Discrimination of Yeast Genes Involved in Methionine and Phosphate Metabolism on the Basis of Upstream Motifs

Didier Gonze, Sylvie Pinloche, Olivier Gascuel, Jacques van Helden

BIOINFORMATICS ORIGINAL PAPER

Vol. 21 no. 17 2005, pages 3490–3500
doi:10.1093/bioinformatics/bti558

Sequence analysis

Discrimination of yeast genes involved in methionine and phosphate metabolism on the basis of upstream motifs

Didier Gonze[1,2], Sylvie Pinloche[3], Olivier Gascuel[3] and Jacques van Helden[1,*]

[1]Service de Conformation des Macromolécules Biologiques et de Bioinformatique, Université Libre de Bruxelles, CP 263, Campus Plaine, Blvd du Triomphe, B-1050 Bruxelles, Belgium, [2]Unité de Chronobiologie Théorique, CP 231, Campus Plaine, Blvd du Triomphe, B-1050 Bruxelles, Belgium and [3]Projet Méthodes et Algorithmes pour la Bioinformatique, LIRMM-CNRS, 161 rue Ada, 34392, Montpellier, France

Received on January 16, 2005; revised on June 15, 2005; accepted on June 27, 2005
Advance Access publication July 5, 2005

ABSTRACT

Motivation: In yeast, methionine and phosphate metabolism are regulated by the complexes Met4p/Met28p/Cbf1p and Pho4p, respectively. The binding sites for these factors share a common core CACGTG. We evaluate our capability to discriminate phosphate- and methionine-responding genes on the basis of putative regulatory elements, despite the similarity between Met4p/Met28p/Cbf1p and Pho4p consensus.

Results: We scanned upstream regions of methionine, phosphate and control genes with position-specific weight matrices for Pho4p, Met4p/Met28p/Cbf1p and Met31p/Met32p, and applied discriminant analysis to classify genes according to matrix matching scores. This analysis showed that matrix scores provided a good discrimination between phosphate, methionine and control genes. The optimal parameters have then been used to predict phosphate and methionine regulation at a genome scale. The genome-scale analysis predicts 37 genes as methionine-regulated and 40 as phosphate-regulated. We compare the predictive results with high throughput data and discuss the difference.

Availability: The programs for sequence retrieval and analysis, as well as the complete data and results, are available on the website on regulatory sequence analysis tools (http://rsat.scmbb.ulb.ac.be/rsat/).

Contact: jvanheld@scmbb.ulb.ac.be

Supplementary information: The complete datasets and results are available at http://rsat.scmbb.ulb.ac.be/rsat/data/published_data/Gonze_MET_PHO/

INTRODUCTION

Living cells respond to changes in their environment by activating or repressing the expression of selected genes. For example, when the intracellular concentration of a given metabolite is too low, a specific transcription factor starts activating the expression of enzymes and transporters involved in the biosynthesis and uptake of this metabolite. Each factor binds to specific sites on the chromosomes and interacts with RNA polymerase to modify the level of expression of the neighbour gene. The specificity of protein–DNA binding is thus the key for restricting the transcriptional response to the appropriate target genes.

*To whom correspondence should be addressed.

The binding specificity of a transcription factor can be described by a pattern, which can be used to detect putative binding sites in new sequences. On the basis of a set of known binding sites, a position-specific scoring matrix (PSSM) can be built to represent the binding specificity of a given transcription factor, and this matrix can be used to scan new sequences to predict putative binding sites (Hertz *et al*., 1990; Hertz and Stormo, 1999; Wasserman and Sandelin, 2004). However, not all matches correspond to effective regulatory elements. Indeed, binding motifs are generally short (typically 5–10 conserved positions) so that, when scanning large sequences, many spurious matches are expected by chance. In addition, the highest specificity does not always correspond to the highest regulatory activity.

A particularly challenging case is the recognition of phosphate- and methionine-responding genes in the yeast *Saccharomyces cerevisiae*. Methionine metabolism is regulated by several transcription factors, with distinct binding sites. The main regulator is the Met4p/Met28p/Cbf1p complex, whose binding consensus is TCACGTGA. Two additional transcription factors have been isolated, Met31p and Met32p, on the basis of their binding to the motif AAACTGTGG (Thomas and Surdin-Kerjan, 1997; Blaiseau *et al*., 1997). Phosphate metabolism is regulated by the transcription factor Pho4p, whose binding sites show two variants : high affinity binding sites are centred on the core CACGTG followed by a short GC-rich region (2–3 nt), whereas medium affinity binding sites have a CACGTT motif followed by a T-rich region (Oshima *et al*., 1996). These two variants of Pho4p binding sites can be summarized with the consensus CACGTKkk, where K means 'G or T', and the two last letters are in lowercase to highlight the fact that they are less conserved.

Thus, Pho4p and Met4p share a common core, CACGTG, and their mutual specificity relies on the flanking bases of this core. Due to this similarity of their consensi, one could expect cross-predictions between Met4p and Pho4p targets. However, we can take benefit from our additional knowledge of phosphate and methionine regulation to establish multi-variate criteria, taking into account the following aspects of the regulation: (1) binding diversity (specificity of the core-flanking bases, two variants of Pho4p binding sites); (2) self-synergy (many genes contain multiple binding sites for either Pho4p or Met4p); (3) heterologous synergy (cooperative regulation by Met4p and Met31p).

In this paper, we apply a discriminant analysis to test whether phosphate- and methionine-responding genes can be discriminated on the basis of putative Met4p, Met31p and Pho4p binding sites. We optimize the parameters on the basis of a set of genes known to be submitted to phosphate and methionine regulation, as well as a control group. Optimal parameters are then used for predicting phosphate and methionine regulation for each gene of the genome. The results of this genome-scale prediction are compared with high-throughput data from ChIP–chip and microarray experiments.

SYSTEMS AND METHODS

Position-specific scoring matrices

Binding sites were collected from TRANSFAC, SCPD and the literature. Altogether, we obtained 21 binding sites for Pho4p, 16 for Met4p and 18 for Met31p. The sizes of the binding sites varied between 16 and 22 bp. The complete data and the sources are provided in the Supplementary material.

PSSMs were constructed by aligning the binding site sequences with the program consensus (matrix width $w = 15$, include both strands as a single sequence) (Hertz *et al*., 1990; Hertz and Stormo, 1999). The resulting matrix M indicates the counts $n_{r,j}$ of each residue r at each position j of the aligned binding sites. This count matrix is automatically converted into weights $W_{r,j}$ by patser:

$$W_{r,j} = \ln\left(\frac{f'_{r,j}}{p_r}\right) = \ln\left(\frac{n_{r,j} + p_r k}{(\sum_{i\in\{\mathrm{A,C,G,T}\}} n_{i,j} + k)\, p_r}\right),$$

where $f'_{r,j} =$ is the relative frequency, corrected with a pseudo-count k, and p_r is the background frequency of the residue r. The position–weight matrix is then used to assign a score X_i to each segment $S_{i,i+w-1}$ of the sequence S.

$$X_i = \sum_{j=1}^{w} W_{r_{i+j-1},j},$$

where r_{i+j-1} is the residue found at position $i + j - 1$ of the sequence S.

High-scoring segments (large X_i) correspond to putative binding sites for the transcription factor. Matrix-based pattern matching was performed with patser (Hertz and Stormo, 1999). This program takes as input a sequence S and a PSSM M of width w, and assigns a score to each position of the sequence as described in Hertz and Stormo (1999). The search was performed on both strands, and the program returned the three top scores per sequence.

DNA–chip data and ChIP–chip data

The phosphate microarray data published by Ogawa *et al.* (2000) was downloaded from http://cmgm.stanford.edu/pbrown/phosphate/ (October 2000). A list of target genes for 106 transcription factors found by a ChIP–chip experiment (Lee *et al*., 2002) were downloaded from http://web.wi.mit.edu/young/regulator_network/ (October 2002). The dataset from Harbison *et al*. (2004) were downloaded from http://web.wi.mit.edu/young/regulatory_code/ (September 2004).

