Réduction de modèle, estimation des paramètres pour une population de génotypes et analyse du contrôle génétique : Cas du métabolisme des sucres dans la pêche

Hussein Kanso

AVIGNON UNIVERSITÉ

THÈSE DE DOCTORAT D'AVIGNON UNIVERSITÉ

École doctorale nº 536
Agrosciences & Sciences

Specialité :
Biostatistiques

Présentée par

Hussein KANSO

Réduction de modèle, estimation des paramètres pour une population de génotypes et analyse du contrôle génétique.

Cas du métabolisme des sucres dans la pêche

Soutenue publiquement le 25 mai 2021 devant le jury composé de :

Mme. Estelle Kuhn	**Rapporteure**	**Directrice de recherche, INRAE, Jouy-en-Josas**
M. Francis Mairet	**Rapporteur**	**Cadre de recherche, IFREMER, Nantes**
M. Eugenio Cinquemani	**Examinateur**	**Chargé de Recherche, INRIA, Grenoble**
Mme. Mathilde Causse	**Examinatrice**	**Directrice de recherche, GAFL, INRAE, Avignon**
Mme. Bénédicte Quilot-Turion	**Directrice de thèse**	**Directrice de recherche, GAFL, INRAE, Avignon**
Mme. Valentina Baldazzi	**Co-encadrante**	**Chargée de recherche, ISA, INRAE, Sophia Antipolis**
M. Mohamed-Mahmoud Memah	**Co-encadrant**	**Chargé de recherche, PSH, INRAE, Avignon**

PSH UR 1115 INRAE GAFL INRA AVIGNON

Résumé

Les modèles génotype-phénotype, permettant de tester les performances de génotypes sous différents climats, sont considérés comme des outils d'avenir pour concevoir de nouvelles variétés. Cependant, des progrès sont encore nécessaires pour inclure le contrôle génétique complexe dans les modèles basés sur les processus et réaliser l'intégration de l'information (métabolisme, contrôle enzymatique, génétique quantitative) du gène à la plante. Dans ce sens, un modèle cinétique du métabolisme des sucres a été développé par Desnoues et al. (2018) pour simuler les concentrations de différents sucres au cours du développement du fruit de la pêche. Les objectifs de mes travaux de thèse sont (a) d'estimer la variabilité inter-génotype des valeurs des paramètres du modèle et (b) d'étudier l'architecture du contrôle génétique des paramètres génotype-dépendants. Pour atteindre ces objectifs, il est nécessaire d'estimer les paramètres influents du modèle pour l'ensemble des 106 génotypes dont certains n'ont que peu de données observées. Le nombre de paramètres et la non-linéarité du modèle rendent la calibration du modèle peu fiable (notamment du fait du grand nombre de paramètres comparé au nombre de données disponibles et des corrélations entre eux) et coûteuse en terme de temps de calcul. Aussi, nous avons développé une stratégie de réduction du modèle initial visant à diminuer le nombre de paramètres et simplifier la structure du modèle tout en maintenant la structure du réseau et l'identité des variables, afin de faciliter leur interprétation biologique. L'estimation des paramètres du modèle réduit ainsi obtenu a été réalisée à l'échelle de la population en utilisant une approche non linéaire du modèle mixte (Baey et al. 2018), qui permet d'estimer simultanément les paramètres individuels pour tous les génotypes, et comparée aux méthodes plus conventionnelles qui procèdent à l'estimation des paramètres individuellement, génotype par génotype. Nous montrons que la fiabilité des estimations est largement améliorée et que les corrélations entre les paramètres estimés sont réduites. La dernière étape a consisté à estimer les liaisons entre les marqueurs du génome et les paramètres génotype-dépendants. Nous avons comparé les résultats d'une analyse dite indépendante (estimation de paramètres puis détection des régions génomiques impliquées) à une méthode d'analyse conjointe (Onogi 2020). Ces travaux de thèse constituent une étape essentielle au développement d'un outil qui prenne en compte un contrôle génétique complexe des caractères pour concevoir des idéotypes.

Mots clés : *Prunus persica*, fruit, modèle cinétique, réduction de modèle, calibration de modèle, variabilité inter-individuelle, modèle mixte non linéaire, optimisation, QTL de paramètres.

Abstract

Genotype-phenotype models are seen as future tools for designing new varieties, as they can help to test the performance of genotypes under different climates and cultural practices. However, progress is still needed to include complex genetic control in process-based models and to achieve the integration of information (metabolism, enzyme control, quantitative genetics) from gene to plant. A kinetic model of sugar metabolism was developed by Desnoues et al. (2018) to simulate the concentrations of different sugars during the development of the peach fruit. The objectives of my thesis work are (a) to estimate the inter-genotype variability of the model parameter values and (b) to study the architecture of genetic control of genotype-dependent parameters. To achieve these objectives, it is necessary to estimate the values of the influential parameters of the model for all 106 genotypes for which few data are available. The number of parameters and the non-linearity of the model make the calibration of the model inaccurate (in particular because of the large number of parameters compared to the number of available data and the correlations between the parameters) and costly in terms of calculation time. Therefore, we have developed a reduction strategy for the initial model to reduce the number of parameters and to simplify the model structure while maintaining the network structure and the identity of the variables to facilitate their biological interpretation. The estimation of the 9 parameters of the reduced model obtained was carried out at population scale using a nonlinear mixed-model (Baey et al. 2018) approach, which makes it possible to simultaneously estimate the values of the parameters of all genotypes. This strategy was compared to more conventional methods which carry out the estimation of the values of the parameters individually, genotype by genotype. We showed that the reliability of the estimates is greatly improved and those correlations between the estimated parameters are reduced. The last step consisted in estimating the links between the genome markers and the genotype-dependent parameters (QTL analyses). We compared the results of a so-called independent analyse (estimation of parameters then detection of the genomic regions involved) with a joint analysis method (Onogi 2020). This PhD work constitutes a crucial step towards the development of a tool that takes into account a complex genetic control of traits to design ideotypes.

Keywords: *Prunus persica*, fruit, kinetic model, model reduction, model calibration, Inter-individual variability, Non-linear mixed model, optimization, Model-based QTL, gene-to-phenotype.

Remerciements

ENFIN! Cette aventure a pris fin *Mardi 25 Mai 2021*. Après plusieurs mois de travail intense, épuisant mais également inoubliable au cours desquels on apprend, on grandit et on mûrit finalement. Cette thèse est le résultat d'un travail d'équipe qui a demandé la collaboration de plusieurs personnes.

Je tiens tout d'abord à remercier les rapporteurs de cette thèse pour leur lecture du manuscrit et l'intérêt qu'ils ont porté à mon travail. Je remercie également les examinateurs pour me faire l'honneur de participer au jury. Je remercie profondément les membres du jury pour leur conseils, suggestions et remarques qui m'ont permis d'améliorer la qualité de la mémoire.

Je souhaite remercier mes encadrants, ce travail n'aurait pas été possible sans leur présence. Merci pour votre confiance, malgré ma faible connaissance dans le domaine de la biologie. Je vous remercie pour le temps que vous m'avez accordé, votre patience et vos participations aux réunions très enrichissantes mais parfois intense!

Un grand merci à M. Olivier Bernard et M. Jean-Luc Gouzé (Inria, BIOCORE, Sophia Antipolis) pour leur implication dans le projet, notamment sur la problématique de simplification du modèle. J'adresse de chaleureux remerciements à Mme. Charlotte Baey (Université de Lille, Laboratoire Paul Painlevé, Villeneuve d'Ascq) pour son aide en statistique. Elle m'a beaucoup appris, j'ai apprécié son enthousiasme et sa sympathie.

J'adresse aussi mes remerciements aux membres du deux Unité GAFL et PSH pour leur accueil qui ont rendu le quotidien agréable. Je n'oublie pas les personnels administratifs et techniques du laboratoire, les soldats inconnus, sans qui tout aurait été plus compliqué.

Bien sûr, atteindre ces objectifs n'aurait pas été possible sans le soutien des amis et la famille. Un grand merci à vous!

Contents

1 Présentation générale

Sommaire

1.1 Introduction

LES fruits sont considérés comme des espèces présentant de multiples bienfaits pour la santé humaine (Slavin et al. 2012). Plusieurs recherches récentes confirment que la consommation de fruits réduit le risque de maladies majeures et retarde l'apparition de troubles liés à l'âge (Vincente et al. 2014). Sources de fibres, minéraux, oligo-éléments, vitamines et antioxydants, ils ont une importance nutritive majeure. Pour favoriser leur consommation et améliorer le régime alimentaire, il est désormais essentiel de les rendre plus appétissants et savoureux.

Les multiples composantes de la qualité des fruits sont influencées par différents facteurs, d'ordres génétiques et environnementaux ainsi que par le positionnement de la récolte par rapport à la maturité du fruit et par la durée de conservation post-récolte. La qualité organoleptique des fruits, primordiale pour le consommateur (Kader 2008), est largement liée aux concentrations en sucres (Vangdal 1985) qui dépendent à la fois de la variété, des pratiques culturales et du climat. Outre la qualité, d'autres critères tels que le rendement ; le calibre et la régularité de production ; la date de maturité et de nombreuses caractéristiques agronomiques sont importants à prendre en compte dans la sélection de nouvelles variétés. Aujourd'hui, des efforts croissants sont déployés pour sélectionner des variétés tolérantes aux stress biotiques et abiotiques.

Dans ce contexte multicritère, la modélisation mathématique apparaît comme un outil puissant permettant de mieux comprendre les liaisons entre les différents mécanismes mis en jeu et les interactions entre le génotype et l'environnement. Ainsi, les modèles génotype-phénotype sont considérés comme des outils prometteurs pour concevoir de nouvelles variétés (Letort et al. 2008 ; Tardieu 2003). Ces modèles peuvent, en effet, aider à tester *in silico* la performance de nouveaux génotypes sous différentes conditions climatiques. Cependant, des progrès sont encore nécessaires pour inclure un contrôle génétique complexe dans les modèles basés sur les processus et réaliser l'intégration de l'information (métabolisme, contrôle enzymatique, génétique quantitative) du gène à l'ensemble de la plante.

Les travaux de cette thèse s'inscrivent dans ce contexte de biologie prédictive où le développement d'outils mathématiques est nécessaire pour accélérer le progrès de la sélection pour des variétés adaptées aux environnements du futur. L'objectif est de progresser vers un modèle intégré de conception d'idéotypes de pêcher. Plus spécifiquement, il s'agit de développer un modèle du métabolisme des sucres adapté à une population de génotypes de façon à déterminer le contrôle génétique des paramètres génotype-dépendant du modèle. Pour cela, des approches complémentaires de mathématiques appliquées, statistiques et génétique ont été mobilisées. Ces travaux permettront dans une seconde étape de construire un modèle intégré cinétique-génétique permettant de raisonner la conception d'idéotypes.

Dans cette thèse, nous nous intéressons essentiellement à la concentration en sucres chez la pêche, en utilisant des données de croissance des fruits et de biochimie acquises en cinétique pour 106 individus d'une population de pêchers (Desnoues et al. 2014) et un modèle cinétique du métabolisme des sucres qui a été développé par Desnoues et al. (2018).

1.2 Le métabolisme des sucres et son contrôle génétique

1.2.1 Les sucres dans la pêche

Si le prix, l'aspect, l'odeur et la fermeté du fruit sont les premiers critères de choix lors de l'achat, son goût reste primordial. Le rapport entre les sucres et les acides détermine largement la douceur, l'acidité et l'intensité des saveurs des fruits (Stevens et al. 1977). Dans les fruits, il existe une grande variabilité des formes de sucres, de leurs proportions et de leurs concentrations selon les espèces. Les principaux sont le saccharose, le glucose, le sorbitol et le fructose. Le fructose possède le pouvoir sucrant le plus important des quatre et influence donc le plus la douceur du fruit.

Parmi les différentes espèces d'arbres fruitiers, le pêcher (*Prunus persica* (L.) Batsch) est, par bien des aspects, une espèce modèle intéressante. Tout d'abord son importance économique le place au deuxième rang mondial en termes de production parmi les Rosacées fruitières, derrière la pomme et au même rang que la poire (https://fr.statista.com/statistiques/673783). De plus, il présente une période juvénile relativement courte (2 à 3 ans) en comparaison aux autres arbres fruitiers (entre 5 et 10 ans pour le pommier, poirier et cerisier). D'un point de vue génomique, des ressources importantes sont disponibles avec le séquençage et l'annotation du génome du pêcher (Verde et al. 2013). Il s'agit d'une espèce diploïde (2n=2x=16) avec un génome de petite taille (265 Mb) soit environ 2 fois le génome d'*Arabidopsis thaliana*. Il existe une grande variabilité du niveau des concentrations en sucres en fonction du génotype, notamment la présence d'une très faible concentration en fructose chez certains pêchers sauvages ou ornementaux. Aussi, une bonne connaissance des mécanismes impliqués dans l'accumulation des sucres au niveau du fruit, principalement des mécanismes entraînant une variabilité des concentrations en sucres, ainsi que leur déterminisme génétique, est essentiel pour guider la création variétale.

Le carbone issu de la photosynthèse foliaire est transloqué via le phloème vers les fruits. Ensuite la majeure partie du métabolisme des sucres dans le fruit se déroule dans le cytosol des cellules mais les sucres sont principalement stockés dans la vacuole. Les principaux facteurs au niveau cellulaire pouvant entraîner une modification importante des concentrations et des proportions relatives en différents sucres sont

(1) les activités enzymatiques qui peuvent moduler la transformation d'un sucre en un autre et (2) la compartimentation cellulaire avec un stockage plus ou moins efficace dans la vacuole en fonction des espèces. Les enzymes, protéines dotées de propriétés catalytiques, permettent de catalyser une réaction précise avec un substrat donné : elles le découpent ou le décomposent en molécules plus simples. Les transporteurs, quant à eux, facilitent le passage d'un soluté d'un compartiment à l'autre : simple canal pour un passage dans le sens du gradient de concentration entre deux compartiments ou au contraire transporteur actif nécessitant de l'énergie pour permettre une accumulation de composés dans la vacuole par exemple. Le métabolisme des sucres dans un fruit fait intervenir de nombreux enzymes et transporteurs dont on ne connaît pas encore bien toutes les caractéristiques biochimiques et le rôle dans le contrôle génétique.

