Nutrilyzer: A Tool for Deciphering Atomic Stoichiometry of Differentially Expressed Paralogous Proteins

Katrin Lotz[1], Falk Schreiber[2,3], Röbbe Wünschiers[4,*]

[1]SunGene GmbH, Correnssstr. 3, D-06466 Gatersleben, Germany

[2] Leibniz Institute of Plant Genetics and Crop Plant Research (IPK) Gatersleben, Correnssstr. 3, D-06466 Gatersleben, Germany

[3]Martin Luther University Halle-Wittenberg, Institute of Computer Science, Von-Seckendorff-Platz 1, 06120 Halle, Germany

[4]University of Applied Sciences Mittweida, Technikumplatz 17, 09648 Mittweida, Germany

Summary

Organisms try to maintain homeostasis by balanced uptake of nutrients from their environment. From an atomic perspective this means that, for example, carbon:nitrogen:sulfur ratios are kept within given limits. Upon limitation of, for example, sulfur, its acquisition is triggered. For yeast it was shown that transporters and enzymes involved in sulfur uptake are encoded as paralogous genes that express different isoforms. Sulfur deprivation leads to up-regulation of isoforms that are poor in sulfur-containing amino acids, that is, methinone and cysteine. Accordingly, sulfur-rich isoforms are down-regulated.

We developed a web-based software, doped Nutrilyzer, that extracts paralogous protein coding sequences from an annotated genome sequence and evaluates their atomic composition. When fed with gene-expression data for nutrient limited and normal conditions, Nutrilyzer provides a list of genes that are significantly differently expressed and simultaneously contain significantly different amounts of the limited nutrient in their atomic composition. Its intended use is in the field of ecological stoichiometry. Nutrilyzer is available at `http://nutrilyzer.hs-mittweida.de`.

Here we describe the work flow and results with an example from a whole-genome *Arabidopsis thaliana* gene-expression analysis upon oxygen deprivation. 43 paralogs distributed over 37 homology clusters were found to be significantly differently expressed while containing significantly different amounts of oxygen.

1 Introduction

Bioinformatical and experimental investigations of microorganisms have revealed a surprising impact of environmental nutrient availability on genome and proteome composition, a research field now known as stoichiogenomics [5].

The effect of nutrient availability on the atomic composition of expressed proteins has already been described for *Saccharomyes cerevisiae* back in 2001 [3]. Proteins involved in nitrogen

*To whom correspondence should be addressed. Email: roebbe.wuenschiers@hs-mittweida.de

transport and metabolism were found to be significantly enriched in amino acids that contain few nitrogen atoms in their side chain. In contrast, in a genome-wide analysis of the nitrogen-fixing cyanobacterium *Nostoc* PCC 7120 an accumulation of nitrogen-rich amino acids was found in the expressed proteome under nitrogen-fixing conditions, where nitrogen is not limited [8]. Only recently it has been shown that ecological nitrogen limitation shapes the DNA composition of plant genomes [1]. For example, well fertilized crop plants do contain more nitrogen in their transcriptome and proteome than wild plants. Elser *et al.* showed that high expressed plant proteins show a lower nitrogen content per amino acid than do low expressed proteins [6]. These findings exemplify the impact environmental nitrogen availability has on plant genomes.

We are particularly interested in whether the carbon, nitrogen, sulfur or oxygen content of paralogous proteins correlate with their expression level under carbon, nitrogen, sulfur or oxygen limitation, respectively. A prominent subgroup of paralogous proteins are isoenzymes that differ slightly in sequence but catalyze the same chemical reaction. In order to facilitate such analyzes for a broad range of organisms we created a generic software pipeline as detailed below.

2 Theoretical Approach

Generally, the work flow is divided into three steps. 1) The gene products of the whole genome are scanned with respect to their calculated amino acid sequences for finding paralogs and their atomic composition. 2) The gene-expression data measured under nutrient limitation and normal conditions are normalized and preprocessed for comparison. 3) After these two initial steps the retrieved data are evaluated using statistical significance tests. The outcome are two lists of genes that a) contain a significantly different atomic composition concerning the limited nutrient and that b) are significantly different expressed under limitation of the same nutrient. Finally, these two lists are brought together by intersection forming the result list of candidates.