Composition of the training sets

For training and evaluation, we selected 16 genes known to respond to a stress in methionine (MET family), 8 genes submitted to phosphate regulation (PHO family) and a control family (CTL family) containing 80 genes, which are supposed to respond neither to methionine nor to phosphate (Table 1). Note that there are less training genes than binding sites for each transcription factor. This is due to the fact that a promoter generally contains multiple binding sites for the same factor (for example, the 21 sites used for building the Pho4p matrix belong to not more than 8 genes).

Table 1. Gene families used as training set for discriminant analysis

Family	Genes
MET	*ECM17, MET1, MET10, MET14, MET16, MET17, MET2, MET28, MET3, MET30, MET6, MET8, MUP3, SAM1, SAM2, ZWF1*
PHO	*PHO5, PHO8, PHO84, PHO81, PHO11, PHO89, PHO86, SPL2*
CTL	*ASN1, BAR1, CAR1, CAR2, CDC19, CHO1, CHO2, CIT1, COX5A, CTT1, CYB2, CYC1, CYC7, CYT1, DAL5, DMC1, ERG11, FAS1, FAS2, GAL1, GAL2, GAL7, GAL80, GAP1, GCY1, GDH1, HEM13, HEM3, HMG1, HOP1, HSF1, HXT9, ILV1, ILV2, IME1, IME2, INO1, LEU1, LEU2, LEU4, LYS1, LYS2, LYS20, LYS21, LYS4, LYS9, MEK1, MEP1, MEP2, MEP3, MER1, OPI3, PDR10, PDR15, PDR3, PDR5, PET9, PUT1, PUT2, REC102, REC114, RED1, RME1, ROX1, SKI8, SNQ2, SOD2, SPO11, SPO13, SPO16, TOP1, UGA1, UGA2, UGA4, URA1, URA3, URA4, YBR184w, YOR1, ZIP1*

Note that the composition of the control (CTL) family could not be based on direct evidences of an absence of response to phosphate or methionine, since scientific articles generally report positive rather than negative results. For this family, we selected genes involved in some well characterized pathways, and which had no apparent reason to interact with phosphate or methionine metabolism. A gene might of course be involved in multiple pathways, and our CTL family is likely to contain some errors.

Discriminant analysis

Leave-one-out evaluation. Before using the discriminant function for predicting phosphate and methionine regulation for all the genes of a genome it is essential to evaluate its accuracy. Classically, the evaluation relies on a testing set, which must be independent from the training set. Given the restricted number of genes with known class membership (8 PHO and 16 MET genes) splitting them into even smaller subsets would strongly bias the training. To circumvent this, we applied the leave-one-out (LOO) procedure (Huberty, 1994): one element is discarded from the training set, a discriminant function is built on the basis of the remainders and this function is used to predict the class of the discarded element. The predicted class is compared with the training class, and the procedure is iterated over all the elements of the training set.

Variable selection. Another classical problem, when working with a very small training set, is the risk of over-fitting: the accuracy of prediction decreases when the number of variables increases [see Huberty (1994) for a detailed discussion]. With our datasets, there is a risk of over-fitting since the number of variables (15 matrix scores, see Results) is higher than the number of elements in some training classes (8 genes in the PHO group). To circumvent this problem, we implemented a forward stepwise procedure, which selects a subset of variables by optimizing the hit rate (Huberty, 1994): the program first compares the rate of error obtained by using each variable alone and selects the most discriminating one. Additional variables are then successively incorporated by selecting, at each step, the variable that returns the smallest rate of error when combined to the variables retained in the preceding steps. For the evaluation of error rates, the forward stepwise variable selection was performed inside the LOO loop in order to prevent a possible bias on the selected variables. For genome-scale prediction, variable selection was performed with all the training objects.

Linear versus quadratic discriminant analyses. We systematically compared the results obtained with the linear discriminant analysis (LDA) and the quadratic discriminant analysis (QDA). Both methods rely on an hypothesis of multinormality, and LDA (but not QDA) assumes that the training classes have the same covariance matrix. Although these hypotheses are rarely satisfied with real data (and certainly not in the case of pattern counts), discriminant analysis generally gives good results, especially LDA, which requires very few parameter estimates.

Principal component analysis. Since we suspected a problem of over-dimensionality, we compared the results of variable selection in the original data space (matrix scores) and in the principal component analysis (PCA)-transformed data. PCA includes the transformation of a p-dimensional

space of variables into a p-dimensional space of components, where each component is a linear combination of the original variables. The contribution (weight) of each variable to each component is calculated in order to maximize the variance associated with the first component. PCA is often used as a method to reduce the dimensionality of a variable space.

Permutation test. The error rate by itself is not sufficient to evaluate the benefit of the discriminant analysis. Indeed, even a random assignation of class membership would still lead to a certain percentage of correct classification by chance. The random expectation depends on the dataset (relative size of the training groups) and on the parameters of the analysis (linear versus quadratic, selected variables). We performed a permutation test to evaluate the random expectation for the error rate. For this test, we randomly permuted the group labels and applied the same discriminant procedure as with the real labels. For each condition, 100 independent permutation tests were performed and the average error rates were calculated.

Availability and supplementary material

Multivariate analysis and figure drawings were performed with the freeware statistical package R (http://cran.r-project.org/). The programs for sequence retrieval and analysis are available on the web site on regulatory sequence analysis tools (http://rsat.scmbb.ulb.ac.be/rsat/) (van Helden *et al*., 2000; van Helden, 2003). The complete datasets and results are available on the same site (http://rsat.scmbb.ulb.ac.be/rsat/data/published_data/Gonze_MET_PHO/).

RESULTS

Position-specific scoring matrices

A position-based matrix was built for each transcription factor: Met4p [Table 2(A)], Met31p [Table 2(B)] and Pho4p [Table 2(C)]. In addition, we made a specific treatment for the Pho4p factor, which shows two variants of binding sites: high affinity sites, containing a CACGTG core, and medium-affinity sites, centred around a CACGTT core (Oshima *et al*., 1996).

Generally, CACGTG-based sites are followed by a short GC-rich region (2–3 bp), whereas CACGTT-based sites are followed by several other Ts (see Supplementary material). Such dependencies between neighbouring positions are not taken into account by PSSM. As a consequence, a matrix built with all the CACGTG- and CACGTT-based sites [Table 2(x)] would assign a very high score to a sequence like CACGTGTTT, despite the fact that this sequence contains a CACGTG core followed by a T-rich region, a situation which has not yet been observed in any experimentally-proven sites.

In order to better reflect the dependency between the core and the flanking region, we built two separate matrices, regrouping the CACGTG-based [Table 2(D)] and CACGTT-based [Table 2(E)] sites, respectively. The drawback is that each of these matrices is based on a very small number of observations, which might reduce its capability to recognize new sites. We thus combined information obtained with the generic (Pho4p, combining all sites) and the two specific (Pho4p.g and Pho4p.t) matrices. This induces some redundancy, and correlations are to be expected between the matching scores obtained with the generic and each specific matrix. Fortunately, inter-column correlations are taken into account by the discriminant analysis, and should thus not provoke any bias. The utilization of three matrices for Pho4p also increases the number of variables, which increases the risk of over-fitting, but this is not problematic since we apply a forward stepwise variable selection.

Each alignment matrix was converted by patser to a weight matrix and used to scan the whole set of yeast upstream sequences for putative matches. Upstream sequences of all yeast genes were retrieved over 800 bp from the start codon. For each PSSM, the three top scores were collected in order to detect multiple binding sites for the same transcription factor. Each upstream sequence is thus characterized by a 15-dimensional (3 scores $\times$ 5 matrices) vector of scores.