1.2.2 L'analyse génétique de la relation phénotype-génotype

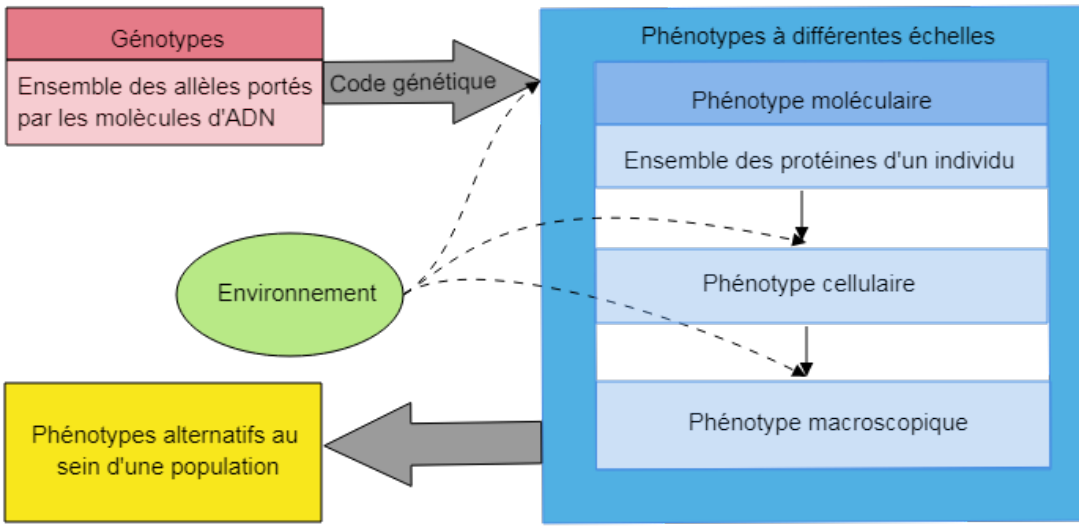


FIGURE 1.1 – Du génotype aux phénotypes à différentes échelles. L'environnement peut moduler les phénotypes à différents niveaux

Tous les individus d'une même espèce ont les mêmes gènes, mais chaque individu a un ensemble spécifique d'allèles. Chez les organismes diploïdes, un individu donné possède, pour chaque gène, un allèle reçu de son père et un allèle reçu de sa mère. Le génotype d'un individu est l'ensemble de ses allèles. En interaction avec l'environnement, il détermine une large part du phénotype d'un organisme. Le phénotype correspond quant à lui à l'ensemble des caractères observables de l'individu, et résulte donc de l'expression des gènes. Le phénotype peut se définir à différentes échelles : moléculaire, cellulaire et macroscopique. Cependant, la relation génotype-phénotype est complexe du fait, d'une part, que plusieurs allèles interviennent directement ou

indirectement dans la réalisation d'un phénotype et, d'autre part, que les facteurs environnementaux, en agissant à différents niveaux du phénotype, peuvent le modifier.

1.2.2.1 Construction de cartes génétiques

Une carte génétique est une représentation du génome. Ainsi, la construction de cartes génétiques à partir de marqueurs moléculaires permet de localiser des loci impliqués dans la variation de caractères quantitatifs (QTL : Quantitative Trait Locus). Un marqueur moléculaire est une séquence polymorphe d'ADN aisément détectable, située à un locus unique et qui se transmet selon les lois mendéliennes. Il est utilisé pour "baliser" le génome. Les marqueurs génétiques permettent de suivre la transmission d'un segment de chromosome d'une génération à l'autre. La construction de cartes génétiques repose sur les propriétés de ségrégation des gènes. En effet, à la méiose, les chromosomes non homologues ségrègent indépendamment. Il en est de même pour les gènes portés par ces chromosomes. En revanche, des gènes portés par un même chromosome ne ségrègent pas indépendamment, et cela d'autant moins qu'ils sont proches. Ce sont ces propriétés que l'on met à profit pour estimer les positions des marqueurs. Lorsque des marqueurs ne ségrègent pas indépendamment dans une descendance, on dit qu'ils sont « liés ». Des tests statistiques simples permettent de décider si deux marqueurs sont indépendants ou non, d'estimer l'ordonnancement de ces marqueurs et leurs distances génétiques, sur différents groupes de liaison. Lorsque le nombre de marqueurs utilisés est suffisamment élevé, le nombre de groupes de liaison correspond au nombre de chromosomes (Morton 1955 ; Vienne 1998).

1.2.2.2 Cartographie de QTL

Disposer d'une telle carte génétique, balisant le génome pour une population ou descendance, permet de déterminer, par cartographie de QTL, les régions du génome vraisemblablement responsables de la variation phénotypique. L'objectif d'une cartographie de QTL est ainsi d'identifier un certain nombre de loci influençant la valeur d'un caractère, leur localisation sur le génome, leur effet ainsi que l'origine parentale de l'allèle favorable ou défavorable. La recherche de QTL se base sur le principe de liaison statistique entre le phénotype et le génotype. Elle repose sur des tests de différences phénotypiques entre classes de descendants selon l'origine grand-parentale de segments chromosomiques hérités d'un parent. La détection de QTL consiste donc pour un marqueur donné M, à observer dans la descendance d'un parent hétérozygote M_1/M_2 (M_1 et M_2 étant 2 allèles différents au marqueur M) s'il existe une différence de trait moyen selon l'allèle du marqueur, M_1 ou M_2, transmis. L'hypothèse sous-jacente est que, si cette différence existe, elle s'explique par la ségrégation des allèles Q_1 et Q_2, en un QTL génétiquement lié au marqueur M. Sous l'hypothèse nulle (absence de QTL), la valeur phénotypique moyenne est indépendante du génotype au marqueur (Soller et al. 1976). En conséquence, l'observation d'une différence significative entre les valeurs moyennes du caractère quantitatif des génotypes au marqueur indique la possibilité d'une liaison du marqueur avec un QTL. Des tests statistiques

appropriés permettent de tester cette hypothèse (Terwilliger 1995; Boitard et al. 2006; Farnir et al. 2002).

Les effets génétiques des QTL et la valeur phénotypique d'un trait quantitatif sont généralement décrits par un modèle linéaire. Comme les localisations des QTL ne sont pas connues a priori, nous utilisons souvent des marqueurs pour représenter les QTL. Certains marqueurs peuvent être liés à un ou plusieurs QTL, et peuvent donc présenter une forte association avec le trait. La plupart des marqueurs, cependant, peuvent ne pas être directement liés au QTL, et donc aucune association ne sera attendue entre ces marqueurs et le trait. La cartographie des intervalles ou l'analyse à un marqueur unique ne fournit pas d'estimations précises des effets du QTL. Cependant, le modèle à régression multiple est un choix raisonnable pour la cartographie des traits quantitatifs (Kao et al. 1999).

Supposons que nous avons une population de n individus, le modèle de régression linéaire est le suivant

$$y^{(k)} = B_0 + \sum_{m=1}^{M} \varkappa_m^{(k)} B_m + e_k \tag{1.1}$$

avec $y^{(k)}$ est le trait quantitative (observation) de l'individu k, B_0 la moyenne de la population, M le nombre de marqueurs du génome entier et $\varkappa_m^{(k)}$ désigne le génotype au marqueur m pour l'individu k et vaut 1 pour Q_1Q_1 et 2 pour Q_1Q_2. B_m représente l'effet du marqueur m sur le trait y et e_k est l'erreur résiduelle qui est supposée suivre une distribution normale de moyenne zéro et de variance constante.

Les méthodes d'ajustement des modèles de QTL (Broman et al. 2009) occupent une place centrale dans toute procédure de cartographie des QTL. L'ajustement d'un modèle pourrait être réalisé en réduisant la somme des carrés des résidus (Residual Sum of Squares en Anglais) :

$$RSS = \sum_{k=1}^{n} (y^{(k)} - \hat{y}^{(k)})^2 \tag{1.2}$$

avec $\hat{y}^{(k)}$ la valeur ajustée par le modèle. Traditionnellement, la détection de QTL est mesurée par un score, le LOD, ou rapport de vraisemblance comparant les hypothèses de présence et d'absence de QTL à chaque position sur le génome.

$$LOD = \frac{n}{2} log_{10}\left(\frac{RSS_0}{RSS}\right), \tag{1.3}$$

avec $RSS_0 = \sum_{k=1}^{n} (y^{(k)} - \overline{y}^{(k)})^2$ la somme des carrés des résidus pour le modèle nul où $\overline{y}^{(k)}$ est la valeur moyenne du trait y pour l'individu k. D'autres méthodes peuvent être utilisées, comme par exemple :

- la cartographie des intervalles avec la méthode de régression de Haley-Knott, telle que décrite par Haley et al. (1992) et Broman et al. (2009) qui repose sur

l'approximation des modèles mixtes et/ou la maximisation de la vraisemblance.

- les méthodes de régression régularisées, telles que la régression de Ridge (Hoerl et al. 1970), le rétrécissement de régression et sélection (LASSO : Least Absolute Shrinkage and Selection Operation) (Tibshirani 1996), le LASSO Bayésien (BL) (Park et al. 2008) ou le LASSO Bayésien Etendu (EBL) (Mutshinda et al. 2010), sont essentiellement des procédures de probabilité pénalisée, où des fonctions de pénalisation appropriées sont ajoutées à la log-vraisemblance négative pour réduire automatiquement les effets parasites. Onogi et al. (2016b) introduit bien la définition des méthodes de régression mentionnées précédemment et propose un Package R qui implémente différentes méthodes de régression.

1.2.3 Le contrôle génétique du métabolisme des sucres

La concentration en sucres est un caractère quantitatif qui est sous contrainte génétique complexe avec plusieurs régions génomiques (loci) impliquées dans leurs variations (par opposition à un caractère qualitatif ou binaire qui est quant à lui contrôlé par un gène unique). Chez la tomate, 95 QTL liés aux sucres ont été identifiés d'après la revue de Labate et al. (2007). Lerceteau-Köhler et al. (2012) ont quant à eux mis en évidence chez la fraise 12 QTL liés aux sucres et 21 liés aux acides. Un grand nombre de QTL sont aussi identifiés chez le pêcher (Sosinski et al. 1997; Dirlewanger et al. 1999; Etienne et al. 2002; Quilot et al. 2004b).

L'annotation du génome offre la possibilité de rechercher des gènes candidats à la fois positionnels (localisés dans l'intervalle de confiance d'un QTL) et fonctionnels (codant pour une protéine dont la fonction peut impacter le caractère) sous-jacents à ces QTL. Des études moléculaires complémentaires permettent ensuite d'identifier le gène et sa fonction avec certitude. Ainsi chez la tomate, un gène codant une invertase pariétale (Fridman et al. 2000) et des gènes codant pour des isoenzymes de la fructokinase (Kanayama et al. 1997; Petreikov et al. 2001) contrôlent les concentrations en fructose et glucose. De même, des transporteurs spécifiques des sucres ont été identifiés chez arabidopsis (Chardon et al. 2013), le riz (Zhou et al. 2014; Eom et al. 2011) et la poire (Zhang et al. 2013).

1.3 La modélisation mathématique du métabolisme

1.3.1 Les modèles du métabolisme des sucres

De nombreuses études ont porté sur la compréhension du métabolisme des sucres chez les plantes mais peu de modèles mathématiques sont aujourd'hui disponibles.

Cependant, un formalisme mathématique a été proposé très tôt par Henri (1903) pour représenter la vitesse d'une réaction enzymatique et l'équation de Michaelis et

Menten (Johnson et al. 2011) qui en découle est encore aujourd'hui utilisée :

$$V = \frac{V_{max}[S]}{K_m + [S]} \tag{1.4}$$

V_{max} correspond à la vitesse limite de la réaction c'est-à-dire la vitesse de synthèse du produit pour une concentration du substrat saturante. Le K_m représente la concentration de substrat pour laquelle la vitesse est égale à la moitié de la vitesse limite (V_{max}).

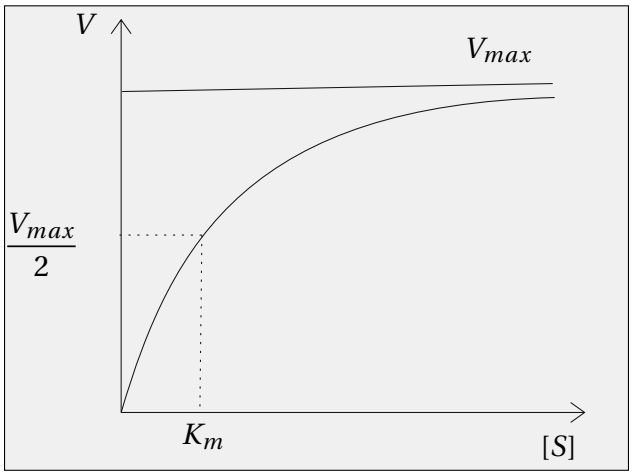


FIGURE 1.2 – Variation de la vitesse d'une réaction enzymatique (V) en fonction de la concentration en substrat ($[S]$)

.

