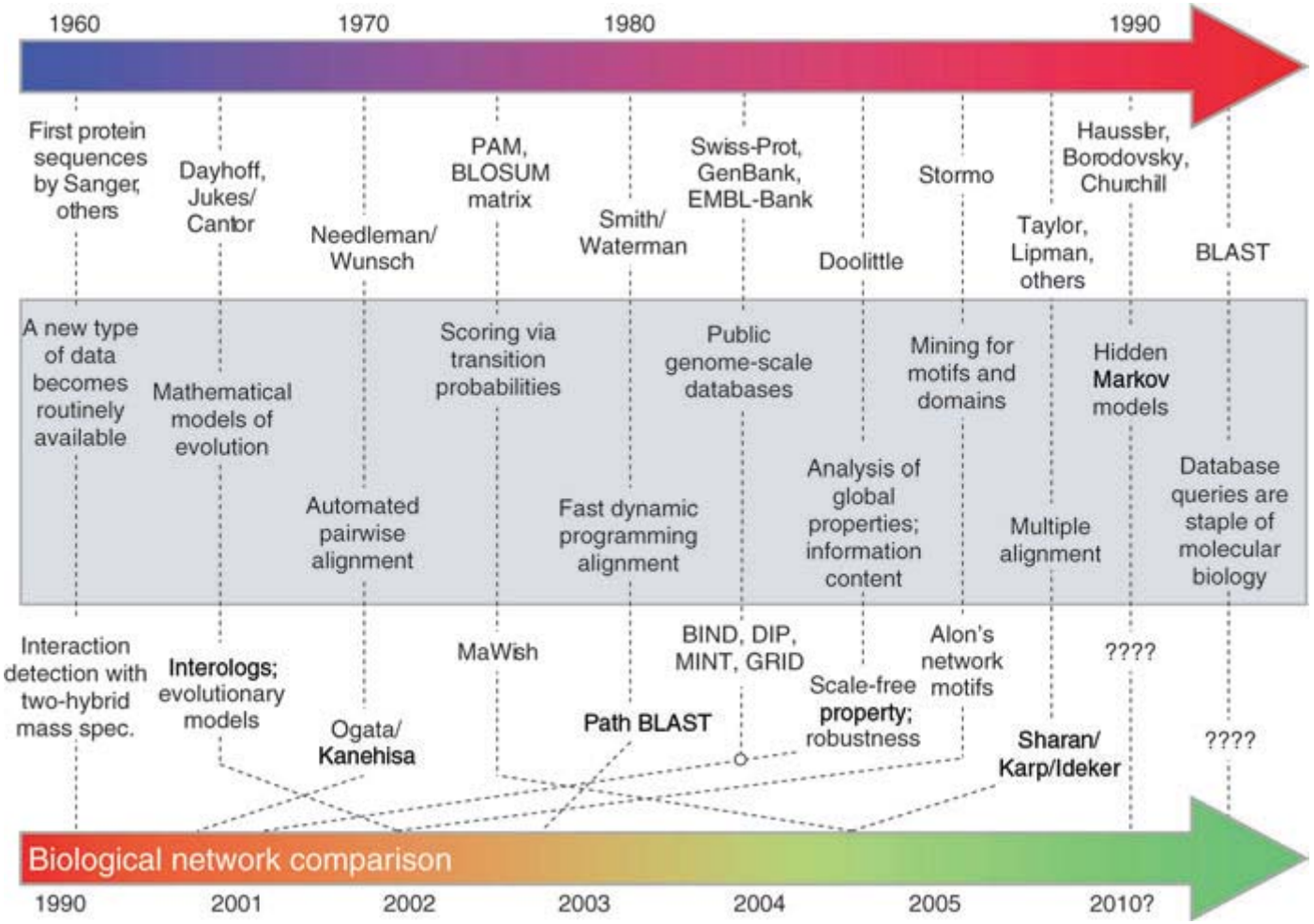
Hierarchical vs. k-means

- Hierarchical clustering:
  - computationally expensive -> relatively small data sets
  - nice visualization, no. of clusters can be selected
  - deterministic
  - cannot correct early ”mistakes” (greedy alg.)
- K-means:
  - computationally efficient -> large data sets
  - predefined no. of clusters
  - non-deterministic -> should be run several times
  - iterative improvement (randomization alg.)

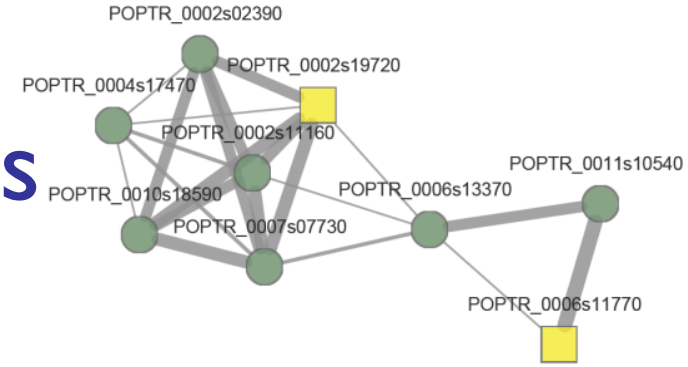
Network representations

- Network: nodes connected by edges
- Nodes represent genes, proteins, metabolites
- Edges represent relationships
  - Protein-protein networks: proteins form a functional complex
  - **Co-expression networks**: expression correlation
  - Gene networks: genes affect the expression of other genes
  - Regulatory network: transcription factors regulate genes by binding DNA motifs in the promoter region
- Network representations are flexible and allow integration of heterogeneous data

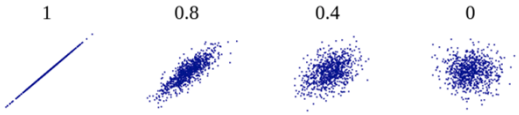
From sequence alignment to network alignment?



Co-expression measures



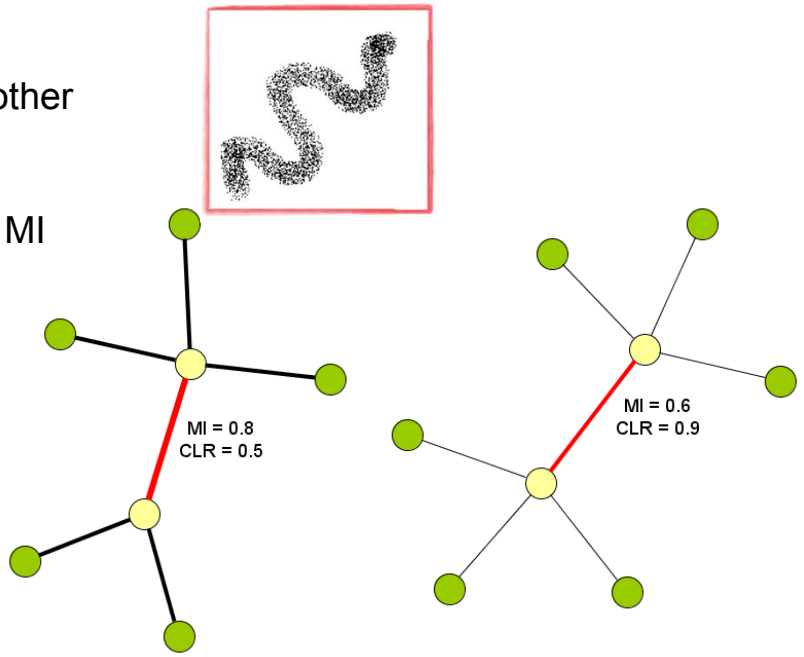
- **Pearson correlation:** measure linear dependency

