Regulatory Sequence Analysis

Classifying genes on the basis of regulatory signals

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Questions

- Let us assume that we dispose of some information about regulatory signals in a set of genes:
  - experimentally measured signals (e.g. databases like TRANSFAC, RegulonDB, SCPD, ...)
  - putative signals predicted by pattern matching
- On this basis, is it possible to predict the regulation of a gene on the basis of its upstream sequence ?
  - Given the low information content of TF binding sites, a consensus motif is expected to be found by chance in many locations.
  - The presence of a single signal is thus generally not sufficient to predict gene regulation.
- However, we can take the multiple motifs into account
  - Multiple occurrences of binding sites for the same TF
  - Binding sites for distinct factors.

Supervised versus non-supervised classification

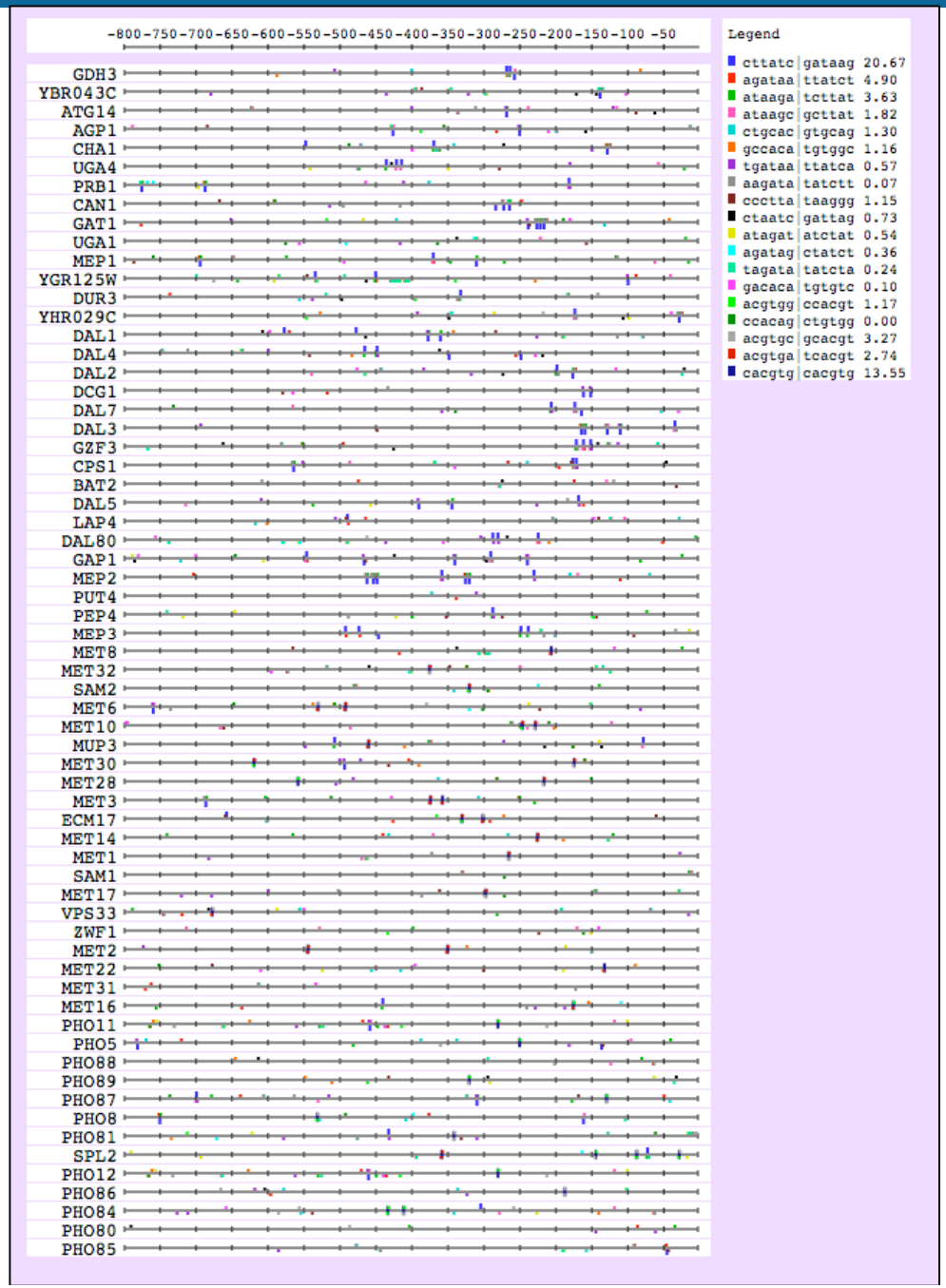
- Approach
  - detect occurrences of one or (preferably) several patterns
  - regroup the matching scores in a multivariate data table
  - apply classification algorithms
- Model system: classification of genes regulated by
  - nitrogen (NIT)
  - methionine (MET)
  - phosphate (PHO)
  - + a set of control sequences, generated randomly
- Two situations
  - We have no a priori idea about functional classes of genes
    → unsupervised classification (clustering)
  - We want to classify genes according to some pre-defined classes, for which we have some training examples
    → supervised classification (discrimination)

Data - pattern counts

- 94 sequences
  - NIT (31 upstream sequences); PHO (13 upstream sequences); MET (20 upstream sequences); RAND (30 random sequences Markov 5)
- 44 patterns
  - Hexanucleotides and dyads involved in the regulation of the MET, PHO and NIT genes.
  - Some of these patterns are very well conserved in the core of the binding site (e.g. CACGTG, CACGTT, ...) whereas some other represent partial conservation of the flanking nucleotides (e.g. ACGTGg, ACGTTt, ...).
  - The data is presented in a multi-variate table, with one row per gene, and one column per pattern.

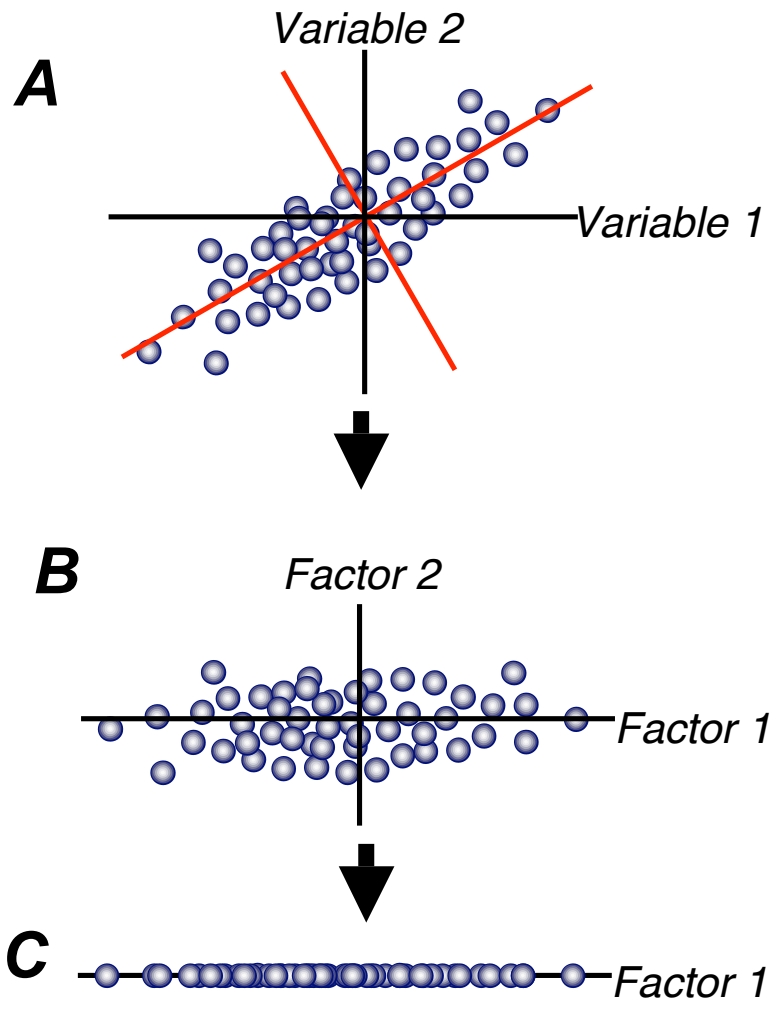
Gene	aaacgt|acgttt	aacgtg|cacgtt	aacn{1}gtg|cacn{1}gtt	aactgt|acagtt	acan{14}tgc|gcan{14}tgt	acan{15}gca|tgcn{15}tgt	acan{6}gca|tgcn{6}tgt	acatct|agatgt	acgn{1}gcg|cgcn{1}cgt	acgn{6}agc|gctn{6}cgt	acgtga|tcacgt	acgtgc|gcacgt	acgtgg|ccacgt	actgtg|cacagt	agataa|ttatct	ataaga|tcttat	atcacg|cgtgat	cacgcc|ggcgtg	cacgtg|cacgtg	cacn{15}ggc|gccn{15}gtg	cacn{1}tga|tcan{1}gtg	cacn{2}gac|gtcn{2}gtg	cagn{2}cgg|ccgn{2}ctg	cagn{7}atc|gatn{7}ctg	ccacag|ctgtgg	cccacg|cgtggg	cccatc|gatggg	ccgcgc|gcgcgg
GDH3	0	0	0	4	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	2	0	0	0	0	0	0
YBR043C	2	0	0	0	0	0	0	2	0	2	0	0	0	2	4	2	0	2	0	0	0	0	0	4	0	0	2	0
APG14	0	0	0	4	0	0	0	0	0	0	0	0	0	0	4	2	0	0	0	0	0	2	0	0	0	0	0	0
AGP1	0	0	2	2	0	0	0	0	0	0	0	0	2	2	4	2	2	0	0	0	0	0	0	2	2	4	0	0
CHA1	0	0	2	2	0	0	4	0	0	2	0	0	0	0	6	6	2	0	0	0	0	0	0	2	2	0	0	2
UGA4	4	0	0	0	0	0	0	2	0	0	0	2	0	2	0	6	0	0	0	0	0	0	0	0	0	0	0	0
PRB1	0	0	2	0	0	0	0	0	0	0	0	0	0	2	4	4	2	2	0	0	0	4	0	0	0	0	0	0
CAN1	0	2	0	0	0	0	2	0	0	4	0	0	0	0	6	2	0	0	0	0	0	0	0	0	2	0	0	0
GAT1	0	0	0	0	2	0	0	0	0	2	2	0	2	0	8	6	0	0	0	0	0	2	2	0	2	0	0	0
UGA1	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	2	0	0	0	2	4	0	0	0
MEP1	2	0	0	0	0	2	0	6	0	0	0	0	0	0	6	14	0	0	0	0	2	0	4	4	2	0	0	0

Feature map of pattern occurrences



The feature map gives the impression that some features are more or less concentrated in some sequence groups, but the separation is not trivial.

Principal component analysis

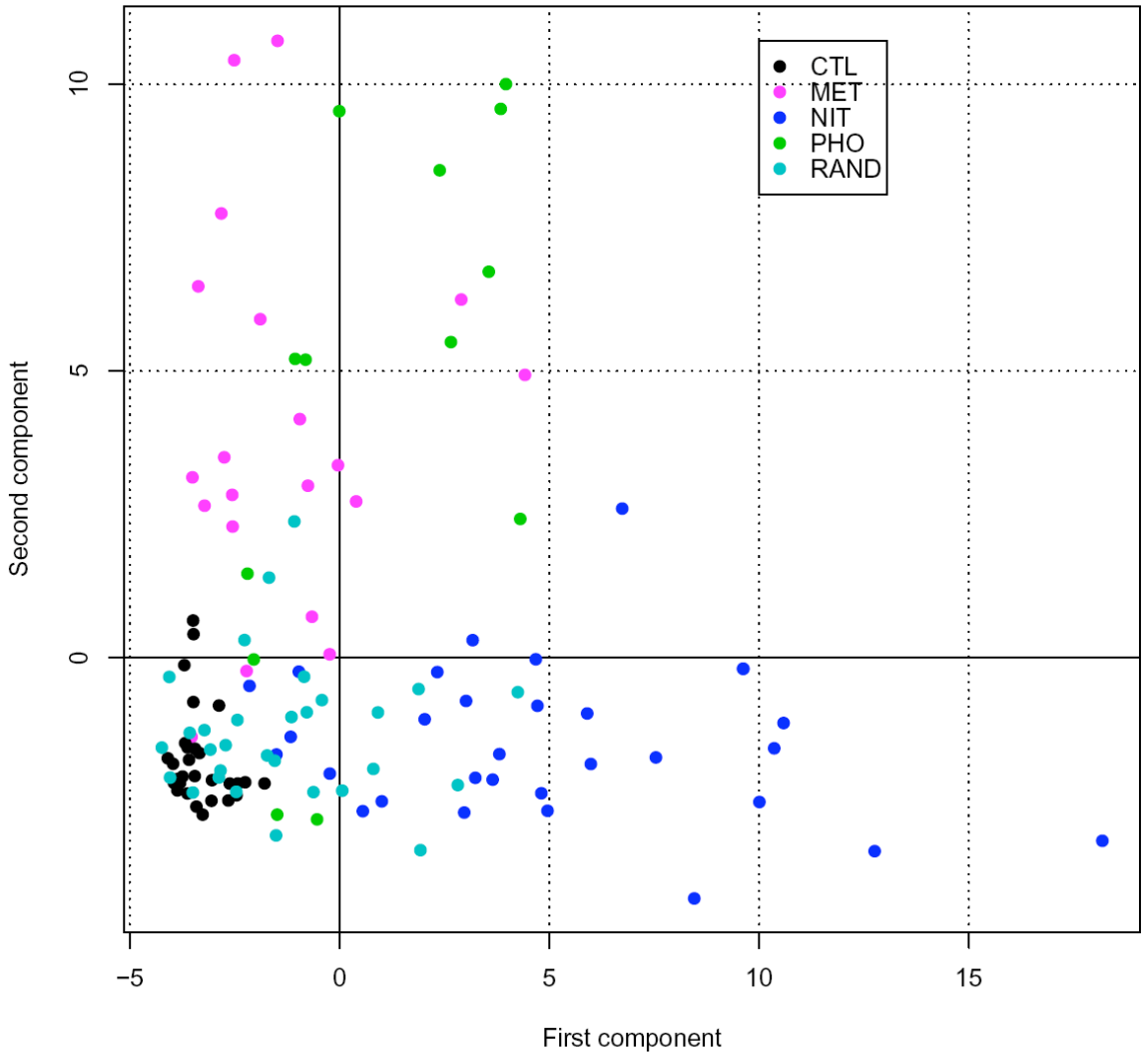


- Multidimensional data
  - n objects, p variables (in this case p=2)
- Principal components
  - n objects, p factors
  - Each component (factor) is a linear combination of variables
- Reduction in dimensions
  - Selection of a subset of principal components
  - q factors, with q < p (in this case, q=1)

Gilbert, D., Schroeder, M. & van Helden, J. (2000). *Trends in Biotechnology* **18**) 487-495.

Principal component plot

- Projection of the 44 dimensions onto a 2D space (Principal Component Analysis)
- Each dot represents one sequence.
- The dimensions represent the first and second components, respectively.
- Note that PCA is suboptimal: the axes with highest variance are not always the most discriminant.



Regulatory Sequence Analysis

Unsupervised classification (clustering)

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Unsupervised classification

- In a first stage, we will apply an unsupervised classification (clustering), i.e. we have no a priori idea about the functional classes.
- For this, we need to choose
  - a clustering algorithm
  - a similarity/dissimilarity metric

Clustering algorithm

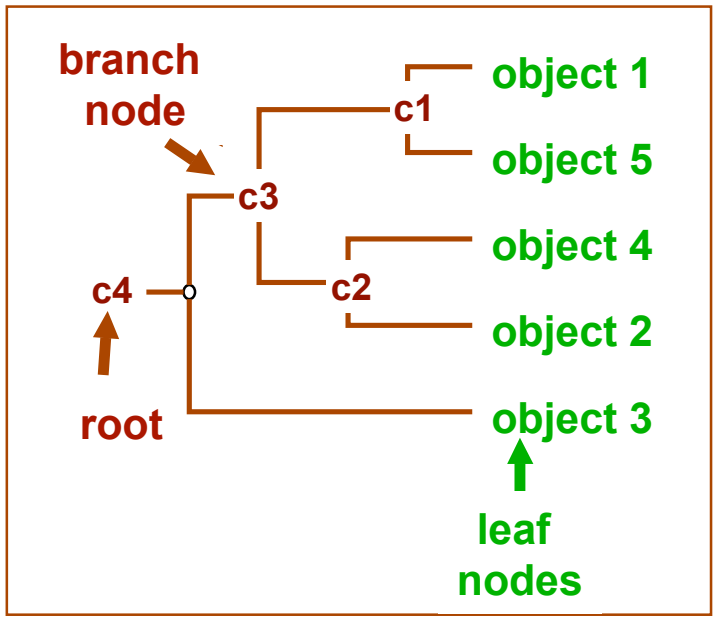
- There is a large variety of clustering algorithms
  - hierarchical
  - k-means
  - self-organizing maps
  - k-nearest neighbours
  - genetic algorithms
  - ...
- In this study, we only applied hierarchical clustering.
  - Besides the choice of the similarity metrics, hierarchical clustering requires to choose an agglomeration rule.

