Machine Learning Techniques to Identify Putative Genes Involved in Nitrogen Catabolite Repression in the Yeast *Saccharomyces cerevisiae*

Kevin Kontos[1] Patrice Godard[2] Bruno André[2]
Jacques van Helden[3] Gianluca Bontempi[1]

[1]Machine Learning Group, Université Libre de Bruxelles (ULB), Belgium

[2]Physiologie Moléculaire de la Cellule, IBMM, ULB, Belgium

[3]Conformation des Macromolécules Biologiques, ULB, Belgium

September 24, 2007 – MLSB 2007

Outline

Outline

Nitrogen Catabolite Repression (NCR)

- Nitrogen is an essential nutrient for all life forms.
- Yeast can use almost 30 distinct nitrogen-containing compounds.
- Yeast utilizes good nitrogen sources in preference to poor ones, and NCR is the mechanism for achieving this selectivity.
- NCR consists in the specific inhibition of transcriptional activation systems of genes needed to degrade poor nitrogen sources.

Which Genes are Involved in NCR?

- All known nitrogen catabolite pathways are regulated by four regulators (Gln3, Gat1, Dal80, and Deh1).
- Biologists would like to discover all genes involved in NCR.
- Current "knowledge":
  - A list of 37 ($= 41 - 4$) annotated NCR genes (ANCR).
  - Three genome-wide experimental and bioinformatics studies [Bar-Joseph et al., 2003, Godard et al., 2007, Scherens et al., 2006].

ML Techniques for Identifying Putative NCR genes

- The proposed approach extends a method [Simonis et al., 2004] which has been successfully used for inferring NCR genes [Godard et al., 2007].
- Classification problem:
  - Input: counts of over-represented motifs (variables) in the non-coding upstream sequences of the yeast genes (samples).
  - Output: the list of annotated NCR genes (ANCR, positive training set) and random gene selections (negative training set).

Our Approach

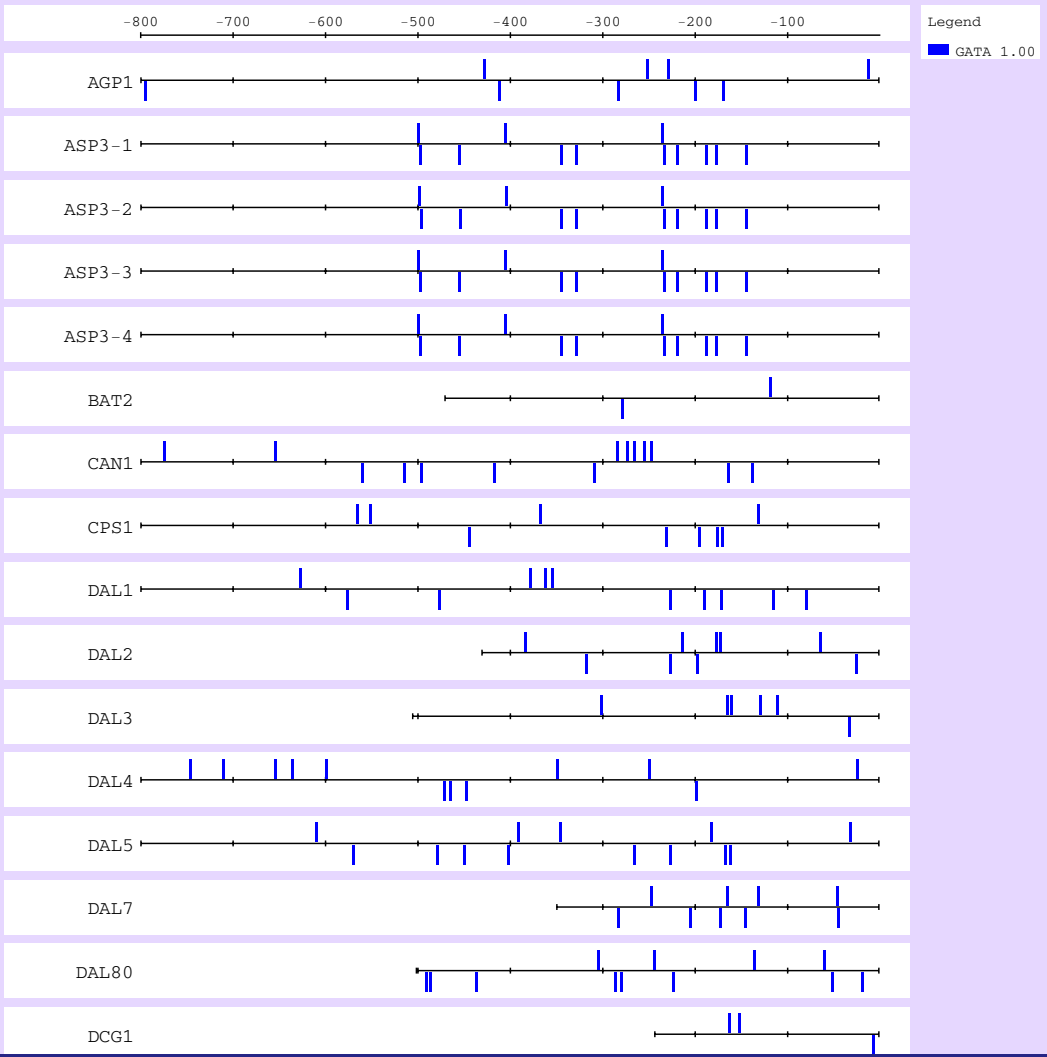
- Focuses on the `GATA` motif in the upstream non-coding sequences.
- Not restricted to counts of pattern occurrences in the upstream non-coding sequences.
- Uses a negative training set.
- Compares different classifiers.