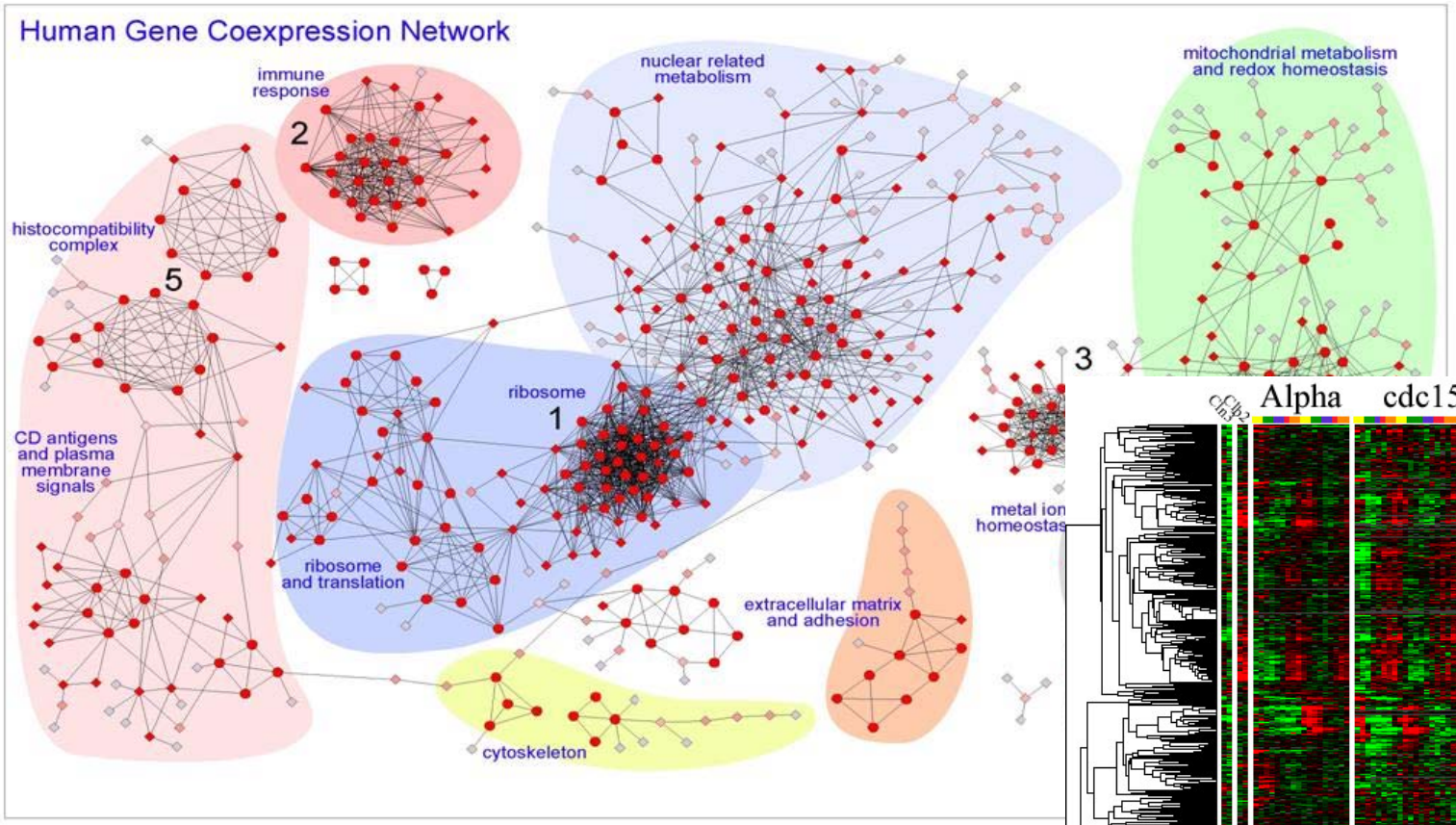


- **Spearman correlation**: measure monotonic trends, more robust to outliers

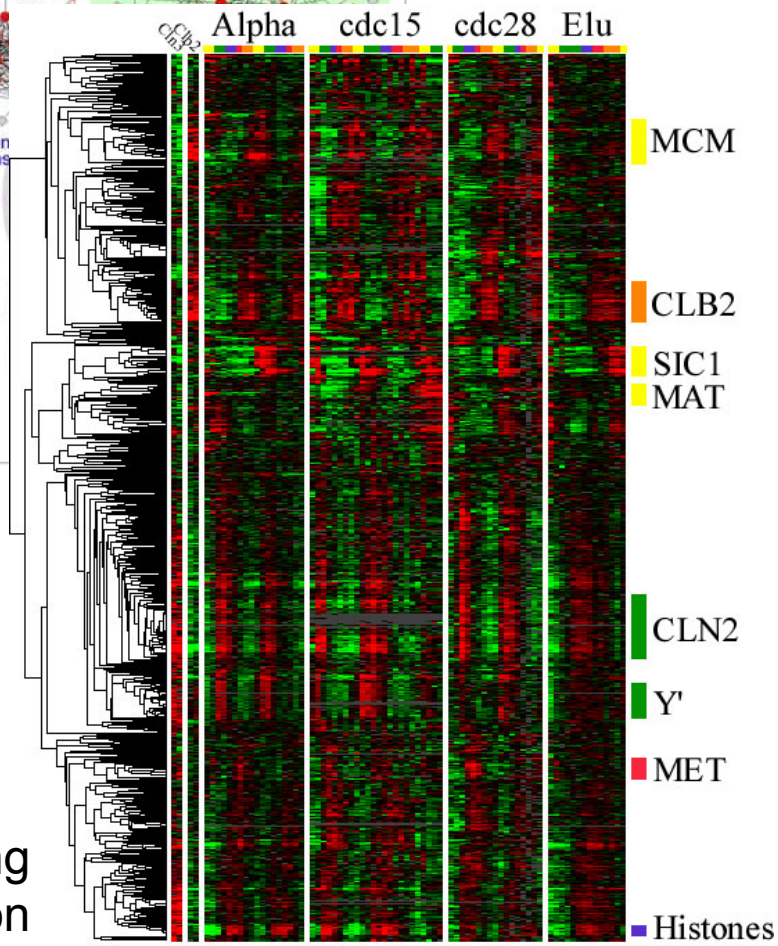
- **Mutual information (MI)**: measure non-monotonic and other more complex relationships

- **Context Likelihood of Relatedness (CLR)**: normalizes MI compared to the neighboring genes