Sur la base des propriétés cinétiques des enzymes qui interviennent dans un réseau métabolique et sa stoechiométrie, différents types de modèles cinétiques ont été proposés. Dans un article de revue, Rohwer (2012) distingue trois types de modèles dédiés : au régime dynamique, au régime stationnaire, et à l'analyse du contrôle métabolique. Il discute également du potentiel de la modélisation cinétique comme outil puissant permettant notamment d'identifier des cibles régulant des réseaux ou de déterminer l'importance de différentes enzymes (y compris les isoformes de la même enzyme) dans une voie.
Dans les modèles dynamiques, l'évolution des concentrations en métabolites est souvent exprimée par des équations différentielles ordinaires comme étant la différence entre la somme des taux de réaction synthétisant le composé et la somme des taux de réaction le dégradant. La résolution de ce système d'équations différentielles permet de simuler les métabolites en fonction du temps. L'un des principaux points limitant réside dans le nombre de données disponibles, par exemple les mesures des capacités enzymatiques dans les échantillons biologiques correspondant aux simulations. En cas d'absence de données sur les paramètres cinétiques des enzymes, il est cependant possible d'estimer ces paramètres manquant par optimisation en comparant les

sorties du modèle avec des données mesurées (généralement des concentrations de métabolites) (voir Section 1.4). Suivant ces principes, des modèles cinétiques ont été développés pour l'accumulation de saccharose lors de la maturation de la tige de la canne à sucre (Uys et al. 2007) et chez Arabidopsis, plante modèle (Nägele et al. 2014), le métabolisme de l'aspartate (Curien et al. 2009) et des glucides (Nägele et al. 2010). Un modèle du métabolisme des sucres constitué d'états stationnaires à différents stades et permettant de simuler l'évolution des flux au cours du temps a été proposé chez la tomate par Beauvoit et al. (2014).

Chez la pêche, le premier modèle simulant l'accumulation de sucres au cours du développement du fruit a été développé par Génard et al. (1996). Le modèle explore la diversité génétique et met l'accent sur les différences observées dans le rapport fructose-glucose (Wu et al. 2012). L'une des limites de ce modèle est qu'il ne prend pas en compte les compartiments cellulaires. Cependant, la compartimentation sous-cellulaire est essentielle pour l'accumulation des sucres (Génard et al. 2014 ; W. Patrick et al. 2013) et peut avoir un effet significatif sur leurs concentrations (Chardon et al. 2013). Ainsi, un modèle du métabolisme de sucres chez la pêche a été récemment développé par Desnoues et al. (2018), en prenant en compte à la fois la compartimentation cellulaire et la cinétique des réactions enzymatiques. C'est ce modèle que nous allons utiliser dans cette thèse (voir Section 1.6.2).

1.3.2 Méthodes de réduction et analyse des modèles

Les modèles dynamiques appliqués à la biologie sont souvent complexes, en raison de leur non-linéarité et de leur taille. Ainsi, leur calibration peut être incertaine et coûteuse en temps de calcul, alors qu'une étude exhaustive des comportements dynamiques du modèle est difficile avec les techniques traditionnelles des systèmes dynamiques et passe souvent par la simulation numérique, à son tour fonction des paramètres du modèle.

Une façon de surmonter les problèmes liés à la complexité des modèles consiste à en simplifier la structure (Okino et al. 1998). Plusieurs techniques de décomposition et de simplification ont été développées ces dernières années afin de réduire la complexité (Gorban et al. 2006 ; Snowden et al. 2017) ou faciliter l'analyse des comportements du modèle en fonction d'objectifs, telles que l'identifiabilité et l'analyse des états d'équilibre.

1.3.2.1 Analyse de sensibilité

L'analyse de sensibilité (AS) est une méthode qui permet d'analyser l'incertitude sur la sortie d'un modèle (numérique ou autre) en l'attribuant à différentes sources d'incertitude en entrée du modèle (Saltelli et al. 2004 ; Turányi 1990 ; Cariboni et al. 2007). Ainsi, l'analyse de sensibilité peut être utilisée pour identifier les paramètres les plus influents du système, à hiérarchiser les entrées à mesurer en priorité ou tout simplement à définir les limites de validité du modèle. Les méthodologies d'AS les

plus courantes peuvent être classifiées en approches locales et globales (Pianosi et al. 2016 ; Saltelli et al. 2008).

Dans les approches locales (également appelées méthodes OAT, one-at-a-time), l'effet de la variation d'un seul facteur est estimé en maintenant tous les autres fixés à leur valeur nominale (Cacuci et al. 2005 ; Borgonovo et al. 2004). Les approches globales estiment, au contraire, l'effet, sur les sorties du modèle, d'un facteur lorsque tous les autres varient, ce qui permet d'identifier les interactions dans des modèles non linéaires et/ou non additifs (Iooss et al. 2015). L'application de l'AS locale exige évidemment une détermination de la valeur nominale pour les facteurs d'entrée. Par contre, l'AS globale permet de surmonter cette limitation, mais nécessite néanmoins de spécifier l'espace de variabilité d'entrée.

Méthodes d'analyse de sensibilité locales

La méthode la plus simple pour obtenir une mesure de sensibilité est de calculer la dérivée de la sortie par rapport au facteur d'entrée d'intérêt. Ainsi, définissons le modèle comme :

$$Y = f(x_1, x_2, \dots x_n)$$

où Y représente la sortie du modèle et $x = (x_1, \dots, x_n)$ est le vecteur des facteurs d'entrée. On peut définir un indice de sensibilité locale comme suit :

$$S_{x_i} = \frac{\sigma_{x_i}}{\sigma_Y} \frac{\partial Y}{\partial x_i}\bigg|_{x_j = x_j^*, j \neq i} \tag{1.5}$$

où σ_Y et σ_{x_i} sont, respectivement, les écarts-types de Y et du facteur x_i et x_j^* la valeur nominale des autres facteurs d'entrée.

Méthodes d'analyse de sensibilité globales

Il existe dans la littérature beaucoup de méthodes de calcul des indices de sensibilité globale. Les méthodes de Sobol (Sobol 1993) et FAST (Fourier Amplitude Sensitivity Test) (Cukier et al. 1973) sont les plus courantes et reposent sur la décomposition complète de la variance de la sortie du modèle.

Ainsi, en reprenant les notations d'avant, la variance de la sortie Y sera décomposée comme

$$Var(Y) = V(x_1) + \dots + V(x_n) + V(x_1, x_2) + \dots + V(x_{n-1}, x_n) + \dots + V(x_1 \dots + x_n) \tag{1.6}$$

en estimant la part de variance (dispersion) induite par chaque entrée perturbée x_i (effets principaux et interactions).

A partir de ce résultat, on définit un indice de sensibilité SI d'un facteur ou d'un groupe de facteurs x_U comme

$$SI_{x_u} = \frac{V(x_u)}{Var(Y)} \tag{1.7}$$

L'indice de sensibilité de x_u, notée SI_{x_u}, est compris entre 0 et 1. La somme des indices est égale à 1. Les facteurs qui ont les indices les plus élevés influencent le plus la réponse Y du modèle. Ils sont par conséquent les facteurs les plus importants dans le modèle, ce sont ceux dont il faudra avoir une connaissance la plus précise possible. La décomposition de la variance permet de calculer ces indices de sensibilité de chaque terme du modèle. Ainsi, l'indice de premier (SI) ordre s'obtient par

$$SI_i = \frac{Var(\mathbb{E}(Y x_u))}{Var(Y)}, \quad i = 1, \ldots, n \tag{1.8}$$

L'indice de sensibilité totale, noté TSI, correspond à tous les effets associés au facteur x_i. C'est donc la somme des indices associés à x_i. Il est calculé comme la variance de Y induite par x_i sachant les autres variables fixées. x_{-i} désigne l'ensemble des facteurs à l'exception de x_i.

$$TSI_i = 1 - \frac{Var(\mathbb{E}(Y x_{-u}))}{Var(Y)}, \quad i = 1, \ldots, n \tag{1.9}$$

Ainsi, les facteurs avec une TSI faible sont des facteurs qui peuvent être fixés sans influencer la sortie Y.

Dans les cas où le modèle contient un grand nombre de facteurs et/ou est trop coûteux en temps de calcul, l'application des méthodes FAST ou Sobol serait aussi très coûteuse en temps de calcul. Dans ces cas, la méthode de criblage développée par Morris (1991) et étendue par Campolongo et al. (2007) peut être utilisée. Cette méthode est moins coûteuse sur le plan des calculs, et peut être utilisée pour identifier des facteurs non influents. Néanmoins, elle ne fournit pas la décomposition complète de la variance, comme dans les méthodes précédentes. Dans le cas des modèles à plusieurs sorties, une extension de l'analyse de sensibilité globale a été développée par Lamboni et al. (2011) pour mesurer la contribution des facteurs d'entrées.

Les méthodes d'analyse de sensibilité sont fréquemment utilisées pour classer les paramètres en fonction de leur impact sur les sorties du modèle. Par exemple, plusieurs études ont utilisé des analyses de sensibilité pour sélectionner les paramètres clés de modèles basés sur des processus, ce qui permet d'identifier les principaux paramètres génétiques et d'adapter facilement le modèle à une nouvelle variété (Quilot-Turion et al. 2012 ; Martre et al. 2015). L'analyse de sensibilité est utilisée aussi pour faciliter la calibration des modèles complexes en réduisant le nombre de paramètres à estimer en identifiant ceux qui sont peu influents et qui peuvent être fixés à leurs valeurs nominales connues (Mathieu et al. 2018).

1.3.2.2 Simplification des modèles dynamiques

Des méthodes de réduction et de simplification des modèles peuvent être employées pour diminuer leur complexité et ainsi permettre une analyse plus fiable de leur comportement. Deux grandes classes de méthodes peuvent être identifiées :

- des approches "structurelles" qui visent à réduire la taille du système à analyser

en se basant essentiellement sur la topologie du réseau et
- des approches "dynamiques" qui visent à simplifier la forme mathématique du système, réduisant les effets non-linéaires ou les nombres de paramètres, ce qui permet d'obtenir des systèmes qui conservent une capacité prédictive assez précise avec une complexité réduite.

Méthodes de simplification structurelle

Les méthodes de décomposition des modèles visent à séparer le système en sous-réseaux ou sous-modèles plus faciles à analyser et à paramétrer. Certains d'entre eux utilisent les propriétés géométrique du réseau pour identifier les réactions pouvant relier de zones fortement connectés du graphe ('modules') (Holme et al. 2003 ; Saez-Rodriguez et al. 2008), alors que d'autres allient topologie et dynamique du système pour définir des sous-unités fonctionnelles (Anderson et al. 2011 ; Sun et al. 2016).

Une autre approche pour simplifier un système d'EDOs est la technique classique de regroupement (lumping) Wei et al. (1969) et Li et al. (1989). Basée sur la séparation dans le temps, elle permet généralement d'éliminer les états non dynamiques (Okino et al. 1998 ; Tomlin et al. 1997). Le regroupement est centré autour de l'identification de groupes de variables qui peuvent être approchées par une seule variable groupée. Cette approche a été appliquée avec succès à différents modèles : un modèle pharmacocinétique Brochot et al. (2005) et Dokoumetzidis et al. (2009), un modèle d'émission de fluorescence en photosynthèse Sunnåker et al. (2010) et un modèle d'hydrotraitement du gazole García et al. (2010).

Méthodes de simplification dynamique

Les modèles de réactions chimiques impliquent souvent un nombre très grand d'espèces chimiques (substrats et produits). Certaines de ces espèces sont des espèces très réactives et peuvent constituer des intermédiaires importants dans le schéma réactionnel. Un grand nombre de réactions élémentaires peuvent se produire entre les espèces ; certaines de ces réactions sont rapides et certaines sont lentes.

L'approche d'approximation quasi-stationnaire (QSSA) est un moyen mathématique pour simplifier les équations différentielles décrivant certains systèmes cinétiques chimiques. La QSSA repose sur l'idée d'étudier la dynamique des réactions lentes, en supposant que les réactions rapides sont quasi-statiques et capables de suivre quasi-instantanément les variations des variables lentes. Par conséquent, les équations différentielles décrivant l'évolution des variables quasi-statiques peuvent être mises à l'équilibre dans le système EDOs (Schauer et al. 1983 ; Heinrich et al. 1996).

Cependant l'identification des variables lentes et rapides n'est pas toujours simple, surtout pour des modèles de grand taille. Récemment, López Zazueta et al. (2018) ont montré que, dans un réseau de réactions de type Michaelis-Menten, les variables quasi-statiques ont une concentration d'un ordre de grandeur inférieur aux variables à dynamique lente et peuvent être considérées à l'équilibre dans le système EDO.

En alternative, un changement de base des variables d'état peut être utilisé pour obtenir un modèle transformé où la séparation des échelles de temps est nettement plus apparente et exploitable. De telles approches peuvent souvent conduire à des modèles de plus faible dimension et plus précis. Ces approches visent à trouver une transformation des variables d'état permettant de découpler les dynamiques rapides et lentes, puis de réduire le système tout en conservant un haut degré de précision du modèle simplifié (Zobeley et al. 2005 ; Surovtsova et al. 2006 ; Vallabhajosyula et al. 2006).

En plus de l'analyse d'échelle de temps, le choix de la façon de modéliser une réaction ou un processus dans un système joue également un rôle primordial au niveau de la complexité du modèle. Les modèles métaboliques sont en effet généralement non linéaires. Dans cette perspective, l'utilisation de cinétique enzymatique simplifiée (Wang et al. 2007 ; Nikerel et al. 2009 ; Schmidt et al. 2008) peut être utile pour éviter les problèmes numériques et améliorer l'identifiabilité du système.

