‘omics data analysis and systems biology

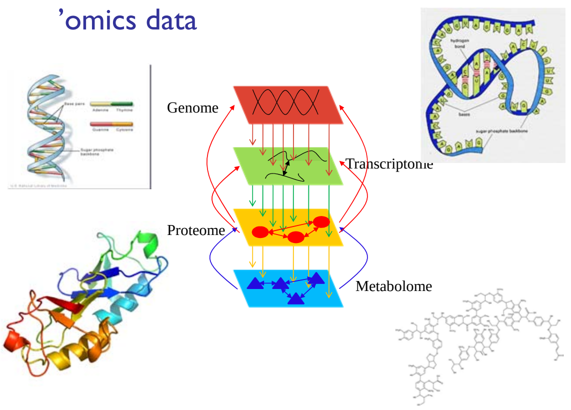
Slides: http://www.trhvidsten.com/Teaching.html

Torgeir R. Hvidsten
Assistant professor in Bioinformatics
Umeå Plant Science Centre (UPSC)
Computational Life Science Cluster (CLiC)

’omics data

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

’omics data

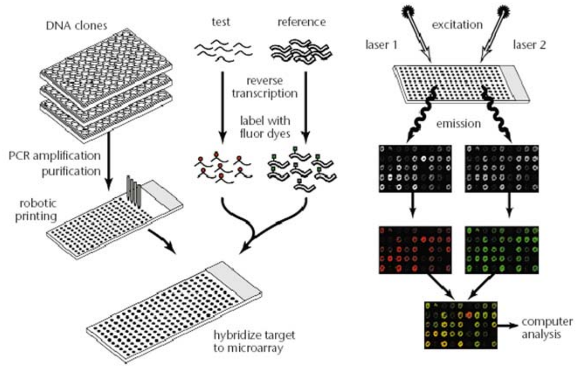


Analysis of ’omics data

1. Preprocessing
2. Browsing the data
3. Model inference and selection
4. Model evaluation
5. Genome annotation quality
6. Result visualization
7. Systems biology

Pre-processing and browsing

Microarray



Hybridization

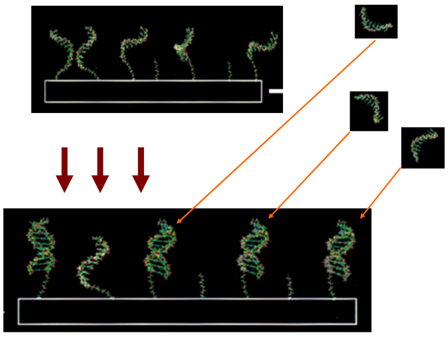
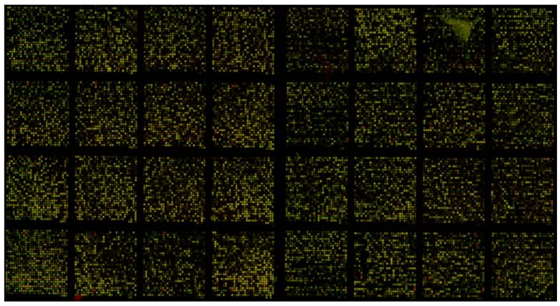


Image after scanning



Microarray 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

2.3/2.4 = “Red/Green”

log-transformation

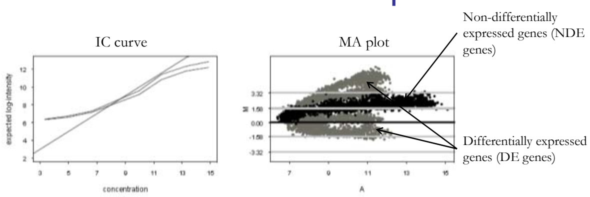
M < 100

Gene/Expr	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	...	EM
G1	-0,47	-3,32	-0,81	0,11	-0,60	-1,36	-1,03	-1,84	-1,00	-0,60	...	-0,94
G2	0,66	0,07	0,20	0,29	-0,89	-0,45	-0,29	-0,29	-0,15	-0,45	...	-0,42
G3	0,14	-0,04	0,00	-0,15	-0,58	-0,30	-0,18	-0,38	-0,49	-0,81	...	-1,12
G4	-0,04	0,00	-0,23	-0,25	-0,47	-0,60	-0,56	-1,09	-0,71	-0,76	...	-0,62
G5	0,28	0,37	0,11	-0,17	-0,18	-0,60	-0,23	-0,58	-0,79	-0,29	...	-0,74
G6	0,54	0,53	0,16	0,14	0,20	-0,34	-0,38	-0,36	-0,49	-0,58	...	-1,47
G7	0,20	0,14	0,00	0,11	-0,34	-0,03	0,04	-0,76	-0,81	-1,12	...	-1,36
G8	0,40	0,43	0,18	0,00	-0,14	0,29	0,07	-0,79	-0,81	-0,92	...	-1,22
G9	0,01	0,46	0,28	-0,34	-0,23	-0,36	-0,45	-0,64	-0,79	-1,22	...	-1,09
...	...	...	...	...	...	...	...	...	...	...	...	...
GN	-0,23	0,04	0,00	-0,30	-0,29	-0,45	-0,97	-2,06	-0,89	-1,22	...	-0,97

N ≈ 10000

log(2.3/2.4) = log(“Red/Green”)

IC-curve and MA plot



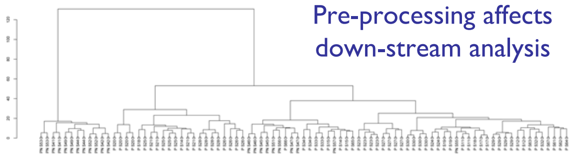
cDNA-microarray experiments where two populations are compared

- IC curve: In spike-in experiments all RNA-abundances are known: the IC-curve plots the expected RNA-abundances against the measured values (*i.e.* the *concentration*)
- MA plot: log-ratios are plotted against the average log-intensities

’omics preprocessing

- **Background correction.** Aims to straighten the lower knee in the IC-curve.
- **Saturation correction.** Aims to straighten the upper knee in the IC-curve.
- **Dye normalization.** Aims to put the IC-curves into a common scale (common slope).