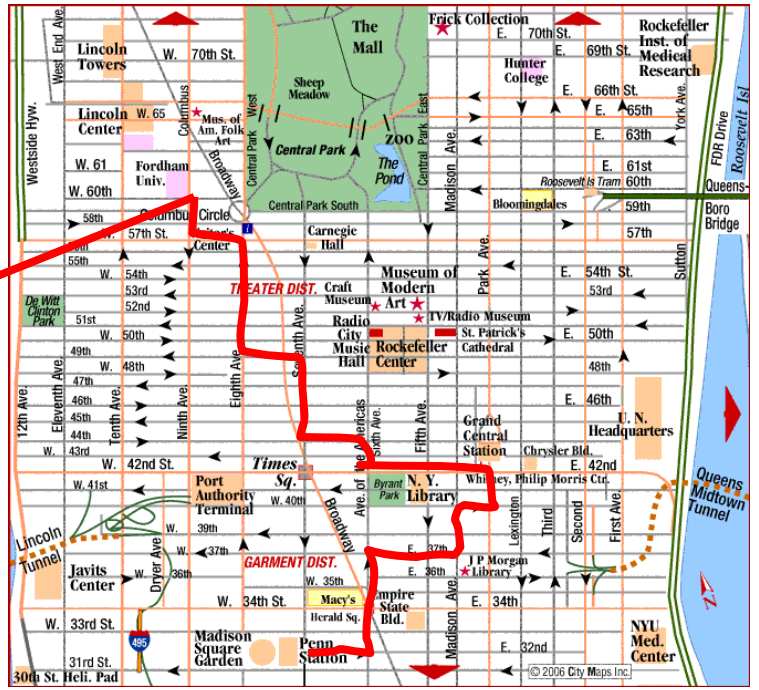


Co-expression network



Hierarchical clustering
of gene expression

Predicting “causality” from expression data: Analogous to establishing whether you are being followed by the car behind you

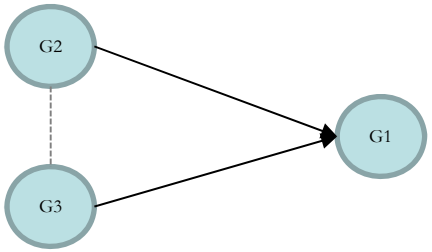


Using array data: with a fuggy rear-view mirror

Using RNA-Seq: with a clear rear-view mirror

Complexity of data analysis

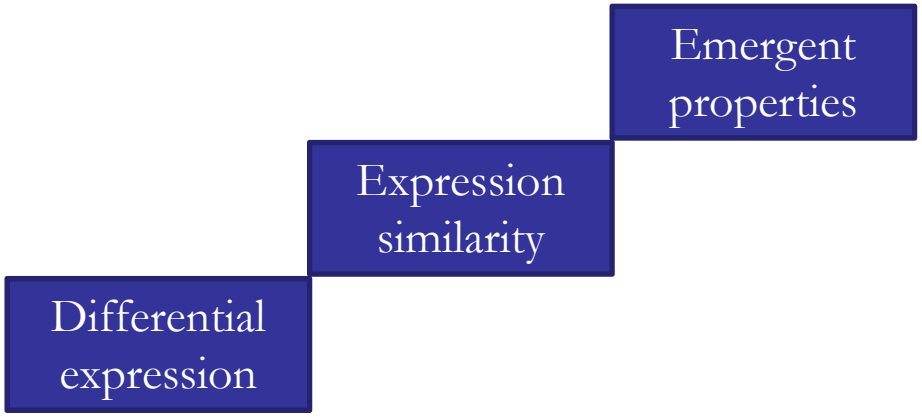
Gene network:



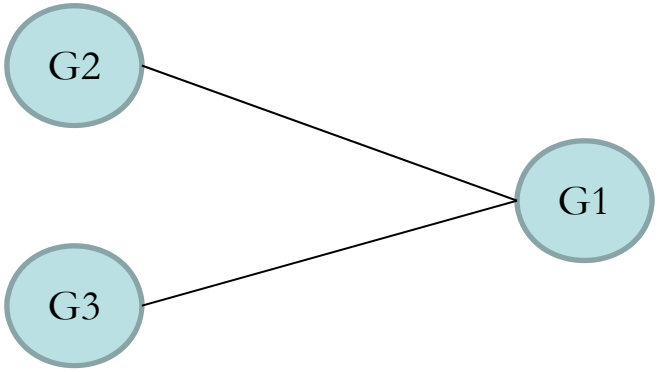
$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

Complexity

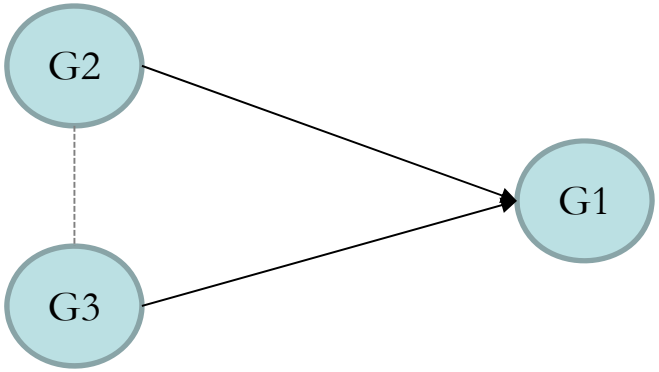
Detecting genes that best predict the expression of our gene of interest



Co-expression networks versus gene networks



Co-expression network:
Expression of G1 correlates with that of G3
Expression of G2 correlates with that of G3

