Extracting yeast stress genes by dependencies between stress treatments

Arto Klami[a,b], Janne Nikkilä[a,b], Christophe Roos[c], Samuel Kaski[a,b]

a Department of Computer Science, University of Helsinki, Finland
b Neural Networks Research Centre, Helsinki University of Technology, P.O. Box 5400, FI-02015 HUT, Finland
c Medicel Oy, Helsinki, Finland

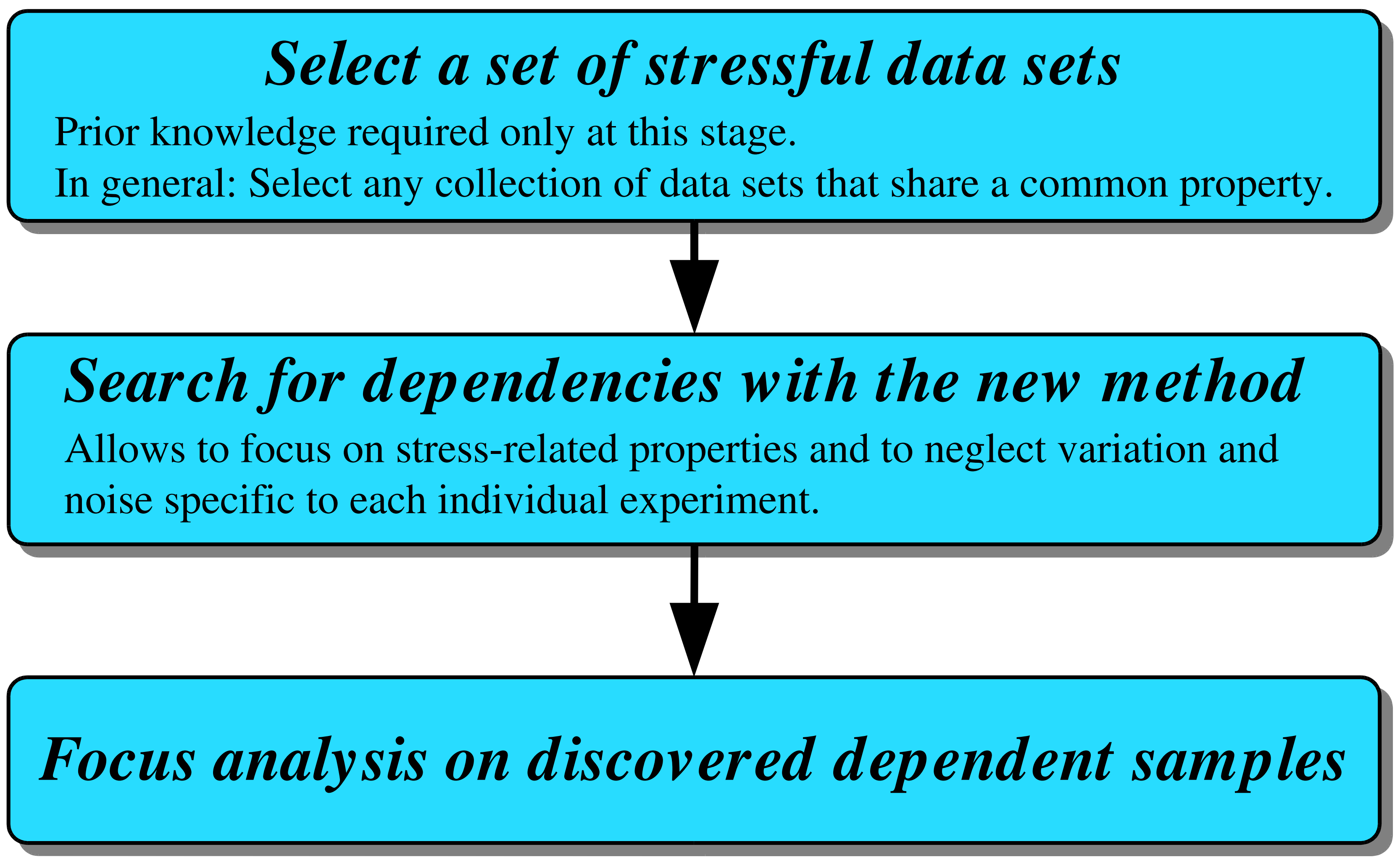
INTRODUCTION

It is much harder to define stress response on gene expression level than it is to name treatments that are stressful for yeast. This inspires a data-driven definition of stress as a response shared by multiple stressful treatments. More specifically, we assume that statistical dependencies between gene expressions in stress data sets are due to stress, because stress is the main common property of the data sets.

We search for environmental stress response genes in yeast with a new non-parametric method that detects genes having notable dependencies. The approach is generally applicable to other analogous problems where it is easier to pick a group of data sets than to specify a model for the biological process.

The method is non-parametric and completely data-driven; nothing needs to be known a priori, on top of collecting the set of stress treatments. Compared to other methods applicable to modeling dependencies, namely Bayes networks or other generative models of data, the method needs less assumptions. It is an exploratory method whose outputs can subsequently be used for building more specific hypotheses and models.

OVERVIEW OF METHOD



Select a set of stressful data sets

Prior knowledge required only at this stage.
In general: Select any collection of data sets that share a common property.

Search for dependencies with the new method

Allows to focus on stress-related properties and to neglect variation and noise specific to each individual experiment.

Focus analysis on discovered dependent samples

METHOD

- Data sets are statistically dependent if their joint variation is not explained by the variations of individual data sets
- Dependency of *variables* is traditionally measured by mutual information, which is zero for independent variables
- Here independent *samples* are sought for, and eventually discarded, based on a *decomposition of mutual information* (details skipped here)
    - Point-wise mutual information, here called *dependency value*, of the *i*th sample is

$$DV(i)=\log\frac{p(\boldsymbol{x}_1(i),\boldsymbol{x}_2(i),\ldots,\boldsymbol{x}_n(i))}{p(\boldsymbol{x}_1(i))\,p(\boldsymbol{x}_2(i))\ldots p(\boldsymbol{x}_n(i))}\quad,$$

where $\boldsymbol{x}_j(i)$ is the *i*th multivariate measurement of the *j*th data set and p(.) denotes the probability

I Low dependency: sample involved in effects specific to few data sets, or small global effects