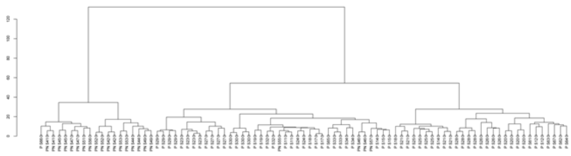
A. without background correction



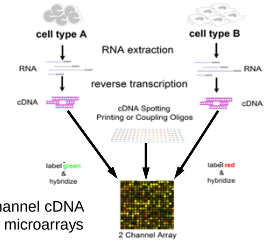
Pre-processing affects down-stream analysis

Hepatitis B virus-positive metastatic hepatocellular carcinomas:
P: 65 patients with metastasis
NP: 22 patients with no metastasis

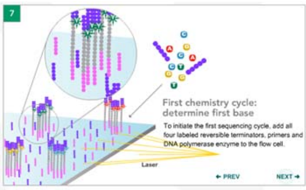
B. with background correction



Microarrays versus RNA-Seq



Two-channel cDNA microarrays

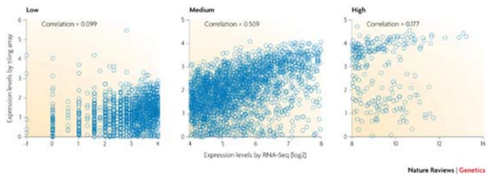


Illumina (HiSeq 2000)
Sequencing by synthesis

Microarrays versus RNA-Seq

RNA-Seq advantages:

- Can detect "all" transcribed regions (including small RNA)
- Can be applied to all organisms
- Broader dynamic range: higher sensitivity and specificity
- Can do much more than just quantifying gene expression (SNP detection, Isoform detection, etc)



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): 3 samples per lane
- 10 - 150 ng of total RNA per wood section requires amplification

