The EuroBiotech Journal
RESEARCH ARTICLE
Bioinformatics / System Biology

Genome-wide BigData analytics: Case of yeast stress signature detection

Zelimir Kurtanjek

Abstract

It has been generally recognized that BigData analytics presently have most significant impact on computer inference in life sciences, such as genome wide association studies (GWAS) in basic research and personalized medicine, and its importance will further increase in near future. In this work non-parametric separation of responsive yeast genes from experimental data obtained in chemostat cultivation under dilution rate and nutrient limitations with basic biogenic elements (C,N,S,P), and the specific leucine and uracil auxothropic limitations. Elastic net models are applied for the detection of the key responsive genes for each of the specific limitations. Bootstrap and perturbation methods are used to determine the most important responsive genes and corresponding quantiles applied to the complete data set for all of the nutritional and growth rate limitations. The model predicts that response of gene YOR348C, involved in proline metabolism, as the key signature of stress. Based on literature data, the obtained result are confirmed experimentally by the biochemistry of plants under physical and chemical stress, also by functional genomics of bakers yeast, and also its important function in human tumorogenesis is observed.

Department of Food Technology and Biotechnology, University of Zagreb, Croatia

Corresponding author: Z. Kurtanjek
E-mail: zelimir.kurtanjek@gmail.com

Published online: 27 October 2017
doi:10.24190/ISSN2564-615X/2017/04.02

Introduction

Great advances in the new generation of sequencing instrumentation (NSG) has led to a rate of data generation in life sciences that exceeds the one for solid state circuit integration known as Moore's law (1). Large data sets with very diverse scales of measurement (single cell, microbiome, individual human, populations), and features (genome level, transcriptome, proteome, metabolome, and phenotypes), followed by their integration into BigData sets led to new challenges for automation of data management and statistical inference in the computer age (2,3). BigData sets are stored, managed and distributed on computer clouds and processed by cloud parallel computing software such as developed by the Hadoop and R projects (4). Commonly, BigData are characterized by 4V properties: volume, velocity, variety, and veracity (6). Especially, omics have become BigData science, which is referred as the 4th paradigm of science fundamentals, similar to 4-th generation of the industrial revolution. Integrated large data sets enable development of computer models of complex systems and data based discoveries in contrast to classical hypothesis driven and biased research goals. Unbiased hypothesis-free concepts of data analysis and the application of parallel multiple and different algorithms, also preferably multi-institutional, can yield unbiased and unexpected novel scientific discoveries, which should create a statistically testable hypothesis as an end result and yield significant advancement in theoretical knowledge (7). Methodologies applied are developed within the frame of computer sciences, the dominant techniques being Artificial Intelligence (AI) and Machine Learning (ML). However, well proven methodologies are still not available. A variety of these techniques, including Elastic Nets, Artificial Neural Networks

(ANNs), Deep Learning Networks, Bayesian Networks (BNs), Support Vector Machines (SVMs), Markov models, Decision Trees and Forests (DTs) have been widely applied in life sciences research for the development of classification, pattern recognition, predictive models, and decision making (2,3). Outcomes of the application of ML models are always associated with "grey" logic and are subject to errors, multiple answers, and sometimes with conflicting conclusions. This makes thorough validation with independent data, and with different experimental designs and instrumental techniques, as the crucial focus in evaluation of the scientific contribution of automated, computer algorithmic inference and knowledge discovery. Open data access policy in the era of BigData life sciences is considered as the paramount for critical validation of computer inferences and new knowledge discovery.

There are numerous studies with application of statistical and ML models for yeast genome-wide responses to various stresses. Yeast BigData from experimental studies of phenotypic growth, microarray experiments, physical and chemical stress tests, genome sequencing, mRNA abundance measurements, protein and metabolome profiles are available in open data literature. For example, yeast functional genomics with ML predictive models are applied for classification of unrecognized Open Reading Frame (ORF) functions (8). The application of Bayesian data integration is used for clustering of evolutionary conservation in yeast (9). Especially systemic and extended comprehensive data from the study of genome-wide transcription of yeast as model organism under various nutritional and growth limitation stress are available (12). Statistical models are developed for dynamic short and long term adaptation. The aim is to elucidate common and global regulatory phenomena that allows yeast to adapt its transcriptomic response (10). However, mechanisms that control system level cell responses are still a challenge and require an integrative and multiscale BigData approach. Machine learning Elastic Net regularization is applied to simplify the complexity of the model by removing the least important features kinases/phosphatases (K/Ps) by integrating with metabolomics measurements and in silico estimated metabolic fluxes (11).