Hierarchical clustering

Distance matrix

	object 1	object 2	object 3	object 4	object 5
object 1	0.00	4.00	6.00	3.50	1.00
object 2	4.00	0.00	6.00	2.00	4.50
object 3	6.00	6.00	0.00	5.50	6.50
object 4	3.50	2.00	5.50	0.00	4.00
object 5	1.00	4.50	6.50	4.00	0.00

Tree representation



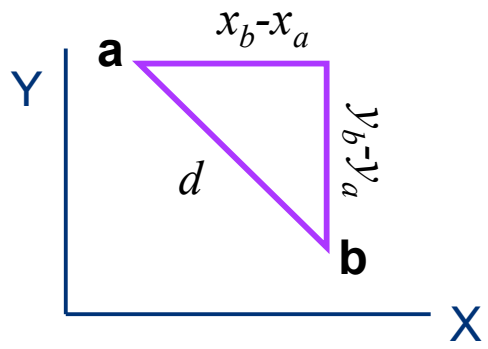
- Hierarchical clustering is an aggregative clustering method
- One needs to define a (dis)similarity metric between two groups. There are several possibilities
  - **Average linkage**: the average distance between objects from groups A and B
  - **Single linkage**: the distance between the closest objects from groups A and B
  - **Complete linkage**: the distance between the most distant objects from groups A and B
  - **Ward clustering**: the dissimilarity between two groups is estimated by the moment of inertia of their elements from the gravity center.
- Algorithm
  - (1) Assign each object to a separate cluster.
  - (2) Find the pair of clusters with the shortest distance, and regroup them in a single cluster
  - (3) Repeat (2) until there is a single cluster
- The result is a tree, whose intermediate nodes represent clusters
  - `N objects → N-1 intermediate nodes
- Branch lengths represent distances between clusters

Classical similarity/dissimilarity metrics

- There are many similarity or dissimilarity metrics, and the choice among them influences drastically the result of the classification.
- Some classical metrics
  - Manhattan distance (=city block distance)
  - Euclidian distance
  - Minkowski metrics
  - correlation coefficient
  - Mahalanobis distance
  - Canberra distance
  - Binary distance
  - chi-square

Euclidian distance

- You are probably familiar with the calculation of Euclidian distance in a 2-dimensional space



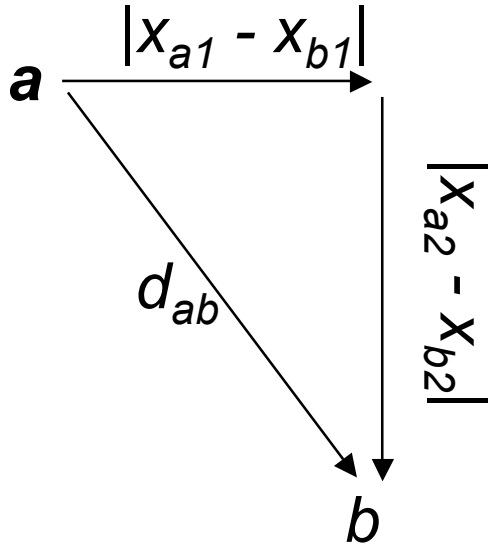
$$d_E = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

- The concept naturally extends to spaces with higher dimension (*p*-dimensional space)

$$d_E = \sqrt{\sum_{i=1}^{p} (x_{ai} - x_{bi})^2}$$

- Two typical applications
  - The distance between genes is calculated in the space of conditions (chips)
  - The distance between tissue types is calculated in the space of genes (spot)

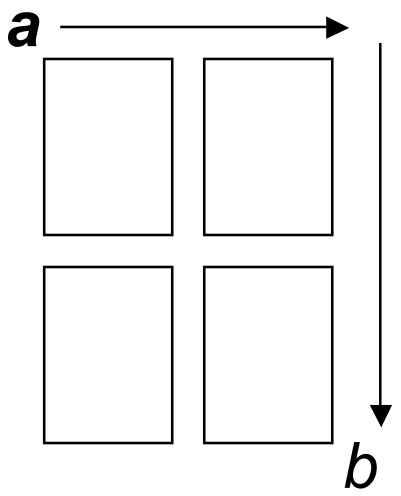
Generalized Euclidian distance



$$D_{ab} = \sqrt{\sum_{i=1}^{p} w_i^2 \left(x_{ai} - x_{bi}\right)^2}$$

- The generalized Euclidian distance between two points is calculated as the Euclidian distance, with a specific weight w_i associated to each dimension i
  - *a,b* two points in the multi-variate space
  - *p* number of dimensions
  - w_i weight if the i^{th} dimension

Manhattan distance



$$D_{ab} = \sum_{i=1}^{p} w_i \left| x_{ai} - x_{bi} \right|$$

- The Manhattan distance between points a and b is the weighted sum of the absolute differences in each dimension.
  - *a,b* two points in the multi-variate space
  - *p* number of dimensions
  - w_i weight if the i^{th} dimension

Minkowski metrics

$$D_{ab} = \sqrt[\lambda]{\sum_{i=1}^{p} w_i^{\lambda} \left| x_{ai} - x_{bi} \right|^{\lambda}}$$

- The Minkowski metrics are a family of dissimilarity metrics, which can be tuned by a parameter (λ).
- This is a generalization, which includes
  - λ=1 Manhattan distance
  - λ=2 Euclidian distance

Correlation coefficient

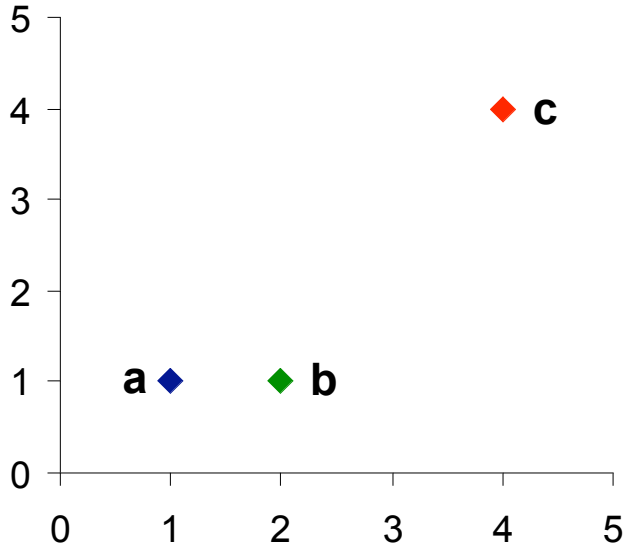
$$S_{ab} = \frac{\sum_{i=1}^{p}\left(x_{ai} - \bar{x}_{a.}\right)\left(x_{bi} - \bar{x}_{b.}\right)}{\sqrt{SSD_a SSD_b}} = \frac{1}{p}\sum_{i=1}^{p} z_{ai} z_{bi}$$

$$\bar{x}_{a.} = \frac{1}{p}\sum_{i=1}^{p} x_{ai} \qquad \bar{x}_{b.} = \frac{1}{p}\sum_{i=1}^{p} x_{bi}$$

$$SSD_a = \sum_{i=1}^{p}\left(x_{ai} - \bar{x}_{a.}\right)^2 \qquad SSD_b = \sum_{i=1}^{p}\left(x_{bi} - \bar{x}_{b.}\right)^2$$

Impact of the distance metrics

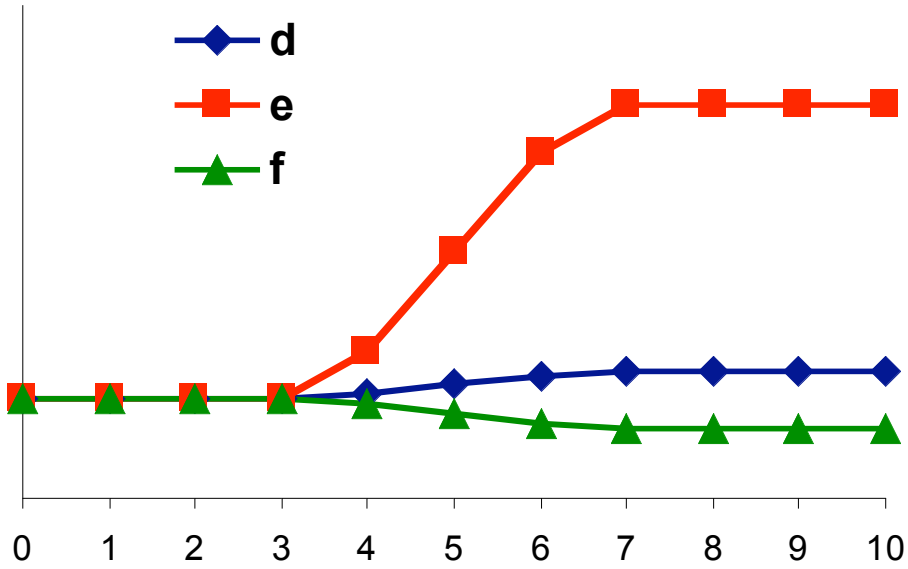
A



Euclidian distances
- **a** close to **b**

Correlation coefficient
- **a** close to **c**

B



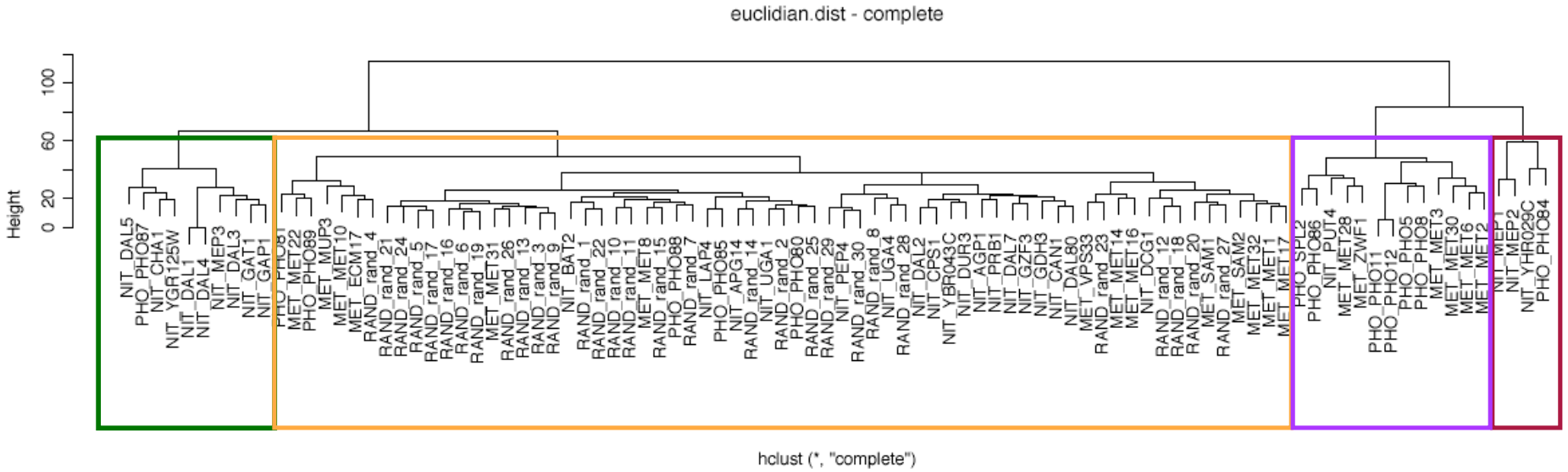
Euclidian distances
- **d** close to **f**

Correlation coefficient
- **d** close to **e**

Comparing sequences on the basis of pattern counts

- The metrics should be appropriate to reflect the characteristics of transcriptional regulation
  - Multiple occurrences of a signal increase the response (e.g. GATA-boxes).
  - Some pathways are regulated by distinct factors (e.g. methionine biosynthesis).
  - Some patterns are more informative than others.
- We need a metric which takes into account the following aspects :
  - count-based comparison (the number of copies of each pattern should be reflected);
  - multi-variate comparison (several distinct patterns are considered);
  - pattern-specific prior probabilities (some patterns are expected to occur by chance more frequently than others).

Clustering - Euclidian distance



- Sequence clustering on the basis of pattern counts
- Distance metric: Euclidian
- Clustering method: UPGMA (complete)
- The four main clusters do not correspond to the prior functional classes
- Genes from different classes are intermingled

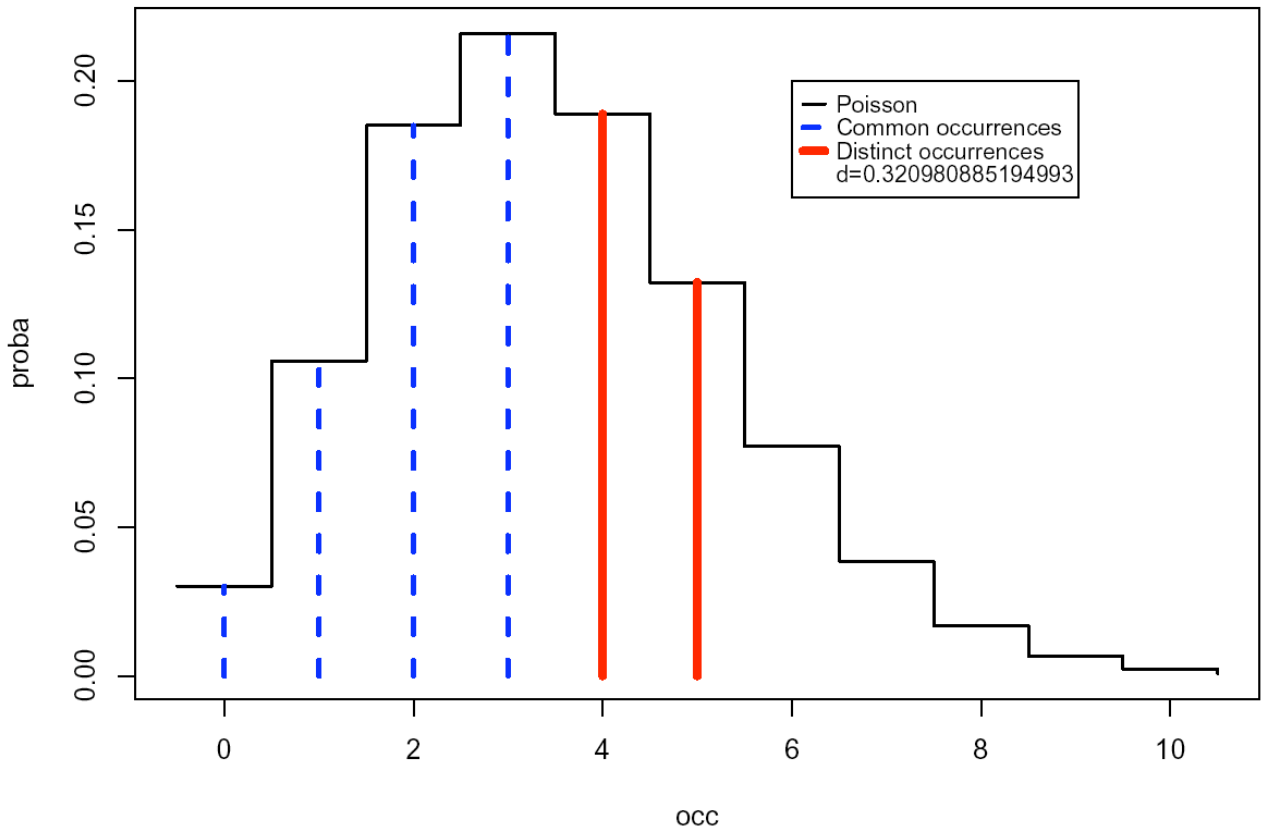
		Known				
		RAND	MET	NIT	PHO	SUM
Predicted	RAND	30	14	18	5	67
	MET	0	6	1	6	13
	NIT	0	0	9	1	10
	PHO	0	0	3	1	4
	SUM	30	20	31	13	94
	Errors	48	51.1%			
	Correct	46	48.9%			

Measuring distance between pattern counts

- The classical distance metrics are not appropriate for measuring distances between sequences en the basis of pattern counts. They fail to represent some of the biological features of these motifs.
- Differential weighting of distinct motifs
  - Some of the classical metrics (correlation, chi squared, …) assign the same weight to each variable.
  - For the other ones (Euclidian, Manhattan, Minkowski, …), a weight can be defined for each variabe (motif), but how to select it ? (by default, all w_i are set to 1)
- Differential weighting of multiple occurrences
  - For a given pattern, the probability is not a linear function of the number of occurrences.
  - Intuitively, it seems reasonable to consider that the difference between 0 and 2 occurrences of a biological signal has not the same impact as the difference between 2 and 4 occurrences. However, for all the metrics described before, the difference between 0 and 2 is the same as the difference between 2 and 4.