1.4 La calibration et sélection d'un modèle dynamique

Le métabolisme des sucres résulte de nombreuses réactions, parfois non linéaires, combinées en réseau, ce qui se traduit par des modèles mathématiques de structure assez complexe. Cette complexité a des conséquences en termes d'identifiabilité des paramètres, de calibration et de confrontation à des données expérimentales. Aussi, des outils mathématiques et statistiques puissants sont nécessaires pour résoudre ces problèmes. A ces difficultés s'ajoute une contrainte majeure liée à l'intégration de l'information génétique dans ces modèles : la difficulté de calibration de ces modèles qui comportent en général un grand nombre de paramètres pour un grand nombre de génotypes (avec un nombre restreint d'observations) (Martre et al. 2015 ; Bertin et al. 2010 ; Barrasso et al. 2019). Ceci d'autant plus qu'en général le nombre d'observations pour chaque individu est assez restreint du fait de la difficulté à produire ces données en cinétique.

Pour résoudre ces problèmes, différentes méthodes peuvent être mobilisées : la vérification de l'identifiabilité des paramètres, la réduction de la complexité du modèle et la sélection d'une méthode de calibration adaptée aux caractéristiques d'une population génétique.

1.4.1 L'identifiabilité des paramètres du modèle dynamique

La question d'identifiabilité d'un modèle se subdivise en deux aspects complémentaires : l'identifiabilité structurelle et l'identifiabilité pratique .

Les paramètres d'un modèle sont **structurellement identifiables** s'il est théoriquement possible de les déterminer d'une façon unique à partir de l'observation des sorties (supposées parfaites) et du système d'équations dynamiques qui le constitue

(Bellman et al. 1970). Ainsi, cette approche peut être appliquée *a priori* sur les équations du modèle et avant la disponibilité des données expérimentales. L'identifiabilité structurelle est une étape importante vers l'évaluation des paramètres du système, car elle définit une condition nécessaire à l'unicité des paramètres. Ainsi, une analyse d'identifiabilité structurelle est fortement conseillée avant de calibrer un modèle, afin de vérifier sa bonne formulation mathématique (Bandara et al. 2009).

L'**identifiabilité pratique** se réfère à la quantification de l'incertitude des valeurs des paramètres lorsqu'ils sont estimés à partir des données expérimentales (Walter et al. 1997). Si l'identification du modèle est structurellement possible, elle est en effet beaucoup plus difficile en pratique et dépend fortement de la qualité des données (Jaqaman et al. 2006). Ainsi, des données éparses ou bruitées peuvent conduire à une estimation peu fiable des valeurs de paramètres ou à l'émergence de corrélations entre paramètres. Contrairement aux approches d'identifiabilité structurelle, l'analyse d'identifiabilité pratique est par définition une analyse *à posteriori*, entièrement basée sur le jeu de données utilisé pour la calibration du modèle. L'acquisition de nouvelles données peut ainsi modifier l'identifiabilité du système (Raue et al. 2009 ; Apgar et al. 2010).

Différentes approches sont proposées en littérature pour analyser l'identifiabilité des paramètres des systèmes biologiques. Certaines méthodes permettent de tester l'identifiabilité globale, propriété de toutes les valeurs de paramètres possibles, c'est-à-dire indépendamment de la valeur réelle du paramètre. D'autres approches permettent de tester l'identifiabilité locale en maintenant un point dans l'espace des paramètres. Parmi les approches existantes : l'approche développée par Villaverde et al. (2016) est basée sur une analyse d'observabilité, décomposition et dérivées de Lie pour montrer l'identifiabilité structurelle des modèles en biologie (génétiques, signalisation, métaboliques et pharmacocinétiques). L'approche DAISY, développée par Bellu et al. (2007), est basée sur l'algèbre différentielle pour effectuer une analyse d'identifiabilité structurelle globale des paramètres pour les modèles dynamiques décrits par des équations polynomiales ou rationnelles. L'approche EAR s'intéresse quant à elle à l'identifiabilité structurelle locale, basée sur l'application du théorème de la fonction inverse au système d'équations algébriques (Pohjanpalo 1978). Enfin, l'approche PL, développée par Raue et al. (2009), gère à la fois l'identifiabilité structurelle et pratique. Elle est basée sur le profil de la vraisemblance. Les paramètres structurellement non-identifiables sont caractérisés par une vraisemblance de profil plat. Pour les paramètres pratiquement non-identifiables, le profil a un minimum qui ne dépasse pas le seuil de la vraisemblance pour les valeurs croissantes et/ou décroissantes du paramètre considéré.

Raue et al. (2014) ont comparé les approches EAR, DAISY et PL. Les auteurs suggèrent d'utiliser EAR pour un grand système ; pour une identifiabilité globale, DAISY est l'approche recommandée ; enfin, pour une identifiabilité pratique ou si les équations incluent des expressions non rationnelles, PL doit être préférée.

1.4.2 Méthodes d'estimation de paramètres

Après la construction et l'évaluation d'un modèle dynamique, la question se pose autour du choix de la méthode d'estimation à utiliser. Quelle méthode choisir et comment? Ce choix est souvent dicté par le type d'observations disponibles, le type de modèle à calibrer, et bien sûr par la nature du problème à traiter.

L'estimation des paramètres consiste à calibrer les valeurs des paramètres à partir de données expérimentales et/ou d'informations issues de l'expertise. Le processus d'estimation des paramètres commence par la définition de la fonction de vraisemblance (ou une fonction objectif), mesurant la distance entre les prédictions du modèle et les observations. Différentes hypothèses peuvent être considérées concernant la relation et l'erreur associée aux données expérimentales, conduisant à différentes formulations possibles de la fonction de vraisemblance (Venter 2010). Deux formulations de la fonction objectif peuvent être considérées pour l'optimisation :

- L'optimisation mono-objectif où un seul critère est à optimiser sous un certain nombre de contraintes d'égalité ou d'inégalité. L'objectif principal est donc de trouver la "meilleure" solution, qui correspond à la valeur minimale ou maximale d'une fonction objectif. Le problème d'optimisation peut être écrit comme suite :

$$\begin{cases} \min f(x) & \\ g_i(x) \leq 0, & i = 1, \ldots, m, \\ h_j(x) = 0, & j = 1, \ldots, p, \\ x^L \leq x \leq x^U & \end{cases}$$

  où $x \in \mathscr{R}^n$ est un vecteur de variables de décision, $g_1, \ldots, g_m$ et $h_1, \ldots, h_p$ sont les fonctions de contraintes d'inégalité et des contraintes d'égalité, respectivement; et $x^L \in \mathscr{R}^n$, $x^U \in \mathscr{R}^n$ sont respectivement les bornes inférieure et supérieure du domaine de recherche. Cet ensemble de contraintes définit l'espace faisable C des variables de décision :

$$C = \left\{ x \in \mathscr{R}^n / h(x) = 0, \;\; g(x) \leq 0 \;\; et \;\; x^L \leq x \leq x^U \right\}$$

  Résoudre ce problème d'optimisation mono-objectif consiste à trouver la solution $x^* \in C$ telle que :

$$\forall x \in C, f(x^*) \leq f(x)$$

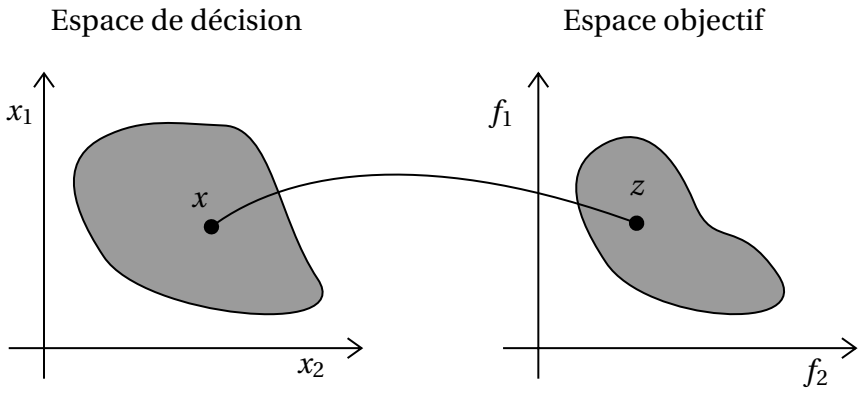


FIGURE 1.3 – Concepts de base : Espace des variables de décision (à gauche) et espace des objectifs (à droite)

- L'optimisation multi-objectif où plusieurs critères (objectifs) doivent être pris en compte. Les objectifs doivent être aussi indépendants les uns des autres que possible pour éviter de se retrouver dans des optima locaux Knowles et al. (2001). Dans ce cas, il n'existe généralement pas une solution optimale unique comme avec une optimisation à un objectif mais un ensemble de solutions. Cet ensemble de solutions, largement connues sous le nom de solutions compromis, non dominées, non inférieures ou Pareto-optimales (exemple la Figure 1.4).

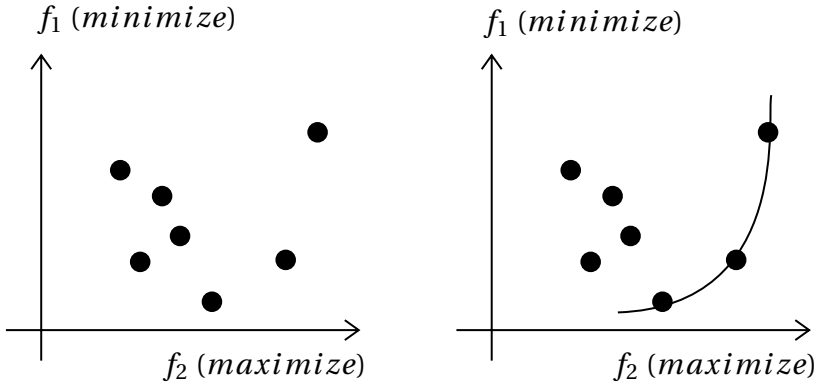


FIGURE 1.4 – Illustration de l'optimalité de Pareto dans l'espace des objectifs

Un problème d'optimisation multiobjectif peut être définit comme suit :

$$\min \{f_1(x), f_2(x), \ldots, f_m(x)\}^T, \ sous\ la\ contrainte,\ x \in C$$

avec m le nombre de fonctions objectif ($m \geq 2$), $x = (x_1, \ldots, x_n)$ le vecteur représentant les variables de décision, C l'ensemble des solutions associé à des contraintes d'égalité, d'inégalité et des bornes explicites comme dans le cas mono-objectif.

On distingue classiquement deux types de méthodes d'optimisation :

- L'optimisation locale recherche une solution qui est la meilleure localement, c'est-à-dire que dans son 'voisinage' aucune solution n'est meilleure qu'elle.

Cette solution est appelée un optimum local. Pour un problème de minimisation, $x^* \in \mathscr{R}^n$ représente un minimum local, s'il existe un voisinage de x^*, $V(x^*)$, tel que :

$$\forall x \in V(x^*) \ on \ a \ que \ f(x^*) \leq f(x)$$

- L'optimisation globale recherche quant à elle la meilleure solution dans un espace de recherche entier, c'est-à-dire que dans tout l'espace de recherche il n'existe aucune solution qui lui soit meilleure tout en respectant les contraintes. Cette solution est appelée optimum global. Par définition, $x^* \in \mathscr{R}^n$ représente un minimum global (cas mono-objectif), si

$$\forall x \in \mathscr{R}^n \ on \ a \ que \ f(x^*) \leq f(x)$$

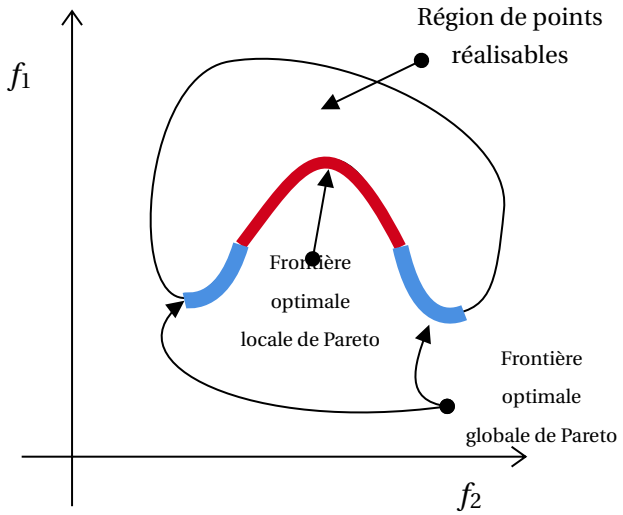


FIGURE 1.5 – Exemple : Ensemble de solutions localement et globalement optimales dans l'espace des objectifs

Cependant, l'optimum global est aussi une solution locale (exemple la Figure 1.5).

Par conséquent, des outils d'optimisation efficaces sont nécessaires pour résoudre des problèmes non linéaires et multi-objectifs et rechercher l'espace de solution multidimensionnel. Les algorithmes d'optimisation bio-inspirés tels que les algorithmes génétiques (GA) et l'optimisation des essaims de particules, sont des outils prometteurs et de plus en plus appliqués dans la conception basée sur des modèles d'idéotypes (Quilot-Turion et al. 2012 ; Semenov et al. 2013) et dans l'optimisation des stratégies de gestion (Soundharajan et al. 2009 ; Grechi et al. 2012). Ces algorithmes sont efficaces pour résoudre des problèmes complexes, surmonter les difficultés rencontrées par les outils d'optimisation traditionnels et ils sont largement utilisés pour résoudre des problèmes académiques et industriels.

1.4.3 Méthodes de sélection des modèles : AIC

Parfois plusieurs formulations d'un même modèle sont possibles. Des méthodes existent pour les comparer par rapport à la qualité de leur ajustement aux données expérimentales, suivant un principe de parcimonie (Burnham et al. 2002). Parmi les indicateurs les plus courants, l'Aikaike Information Criterium (AIC) permet de classifier plusieurs modèles en se basant sur leur vraisemblance aux données expérimentales, avec une pénalité pour les modèles ayant le plus de paramètres libres (à estimer).