Outline

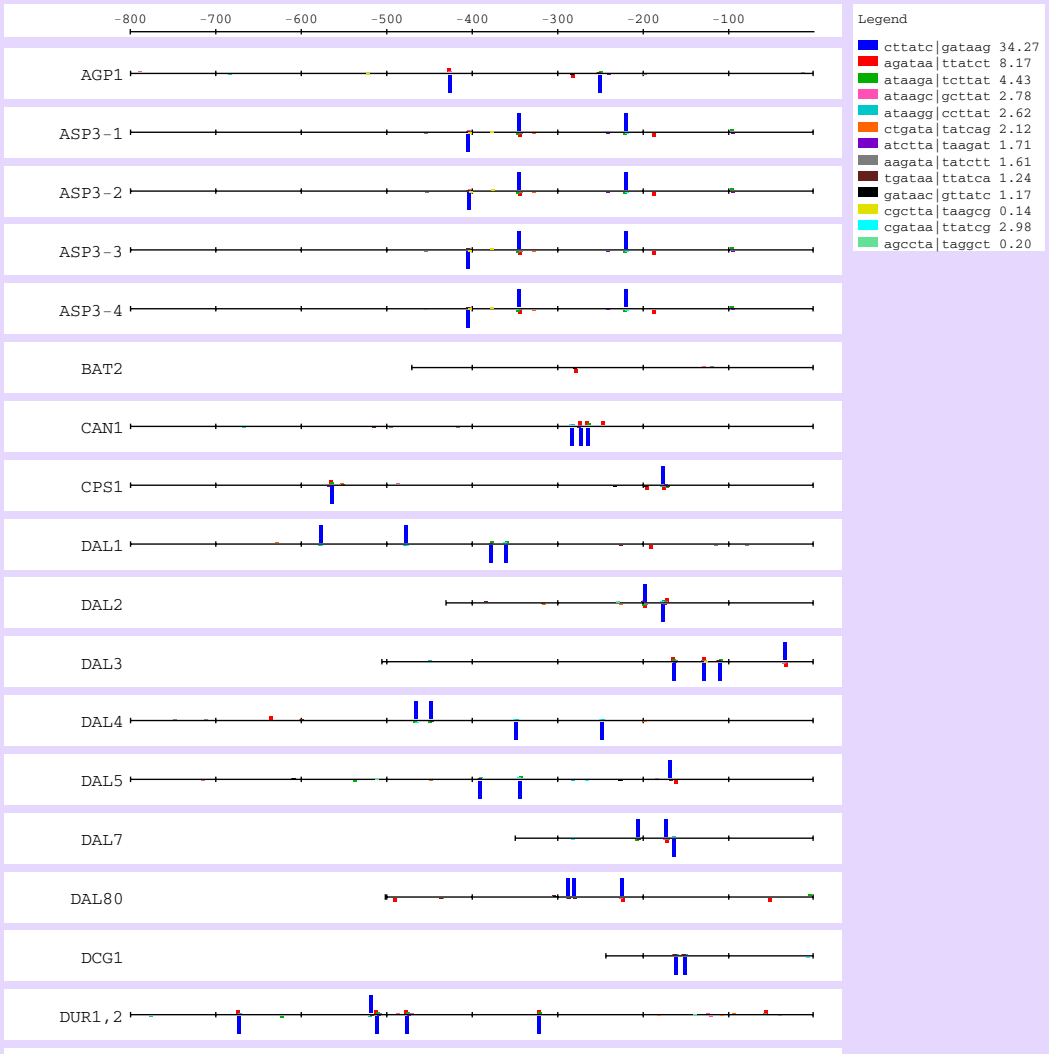
GATA Box

- The promoter regions of NCR target genes typically contain several 5'-GATA-3' core sequences recognized by the GATA family transcription factors.
- GATAAG, GATAAH, GATTA ...

GATA Box


-800
-700
-600
-500
-400
-300
-200
-100
Legend
GATA 1.00
AGP1
ASP3-1
ASP3-2
ASP3-3
ASP3-4
BAT2
CAN1
CPS1
DAL1
DAL2
DAL3
DAL4
DAL5
DAL7
DAL80
DCG1

GATA Box


-800
-700
-600
-500
-400
-300
-200
-100
AGP1
ASP3-1
ASP3-2
ASP3-3
ASP3-4
BAT2
CAN1
CPS1
DAL1
DAL2
DAL3
DAL4
DAL5
DAL7
DAL80
DCG1
DUR1,2
Legend
cttatc|gataag 34.27
agataa|ttatct 8.17
ataaga|tcttat 4.43
ataagc|gcttat 2.78
ataagg|ccttat 2.62
ctgata|tatcag 2.12
atctta|taagat 1.71
aagata|tatctt 1.61
tgataa|ttatca 1.24
gataac|gttatc 1.17
cgctta|taagcg 0.14
cgataa|ttatcg 2.98
agccta|taggct 0.20

Upstream Sequence Variables

Based on biological expertise, we considered 586 variables:

Abbreviation	Description
NUM	Number of GATA boxes.
LG	Sequence length.
F-POS, L-POS	Start positions of the first and of the last GATA box.
M-POS, MI-POS, SD-POS	Mean, median and standard deviation of the start positions of all the GATA boxes.
1-GAP, 2-GAP, 3-GAP, B-GAP	First, second and third smallest, and biggest gaps.
M-GAP, MI-GAP, SD-GAP	Mean, median and standard deviation of the gaps.
k-MINDIST	Minimum distance including k ($2 \leq k \leq 5$) GATA boxes.
UP-k-MER	`k{1,3}GATA`
DOWN-k-MER	`GATAk{1,3}`
GAP-k-MER	`k{1,2}GATAk{1,2}`

Training Sets

- As a positive training set, we used a set of 37 genes previously annotated as NCR-responding (ANCR).
- The negative training set is composed of 89 manually-selected genes known to be insensitive to NCR, most of which being involved in house-keeping cellular functions unrelated to nitrogen metabolism (NNCR).

Overview

The approach consists in formulating the problem of inferring putative NCR genes as a classification problem.