Poisson-based similarity and dissimilarity

- Let us take a simple example:
  - Sequence a contains 3 occurrences of a motif
  - Sequence b contains 5 occurrences of the same motif
- We have thus 3 common and 2 distinct occurrences.



Poisson-based similarity metric

- The probability to observe at least *x* occurrences of pattern *i* in common between sequences *a* and *b* is the joint probability of observing at least *x* occurrences in sequence *a* and at least *x* occurrences in sequence *b*.

$$C_i^{ab} > 0 \qquad P(x \geq C_i^{ab}) = \left[1 - F\left(C_i^{ab} - 1, m_i\right)\right]^2$$

$$C_i^{ab} = 0 \qquad P(x \geq C_i^{ab}) = 1$$

- Lower probabilities correspond to higher similarities. The probability of common occurrences can be converted in a similarity metrics.

$$s_i^{ab} = 1 - P(x \geq C_i^{ab})$$

Multi-variate Poisson-based similarity

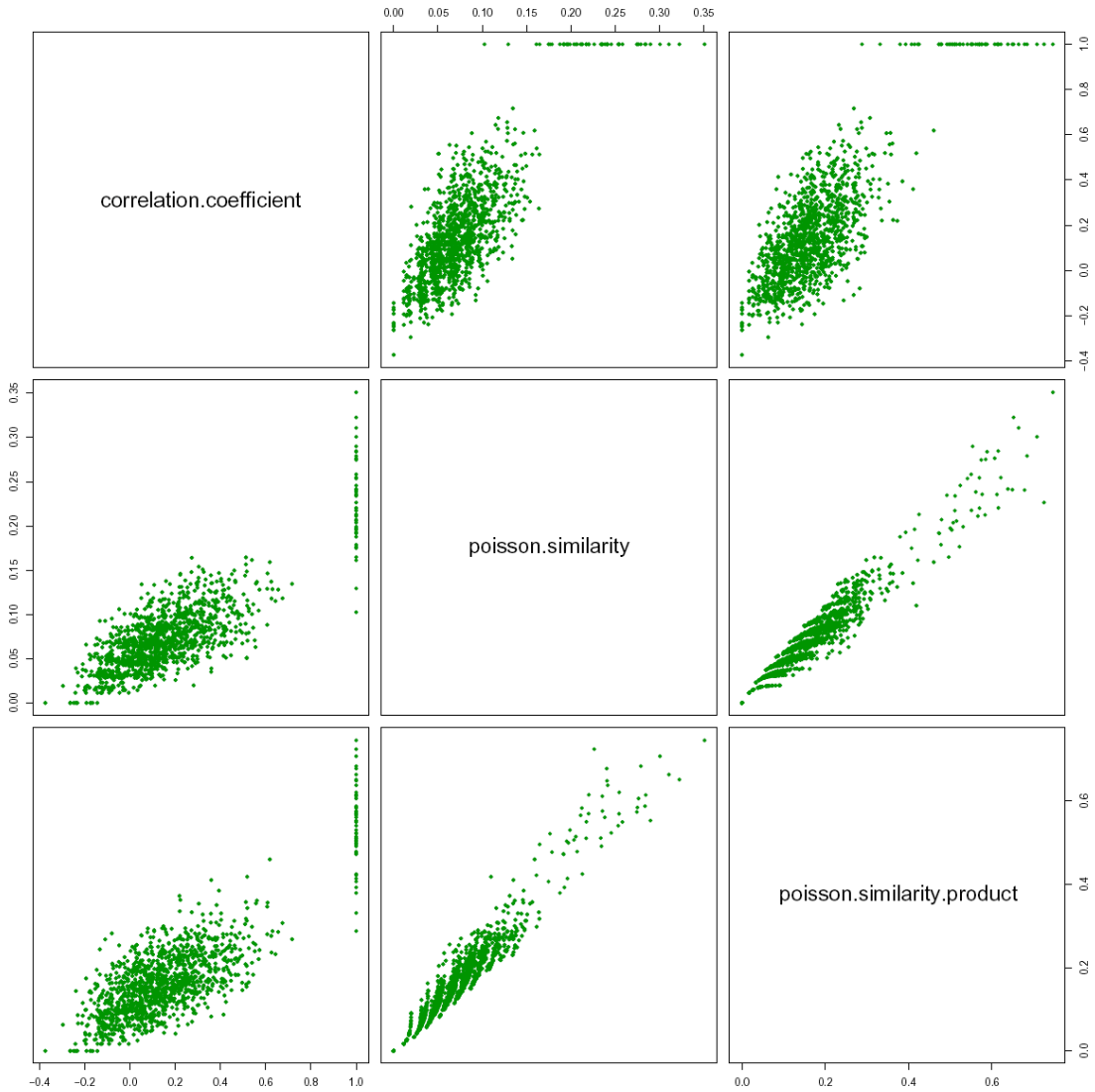
- A multi-variate similarity metric can be calculated as the average of single-variate metrics :

$$S_{add}^{ab} = \frac{1}{p}\sum_{i=1}^{p} s_i^{ab}$$

  - Alternatively, one can consider the geometric mean, which reflects the joint probability of common occurrences for the different patterns :

$$S_{prod}^{ab} = 1 - \sqrt[p]{\prod_{i=1}^{p} P(x \geq C_i^{ab})}$$

Scoring sequences with equal number of occurrences



Metric comparison, with random sequences and random patterns

- If two sequences have exactly the same number of occurrences, classical similarity metrics do not indicate ***how much*** they are similar.
  - If $N^a = N^b$, the correlation coefficient = 1, irrespective of the actual number of occurrences found in common
- On the contrary, Poisson-based similarity metrics assign a higher score to two sequences if they both have 6 occurrences of the motif than it they both have 1 occurrence.
- In addition, the score will be higher if the pattern is rare than if it is frequent.
- The Poisson similarity is thus likely to better reflect the biological properties of the regulatory signals.

Properties of the Poisson-based similarity

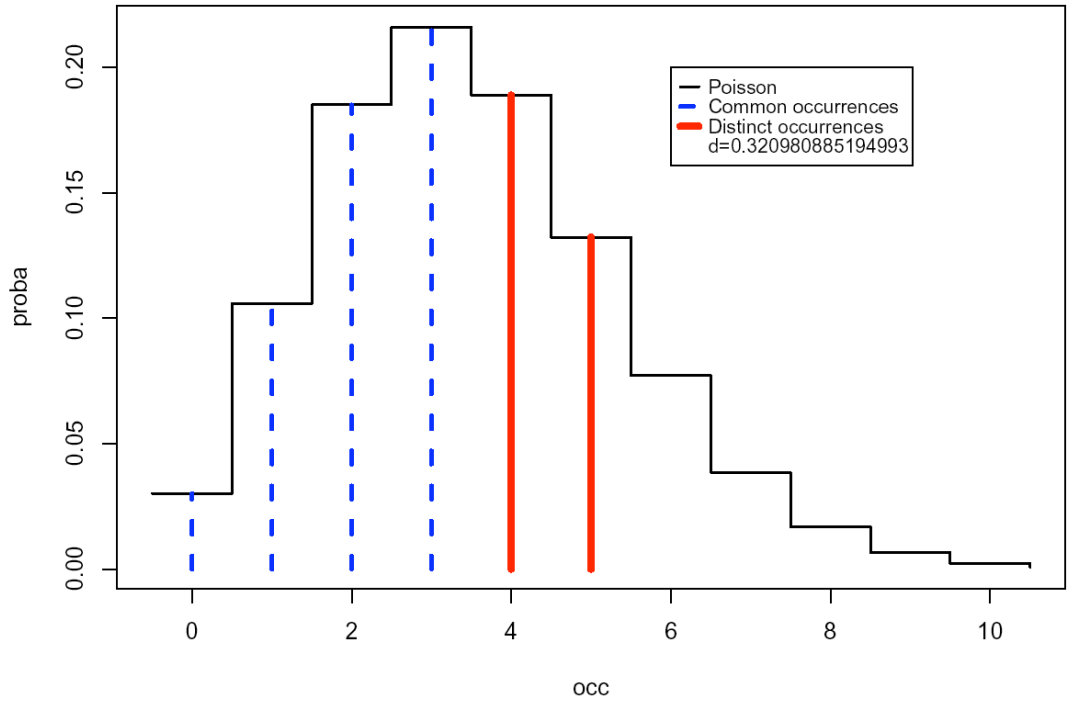
- The Poisson-based similarity metrics fit well with the intuitive concept of similarity, in terms of pattern occurrences :
  - If two sequences do not have a single common site, their similarity is 0.
  - The score increases when multiple copies of a given pattern are found in both sequences.
  - The score increases when several patterns are common to both sequences.
  - Patterns with low prior probabilities contribute more than those with high prior probabilities.
- However, these metrics are based on the counts of common occurrences only, and do not reflect the differences, since occurrences found in gene a but not in gene b do not affect the score.

Poisson-based dissimilarity

- A Poisson-based dissimilarity metric can be defined, by calculating the sum of probabilities of distinct occurrences, i.e. occurrences found in one sequence but not the other one.

$$d^{ab}_{distinct_i} = \left| F(N^b_i, m_i) - F(N^a_i, m_i) \right|$$

$$D^{ab}_{distinct} = \frac{1}{p} \sum_{i=1}^{p} d^{ab}_i$$

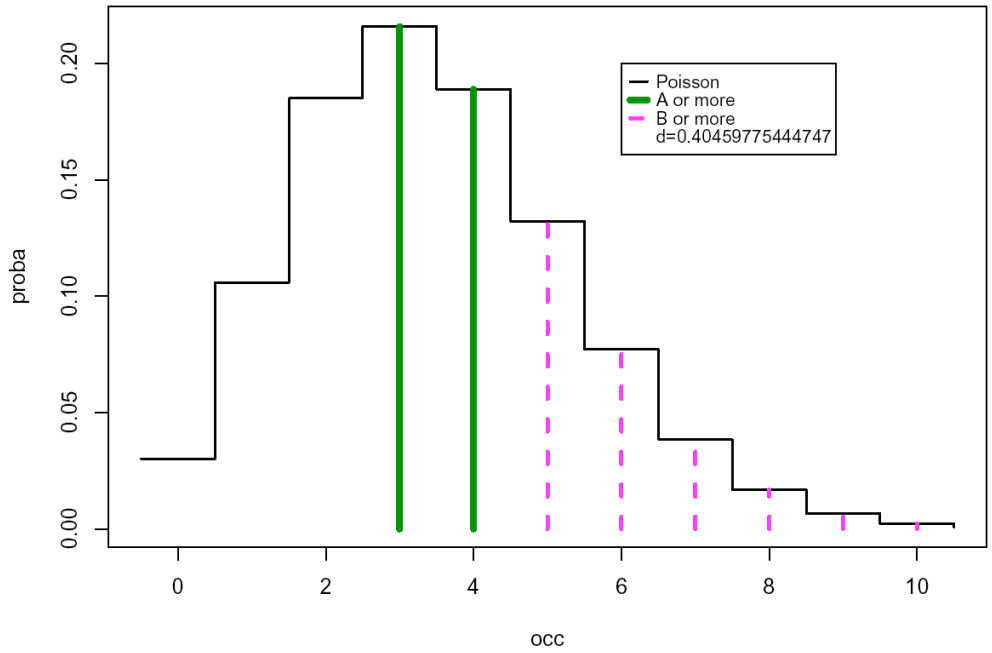


Poisson-based dissimilarity based on over-representation

- Rather than occurrence probabilities (right tail of the Poisson distribution), one could consider their P-values (left tail), which represent their degree of over-representation.
- The dissimilarity metrics can be calculated as the difference between the left tails of occurrences found in sequences a and b, respectively.

$$d_{over_i}^{ab} = \left| P(x \geq N_i^a) - P(x \geq N_i^b) \right| = \left| F(N_i^b - 1, m_i) - F(N_i^a - 1, m_i) \right|$$

$$D_{over}^{ab} = \frac{1}{p} \sum_{i=1}^{p} d_i^{ab}$$



Comparison of the metrics with random sequences

	correlation.coefficient	canberra.distance	binary.distance	poisson.mixed.over	park.similarity	poisson.similarity	poisson.similarity.product	poisson.mixed.distinct.product	poisson.mixed.over.product	poisson.dissimilarity.distinct	poisson.dissimilarity.over	manhattan.distance	euclidian.distance
correlation.coefficient		-0.76	-0.81	-0.80	0.79	0.72	-0.71	0.60	0.28	-0.72	-0.67	-0.56	-0.56
canberra.distance	-0.76		**0.98**	0.91	-0.89	-0.85	0.82	-0.71	-0.35	0.76	0.74	0.62	0.60
binary.distance	-0.81	**0.98**		0.91	-0.89	-0.88	0.84	-0.73	-0.38	0.75	0.72	0.57	0.54
poisson.mixed.over	-0.80	0.91	0.91		**-0.98**	-0.76	0.74	-0.58	-0.12	0.89	0.92	0.78	0.68
park.similarity	0.79	-0.89	-0.89	**-0.98**		0.76	-0.74	0.60	0.15	-0.85	-0.89	-0.79	-0.72
poisson.similarity	0.72	-0.85	-0.88	-0.76	0.76		**-0.96**	0.92	0.71	-0.50	-0.45	-0.29	-0.29
poisson.similarity.product	-0.71	0.82	0.84	0.74	-0.74	**-0.96**		**-0.98**	-0.76	0.46	0.43	0.28	0.30
poisson.mixed.distinct.product	0.60	-0.71	-0.73	-0.58	0.60	0.92	**-0.98**		0.87	-0.25	-0.24	-0.08	-0.13
poisson.mixed.over.product	0.28	-0.35	-0.38	-0.12	0.15	0.71	-0.76	0.87		0.18	0.26	0.36	0.24
poisson.dissimilarity.distinct	-0.72	0.76	0.75	0.89	-0.85	-0.50	0.46	-0.25	0.18		0.93	0.89	0.78
poisson.dissimilarity.over	-0.67	0.74	0.72	0.92	-0.89	-0.45	0.43	-0.24	0.26	0.93		0.91	0.76
manhattan.distance	-0.56	0.62	0.57	0.78	-0.79	-0.29	0.28	-0.08	0.36	0.89	0.91		0.94
euclidian.distance	-0.56	0.60	0.54	0.68	-0.72	-0.29	0.30	-0.13	0.24	0.78	0.76	0.94	

Evaluation of the biological relevance

- After having defined the different variants of Poisson-based metrics, we can apply them to our biological example (NIT, PHO, MET and RAND genes), in order to evaluate their respective capability to classify genes according to their regulation.
- For each metric and for each agglomeration rule (single, average, complete, Ward)
  - Apply hierarchical clustering.
  - Prune the tree to select the 4 topmost branches.
  - Compare the 4 topmost branches with the prior class (NIT, MET, PHO or RAND), and calculate the hit rate (% of genes assigned to the correct class).
- Sort the results according to the hit rate.