II Average dependency: independent sample

III High dependency: sample involved in large effect common to most data sets

- Probabilities estimated with leave-one-out Parzen kernel estimators

$$\hat{p}(x_j(i))=\sum_{k\neq i}K(x_j(i),x_j(k))$$

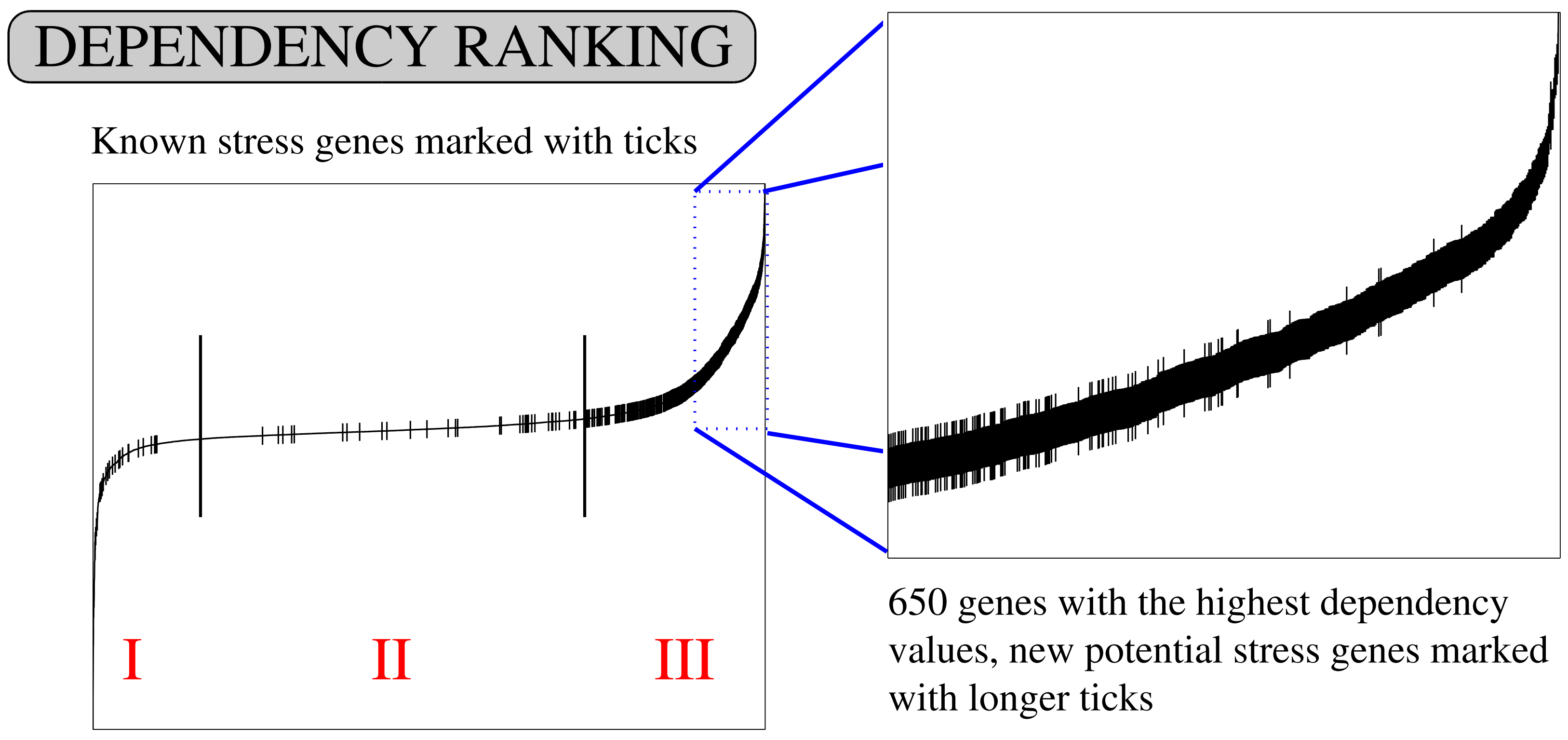
- Gaussian kernels, independent variables within each data set
- Bootstrap to increase robustness and to estimate confidence intervals for the dependency values
- Detection of independent samples by testing whether the dependency value differs significantly from independence

More information: ***http://www.cis.hut.fi/projects/mi/***

FINDING YEAST STRESS GENES

- 17 data sets from different stress treatments, obtained from two different sources (Gasch et al. 2000, Causton et al. 2001)
- Heat (3), acid, alkali, peroxide, NaCL, sorbitol (2), H202, menadione, dtt (2), diamide, hypo-osmotic, aminoacid starvation, and nitrogen depletion
- Each data set is a short time series of logarithmic expression values, translated so that the value at the beginning of the series is zero
- 6013 common genes and 108 dimensions in total