Table 2. PSSMs used to describe the transcription factor binding sites

(A) met4p matrix

A	7	9	0	0	**16**	0	1	0	0	**11**	6	9	6	1	8
C	5	1	4	**16**	0	**15**	0	0	0	3	5	5	0	2	0
G	4	4	1	0	0	0	**15**	0	**16**	0	3	0	0	2	0
T	0	2	**11**	0	0	1	0	**16**	0	2	2	2	10	11	8
	V	R	Y	C	A	C	G	T	G	A	M	M	W	T	W

(B) met31p matrix

A	3	6	9	6	**14**	**18**	**16**	**18**	2	0	0	0	1	3	8
C	8	3	3	2	3	0	1	0	**13**	2	0	1	0	3	6
G	4	3	4	8	0	0	1	0	2	0	**17**	1	**17**	**11**	1
T	3	6	2	2	1	0	0	0	1	**16**	1	**16**	0	1	3
	C	W	A	R	A	A	A	A	C	T	G	T	G	G	M

(C) pho4p matrix

A	0	4	4	1	1	**21**	0	0	0	0	2	2	6	1	7
C	2	7	12	6	**20**	0	**20**	0	1	0	5	5	8	4	6
G	5	1	2	**11**	0	0	0	**21**	0	**15**	**8**	**7**	2	11	2
T	14	9	3	3	0	0	1	0	**20**	**6**	**6**	**7**	5	5	6
	T	Y	C	S	C	A	C	G	T	K	K	K	M	G	H

(D) pho4p.g matrix

A	3	0	2	3	2	0	**14**	0	0	0	0	2	1	5	0
C	2	3	4	7	2	**14**	0	**13**	0	0	0	4	5	6	5
G	3	3	1	2	**8**	0	0	0	**14**	0	**14**	**8**	**6**	1	5
T	6	8	7	2	2	0	0	1	0	**14**	0	0	2	2	4
	T	T	Y	C	G	C	A	C	G	T	G	S	S	M	B

(E) pho4p.t matrix

A	6	0	3	2	1	0	**7**	0	0	0	0	1	2	1	2
C	0	2	2	5	2	**7**	0	**7**	0	1	0	0	0	2	1
G	0	0	0	0	**4**	0	0	0	**7**	0	0	2	1	2	3
T	1	5	2	0	0	0	0	0	0	**6**	**7**	**4**	**4**	2	1
	A	Y	H	M	S	C	A	C	G	T	T	K	W	B	R

The matrix Pho4p.g. is restricted to the Pho4p binding sites containing the CACGTG core. The matrix Pho4p.t. is restricted to the Pho4p binding sites containing the CACGTT core. Below each matrix is indicated the IUPAC consensus.

Comparison of matrix scores on upstream sequences of the training set

Figure 1 plots the scores assigned to the training genes with the different matrices. A simple visual inspection already reveals interesting properties of the PSSM.

Most PHO genes have a high score ($\geq$10) with the Pho4p matrix (Fig. 1A), whereas such a score is observed only once for a MET gene. Reciprocally, most MET genes have a high Met31p score, which is rarely observed in PHO genes. The Pho4p and Met31p matrix thus provide a reasonably good, but not perfect, separation between PHO and MET genes. A few MET genes are mixed with the CTL genes. Figure 1B shows that there is no apparent correlation between Pho4p and Met4p scores, despite the fact that they share the same binding core CACGTG. The difference between flanking residues (compare Table 2A and C) are thus informative.

Figure 1C shows that most MET genes have a high scoring match for both Met31p and Met4p, highlighting the cooperative effect of the two factors. However, some MET genes seem to have binding sites for Met4p alone, or Met31p alone. As expected, PHO genes

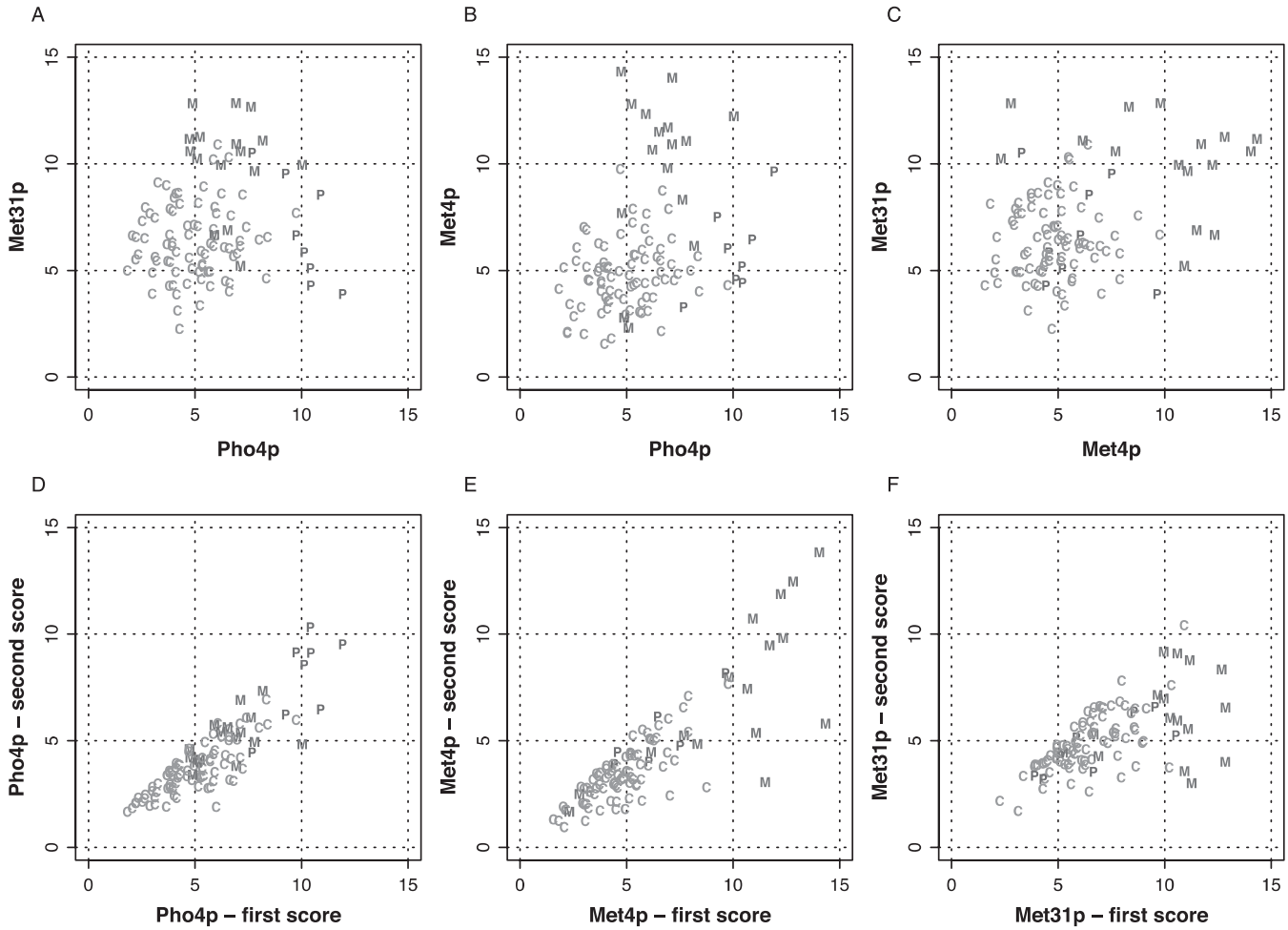

Fig. 1. Comparisons of PSSM scores for upstream sequences of the training set. The letter indicates the gene family (P, PHO; M, MET; C, CTL).

are mixed with CTL as shown in Figure 1C. The combination of Met4p and Met31p matrices separates thus reasonably well MET from non-MET genes.