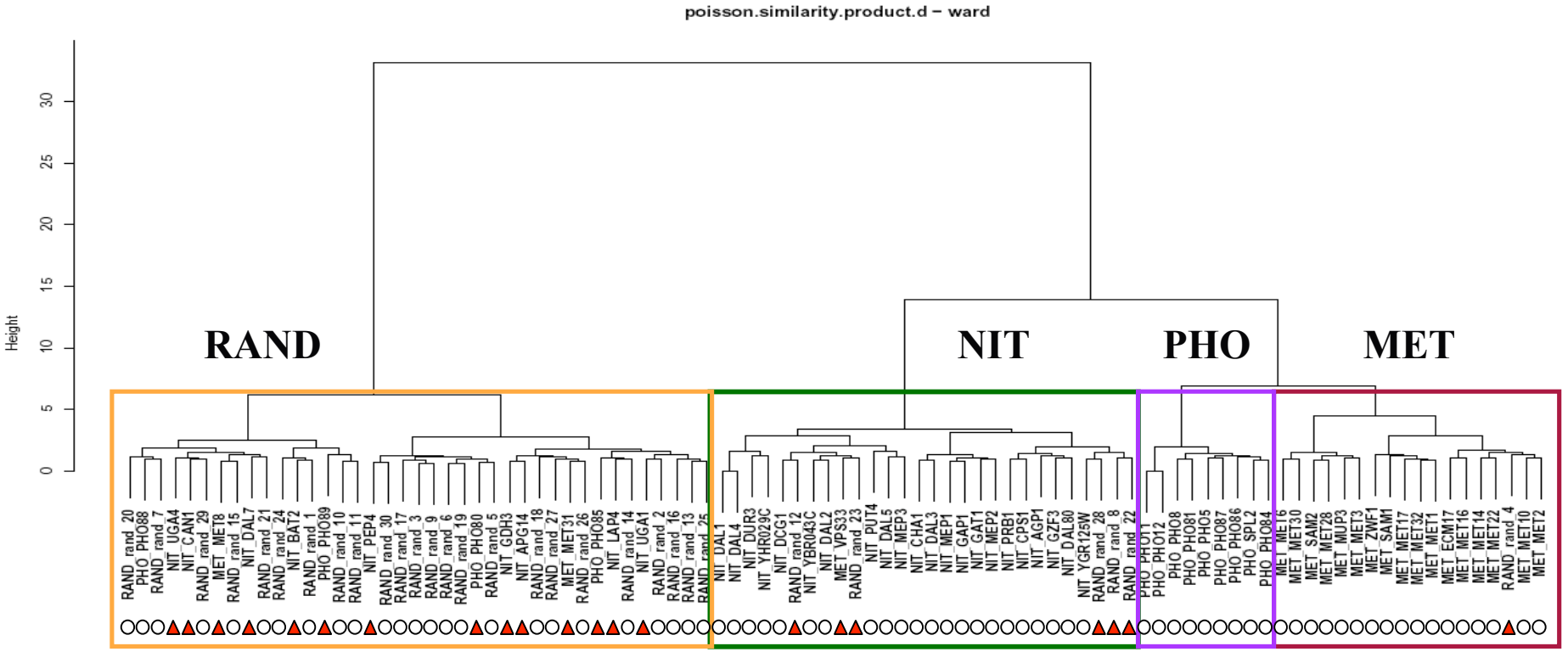
Results with MET-PHO-NIT genes + random sequences

metric	model	clustering method	MET >NIT	NIT >NIT	PHO >NIT	RAND >NIT	MET >MET	NIT >MET	PHO >MET	RAND >MET	MET >PHO	NIT >PHO	PHO >PHO	RAND >PHO	MET>RAND	NIT > RAND	PHO > RAND	RAND > RAND	EXTERNAL	TRUE	FALSE	hit rate
poisson.similarity.product.d	product	ward	1	22	0	5	17	0	0	1	0	0	9	0	2	9	4	24		72	22	76.6%
poisson.mixed.distinct.d	additive	complete	0	25	0	5	10	0	0	0	0	0	8	0	10	6	5	25		68	26	72.3%
poisson.mixed.over.d	additive	ward	3	21	1	7	15	1	1	1	0	0	8	0	2	9	3	22		66	28	70.2%
poisson.mixed.distinct.d	additive	ward	0	19	0	5	16	1	2	3	0	0	8	0	4	11	3	22		65	29	69.1%
poisson.mixed.distinct.product.d	product	complete	0	15	0	0	12	5	1	1	5	0	8	0	3	11	4	29		64	30	68.1%
poisson.similarity.d	additive	ward	4	25	2	14	15	1	1	0	0	0	8	0	1	5	2	16		64	30	68.1%
poisson.dissimilarity.distinct	additive	ward	4	22	2	12	15	0	1	0	0	0	8	0	1	9	2	18		63	31	67.0%
poisson.mixed.over.product.d	product	ward	3	23	1	9	11	0	0	1	3	0	9	2	3	8	3	18		61	33	64.9%
correlation.coefficient.d	additive	ward	1	26	2	8	10	1	1	2	1	1	8	6	8	3	2	14		58	36	61.7%
poisson.dissimilarity.over	additive	ward	9	16	1	4	8	0	1	0	0	0	8	0	3	15	3	26		58	36	61.7%
poisson.similarity.d	additive	complete	13	25	2	13	6	0	0	0	0	0	8	0	1	6	3	17		56	38	59.6%
correlation.coefficient.d	additive	complete	0	26	3	11	9	1	0	2	1	1	8	5	10	3	2	12		55	39	58.5%
poisson.similarity.product.d	product	complete	0	12	0	0	12	10	1	9	6	0	9	0	2	9	3	21		54	40	57.4%
manhattan.dist	additive	ward	0	10	0	0	10	13	3	4	8	0	7	0	2	8	3	26		53	41	56.4%
park.similarity.d (pruning 5)	additive	ward	4	14	2	9	11	1	0	2	0	0	8	0	1	15	2	19	6	52	42	55.3%
poisson.dissimilarity.over	additive	complete	5	31	4	29	13	0	1	1	0	0	8	0	2	0	0	0		52	42	55.3%
poisson.mixed.distinct.product.d	product	ward	0	20	0	1	14	0	9	0	5	7	2	14	1	4	2	15		51	43	54.3%
poisson.dissimilarity.distinct	additive	complete	5	31	4	30	11	0	1	0	0	0	8	0	4	0	0	0		50	44	53.2%
poisson.mixed.over.d	additive	complete	3	28	3	26	13	1	1	1	2	2	9	3	2	0	0	0		50	44	53.2%
park.similarity.d	additive	ward	5	29	4	28	11	1	0	2	0	0	8	0	4	1	1	0		48	46	51.1%
euclidian.dist	additive	ward	0	6	0	0	12	16	3	8	6	4	7	0	2	5	3	22		47	47	50.0%
mahalanobis.dist	additive	complete	6	16	5	12	9	7	0	1	5	7	7	2	0	1	1	15		47	47	50.0%
euclidian.dist	additive	complete	0	9	1	0	6	1	6	0	0	3	1	0	14	18	5	30		46	48	48.9%
mahalanobis.dist	additive	ward	6	15	5	9	11	11	2	5	3	4	5	1	0	1	1	15		46	48	48.9%
park.similarity.d	additive	complete	15	30	4	30	5	1	1	0	0	0	8	0	0	0	1	0		43	52	45.3%
poisson.mixed.over.product.d	product	complete	5	27	9	24	2	4	1	6	0	0	2	0	0	0	1	0		31	50	38.3%
manhattan.dist	additive	complete	12	22	6	30	6	9	1	0	0	0	6	0	2	0	0	0		34	60	36.2%

Clustering - Poisson-based distance metrics

- Metric: Poisson similarity product;
- Clustering: Ward hierarchical.
- Red triangles below the tree indicate errors
- Most errors consist in false negative.

		Known				
		RAND	MET	NIT	PHO	SUM
Predicted	RAND	24	2	9	4	39
	MET	1	17	0	0	18
	NIT	5	1	22	0	28
	PHO	0	0	0	9	9
	SUM	30	20	31	13	94
	Errors	22	23.40%			
	Correct	72	76.60%			



van Helden, J. 2004. Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics 20: 399-406.8

Summary: clustering on the basis of pattern counts

- The choice of the distance metric and clustering method is crucial
- Classical distance metric give very bad results
- Poisson-based metric bring a sensible improvement
- Weaknesses
  - Dependency between variables are not (yet) taken into account
  - This is an unsupervised approach. We could obtain better results by taking advantage of our prior knowledge of the functional classes in order to train a program, which could then be used to classify new genes.
- Reference
  - van Helden, J. (2004). Metrics for comparing regulatory sequences on the basis of pattern counts. Bioinformatics, 2004 20(3):399-406.

Regulatory Sequence Analysis

Supervised classification (discriminant analysis)

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Study case

- Evaluation on a challenging case : given the similarity of Met4p (tCACGTGa) and Pho4p (CACGTGgg or CACGTttt) binding sites, can we distinguish their respective target genes ?
- Genome-scale classification : having at hand the complete yeast genome, can we classify genes according to their predicted regulatory responses ?

Discrimination between MET and PHO genes

- On the basis of upstream motifs, can we predict the regulation of a gene ?
  - Pho4p binds CACGTGgg and CACGTTtt (CACGTkkk)
  - Met4p binds tCACGTGa
  - Met31p binds AAAACTGTGG
- Clues
  - Combinatorial aspect : several MET genes are regulated by both Met4p and Met31p
  - Multiple motifs : many PHO genes have several Pho4p sites
- Approach
  - Build position weight matrices reflecting the specificity of each factor
  - For each upstream region, find the 3 top scores obtained with the different matrices
  - Define a training set with known PHO, MET and control genes
  - Apply discriminant analysis

Met4p binding sites

gene	start	end	sequence
MET3	-367	-349	GAAAAG**TCACGTG**TAATTT
MET3	-384	-366	AAAAGG**TCACGTGA**CCAGA
MET14	-235	-217	CTAATT**TCACGTGA**TCAAT
MET16	-185	-167	ATCATT**TCACGTG**GCTAGT
ECM17	-311	-293	ATTTCA**TCACGTG**CGTATT
ECM17	-339	-321	.TTTGTC**CACGTGA**TATTTC
MET10	-255	-237	.CCACAC**CACGTGA**GCTTAT
MET10	-237	-219	.TAGAAG**CACGTGA**CCACAA
MET2	-360	-342	GTATTT**TCACGTGA**TGCGC
MET2	-554	-536	TAATAA**TCACGTGA**TATTT
MET17	-306	-288	.AAATGG**CACGTGA**AGCTGT
MET17	-332	-314	TTGAGG**TCAC**A**TGA**TCGCA
MET6	-540	-522	GCCACA**TCACGTGCA**CATT
MET6	-502	-484	AATATT**TCACGTGA**CTTAC
SAM2	-329	-311	.TCTACC**CACGTGA**CTATAA
SAM2	-381	-363	.TCTTCA**CA**T**GTGA**TTCATC

A	13	11	3	3	2	0	16	0	1	0	0	12
C	1	0	0	3	0	16	0	15	0	0	0	0
G	1	1	4	4	4	0	0	0	15	0	16	4
T	1	4	9	6	10	0	0	1	0	16	0	0

Pho4p binding sites

gene	start	end	sequence
PHO5	-260	-242	..GCACTCA**CACGTGGG**ACTA
PHO5	-260	-245	..GCACTCA**CACGTGGG**A
PHO5	-262	-239	TGGCACTCA**CACGTGGG**ACTAGCA
PHO8	-540	-522	...TCGGGC**CACGTGC**AGCGAT
PHO8	-736	-718	...ATATTAAG**CGTGCG**GGTAA
PHO81	-350	-332	...TTATGG**CACGTGCG**AATAA
PHO84	-421	-403	..TTTCCAG**CACGTGGG**GCGG
PHO84	-442	-425	...TAGTTC**CACGTGG**ACGTG
PHO84	-879	-874	.aaaagtgt**CACGTG**ataaaaat
PHO84	-267	-250	....TTAAAA**ACGTGCG**TATTA
PHO84	-592	-575	....TTACG**CACGTT**GGTGCTG
PHO5	-368	-349	...AATTAG**CACGTTTT**CGCATA
PHO5	(?)	(?)	..AAATTAG**CACGTTT**CGC
PHO5	-370	-347	.TAAATTAG**CACGTTTT**CGCATAGA

Pho4p binding specificity - matrix descriptions

C

Pho4p

A	14	0	5	7	6	0	26	0	0	0	0	3
C	2	8	5	16	6	26	0	26	0	1	0	4
G	4	2	1	1	12	0	0	0	26	0	16	12
T	6	16	15	2	2	0	0	0	0	25	10	7

D

Pho4p.cacgtg

A	2	17	0	0	0	0	2	1	8	5	5	13
C	16	0	18	0	0	0	6	3	4	5	0	1
G	0	1	0	18	0	18	9	12	2	5	2	1
T	0	0	0	0	18	0	1	2	4	3	11	3

E

Pho4p.cacgtt

A	7	0	2	5	1	0	8	0	0	0	0	1
C	0	1	1	3	3	8	0	8	0	0	0	0
G	0	0	0	0	4	0	0	0	8	0	0	2
T	1	7	5	0	0	0	0	0	0	8	8	5

Met31p binding sites

gene	start	end	sequence
MET14	-202	-182	CCTC**AAAAA**A**TGTGG**CAATGG
MET2	-313	-293	TGC**AAAAAA**T**TGTGG**ATGCAC
MET17	-227	-207	TCATG**AAAACTGTG**TAACATA
MET6	-313	-293	GTCGC**AAAACTGTGG**TAGTCA
SAM2	-306	-286	GCTTG**AAAACTGTGG**CGTTTT
SAM1	-283	-263	ACAGG**AAAACTGTGG**TGGCGC
MET19	-173	-153	ATAAGC**AAACTGTGG**GTTCAT
MUP3	-188	-168	CGG**AAAAAACTGTGG**CGTCGC
MET8	-184	-164	GG**AAAAAAA**A**TGTG**AAAATCG
MET1	-232	-212	CATAAT**AAACTGTG**AACGGAC
MET3	-259	-239	ACAAAG**CCACAGTTTT**ACAAC
MET28	-159	-139	CTAACA**CCACAGTTTT**GGGCG
MET8	-434	-414	TCTTGT**CCGCAGTTTT**ATCTG
MET30	-168	-148	GGGAAG**CCACAGTTT**GCGCGG
MET6	-405	-385	CTATCGAA**CTCGTTT**AGTCGC

A	5	11	14	14	14	2	0	0	0	0	2	5
C	2	2	0	0	0	11	0	0	1	0	0	5
G	5	0	0	0	0	0	0	14	0	14	11	1
T	2	1	0	0	0	1	14	0	13	0	1	3

Matching upstream regions with multiple matrices

- Each one of the 6309 upstream regions is scanned with each one of the 5 matrices
- For each matrix and gene, the 3 top scores are retained
- This results in a 15-variate table, where each gene is characterized by 15 scores

Matrix	Pho4p			Pho4p.cacgtg			Pho4p.cacgtt			Met4p			Met31p		
	top 1	top 2	top 3	top 1	top 2	top 3	top 1	top 2	top 3	top 1	top 2	top 3	top 1	top 2	top 3
gene 1	$x_{1,1}$	$x_{2,1}$	…	…	…	…	…	…	…	…	…	…	…	…	…
gene 2	$x_{1,2}$	$x_{2,2}$	…	…	…	…	…	…	…	…	…	…	…	…	…
gene 3	$x_{1,3}$	$x_{2,3}$	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
gene i	$x_{1,i}$	$x_{2,i}$	…	…	…	…	…	…	…	…	…	…	…	…	…
gene i+1	$x_{1,i+1}$	$x_{2,i+1}$	…	…	…	…	…	…	…	…	…	…	…	…	…
gene i+2	$x_{1,i+2}$	$x_{2,i+2}$	…	…	…	…	…	…	…	…	…	…	…	…	…
…	…	…													
gene 6308	$x_{1,n-1}$	$x_{2,n-1}$	…	…	…	…	…	…	…	…	…	…	…	…	…
gene 6309	$x_{1,n}$	$x_{2,n}$	…	…	…	…	…	…	…	…	…	…	…	…	…

Data - matrix scores

- matrix scores: top scores obtained with PSSM for Met4p, Met31p, Pho4p, Pho4p.cacgtg, and Pho4p.cacgtt; 5 matrices, 3 top scores per matrix.