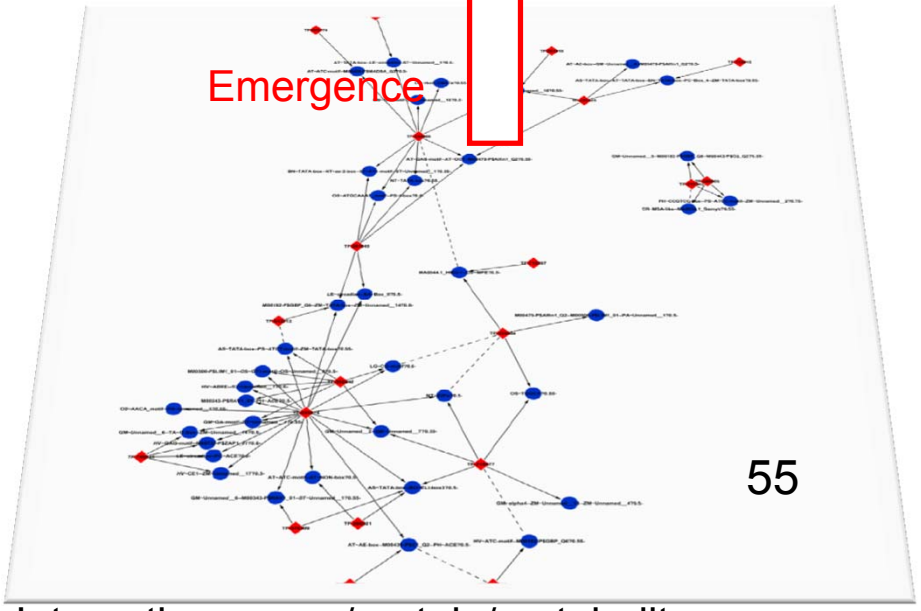


Gene network/regulatory networks:
G1 and G2 predict the expression of G3

Emergent properties

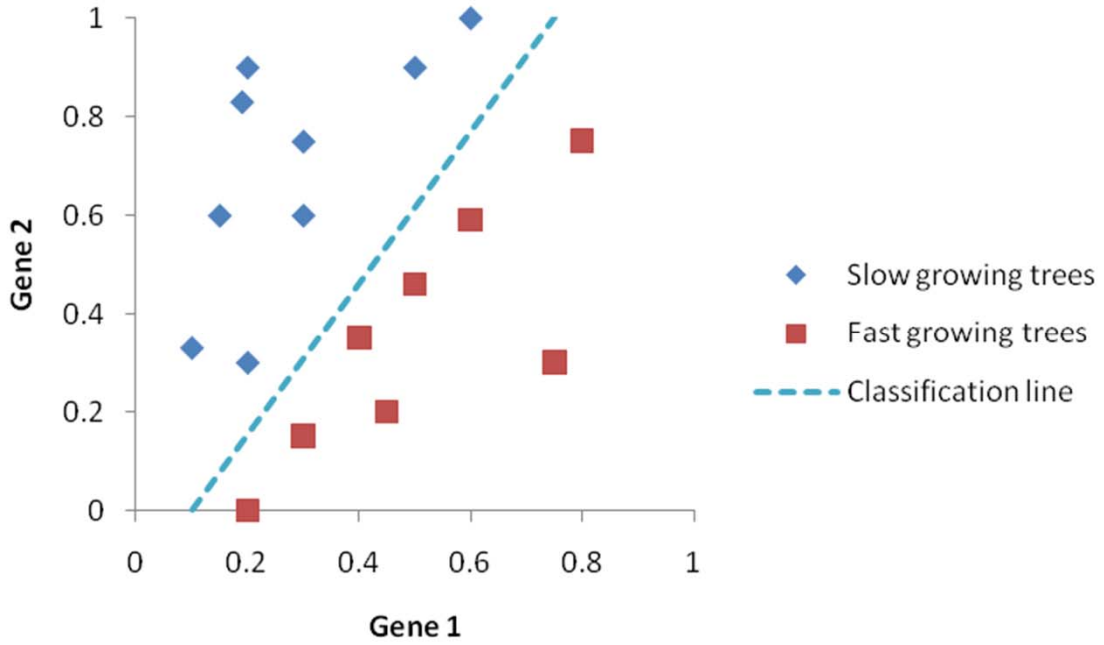
Phenotypes

Emergence

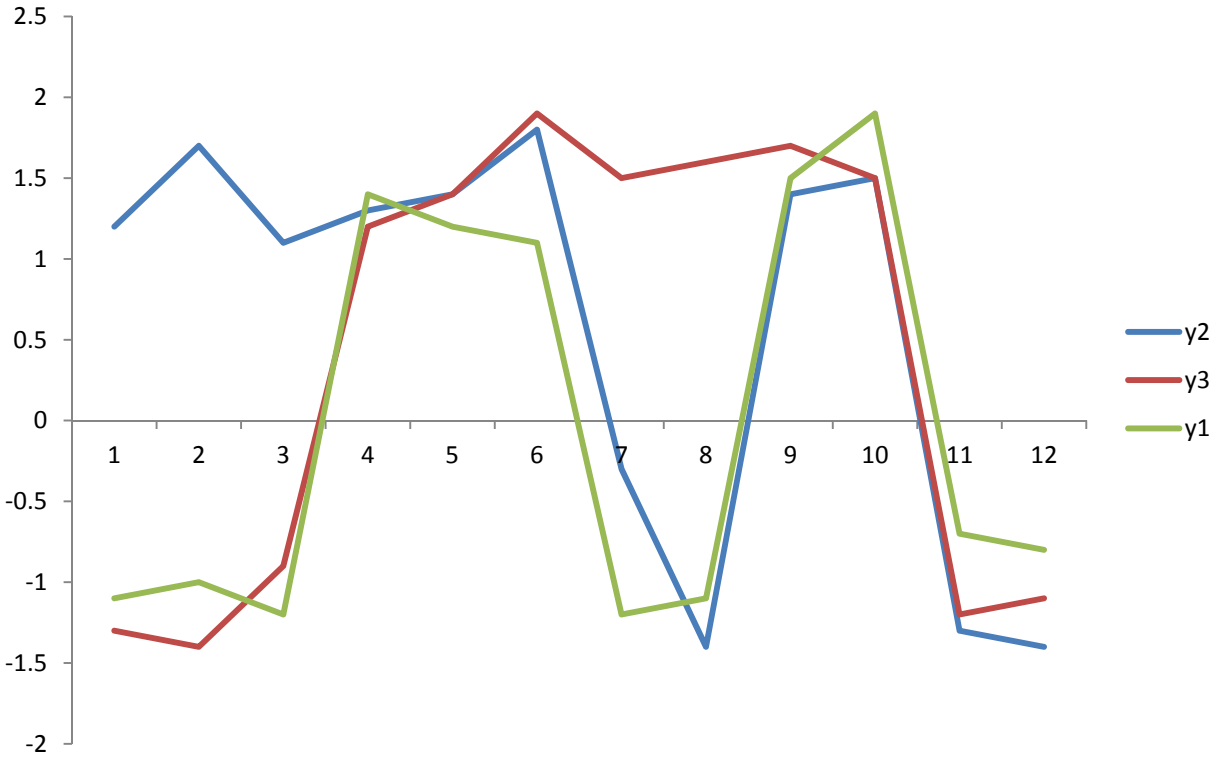


Interacting genes/protein/metabolites

Emergent properties: differential expression



Emergent properties: AND logics in regulation



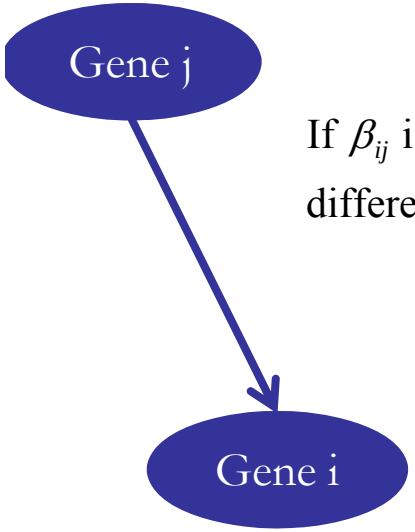
Example: Three genes

α = -0.46
β_{12}= 0.43
β_{13}=0.50

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

Expr	y_2	y_3	y_1	y_1 predicted	
Cond. A	1.2	-1.3	-1.1	$a + \beta_{12} \cdot 1.2 - \beta_{13} \cdot 1.3$	-0.594
Cond. B	1.7	-1.4	-1	$a + \beta_{12} \cdot 1.7 - \beta_{13} \cdot 1.4$	-0.429
Cond. C	1.1	-0.9	-1.2	$a + \beta_{12} \cdot 1.2 - \beta_{13} \cdot 0.9$	-0.437
Cond. D	1.3	1.2	1.4	$a + \beta_{12} \cdot 1.3 + \beta_{13} \cdot 1.2$	0.699
Cond. E	1.4	1.4	1.2	$a + \beta_{12} \cdot 1.4 + \beta_{13} \cdot 1.4$	0.842
Cond. F	1.8	1.9	1.1	$a + \beta_{12} \cdot 1.8 + \beta_{13} \cdot 1.9$	1.264
...	...	...	...	...	...

Correlation: 0.78



If β_{ij} is significantly different from 0!

Choose α, $\beta_{12}$and β_{13} so that the correlation between observed (y_1) and predicted (y_1 predicted) expression is maximized!

Overfitting and Occam's razor

Experiments/samples (data):

$$x = 7y$$
$$y = 3 + x$$

Has a unique solution: x=-3.5, y=-0.5

Gene interaction (fitted β):

$$x = 7y$$
$$y = z + x$$

Has many solutions: z=3, x=-3.5, y=-0.5
z=6, x=-7, y=-1
...

