Analyzing Genetic Control Of Traits In Ecophy-genetic Modelling

Omar Belatmane, Florent Bonneu, Daniel Gourion, M. Memmah, Bénédicte Quilot-Turion

TISS1 Tersys Implanteus 1st Summer School
5 et 6 juillet 2022 – Pôle Agrosciences, Avignon Université

Belatmane Omar [1], Bonneu Florent [1], Gourion Daniel [1]
Memah Mohamed [2] ,Quilot-Turion Bénédicte [3]

Analyzing Genetic Control Of Traits In Ecophy-genetic Modelling.

[1] Avignon University – LMA [2] INRAE PSH [3] INRAE GAFL

1- Introduction

Predicting genotype-to-phenotype relationships under contrasting environments is a big challenge for plant biology. Integrating genetic information into a dynamical model is the key to understand the variations among genotypes.

Several models have been proposed, one of which is to simultaneously estimate parameters for each genotypes and the relationships between genomic markers and genotype-dependent parameters. This new "joint analysis" may replace what is known as independent analysis.

Framework :

MODEL: System of ODE kinetic model of sugar metabolism to model the concentrations of different sugars during the development of the peach fruit.
POPULATION : We have 106 Genotypes, and measure sugars [Sucrose, Sorbitol, Fructose, Glucose], at 6 different times.
PARAMETERS : 9 parameters denoted [LHx ,RAI ,KFk ,KHk ,OCp ,RSS,TAS ,TPF ,RSO].

Objectives :

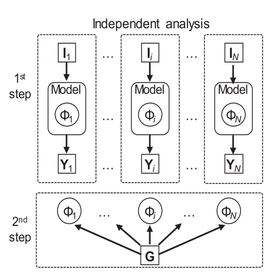
- Estimate the 9 unknown parameters of the mathematical model.
- Adjust and implement the joint study approach of sugar metabolism in peaches.

In genetic analysis, **association mapping**, is a method of mapping quantitative trait loci (QTLs) that takes advantage of linkage imbalance to relate phenotypes to genotypes, uncovering genetic associations.

Two approaches can be considered:

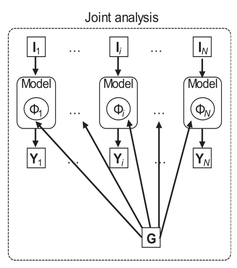
Independent analysis

- Step 1: Estimation of **model parameters** per genotype.
- Stage 2: Use of these estimation parameters for **association mapping**.



Joint analysis

- A single step: **Parameter estimation for all genotypes** and **genetic analyses on the parameters** are carried out simultaneously using a hierarchical statistical model.



Figures was taken from : Akio Onogi (2020)

2- Independent Analysis

2-1 Continuity of parameters

The purpose of this study is to explore the continuity the ODE model solutions relatively to the 9 unknown parameters:

To illustrate our exploration of continuity we choose randomly one genotype [E14] and one parameter [RAI] and represent below the evolution of solutions for different parameter value [0.001, 0.012, … , 0.1].

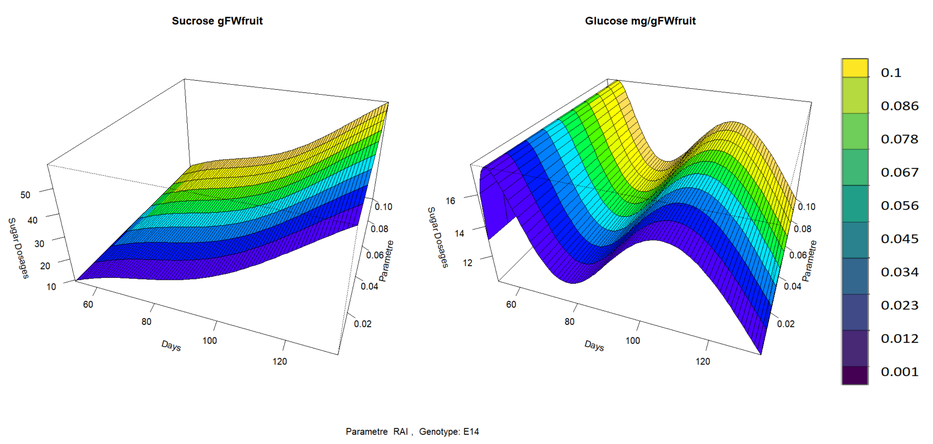


Figure 1: Resolutions of System EDO with various RAI parameter for genotype E14.

❑ These results give hope that the solution of the ODE is regular enough depending on the parameters to allow an efficient optimization process.

2-2 Parameter optimization

During this work we compare 3 types of initial points :

- **Initial_1** = Represent a set of parameters obtained in a previous work [by Mohamed Memah and Quilot-Turion Bénédicte]: same vector for all Genotypes.
- **Initial_2** = **[0.45,0.03,34.67,236.04,943.88,449.44,184.01,25.71,2.21]** Population-adjusted parameters, obtained in a previous work [Doctoral thesis by Hussein KANSO]: same vector for all Genotypes.
- **Initial_3** = **((Upper bound+ Lower bound)/2)** represents the center point of each parameter interval value : same vector for all Genotypes.

