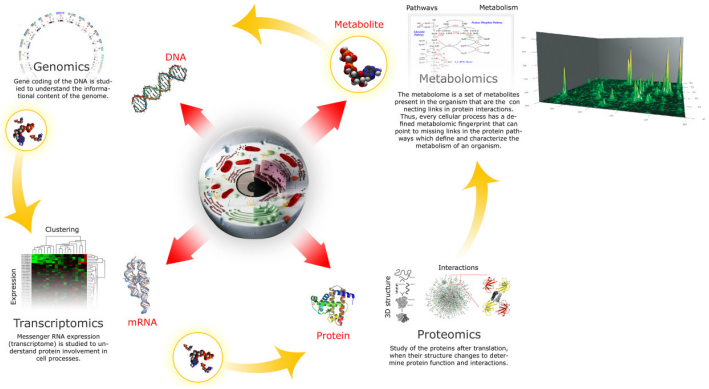
MetaboStat: post ion-identification analysis of LC/GC-MS data

Brodsky Leonid[1,2], Rogachev Ilana[1], Venger Ilya[1], Malitsky Sergey[1] and Aharoni Asaph[1]

[1] Department of Plant Sciences, Weizmann Institute of Science, P.O. Box 26, Rehovot 76100, Israel
[2] Institute of Evolution, University of Haifa, Mount Carmel, Haifa 31905, Israel



Abstract

The peak-picking/peak-alignment programs such as XCMS or MarkerLynx carry out identification of ion-profiles, where each profile is a series of ion intensities across biological samples. After identification of ions and following the behavior of their intensities across samples, the researcher has to interpret the results from several perspectives: quality control (QC), biostatistics, and structural identification of metabolites. Based on the profiles of ion intensities our MetaboStat program helps to answer these questions:

Detection of trustable ions : ion-profiles that are robust under random perturbation of parameters of the peak-picking/peak-alignment procedure;

Detection of "metabolites". Each metabolite is a group of ions that are (i) eluted at the same narrow retention time interval; (ii) have highly correlated profiles; and (iii) the group is enriched with the isotope series.

Estimation of the statistically significant effects of applied biological factors and their interactions for each ion and each metabolite.

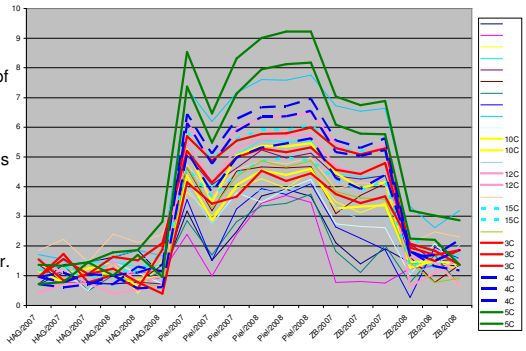
As an input MetaboStat takes the multiple detection of ion-profiles by any peak-picking/peak-alignment program (XCMS by default) under variations of the program's parameters.

The algorithmic core of MetaboStat exploits the following ideas:

- Quantile normalization of ion-intensities inside each group of biological replicates.
- Clustering of ion-profiles in the PCA space. The nested clustering is generated by a recursive identification of the local neighborhoods in the PCA-space that are enriched by ion-profiles. Under different initial limitations this procedure is applied to the identification of the trustable ions and to detection of the metabolites as ion groups.
- The ANOVA-equivalent multiple linear regression analysis of ion intensities against orthogonal contrasts is used for the detection of the per-ion significant biological effects and their interactions.
- The significant effects for metabolites (groups of ions) are identified via Wald statistics based "voting" of individual ion-profiles of the metabolite either "for" or "against" significance of every effect (contrast) of the ANOVA model.

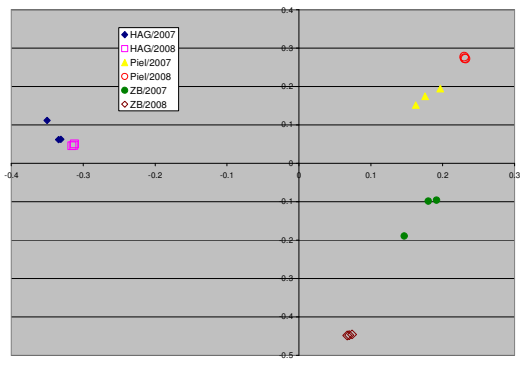
Metabolites as clusters of ions

Every metabolite is detected as a cluster of ions, eluting at the same retention time, and having similar intensity profiles across samples. Here five series of isotopes of five ions (10C, 12C, 15C, 3C, 4C, 5C) are members of the cluster.

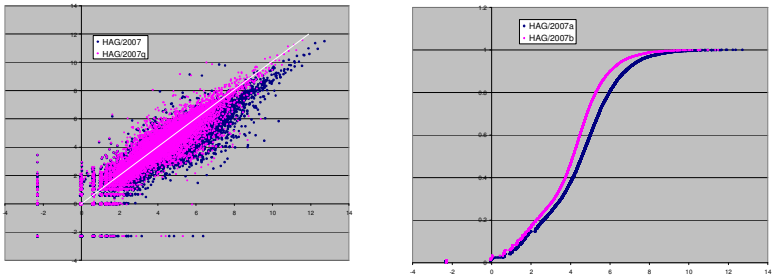


Distribution of samples in the PCA space

MetaboStat performs the non-supervised grouping of samples in the PCA space. Particularly, the program generates the visual pictures for distributions of samples in a variety of two dimensional PCA subspaces. The pictures could be generated under different filtering of ions based on the level of impact of biological factors on the ions.