Variable Selection

Since the number of variables is greater than the number of samples, a variable selection step is performed to improve prediction performance.

We compared two variable selection methods:

- Filter method.
- Wrapper method.

Filter Method

A filter method based on the Gram-Schmidt orthogonalization procedure.

The ranking of variables through orthogonalization has many interesting features:

- It is computationally fast.
- It takes into account the collinearity between variables.
- It allows an incremental construction of the model, so that training can be terminated without using all variables.

Wrapper Method

- A wrapper method consisting of a forward stepwise procedure where the prediction performance is assessed by means of stratified 10-fold cross-validation.
- By using the prediction performance of a given learning machine to assess the relative usefulness of subsets of variables, wrappers offer a simple and powerful way to address the problem of variable selection.
- A greedy search strategy, such as forward selection, is both computationally advantageous and robust against overfitting.

Performance Assessment

- Each combination of variable selection strategy and classifier is assessed through stratified 10-fold cross-validation.
- The performance measure used is the balanced error rate (BER), defined as the average of the errors on each class.
- The threshold on the corrected posterior probability is fixed at 0.5.

Posterior Probability Correction

- The a priori probabilities of the training set (37 ANCR and 89 NNCR) do not reflect the expected a priori probabilities of the target classes ($\sim$ 200 NCR genes out of $\sim$ 6000).
- We adjusted the posterior probabilities returned by the classifiers by setting the "expected" priors to $200/N$ (where $N = 5869$ is the total number of genes considered).