L'AIC est défini comme

$$AIC(\mathscr{M}(p)) = -2\,\ell(\mathscr{M}(p)) + 2n_p \tag{1.10}$$

où n_p est le nombre de paramètres à estimer p et $\ell(\mathscr{M}(p))$ le logarithme de la vraisemblance du modèle.

Lorsque la taille de l'échantillon est petite par rapport au nombre de paramètres, il est préférable d'utiliser la version corrigée de AIC noté par AIC_C :

$$AIC_C(\mathscr{M}(p)) = AIC(\mathscr{M}(p)) + \frac{2n_p(n_p+1)}{N-n_p-1} \tag{1.11}$$

avec N la taille de l'échantillon. Lorsque l'on compare différents modèles, on retient celui pour lequel ces critères sont minimaux.

1.5 L'intégration du contrôle génétique dans les modèles dynamiques

Prédire les relations génotype-phénotype dans des environnements contrastés est un défi majeur pour la biologie et la sélection végétale. Des modèles écophysiologiques basés sur des processus qui simulent les fonctions physiologiques de la culture / plante / organe en tenant compte de divers facteurs affectant la croissance et le développement ont été développés pour prédire le rendement ou la qualité des cultures dans des environnements fluctuants. Cependant, ils sont généralement calibrés pour un seul génotype, ce qui limite la portée de leur utilisation. L'intégration du contrôle génétique dans de tels modèles est devenue une priorité (Boote et al. 2001 ; Hoogenboom et al. 2004).

A ce jour, peu de modèles écophysiologiques incluent des contrôles génétiques. La plupart d'entre eux utilisent la modélisation basée sur la recherche de QTL de paramètres génotype-dépendant du modèle. Dans ce cas, les génotypes sont définis par un jeu de paramètres dont les valeurs dépendent de leur combinaison allélique aux QTL contrôlant ces paramètres. Des résultats prometteurs ont été obtenus avec des modèles assez simples décrivant l'élongation des feuilles (Reymond et al. 2003), le développement d'une céréale (Technow et al. 2015 ; Yin et al. 2000), la croissance

précoce de *Medicago truncatula* (Brunel et al. 2009), l'absorption d'azote et la croissance et l'architecture des racines du blé (Laperche et al. 2006) et la croissance et la douceur des fruits de la pêche (Quilot et al. 2005).

L'analyse génétique, telle que la cartographie de QTL, des paramètres estimés a le potentiel de révéler l'architecture génétique qui sous-tend les processus biologiques. Dans ces analyses, deux approches peuvent être considérées. La première est l'approche indépendante ou 'two-steps' et la deuxième l'approche jointe ou 'one-step'.

1.5.1 L'approche indépendante ou 'two-step'

Dans cette approche, les paramètres génétiques (paramètres dépendant du génotypes) d'un modèle sont d'abord estimés pour chaque génotype (voir Section 1.4.2), puis les paramètres estimés sont soumis à une cartographie de QTL (voir Section 1.2.2.2). Cette approche profite de la disponibilité d'algorithmes d'optimisation efficaces. Cependant, elle présente l'inconvénient de sa mauvaise prise en compte de l'incertitude dans l'optimisation des paramètres du modèle dans la construction des modèles de cartographie de QTL parce que ces processus sont effectués séparément.

1.5.2 Les approches jointes ou 'one-step'

Dans cette approche, l'estimation des paramètres pour tous les génotypes et les analyses génétiques sur les paramètres sont effectuées simultanément en utilisant un modèle statistique hiérarchique.

L'approche de cartographie fonctionnelle "Functional mapping" (Wu et al. 2006; Ma et al. 2002) fait la synthèse entre la modélisation du phénotype et l'analyse QTL. La cartographie fonctionnelle intègre les principes biologiques des traits dans un modèle mixte fini, permettant l'interprétation biologique des QTL détectés.

Cette approche présente plusieurs avantages, comme la prise en compte de l'incertitude sur les paramètres du modèle dans l'analyse QTL. En plus, elle offre une stabilité en modélisant le développement des caractères et les autocorrélations entre différents points dans le temps. Cela permet d'améliorer la puissance statistique pour détecter les QTL significatifs (Wei et al. 2018; Wu et al. 2006). Malgré ces propriétés prometteuses, certains auteurs ont identifié des inconvénients dans l'approche de cartographie fonctionnelle. D'abord, la combinaison d'un modèle complexe fondé sur les processus et de la théorie de la cartographie des QTL peut s'avérer impossible. De plus, les valeurs attendues de QTL à tous les pas de temps (pour les modèles cinétiques) pour différents génotypes et tous les éléments de la matrice de covariance doivent être estimés, ce qui entraîne des difficultés de calcul considérables (Kwak et al. 2014).

Onogi et al. (2016a) et Onogi (2020) proposent une approche qui intègre le modèle dans un cadre bayésien. Cette méthode permet d'estimer simultanément les

paramètres du modèle et l'effet des marqueurs du génome sur les paramètres. La méthode d'intégration du modèle tend à fournir des prédictions plus précises que les méthodes classiques indépendantes. En revanche, cette méthode est appliquée sur une architecture génétique relativement simple et les modèles utilisés dans ces études sont également peu complexes, composés de seulement quelques fonctions. De façon comparable, Technow et al. (2015) ont proposé une approche statistique en utilisant une technique bayésienne approximative pour combiner la cartographie de QTL avec le modèle de culture afin de prédire le rendement du maïs. En utilisant des marqueurs à l'échelle du génome, les auteurs ont défini les distributions préalables de quatre caractères mesurables qui sont les entrées d'un modèle de croissance du maïs et ont prédit le rendement du grain.

1.5.3 La conception d'idéotypes assistée par modèles

Les modèles intégrés écophy-génétiques peuvent être utilisés pour simuler la variation du génotype (G) dans différentes conditions Environnementales (E), de Pratiques culturales (P) et finalement capturer les interactions entre eux (c'est-à-dire les interactions GxExP). Ils peuvent également être utilisés pour concevoir des « idéotypes », c'est-à-dire des plantes réelles ou virtuelles exprimant un phénotype idéal adapté à un environnement biophysique particulier et à une conduite de culture donnée (Tardieu 2003). Des tentatives d'utilisation des modèles intégrés pour optimiser les génotypes ont été conduites ces dernières années (Letort et al. 2008) mais ont révélé des limites importantes, notamment liées au réalisme des solutions proposées (Quilot-Turion et al. 2016) et aux antagonismes entre certains traits (Da Silva et al. 2014). La Figure 1.6 illustre la démarche d'une conception assistée par modèle d'idéotypes variétaux.

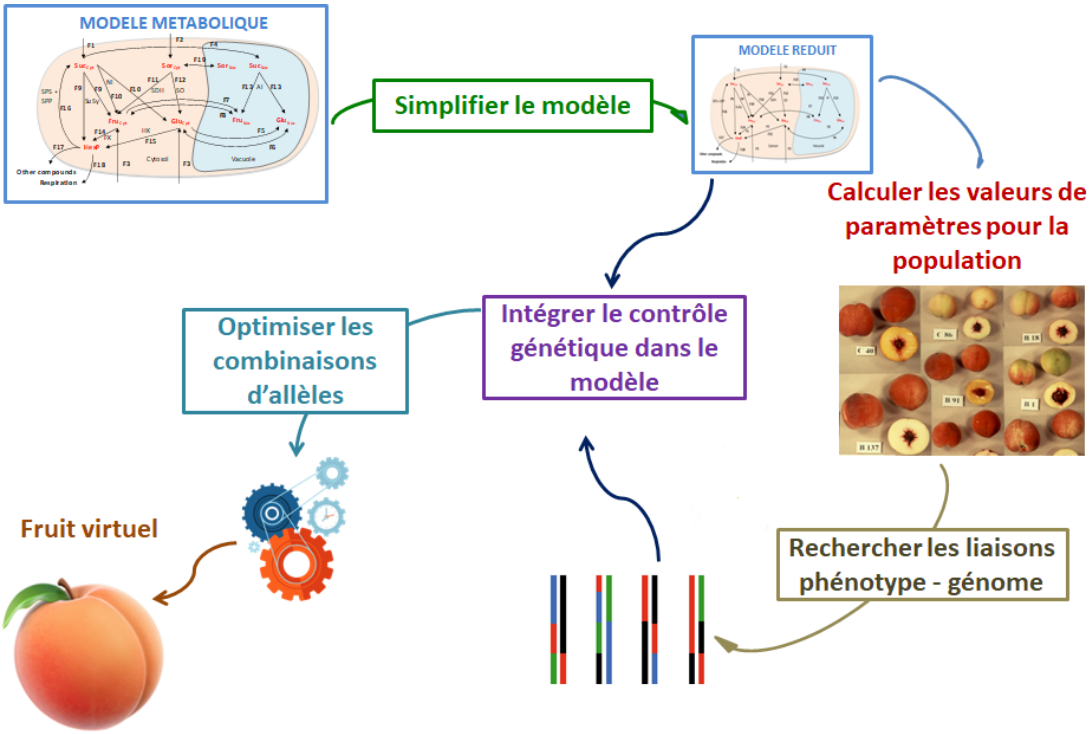


FIGURE 1.6 – Représentation graphique de la méthodologie de conception d'idéotypes

1.6 Présentation de la thèse

1.6.1 Objectifs et démarche

L'objectif de cette thèse est d'intégrer le contrôle génétique du métabolisme des sucres au modèle cinétique développé par Desnoues et al. (2018) et présenté ci-après. Ce modèle présentait deux inconvénients majeurs : i) un nombre de paramètres à calibrer important par rapport aux données disponibles et ii) un temps d'intégration long du fait de la non-linéarité et de fonctions d'entrée dépendant du temps. Ensemble, ces problèmes entravent l'utilisation du modèle pour un large panel de génotypes pour lesquels peu de données sont disponibles. Aussi est-il nécessaire, dans une première étape, de simplifier le modèle de façon à l'adapter à la spécificité des études génétiques. Ce modèle réduit doit être adapté à l'ensemble de la diversité génétique attendue, il doit conserver la structure du réseau et l'identité des variables de façon à faciliter leur interprétation biologique.

L'approche proposée pour résoudre ce problème est basée sur la combinaison et l'évaluation systématique de différentes méthodes de réduction. Ainsi, nous avons combiné l'analyse de sensibilité multivariée, la simplification structurelle et les approches basées sur l'échelle de temps pour simplifier le nombre et la structure des équations différentielles ordinaires du modèle. Ces travaux sont présentés dans le

Chapitre 2 de ce manuscrit et ont fait l'objet d'un article publié dans le journal 'Mathematical Biosciences' (Kanso et al. 2020).

La seconde étape de la démarche vise à estimer les paramètres du modèle réduit, issu de l'étape précédente, pour une population de 106 génotypes de pêcher. Les modèles cinétiques souffrent généralement de problèmes numériques et d'identifiabilité qui entravent leur application dans le contexte des études génétiques. Aussi, différentes méthodes ont été mobilisées pour surmonter cette limite et sont présentées dans le Chapitre 3.

Enfin, la dernière étape réalisée dans cette thèse consiste à explorer l'architecture du contrôle génétique (nombre de QTL et effets) des paramètres génotype-dépendant du modèle réduit.

1.6.2 Le modèle cinétique du métabolisme des sucres chez la pêche

Le modèle développé par Desnoues et al. (2018) est utilisé dans le cadre de notre étude. Dans ce modèle, le fruit est considéré comme une cellule ayant deux compartiments intracellulaires, le cytosol et la vacuole, dont la proportion a été établie selon les résultats des analyses cytologiques.

Le carbone entre dans le fruit par la sève de la plante où il est ensuite métabolisé par un réseau cellulaire complexe, comprenant des réactions enzymatiques et des mécanismes de transport entre le cytosol et la vacuole. Le modèle mathématique est construit sous la forme d'un ensemble d'équations différentielles ordinaires paramétriques (EDOs) et il comprend 10 variables : la concentration de quatre sucres principaux (saccharose, glucose, fructose et sorbitol) dans les deux compartiments (voir Figure 1.7), une variable hexoses-phosphates, comprenant glucose-1-phosphate, glucose-6-phosphate, fructose-6-phosphate et UDP-glucose et une variable pour les autres composants.