Mapping reads to reference genome

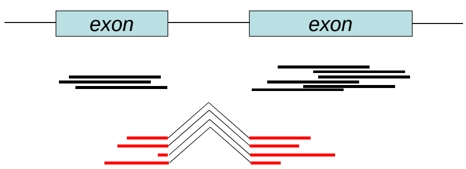
Raw data:

```
@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

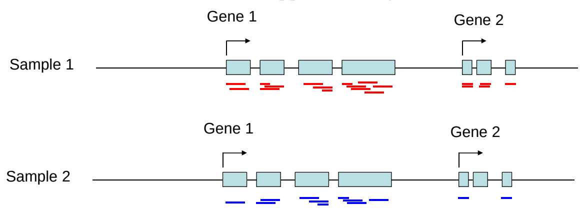
reference

Aligned/
Unaligned
reads



Quantifying expression

- Count the number of reads mapped to each gene



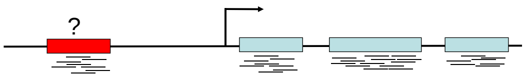
RPKM = **R**eads **P**er **K**ilobase of exon model per **M**illion mapped reads

	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

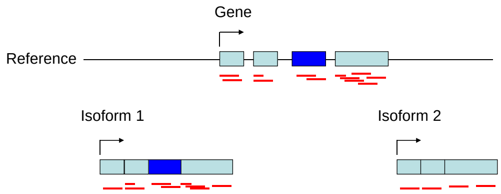
- Detect regions outside known gene models



- Go through whole genome
  - Sliding window or similar
  - Search for regions with high coverage
  - Do semi-*de novo* transcript assembly

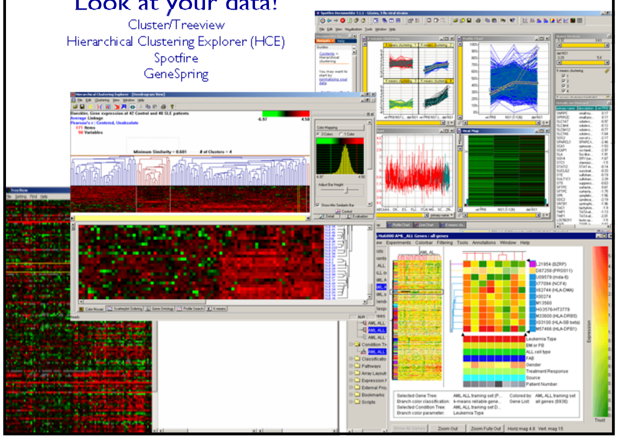
Isoform detection (splicing variants)

- Detected by bethods that reconstruct entire transcripts

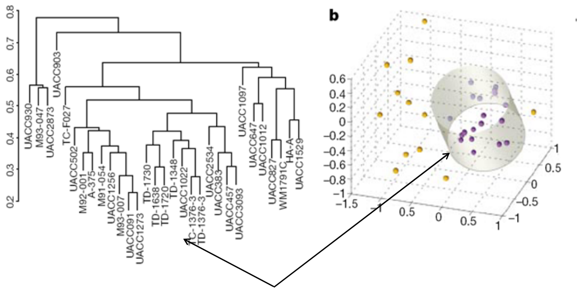


Look at your data!

Cluster/Treeview
Hierarchical Clustering Explorer (HCE)
Spotfire
GeneSpring



Hierarchical clustering and principle component analysis (PCA)



19 melanomas of all 31 cutaneous melanoma samples
(Bitter et al. *Nature.* 406: 536, 2000)

Model inference and selection

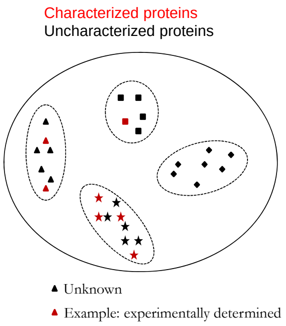
Model inference methods

- Unsupervised learning (clustering, class discovery); used to "discover" natural groups of genes/experiments e.g.
  - discover subclasses of a form of cancer that is clinically homogenous
- Supervised learning; used to "learn" a model of a set of predefined classes of genes/experiments e.g.
  - diagnosis of cancer/subclasses of cancer

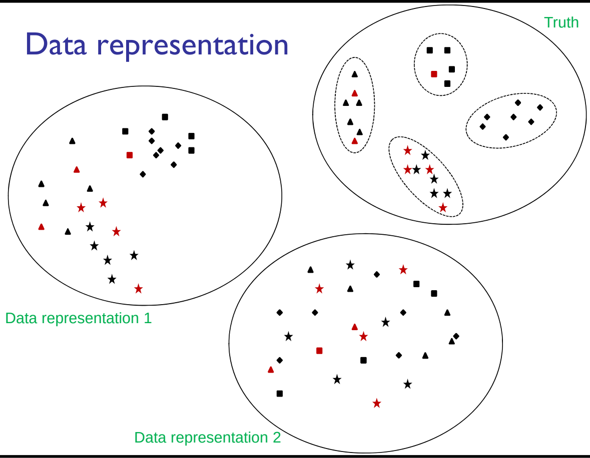
The machine learning strategy …

… iteratively uses experiments to provide representative examples and computational models to provide experimentalists with new, testable hypotheses

- Clustering
- Nearest neighbor predictors
  - evolutionary link
  - need few examples
- Model inducers
  - more powerful
  - interpretable models



Data representation



Clustering analysis

Need to define;

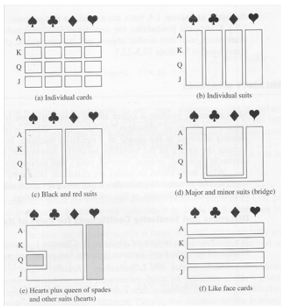
- measure of similarity
- algorithm for using the measure of similarity to discover natural groups in the data

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

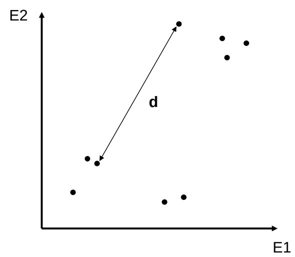
Example: $10^{500}/10! = 2.756 \times 10^{493}$

Measure of similarity

What is similar?

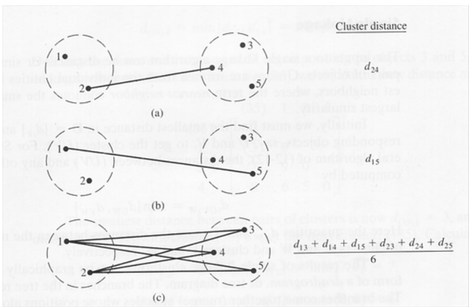


Euclidean distance



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

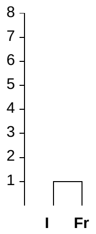
Distance: Frequency of numbers with different first letter e.g.

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

Inter-cluster strategy: SINGEL LINKAGE

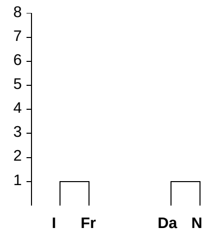
Iteration 1

	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



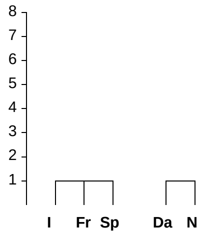
Iteration 2

	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



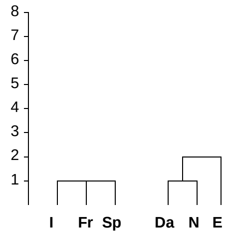
Iteration 3

	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



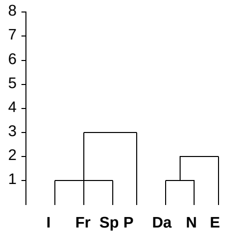
Iteration 4

	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



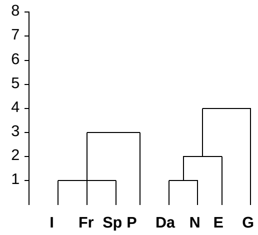
Iteration 5

	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



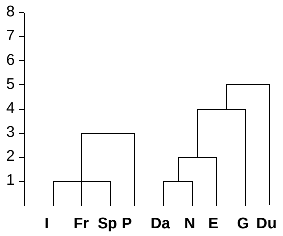
Iteration 6

	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



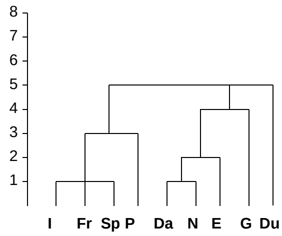
Iteration 7

	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



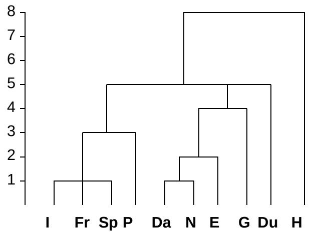
Iteration 8

	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



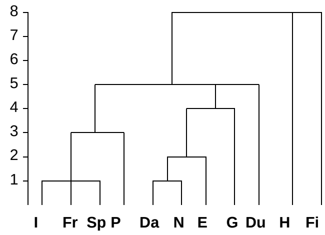
Iteration 9

	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

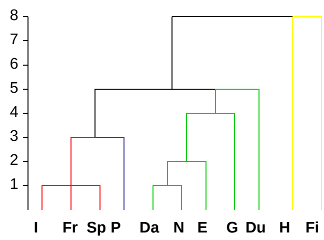
	Fi H	P Sp I Fr Du G E Da N
Fi H	0	
P Sp I Fr Du G E Da N	8	0



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: Decision tree learning

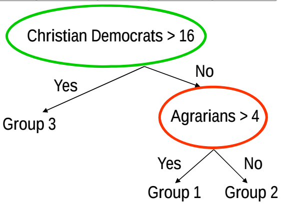