		Pho4p									Met4p			Met31p		
;orf	gene	Pho4p.cacgtg.1	Pho4p.cacgtg.2	Pho4p.cacgtg.3	Pho4p.cacgtk.1	Pho4p.cacgtk.2	Pho4p.cacgtk.3	Pho4p.cacgtt.1	Pho4p.cacgtt.2	Pho4p.cacgtt.3	Met4p.1	Met4p.2	Met4p.3	Met31p.1	Met31p.2	Met31p.3
YAL001C	TFC3	3.79	3.42	2.28	3.27	0.17	0.14	4.38	4.17	3.41	7.88	6.91	2.73	5.39	3.55	3.55
YAL002W	VPS8	9.31	6.18	5.16	9.59	4.72	4.22	7.75	6.36	5.79	5.69	4.68	3.37	3.55	3.14	2.09
YAL003W	EFB1	6.66	3.23	2.39	6.02	4.75	2.77	4.04	3.84	3.09	4.3	3.94	3.8	4.64	4.54	3.78
YAL004W	YAL004W	2.39	1.69	1.07	3.05	2.98	1.96	5.66	2.11	1.52	3.14	3.01	2.65	3.37	3.11	2.96
YAL005C	SSA1	6.66	3.23	2.39	6.02	4.75	2.77	4.04	3.84	3.09	4.3	3.94	3.8	5.15	4.64	4.54
YAL007C	ERP2	5.02	3.6	2.28	2.11	1.16	1.16	4.75	3.18	2.65	4.19	4.13	3.44	7	3.55	2.48
YAL008W	FUN14	6.27	4.92	2.94	4.65	3.66	3.47	3.26	2.82	2.8	4.26	4.19	3.81	4.54	3.78	2.35
YAL009W	SPO7	6.11	3.69	0.65	5.37	4.99	3.81	8.51	5.23	3.34	7.42	3.38	3.28	6.01	2.48	2.2
YAL010C	MDM10	6.27	1.7	1.58	3.47	1.43	0.75	3.01	2.82	1.98	4.26	4.19	2.49	3.78	2.35	2.35
YAL011W	YAL011W	4.2	3.02	2.61	5.68	3.56	3.23	5.51	5.08	4.85	1.93	1.63	1.29	4.06	2.96	2.79
YAL012W	CYS3	4.78	4	3.73	5.43	3.72	3.55	6.5	5.34	3.62	9.34	5.92	4.88	9.43	3.96	3.07
YAL013W	DEP1	5.41	4.83	3.3	6.81	2.43	2.11	4.73	4.54	1.71	3.35	2.82	2.62	6.59	2.35	2.31
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
YPR203W	YPR203W	3.33	3.31	2.73	6.75	4.72	4.45	6.94	3.4	2.94	4.89	3.76	3.28	5.98	5.39	4.91

Training and testing sets

- There is a subset of objects for which the class is n-known a priori, on the basis of external information (e.g. biological knowledge)
- These classes will be used as criterion variable.
  - To train the program (training set)
  - To test the program (testing set)
- Note : the training and testing sets might contain some errors.

Phosphate-responding genes

#	ORF	Gene name	Family
1	YBR093C	PHO5	PHO
2	YDR481C	PHO8	PHO
3	YAR071W	PHO11	PHO
4	YHR215W	PHO12	PHO
5	YOL001W	PHO80	PHO
6	YGR233C	PHO81	PHO
7	YML123C	PHO84	PHO
8	YPL031C	PHO85	PHO
9	YJL117W	PHO86	PHO
10	YCR037C	PHO87	PHO
11	YBR106W	PHO88	PHO
12	YBR296C	PHO89	PHO
13	YHR136C	SPL2	PHO

Methionine-responding genes

#	ORF	Gene name	Family
14	YBR213W	MET8	MET
15	YDR253C	MET32	MET
16	YDR502C	SAM2	MET
17	YER091C	MET6	MET
18	YFR030W	MET10	MET
19	YHL036W	MUP3	MET
20	YIL046W	MET30	MET
21	YIR017C	MET28	MET
22	YJR010W	MET3	MET
23	YJR137C	ECM17	MET
24	YKL001C	MET14	MET
25	YKR069W	MET1	MET
26	YLR180W	SAM1	MET
27	YLR303W	MET17	MET
28	YLR396C	VPS33	MET
29	YNL241C	ZWF1	MET
30	YNL277W	MET2	MET
31	YOL064C	MET22	MET
32	YPL038W	MET31	MET

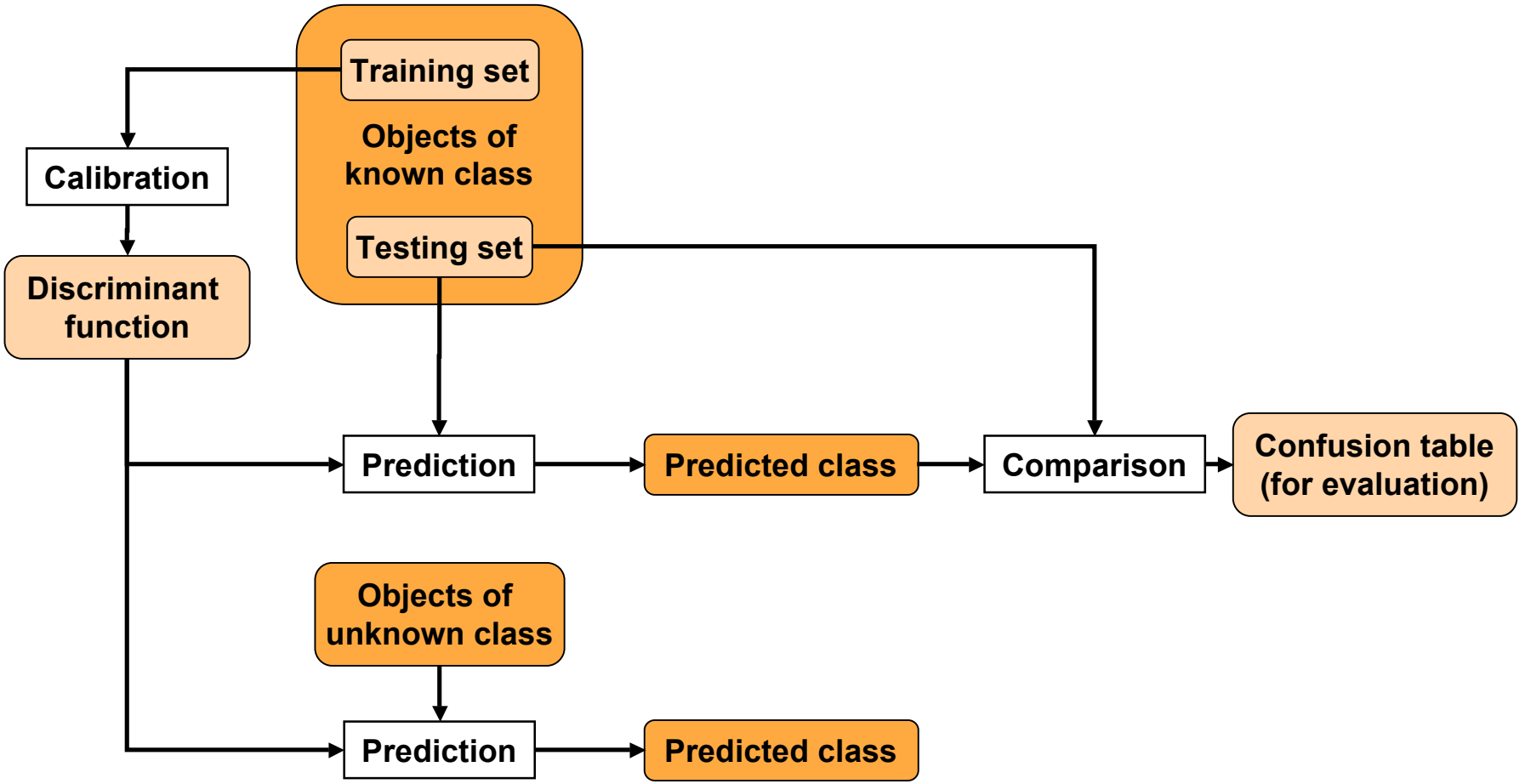
Control genes

#	ORF	Gene name	Family
33	YAL038W	CDC19	CTL
34	YBL005W	PDR3	CTL
35	YBL005W-A	YBL005W-A	CTL
36	YBL005W-B	YBL005W-B	CTL
37	YBL030C	PET9	CTL
38	YBR006W	UGA5	CTL
39	YBR018C	GAL7	CTL
40	YBR020W	GAL1	CTL
41	YBR115C	LYS2	CTL
42	YBR184W	YBR184W	CTL
43	YCL018W	LEU2	CTL
44	YDL131W	LYS21	CTL
45	YDL182W	LYS20	CTL
46	YDL205C	HEM3	CTL
47	YDL210W	UGA4	CTL
48	YDR011W	SNQ2	CTL
49	YDR044W	HEM13	CTL
50	YDR234W	LYS4	CTL
51	YDR285W	ZIP1	CTL
…	…	…	…
112	YPR065W	ROX1	CTL
113	YPR138C	MEP3	CTL
114	YPR145W	ASN1	CTL

Approach

- Extract upstream sequences for each one of the 6000 yeast genes
- Use position-weight matrices to predict putative regulatory elements
- Use genes with known PHO or MET regulation, plus a control group (CTL) as training set
  - Build a classification rule (train a classifier) based on the training set.
  - Evaluate the accuracy of the classification rule.
  - Select the best classification method and parameters.
- Apply the classification to each one of the 6000 yeast genes, in order to predict the MET- and PHO-regulated genes

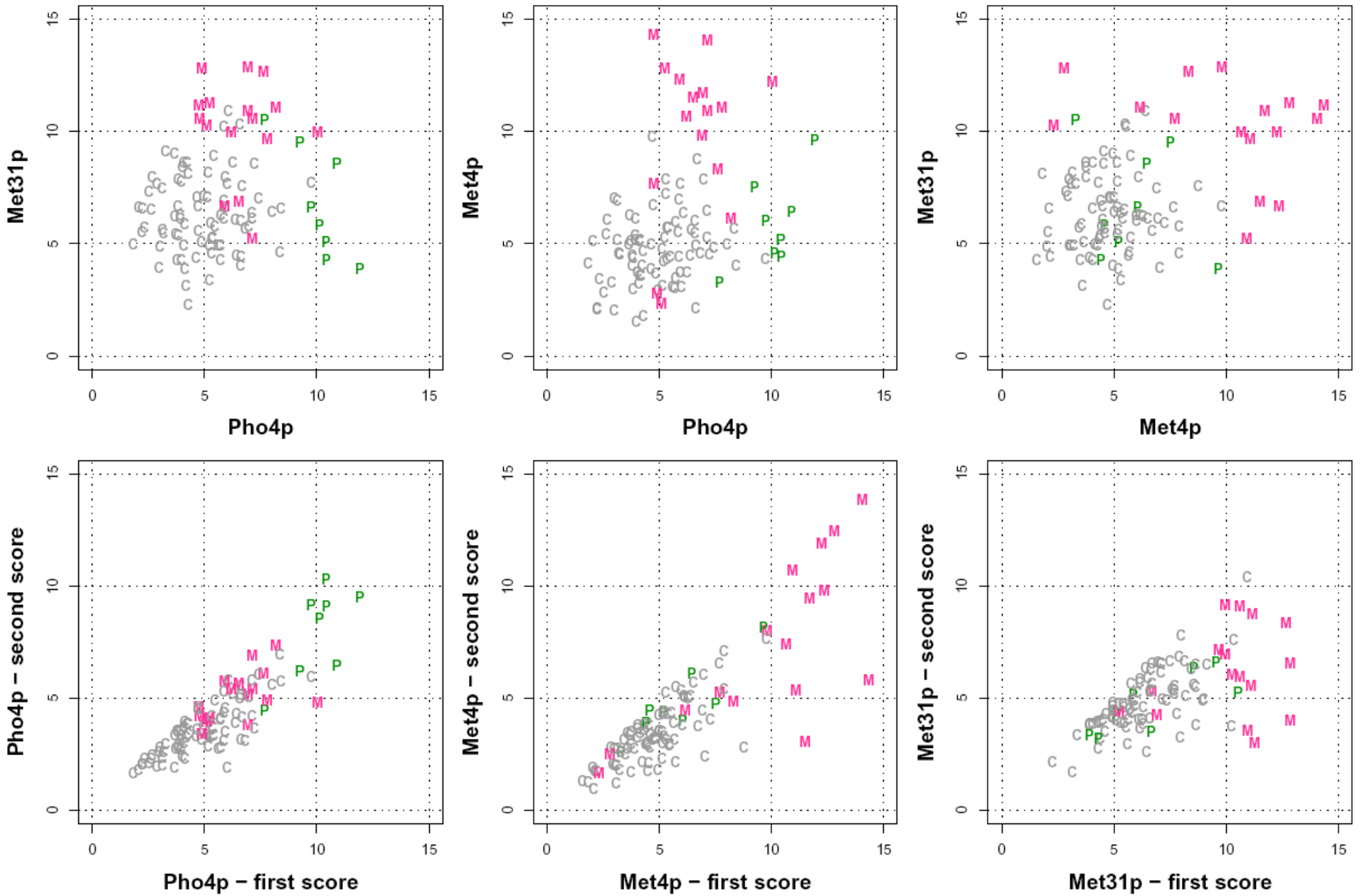
Discriminant analysis



Difficulties

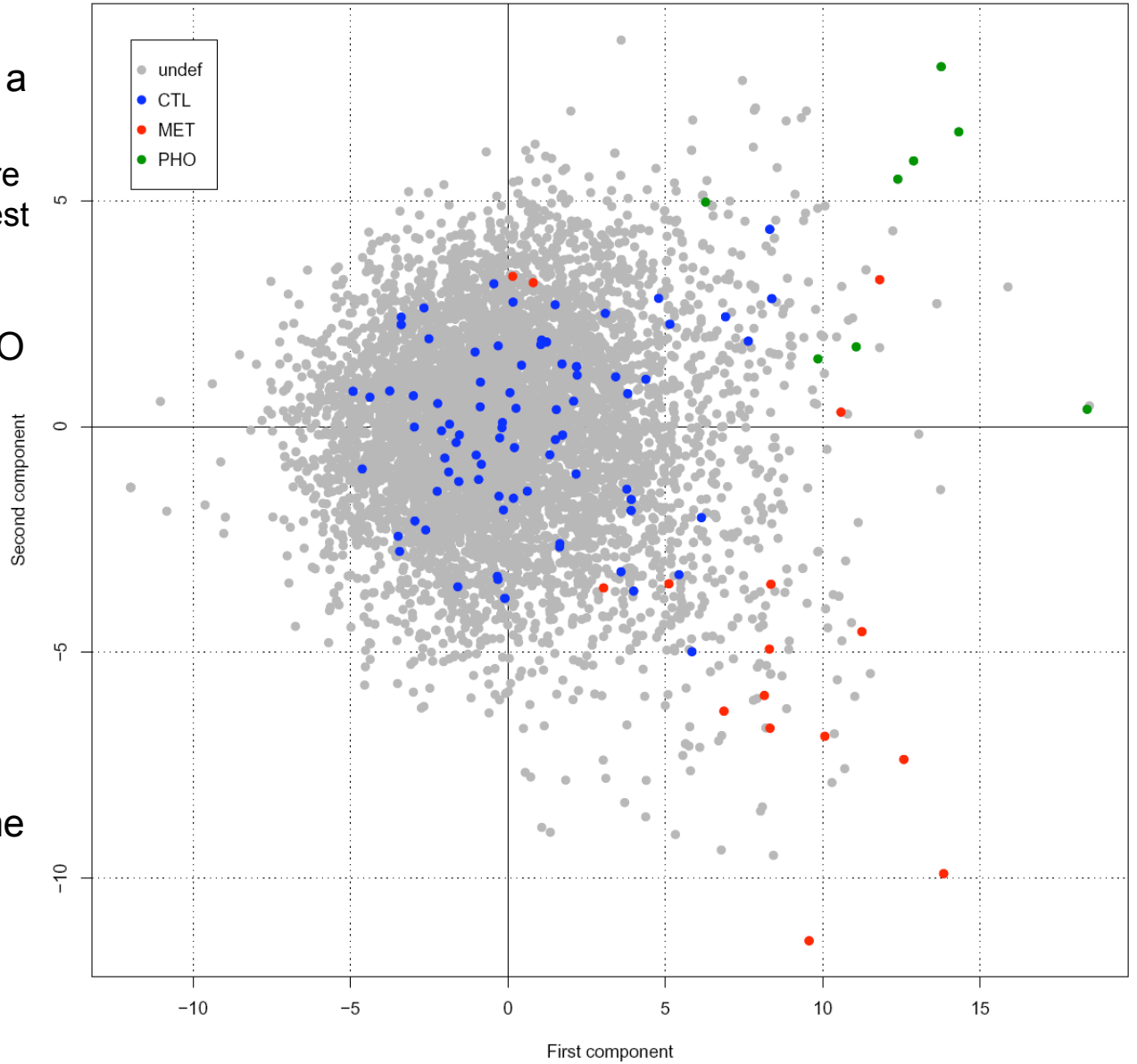
- The training sets are very small
  - 13 PHO genes
  - 19 MET genes
  - 82 control genes (supposed to respond neither to phosphate nor to methionine)
- Over-fitting
  - The number of variables (15 matrix scores, or 44 pattern counts, depending on the considered data set) is higher than the number of elements in some classes of the training set
- Size of the prediction set
  - After training and evaluation, the discriminant function will b used to classify each one of the 6300 yeast genes. Even a small error rate (e.g. 1%) would lead to an important number of false predictions (60 false positives).

MET and PHO predicted sites



Principal Component Analysis - matrix scores

- The 15-dimensional score space can be projected onto a 2-dimensional space
  - The two first components are the directions with the highest variance in the 15-dimensional space
- Previously characterized PHO and MET genes are labelled
- The CTL genes are genes with known regulation, and supposedly not regulated by MET or PHO. As expected, they are mixed with the unlabelled genes.
- Most PHO and MET genes are projected in different angles of the space, but some of them are mixed with other groups.