Experiment and Data Set

The objective of the experimental study (12) and this data analysis is to use BigData analytics methodologies for systematic identification and quantitative relations (models) of the most responsive genes in exponentially growing yeast chemostat populations limited by dilution rate and basic biotic (C,S,N,P) and auxothropic (leucin, uracil) nutrients. Especially is sought to determine global stress response (gene expression) independent of the specific environmental limiting factors.

Analyzed data are the result of the experimental work of M. Brauer *et al.* (12) a study of genome-wide responses of yeast at steady state conditions in a chemostat with reactor volume 0.5 L during 36 continuous culture conditions at six different limitations and six dilution rates, i.e. at average exponential biomass specific growth rates μ (1/h) =(0.05, 0.1, 0.15, 0.2, 0.25, 0.3). The cells were cultured on minimal defined growth medium and limited by the four basic biotic nutrients (C, N, P, S) as $C_6H_{12}O_6$, $(NH_4)_2SO_4$, $KH_2(PO_4)_2$, $MgSO_4{\cdot}7H_2O$ and two auxothropic components (L, U) leucin and uracil. For basic limitations (G, N, P, and S), was used strain DBY10085 (relevant genotype Mata MAL2-8C), which is the prototrophic CEN.PK derived strain described by van Dijken et al. (13) . For limitations with uracil (U) or leucine (L), was used nonreverting mutant versions of the same strain,i.e., DBY9492 (ura3-52) or DBY9497 (leu2-3leu2-11), respectively (12).

The genome-wide responses of 5337 genes measured as mRNA abundance data were recorded. Details on the experimental procedure and the experimental data are available as a text file (csv) from (12). After upload of the data set missing data were imputed by linear interpolation and stored as a working data set in an excel file for further analysis. The excel data set was imported to *R* software (5) and saved for analysis by various BigData analytical methods.

Methods

Regularized algorithms for Elastic Nets (EN) and Boosted Random Forest Decision Trees (BRFDT) were applied for extraction of the most probable important features, i.e. gene expressions under specific experimental conditions (2,3). The ill-condition properties of BigData models are regularized by appended objective function *F* of statistical criteria of minimization of sum of square errors (SSE), between model predictions and experimental data (*x*), with additional "cost" function *h*(*x*):

$$\min F(x;\beta,\lambda) = SSE(x;\beta) + \lambda \cdot h(x;\beta) \qquad /1/$$

For EN model the important features are associated with the indices of significant parameters β_i for which significantly holds $\beta_i^2 > 0$. The crucial parameter for determination of model dimension λ is selected by cross-validation curve with the untrained data set.

$$h(x;\beta) = \sum_{i=1}^{i=N} \left(|\beta_i| + \beta_i^2 \right) \qquad /2/$$

Interactions, synergism and antagonism of genes, is modeled by the nonparametric decision tree model (BRFDT) for which the cost is proportional to tree complexity defined by number of tree leafs NL and the leaf scores ω_i (15-18):

$$h(BRDTF,\omega) = \gamma \cdot N_L + \frac{1}{2} \cdot \lambda \cdot \sum_{i,k}^{N_L} \omega_i^2 \qquad /3/$$

The integral data set was restructured into separate 36 sub data sets corresponding to individual experimental conditions

and for their later comparative analysis. First each sub set was subject to nonparametric data interval separation according to J.W. Tukey (19). The separation of the subsets of upregulated genes $S^{+}_{\mu,n}$ at specific growth rate μ under nutrient limitation n is defined by:

$$S^{+}_{\mu,n} = \left\{ x_{\mu,n} \middle| x_{\mu,n} \geq Q^{+}_{\mu,n} \right\} \qquad /4/$$

where $Q^{+}_{\mu,n}$ are the upper quantiles. The subsets of downregulated genes corresponding to the same specific growth rate and nutritional limitation are defined analogously as:

$$S^{-}_{\mu,n} = \left\{ x_{\mu,n} \middle| x_{\mu,n} \geq Q^{-}_{\mu,n} \right\} \qquad /5/$$

By combination of the upregulated and downregulated sets of genes new sets of upregulated and downregulated genes are constructed corresponding to the family of experiments at constant specific growth rate and various limitations and sets of genes at variable growth rate and constant nutrient limitation. Venn set operations of intersection and union are then applied. Hence, for a specific n = (C, N, P, S, L, U) limiting nutrient the set of upregulated genes in the growth range from 0.05 to 0.3 h^{-1} is determined by:

$$G^{+}_{n} = S^{+}_{0.3,n} \bigcap \left(S^{+}_{0.25,n} \left(\bigcap S^{+}_{0.2,n} \bigcap \left(S^{+}_{0.15,n} \bigcap \left(S^{n}_{0.1,n} \bigcap S^{+}_{0.05,n} \right) \right) \right) \right) \qquad /6/$$