Figure 1D–F show the two top scoring matches obtained with the Pho4p, Met4p and Met31p matrices, respectively, on each sequence of the training set. It is striking that seven of the eight PHO genes have two very high-scoring matches (Fig. 1D), suggesting the presence of multiple Pho4p binding sites in their upstream sequences. Similarly, most MET genes have at least two very good matches for the Met4p matrix (Fig. 1E). The effect is less pronounced for the Met31p matrix (Fig. 1F): the majority of the MET genes have a very high (>10) first score, but a low (<7) second score.

In summary, PSSMs seem reasonably specific for their expected gene families (Met31p and Met4p matrices are specific for MET genes, and the Pho4p matrices for PHO genes), but each matrix only provides a partial information on the way a gene is regulated. Thus, for classifying genes on the basis of their upstream sequence motifs, one would like to combine information provided by all the matrices. The problem is obviously to find an optimal criterion for weighting the different matrix scores. This can be done with the discriminant analysis, as shown in the next section.

Discriminant analysis with matrix scores

In order to evaluate whether PHO, MET and CTL genes can be discriminated on the basis of upstream motifs, we applied LDA and QDA. We performed two separate analyses for predicting phosphate (PHO against MET + CTL), and methionine (MET against PHO + CTL) regulation, respectively. The evaluation was performed with a LOO test. To prevent the risk of over-fitting, we applied a forward stepwise variable selection.

Error rates were calculated as a function of the number of variables, with different discriminant methods (linear or quadratic). We also tested the effect of data transformation by PCA.

Figure 2 summarizes the results of this evaluation. Each curve represents the rate of error for one discriminant method as a function of the number of selected variables.

For both PHO and MET predictions, better results are obtained with LDA than with QDA. Let us consider the PHO against CTL + MET classification (Fig. 2A): with LDA, the error rate decreases when the 5 first variables are incorporated, after which it remains constant until the 10th variable is incorporated. Optimal discrimination (1% errors) is obtained with 5–9 variables. Under the same condition, a random classification would return 7.7% of errors

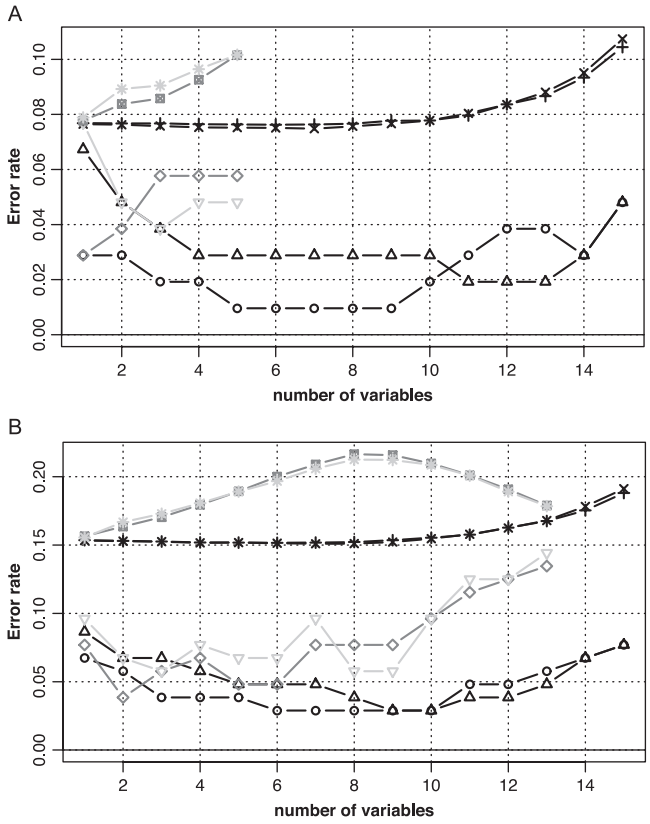

Fig. 2. Error rates obtained with different discriminant analysis approaches (LDA or QDA), on real and permuted data, as a function of the number of variables selected by the forward stepwise procedure. (**A**) PHO against MET + CTL. (**B**) MET against PHO + CTL. Symbols: open circles, LDA on real data; open triangles, LDA on PCA-tranformed data; plus signs, LDA on real data with permuted labels; crosses, LDA on PCA-tranformed data with permuted labels; open diamonds, QDA on real data; open inverted triangles, QDA on PCA-transformed data; squares, QDA on real data with permuted labels; asterisks, QDA on PCA-transformed data with permuted labels.

(LDA permutation curve). The increases of the error rate with the last variables suggests an effect of over-fitting. This is not surprising since the training uses more variables (15) than objects in the PHO training group (eight genes). A similar behaviour is observed for the error curves of the MET against PHO + CTL discrimination (Fig. 2B). It is interesting to note that the error rate increases with the number of variables even with the permuted dataset, which is typical of a situation of over-fitting: when the training is done with more variables than objects, the discriminant function is over-fitted to the training set, even if there is nothing to be learned from it (permuted dataset), and tends to misclassify new objects.

The order of incorporation of the variables is indicative of the information contained in the matrix scores. Not suprisingly, for the PHO against MET + CTL discrimination, the seven first variables correspond to the different scores of Pho4 matrices [Table 3(A)]. The eighth and ninth variables (which neither decrease nor increase the error rate) are the second and third top scores of the Met31 matrix. Consistently, a different subset of variables is used for the MET against PHO + CTL discrimination [Table 3(B)]: the first three

Table 3. Variables selected by the stepwise forward procedure

Rank	Matrix	Score
(A) PHO against MET + CTL		
1	Pho4p	1st
2	Pho4p.t	2nd
3	Pho4p	2nd
4	Pho4p.g	2nd
5	Pho4p.g	1st
6	Pho4p.g	3rd
7	Pho4p.t	3rd
8	Met31p	2nd
9	Met31p	3rd
(B) MET against PHO + CTL		
1	Met4p	1st
2	Met31p	1st
3	Met31p	3rd
4	Pho4p	3rd
5	Pho4p	1st
6	Met31p	2nd
7	Pho4p	2nd
8	Pho4p.t	2nd
9	Pho4p.g	1st
10	Pho4p.t	1st

selected variables are the top scores for the Met4 matrix and the first and third scores of Met31. The next variables are mostly Pho4 scores.

In general, QDA is more powerful than LDA, but in our case the results are opposite. This probably results from the over-fitting problem: the dimensionality of the discrimination criterion increases quadratically with the number of selected variables. QDA could not incorporate more than five variables due to the restricted size of the smallest training class (eight genes for the PHO group). With QDA, over-fitting is already perceptible in the permutation test when the second (Fig. 2A) or third (Fig. 2B) variable is incorporated.

Beyond the raw error rate, it is important to evaluate the types of prediction errors. Table 4 summarizes the number of correct and wrong assignations obtained by LDA for each class of the training set. For the two-group classifications (PHO against MET + CTL and MET against PHO + CTL), all errors consist of 'false negatives', i.e. one PHO and three MET genes are assigned to the CTL group, giving error rates of 1.0 and 2.9%, respectively.

We also tested the direct three-group classification (PHO against MET against CTL), but this raised an error rate of 5.8% [Table 4(E)]. In addition, from the biological point of view, the regulation of a given gene might be achieved by multiple transcription factors. There is thus no reason to impose a priori that a gene belongs to only one class. Indeed, from their experimental observations, O'Connell and Baker (1992) have postulated a possible cross-regulation between phosphate and sulfate metabolism in yeast.

The permutation test shows that the random expectation for the error rate is markedly higher for the three-group classification (23.1%) than for the two-group classifications (15.4% for MET and 7.6% for PHO). The confusion tables [Table 4(B, D and F)] show that when the program is trained with randomly permuted labels the classifier tends to assign all objects to the most frequent class (CTL).