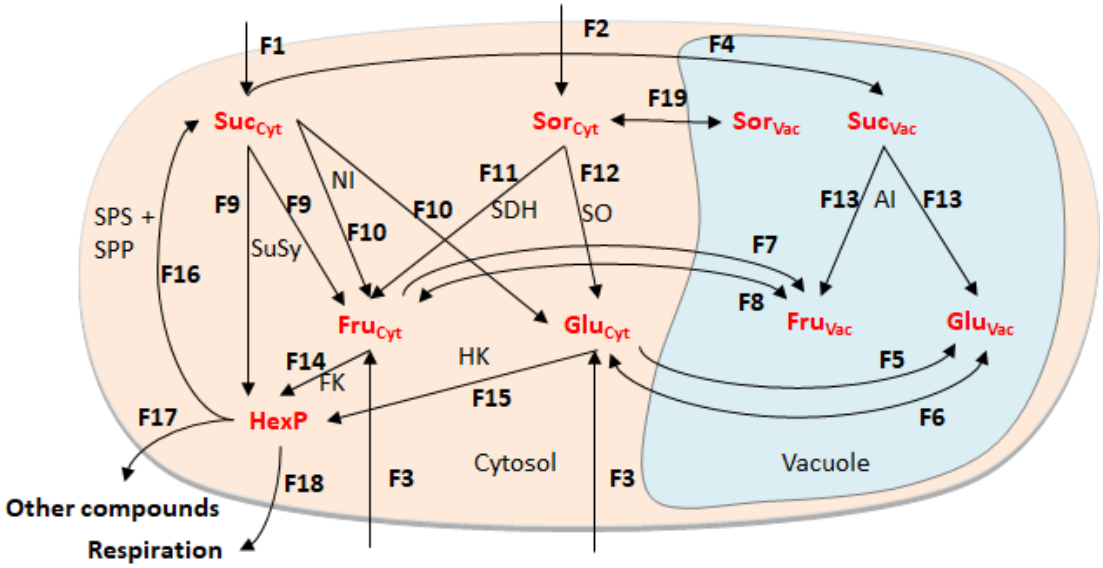


FIGURE 1.7 – Réseau schématique du modèle d'accumulation des sucres dans la pêche de Desnoues et al. (2018)

Le réseau métabolique inclue 19 réactions (voir Figure 1.7). Le carbone contenu dans la sève de la plante pénètre dans le fruit sous forme de saccharose et de sorbitol. Sous l'action de l'invertase, le saccharose est partiellement transformé en glucose et fructose dans l'apoplasme. Ensuite les 4 sucres entrent dans le cytosol via les flux F_1 (saccharose), F_2 (sorbitol) et F_3 (glucose et fructose). A l'intérieur du fruit, les 4 sucres sont métabolisés dans le cytosol et peuvent être transportés dans la vacuole par l'action de transporteurs actifs (flux F_4, F_5, F_7 F_{19}) ou passifs (flux F_6, F_8). Dans le cytosol, la sucrose synthase (SuSy), la sorbitol déshydrogénase (SDH), l'invertase neutre (NI) et la sorbitol oxydase (SO) permettent la transformation du saccharose et sorbitol en fructose, glucose, qui sont ensuite hydrolysés via la fructokinase (FK) et l'hexokinase (HK) en hexoses phosphates. Les hexoses phosphates ainsi produits peuvent être utilisés pour la re-synthèse du sucrose (F_{16}) et d'autres composés structurels ou consommés par la respiration.

Les réactions enzymatiques ont été représentées par une équation de Michaelis-Menten (MM) irréversible (Eq. 1.4). Dans le cas de l'acide invertase (AI), la régulation allostérique par son produit a été prise en compte et décrite comme une inhibition compétitive :

$$F_{13}(x_6, x_8, x_9) = \frac{v_1}{(1 + \frac{x_8(t)+x_9(t)}{p_2})p_{23} + x_6(t)} x_6(t) \tag{1.12}$$

où x_8, x_9 et x_6 sont les concentrations respectives du fructose, glucose et saccharose dans la vacuole, p_2 la constante d'inhibition, V_1 la capacité enzymatique et p_{23} l'affinité de l'acide invertase (AI).

Les échanges entre le cytosol et la vacuole sont possibles selon deux mécanismes principaux : le transport actif ou passif. Le transport actif est généré par des transpor-

teurs spécifiques et permet le stockage des sucres dans les vacuoles indépendemment de leur gradient de concentration. Comme pour les réactions enzymatiques, ce type de transport a été représenté par une équation irréversible de Michaelis-Menten (Eq. 1.4). Le transport passif en revanche facilite l'écoulement des molécules suivant leur gradient de concentration grâce à l'existence de canaux protéiques spécifiques. Ce mode de transport a été représenté par des fonctions linéaires simples du gradient de concentration. Quelque soit le mode de transport, la densité des transporteurs et des canaux protéiques par unité de surface a été supposée constante, de sorte que le transport augmente proportionnellement à la surface du tonoplaste pendant le développement du fruit. Par exemple, dans le cas du transport passif du glucose (flux $F6$), l'equation considerée est la suivante :

$$F_6(x_4, x_9, S) = p_{10}\,(x_9(t) - x_4(t))\,S(t)$$

où p_{10} est le taux de transport par unité de surface, x_4 et x_9 sont les concentrations en glucose respectivement dans le cytosol et dans la vacuole et S(t) est la surface tonoplastique, supposée sphérique.

Le modèle final comprend trente paramètres. Les capacités enzymatiques des enzymes (V_{max}) ont été mesurées expérimentalement dans une étude menée par Desnoues et al. (2014). En accord avec les données expérimentales, certaines capacités enzymatiques varient avec le temps ou dépendent du phénotype en fructose de l'individu considéré, alors que d'autres sont supposées constantes. La plupart des valeurs de K_m ont été fixées d'après la littérature sur la pêche ou les fruits, à l'exception des K_m de l'hexokinase (HK), de la fructokinase (FK) et de la saccharose synthase (Susy) qui ont été estimées numériquement. Les six paramètres liés à la cinétique du transport vacuolaire ont été supposés constants et ont été estimés numériquement. Dans l'ensemble, 14 paramètres ont été estimés lors de la calibration du modèle original. Les tableaux suivants décrivent les équations du modèle original.

TABLE 1.1 – Taux des réactions du modèle

Processus	Équations
Flux d'entrée	$I(t) = \sigma_f \frac{dDW}{dt} + R(t) = (\sigma_f + q_g)\frac{dDW}{dt} + q_m DW Q_{10}^{\frac{(T-20)}{10}}$ $R(t) = q_m DW Q_{10}^{\frac{(T-20)}{10}} + q_g \frac{dDW}{dt}$ $DW = DW(t_0) + w_1(1 - e^{-w_2 t}) + \frac{w_3}{1+e^{-w_4(t-w_5)}}$ $F_1(I) = \lambda \lambda_{suc}(t) I(t)$ $\lambda_{suc}(t) = \frac{p_1 t}{t_{max}}$ $F_2(I) = (1-\lambda) I(t)$ $F_3(I) = \frac{\lambda}{2}(1 - \lambda_{suc}(t)) I(t)$
Métabolisme	$F_9(v_2, x_1) = \frac{v_2(t)}{p_5 + x_1(t)} x_1(t)$ $F_{10}(x_1) = \frac{v_3}{p_{21} + x_1(t)} x_1(t)$ $F_{11}(v_4, x_2) = \frac{v_4(t)}{p_{22} + x_2(t)} x_2(t)$ $F_{12}(v_5, x_2) = \frac{v_5(t)}{p_{13} + x_2(t)} x_2(t)$ $F_{13}(x_6, x_8, x_9) = \frac{v_1}{(1 + \frac{x_8(t) + x_9(t)}{p_2}) p_{23} + x_6(t)} x_6(t)$ $F_{14}(v_6, x_3) = \frac{v_6(t)}{p_3 + x_3(t)} x_3(t)$ $F_{15}(v_7, x_4) = \frac{v_7(t)}{p_4 + x_4(t)} x_4(t)$ $F_{16}(x_5) = p_7 x_5(t)$ $F_{17}(x_5) = p_6 x_5(t)$ $F_{18}(R) = R(t)$
Transport	$F_4(S, x_1) = p_8 x_1(t) S(t)$ $F_5(S, x_3, x_4) = \frac{p_{11}}{p_{19} + x_3(t) + x_4(t)} x_4(t) S(t)$ $F_6(S, x_4, x_9) = (x_9(t) - x_4(t)) p_{10} S(t)$ $F_7(S, x_3, x_4) = \frac{p_{12}}{p_{20} + x_3(t) + x_4(t)} x_3(t) S(t)$ $F_8(S, x_3, x_8) = (x_8(t) - x_3(t)) p_9 S(t)$ $F_{19}(S, x_2, x_7) = p_{14}(x_7(t) - x_2(t)) S(t)$ $S(t) = (4\pi)^{\frac{1}{3}} (V_2)^{\frac{2}{3}}$ avec V_2 est le volume de vacuole

TABLE 1.2 – Équations du modèle original

Équations du modèle
$\frac{dx_1}{dt} = F_1 + F_{16} - F_{10} - F_4 - F_9$
$\frac{dx_2}{dt} = F_2 - F_{11} - F_{12} + F_{19}$
$\frac{dx_3}{dt} = F_3 + F_8 + \frac{1}{2}F_9 + \frac{1}{2}F_{10} + F_{11} - F_7 - F_{14}$
$\frac{dx_4}{dt} = F_3 + F_6 + \frac{1}{2}F_{10} + F_{12} - F_5 - F_{15}$
$\frac{dx_5}{dt} = \frac{1}{2}F_9 + F_{14} + F_{15} - F_{17} - F_{16} - F_{18}$
$\frac{dx_6}{dt} = F_4 - F_{13}$
$\frac{dx_7}{dt} = -F_{19}$
$\frac{dx_8}{dt} = F_7 + \frac{1}{2}F_{13} - F_8$
$\frac{dx_9}{dt} = F_5 + \frac{1}{2}F_{13} - F_6$
$\frac{dx_{10}}{dt} = F_{17}$

TABLE 1.3 – Variables du modèle et leur localisation

x_1	Saccharose	Cytosol
x_2	Sorbitol	Cytosol
x_3	Fructose	Cytosol
x_4	Glucose	Cytosol
x_5	Hexose phosphate	Cytosol
x_6	Saccharose	Vacuole
x_7	Sorbitol	Vacuole
x_8	Fructose	Vacuole
x_9	Glucose	Vacuole
x_{10}	Autres composés	Cytosol

TABLE 1.4 – Description des paramètres du modèle original (FRU : Individu ayant un type fructose standard, SSFRU : Individu ayant un type faible fructose).

p_i	Paramètre	Description	Référence	Valeur	Unité
p_1	λ_{Suc}	Proportion de saccharose hydrolysée dans l'apoplasme	Estimé	0-1	
p_8	$TactifSuc$	Coefficient de transport du saccharose (importation active) du cytosol à la vacuole	Estimé	0-400	$mggFW^{-1}day^{-1}$
p_{10}	$TpassifGlu$	Coefficient de transport passif du glucose entre le cytosol et la vacuole et dans le sens inverse	Estimé	0-150	$mggFW^{-1}day^{-1}$
p_9	$TpassifFru$	Coefficient de transport passif du fructose entre le cytosol et la vacuole et dans la direction opposée	Estimé	0-150	$mggFW^{-1}day^{-1}$
$v_2(t)$	V_{susy}	Activité de saccharose synthase (susy) pour transférer le saccharose en fructose et hexoses phosphate	Desnoues et al. (2014)	$6.5–3.8e^{-2}t+2.7e^{-3}t^2–1.3e^{-5}t^3$	$mggFW^{-1}day^{-1}$
p_5	K_{susy}	Affinité de saccharose synthase (susy)	Estimé	0-200	$mggFW^{-1}$
v_3	V_{ni}	Activité de neutral invertase (ni) pour transférer le saccharose en glucose et fructose	Desnoues et al. (2014)	FRU=10 et SSFRU=6.8	$mggFW^{-1}day^{-1}$
p_{21}	K_{ni}	Affinité de neutral invertase (ni)	Vorster et al. (1998)	3.42	$mggFW^{-1}$
v_4	V_{sdh}	Activité de sorbitol déshydrogénase pour transférer le sorbitol en fructose	Desnoues et al. (2014)	$2+0.3t–3.1e^{-3}t^2+8.7e^{-6}t^3$	$mggFW^{-1}day^{-1}$
p_{22}	K_{sdh}	Affinité de sorbitol déshydrogénase (sdh)	Oura et al. (2000)	17.54	$mggFW^{-1}$
v_5	V_{so}	Activité de sorbitol oxydase (so) pour transférer le sorbitol en glucose	Desnoues et al. (2014)	$4.9+0.2t–3.1e^{-3}t^2–1.3e^{-5}t^3$	$mggFW^{-1}day^{-1}$
p_{13}	K_{so}	Affinité de sorbitol oxydase (so)	Estimé	0-300	$mggFW^{-1}$
v_1	V_{ai}	Activité d'invertase acide (ai) pour transférer le saccharose en glucose et fructose	Desnoues et al. (2014)	6.5	$mggFW^{-1}day^{-1}$
p_{23}	K_{ai}	Affinité d'invertase acide (ai)	Moriguchi et al. (1991)	1.43	$mggFW^{-1}$
p_2	Ki_{AI}	Affinité d'inhibitrice de l'invertase acide	Estimé	0-10	$mggFW^{-1}$
v_6	V_{fk}	Activité de fructokinase (fk) pour transférer le fructose en phosphate d'hexose	Desnoues et al. (2014)	FRU : $75–1.9t+2e^{-2}t^2–6.4e^{-5}t^3$ SSFRU : $61–1.4t+1.5e^{-2}t^2–4.7e^{-5}t^3$	$mggFW^{-1}day^{-1}$
p_3	K_{fk}	Affinité de fructokinase (fk)	Estimé	0-30	$mggFW^{-1}$
v_7	$Vhk(t)$	Activité de l'hexokinase (hk) pour transférer le glucose au phosphate d'hexose	Desnoues et al. (2014)	FRU : $83.7–2.2t+2.5e^{-2}t^2–8.1e^{-5}t^3$ SSFRU : $88.7–2.4t+2.7e^{-2}t^2–8.7e^{-5}t^3$	$mggFW^{-1}day^{-1}$
p_4	Khk	Affinité avec l'hexokinase	Estimé	0-300	$mggFW^{-1}$
p_7	$ReSyntSuc$	Coefficient de la fonction de transfert entre le phosphate d'hexose et le saccharose	Estimé	0-300	day^{-1}
p_6	$OthComp$	Coefficient de la fonction de transfert entre le phosphate d'hexose et d'autres composés	Estimé	450-1500	day^{-1}
p_{14}	$TpassifSor$	Coefficient de transport passif du sorbitol entre le cytosol et la vacuole	Estimé	0-150	$mggFW^{-1}day^{-1}$
σ_f	$PropCdw$	Concentration de carbone dans le mésocarpe	Génard et al. (2010)	0.44	$gCgDW^{-1}$
p_{17}	q_g	coefficient de respiration de croissance	Génard et al. (2010)	0.084	$gCgDW^{-1}$
p_{18}	q_m	Coefficient de maintien respiratoire	Génard et al. (2010)	2.76e-4	$gCgDW^{-1}day^{-1}$
p_{16}	Q_{10}	Rapport de température de la respiration d'entretien	Génard et al. (2010)	1.9	
p_{15}	λ	Proportion de saccharose dans sève	Desnoues et al. (2018)	0.65	
p_{11}	$VmtactifFru$	Importation active de fructose (activité)	Estimé	0-150	$mggFW^{-1}day^{-1}$
p_{12}	$VmtactifGlu$	Importation de glucose actif (activité)	Estimé	0-150	$mggFW^{-1}day^{-1}$
p_{19}	$KmtactifGlu$	Importation de glucose actif (affinité)	Shiratake et al. (1997)	0.054	$mggFW^{-1}$
p_{20}	$KmtactifFru$	Importation active de fructose (affinité)	Shiratake et al. (1997)	0.288	$mggFW^{-1}$

1.6.3 Le matériel végétal

La population utilisée dans cette thèse est issue d'un croisement interspécifique entre un pêcher sauvage, *P. davidiana* (clone P1908) et une variété commerciale de nectarine à chair jaune Summergrand® (*P. persica*). Suite à ce croisement, un hybride SD40 présentant un bon niveau de résistance à l'oïdium a été sélectionné pour effectuer un rétrocroisement avec la variété Summergrand donnant lieu à la famille BC1. Un mélange des pollens de cette famille a servi à féconder la variété commerciale Zéphyr® (*P. persica*), une nectarine à chair blanche. C'est la population appelée BC2 (pour Back Cross 2) issue de ce dernier croisement qui a fait l'objet de cette étude.