Similarly, the set for down regulated genes is given by:

$$G^{-}_{n} = S^{-}_{0.3,n} \bigcap \left(S^{-}_{0.25,n} \left(\bigcap S^{-}_{0.2,n} \bigcap \left(S^{-}_{0.15,n} \bigcap \left(S^{-}_{0.1,n} \bigcap S^{-}_{0.05,n} \right) \right) \right) \right) \qquad /7/$$

The summary of effects G_n, upregulated and downregulated genes, for limitation with nutrient n are obtained by:

$$G_n = G^{+}_{n} \bigcup G^{-}_{n} \qquad /8/$$

To determine a global stress signature under all the nutrient and growth limitations defined is the total set G_T of responsive genes:

$$G_T = G_C \bigcup \left(G_N \bigcup \left(G_P \bigcup \left(G_S \left(\bigcup G_{leucine} \bigcup G_{uricil} \right) \right) \right) \right) \qquad /9/$$

Results

Graphical Venn diagrams of the responsive genes to the limitations are presented on **Fig. 1-2** (19,20). Separated are the upregulated and downregulated genes illustrating relative balance between them. It can be observed that there is relatively little overlapping of the sets among genes responsive to the specific nutrient limitations. There is no overlapping of the sets of responsive genes due to limitations between the basic biogenic elements (basic nutrients) and leucine and uracil (not presented). It also shows that the number of responsive genes for the basic biogenic elements (basic nutrients) limitations considerably exceeds the ones with leucine and uracil. The greatest effect of basic the biogenic elements (natural nutrients) limitation on number of responsive genes occurs under phosphorus and sulphur limitations, while the least effect is observed for glucose. The cumulative effects (**Fig. 2**) clearly show that the yeast cell metabolism must adapt to phosphor limitation by greatest rearrangements of metabolic fluxes inferred from the maximum number of the upregulated and downregulated genes.

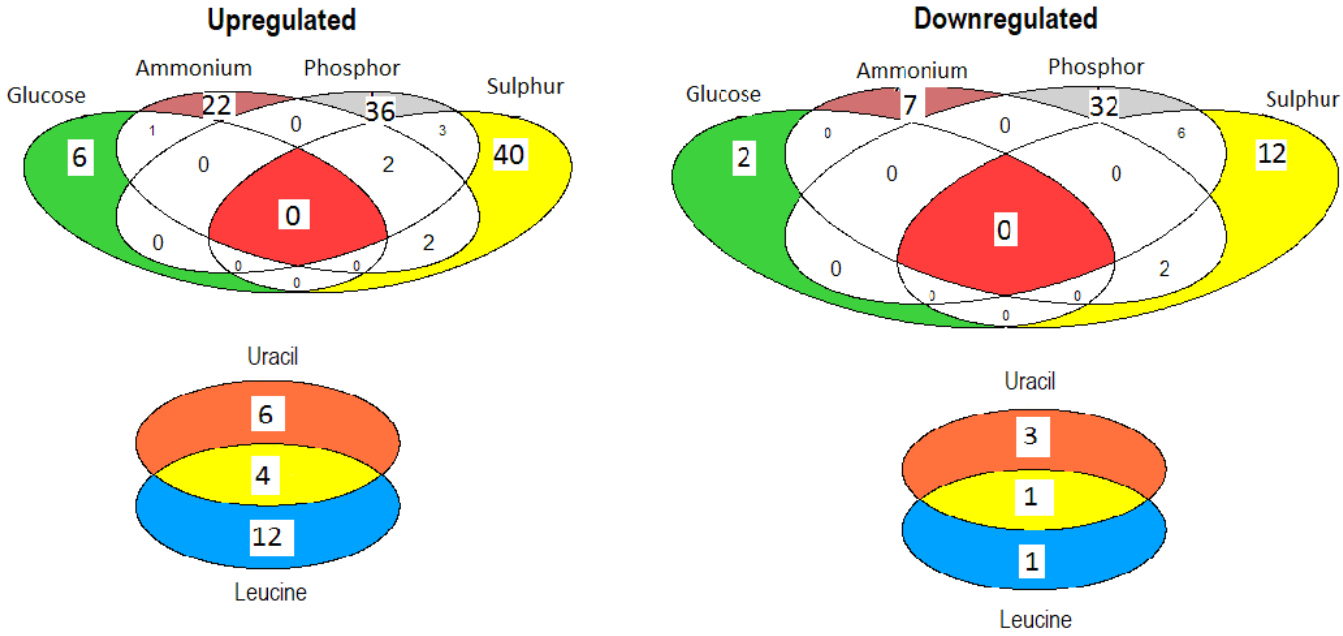

Figure 1. Venn diagrams of sets of responsive upregulated and downregulated genes for basic biogenic and specific limitations (C,N,S,P, uracil, leucine) in the range of biomass specific growth rate $\mu(h^{-1}) \in [0.05 - 0.3]$.