2.1 Step 1: Scouting for Paralogs and Atomic Content Analysis

At first, all paralogous genes in the whole genome have to be identified. By definition, paralogous genes are homologous genes within the same organism. Such paralogous genes share a high sequence similarity and often have similar functions. Finding clusters of paralogs is achieved by sequence alignments using the BLAST algorithm [2].

After identifying the clusters of paralogous genes the atomic composition of their products with regards to the limited nutrient (only carbon, nitrogen, sulfur or oxygen are considered) is calculated by counting the occurrence of particular chemical elements in each amino acid side chain. Because of the differing length of the sequences of all paralogous proteins the absolute values of the nutrient's occurrence have to be normalized to be comparable among each other. Hence every value is divided by sequence length as follows:

$$ac_n = ac/sl$$

where ac denotes the value of the atomic composition for an element, ac_n defines the value after normalization and sl is the sequence length. These normalized values were used for further analysis.

2.2 Step 2: Gene-Expression Data Analysis

The gene-expression data were obtained from DNA-microarray experiments performed under nutrient limitation and control conditions. These data have been preprocessed in order to be comparable among each other. Preprocessing was done by normalization such that all arrays show the same distribution, an equal mean of 0 and an equal standard deviation of 1.

2.3 Step 3: Statistical Evaluation and Formation of the Result

Since the atomic composition of all proteins is not normally distributed, the non-parametric Wilcoxon rank sum test was used to identify significant differences. In a first step all atom counts for the limited nutrient of all gene products are arranged according to size. Subsequently, every value gets a rank and the significance level is calculated according to the designated rank. Values at top and bottom of this ordered list are significantly different from the whole.

Next, the gene-expression data is statistical evaluated. Since the preprocessed data are normal distributed the parametric Student's T-test was used. That way all genes can be denoted that are significantly different expressed under nutrient limitation.

Finally the results of these two significance tests were merged by composing an intersection. The result is a list of paralogus genes that are both significantly different expressed under nutrient limitation and significantly different with respect to the amount of the limited nutrient in their amino acid side chains.

3 Implementation

For Nutrilyzer the model-view-controler (MVC) architecture is used: the model contains all parts for storing the data and the respective logic implemented as Java-Beans. The view deals with data visualization and user interaction and is realized as HTML- and JSP-pages. The controler is responsible for the connection of model and view implemented as Java Servlet. A database (Genomeclust) is also part of Nutrilyzer and contains the sequence and cluster information.

The implementation is web-based. The main programming language is Java 5 using the Spring Web MVC and Spring Web Flow frameworks. The software runs under an Apache Tomcat web server (version 6, [10]) on an openSUSE (version 10.3) operating system. Furthermore, an Oracle database (Express Edition, version 10.20.0.1.0, [14]) is used for storing genome and cluster information. For finding all paralogs from the whole genome the program `blastclust` from the NCBI Toolbox is used. The significance tests are performed using the statistical language R (version 2.7.0, [11]) from the R-project and Bioconductor (version 2.2, [13]), an additional library of R. The software can be requested from the authors.

4 Examplifying Use Case

The chosen microarray experiment analyzes the impact of oxygen deprivation on gene expression in the higher plant *Arabidopsis thaliana* and is deposited in the GEO database [9] under

the identification number *GSE2218* [4]. For stress treatment, 7 day old seedlings were exposed to low light supply. The environment contains a very low oxygen concentration of 0.002%. Due to photosynthesis the environmental conditions can be denoted as hypoxia rather than anoxia. After 12 h of stress treatment RNA from cell extract was isolated and hybridized to DNA-microarrays.

4.1 Using Nutrilyzer

Nutrilyzer is user-friendly and self-explanatory: at the first page all required data and parameters are set. As shown in Figure 1 you can choose an organism, select cluster parameters, input gene-expression data for the same organism and enter the nutrient, which is limited in the gene-expression data. The genomes from *Nostoc* PCC 7120, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* are preprocessed, however, any annotated genome can be uploaded in FASTA format. For gene-expression data only Affymetrix CEL file format is supported.