Country	Communists	Socialists	Greens	Social Democrats	Liberals	Agrarians	Subnational, regional and ethnic parties	Christian Democrats	Conservatives	Extreme Right
Norway	0	7	0	38	4	8	0	9	24	6
Sweden	6	0	2	43	10	17	0	2	18	1
Denmark	4	9	0	33	13	14	0	3	15	9
Finland	15	0	2	24	3	25	5	3	21	0
Iceland	0	18	3	16	4	22	0	0	36	0
UK	0	0	9	39	15	0	4	0	42	0
Netherlands	2	5	0	30	23	0	0	37	0	0
Belgium	2	0	4	27	19	0	14	31	0	2
Luxembourg	6	1	3	31	21	0	0	34	0	1
Switzerland	2	2	7	22	23	11	0	22	3	5
Austria	1	0	2	48	0	0	0	41	0	8
Germany	1	0	3	40	9	0	0	46	0	1
France	15	2	2	28	20	0	0	0	25	5
Italy	29	0	3	15	4	0	3	35	2	6
Greece	10	0	0	39	6	0	0	0	44	0
Spain	8	0	0	39	16	0	10	0	21	0
Portugal	15	0	1	31	38	0	0	1	11	0

Class knowledge:

Group 1: Nordic countries

Group 2: UK, France, Greece, Spain, Portugal

Group 3: Benelux countries, Switzerland, Austria, Italy, Germany



Example: Decision tree learning

Some concepts:

1. Data: Observations collected from the real world (e.g. the voting pattern in Sweden). Observations consist of a number of features (e.g. communist votes)
2. Examples: Observations labeled with class information (e.g. Sweden belong to group 1).
3. Model: A general representation of the data (e.g. the decision tree)

Models are induced!

1. Induction: Using specific information/data to arrive at general knowledge (e.g. from examples to a decision tree).
2. Deduction: Using general knowledge to say something about a specific case (e.g. using a decision tree to predict the group of a new country).

Models can be predictive and/or descriptive.

Prior Probability

- w - state of nature, e.g.
  - w_1 the object is a fish, w_2 the object is a bird, etc.
  - w_1 this course is good, w_2 this course is bad
  - etc.
- *A priori* probability (or prior) $P(w_j)$

Class-conditional probability

- Observation x, e.g.
  - The objects has wings
  - The 10 minutes of the lecture was interesting
- Class-conditional probability $p(x|w)$

Bayes decision rule

Suppose the priors $P(w_j)$ and conditional densities $p(x|w_j)$ are known

likelihood prior

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j) P(\omega_j)}{p(x)}$$

posterior evidence

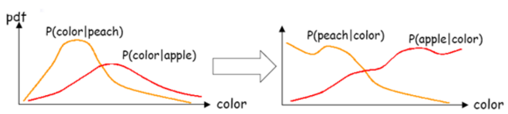
Bayes decision rule:

Two classes: If $P(w_1|x) > P(w_2|x)$ then choose w_1, else choose w_2.

In general: Choose

$$w^* = \arg\max_i P(w_i \mid x)$$

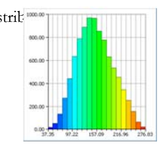
Example



- Bayes Decision Rule
  - If P(apple | color) > P(peach | color) then choose apple
- Note that the evidence p(color) is only necessary for normalization purposes; it does not affect the decision rule

So, what about the data?

- Use examples to estimate the probability distrib
  - $P(w_j)$ is easy.
  - $p(x|w_j)$: Histogram!



- One feature: bins are rectangles, Two features: cubes, n-features: hyper-cubes.
- More dimensions/features require more training data: Curse of dimensionality!
  - If we need 10 observations when we have one feature (to get a good histogram), then we need 10^n observations when we have n-features!
- If the true probability distributions are known, then Bayes decision rule is optimal (minimizes error rate).

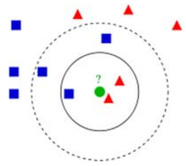
Feature selection

Feature selection is used to deal with the curse of dimensionality

- Ranking methods: compute the discriminatory capability of each feature and select the best ones
- Wrapper methods: select a subset of features, induce a model and use it's prediction performance as fitness. Repeat. Computationally expensive!
- Dimensionality reduction: map your features into a smaller features space (e.g. PCA)

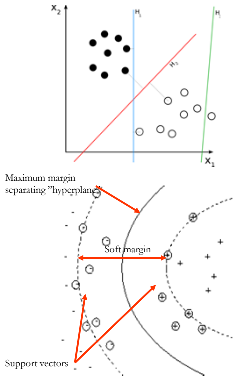
k-nearest neigboor

- The simplest of all machine learning algorithms.
- Each observation is a point in the n-dimensional space spanned by the features.
- An observation is assigned to the class most common amongst its k nearest neighbors.
- "Nearest" can be defined differently: Euclidean distance, correlation, etc.
- Lazy learning where the function is only approximated locally and all computation is delayed until classification.