OBJECTIVE FUNCTION : Sum of RMSE normalized by the Standard Deviation of all population per sugar, i.e. $f(x) = \sum_{i=1}^{4} \sqrt{\frac{1}{6}\sum_{j=1}^{6} \frac{(y_{i,j}-\widetilde{y_{i,j}})^2}{\sigma_i^2}}$, where i represents sugar, j represents the date of measure, σ_i^2 the standard deviation of sugar i, $y_{i,j}$ and $\widetilde{y_{i,j}}$ are respectively the observed and predicted values of our ODE model for sugar i at day j.

A variety of optimization methods including BFGS, interior points, simulated annealing have been applied to find optimized model parameters. In the following results, we choose the SANN method corresponding to the simulated annealing method.

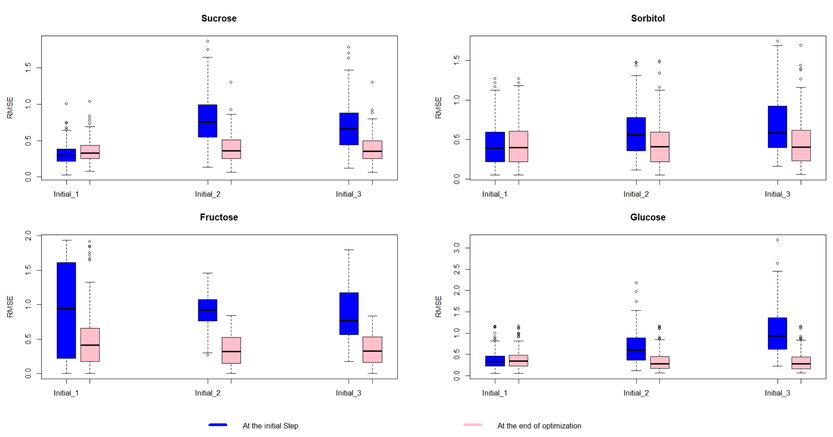


Figure 2: Boxplot of RMSE (Root-Mean-Square-Error) per sugar depending on the initial conditions used for the parameters in the SANN method.

❑ The optimization process drastically reduces the error for the three initial points. For some genotypes, these initial points provide very different local optima with approximately the same value of objective function.

2-3 QTL detection :

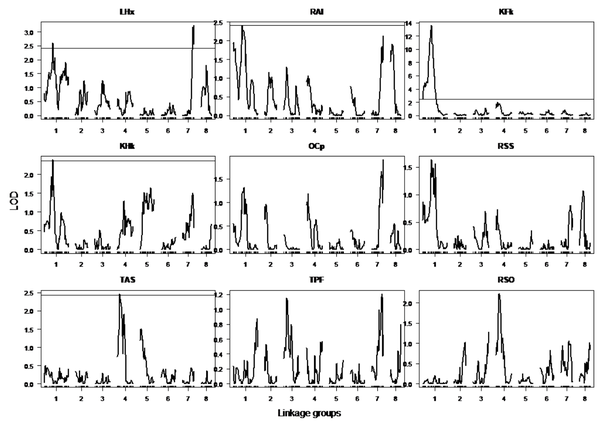


Figure 3: LOD profiles and QTL detected from 340 SNPs taken from 8 chromosomes for the 9 parameters of the model estimated using the Initial_3 set of values in SANN.
The horizontal lines represent the LOD thresholds used to detect QTLs.

3- Joint Analysis

3-1- Mathematical Model

- $Y_i = m(I_i, \Phi_i) + \varepsilon_i$, $\quad \varepsilon_i \sim Normal[0, \Sigma_\varepsilon^2]$

i : Index of Genotypes i.e. i = 1, … , 106
m : ODE model of sugar metabolism
Φ : Model parameters
I : Inputs
ε : Error term (assumed to be normally distributed)
Yi : Vector combine the 6 mesures per sugar

$$\Sigma_\varepsilon^2 = \begin{bmatrix} \sigma_{\varepsilon,1}^2 & 0 \cdots 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 \cdots 0 & \sigma_{\varepsilon,D}^2 \end{bmatrix}$$, D =24 is the dim(Y_i)

- $\varphi_{i,j} \sim Normal\left[f_j(G_i, \Theta_j, \Psi_j), {}^1/_{\tau_{0,j}^2}\right]$ Prior distribution of j*th* parameter for i*th* Genotype

j : Position in the vector of model parameter
$\varphi_{i,j}$: jth parameter of Φ_i.
f_j : Function linking the marker genotypes to the jth parametre.
G_i : Marker genotype, in our case we use SNP which is a short DNA sequence.
Θ_j, Ψ_j : Parameters and Hyperparameters of f_j respectively.
τ_j^2 : Residual precision.