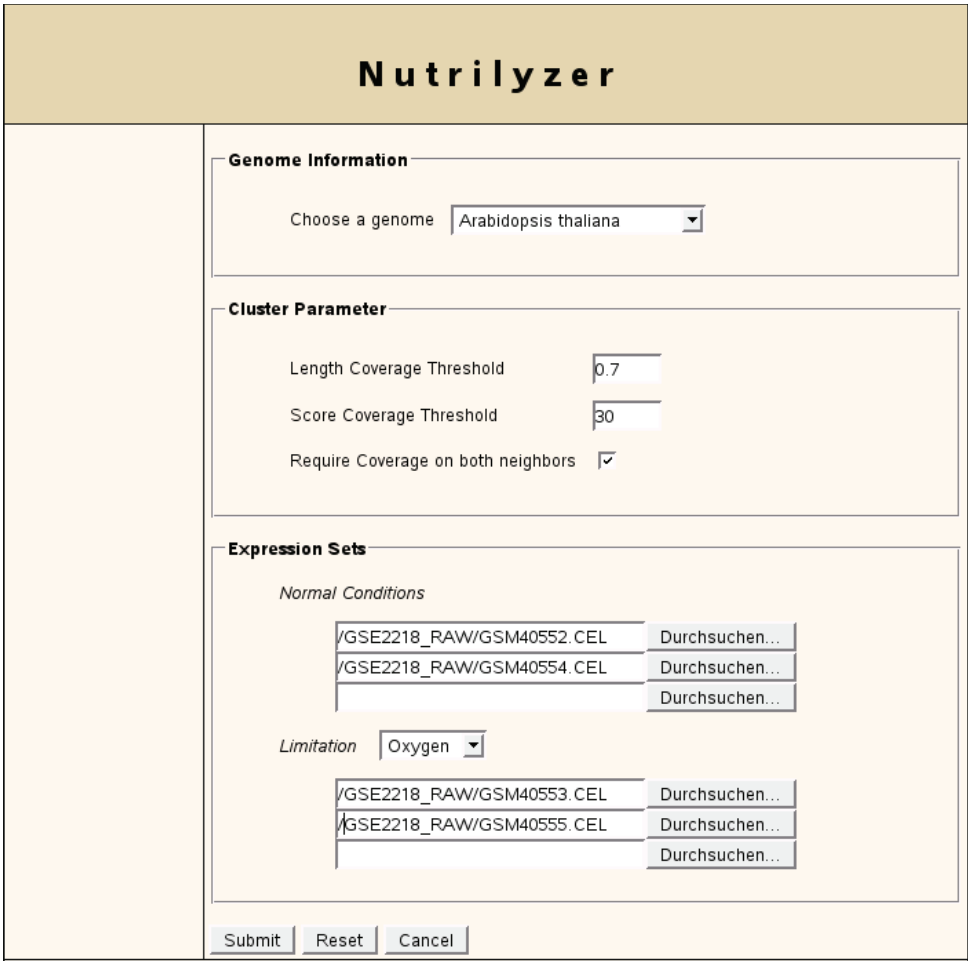
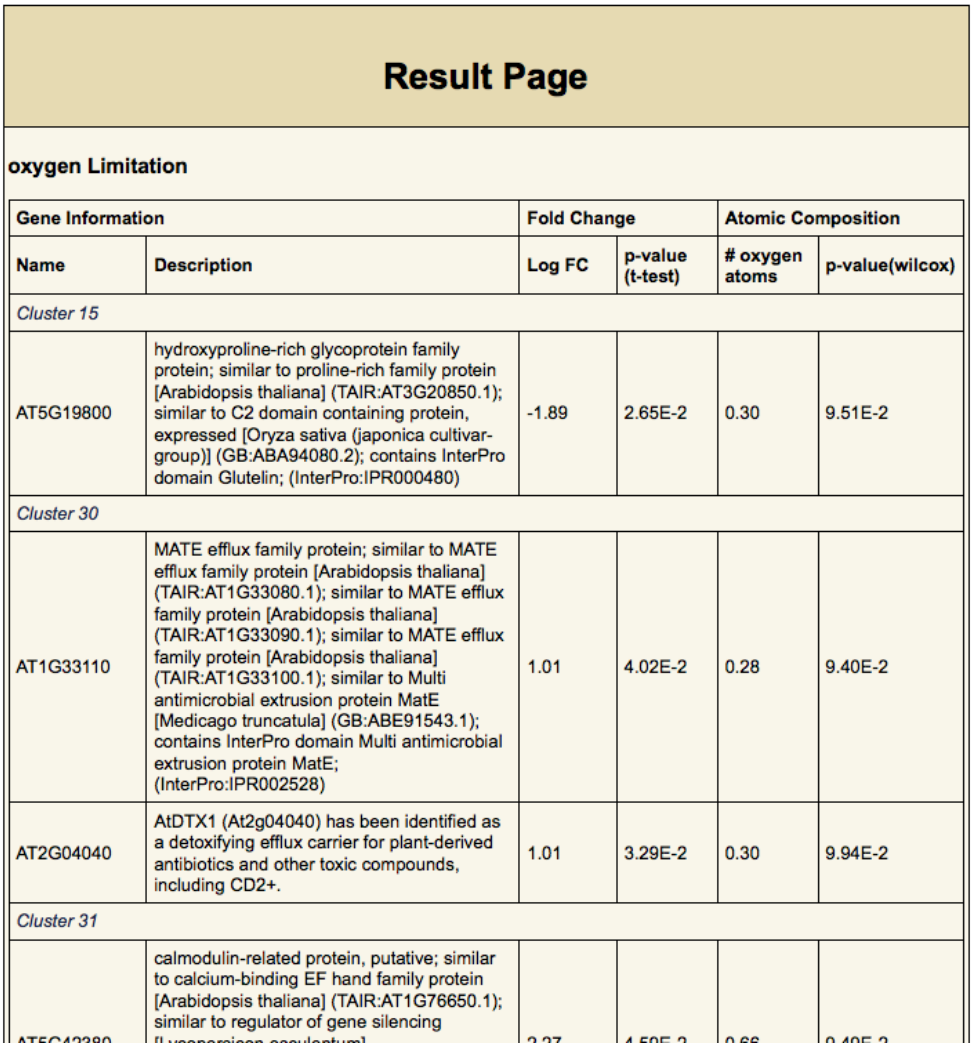