Linear versus non-linear classifiers

- Linear: Finds a hyperplane that separates the classes
  - In two dimensions: $w_0 + w_1 \cdot x_1 + w_2 \cdot x_2$
  - Use the examples $\boldsymbol{x}$ to estimate $\boldsymbol{w}$
- Non-linear: Support vector machines uses the kernel trick:
  - The kernel maps the observations into a higher dimensional space where the problem is linearly separable

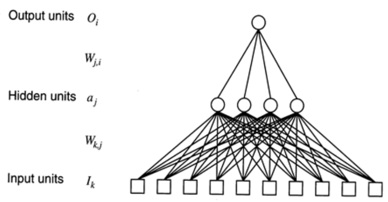


Artifical neural networks

- Inspired by how the brain works – a mathematical model of the operation of the brain
- Brain versus computers:
  - serial versus parallell computing
  - even though a computer is much faster in raw swithcing speed, the brain is faster at what it does
- An ANN is a number of nodes (units) connected by links. Each link is associated with a numerical weight.
  - Training set: $(x_1, f(x_1)), (x_2, f(x_2)), ..., (x_n, f(x_n))$
  - Learning in an ANN is reduced to the process of using the training data to tune the weights so that the network represents the function f

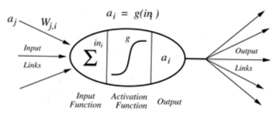
Network structure

- Feed-forward network: all units are connected to all units in the next layer
  - One (sufficiently large) hidden layer can represent any continuous function
  - More hidden layers can even represent discontinuous functions

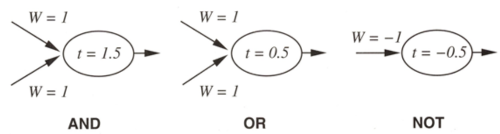


- Recurrent network: feed back loops, internal states (memory):
  - E.g. The brain is clearly a recurrent network

Boolean functions



- Units can represent the basic logical gates
- Thus, units can build networks that can represent any Boolean function

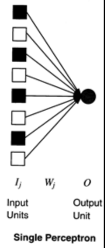


Optimal network structures, overfitting and Occam's razor

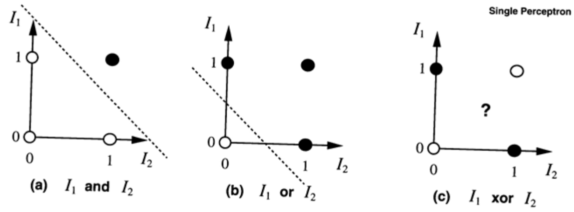
- Too small network: the network will be incapable of representing the desired function
- Too large network: the network can memorize all the examples by forming a lookup table: Overfitting!

- Every algorithm involved with classification runs the risk of overfitting the data
  - The alg. learns the errors (noise) in the data as well as the underlying structure of the processes that created the data
  - Occurs because the alg. tries to reduce the classification error on the training data
  - A model X is overfitted if there exists a model Y that do better on the unseen test set, but worse on the training set
- To identify this phenomenon:
  - Use training/test sets
  - Choose the simples model that explains the data! Occam's razor

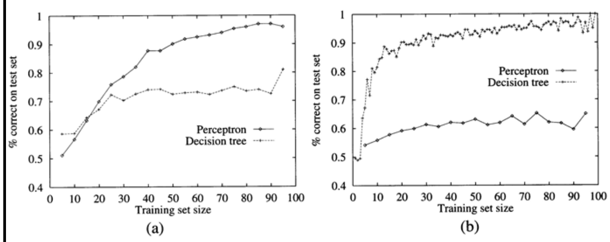
Perceptrons



- Perceptrons: single-layer, feed-forward networks
  - Majority function: outputs 1 if a majority of the n inputs are 1 (would require a decision tree with $O(2^n)$ nodes)
- A perceptron can only represent a function if there is a line that separates all the white dots (0s) from the black dots (1s), i.e. functions that are linearly separable



Perceptrons versus decision trees: Example



(a) **Majority function**
(b) **Waiting problem**

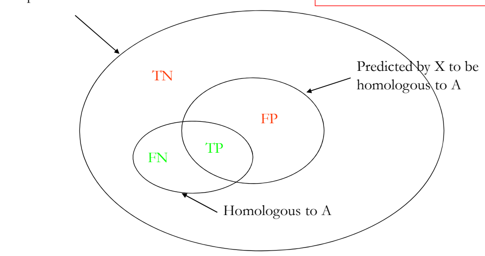
Example	Attributes										Goal
	Alt	*Bar*	*Fri*	*Hun*	*Pat*	*Price*	*Rain*	*Res*	*Type*	*Est*	*WillWait*
X_1	Yes	No	No	Yes	Some	$$$	No	Yes	French	0–10	Yes
X_2	Yes	No	No	Yes	Full	$	No	No	Thai	30–60	No
X_3	No	Yes	No	No	Some	$	No	No	Burger	0–10	Yes
X_4	Yes	No	Yes	Yes	Full	$	No	No	Thai	10–30	Yes
X_5	Yes	No	Yes	No	Full	$$$	No	Yes	French	>60	No
X_6	No	Yes	No	Yes	Some	$$	Yes	Yes	Italian	0–10	Yes
X_7	No	Yes	No	No	None	$	Yes	No	Burger	0–10	No
X_8	No	No	No	Yes	Some	$$	Yes	Yes	Thai	0–10	Yes
X_9	No	Yes	Yes	No	Full	$	Yes	No	Burger	>60	No
X_{10}	Yes	Yes	Yes	Yes	Full	$$$	No	Yes	Italian	10–30	No
X_{11}	No	No	No	No	None	$	No	No	Thai	0–10	No
X_{12}	Yes	Yes	Yes	Yes	Full	$	No	No	Burger	30–60	Yes

Model evaluation

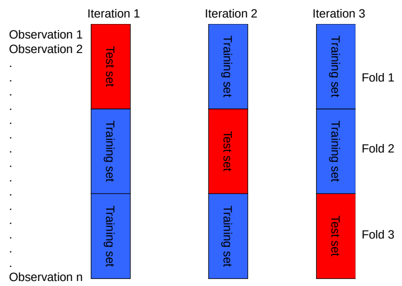
Method power

You want to find homologous proteins to a specific protein A using some computational method X:

Sensitivity: TP/(TP+FN)
Specificity: TN/(TN+FP)



Cross validation