Table 4. Confusion tables on the training set with the optimal linear discriminant functions

Pred	(A)[a] Training		(B)[b] Training (permuted)	
	CTL + PHO	MET	CTL + PHO	MET
CTL + PHO	88	3	87.56 ± 0.76	15.17 ± 1.33
MET	0	13	0.44 ± 0.76	0.83 ± 1.33
Pred	**(C)[c] Training**		**(D)[d] Training (permuted)**	
	CTL + MET	PHO	CTL + MET	PHO
CTL + MET	96	1	95.84 ± 0.48	7.74 ± 0.58
PHO	0	7	0.16 ± 0.48	0.26 ± 0.58

Pred	(E)[e] Training			(F)[f] Training (permuted)		
	CTL	MET	PHO	CTL	MET	PHO
CTL	79	4	0	80	16	8
MET	1	12	1	0	0	0
PHO	0	0	7	0	0	0

Permutation tests were repeated 100 times and the confusion tables indicate average ± SD.
[a]MET against PHO + CTL; Error rate: 2.9%.
[b]Permutation test (MET against PHO + CTL); Error rate: 15% ± 0.8%.
[c]PHO against MET + CTL discrimination; Error rate: 1.0%.
[d]Permutation test (PHO against MET + CTL); Error rate: 7.6% ± 0.3%.
[e]Three-group classification (PHO against MET against CTL); Error rate: 5.8%.
[f]Permutation test (PHO against MET against CTL); Error rate: 23.1%.

The types of errors returned by QDA (Table 5) show similar trends as with LDA (Table 4), but the results are slightly poorer.

Data transformation by PCA did not improve the performances of the classification (Fig. 2).

In summary, two-group classifications with LDA provides a quite stringent discrimination criterion (few false positives), which is essential for the genome-scale prediction.

Genome-scale prediction of methionine and phosphate responding genes

We used the optimal method (LDA) and the variables selected by the stepwise procedure for predicting methionine and phosphate regulation in all yeast genes. This resulted in the prediction of 40 phosphate- and 37 methionine-regulated genes.

Among the PHO-predicted genes (Table 6), all training genes but *PHO8* are recovered. The non-training gene predicted with the highest probability is *STB3*. Almost nothing is known about this gene except that its product binds to Sin3p, a global transcription factor affecting expression of many genes including *INO1*, which itself is involved in phospholipid biosynthesis (Slekar and Henry, 1995). Also predicted with a high probability are *VTC4*, which was recently proposed as a target for *PHO4* (Huang *et al*., 2002), *PHO87*, coding for a phosphate permease, *PHM6*, which might have a role in phosphate metabolism and is regulated by phosphate (Standford Genome Database) and *PMP2*, coding for an H^+-ATPase subunit. These genes are interesting candidates for the experimental analysis.

Among the MET-predicted genes (Table 7), 11 of the 16 training genes are recovered. There are thus five false negatives in the

Table 5. Confusion tables on the training set with the optimal quadratic discriminant functions

Pred	(A)[a] Training		(B)[b] Training (permuted)	
	CTL + PHO	MET	CTL + PHO	MET
CTL+PHO	86	4	87.32 ± 1.00	15.27 ± 1.14
MET	2	12	0.68 ± 1.00	0.73 ± 1.14
Pred	**(C)[c] Training**		**(D)[d] Training (permuted)**	
	CTL + MET	PHO	CTL + MET	PHO
CTL + MET	94	1	95.8 ± 0.49	7.71 ± 0.7
PHO	2	7	0.2 ± 0.49	0.29 ± 0.7

Pred	(E)[e] Training			(F)[f] Training (permuted)		
	CTL	MET	PHO	CTL	MET	PHO
CTL	79	3	0	79	15	8
MET	1	12	1	1	1	0
PHO	0	1	7	0	0	0

[a]MET against PHO + CTL; Error rate: 5.8%.
[b]Permutation test (MET against PHO + CTL); Error rate: 15.3% ± 0.6%.
[c]PHO against MET + CTL discrimination; Error rate: 2.8%.
[d]Permutation test (PHO against MET + CTL); Error rate: 7.6% ± 0.5%.
[e]Three-group classification (PHO against MET against CTL); Error rate: 5.8%.
[f]Permutation test (PHO against MET against CTL); Error rate: 23.1%.

genome-scale prediction, i.e. two more than in the LOO evaluation. This is due to a change in prior probabilities: for the evaluation, the frequencies of training classes were used as priors, whereas during genome-scale predictions we intentionally reduced MET prior probability to 1% in order to minimize false positives.

Some of the predicted genes are interesting targets to look for in effective binding to *MET4* and/or *MET31/MET32*. In particular, genes predicted with a high probability include *CYS3* and *CYS4*, both involved in cystathionine metabolism, *MUP1*, coding for a methionine permease and *MET32*. This latter gene is particularly interesting since it codes for a transcription factor involved in methionine regulation. Its classification among MET genes suggests that it is itself regulated at the transcriptional level, and this might be mediated by Met4p, Met31p or by auto-activation. We did not find any evidence for such a regulation in the literature, but the *MET32* upstream sequence contains a high scoring site for Met31p/Met32p. In constrast, the gene *MET31* is not classified in the MET group by the discriminant procedure and its upstream region does not seem to contain any match for Met31p/Met32p or Met4p. This raises the intriguing hypothesis that the two homologous genes *MET31* and *MET32* could be regulated differently, and this might enable the cell to activate methionine biosynthesis in response to different conditions.

Comparison with experimental genome-scale detection of transcription factor target genes

In order to assess the reliability of our predictions, we compared the results of the discriminant procedures with high-throughput experiments reporting the binding of transcription factors to DNA (Lee *et al*., 2002) and the transcriptional response to phosphate stress (Ogawa *et al*., 2000).

