BIOINFORMATIC APPROACHES FOR THE GENETIC AND PHENOTYPIC CHARACTERIZATION OF A SACCHAROMYCES CEREVISIAE *WINE YEAST COLLECTION*

R. Franco-Duarte[1], L. Umek [2], B. Zupan [2,3], D. Schuller[1]

1 Molecular and Environmental Biology Centre (CBMA), Universidade do Minho, Braga, Portugal
2 Faculty of Computer and Information Science, University of Ljubljana, Slovenia
3 Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, Texas, USA

Introduction

As for many other laboratory model organisms, understanding of the ecological, evolutionary and population genetic features that shaped the biology of *S. cerevisiae* is underscored by a wealth of knowledge on molecular and cellular biology, mainly obtained from a very limited number of reference laboratory strains. In the last few years, yeast researchers became more interested in identifying genomic variability between wild-type yeast strains from different ecological niches or strains that are used for different technological applications. Phenotypic variation among wine yeasts is well-known among strains and was recognized by winemakers before being appreciated by geneticists. In this context we constituted a *S. cerevisiae* wine yeast strain collection (350 strains, obtained from winemaking environments in Portugal) for the conservation of biodiversity, sustainable development of genetic resources and for an equitable sharing of genotypic and phenotypic data.

The objective of the present study was the development of bioinformatic approaches (based on standard and not-so-standard machine learning and data mining algorithms (Orange software) to find links between genetic and phenotypic similarities within a group of 103 *S. cerevisiae* strains that belong to the above mentioned strain collection.

Materials and Methods

Strain collection
The strains used in this study were obtained from winemaking environments in the Vinho Verde Region in Portugall, within our previous studies related to *S. cerevisiae* ecology and biodiversity [1, 2].

Molecular identification
Isolated strains were analysed by mitochondrial DNA restriction patterns (mtDNA RFLP) [3]. Strains with identical mtDNA RFLP patterns were grouped and one representative strain was further characterised by analysis of 10 *S. cerevisiae* specific microsatellite loci [4, 5]. The equivalent discriminatory power of mtDNA RFLP and microsatellite analysis has been previously reported [6].

Strain Selection
Due to the high number of strains, neural networks (Kohonen self-organizing maps) were applied within the JATOON software package [7] to choose a genetically most diverse sub-set of 103 strains based on microsatellite data. These were used for phenotypic screens and data analysis.

Phenotypic tests
Phenotyping included the evaluation of 19 traits used in yeast taxonomy [8] or for wine yeast strain selection. Yeast cells were withdrawn from frozen aliquots (glycerol, 30%, v/v, at -80ºC) and were pre-inoculated in 10 ml of the culture medium to be used (see table) and incubated (24h, 30ºC, 200rpm). Cell density was determined (OD_{640}) and adjusted to 1.0. From this suspension, 15 µl were inoculated in 4 or 8 replicate wells of a 96-well microplate containing 135uL culture medium, so that the final cellular density was 0.1 OD. Growth conditions were used as indicated in the table.

Computational analysis
All the computational analysis was performed using the Orange software [8]. Predictive accuracy was measured for each test performed, using area under receiver operating characteristics score (AUC), which estimates the probability that the predictive model would correctly distinguish between two different locations or groups of strains for a specific phenotype.

	Parameters	Culture medium	Incubation temperature	Incubation time	Agitation
Carbon Sources	Glucose	YNB (0,67% w/v) + carbon source (2% w/v)	30ºC	22h	200rpm
	Ribose				
	Arabinose				
	Saccarose				
	Galactose				
	Raffinose				
	Maltose				
	Glycerol				
	Potassium acetate				
Nitrogen Sources	Peptone	YNB without nitrogen (0,67% w/v) + nitrogen source (0,05% w/v)	30ºC	22h	200rpm
	Ammonium sulfate				
	Imidazole				
	Urea				
Stress conditions	Growth in wines	Vinho Verde Wine	18ºC	3 weeks	none
	Growth in ethanol	MS medium + ethanol (6% v/v)	18ºC	3 weeks	none
	Temperature	MS medium	4ºC, 18ºC, 30ºC, 37ºC, 42ºC	4ºC – 3 weeks Others – 22h	4ºC – none Others – 200rpm