- k-fold cross validation: k iterations
- Leave-one out cross validation: n iterations

Evaluation

- ➢ Classifications can be
  - True positives (TP)
  - False negatives (FN)
  - True negatives (TN)
  - False positives (FP)
- ➢ Evaluation measures:
  - accuracy = (TP+TN)/(TP+FN+TN+FP)
  - sensitivity = TP/(TP+FN)
  - specificity = TN/(TN+FP)
- ➢ Confusion matrix:

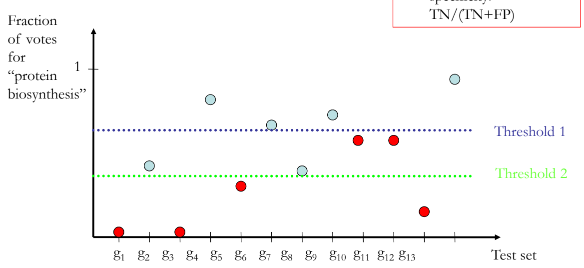
		Predicted	
		Class 0	Class 1
Actual	Class 0	TN	FP
	Class 1	FN	TP

Threshold selection

Gene with function "protein biosynthesis"

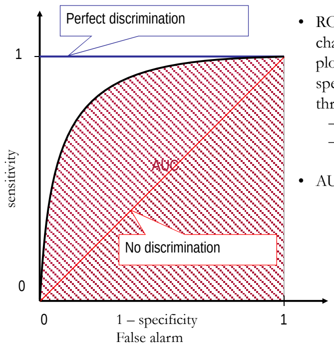
Gene with a different function

sensitivity: TP/(TP+FN)
specificity: TN/(TN+FP)



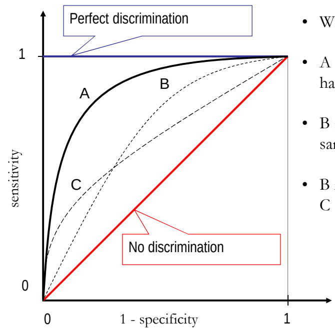
Sensitivity = 2/3, Specificity=1
Sensitivity = 1, Specificity=2/3

ROC analysis and classifier evaluation



- ROC: Receiver operating characteristics curve results from plotting sensitivity against specificity for all possible thresholds
  - sensitivity: TP/(TP+FN)
  - specificity: TN/(TN+FP)
- AUC: Area under the ROC curve

ROC analysis and classifier evaluation



- Which ROC curve is better?
- A dominants B and C and clearly has a higher AUC
- B and C have approximately the same AUC
- B is better for some thresholds, C for others

Machine learning summary

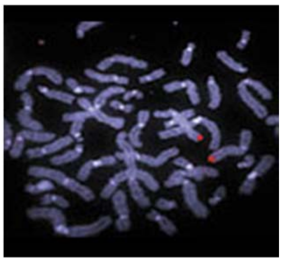
- ➢ Machine learning allows models with predictive and descriptive capabilities to be induced from examples
- ➢ Evaluation: training set, test set, cross validation, …
- ➢ Different approaches have different strengths and weaknesses
  - Linear versus non-linear
  - Interpretable versus black box
  - Regression versus classification

Machine learning summary cont.

- ➢ Overfitting: you select a model A over a model B when A performs better on the training set, but worse on the unseen test set
  - Stop before overfitting occurs (e.g. before the decision tree is to long or when the performance of the neural network no longer improves)
  - Occam's razor: Select the simplest model that explains the data (do not use non-linear methods on a linearly separable problem)
- ➢ Course of dimentionality
  - Rule of tumb: You need more observations than features
  - Use dimentionality reduction methods (e.g. PCA) or feature selection (on the traing set!)

Genome annotation quality

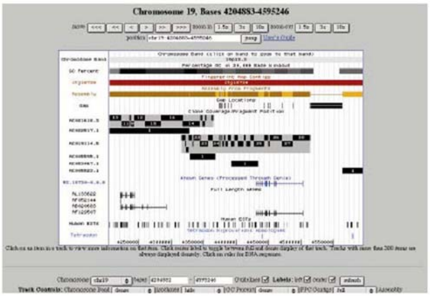
How to get from here…



UPSC

Stefan Jansson 2009

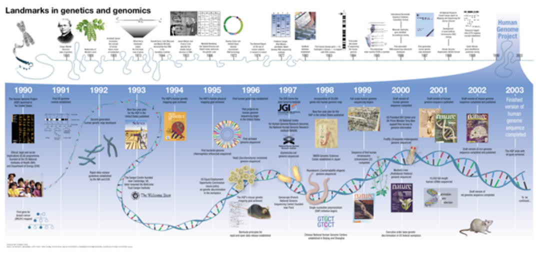
…to here?



UPSC

Stefan Jansson 2009

Human genome project timeline



UPSC

Stefan Jansson 2009

ESTs/RNASeq – A rapid gateway into the genome

- Only expressed parts of genes
- Necessary for genome annotation
- Short and incomplete
- Often bad quality and sometimes with cloning artifacts

UPSC

Stefan Jansson 2009

Whole genome shotgun sequencing

- 2, 10 and 50 kbp libraries
- Sequenced from both ends
- Sequence “mates”
- 8-fold coverage
- NOW: more and more use of short reads

UPSC

Stefan Jansson 2009

Full genome sequencing: Reconstruct the chromosomes…

UP

2009

Whole genome assembly and mapping

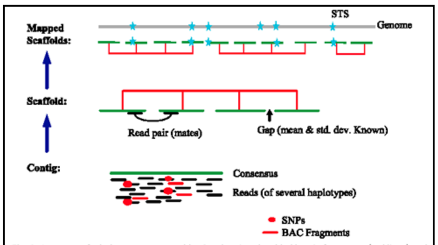
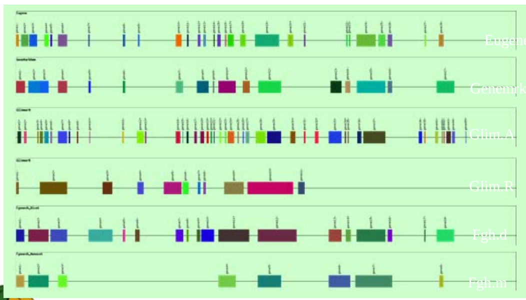


Fig. 3. Anatomy of whole-genome assembly. Overlapping shredded bactig fragments (red lines) and internally derived reads from five different individuals (black lines) are combined to produce a contig and a consensus sequence (green line). Contigs are connected into scaffolds (red) by using mate pair information. Scaffolds are then mapped to the genome (gray line) with STS (blue star) physical map information.

UPSC

Stefan Jansson 2009

Where are the genes?



UPSC

Stefan Jansson 2009

Algorithms must be trained for:

- Splice sites
- Exon/intron lengths
- Codon usage
- GC frequencies exons/introns
- Trancription start sites
- Polyadenylation sites
- UTR (untranlated region) lengths

and predicted exons must be joined to genes (ESTs necessary)

UPSC

Stefan Jansson 2009

Annotation:

- Comparisions to databases
- What is significant similarity (on protein level or on DNA level)?