Cette population s'apparente à un double back cross si l'on considère que les allèles proviennent soit de *P. persica*, soit de *P. davidiana*. Cependant ce n'est pas un double back cross au sens strict car les deux parents *P. persica* utilisés sont différents, bien que très proches comparés à *P. davidiana*. Dans la population BC2, tous les génotypes portent un des deux allèles de la variété Zéphyr alors que l'allèle de *P. davidiana* est présent chez environ ¼ des individus.

La population BC2 a été caractérisée pour différents critères liés à la qualité du fruit en cinétique au cours du développement du fruit (Desnoues et al. 2014). Pour la majorité des génotypes, la concentration en glucose est équivalente à la concentration en fructose à maturité. De manière intéressante, environ un quart des génotypes présentent une concentration en fructose très faible et une concentration en glucose variable, parfois importante avec un rapport fructose sur glucose faible. Nous qualifierons ces derniers génotypes comme ayant 'peu de fructose' (ou 'low-fructose-to-glucose-ratio' genotypes) en opposition avec les génotypes au ratio équilibré appelés génotypes 'standards' (ou 'standard-fructose-to-glucose-ratio' genotypes) dans le reste du manuscrit. Cette particularité fait de cette population un matériel privilégié pour l'étude du métabolisme des sucres et plus particulièrement l'effet d'une modification ponctuelle du taux de fructose sur l'ensemble du métabolisme des sucres.

1.6.4 Laboratoires d'accueil et collaborations

J'ai réalisé cette thèse conjointement au sein des unités GAFL (INRAE UR1052 Génétique et Amélioration des Fruits et Légumes, Avignon) et PSH (INRAE UR1115 Plantes et Systèmes de Culture Horticoles, Avignon).

Pendant cette thèse, deux collaborations ont été mises en place dans le but de réduire le modèle cinétique et de calibrer le modèle sur une population génétique des pêchers. La première, avec Olivier Bernard et Jean-Luc Gouzé (Inria, BIOCORE, Sophia Antipolis), a permis de réaliser une simplification du modèle cinétique. La seconde, avec Charlotte Baey (Université de Lille, Laboratoire Paul Painlevé, Villeneuve d'Ascq), visait à modéliser les variabilités intra et inter-individuelle et à calibrer le modèle pour 106 génotypes.

1.6.5 Organisation du manuscrit

La suite du manuscrit est organisée en quatre chapitres :

- le chapitre 2 correspond à une publication, accompagnée de résultats complémentaires, décrivant une méthodologie pour réduire un modèle de métabolisme des sucres dans la pêche afin de l'appliquer à une population génétique.
- le chapitre 3 correspond à une comparaison de plusieurs approches de calibration du modèle cinétique réduit pour une population génétique de pêchers. La partie principale de ce chapitre correspond à un article en cours de finalisation.
- le chapitre 4 correspond à la détection de QTL relatifs aux paramètres génétiques contrôlant le modèle réduit ainsi que l'influence des estimations sur ce dernier. Une publication est également envisagée pour valoriser ces résultats.
- le chapitre 5 est une discussion générale qui restitue les principaux résultats de la thèse au regard de la bibliographie et propose des perspectives de ce travail à court et plus long termes.

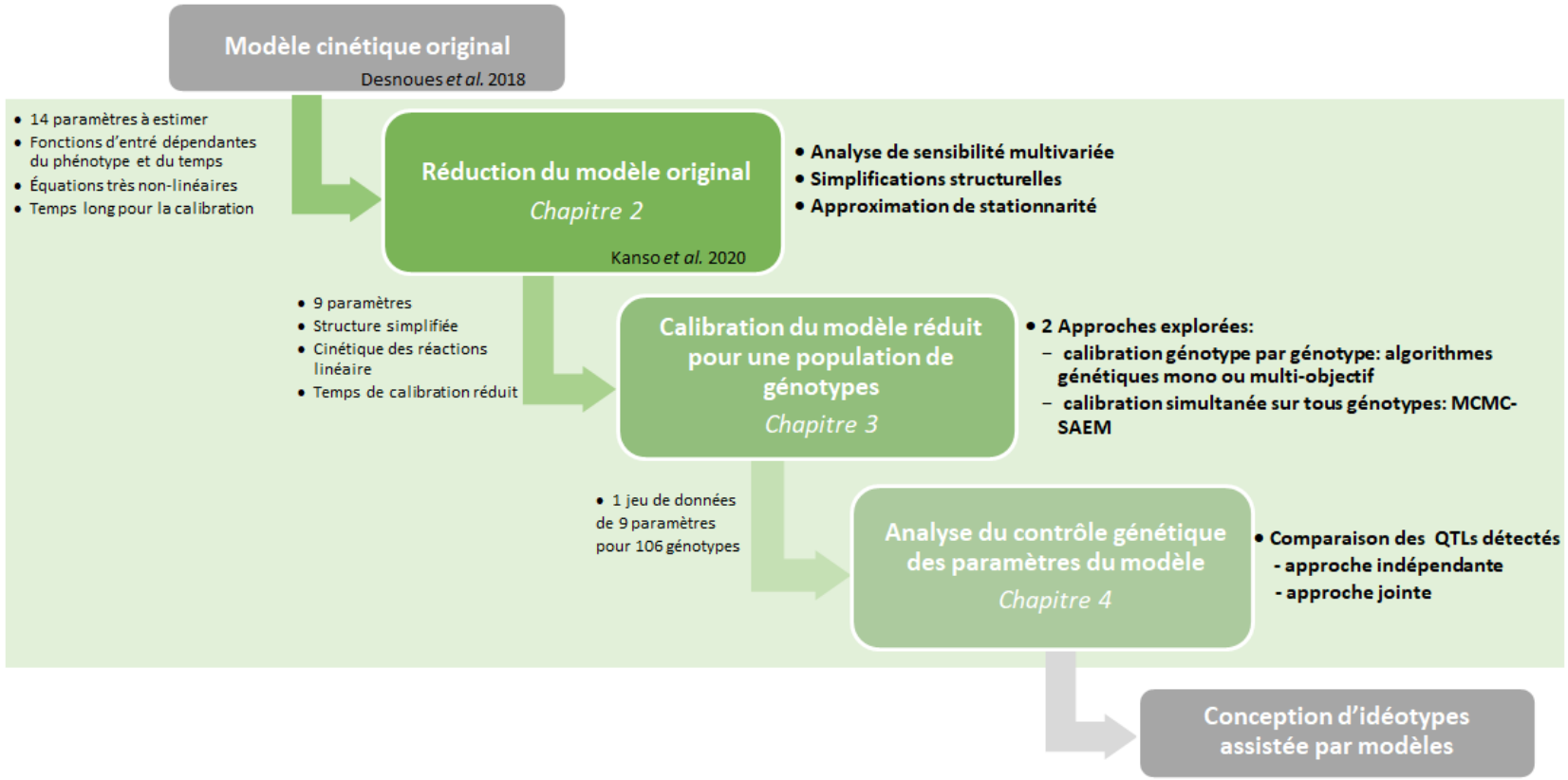


FIGURE 1.8 – Schéma représentant les différents axes de recherche suivis dans cette thèse

2 Réduction du modèle de métabolisme des sucres de la pêche

Sommaire

L'OBJECTIF de ce chapitre est de construire un modèle réduit, robuste et unique qui puisse être utilisé à l'échelle d'une population génétique de 106 génotypes. Le modèle développé par Desnoues et al. (2018) a été utilisé dans cette thèse. Il permet la prédiction des quatre sucres majoritaires : saccharose, sorbitol, fructose et glucose pour 10 genotypes seulement. En effet, ce modèle présente plusieurs inconvénients majeurs qui empêchent son utilisation pour l'ensemble de la population génétique de 106 génotypes, pour lesquels peu de données sont disponibles (six points de données ou moins par sucre) : le nombre de paramètres à estimer, son temps d'intégration qui peut être coûteux en raison de la non-linéarité et des fonctions d'entrée dépendantes du temps et de la qualité des observations. C'est un problème majeur limitant son utilisation pour l'analyse du contrôle génétique du métabolisme des sucres chez la pêche qui nécessite un modèle unique et fiable pour l'appliquer à une grande population génétique.

C'est pourquoi, une première méthode a été testée qui consistait en un changement de coordonnées. Le résultat obtenu n'a pas été retenu car il impliquait une perte de signification biologique des variables et des paramètres du modèle. Nous avons donc ensuite tenté une autre stratégie pour simplifier le modèle cinétique. L'ensemble des résultats est présenté sous la forme d'un article paru dans 'Mathematical Bioscience'. Cette stratégie est basée sur la combinaison et l'évaluation systématique de différentes méthodes de réduction à l'aide de plusieurs critères d'évaluation. Ainsi, nous avons combiné l'analyse de sensibilité, la simplification structurelle et les approches basées sur l'échelle de temps pour simplifier le nombre et la structure des équations différentielles ordinaires du modèle.

2.1 Réduction de l'ensemble des paramètres

Dans cette section, nous appliquons la méthode développée par Lemaire et al. (2012) et Boulier et al. (2009) pour la réduction du nombre de paramètres d'un système d'équations différentielles ordinaires . Cette méthode est basée sur les symétries de Lie d'un système d'équations différentielles (Stephani 1989; Bluman et al. 2013). L'objectif de la méthode est d'éliminer certains paramètres en construisant un système réduit à l'aide d'une transformation des coordonnées. Le système réduit est équivalent à l'ancien en dynamique avec moins de paramètres, facilitant ainsi l'étude et la calibration du modèle.

La méthode reçoit en entrée le système d'équations et la liste des paramètres du modèle. L'algorithme prend chaque paramètre dans la liste et essaie de l'éliminer avant d'envisager le suivant. Il peut arriver que certains de ces paramètres ne puissent pas être éliminés à cause de la structure du système. La sortie de l'algorithme est composée de trois objets : le système réduit d'Équations Différentielles ordinaires, le changement

de coordonnées qui donne les relations exactes entre les variables d'origine et les variable du système réduit et la liste des paramètres qui ont été éliminés.

Suite à l'application de l'algorithme à notre système d'équations, 5 paramètres (p_1, p_5, p_8, p_2 et p_{13}, voir Tableau 1.4) ont pu être éliminés, réduisant ainsi le nombre totale de 14 à 9. En revanche, l'élimination des paramètres est fait par un changement de coordonnées qui résulte sur un changement de la signification biologique des paramètres. Aussi le modèle obtenu perd tout intérêt pour une étude génétique fonctionnelle. Par exemple, le paramètre p_3 qui désigne l'affinité de fructokinase a été transformé en $\frac{p_3}{p_2}$, où p_2 est l'affinité d'inibitrice de l'invertase acide. Par contre, il y a aucune liaison biologique entre ces deux paramètres. En conséquences, ce résultat n'a pas été retenu pour la suite de la thèse.

2.2 Reducing a model of sugar metabolism in peach to catch different patterns among genotypes

Abstract

Several studies have been conducted to understand the dynamic of primary metabolisms in fruit by translating them into mathematics models. An ODE kinetic model of sugar metabolism has been developed by Desnoues et al. (2018) to simulate the accumulation of different sugars during peach fruit development. Two major drawbacks of this model are (a) the number of parameters to calibrate and (b) its integration time that can be long due to non-linearity and time-dependent input functions. Together, these issues hamper the use of the model for a large panel of genotypes, for which few data are available. In this paper, we present a model reduction scheme that explicitly addresses the specificity of genetic studies in that : i) it yields a reduced model that is adapted to the whole expected genetic diversity ii) it maintains network structure and variable identity, in order to facilitate biological interpretation. The proposed approach is based on the combination and the systematic evaluation of different reduction methods. Thus, we combined multivariate sensitivity analysis, structural simplification and timescale-based approaches to simplify the number and the structure of ordinary differential equations of the model. The original and reduced models were compared based on three criteria, namely the corrected Aikake Information Criterion (AIC_C), the calibration time and the expected error of the reduced model over a progeny of virtual genotypes. The resulting reduced model not only reproduces the predictions of the original one but presents many advantages including a reduced number of parameters to be estimated and shorter calibration time, opening new promising perspectives for genetic studies and virtual breeding. The validity of the reduced model was further evaluated by calibration on 30 additional genotypes of an inter-specific peach progeny for which few data were available.