Classification rules

- New units can be classified on the basis of rules based on the calibration sample
- Several alternative rules can be used
  - Maximum likelihood rule: assign unit u to group g if

$$f(X \mid g) > f(X \mid g') \qquad for\ g' \neq g$$

  - Inverse probability rule: assign unit u to group g if

$$P(X \mid g) > P(X \mid g') \qquad for\ g' \neq g$$

  - Posterior probability rule: assign unit u to group g if

$$P(g \mid X) > P(g' \mid X) \qquad for\ g' \neq g$$

Posterior probability rule

- The posterior probability can be obtained by application of Bayes' theorem

$$P(g \mid X) = \frac{P(X \mid g)P(g)}{P(X)}$$

$$P(g \mid X) = \frac{P(X \mid g)\pi_g}{\sum_{g'=1}^{k} P(X \mid g')\pi_{g'}}$$

Where

- X is the unit vector
- g is a group
- k is the number of groups
- p_g is the prior probability of group g

p_g can be specified by the user or estimated from the training frequencies.

Linear versus quadratic classification rule

- Under assumption of multivariate normality
  - There is one covariance matrix per group g.
  - When all covariance matrices are assumed to be identical, the classification rule can be simplified to obtain a linear function
    -> **Linear Discriminant Analysis** (**LDA**)
  - When the groups have not the same covariance matrix,
    -> **Quadratic Discriminant Analysis** (**QDA**) is more appropriate.

Evaluation of the discriminant function - confusion table

- The results of the evaluation are summarized in a **confusion table**, which contains the count of the predicted versus known class.
- The confusion table can be used to calculate the accuracy of the predictions.
- With linear discriminant analysis, the error rate is even higher than with hierarchical clustering ! This is due to a problem of over-fitting: there are more variables (15) than objects in some training classes (13 for PHO)

Internal validation (biased)

		Known			
		PHO	MET	CTL	Sum
Predicted	PHO	10	0	1	11
	MET	0	16	1	17
	CTL	3	3	80	86
	Sum	13	19	82	114

Error rate 0.07
Hit rate 0.93

Leave-one-out validation

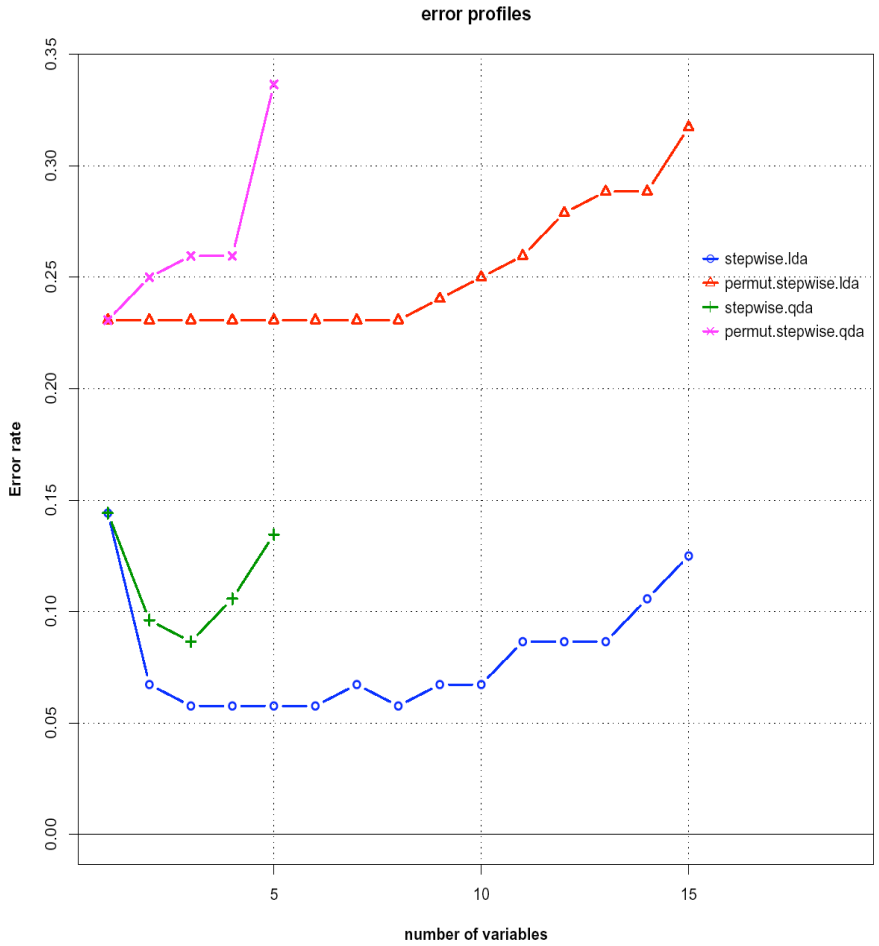
		Known			
		PHO	MET	CTL	Sum
Predicted	PHO	6	1	2	9
	MET	3	13	1	17
	CTL	4	5	79	88
	Sum	13	19	82	114

Error rate 0.14
Hit rate 0.86

Selection of variables

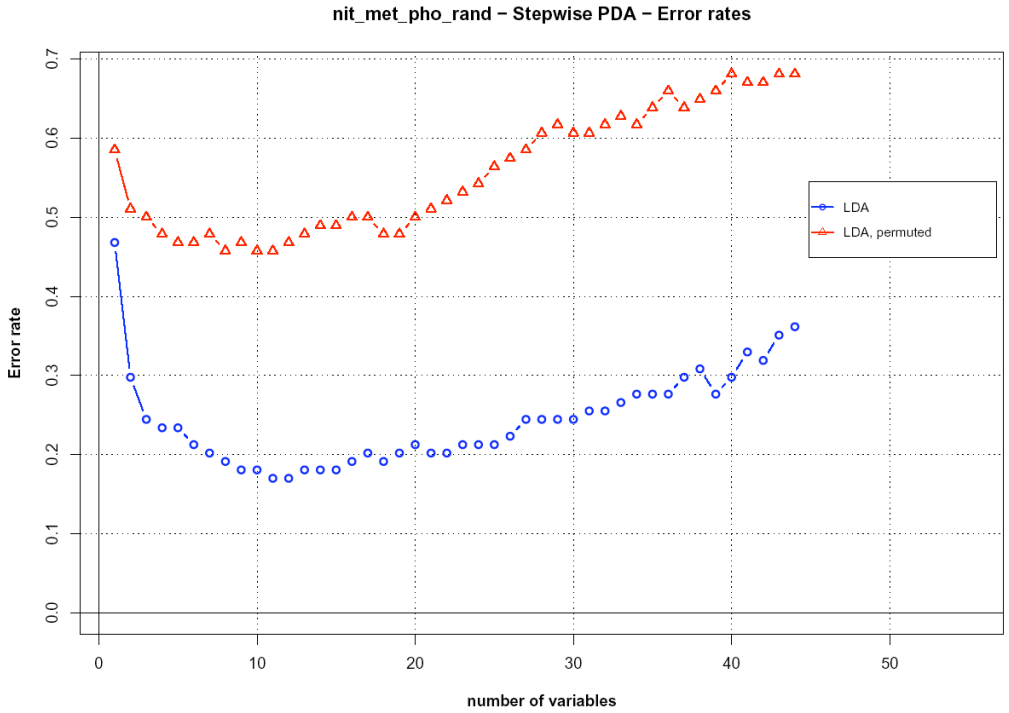
- When there are too many variables, the classification is less accurate.
- In particular, the number of variables must be much smaller than the number of elements in the training groups.
- In our case, we have 15 variables, but the PHO group contains only 13 genes.
- We select the best subset of variables via a stepwise procedure

Stepwise discriminant analysis - error rate



- Even a random classification would still assign some objects to the correct group by chance.
- The random rate of correct assignation depends on
  - The relative size of the groups
  - The structure of the data
  - The number of variables
- The expected error rate can be estimated with a permutation test
  - The method is applied to the real data set, but the training labels are randomly assigned.

Error rate - pattern counts



Predicted \ Known	MET	NIT	PHO	RAND	SUM
MET	15	0	1	1	17
NIT	0	26	0	1	27
PHO	0	0	9	0	9
RAND	5	5	3	28	41
SUM	20	31	13	30	94
Errors	16	17.02%			
Correct	78	82.98%			

- Genes : NIT, PHO, MET + random sequences (RAND)
- 44 variables (pattern counts)
- Optimum: 7 variables
- Best variables
  - cttatc.gataag
  - cacgtg.cacgtg
  - aacgtg.cacgtt
  - acgngcg.cgcncgt
  - acgtga.tcacgt
  - ctgata.tatcag
  - agataa.ttatct
  - atcacg.cgtgat
  - $acan_{14}tgc.gcan_{14}tgt$
  - aacngtg.cacngtt
  - $cacn_2gac.gtcn_2gtg$
  - $cagn_2cgg.ccgn_2ctg$

Predicted versus training class

- Each gene is plotted in a plane where the axes represent the two linear discriminant functions.
- The colour indicates the training class, the letter the predicted class.
- Misclassifications
  - One PHO gene (green) and one CTL gene (blue) are classified as MET
  - Four MET genes are classified as CTL.

Pattern profiles of misclassified genes

- Each column represents a matrix score
  - 1-3: Pho4p
  - 4-6: Pho4p.cacgtg
  - 7-9: Pho4p.cacgtt
  - 10-12: Met4p
  - 13-15: Met31p
- The MET genes which were classified as CTL have
  - no good match for Met4p matrix.
  - a good match for Met31p matrix.
- The CTL and PHO genes classified as MET have a quite good match for Met4p matrix.

Optimal conditions

- Pattern detection: 3 top scores for 5 position-weight matrices
- Linear Discriminant Analysis
- Forward selection procedure
- External validation with the leave-one-out method

3 group classification

Predicted \ Known	PHO	MET	CTL	SUM
PHO	7	0	0	7
MET	1	12	1	14
CTL	0	4	79	83
SUM	8	16	80	104
Errors	6	5.77%		
Correct	98	94.23%		

PHO against others

Predicted \ Known	PHO	CTL	SUM
PHO	7	0	7
CTL	1	96	97
SUM	8	96	104
Errors	1	0.96%	
Correct	103	99.04%	

MET against others

Predicted \ Known	MET	CTL	SUM
MET	13	0	13
CTL	3	88	91
SUM	16	88	104
Errors	3	2.88%	
Correct	101	97.12%	

Gonze, D. et al.. 2005. Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs. *Bioinformatics* **21:** 3490-3500.

Comparison of predicted and prior class

- Letters indicate the predicted class
- Colors indicate the prior class

Profiles by predicted class

"Misclassified units"

Choice of the prior probabilities

- The classes may have different proportions between the sample and the population
- For example, we could decide, on the basis of our biological knowledge, that it is likely to have 1% rather than 11% of yeast gene responding to phosphate.

		Population	
Class	**Sample**	**Priors from sample**	**Arbitrary priors**
PHO	**13** ***11%***	**659** ***11%***	**58** ***1%***
MET	**19** ***17%***	**964** ***17%***	**58** ***1%***
CTL	**82** ***72%***	**4160** ***72%***	**5667** ***98%***
TOTAL	**114**	**5783**	**5783**

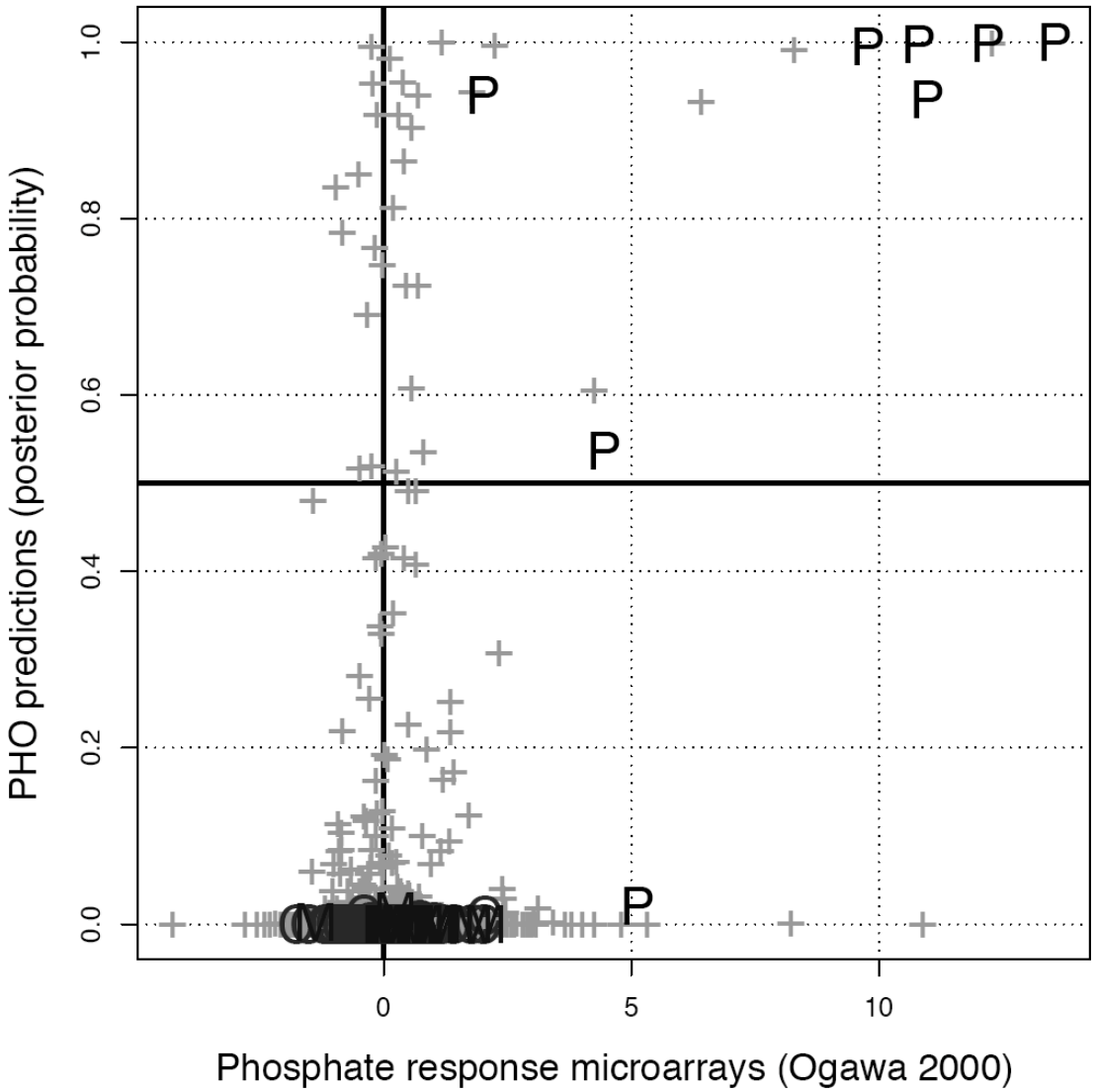
Discriminant analysis - prediction

- 138 genes predicted as methionine-regulated
- 64 genes predicted as phosphate-responding