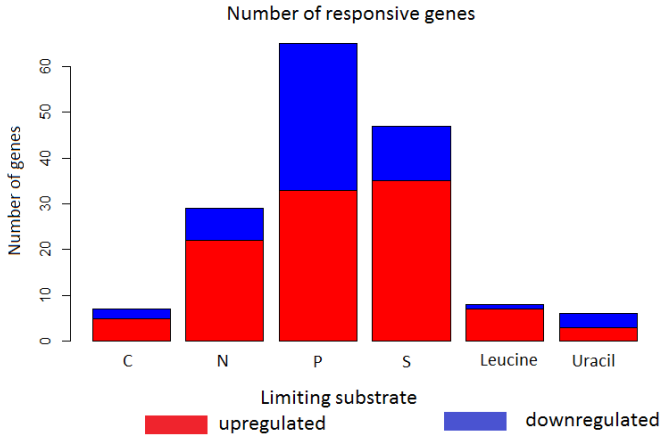

Figure 2. Cumulative effects of the nutritional limitations on the responsive upregulated and downregulated genes.

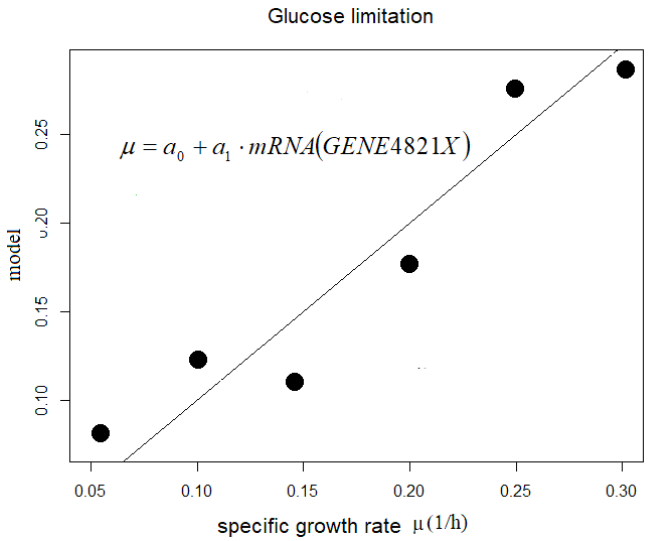

Figure 3. Prediction of the specific growth rate under glucose limitation based on mRNA abundance of the most responsive gene (GENE4821X).

When genome-wide expressions are measured, as mRNA abundance, it is observed that most of the data, about 30 %, are linearly related to the specific biomass growth, either with positive of negative correlation under all limitations (12). They are the "master regulators" that are responsible for maintaining homeostasis under various limitations at steady state conditions (chemostat). Based on genome-wide correlation analysis with the growth rate under all of the limitations, in **Table 1**. listed are the first three genes with the highest correlation, positive and negative. The annotated biological functions are given by M.J. Brauer *at al.* (12) and Funspec Yeast Data Base (14). The table also includes the genes GENE4000X (SNF1) and GENE3004L. SNF1 gene is considered as the master gene which regulates catabolism and anabolism, and regulates the cellular growth and development in coordination with other signaling pathways (21). Also included is the correlation with GENE3004X which is associated with neutral amino acid transmembrane transporter activity, proline catabolism and L-proline permease (12, 14).