$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$

n=~30 000 for humans
n > 20 000 for plants

Occam's razor: The simplest model that best explains the data should be chosen

- Require weighting model complexity (no. parameters) against model fit (p-value)
- Example: multiple hypotheses correction (significance threshold = 0.05/n)

Linear versus non-linear models

- Linear model:

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$$

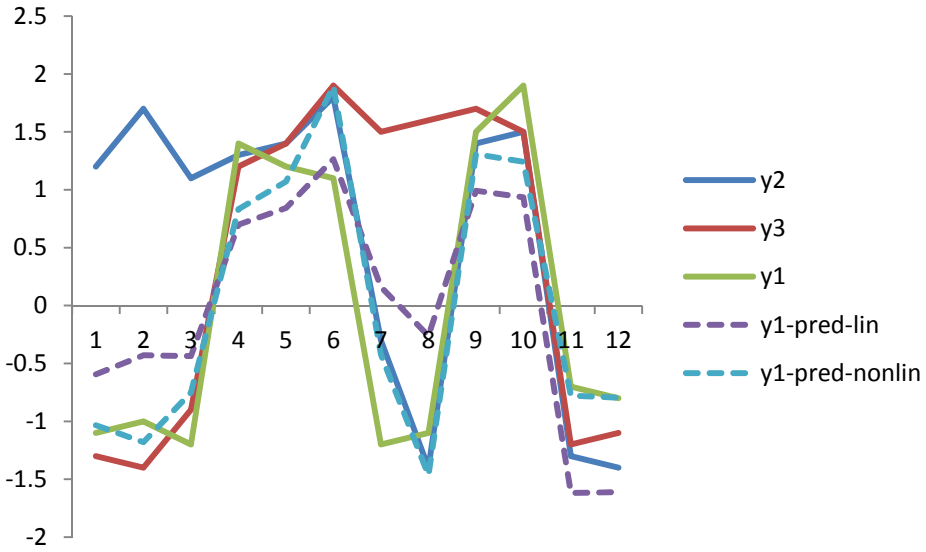
- Non-linear model:

$$y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3 + \beta_{123} y_2 y_3$$

$\beta_{123} > 0$: synergistic interactions

$\beta_{123} < 0$: competitive relationship

AND - logic



Linear model:

α = -0.46

β_{12}= 0.43

β_{13}=0.50

Non-linear model:

α = -0.55

β_{12}= 0.37

β_{13}=0.27

β_{123}=0.37

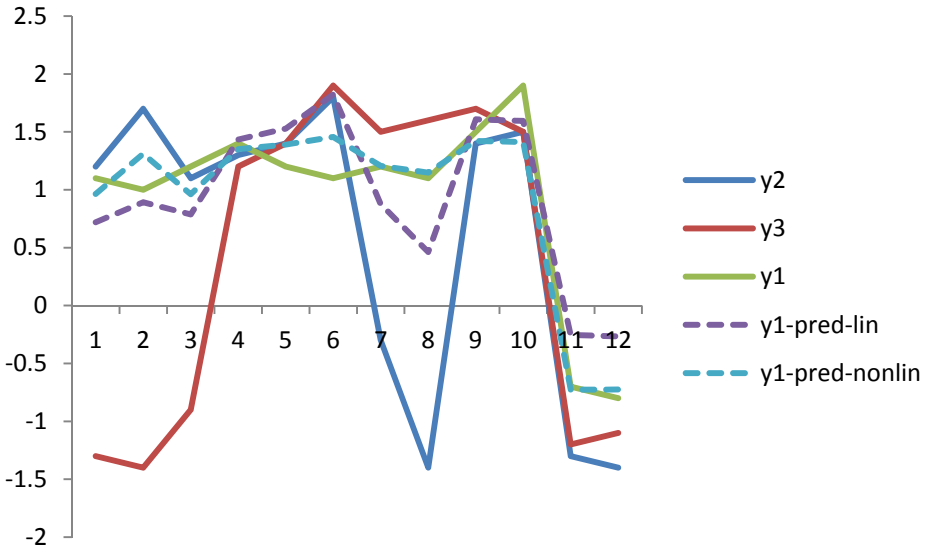
Correlation between observed and predicted:

linear model:	0.77 (P < 0.0029)
non-linear model:	0.91 (P < 1.2e-5)

Correlation between gene 1 and

gene 2:	0.55 (P < 0.061)
gene 3:	0.65 (P < 0.022)

OR - logic



Linear model:

α = 0.59

β_{12}= 0.40

β_{13}=0.27

Non-linear model:

α = 0.64

β_{12}= 0.43

β_{13}=0.40

β_{123}=-0.21

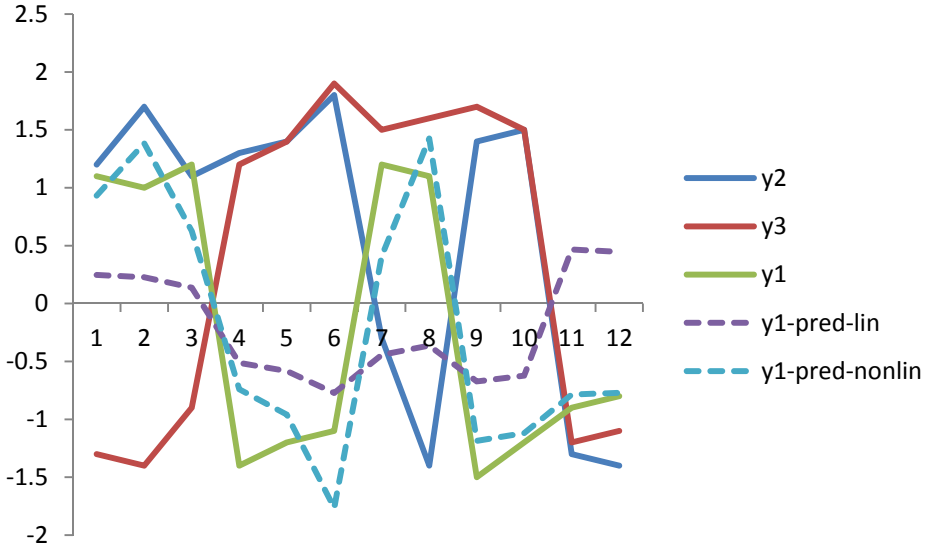
Correlation between observed and predicted:

linear model:	0.85 (P < 4.5e-4)
non-linear model:	0.96 (P < 7.9e-7)

Correlation between gene 1 and

gene 2:	0.72 (P < 0.0086)
gene 3:	0.60 (P < 0.041)

XOR - logic



Linear model:

α = -0.02

β_{12}= -0.10

β_{13}=-0.30

Non-linear model:

α = 0.11

β_{12}= -0.01

β_{13}=0.03

β_{123}=-0.56

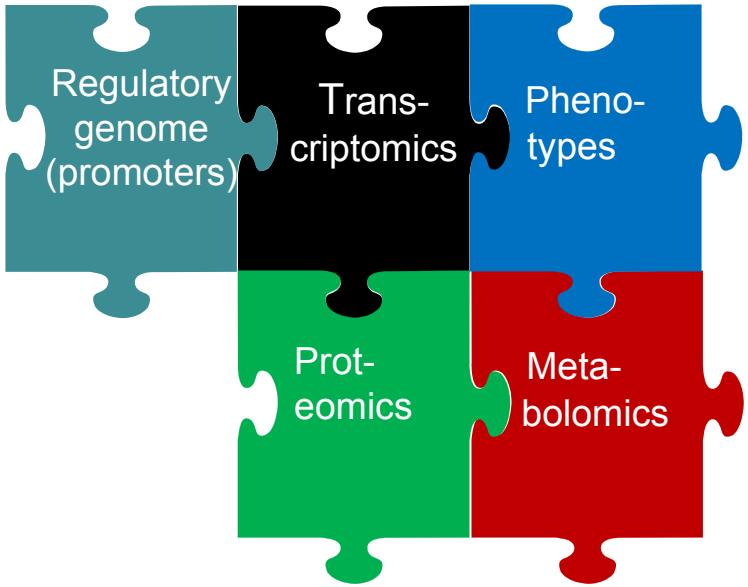
Correlation between observed and predicted:

Linear model:	0.40 (P < 0.19)
Non-linear model:	0.92 (P < 1.83e-5)

Correlation between gene 1 and

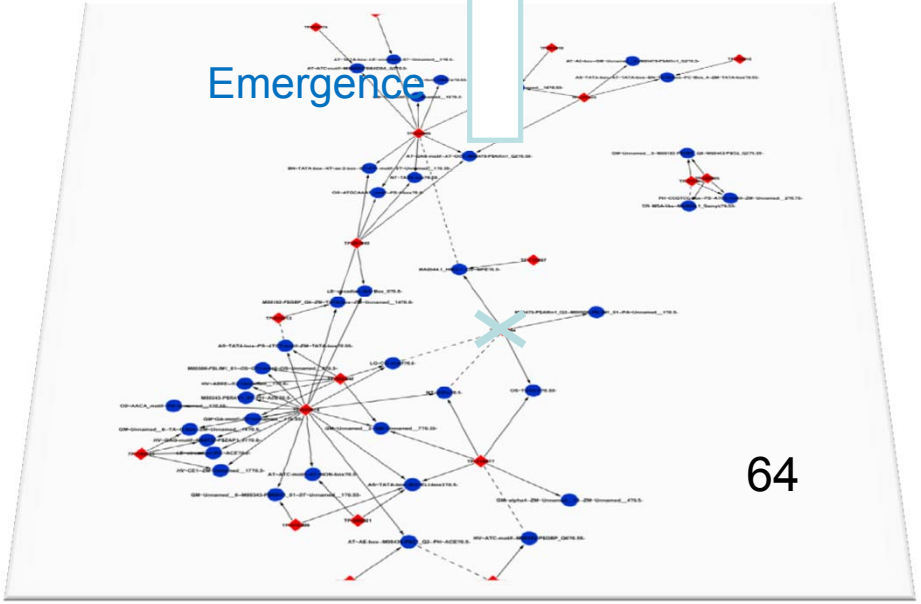
gene 2:	-0.19 (P < 0.56)
gene 3:	-0.39 (P < 0.21)

Data integration



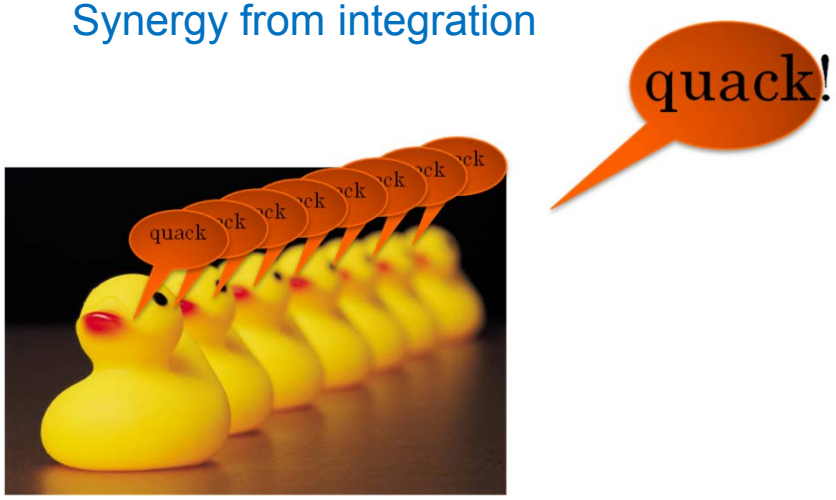
Phenotypes

Emergence



Interacting genes/protein/metabolites

Synergy from integration



Integration

Many different approaches:

- One model: use all data to infer one model
  - the ultimate goal, but can result in even more parameters
  - e.g. using transcriptomics and proteomics data on the same samples

- Consensus model: infer one model per dataset and take the intersection
  - low sensitivity (no novel findings)/high specificity
  - e.g. one network for transcriptomics, one for proteomics: consensus network with edges that exists in both networks

- Conditional integration: only combine congruent data

"One model" integration – same constraints, more parameters

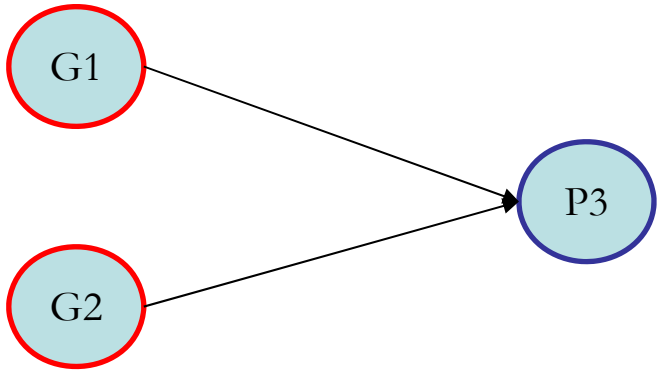
Samples

Genes

0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06

Proteins

0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06



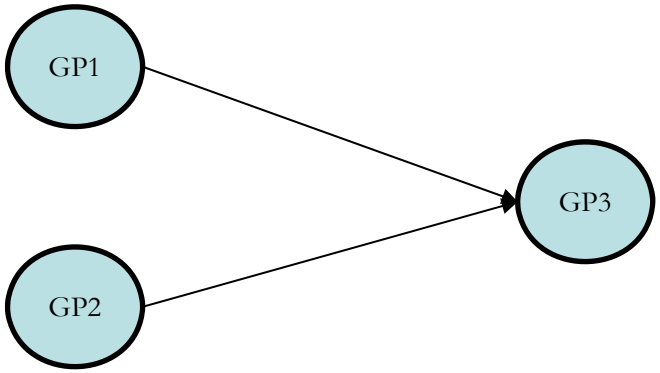
"One model" integration –
more constraints, same parameters

Samples

Genes/Proteins

0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06

0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06



"Consensus model" integration – same constraints, same parameters

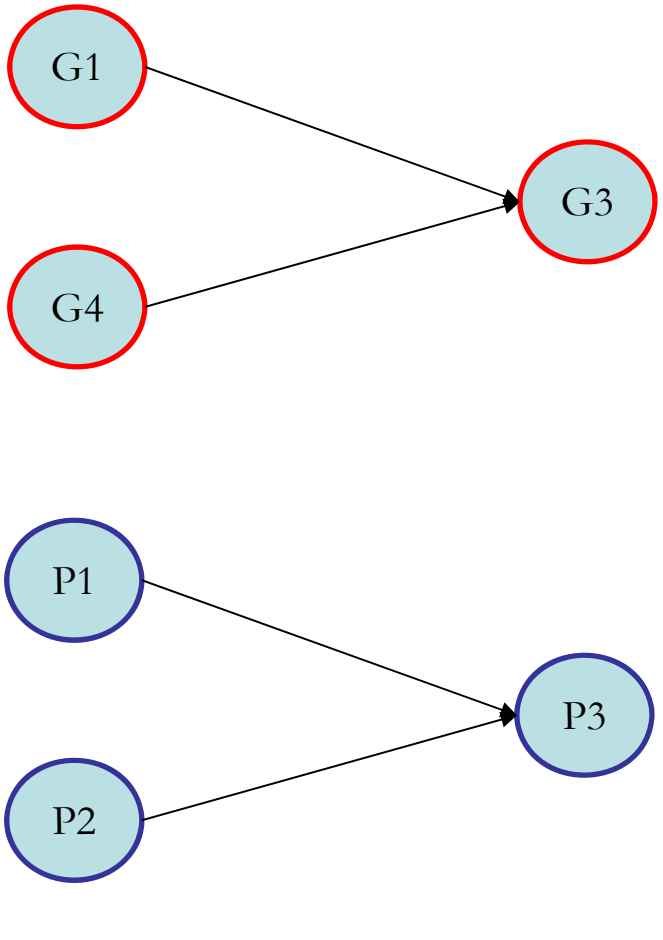
Samples

Genes

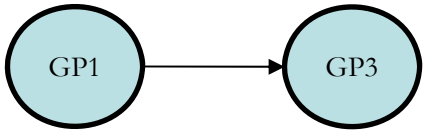
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06