RESULTS

1 *S. cerevisiae* strain collection

Minho	2001	2002	2003	2006	Total
Nº of samples collected	36	18	36	84	174
Nº of spontaneous fermentations	19	12	23	30	84
Nº of isolates	570	360	690	900	2520
Nº of *S. cerevisiae* strains	110	56	137	47	350
Percentage of *S. cerevisiae* among the total fermentative flora	100%	100%	100%	100%	

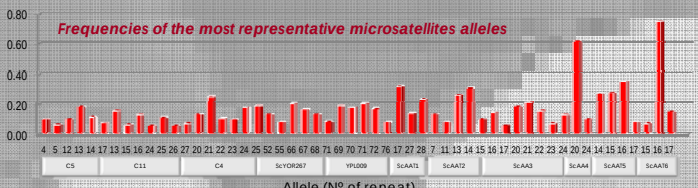
- *174 grape samples were collected during the harvests of four years in three vineyards (A, C and P) of the Vinho Verde Wine Region*
- *2520 S. cerevisiae isolates were obtained from the final stages of fermentation and 350 S. cerevisiae strains were delimitated, based on mitochondrial DNA restriction fragment length polymorphism analysis*

2 Genetic characterization by a set of highly polymorphic microsatellites

- *Each of the 350 strains was characterized by a set of 6 polymorphic microsatellites (ScAAT1 – ScAAT6).*
- *Based on allelic information and using Kohonen self-organizing maps (JATOON software), a more restricted a subset of 103 strains was obtained, that were genetically most diverse.*

Each neuron ( ) represents a group of strains that share genetic similarities

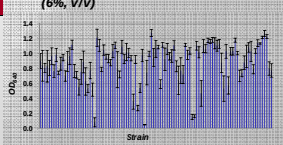
- *The genetic characterization of these strains was expanded to five additional microsatellite loci.*
- *The microsatellite markers revealed a high degree of variability (171 alleles), and 32 alleles had an allelic frequency above 0.1.*



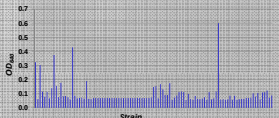
3 Phenotypic characterization

Culture media containing		Growth (final OD_{640}) Range	Average	Taxonomy [9]
Carbon source	Glucose	0,9-1,4	1,2	+
	Ribose	0,1-0,6	0,2	-
	Arabinose	0,1-0,5	0,1	-
	Saccarose	0,4-1,5	1,1	v
	Galactose	0,1-1,5	1,0	v
	Raffinose	0,2-1,2	0,7	v
	Maltose	0,2-1,4	1,0	v
	Glycerol	0,1-0,4	0,2	v
	Potassium acetate	0,1-0,4	0,1	v
Nitrogen source	Peptone	0,6-1,4	1,3	+
	Ammonium Sulfate	0,3-1,4	1,0	v
	Imidazole	0,2-1,2	0,6	v
	Urea	0,3-1,4	1,1	v
Temperature	4 ºC	0,1-0,3	0,1	
	18 ºC	0,2-1,4	1,1	
	30 ºC	0,6-1,4	1,0	+
	37 ºC	0,7-1,5	1,0	v
	42 ºC	0,1-0,3	0,1	
Stress	Ethanol 6% (v/v)	0,1-1,3	0,9	
	Wines	0,1-0,6	0,1	

Growth in the presence of ethanol (6%, v/v)



Growth in finished wines (ethanol content 12%, v/v)



\+ positive
\- negative
v variable, depending on the strain

- *The phenotypic diversity of 103 S. cerevisiae strains was assessed using 22 physiological tests, being 15 used for the conventional identification of yeasts.*
- *The results generally matched with taxonomic data. However, one and six strains were identified that were capable to consume arabinose and ribose, respectively (O.D. > 0.4).*
- *Variation was also apparent for phenotypic traits that are used for wine yeast strain selection such as the finding of strains that show a very low / very high ethanol tolerance.*

4 Computational approaches

Prediction of geographical location

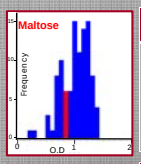
existent \ predicted	A	C	P
A	23	4	6
C	3	14	8
P	11	6	28