The key most responsive genes under each individual specific nutrient limitations are presented in **Table 2**. Applied are linear models:

Table 1. The most linearly correlated mRNA abundances and YDR477W ("master gene") and YOR348C („regulator of transmembrane transport") with the biomass specific growth rate, positive and negative, based on the complete data set of the nutrient limitations

R (mRNA/μ)	Gene ID name	Biological and molecular function
0.875	1739X YDR489W	SLD5 DNA-dependent DNA replication DNA binding DNA replication preinitiation complex
0.875	5385X YBR085W	AAC3 anaerobic respiration ATP, ADP antiporter activity heme transport
0.875	2219X YDL064W	UBC9 G2/M transition of mitotic cell cycle ubiquitin-like conjugating enzyme activity, mitotic spindle elongation
-0.917	4175X YGL156W	AMS1 carbohydrate metabolism, alphamannosidase activity
-0.902	5448X YDR043C	NRG1 regulation of transcription from, glucose metabolic process RNA polymerase, promoter DNA binding
-0.890	4726X YKL151C	Biological process unknown, molecular function unknown
-0.400	4000X* YDR477W	SNF1 protein amino acid phosphorylation, regulation of carbohydrate metabolic process, AMP-activated protein kinase activity
-0.262	GENE3004X** YOR348C	neutral amino acid transmembrane transporter activity, proline catabolism, L-proline permease activity

Table 2. Most important responsive gene determined by Elastic Nets (EN) for each of the specific nutrient limitations

R (mRNA/μ)	Limitation	Gene	Biological and molecular function
+0.893	glucose	4821X YOR374W	ALD4 ethanol metabolism, aldehyde dehydrogenase (NAD) activity
+0.983	ammonium	4616X YPR108W	RPN7 ubiquitin-dependent protein catabolism, proteasome regulatory, structural molecule activity
+1.000	phosphorus	3578X YPR169W	protein monoubiquitination, molecular function, ribosomal large subunit biogenesis
-0.989	sulphur	4197X YDR059C	UBC5 endocytosis, ubiquitin conjugating enzyme activity, acid-amino acid ligase activity, post-translational protein modification
-0.968	leucine	1532X YGL141W	HUL5 ER-associated protein catabolic process, protein monoubiquitination, ubiquitin-protein ligase activity
+0.926	uracil	1460X YOR154W	protein folding in endoplasmic reticulum, molecular function unknown

$$\mu = \beta_0 + \sum_{i=1}^{N_S} \beta_i \cdot mRNA(gene_i) \qquad /10/$$

where N_S are number of most responsive genes, values given in **Fig. 2**, for each of the limitations determined by regularization method given in /8/. The model complexity parameters λ are evaluated separately for each individual limitation by cross-validation. About four to five genes are selected as the optimal dimensions for each of the limitations. In **Table 2** are given the genes with the highest correlation with the biomass specific growth rate. Positive correlation coefficients R correspond to up-regulation and negative to down-regulation under cell adaptation pressure. Biological and molecular functions are given according to M. Brauer (12) and updated by on-line gene data bank Funspec (14).

For determination of a gene as a global stress signature, i.e. under all basic biogenic elements (basic nutrients) , leucine and uracil limitations, among the set of all responsive genes, applied is a model of random forest of decision trees (14-17):

$$\mu = \frac{1}{N} \cdot \sum_{i=1}^{N} DT_i(gene_1 \cdots gene_{N_S}) \qquad /11/$$

where N is the number of individual regression decision trees DT_i and N_s is the number of all responsive genes. Gene importance on the specific growth rate, under all of the nutrient limitations, is inferred from extensive bootstrap perturbation and permutation on "out of box" subsets. The results are depicted in **Fig. 4** shows a prominent role of the YOR348C gene. According to the database (14) its biological functions are: neutral amino acid transport (p-value 0.00015), ammonia