PHO predictions

#	ORF	name	training class	predicted class	posterior probabilities			Description
					CTL	MET	PHO	
1	YBR093C	PHO5	PHO	PHO	0.00%	0.00%	100.00%	repressible acid phosphatase precursor
2	YML123C	PHO84	PHO	PHO	0.01%	0.01%	99.99%	high-affinity inorganic phosphate/H+ symporter
3	YBR296C	PHO89	PHO	PHO	0.02%	0.00%	99.98%	Na+-coupled phosphate transport protein, high affinity
4	YEL017W	YEL017w		PHO	0.04%	0.01%	99.94%	hypothetical protein
5	YHR137W	ARO9		PHO	0.08%	0.11%	99.81%	aromatic amino acid aminotransferase II
6	YHR136C	SPL2	PHO	PHO	0.10%	0.13%	99.77%	suppressor of plc1-delta
7	YGR233C	PHO81	PHO	PHO	0.17%	0.06%	99.77%	cyclin-dependent kinase inhibitor
8	YAR071W	PHO11	PHO	PHO	0.25%	0.00%	99.75%	secreted acid phosphatase
9	YHR215W	PHO12		PHO	0.25%	0.00%	99.75%	secreted acid phosphatase
10	YDR303C	RSC3		PHO	0.35%	0.01%	99.65%	similarity to transcriptional regulator proteins
11	YDL202W	MRPL11		PHO	0.60%	0.04%	99.36%	ribosomal protein of the large subunit, mitochondrial
12	YAR070C	YAR070c		PHO	0.67%	0.01%	99.32%	hypothetical protein
13	YDR281C	PHM6		PHO	0.96%	0.01%	99.03%	hypothetical protein, has a role in phosphate metabolism
14	YER073W	ALD5		PHO	1.06%	0.00%	98.94%	aldehyde dehydrogenase (NAD+), mitochondrial
15	YKR050W	TRK2		PHO	1.31%	0.01%	98.68%	moderate-affinity potassium transport protein
16	YKR048C	NAP1		PHO	1.39%	0.05%	98.56%	nucleosome assembly protein I
17	YMR253C	YMR253c		PHO	0.37%	1.10%	98.54%	strong similarity to YPL264c
18	YMR255W	GFD1		PHO	0.37%	1.10%	98.54%	protein of the nuclear pore complex
19	YNL113W	RPC19		PHO	1.84%	0.05%	98.10%	DNA-directed RNA polymerase I,III 16 KD subunit
20	YNL115C	YNL115c		PHO	1.84%	0.05%	98.10%	weak similarity to S.pombe hypothetical protein SPAC23C11
21	YDR310C	SUM1		PHO	1.69%	1.24%	97.07%	suppressor of SIR mutations
22	YDR311W	TFB1		PHO	1.69%	1.24%	97.07%	TFIIH subunit (transcription initiation factor), 75 kD
23	YJR059W	PTK2		PHO	2.88%	0.07%	97.05%	involved in polyamine uptake
24	YCR037C	PHO87		PHO	0.04%	3.25%	96.71%	member of the phosphate permease family
25	YJR058C	APS2		PHO	3.88%	0.10%	96.02%	AP-2 complex subunit, sigma2 subunit, 17 KD
26	YCR098C	GIT1		PHO	4.14%	0.18%	95.67%	glycerophosphoinositol transporter
27	YOR347C	PYK2		PHO	4.97%	0.01%	95.02%	pyruvate kinase, glucose-repressed isoform
28	YCL054W	SPB1		PHO	6.00%	0.21%	93.79%	required for ribosome synthesis, putative methylase
29	YHR079C	IRE1		PHO	6.40%	0.00%	93.60%	protein kinase
30	YAL002W	VPS8		PHO	9.54%	0.07%	90.39%	vacuolar sorting protein, 134 kD

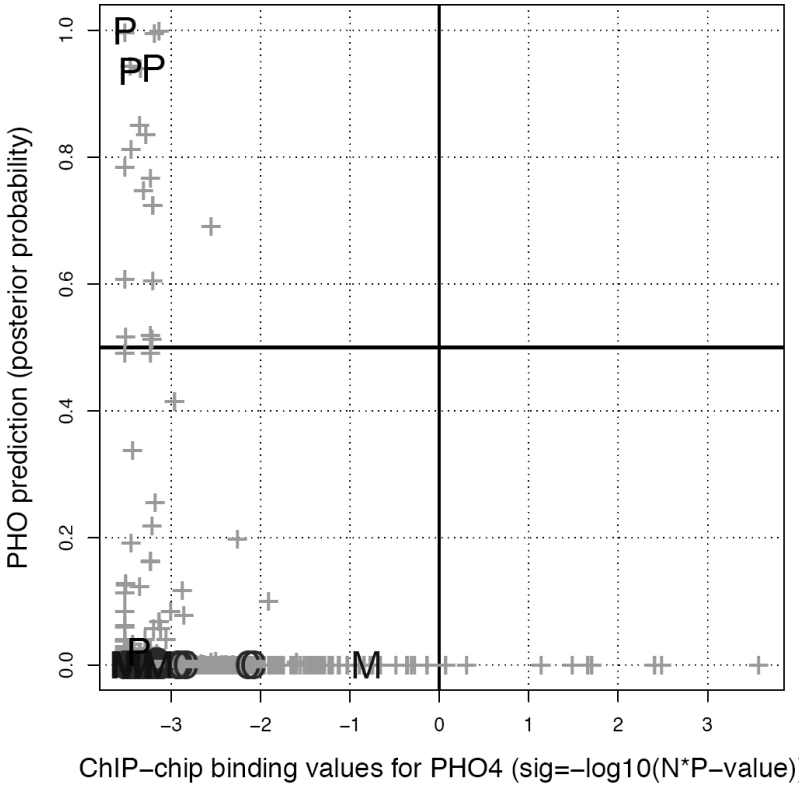
PHO predictions versus microarray data

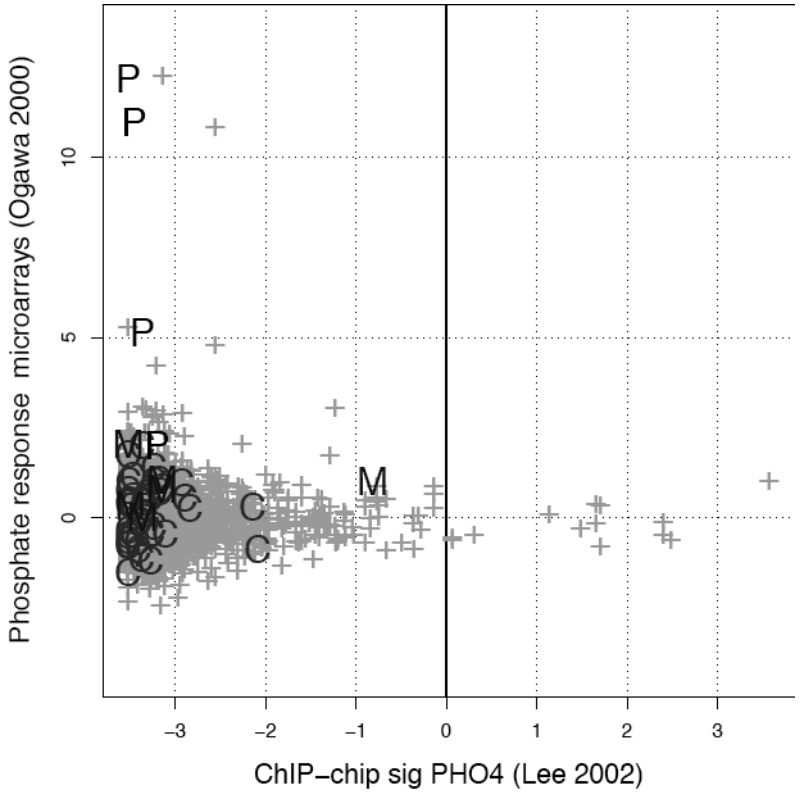


- PHO predictions include most (but not all) of the phosphate-responding genes (microarray data, Ogawa 2000)
- There are many additional predictions which are not detected by microarrays
  - False positives ?
  - Genes responding to different conditions ?

PHO predictions versus Chip-CHIP data (Lee, 2002)

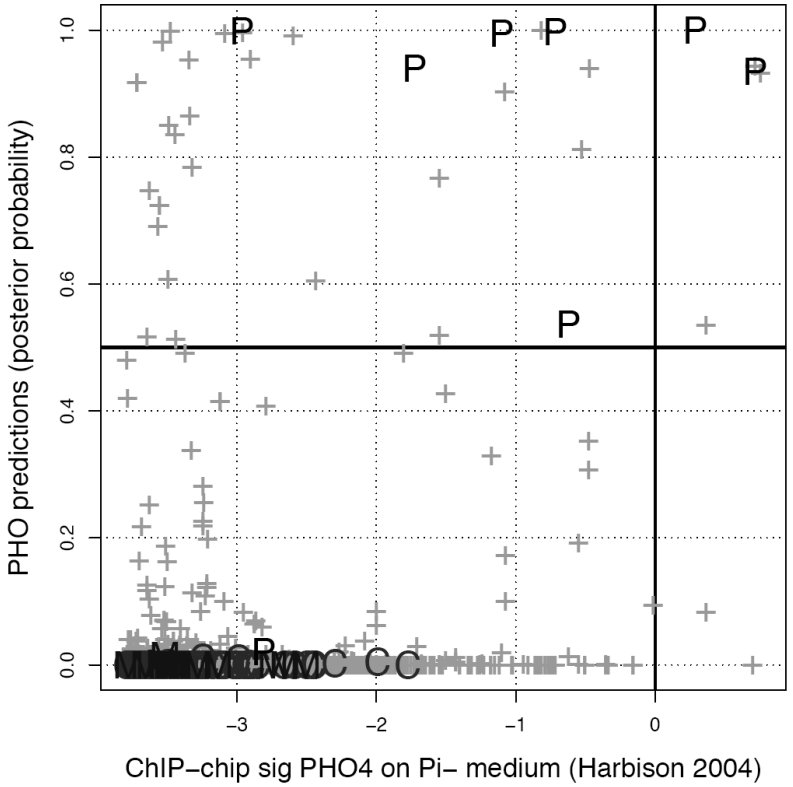
- There is not a single common gene between our PHO predictions and the Pho4p-bound promoters detected with the ChIP-chip technology by Lee et al, 2002)
- However, Lee results for Pho4p fail to detect
  - genes known to be regulated by Pho4p
  - Genes responding to phosphate in Ogawa (2000)
- Problem with the ChIP-chip experiment
  - was performed in rich medium -> ***Pho4p is inactive !!!***

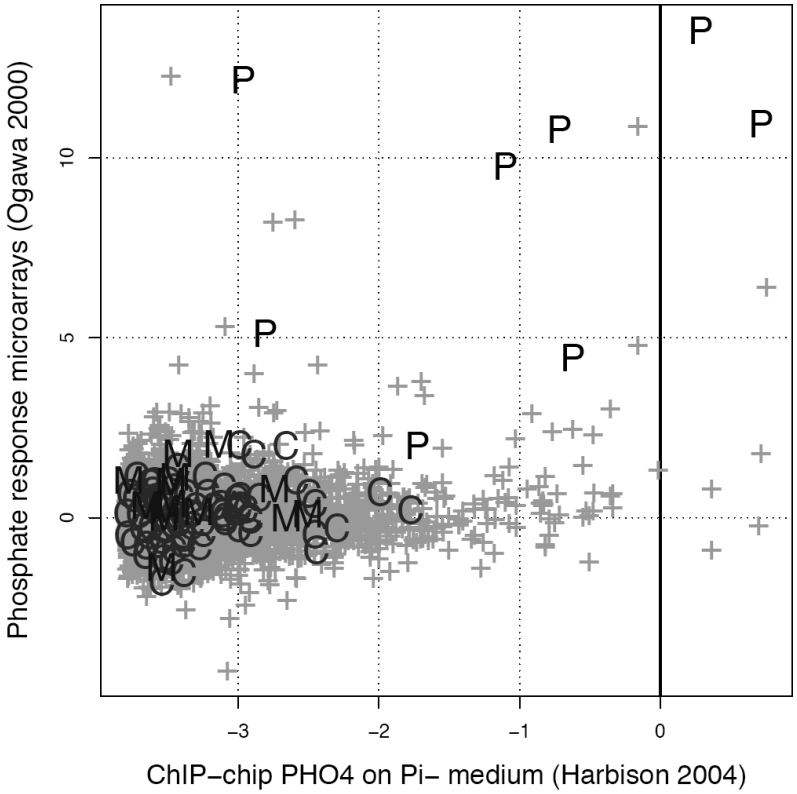




PHO predictions versus Chip-CHIP data (Harbison, 2004)

- In 2004, the same group performed new experiments with different environmental conditions (Harbison, 2004)
- There is a slightly better (but far from perfect) correspondence between ChIP-chip results and
  - Our PHO predictions
  - microarray data (Ogawa *et al.*, 2000)
  - Annotated Pho4p target genes (P on the plots)



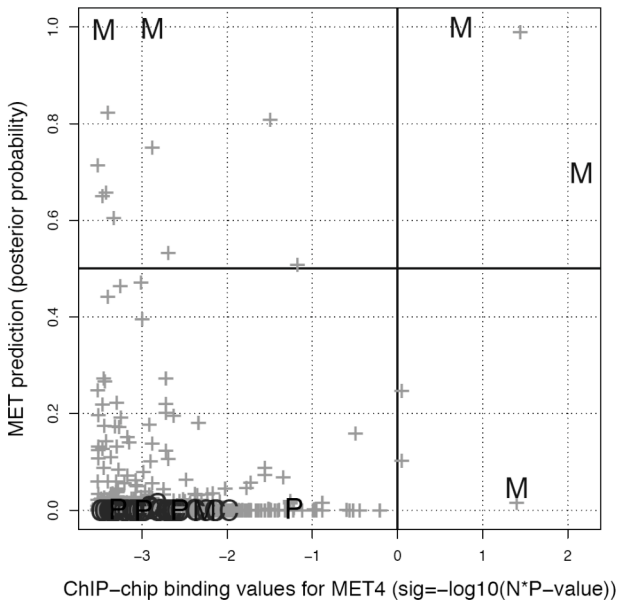


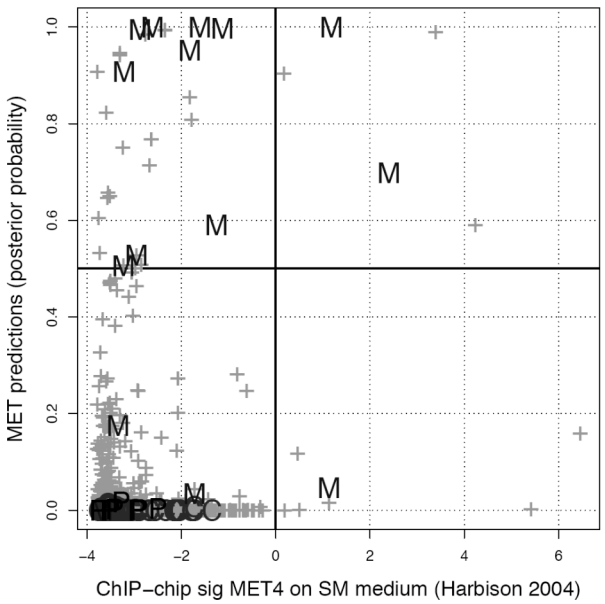
MET predictions

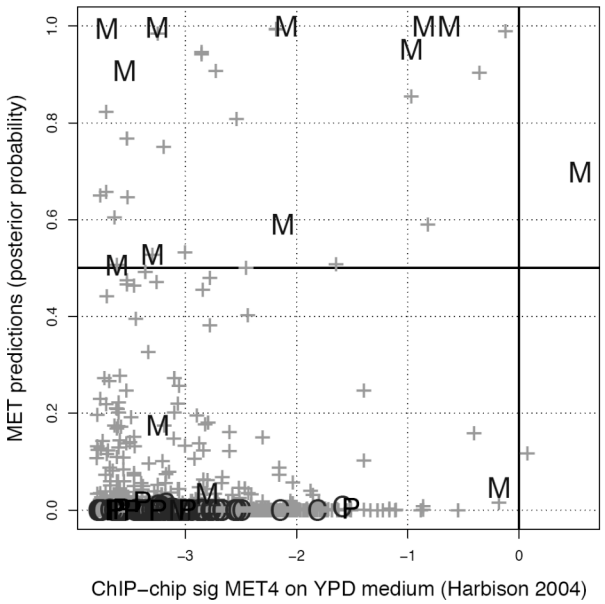
#	ORF	name	training class	predicted class	CTL	MET	PHO	Description
1	YKL016C	ATP7		MET	0.01%	99.99%	0.00%	F1F0-ATPase complex, FO D subunit
2	**YNL277W**	**MET2**	**MET**	**MET**	**0.01%**	**99.98%**	**0.00%**	**homoserine O-acetyltransferase**
3	**YBR213W**	**MET8**	**MET**	**MET**	**0.02%**	**99.98%**	**0.00%**	**siroheme synthase**
4	**YKL001C**	**MET14**	**MET**	**MET**	**0.02%**	**99.98%**	**0.00%**	**ATP adenosine-5prime-phosphosulfate 3prime-phosphotransferase**
5	YLR149C	YLR149c		MET	0.04%	99.96%	0.00%	weak similarity to hypothetical protein SPCC4G3.03 S. pombe
6	**YER091C**	**MET6**	**MET**	**MET**	**0.05%**	**99.95%**	**0.00%**	**5-methyltetrahydropteroyltriglutamate--homocysteine methyltransferase**
7	YER092W	IES5		MET	0.05%	99.95%	0.00%	weak similarity to tryptophan synthase beta subunit - Aquifex aeolicus
8	YLR150W	STM1		MET	0.06%	99.94%	0.00%	specific affinity for guanine-rich quadruplex nucleic acids
9	YDL074C	BRE1		MET	0.09%	99.91%	0.00%	weak similarity to spindle pole body protein NUF1
10	**YJR010W**	**MET3**	**MET**	**MET**	**0.09%**	**99.91%**	**0.00%**	**sulfate adenylyltransferase**
11	YER125W	RSP5		MET	0.16%	99.84%	0.00%	hect domain E3 ubiquitin-protein ligase
12	YGR154C	YGR154c		MET	0.24%	99.75%	0.01%	strong similarity to hypothetical proteins YKR076w and YMR251w
13	YLL060C	GTT2		MET	0.26%	99.74%	0.00%	glutathione S-transferase
14	**YIL046W**	**MET30**	**MET**	**MET**	**0.22%**	**99.74%**	**0.04%**	**involved in regulation of sulfur assimilation genes and cell cycle progression**
15	YIL047C	SYG1		MET	0.22%	99.74%	0.04%	member of the major facilitator superfamily
16	YOR367W	SCP1		MET	0.26%	99.74%	0.00%	similarity to mammalian smooth muscle protein SM22 and chicken calponin alpha
17	YFL018C	LPD1		MET	0.27%	99.69%	0.04%	dihydrolipoamide dehydrogenase precursor
18	YHR001W-A	QCR10		MET	0.32%	99.67%	0.00%	ubiquinol--cytochrome-c reductase 8.5 kDa subunit
19	YML122C	YML122c		MET	0.24%	99.66%	0.10%	hypothetical protein
20	YER091C-A	YER091c-a		MET	0.43%	99.56%	0.01%	hypothetical protein - identified by SAGE
21	YIL074C	SER33		MET	0.44%	99.52%	0.04%	3-phosphoglycerate dehydrogenase
22	YFL017W-A	SMX2		MET	0.51%	99.46%	0.03%	snRNP G protein (the homologue of the human Sm-G)
23	YPL250C	ICY2		MET	0.05%	99.43%	0.52%	weak similarity to YMR195w
24	YDL059C	RAD59		MET	0.57%	99.36%	0.07%	recombination and DNA repair protein
25	YGR204W	ADE3		MET	0.58%	99.30%	0.12%	C1-tetrahydrofolate synthase (trifunctional enzyme),cytoplasmic
26	YGR155W	CYS4		MET	0.85%	99.12%	0.04%	cystathionine beta-synthase
27	YDL058W	USO1		MET	0.95%	99.04%	0.01%	intracellular protein transport protein

MET predictions versus chip-chip data

- We compared MET predictions with ChIP-chip data
  - Lee (2002): rich medium
  - Harbison (2004): SM medium
  - Harbison (2004): YPD medium
- The correspondences are rather poor
- Even though our predictions contain a rate of false positives, and miss some MET genes, the correspondence with annotated MET is better than for genes detected experimentally with the ChIP-chip method !