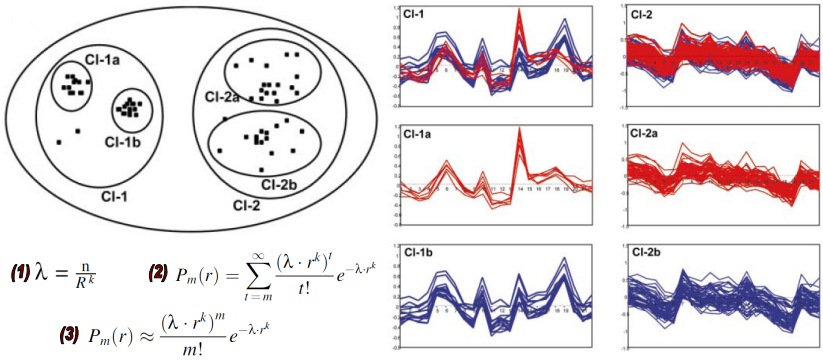


Quantile normalization of replicates



Quantile normalization provides the same cumulative distributions (accumulated quantiles) of the replicate samples. As a result, the ion's intensities of replicates (their trends) match each other in a more correct way. The match of two replicates before (dark blue) and after (pink) quantile normalization is presented.

Poisson distribution based nested clustering of mass-ion intensity profiles in the PCA space



(1) $\lambda = \frac{n}{R^k}$ **(2)** $P_m(r) = \sum_{t=m}^{\infty} \frac{(\lambda \cdot r^k)^t}{t!} e^{-\lambda \cdot r^k}$

(3) $P_m(r) \approx \frac{(\lambda \cdot r^k)^m}{m!} e^{-\lambda \cdot r^k}$

The nested clustering of ion-profiles is generated by the recursive identification of local neighborhoods in the PCA-space that are enriched by ion-profiles. Under different initial limitations this procedure is applied to the detection of trustable ions, detection of the metabolites as ion groups, and clustering of metabolites. The idea of clustering is as follows. Under the hypothesis on uniform distribution of points (profiles) in a volume of the k-dimensional space, and having the average density of points in the initial volume (formula 1), the probability for number of points in any sub-volume will be calculated according to Poisson distribution (formula 2). This probability is well approximated by formula 3.The greedy clustering procedure detects the most enriched by profiles sub-volume (minimum probability according to formula 3) as the first cluster, the next most enriched sub-volume as the second one, and so on. Every sub-volume of the initial clustering (Cl-1 and Cl-2 of the figure above) will be taken as the initial volume, its density (formula 1) will be recalculated, and the nested enriched sub-volumes (like Cl-1a and Cl-1b) will be detected.

Regression based ANOVA for ions and metabolites

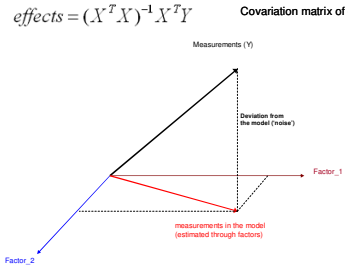
Model of true factor effects

Stage	Year	A	Stage(-)	Stage(+)	Year(-)	Year(+)	Stage(-):Year(-)	Stage(-):Year(+)	Stage(+):Year(-)	Stage(+):Year(+)	
green	2007	1	1		1		1				
green	2008	1	1			1		1			
red	2007	1		1	1				1		
red	2008	1		1		1				1	
			1	1							= 0
					1	1					= 0
							1	1			= 0
							1	1	1		= 0
							1	1	1	1	= 0

Model of 'contrasts'

Stage	Year	A	Stage(-/+)	Year(-/+)	Stage:Year
green	2007	1	1	1	1
green	2008	1	1	-1	-1
red	2007	1	-1	1	-1
red	2008	1	-1	-1	1

$effects = (X^T X)^{-1} X^T Y$ Covariation matrix of effects: $(X^T X)^{-1}$



ANOVA estimations of how significant are the main effects of the biological factors and their interactions for behavior of intensities of individual ions across samples are performed through the multiple linear regression model of contrasts. The significance of effects is evaluated by a *t*-test significance of the contrast parameter estimations. The significant effects for metabolites (groups of ions) are identified via the chi-square Wald statistics based "voting" of individual ion-profiles of the metabolite either "for" or "against" significance of every contrast of the regression-based ANOVA model.

MetaboStat: future developments

- Isotope-based peak detection and RT warping
- Isotope and clustering-based detection of metabolites
- Identification of elemental composition for metabolites
- Annotation of metabolites: structural formulae and metabolic pathways the metabolites participate in
- Supervised discrimination of samples (Discriminant Analysis, SVM, decision tree/forest)
- Integration and networking: association of metabolites, transcripts, and protein interactions