Proteins

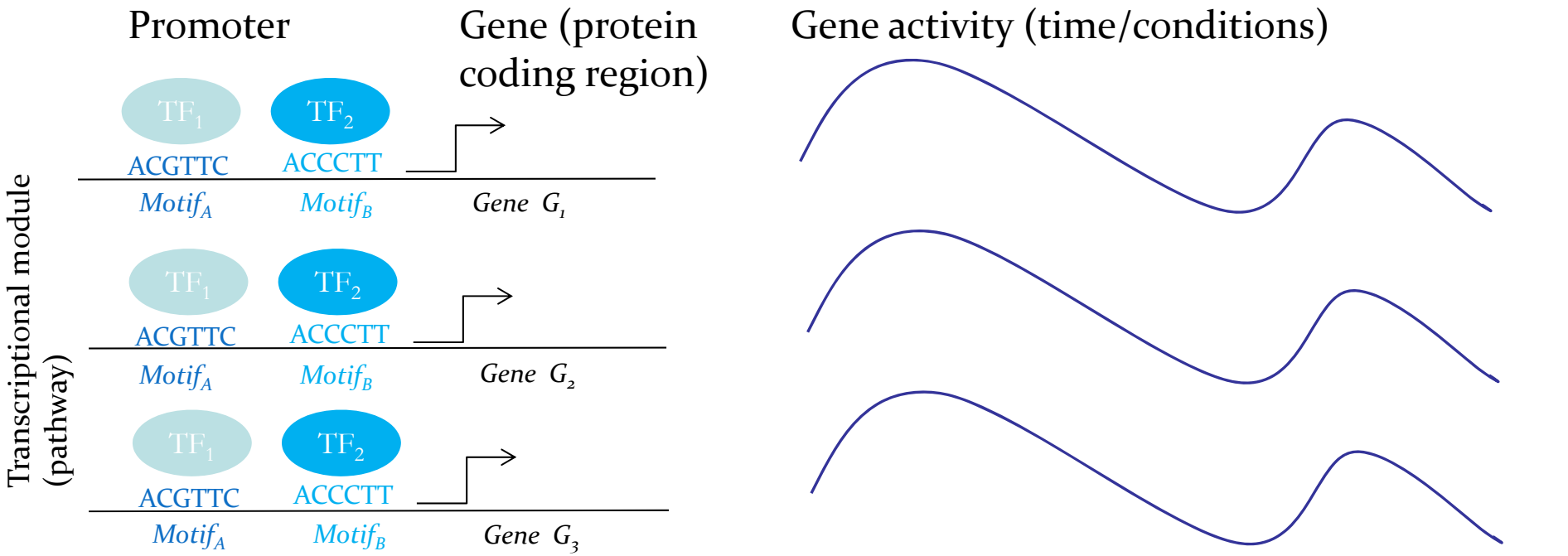
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
-0.47	-3.32	-0.81	0.11	-0.60	-1.36	-1.03	-1.84
0.66	0.07	0.20	0.29	-0.89	-0.45	-0.29	-0.29
0.14	-0.04	0.00	-0.15	-0.58	-0.30	-0.18	-0.38
-0.04	0.00	-0.23	-0.25	-0.47	-0.60	-0.56	-1.09
0.28	0.37	0.11	-0.17	-0.18	-0.60	-0.23	-0.58
0.54	0.53	0.16	0.14	0.20	-0.34	-0.38	-0.36
0.20	0.14	0.00	0.11	-0.34	-0.03	0.04	-0.76
0.40	0.43	0.18	0.00	-0.14	0.29	0.07	-0.79
0.01	0.46	0.28	-0.34	-0.23	-0.36	-0.45	-0.64
…	…	…	…	…	…	…	…
-0.23	0.04	0.00	-0.30	-0.29	-0.45	-0.97	-2.06



Consensus



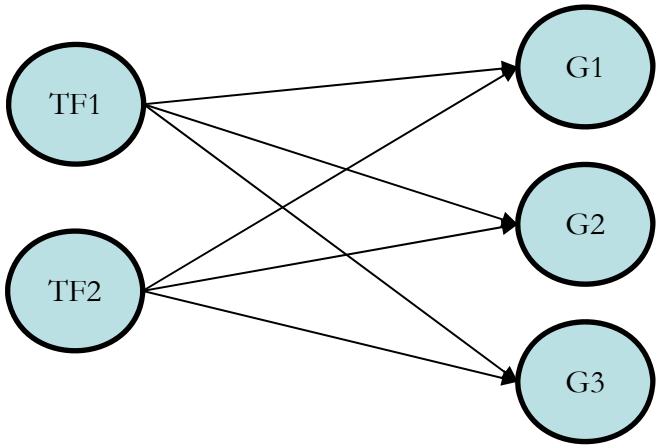
Integration: gene regulatory networks



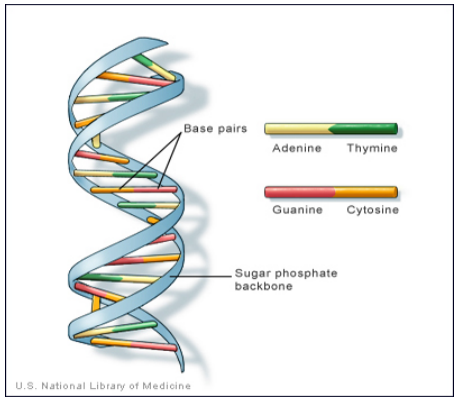
Data integration: Expression + promoter information

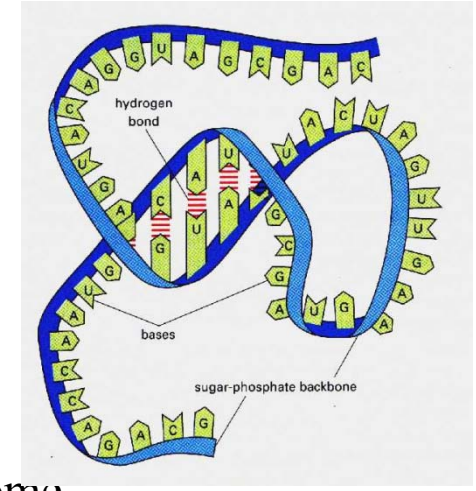
Chip-Seq could provide further information on