Figure 1: Screenshots of Nutrilyzer: Start page. Graphical user interface to select organism and cluster parameters, to input gene-expression data and choose limited nutrient.

After submission computation of the results begins. The results are shown at the result page, see Figure 2. All listed paralogs are significantly different expressed under nutrient limitation and are also significant in their atomic composition concerning this nutrient. These genes are ordered by their paralogous clusters and the following information is available for each gene:



Result Page

oxygen Limitation

Gene Information		Fold Change		Atomic Composition	
Name	**Description**	**Log FC**	**p-value (t-test)**	**# oxygen atoms**	**p-value(wilcox)**
Cluster 15					
AT5G19800	hydroxyproline-rich glycoprotein family protein; similar to proline-rich family protein [Arabidopsis thaliana] (TAIR:AT3G20850.1); similar to C2 domain containing protein, expressed [Oryza sativa (japonica cultivar-group)] (GB:ABA94080.2); contains InterPro domain Glutelin; (InterPro:IPR000480)	-1.89	2.65E-2	0.30	9.51E-2
Cluster 30					
AT1G33110	MATE efflux family protein; similar to MATE efflux family protein [Arabidopsis thaliana] (TAIR:AT1G33080.1); similar to MATE efflux family protein [Arabidopsis thaliana] (TAIR:AT1G33090.1); similar to MATE efflux family protein [Arabidopsis thaliana] (TAIR:AT1G33100.1); similar to Multi antimicrobial extrusion protein MatE [Medicago truncatula] (GB:ABE91543.1); contains InterPro domain Multi antimicrobial extrusion protein MatE; (InterPro:IPR002528)	1.01	4.02E-2	0.28	9.40E-2
AT2G04040	AtDTX1 (At2g04040) has been identified as a detoxifying efflux carrier for plant-derived antibiotics and other toxic compounds, including CD2+.	1.01	3.29E-2	0.30	9.94E-2
Cluster 31					
	calmodulin-related protein, putative; similar to calcium-binding EF hand family protein [Arabidopsis thaliana] (TAIR:AT1G76650.1); similar to regulator of gene silencing				

Figure 2: Nutrilyzer Result Page (part). Table including all paralogs that are significantly different expressed under nutrient limitation and also contain a significant different amount of this nutrient in their amino acid sequence.

- Description / function
- Logarithmized fold change
- p-value of T-test
- Relative amount of limited nutrient in amino acid sequence
- p-value of Wilcoxon rank sum test

The results can also be downloaded as tab-separated file. Furthermore, information about all genes of a cluster can be retrieved.