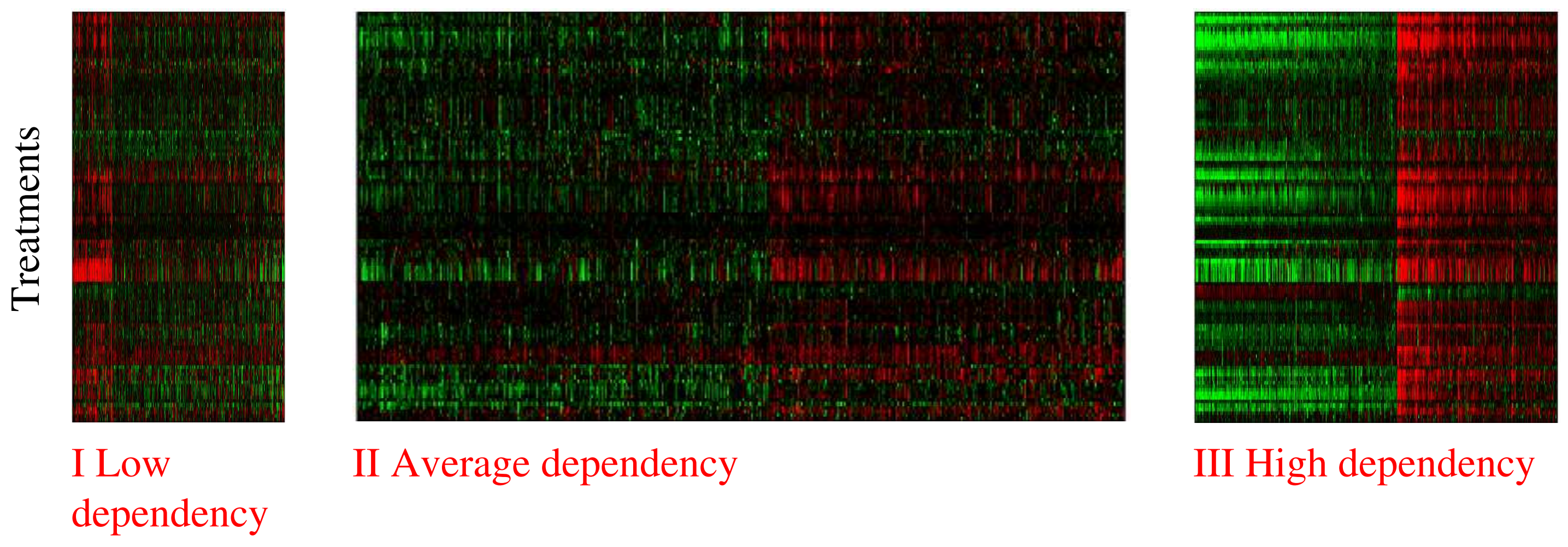
DEPENDENCY RANKING



Known stress genes marked with ticks

650 genes with the highest dependency values, new potential stress genes marked with longer ticks

EXPRESSIONS GROUPED BY DEPENDENCIES



RESULTS

1. Most (784 of 868) environmental stress genes (ESR) found by Gasch et al. (2000) are among the genes with the highest dependency values
2. Some of the missing ESR genes recognized as potential measurement errors: YMR234W and YGL189C are up-regulated in Causton experiments and down-regulated in Gasch experiments
3. New potential ESR genes in group III, 97 most promising ones (within 650 highest dependency values) were studied more closely using the SGD (http://www.yeastgenome.org/):

- 14 associated to stress defense
- 13 chromatin structure and transcription
- 17 ribosomal proteins (coordinately regulated in cell growth)
- 8 cell division
- 5 membrane component metabolism
- 25 unknown function
- 15 other functional classification

CONCLUSIONS

A new, simple data-driven method for extracting yeast stress genes based on dependencies between data sets was introduced. The method found previously detected stress genes with good accuracy, and identified a set of new genes that can have a role in stress response. About half of the 97 new stress genes studied more closely have already been attributed other primary functions than stress response, but our analysis suggested that these genes could also be involved in environmental stress response. The same new annotation was suggested to 25 genes of an unknown function.

REFERENCES

- Causton,H.C, Ren,B., Koh,S.S., Harbison,C.T., Kanin,E., Jennings,E.G., Lee,T.I, True,H.L., Lander,E.S., and Young,R.A. (2001) Remodeling of yeast genome expression in response to environmental changes, *Mol. Biol. Cell*, ***12***, 323-337
- Gasch,A.P., Spellman,P.T., Kao,C.M., Carmel-Harel,O., Eisen,M.B., Storz,G., Botstein,D., and Brown,P.O. (2000) Genomic expression programs in the response of yeast cells to environmental changes, *Mol. Biol. Cell*, ***11***, 4241-4257