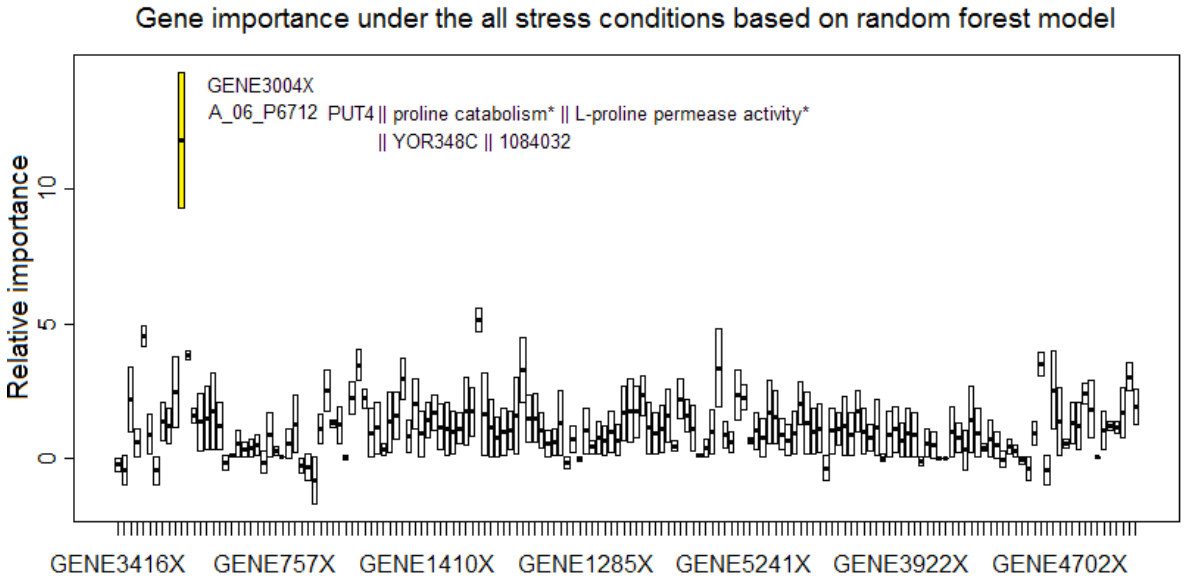

Figure 4. Determination of the most important responsive genes by the random forest model under conditions of all nutrient limitations.

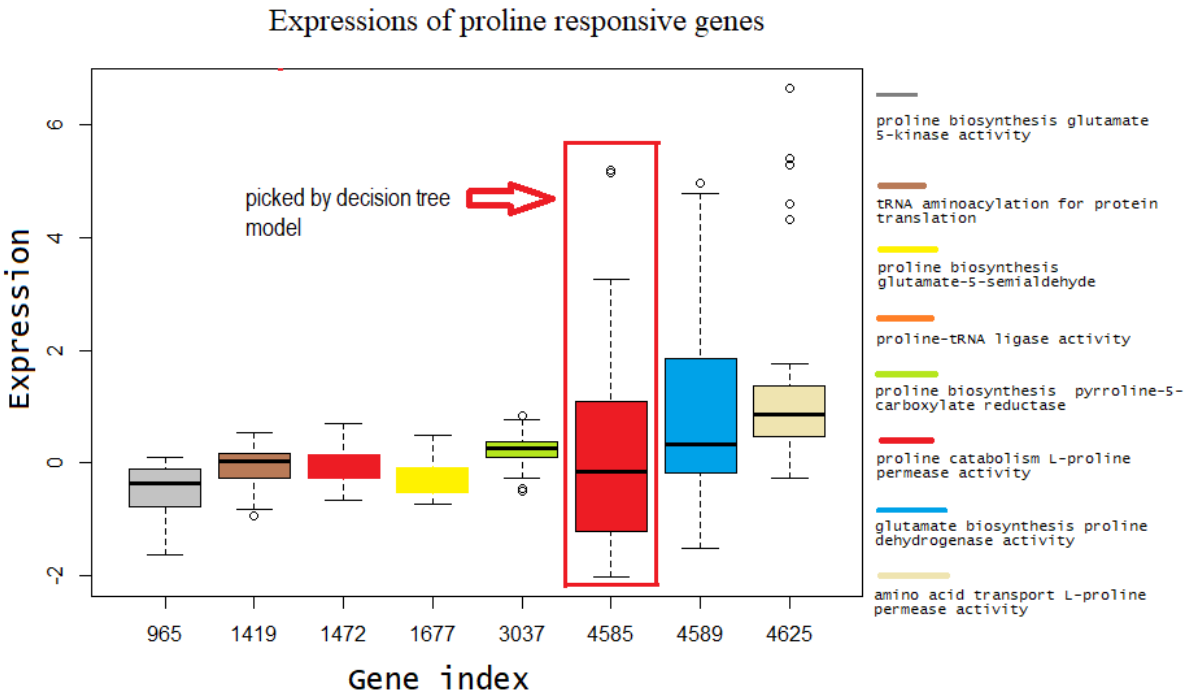

Figure 5. Quantiles of mRNA abundance of the genes related to proline metabolism under all of the nutrition limitations.

assimilation cycle (p-value 0.0007), proline catabolism (p-value 0.003), and proline transport (p-value 0.0003). Its molecular function is in neutral amino acid transporter activity. The simulation data are presented as gene corresponding medians and upper and lower error boundaries in evaluation of relative importance. In **Fig. 5** are presented "box and whiskers" plots, i.e. quantile distributions of the experimental data of mRNA abundances, for the genes directly related to proline metabolism. As presented, the gene selected by the model as the one with highest importance covers the largest range of expressions change from -2 to 5.

Detection of the gene that has the highest response under all six experiments of limitations can be considered as a common cell stress signature (reaction to stresses), which does not imply its key metabolic function.