AUC = 0.805
Classification accuracy= 0.633

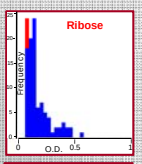
- *Various prediction methods were tested using 10-fold cross validation; best performance was achieved with naive Bayesian classifier;*
- *The table shows a prediction of geographical location (vineyards A, C and P), based on genetic data (microsatellites). A correct assignment of a strain to the respective vineyard was obtained for 68%, 56% and 62% of strains from vineyards A, C and P, respectively;*
- *The area under receiver operating characteristics score (AUC) is high, and classification accuracy is well above the default accuracy of 0.45, which would be achieved by classifying to the majority class (location P).*

Phenotype analysis

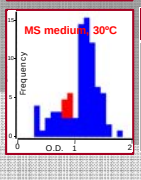
One phenotype at the time was analyzed by the Orange software [8], and subgroups were then analyzed through hierarchical clustering. The success of group-characterization was measured through leave-one-out and using AUC score. Subgroups with AUC scores above 0.75 are shown below.



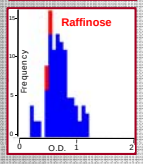
Strains	$O.D._{640nm}$	AUC
31 330 222 315 189 94	0,8	0,845



Strains	$O.D._{640nm}$	AUC
142 175 237 172 270 276	0,1	0,784



Strains	$O.D._{640nm}$	AUC
50 205 260 363 289 301 383	0,9	0,816



Strains	$O.D._{640nm}$	AUC
124 159 343 330 383 333	0,5	0,771

- *Tables show strain subgroups with identical or very similar growth rate that also share similarities regarding microsatellite allelic combinations (marked by red bars in the histograms).*
- *All 6 strains with intermediate growth in maltose (average final O.D. = 0.8) containing culture medium showed microsatellite allelic similarities. These strains are genetically not related, they were obtained in different vineyards.*
- *Subgroups of strains with similar growth characteristics in MS medium (average final D.O. = 0.9), YNB containing ribose (average final DO = 0.1) and raffinose (average final DO = 0.5) showed similarities regarding microsatellite allelic patterns.*

Conclusions

- Strains are genetically variable, demonstrated by the high number (171) of microsatellite alleles among 103 strains. This variation is also apparent for phenotypic traits.
- Bayesian classifier can assign, with high probability, a strain to the vineyard from where it was isolated.
- Hierarchical methods showed that groups of strains that share growth patterns for some culture media (MS medium or YNB containing maltose, ribose or raffinose), can be also grouped based on their microsatellite similarities.
- Our study shows the potential of computational approaches to obtain indications about strain characteristics from microsatellite allelic data.

References

[1] Schuller, D. et al., 2005. FEMS Microbiol Ecol 51, 167-177.
[2] Valero, DE. Et al., 2005. FEMS Yeast Res. 5. 959-969.
[3] López V., et al., 2003. Int J Food Microbiol 68:75-81.
[4] Legras, J. L., et al., 2005. Int. J Food Microbiol. 102, 73-83.
[5] Pérez, M.A., et al., 2001. Lett. Appl. Microbiol. 33, 461-466.
[6] Schuller, D., et al., 2004. FEMS Microbiol Lett. 231(1) 19-26.
[7] Aires-de-Sousa and Aires-de-Sousa, 2003, Bioinformatics. 19(1):30-6.
[8] Demsar J, Zupan B, Leban G (2004) Orange: From Experimental Machine Learning to Interactive Data Mining. White Paper. Faculty of Computer and Information Science, University of Ljubljana
[9] Barnett, J.A., Payne, R.W. and Yarrow, D. 2000. Yeasts: Characteristics and identification. 3rd Ed., Cambridge University Press, Cambridge, UK.

Acknowledgements

This study was financially supported by the programs POCI 2010 (FEDER/FCT, POCTI/AGR/56102/2004) and AGRO (ENOSAFE, Nº 762).
The authors thank João Aires-de-Sousa for support and training with the JATOON Kohonen SOMs.
Magda Silva Graça is gratefully acknowledged for the operation of the DNA sequencer.