Classifiers

We compared three classifiers:

- Naïve Bayes (NB)
- *k*-nearest-neighbors (KNNs)
- Linear support vector machines (SVMs)

Outline

"Gold Standard"

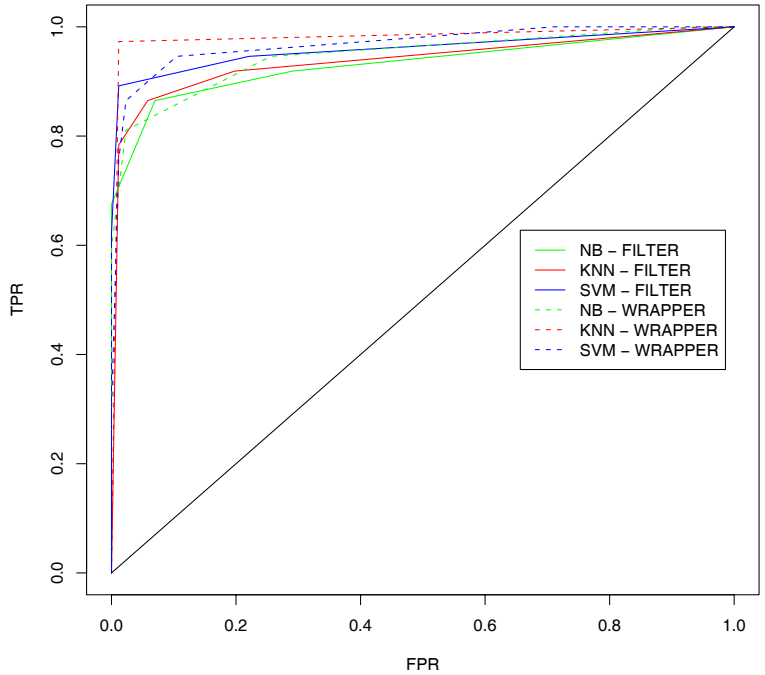
Training sets: ANCR and NNCR.

Test sets:

1. ANCR and NNCR.

Prediction of the genes used to train the classifiers (37 ANCR and 89 NNCR) is performed through leave-one-out.

ROC Curves – ANCR+NNCR



VS	CLASS	AUC
Filter	NB	0.93
	KNN	0.90
	SVM	0.93
Wrapper	NB	0.95
	KNN	0.97
	SVM	0.95

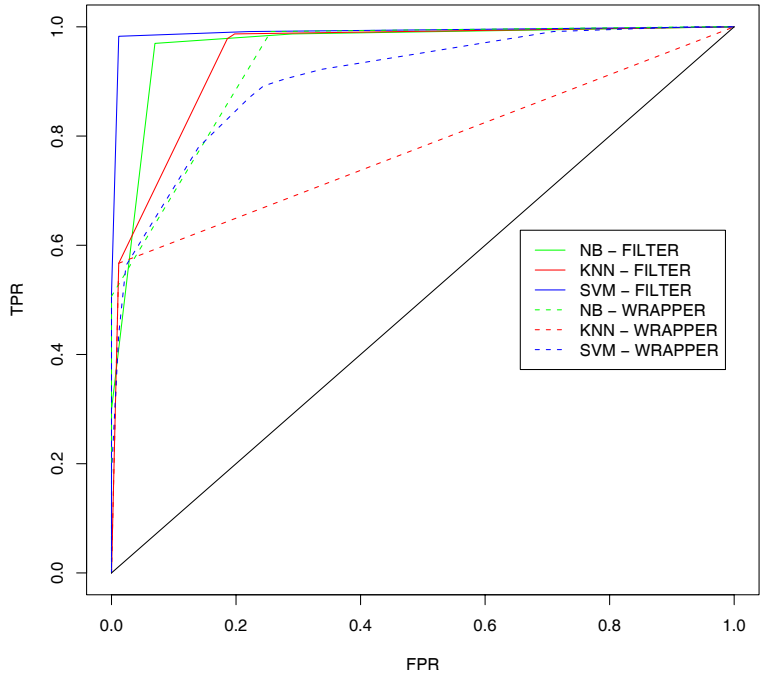
"Gold Standard" (cont.)