Discussion

Application of BigData analytics for inference from genome-wide studies enables the power of non-parametric methodology to be used and captures the nonlinear effects of interaction and features of importance by elastic nets and decision trees models. Use of non-parametric quantile analysis provides effective separation of non-responsive (in homeostasis) from highly responsive genes in each individual stress experiment. By applying Venn set algebra it can be shown that most of the responsive genes for individual limitations do not overlap, indicating distinct adaptations to each of the limitations.

Especially, by comparison is shown that intersections between responsive genes for basic biotic and specific auxothropic limitations are empty, i.e. reflects basically different mechanism of adaptation to stress induced by the basic biotic and auxothropic nutrient deprivations. Of course, this imply (which is also observed in the complete data set) that prolin related gene expression under specific auxothropic limitations are not pronounced, although they follow the same general trend, as for the basic biogenic limitations, i.e. negative correlation of prolin metabolism with growth rate.

Although on a genome-wide scale most genes are linearly correlated with the biomass specific growth rates, information on key genes responding to specific limitations is extracted by the regularized linear models and cross validation of elastic nets. Dimensions obtained of feature spaces are relatively low, with three to four genes. Due to the relatively low number of degree of freedoms, only the one dimensional models are extracted indicating the gene signatures for each of the limitations (**Table 2**).

The result for determination of stress signature under all of the nutrient limitations by the random forest decision tree model indicates a prominent role of proline. Proline has been referred to as a stress signal in numerous studies with plants (22-24).

Biochemical analysis of plants under stressful conditions (chemical, physical, temperature) shows that plants accumulate proline. Besides its role as an osmolyte regulator, it has a role as a metal chelator, and a signaling molecule. Proline is also utilized by different organisms to regulate metabolism perturbations caused by various environmental stresses. It is recognized from studies that proline affects signaling pathways through the mechanism of increased production of reactive oxygen species (ROS) in mitochondria (22). Besides plants (23, 24), the same biological function has been observed in worms, apoptosis, tumor suppression, and cell survival in animals. In human tumorogenesis and tumor development it has been found that proline degradative pathway, plays a special role in (25).

Most of the studies on proline role in yeast has been published by Takagi H. *et al* (26-28). They proved that modification of the metabolic pathway by self cloning of trehalose and proline significantly increased tolerance by baking yeast to stress during different phases of technology in bread making. Also, applied were methods of phenomic and functional genomics to identify the key genes, resulting with the identification of V-ATPase as the main controller of the yeast protective mechanism resulting in proline accumulation and improved viability under stress conditions.

Conclusion

Genome-wide studies using the BigData analytical methods of computer age inference has great potential for practical applications but also for discovery of fundamental biological knowledge. The main aspect of the improvement of inference is to reduce ill-conditioned aspect of models with high dimension feature space and usually small dimension of row (sample) space. Use of non-parametric data separation enables us to untangle various factors in large data sets and provide crisp focus on key variables. Interactions of genes under external stresses are discrete phenomena and due to the main features of decision for discontinuous and non-linear models are appropriate to capture the complexity of interactions contained in big data sets.

In this work elastic nets for determination of the main yeast responsive genes for basic biogenic elements (basic nutrients C,S,N,P) and leucine and uracil have been successfully applied.

The random forest decision tree model was applied for the extraction of the main stress signature (responsive gene) from the whole set of limitations. Bootstrap and perturbation of factors of the model provides quantiles and importance levels.

The results clearly indicate that the gene responsible for proline metabolic functions is the dominant signature. This result is confirmed by numerous published literature studies on stress effects on the biochemistry of plants and yeast functional genomics.

Also, proline, as a unique proteogenic secondary amino acid, has its own metabolic system with special features as a stress substrate in the microenvironment of inflammation and tumorigenesis (29).