- In Whole Genome Regression f_j is written : $f_j(G_i, \Theta_j, \Psi_j) = 1\mu_j + G_i^T B_j$

μ_j : Intercept.
B_j : Regression coefficients.

3-2- Algorithm & Method :

WGR Regression BayesC Model :

$$\mu_j \sim \text{cte}, \quad B_j \sim \begin{cases} Normal(0, \sigma^2), & \rho_j = 1 \\ 0, & \rho_j = 0 \end{cases}, \quad \rho_j \sim \text{Bernoulli}(\kappa), \quad \sigma^2 \sim Scaled\ inverse\ chi\ square(\vartheta, S^2)$$

Regression Parameters: $\{\mu_j, B_j, \rho_j, \tau_{0,j}^2, \sigma^2\}$; Regression Hyperparameters: $\{\vartheta, S^2, \kappa\}$

Algorithm :

The statistical framework is provided by the R package GenomeBasedModel [GBM], which was designed by Akio Onogi. It builds a hierarchical Bayesian model that uses genome-wide SNPs for the prior distributions of the genotype-specific parameters of statistical models in biology. The function GenomeBasedModel regresses the model parameters on SNPs and infers the model parameters and SNP effects on the parameters jointly.
The parameters of the GBM function have been reduced, therefore they have an impact on the quality of the result, for example :

- Maximum number of MPI-GWR loops = 2
- Number of MCMC iterations in MPI = 50
- Number of MCMC iterations in MPI, which is added after convergence = 200

3-3- QTL detection :

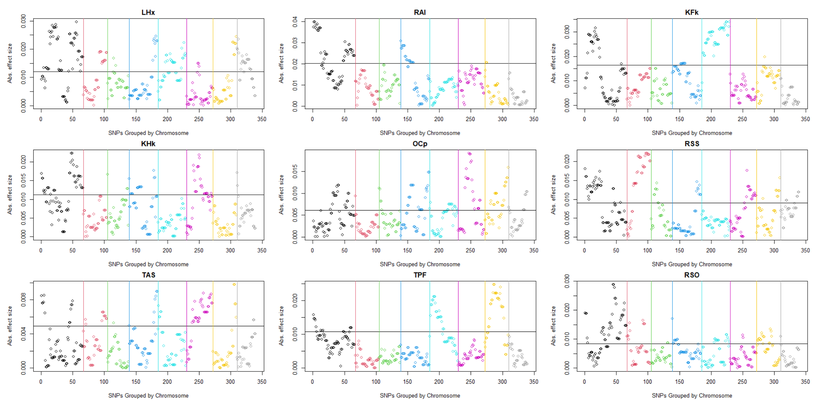


Figure 4: Absolut effect size of 340 SNPs from 8 chromosomes per Parameter.
The horizontal lines represent the thresholds using K-means with 2 clusters.

4- Conclusion :

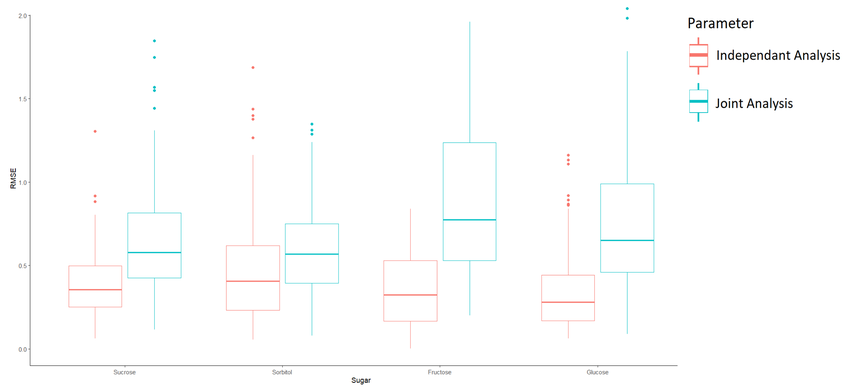


Figure 5: Boxplot of RMSE in the independent (SANN with Initial_3 set of values) and joint analysis (GBM) per sugar.

❑ The current implementation of GBM gives poorer minimization results of the objective function than SANN method. Further simulations with different settings are needed to conclude.

5- Perspective :

- Run the GBM function with more MCMC iterations to improve the estimation of ecophysiological parameters and QTL detection parameters.
- Use of machine learning methods for QTL detection without ecophysiological model.

6- References :

- GenomeBasedModel R-package:
- Hussein KANSO (2021) -doctoral thesis : Model reduction, parameter estimation for a genotype population and analysis due Genetic control *Cas du métabolisme des sucres dans la pêche*.
- Akio Onogi (2020)- Connecting mathematical models to genomes: joint estimation of model parameters and genome-wide marker effects on these parameters, *Bioinformatics, 36(10), 3169–3176*