- What if the other databases are wrong? (which they are)
- There is no ”best database”

UPSC

Stefan Jansson 2009

A typical genome (Arabidopsis)

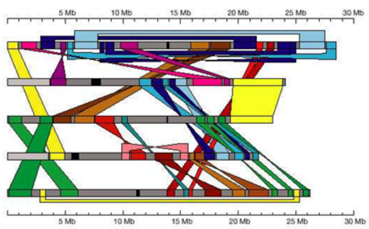


Figure 4 Segmentally duplicated regions in the *Arabidopsis* genome. Individual chromosomes are depicted as horizontal grey bars (with chromosome 1 at the top), centromeres are marked black. Coloured bands connect corresponding duplicated segments. Similarity between the rDNA repeats are excluded. Duplicated segments in reversed orientation are connected with twisted coloured bands. The scale is in megabases.

UPSC

Stefan Jansson 2009

A typical eukaryotic genome

- 15 - 50 000 genes
- Most in dispersed gene families
- Duplications
- Many repetitive sequences
  (e g microsatellites of 1-6 base pairs
- Many pseudogenes
- Centromers and telomers

UPSC

Stefan Jansson 2009

Annotation quality

- 1/3 of all genes are typially ”unknown”

For the rest, some kind of function can be assigned but
- 1/3 has a good annotation
- 1/3 has an inprecise annotation
- 1/3 has a bad annotation

Curation is needed, but who will do it?

UPSC

Stefan Jansson 2009

Classification

- According to function?
- According to biochemical pathway?
- According to Gene Ontology?

UPSC

Stefan Jansson 2009

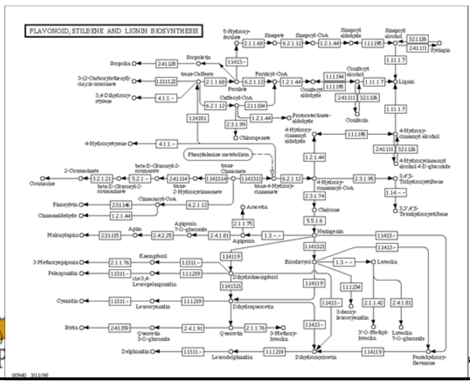
Main classes in MIPS

- Metabolism
- Energy
- Cell growth, division and DNA synthesis
- Transcription
- Protein synthesis
- Protein destination
- Transport facilitation
- Cellular transport
- Cellular communication/signal transd.
- Cell rescue defence, death and aging etc.

UPSC

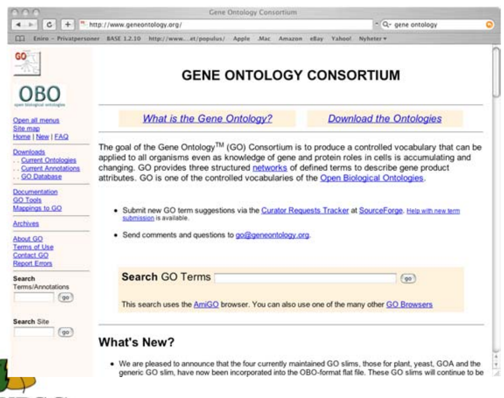
Stefan Jansson 2009

Biochemical pathways - KEGG



UP

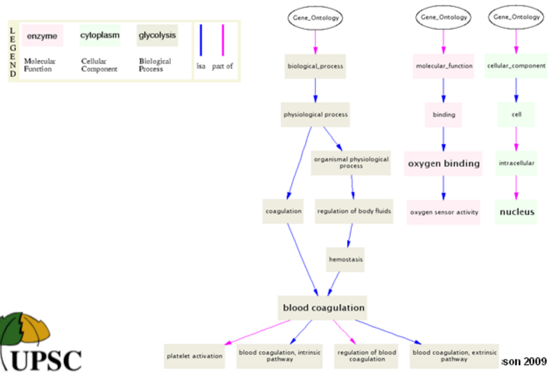
son 2009



UPSC

Stefan Jansson 2009

Gene Ontology



son 2009

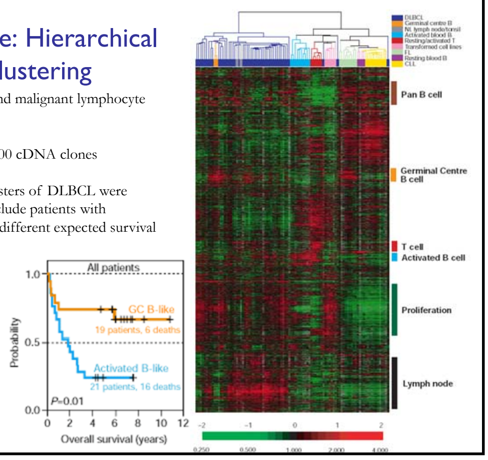
Result visualization

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!

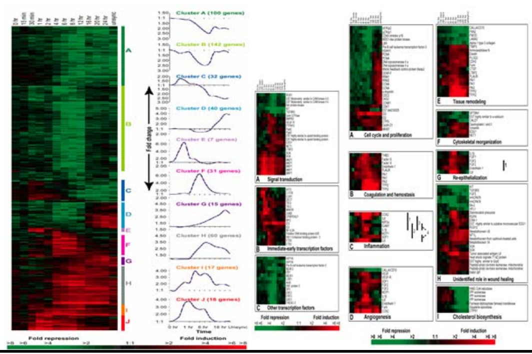


Alizadeh et al., Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling, *Nature*, 403:503-511, 2000.

Example: Hierarchical clustering

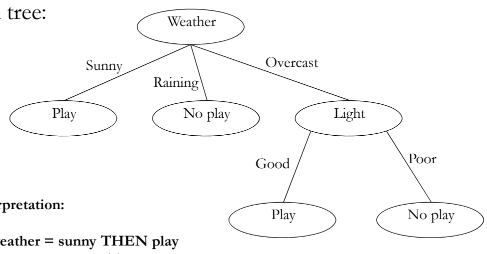
Expression clusters

Functional clusters



Supervised learning: Cricket game

Decision tree:



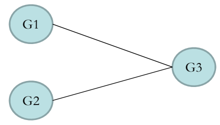
Interpretation:

IF weather = sunny THEN play
IF weather = raining THEN no play
IF weather = overcast AND light = good THEN play
IF weather = overcast AND light = poor THEN no play

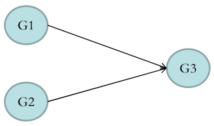
Network representations

- Network: nodes connected by edges
- Nodes represent genes, proteins, metabolites
- Edges represent relationships
  - Co-expression networks: expression correlation
  - Protein-protein networks: proteins form a functional complex
  - 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

Co-expression networks versus gene networks

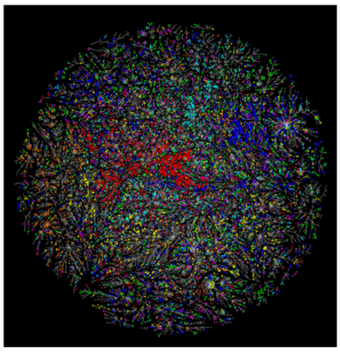


Co-expression network:
Expression of G1 correlates with that of G3
Expression of G2 correlates with that of G3



Gene network:
The expression of G3 can be predicted from that of G1 *and* G2

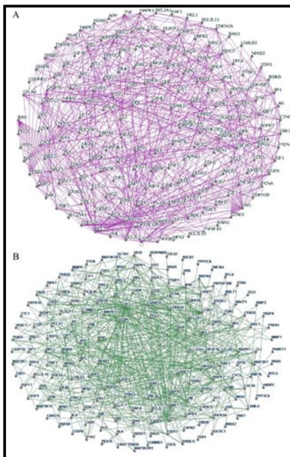
Co-expression network in aspen trees



Based on a UPSC collection of over 1000 cDNS microarrays

A Grönlund, RP Bhalerao, J Karlsson. Modular gene expression in Poplar: a multilayer network approach. New Phytologist, 2009.

Global protein-protein interactions of apoptosis in cancerous and normal cells



- (A) Apoptotic protein-protein interaction network in HeLa cells
- (B) Apoptotic protein-protein interaction network in normal primary lung fibroblasts
- Two-hybrid data sets, four online databases and microarray data

Chu and Chen BMC Systems Biology 2008 2:56

Regulatory network in Arabidopsis

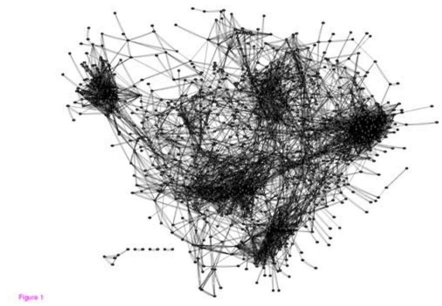
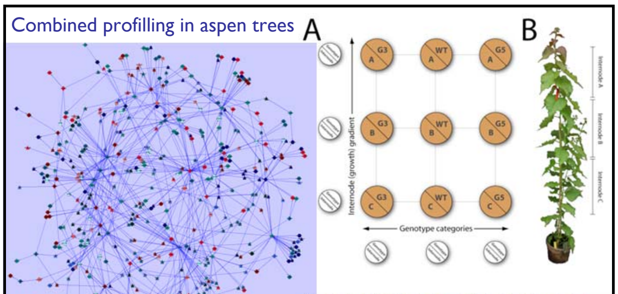
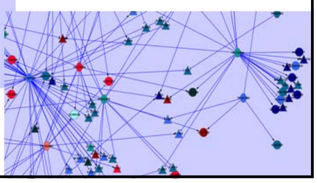


Figure 1