Training sets: ANCR and NNCR.

Test sets:

2. ANCRext (ANCR set extended with the putative NCR genes provided by [Bar-Joseph et al., 2003], [Godard et al., 2007], and [Scherens et al., 2006]) and NNCR.

ROC Curves – ANCRext+NNCR



VS	CLASS	AUC
Filter	NB	0.95
	KNN	0.91
	SVM	0.98
Wrapper	NB	0.91
	KNN	0.66
	SVM	0.88

Negative Control

- We wish to determine whether the results are significant or not.
- In addition, the variable selection procedure presents a risk of overfitting the selected variables to the training set.
- An empirical way to estimate the random rate of correct classification is to run the same procedure but with randomized data sets obtained by randomly sampling the labels of the training set.

Gene Set Comparisons

- For each combination of variable selection method and classifier, we compute the F-measure of the overlap of the predicted NCR genes and the sets of [Bar-Joseph et al., 2003], [Godard et al., 2007], and [Scherens et al., 2006], respectively.

$$F\text{-measure} = \frac{2pr}{p+r} \ ,$$

where p=precision= $\frac{TP}{TP+FP}$ and r=recall= $\frac{TP}{TP+FN}$

Gene Set Comparisons (cont.)

To assess the significance of the overlap between two sets, overlapping *P*-values are computed on the basis of the cumulative distribution function of the hypergeometric distribution.

Gene Set Comparisons (cont.)

VS	CLASS	F-measure (P-value)		
		[Bar-Joseph et al., 2003]	[Godard et al., 2007]	[Scherens et al., 2006]
Filter	NB	0.05 (2.9×10^{-16})	0.09 (3.5×10^{-7})	0.06 (2.4×10^{-13})
	KNN	0.06 (9.4×10^{-9})	0.09 (4.8×10^{-5})	0.07 (1.1×10^{-7})
	SVM	0.11 (1.5×10^{-13})	0.15 (9.0×10^{-10})	0.14 (8.2×10^{-14})
Wrapper	NB	0.07 (9.1×10^{-11})	0.11 (7.7×10^{-18})	0.08 (4.33×10^{-16})
	KNN	0.12 (7.7×10^{-14})	0.20 (7.0×10^{-28})	0.16 (5.2×10^{-26})
	SVM	0.13 (8.87×10^{-11})	0.16 (7.18×10^{-14})	0.13 (2.62×10^{-11})

Analysis of the Selected Variables

- The improvement of prediction performance with variable selection is confirmed by the preliminary results obtained with classifiers trained using all variables (data not shown).
- The top selected variables are k-mers (UP-k- MER, DOWN-k-MER and GAP-k-MER).
- `GATAAG`, `TAGATAA`, `GATAGG`, `GTAGATA`

Outline

Conclusion

- All classifiers were able to identify significant number of genes identified as NCR-responding genes in three experimental and bioinformatics studies.
- Variable selection improves performance.
- Interesting selected variables.

Future Work

- Analysis of selected variables.
- Comparison to [Godard et al., 2007]'s results.
- One-class classification (outlier detection).
- The robustness assessment of our approach with respect to the negative training set (extending this set with randomly selected genes).
- Experimentally testing the NCR-sensitivity of the putative NCR genes identified.

Thank you for your attention!

Bar-Joseph, Z., Gerber, G., Lee, T., Rinaldi, N., Yoo, J., Robert, F., Gordon, D., Fraenkel, E., Jaakkola, T., Young, R., et al. (2003).

Computational discovery of gene modules and regulatory networks.

Nature Biotechnology, 21(11):1337–1342.

Godard, P., Urrestarazu, A., Vissers, S., Kontos, K., Bontempi, G., van Helden, J., and André, B. (2007).

Effect of 21 different nitrogen sources on global gene expression in the yeast *Saccharomyces cerevisiae*.

Molecular and Cellular Biology, 27(8):3065–3086.

Scherens, B., Feller, A., Vierendeels, F., Messenguy, F., and Dubois, E. (2006).

Identification of direct and indirect targets of the Gln3 and Gat1 activators by transcriptional profiling in response to nitrogen availability in the short and long term.

FEMS Yeast Research, 6(5):777–791.

Simonis, N., Wodak, S. J., Cohen, G. N., and van Helden, J. (2004).

Combining pattern discovery and discriminant analysis to predict gene co-regulation.

Bioinformatics, 20(15):2370–2379.