Summary - discriminant analysis

- Discriminant analysis is based on a set of quantitative predictor variables, and a single nominal criterion variable.
- A sample is used to build a set of discriminant functions (calibration), which is then used to assign additional units to classes (prediction).
- The discriminant function can be either linear or quadratic. Linear discriminant analysis relies on the assumption that the different classes have similar covariance matrices.
- The accuracy of the discriminant function can be evaluated in different ways.
  - On the whole sample (internal approach)
  - Splitting of the sample into training and testing set (holdout approach)
  - Successively discard each sample unit, build a discriminant function and predict the discarded unit (leave-one-out)
- The efficiency decreases with the p/N ratio. When this ratio is too low, there is a problem of over-fitting.
- Stepwise approaches consist in selecting the subset of variables which raises the highest efficiency.

Summary - Gene classification

- To some extend, it is possible to classify genes according to regulatory signals, but there are different sources of errors
  - a single pattern is poorly informative
  - combining multiple patterns returns however interesting results.
- Unsupervised classification (clustering)
  - simple counts of selected patterns already return some interesting results
  - the choice of an appropriate metrics is critical
- Supervised classification (discriminant analysis)
  - training sets are generally small, when they exist
  - if there are many variables, feature selection is necessary to avoid over-fitting
  - if correctly used, it is always more accurate than unsupervised classification

Evaluation with annotated regulons

- All yeast regulons from TRANSFAC + additional annotation from aMAZE
- Pattern discovery with oligo-analysis + dyad-analysis
- Discriminant analysis was applied to regulons where a motif with sig >= 1 was detected.

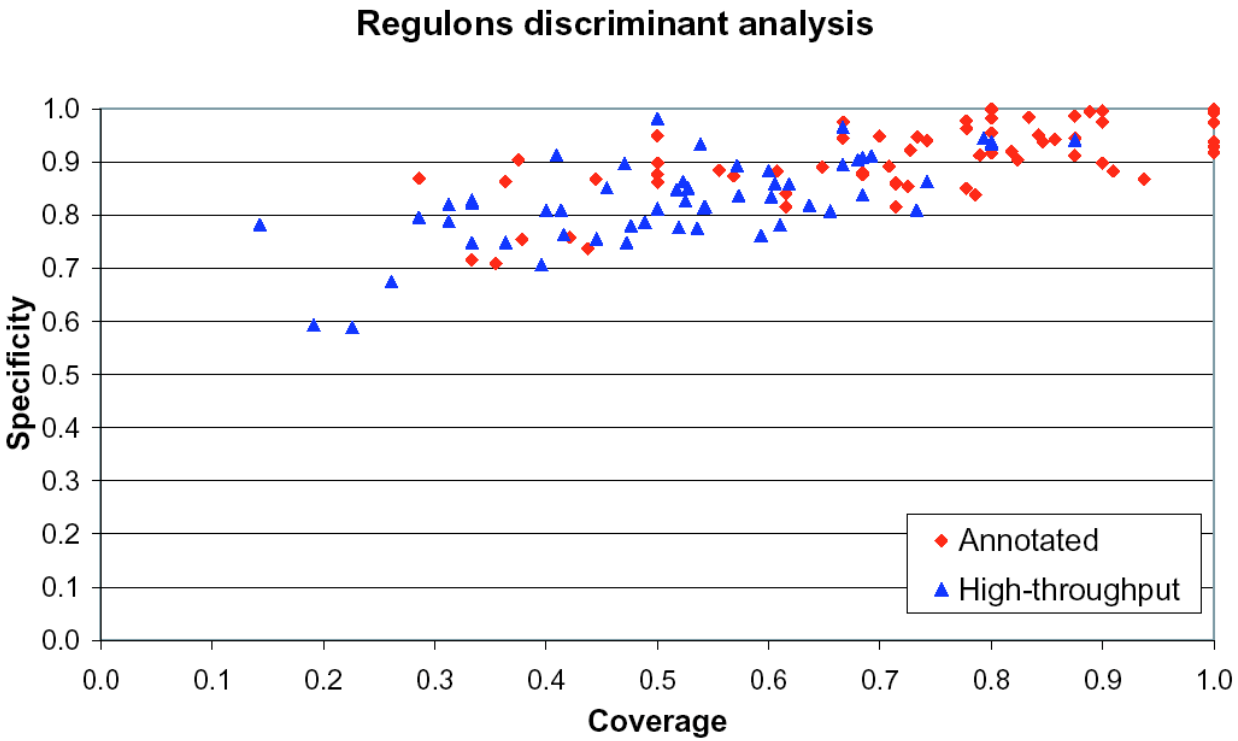
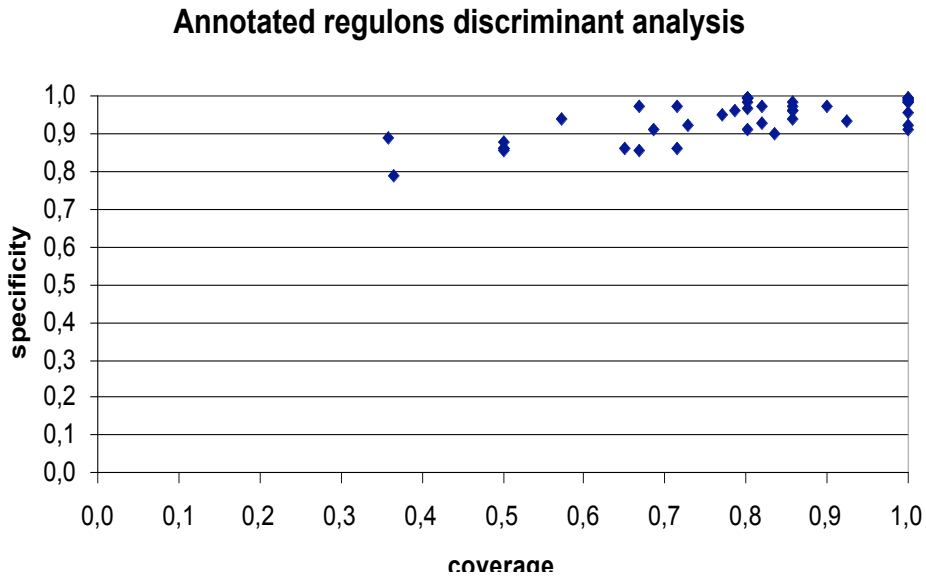


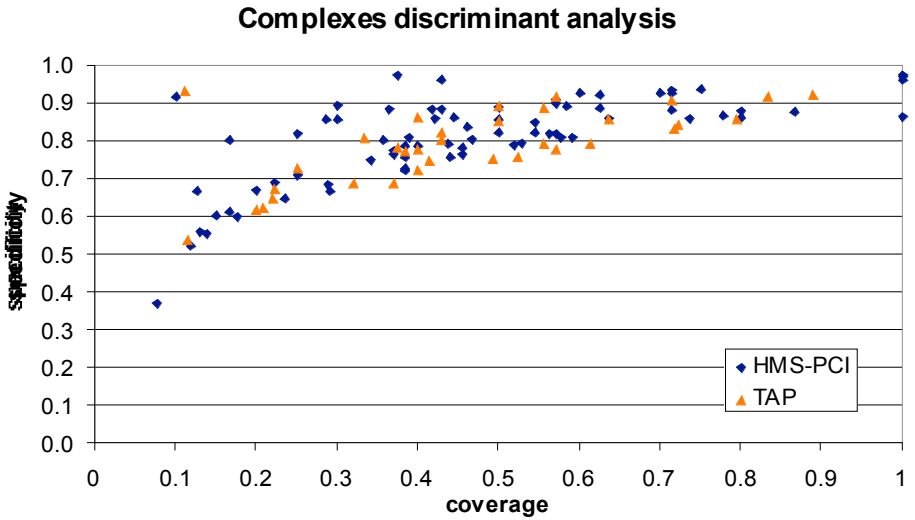
Figure 2

Cov= TP/(TP+FN)
Spec=TP/(TP+FP)

Simonis et al. (2004). Transcriptional regulation of protein complexes in yeast. Genome Biol 5: R33.

Analysis of protein complexes





Simonis et al. (2004). Transcriptional regulation of protein complexes in yeast. Genome Biol 5: R33.

Replication fork complexes

COMPLEX	GENE	ORF	P(da)
DNA polymerase alphal primase complex	*POL1*	*YNL102W*	0.95155315
	PRI2	*YKL045W*	0.92757655
	POL12	*YBL035C*	0.8740073
	PRI1	***YIR008C***	***0.14922267***
DNA polymerase deltaIII	*POL32*	*YJR043C*	0.88528264
	HYS2	*YJR006W*	0.79712089
	CDC2	*YDL102W*	0.79079457
DNA polymerase epsilonII	*POL2*	*YNL262W*	0.85845235
	DPB2	*YPR175W*	0.7575824
	DPB3	*YBR278W*	0.64557167
Exonucleases	*RAD27*	*YKL113C*	0.99432872
PCNA	*POL30*	*YBR088C*	0.69668482
Replication factor A complex	*RFA1*	*YAR007C*	0.98673268
	RFA2	*YNL312W*	0.87875323
	RFA3	*YJL173C*	0.73210048
Topoisomerases	*TOP1*	*YOL006C*	0.82395424
	TOP2	*YNL088W*	0.73032433
DNA helicases	*ECM32*	*YER176W*	0.23499023
	DNA2	*YHR164C*	0.05406037
DNA ligases	*CDC9*	*YDL164C*	0.03708159
DNA polymerase betaIV	*POL4*	*YCR014C*	0.03919368
DNA polymerase gamma	*MIP1*	*YOR330C*	0.03182497
DNA polymerase zeta	*REV7*	*YIL139C*	0.05724429
	REV3	*YPL167C*	0.04090946
Replication factor C complex	***RFC4***	***YOL094C***	***0.50870756***
	RFC5	*YBR087W*	0.33589936
	RFC3	*YNL290W*	0.26089619
	RFC2	*YJR068W*	0.10584985
	RFC1	*YOR217W*	0.05246123
RNase H1	*RNH1*	*YMR234W*	0.02989109

- The replication fork complex regroups 30 genes regrouped in 14 subunits.
- Discriminant analysis classifies the genes in two groups.
  - Genes predicted as regulated by the discovered motifs belong to 7 of the sub-units (15 out of 16 genes)
  - Genes predicted as non-regulated by the discovered motifs belong to the 7 other subunits (13 out of 14 genes).

Simonis et al. (2004). Transcriptional regulation of protein complexes in yeast. Genome Biol 5: R33.

Supplementary material

Jacques.van.Helden@ulb.ac.be
Université Libre de Bruxelles, Belgique
Laboratoire de Bioinformatique des Génomes et des Réseaux (BiGRe)
http://www.bigre.ulb.ac.be/

Discriminant analysis - validation

Internal validation

Leave-one-out

Phosphate microarray data - Prediction phase

Flexible discriminant analysis - pattern counts

- Since we know in advance which genes belong to which family, we can use this information to train a program.
- This is called **supervised classification**.
- There are multiples approaches to supervised classification.
- Results obtained by Flexible Discriminant Analysis (**FDA**)

Confusion table

		Known			
		MET	NIT	PHO	SUM
Predicted	MET	20	0	0	20
	NIT	0	31	1	32
	PHO	0	0	12	12
	SUM	20	31	13	64
	Errors	1	1.56%		
	Correct	63	98.44%		

Flexible discriminant analysis - matrix scores

- Discrimination between MET and PHO genes
- Validation on the basis of the training set (Leave-one-out approach)

Confusion table

		Known			
		CTL	MET	PHO	SUM
Predicted	CTL	80	3	3	86
	MET	1	16	0	17
	PHO	1	0	10	11
	SUM	82	19	13	114
	Errors	8	7.02%		
	Correct	106	92.98%		

Multivariate data with a nominal criterion variable

- One disposes of a set of objects (the **sample**) which have been previously assigned to predefined classes.
- Each object is characterized by a series of quantitative variables (the **predictors**), and its class is indicated in a separated column (the **criterion variable**).

	Predictor variables				**Criterion variable**
	score 1	**score 2**	**…**	**score 15**	**class**
gene 1	$x_{1,1}$	$x_{2,1}$	…	$x_{p,1}$	**PHO**
gene 2	$x_{1,2}$	$x_{2,2}$	…	$x_{p,2}$	**PHO**
gene 3	$x_{1,3}$	$x_{2,3}$	…	$x_{p,3}$	**PHO**
…	…	…	…	…	**…**
gene i	$x_{1,i}$	$x_{2,i}$	…	$x_{p,i}$	**MET**
gene i+1	$x_{1,i+1}$	$x_{2,i+1}$	…	$x_{p,i+1}$	**MET**
gene i+2	$x_{1,i+2}$	$x_{2,i+2}$	…	$x_{p,i+2}$	**MET**
…	…	…			
gene n-1	$x_{1,n-1}$	$x_{2,n-1}$	…	$x_{p,n-1}$	**CTL**
gene n	$x_{1,n}$	$x_{2,n}$	…	$x_{p,n}$	**CTL**

Discriminant analysis - calibration and prediction

- **Calibration phase**
  - The sample is used to build a discriminant function
- **Prediction phase**
  - The discriminant function is used to predict the value of the criterion variable for new objects

	Predictor variables				Criterion variable
	score 1	**score 2**	**…**	**score p**	**class**
gene 1	x_{11}	x_{21}	…	x_{p1}	**PHO**
gene 2	x_{12}	x_{22}	…	x_{p2}	**PHO**
gene 3	x_{13}	x_{23}	…	x_{p3}	**MET**
…	…	…	…	…	**…**
gene ntrain	x_{1n}	x_{2n}	…	x_{pn}	**CTL**

	Predictor variables				Criterion variable
	score 1	**score 2**	**…**	**score p**	**class**
gene 1	x_{11}	x_{21}	…	x_{p1}	**?**
gene 2	x_{12}	x_{22}	…	x_{p2}	**?**
gene 3	x_{13}	x_{23}	…	x_{p3}	**?**
…	…	…	…	…	**…**
gene npred	x_{1n}	x_{2n}	…	x_{pn}	**?**

Regulatory motif profiles

- Each unit on the X axis represents one matrix (5 matrices, 3 scores per matrix)
- Y axis gives the top scores
- **MET genes** have higher scores in columns 10-15 (Met4p and Met31p matrices)
- **PHO genes** have higher scores in columns 1-9 (Pho4p matrices)