Table 6. Predicted PHO genes

ORF	Train	proba.PHO	Name	Description
YML123C	PHO	9.999987e−1	*PHO84*	High-affinity inorganic phosphate/H+ symporter
YDR169C	NA	9.995144e−1	*STB3*	SIN3 protein-binding protein
YAR071W	PHO	9.984716e−1	*PHO11*	Secreted acid phosphatase
YHR215W	NA	9.984716e−1	*PHO12*	Secreted acid phosphatase
YBR296C	PHO	9.975738e−1	*PHO89*	Na+-coupled phosphate transport protein, high affinity
YAR070C	NA	9.961708e−1	*YAR070c*	Hypothetical protein
YBR093C	PHO	9.946229e−1	*PHO5*	Repressible acid phosphatase precursor
YHR168W	NA	9.944950e−1	*YHR168w*	Similarity to GTP-binding proteins
YJL012C	NA	9.916370e−1	*VTC4*	Similarity to YPL019c and YFl004w
YCR037C	NA	9.814537e−1	*PHO87*	Member of the phosphate permease family
YEL017C-A	NA	9.780367e−1	*PMP2*	H+-ATPase subunit, plasma membrane
YAL002W	NA	9.587233e−1	*VPS8*	Vacuolar sorting protein, 134 kD
YKR050W	NA	9.540507e−1	*TRK2*	Moderate-affinity potassium transport protein
YKR048C	NA	9.535089e−1	*NAP1*	Nucleosome assembly protein I
YHR137W	NA	9.434697e−1	*ARO9*	Aromatic amino acid aminotransferase II
YGR233C	PHO	9.404826e−1	*PHO81*	Cyclin-dependent kinase inhibitor
YCR098C	NA	9.394289e−1	*GIT1*	Glycerophosphoinositol transporter
YHR136C	PHO	9.347150e−1	*SPL2*	Suppressor of plc1-delta
YDR281C	NA	9.324760e−1	*PHM6*	Hypothetical protein, has a role in phosphate metabolism
YJL209W	NA	9.176324e−1	*CBP1*	Apo-cytochrome b pre-mRNA processing protein
YJL211C	NA	9.176324e−1	*YJL211c*	Questionable ORF
YER073W	NA	9.035125e−1	*ALD5*	Aldehyde dehydrogenase (NAD+), mitochondrial
YEL017W	NA	8.647403e−1	*YEL017w*	Hypothetical protein
YDR041W	NA	8.505285e−1	*RSM10*	Component of the mitochondrial ribosomal small subunit
YPL068C	NA	8.354381e−1	*YPL068c*	Hypothetical protein
YNL064C	NA	8.287782e−1	*YDJ1*	Mitochondrial and ER import protein
YER017C	NA	8.124987e−1	*AFG3*	Protease of the SEC18/CDC48/PAS1 family of ATPases (AAA)
YNL061W	NA	7.846272e−1	*NOP2*	Nucleolar protein
YDR054C	NA	7.668280e−1	*CDC34*	E2 ubiquitin-conjugating enzyme
YER019W	NA	7.472275e−1	*ISC1*	Weak similarity to human and mouse neutral sphingomyelinase
YDR310C	NA	7.243479e−1	*SUM1*	Suppressor of SIR mutations
YDR311W	NA	7.243479e−1	*TFB1*	TFIIH subunit (transcription initiation factor), 75 kD
YNL063W	NA	6.909194e−1	*YNL063w*	Weak similarity to Mycoplasma protoporphyrinogen oxidase
YAR064W	NA	6.075119e−1	*YAR064w*	Hypothetical protein
YFL004W	NA	6.052613e−1	*VTC2*	Putative polyphosphate synthetase
YJL117W	PHO	5.368084e−1	*PHO86*	Inorganic phospate transporter
YML121W	NA	5.348837e−1	*GTR1*	GTP-binding protein
YDR055W	NA	5.197217e−1	*PST1*	Strong similarity to SPS2 protein
YEL045C	NA	5.166707e−1	*YEL045c*	Weak similarity to cytochrome c oxidase III of *Trypanosoma brucei* kinetoplast
YDR303C	NA	5.137511e−1	*RSC3*	Similarity to transcriptional regulator proteins

Lee *et al*. (2002) applied a ChIP–chip approach to detect binding between 106 yeast transcription factors and all the yeast intergenic regions, and characterized the reliability of each measurement by a P-value. From this dataset, we selected the ChIP–chip experiment P-values for Pho4p, Met4p and Met31p.

Ogawa *et al*. (2000) performed eight DNA chip experiments to test the transcriptional response of yeast to various phosphate stress conditions (low concentrations or PHO mutants) and selected 21 genes showing a consistent transcriptional response (at least 2-fold regulation in at least five of the eight experiments).

We first compared microarray data from Ogawa *et al*. (2000) with ChIP–chip provided by Lee *et al*. (2002). Suprisingly, the comparison reveals a striking discrepancy between these two datasets (Fig. 3A): genes showing transcriptional response are not detected by the ChIP–chip experiment, and reciprocally. Some PHO training genes appear among the regulated genes, but none of them is detected in the ChIP–chip experiment.

A closer analysis of the ChIP–chip experiment reveals that 62 genes have a P-value $< 10^{-3}$, but that none of these genes corresponds to known Pho4p target genes (Zhu and Zhang, 1999), and, in addition, the upstream sequences of the genes detected by ChIP–chip do not contain the Pho4p binding motif (Simonis *et al*., 2004). The most likely reason for the absence of PHO genes is that Lee and co-workers used the same rich medium for all their experiments. Since it is well known that, in presence of phosphate, Pho4p is inactivated by sequestration in the cytoplasm (Oshima *et al*., 1996) there was not much chance to detect real Pho4p targets in the experimental conditions used, and the 62 reported genes are thus likely to be experimental artefacts. For a similar reason, the Met4p ChIP–chip experiment should also be considered with caution since Met4p

Table 7. Predicted MET genes

ORF	Train	proba.MET	Name	Description
YNL277W	MET	9.996774e−1	*MET2*	Homoserine *O*-acetyltransferase
YJR010W	MET	9.994956e−1	*MET3*	Sulfate adenylyltransferase
YKL001C	MET	9.993998e−1	*MET14*	ATP adenosine-5′-phosphosulfate 3′-phosphotransferase
YLR303W	MET	9.976683e−1	*MET17*	*O*-acetylhomoserine sulfhydrylase
YGR155W	NA	9.950779e−1	*CYS4*	Cystathionine beta-synthase
YDR502C	MET	9.945377e−1	*SAM2*	*S*-adenosylmethionine synthetase 2
YGR154C	NA	9.928820e−1	*YGR154c*	Strong similarity to hypothetical proteins YKR076w and YMR251w
YAL012W	NA	9.901753e−1	*CYS3*	Cystathionine gamma-lyase
YER125W	NA	9.852316e−1	*RSP5*	hect domain E3 ubiquitin-protein ligase
YIL074C	NA	9.554347e−1	*SER33*	3-phosphoglycerate dehydrogenase
YIR017C	MET	9.524668e−1	*MET28*	Transcriptional activator of sulfur amino acid metabolism
YDR253C	NA	9.470668e−1	*MET32*	Transcriptional regulator of sulfur amino acid metabolism
YDR254W	NA	9.416109e−1	*CHL4*	Chromosome segregation protein
YHL036W	MET	9.082197e−1	*MUP3*	Very low affinity methionine permease
YHL038C	NA	9.082197e−1	*CBP2*	Apo-cytochrome b pre-mRNA processing protein 2
YGR055W	NA	9.036226e−1	*MUP1*	High affinity methionine permease
YIR018W	NA	8.547088e−1	*YAP5*	Involved in transcription activation
YOR284W	NA	8.228867e−1	*YOR284w*	Weak similarity to *Methanococcus jannaschii* hypothetical protein MJ0694
YPL250C	NA	8.090201e−1	*ICY2*	Weak similarity to YMR195w
YJL186W	NA	7.677502e−1	*MNN5*	Putative mannosyltransferase
YJL187C	NA	7.677502e−1	*SWE1*	Ser/tyr dual-specifity protein kinase
YNL259C	NA	7.503313e−1	*ATX1*	Antioxidant protein and metal homeostasis factor
YAR064W	NA	7.140508e−1	*YAR064w*	Hypothetical protein
YFR030W	MET	6.990142e−1	*MET10*	Sulfite reductase flavin-binding subunit
YFR049W	NA	6.579113e−1	*YMR31*	Ribosomal protein, mitochondrial
YOR367W	NA	6.502742e−1	*SCP1*	Similarity to mammalian smooth muscle protein SM22 and chicken calponin alpha
YMR061W	NA	6.471349e−1	*RNA14*	Component of pre-mRNA 3′-end processing factor CF I
YNL260C	NA	6.046290e−1	*YNL260c*	Weak similarity to hypothetical protein *Schizosaccharomyces pombe*
YER091C	MET	5.899951e−1	*MET6*	5-methyltetrahydropteroyltriglutamate–homocysteine methyltransferase
YER092W	NA	5.899951e−1	*IES5*	Weak similarity to tryptophan synthase beta subunit—*Aquifex aeolicus*
YDL208W	NA	5.322103e−1	*NHP2*	Nucleolar rRNA processing protein
YKR068C	NA	5.281643e−1	*BET3*	Involved in targeting and fusion of ER to golgi transport vesicles
YKR069W	MET	5.280325e−1	*MET1*	Siroheme synthase
YOR136W	NA	5.078566e−1	*IDH2*	Isocitrate dehydrogenase (NAD+) subunit 2, mitochondrial
YIL046W	MET	5.066017e−1	*MET30*	Involved in regulation of sulfur assimilation genes and cell cycle progression
YIL047C	NA	5.066017e−1	*SYG1*	Member of the major facilitator superfamily
YMR301C	NA	5.010956e−1	*ATM1*	ATP-binding cassette transporter protein, mitochondrial

is inactivated by methionine (Kuras and Thomas, 1995; Thomas and Surdin-Kerjan, 1997). These ChIP–chip data also show a large inconsistency between target genes detected for Met4p and Met31p (Supplementary material).