J. Carrera , G. Rodrigo , A. Jaramillo and S. F Elena. Reverse-engineering Arabidopsis thaliana transcriptional network under changing environmental conditions. Genome Biology, 10:R96, 2009.

Combined profilling in aspen trees

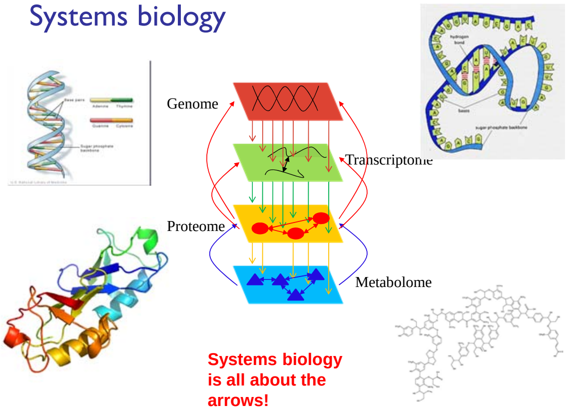




Three platforms to explain regulation in trees:
1. Transcriptomics: **G**enes and **T**ranscription **F**actors
2. Metabolomics: **M**etabolites
3. Proteomics: **P**roteins

Systems biology

Systems biology



Holistic versus reductionistic

- Traditionally:
  - Can biology be reduced to chemistry?
  - Can chemistry be reduced to physics?
- Operationally:
  - Are the assumptions/simplifications in the scientific method reasonable?
  - E.g. can the regulatory mechanism of this cluster be found by considering candidate transcription factors one by one?
  - E.g. can the expression difference between slow and fast growing trees be found by finding (individual) differentially expressed genes?

Does interactions matter?
- AND logics in regulation

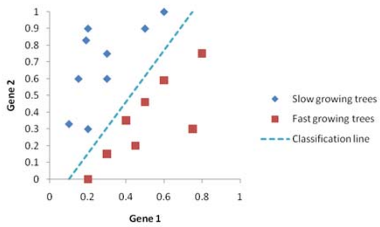
Correlation between the gene and

regulator 1:	0.55 ($P < 0.06$)
regulator 2:	0.65 ($P < 0.02$)

Correlation between gene and model prediction:

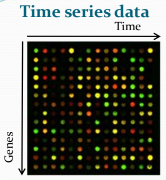
Linear model:	0.77 ($P < 0.003$)
Non-linear model:	0.91 ($P < 1.2E-05$)

Does interactions matter?
- differential expression



Inferring regulatory mechanism

Time series data

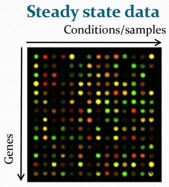


For each gene i :

$$\frac{dy_i}{dt} = \alpha_i - \partial_i y_i + \sum_j \beta_{ij} y_j$$

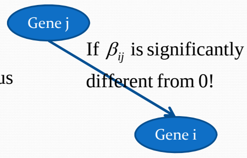
where α_i is its transcription rate,
∂_i the degradation coefficient,
and β_{ij} is the regulatory effect that gene j has on gene i.

Steady state data



$\frac{dy_i}{dt} = 0$ and $\partial_i = 1$, thus

$$\boxed{y_i = \alpha_i + \sum_j \beta_{ij} y_j}$$



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

Correlation: 0.78

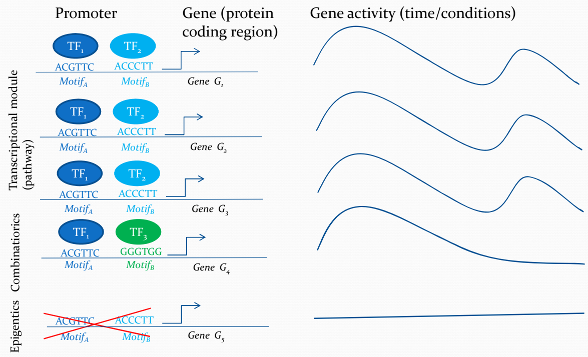
Choose α, β_{12} and β_{13} so that the correlation between observed (y_1) and predicted (y_1 predicted) expression is maximized!

Two types of networks inferred from expression data

- Gene networks: describe the effect that genes have on the expression of one gene (direct or indirect regulation)
- Regulatory network: describe transcription factors regulating genes by binding DNA motifs in the promoter region (physical regulation)

- Gene networks cannot distinguish direct and indirect effect (e.g. the framework on the two previous slides)
- Regulatory networks describe causality: need to incooperate promoter information and knowledge of transcription factors

Gene regulatory networks



Linear versus non-linear models

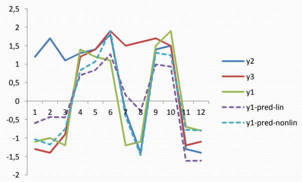
- Linear model: $y_1 = \alpha + \beta_{12} y_2 + \beta_{13} y_3$

- 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

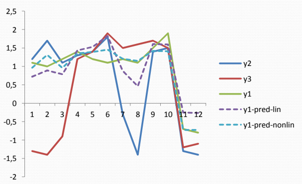


Linear model:
α = -0.46
β_{12}= 0.43
β_{13}=0.50

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
Non-linear model: 0.91
Correlation between gene 1 and
gene 2: 0.55
gene 3: 0.65

OR - logic

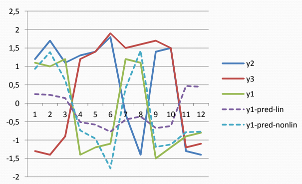


Linear model:
α = 0.59
β_{12}= 0.40
β_{13}=0.27

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
Non-linear model: 0.96
Correlation between gene 1 and
gene 2: 0.72
gene 3: 0.60

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
Non-linear model: 0.92
Correlation between gene 1 and
gene 2: -0.19
gene 3: -0.39

Overfitting and the course of dimensionality

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

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

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

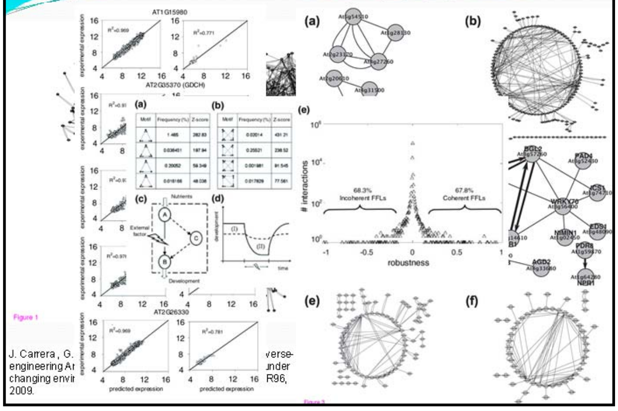
Has many solutions: z=3, x=-3.5, y=-0.5
z=6, x=-7, y=-1
...

i.e. we need more samples than genes in order to solve:

$$\boxed{y_i = \alpha_i + \sum_j \beta_{ij} y_j}$$

there are ~45 000 genes in *Populus* ...
and even ~2500 transcription factors ...

Regulatory network of Arabidopsis



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