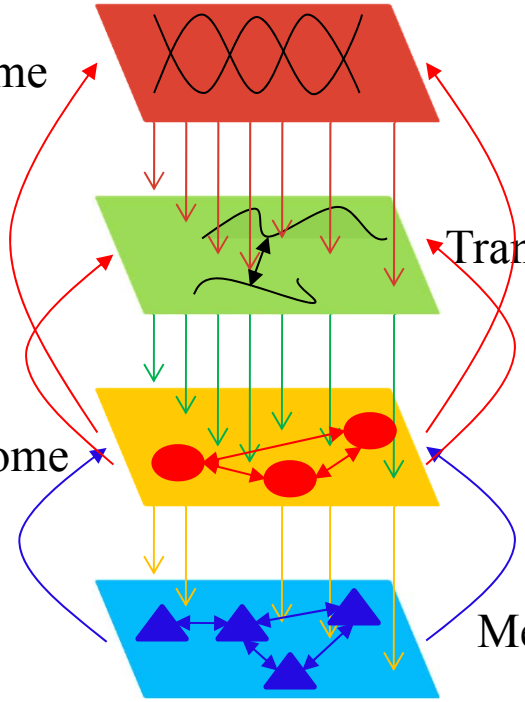
- Open chromatin
- Physical interactions between TFs and DNA
- …

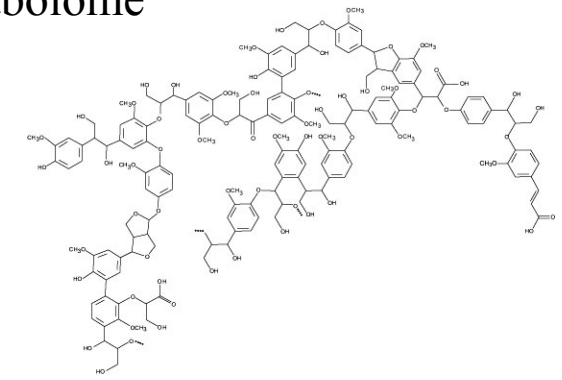


Next-generation genomics









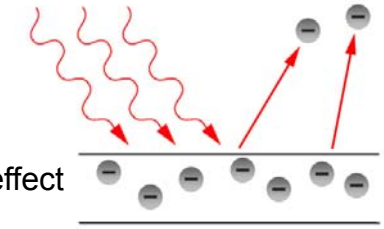
emergent properties
+
integration

Ernest
Rutherford

“All science is either physics or stamp collecting”

Physics

Photoelectric effect

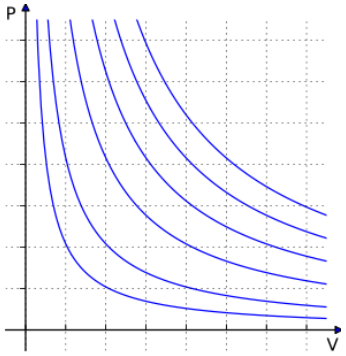


$$E = hv$$

E = energy
H = Planck's constant
v = frequency of light radiation

Chemistry

Ideal gas law

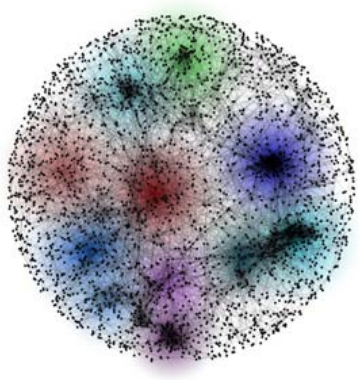


$$PV = nRT$$

P = absolute pressure
V = volume of the vessel
n = number of moles of gas
R = ideal gas constant
T = absolute temperature

Biology

~~Gene interactions~~



$$y_i = \alpha_i + \sum_{j=0}^{n} \beta_{ij} y_j$$

y_i = gene expression of gene i
n = number of genes
α = transcription rate
β_{ij} = effect of gene j on gene i

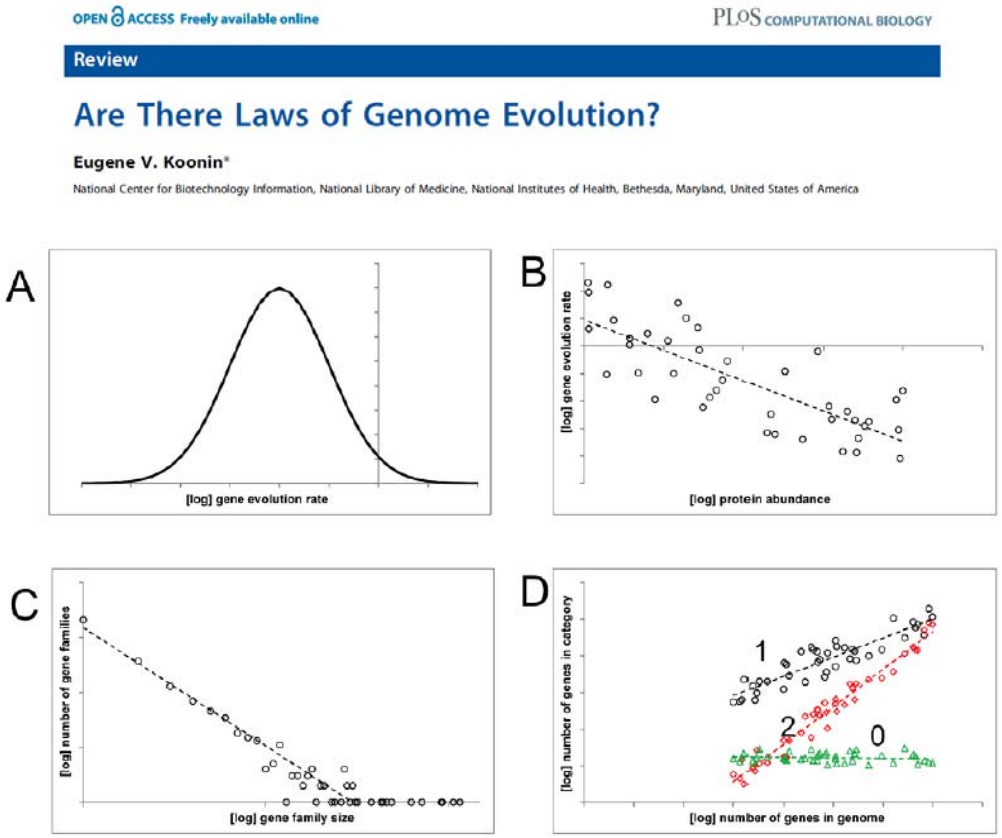
Scale free networks

$$P(k) \sim k^{-\gamma}$$

k = node degree
$P(k)$ = degree distribution
γ = degree exponent

Laws of genome evolution

A. Log-normal distribution of the evolutionary rates between orthologous genes
B. Negative correlation between gene sequence evolution rate and expression level (or protein abundance)
C. Power law–like distributions of membership in paralogous gene families and node degree in biological networks
D. Distinct scaling of functional classes of genes with genome size



0. No dependence: translation
1. Linear dependence: enzymes
2. Quadratic dependence: regulation/signaling

Koonin. PLoS Computational Biology 7:e1002173, 2011.

Summary: Systems biology

- Traditional methods treat and visualize genes as independent entities (reductionistic):
  - Hierarchical clustering
  - Co-expression networks
- Systems biology treat and visualize genes in the context of other genes (holistic)
  - Gene networks
  - Gene regulatory networks

Some freely available tools

- R contains packages for most methods discussed here
- Hierarchcial clustering: MeV (MultiExperiment Viewer)
- Machine learning: RapidMiner
- Networks: Cytoscape