In a more recent study, Harbison *et al*. (2004) detected the binding regions of yeast transcription factors in different culture media. This study includes detection of PHO4 targets in phosphate-poor medium (Pi−) and of MET4 targets in methionine-poor medium (SM). Figure 3B shows a better consistency between ChIP–chip detection (Pho4p targets in Pi- medium) and expression data, although many genes are detected with ChIP–chip but not by microarray experiments.

Not suprizingly, we found not a single common gene between our PHO predicted genes and the Pho4p targets detected by Lee *et al*. (2002) in the rich medium (Fig. 3C). The comparison with Harbison *et al*. (2004) results is more informative: several genes are detected by both our motif-based PHO predictions and the ChIP–chip detection (Fig. 3D), among which a good fraction is of the annotated Pho4p targets.

The comparison between our PHO predictions and the gene expression data (Fig. 3E) shows that almost all the genes showing a high transcriptional response are detected by the discriminant analysis (top-right corner). These genes include most of the training genes (*PHO12*, *PHO84*, *PHO86*, *SPL2*, *PHO89*, *PHO5* and *PHO11*) as well as some of our *de novo* predictions (*VTC2*, *VTC4* and *PHM6*). However, many other predicted genes do not show any response to phosphate in Ogawa's experiments. These genes include one of the training genes, *PHO81*, suggesting that some real phosphate genes can escape detection in Ogawa's experiment. However, motif-based predictions also contain a certain rate of false positive (see Discussion).

Genes classified as MET by the discriminant procedure poorly correspond to those detected by the ChIP–chip data in methionine-poor medium (Fig. 3F). Only five genes are detected by both our

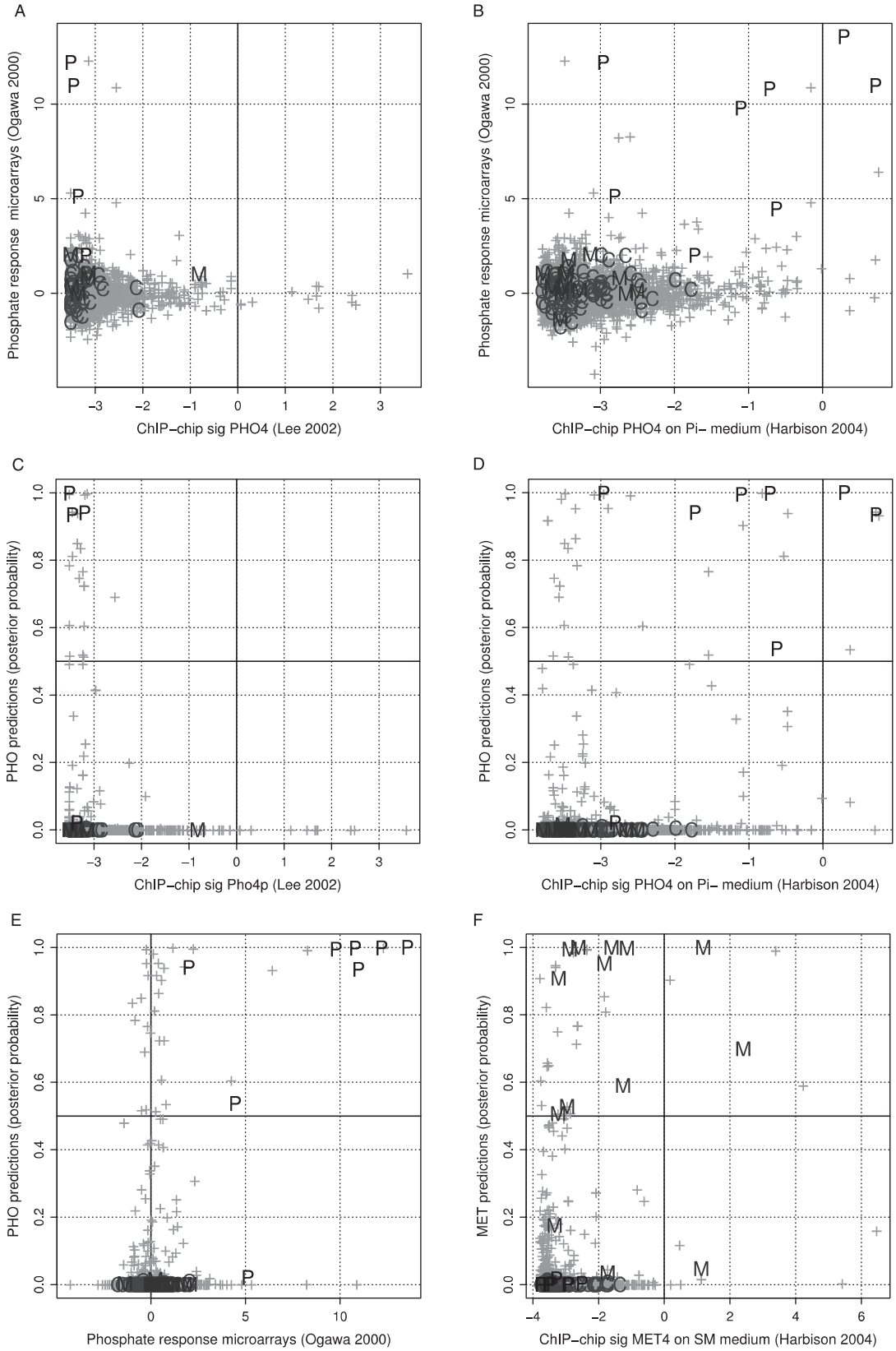

Fig. 3. Comparison of matrix score based predictions with high-throughput data on DNA binding (Lee *et al*., 2002) and on transcriptional response to phosphate Ogawa *et al*. (2000). Genes used in the training set are labelled (P, PHO; M, MET; C, CTL). Note that some training genes are missing on the graph because the ChIP–chip dataset contains 3295 genes and Ogawa's data 5783 genes among the 6345 genes in the complete yeast genome.

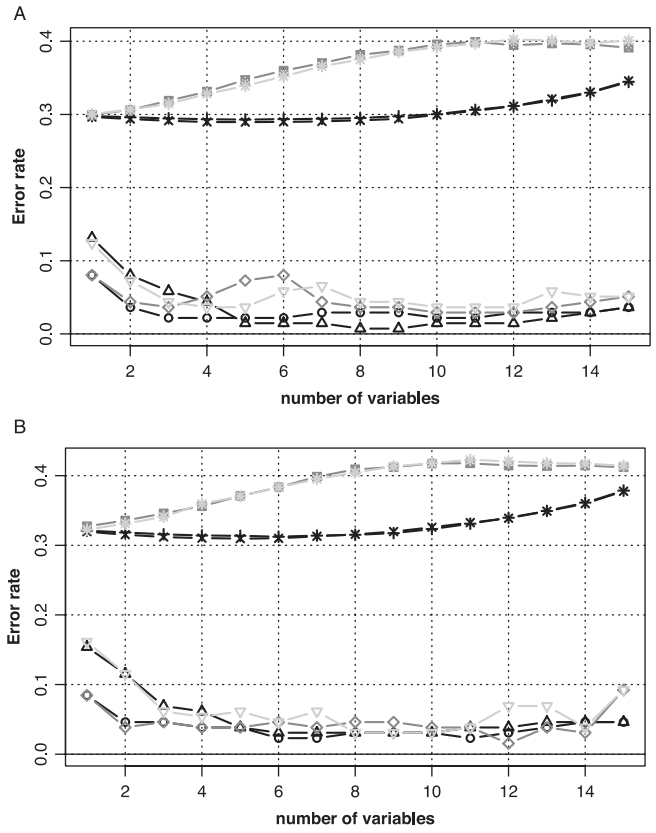

Fig. 4. Error rates obtained with the expanded training groups. See Figure 2 for legends.

predictions and the ChIP–chip experiment, including two of the MET training genes. Nine other training genes are detected by our predictions but not in the ChIP–chip experiment, whereas one is detected by the ChIP–chip and not predicted by our method.