4.2 Interpreting Results

Arabidopsis thaliana can be exposed to oxygen deprivation. The decreased oxygen availability limits the production of ATP by mitochondrial respiration. Instead, tissues predominantly

Table 1: Results for *A. thaliana*. Proteins with significantly up-regulated expression (column FoldChange) and significant low oxygen content in the amino acid side chains (column O_2) under oxygen limited growth. Only results from clusters with more than 10 entries are shown. Proteins with a fold change of +1 are expressed more than twice as strong as in the control.

Cluster			*A. thaliana* Protein			
ID	Size	AvgO_2	ID	Description	FoldChange	O_2
030	49	0.35	AT1G33110	multi antimicrobial extrusion (MATE) protein	+1.01	0.28
030	49	0.35	AT2G04040	detoxifying efflux carrier	+1.01	0.30
033	47	0.35	AT1G67300	major facilitator superfamily protein	+1.01	0.30
051	33	0.42	AT5G20230	Al-stress-induced gene	+2.83	0.26
088	23	0.34	AT4G15620	uncharacterised protein	+1.27	0.26
092	22	0.31	AT1G62510	bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin superfamily protein	+2.27	0.26

produce ATP and regenerate NAD by glycolysis and fermentation. The responds of the plant can include an adaptation of energy consumption as well as a complex regulation of the gene-expression on transcriptional and post-transcriptional level. Several ANPs (anaerobic polypeptides) and ASPs (anaerobic stress proteins) were identified under oxygen deprivation. These proteins include enzymes that are involved in sucrose breakdown, glycolysis and fermentation.

Selected significant paralogous genes obtained from the program are listed in Table 1. There are a total of 43 paralogs that are significantly different expressed under oxygen limitation and also contain significantly different amount of oxygen in their amino acid sequences. The specific role of the identified proteins under oxygen deprivation remains to be elucidated and it is not our intention to discuss the biological results in detail, here. The shown example was obtained with only one setting in the parameter space. Parametric walks would be necessary in order to determine result stability.

5 Conclusion

Nutrilyzer provides an easy-to-handle user interface to extract significantly differently expressed and composed paralogous genes and gene products from a given annotated genome and given genome-wide gene-expression data. Its intended use is in the field of ecological stoichiometry and stoichiogenomics. The software is made available at `http://nutrilyzer.hs-mittweida.de`.

Acknowledgements

The authors like to thank Justin Waghray for transferring the original program written by KL to a virtual machine that in turn is administrated by Marcel Scheuche, University of Applied Sciences Mittweida, Germany. RW likes to express his gratitude to the Saxon State Ministry of Science and the Arts, Germany for financial support.

References

[1] C. Acquisti, J. J. Elser and S. Kumar. Ecological nitrogen limitation shapes the DNA composition of plant genomes. *Molecular Biology and Evolution*, 26:953–956, 2009.

[2] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[3] P. Baudouin-Cornu, Y. Surdin-Kerjan, P. Marliere and D. Thomas. Molecular evolution of protein atomic composition. *Science*, 293:297–300, 2001.

[4] C. Branco-Price, R. Kawaguchi, R. B. Ferreira and J. Bailey-Serres. Genome-wide Analysis of Transcript Abundance and Translation in *Arabidopsis* Seedlings Subjected to Oxygen Deprivation. *Annals of Botany*, 96:647–660, 2005.

[5] J. J. Elser, C. Acquisti and S. Kumar. Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition. *Trends in Ecology and Evolution*, 26:38–44, 2011.

[6] J. J. Elser, W. F. Fagan, S. Subramanian and S. Kumar. Signatures of ecological resource availability in the animal and plant proteomes. *Molecular Biology and Evolution*, 23:1946–1951, 2006.

[7] D. Hawkins. Biomeasurement. *Oxford University Press*, 2005.

[8] R. Wünschiers. Nitrogen availability for nitrogen fixing cyanobacteria upon growth on dinitrogen. *African Journal of Biotechnology*, 5:1969–1972, 2006.

[9] NCBI Gene Expression Omnibus Database: `http://www.ncbi.nlm.nih.gov/geo/`

[10] Apache Tomcat: `http://tomcat.apache.org/`

[11] The R Project: `http://www.r-project.org/`

[12] SUN JavaServer Pages Technology: `http://java.sun.com/products/jsp/`

[13] Bioconductor - Open source software for bioinformatics: `http://www.bioconductor.org/`

[14] Oracle: `http://www.oracle.com/`