References

1. O'Driscoll A, Daugelaite J, Sleator RD. "Big data" Hadoop and cloud computing in genomics. J Biomed Inform 2013; 46(5): 774-781.
2. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edition, Springer, New York, 2016.
3. Efron B, Hastie T. Computer Age Inference: Algorithms, Evidence, and Data Science, Cambridge University Press, New York, 2016.
4. Prajapati V, Big Data Analytics With R and Hadoop, Packt Publishing Limited, Birmigham, UK, 2013.
5. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Austria, 2017, URL https://www.R-project.org/.
6. Chen H, Chiang RHL, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Managment Information Systems Q 2012; 36(4): 1165-1188.
7. Robinson SW, Fernandes M, Husi H. Current advances in systems and integrative biology. CSBJ, Computational and Structural Biotechnology Journal 2014; 11(18): 35–46.
8. Clare A, "Machine learning and data mining for yeast functional genomics", PhD Thesis, 2003, University of Wales, Aberystwyth, UK.
9. Huttenhower C., Mutungu K.M., Indik N., Yang W., Schroeder M., Forman J.J., Troyanskaya O.G., Coller H. Detailing regulatory networks through large scale data integration. Bioinformatics 2009; 25(24): 3267-3274.
10. Taymaz-Nikerel H, Cankorur-Cetinkaya A, Kirdar B. Genome-Wide Transcriptional Response of *Saccharomyces cerevisiae* to Stress-Induced Perturbations. Front Bioeng Biotechnol 2016; 4(17)
11. Goncalves E, Nakic ZR, Zampieri M, Wagih O, Ochoa D, Sauer U, Beltrao P, Saez Rodriguez J. Systemic Analysis of Transcriptional and Post-transcriptional Regulation of Metabolism in Yeast. PLOS Computational Biology 2017; 13(1)
12. Brauer MJ, Huttenohwer C, Airoldi M, Rosenstein R, Matese C, Gresham D, Boer VM, Troyanskaya OG, Botstein F. Coordination of Growth Rate, Cell Cycle, Stress Response and Metabolic Activity in Yeast. MBoC, Molecular Biology of the Cell 2008; 19: 352-367.
13. van Dijken JP *et al*. An interlaboratory comparison of physiological and genetic properties of four *Saccharomyces cerevisiae* strains. EMT, Enzyme Microb Technol 2000; *26*(9-10): 706–714.
14. Funspec, Yeast Data Base, http://funspec.med.utoronto.ca/
15. Liaw A, Wiener M, Classification and Regression by random Forest. R News 2002;. 2(3); 18-22.
16. Chen T, Tong H, Benesty M, Khotilovich V. ,Yuan Tang (2017). xgboost: Extreme Gradient Boosting. https://CRAN.Rproject.org/package=xgboost
17. Simon N, Friedman J, Hastie T, Tibshirani R, Journal of Statistical Software, 2011, 39(5), 1-13. URL http://www.jstatsoft.org/v39/i05/.
18. Meinshausen N., Quantile Regression Forests, 2016; https://CRAN.R-project.org/package=quantregForest
19. McGill R, Tukey JW, Larsen WA. Variations of Box Plots, AM STAT. The American Statistician 1978; (32): 12–16.
20. Gregory R. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson R, Moeller S, Schwartz M, Venables B, 2016, URL https://CRAN.R-project.org/package=gplots
21. Zhang J, Vemuri G, Nielsen J, Systems biology of energy homeostasis in yeast, Curr Opin Microbiol 2010; 13(3); 382-388.
22. Hayat S, Hayat Q, Alyemeni MN, Wani AS, Pichtel J, Ahmad A. Role of proline under changing environment, Plant Signal Behav 2012; 7(11); 1456-1466.
23. Liang X, Zhang L, Natarajan SK, Becker DF, Proline mechanism of stress survival. Antioxid Redox Signal 2013; 19(9); 998-1011.
24. Morosan M, Al Hassan M, Naranjo MA, López-Gresa MP, Boscaiu M, Vicente O. Comparative analysis of drought responses in *Phaseolus vulgaris* (common bean) and *P. coccineus* (runner bean) cultivars The EuroBiotech Journal 2017; 1(3); 247-253.
25. Liu W, Phang JM, Proline dehydrogenase (oxidase) in cancer. Biofactors 2012 ; 38(6): 398-406.
26. Shima J, Takagi H, A New Simple Method for Isolating Multistress-Tolerant Semidominant Mutants of Saccharomyces cerevisiae by One-Step Selection under Lethal Hydrogen Peroxide Stress Condition; Biotechnol Appl Biochem 2009; 53; 155-164.
27. Kaino T, Takagi H. Proline as a Stress Protectant in the Yeast *Saccharomyces cerevisiae*, Biosci Biotechnol Biochem 2009: 73(9); 2131-2135.
28. Tsolmonbaatar A, Hashida K, Sugimoto Y, Furukawa S, Takagi H. Isolation of baker's yeast mutants with proline accumulation that showed enhanced tolerance to baking associated stresses, Int J Food Microbiol 2016; 238; 233-240.
29. Phang JM, Pandhare J, Liu Y, The metabolism of proline as microenvironmental stress substrate, J Nutr 2008; 138(10); 2008S-2015S.