Discriminant analysis with an expanded training set

It is somewhat surprising to notice that the simplest method (LDA) gives better performances than more elaborate treatments (QDA, PCA-transformation before LDA or QDA) (Fig. 2). This effect might result from the small number of training objects, in particular for the PHO group (eight genes). In order to test this possibility, we performed the same analysis with a larger training set, regrouping the original training set (proven *PHO* and *MET* genes) with those predicted by our LDA procedure.

On the error curves obtained with the expanded training set (Fig. 4), QDA indeed gives better results than with the small training sets, and its performances are similar to LDA. The differences between PCA-transformed and non-transformed data are also less sensitive with the expanded dataset, and, for PHO against CTL + MET classification, the best results are obtained with LDA on the PCA-transformed data.

DISCUSSION

The first question addressed in this paper was whether PSSMs would allow us to distinguish PHO and MET genes, despite the high similarity between the consensus of their main regulator (TCACGTGA for Met4p and CACGTKkk for Pho4p). Discriminant analysis using the LOO evaluation shows that there is no confusion between the training PHO and MET genes. The PSSMs are thus sufficiently informative, despite the small number of sites used for building them (no more than seven sites for Pho4p.t). An essential reason for the absence of cross-predictions is the integration of multiple criteria. For the MET family, two distinct PSSMs were used to reflect the synergic regulation by multiple transcription factors (Met4p and Met31p, respectively). The comparison of scores obtained with Met31 and Met4 matrices shows that some genes are essentially regulated by Met4p, some by Met31p and some by both factors. The combination of matrices thus provides better information to distinguish MET sequences from the PHO and CTL groups. We also used multiple scores to reflect the binding diversity due to the specificity of the core-flanking bases (e.g. Pho4p.t and Pho4p.g matrices) and self-synergy assured by multiple binding sites (selection of the three top scores for each sequence). This multi-variate representation gives better classification than any matrix taken alone. However, it is essential to restrict the number of predictive variables in order to avoid the trap of over-fitting (Fig. 2).

The second question was whether the classification based on putative binding motifs would allow to predict phosphate or methionine response at a genome-scale. These genome-scale predictions should be taken with caution, for various reasons. (1) The number of genes is so large (6345) that even a small risk of error would result in many misclassifications. To avoid this, we deliberately chose low prior probabilities (1%) for PHO and MET classes, and this has a cost in terms of sensitivity (we miss 5 out of the 16 MET genes). The predictions can thus certainly not be considered as exhaustive. (2) When two neighbour genes are transcribed in divergent directions they share the same intergenic region. In such cases, we cannot predict whether the binding sites are involved in the regulation of the gene at their left, at their right, or both. (3) The binding of a transcription factor is not always sufficient to confer a transcriptional regulation. Note that the restrictions (2) and (3) also apply for the interpretation of ChIP–chip data, since this method detects the binding of a transcription factor to an intergenic region.

Having in mind all the restrictions above, genome-scale predictions can nevertheless give some useful information. It is interesting to note that among the 6345 yeast genes not a single one is predicted as both phosphate and methionine regulated. Another observation is that, despite the simplicity of its underlying model, LDA gives better results than QDA. Paradoxically, our attempts to use more sophisticated methods (SVM) resulted in a lower hit rate (not shown). It is well known that each classifier has its own range of applications, and apparently with the type of data treated here good results can be obtained with one of the simplest classifiers. It is likely that the accuracy of LDA observed in this case comes from the very small size of the training set. In cases where larger training sets are available, other methods that rely on the estimation of more parameters (SVM, QDA) might become more efficient. This hypothesis seems to be supported by our observation that the differences between LDA and QDA are reduced when the classification is performed with the expanded training set combining annotated and predicted genes.

The comparison between our predictions and the gene expression and ChIP–chip data suggests that pattern-based prediction of gene regulation can be very helpful as a complement to high-throughput data. The high rate of false-positives returned by some methods can be reduced by selecting the most consistent results (intersection between sets of genes detected by the different methods).

Alternatively, sequence-based predictions could be used to detect potential targets, which for some biological (culture conditions) or technical (noise) reason escaped the detection by high-throughput methods. These predicted targets could then be submitted to a more precise experimental characterization.

ACKNOWLEDGEMENTS

The authors are grateful to Jerry Hertz, who kindly adapted his *patser* program in order to report multiple top matches per sequence. D.G. is Chargé de Recherches du Fonds National Belge de la Recherche Scientifique. This research was partly funded by the Actions de Recherche Concertée de la Communauté Française de Belgique (contracts ARC-02/07-291 and ARC-04/09-307). S.P. and O.G. were supported by the Fench Réseau National des Génopoles (RNG).

Conflict of Interest: none declared.

REFERENCES

Blaiseau,P.L. *et al*. (1997) Met31p and Met32p, two related zinc finger proteins, are involved in transcriptional regulation of yeast sulfur amino acid metabolism. *Mol. Cell. Biol.*, **17**, 3640–3648.

Harbison,C.T. *et al*. (2004) Transcriptional regulatory code of a eukaryotic genome. *Nature*, **431**, 99–104.

Hertz,G.Z. and Stormo,G.D. (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, **15**, 563–577.

Hertz,G.Z. *et al*. (1990) Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput. Appl. Biosci.*, **6**, 81–92.

Huang,D. *et al*. (2002) Dissection of a complex phenotype by functional genomics reveals roles for the yeast cyclin-dependent protein kinase Pho85 in stress adaptation and cell integrity. *Mol. Cell. Biol.*, **22**, 5076–5088.

Huberty,C.J. (1994) *Applied Discriminant Analysis*. Wiley series in Probability and Mathematical Statistics, John Wiley & Sons, New York.

Kuras,L. and Thomas,D. (1995) Functional analysis of Met4, a yeast transcriptional activator responsive to *S*-adenosylmethionine. *Mol. Cell. Biol.*, **15**, 208–216.

Lee,T.I. *et al*. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

O'Connell,K.F. and Baker,R.E. (1992) Possible cross-regulation of phosphate and sulfate metabolism in *Saccharomyces cerevisiae*. *Genetics*, **132**, 63–73.

Ogawa,N. *et al*. (2000) New components of a system for phosphate accumulation and polyphosphate metabolism in *Saccharomyces cerevisiae* revealed by genomic expression analysis. *Mol. Biol. Cell*, **11**, 4309–4321.

Oshima,Y. *et al*. (1996) Regulation of phosphatase synthesis in *Saccharomyces cerevisiae*—a review. *Gene*, **179**, 171–177.

Simonis,N. *et al*. (2004) Combining pattern discovery and discriminant analysis to predict gene co-regulation. *Bioinformatics*, **20**, 2370–2379.

Slekar,K.H. and Henry,S.A. (1995) SIN3 works through two different promoter elements to regulate INO1 gene expression in yeast. *Nucleic Acids Res.*, **23**, 1964–1969.

Thomas,D. and Surdin-Kerjan,Y. (1997) Metabolism of sulfur amino acids in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, **61**, 503–532.

van Helden,J. (2003) Regulatory sequence analysis tools. *Nucleic Acids Res.*, **31**, 3593–3596.

van Helden,J. *et al*. (2000) A web site for the computational analysis of yeast regulatory sequences. *Yeast*, **16**, 177–187.

Wasserman,W.W. and Sandelin,A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.