Keyword

model reduction, sensitivity analysis, structural simplification, quasi-steady-state, peach fruit, kinetic model, model calibration, gene-to-phenotype.

2.2.1 Introduction

Plants are sessile organisms endowed with the capacity to alter their development, physiology, and morphology depending on the context. Plant phenotype is the result of the interaction between the environment, cultural practices and plant's genetic background (genotype). In the context of agronomy, increasing efforts have been made to select varieties that better meet consumers' expectations. Today it is clear that future breeding should account for complex plant phenotypes, responding to a

large panel of criteria, including increased yield, abiotic and biotic stress tolerance, and quality of food products.

Genotype-phenotype models have been considered as the tools of the future to design new genotypes since they can help to test the performance of new genotypes (G) under different Environments (E) x Management (M) conditions. The challenge is to build ecophysiological models that integrate genetic information associated to specific processes (traits). In general, genotypes are defined by a set of parameters, which depends on gene expression or allelic combination, depending on the genetic complexity of the considered trait as well as the available information (White et al. 2003). Genetic-improved ecophysiological models can then be used to capture GxExM interactions. They can also be used to design "ideotypes" i.e. real or virtual plant cultivars expressing an ideal phenotype adapted to a particular biophysical environment, crop management, and end-use (Letort et al. 2008; Tardieu 2003). For this, it is necessary to combine the genetic-improved ecophysiological model with a multi-objective optimization algorithm to identify the best genotypes for specific conditions (Quilot-Turion et al. 2016).

Construction of gene-to-phenotype models is challenging. First, the approach requires that a sole and unique model can reproduce the behavior of all genotypes, in multiple environments, the diversity observed being supported by different sets of parameters. Second, calibration of the models for a large number of genotypes is generally difficult, due to a large number of parameters (typically from 50 to 200 in whole-plant ecophysiological models) along with a restricted number of observations (Martre et al. 2015; Bertin et al. 2010). Due to the model complexity and non-lineairities, evolutionary and bio-inspired algorithms are increasingly used both for parameter estimation and ideotype design. These methods can explore high-dimensional parameter space efficiently but they rely on a large number of model evaluations, that can rapidly increase the computational time required to find a solution. Third, the genetic architecture of complex traits can be very complex, due to epistatic and pleiotropic effects. In this sense, the presence of biologically-meaningful parameters can considerably help the interpretation of the resulting genetic architecture, facilitating the breeding process. Ideally, most the model is close to omics data, the easier the linkage between the parameters and the underlying physiological processes.

Kinetic modeling has been successfully applied to several metabolic pathways in plants (Curien et al. 2009; Nägele et al. 2014; Beauvoit et al. 2014). In this spirit, a kinetic model of sugar metabolism has been developed in (Desnoues et al. 2018) to simulate the accumulation of different sugars during peach fruit development. The model correctly accounts for annual variability and the genotypic variations observed in ten genotypes derived from a larger progeny of inter-specific peach cross. At term, the objective of the research is to integrate the genetic control of sugar metabolism in this kinetic model and develop a methodology to design ideotypes by virtual breeding. To achieve this, it is necessary to estimate accurately the values of the influential parameters of the model for the whole progeny of 106 genotypes for which few data

are available. Unfortunately, the size of the parameter space and the non-linearity of the reaction rates make the calibration of the model unreliable and time-consuming.

One way to face these weaknesses is to reduce the complexity of the model (Okino et al. 1998). Several reductions and approximation approaches exist in the literature, each one addressing a specific aspect of model complexity (Gorban et al. 2006; Snowden et al. 2017). A number of methods, such the lumping method (Wei et al. 1969; Sunnåker et al. 2011) or the classical quasi-steady-state (QSS) approaches, aim at reducing the number of variables based on chemical or time-scale considerations (Schauer et al. 1983; Heinrich et al. 1996). Methods from sensitivity analysis may help to reduce the parameter space by identifying non-influential parameters, whose values can be fixed by broad literature data (Turányi 1990; Cariboni et al. 2007; Vanuytrecht et al. 2014; Saltelli et al. 2008). Last but not least, the structure of the model itself can be simplified. Methods for model decomposition (Holme et al. 2003; Anderson et al. 2011; Sun et al. 2016) aim to separate the system into sub-networks or sub-models, that are easier to analyze and parameterize. The choice of reaction kinetics is also very important for model complexity. In this perspective, the use of simplified enzyme kinetics (Wang et al. 2007; Nikerel et al. 2009; Schmidt et al. 2008) may be useful to avoid the emergence of numerical and identifiability issues.

Different reduction methods can be combined together. In (Liebermeister et al. 2005) for instance, model decomposition is associated to variable transformation, resulting in a low-dimensional description of the "exterior" part of the system, whereas in (Sunnåker et al. 2011) time scale analysis is used to identify a cluster of fast variables to be lumped together.

In the work of Apri et al. (2012) different reduction steps (parameter removal, node removal, variable lumping) are sequentially tested following a practical scheme : at each step, if the reduced model, after parameter re-estimation, can reproduce some target outputs, the modification is selected, and rejected otherwise. From the point of view of genetic applications, a major drawback of the approach of Apri et al. (2012) is that the selection of acceptable reduction results depends on the specific target dynamics.

As a consequence, different target outputs (i.e. genotypes) can give rise to reduced models with different structures or parameters number, making their comparison difficult in the perspective of genetic studies.

The objective of this work was to provide a method to build a reduced model that is adapted to the specificity of genetic studies in that : i) it yields a reduced model that is adapted to the whole expected genetic diversity ii) it maintains network structure and variable identity, in order to facilitate the biological interpretation of the reduced model.

Similarly to the approach of Apri et al. (2012), our reduction strategy tests different methods in several *parallel* steps that, if retained, are combined together into a final reduced model (Fig. 2.1).

First, multivariate sensitivity analysis was attempted to reduce the parameter space (Lamboni et al. 2009). Second, we tried to simplify the structure of the model by re-

ducing non-linearity and time-dependent forcing, and finally, a quasi-steady-state approximation based on time-scale separation was tested to reduce the size of the system. Particular attention was devoted to the systematic evaluation of the different reduction methods. Three main criteria were used to assess the interest of the reduction : i) the corrected AIC value, evaluating the relative gain between model simplification and loss of accuracy over an experimental dataset, ii) the calibration time, as a measure of model efficiency, iii) the expected error between the original and the reduced model over a population of virtual genotypes, as a measure of the reliability of the simplification scheme.

As a case study, the proposed reduction scheme was applied to the model of sugar metabolism proposed by Desnoues et al. (2018). The resulting reduced model correctly reproduces data on the original 10 genotypes with only 9 estimated parameters (out of 14 in the original model) and a gain in calibration time over 40%. In addition, the reduced model was successfully calibrated on 30 new genotypes of the same inter-specific peach progeny, for which fewer data points were available.

The paper is organized as follows. In the next section, we briefly present the original model of sugar metabolism developed by Desnoues et al. (2018). Section 2.2.3 is devoted to the description of the individual reduction methods, whereas Sections 2.2.4 and 2.2.5 present, respectively, the datasets and the numerical methods used for the assessment of the proposed model reduction. The results of the application of our reduction scheme to the model of sugar metabolism are reported in Section 2.2.6. A general discussion on the advantages and limitations of our approach closes the paper.

2.2.2 Description of the peach sugar model

The model developed by Desnoues et al. (2018) describes the accumulation of four different sugars (sucrose, glucose, fructose, and sorbitol) in peach fruit during its development over a progeny of ten peach genotypes with contrasting sugar composition. The fruit was assumed to behave as a single big cell with two intra-cellular compartments, namely the cytosol and the vacuole. Carbon enters the fruit from the plant sap which is transformed by a metabolic network, including enzymatic reactions and transport mechanisms between the cytosol and the vacuole.

The developed dynamical model made explicit use of experimental data to describe the evolution of the sub-cellular compartment (due to fruit growth) and enzyme activities (due to fruit developmental program) over time. To this aim, measured fruit dry and fresh masses and enzyme activities were represented by genotype-specific temporal functions and provided as input to the model.

From a mathematical point of view, the model can be described as a set of parametric

ordinary differential equations :

$$\frac{dx}{dt} = f(x(t), I(t), v(t), p), \tag{2.1}$$

$$x(t_0) = x_0, \tag{2.2}$$

where t is the independent time variable in days after bloom (DAB); $x \in \mathbb{R}^{10}$ is the concentration vector of metabolites in the corresponding intra-cellular compartment and $x_0 \in \mathbb{R}^{10}$ in Eq.(2.2) is the vector of the corresponding initial values. $I \in \mathbb{R}$ is the time-dependent input of carbon from the plant and $v \in \mathbb{R}^7$ is the vector of time-dependent measured enzymatic activities; $p = (p_1, \ldots, p_{23})$ is the vector of parameters defining the rate reactions where $p_1, \ldots, p_{14}$ have to be estimated and $p_{15}, \ldots, p_{23}$ are fixed from literature data. $f(x(t), I(t), v(t), p)$ of Eq.(2.1) describes the change in compounds concentrations. Equations of the reduced and original model are introduced in Appendix 2.2.8.1.

2.2.3 Model reduction methods

In this section, we present a reduction scheme explicitly dedicated to genetic studies that combines different methods in several parallel steps as shown in (Fig. 2.1) and explained in the next subsections.

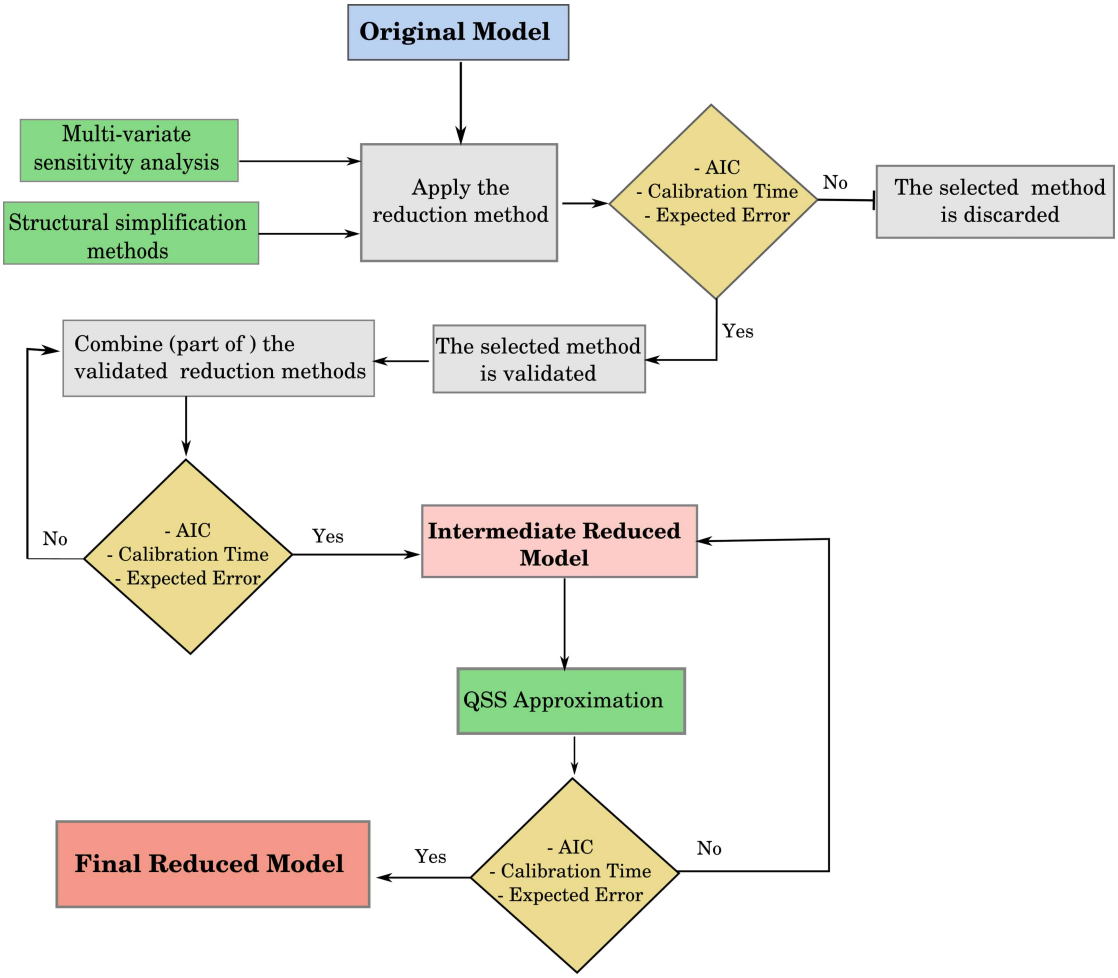


FIGURE 2.1 – Graphical representation of the proposed model reduction scheme. Yellow diamonds represent model evaluation steps by means of our 3 criteria : the corrected AIC value, calibration time and expected error over a virtual population. The tested reduction methods are indicated in green. Multivariate sensitivity analysis and three structural simplification methods are independently applied to the original model and evaluated. The validated methods are then combined into an intermediate reduced model whose performances are again submitted to evaluation. Finally, the application of a QSS approximation over the intermediate reduced model is tested to yield the final reduced model.

2.2.3.1 Multivariate sensitivity analysis

Generally, in the case of complex models, estimating parameters requires a lot of effort and is known to be a difficult and challenging task. In particular, it is tricky to determine which parameters can be fixed. The global sensitivity analysis methods allow to explore the influence of each parameter on model outputs and thus to identify the

key parameters that affect model performance and play important roles in model parameterization, calibration and optimization (Saltelli et al. 2008). Multivariate sensitivity is a method developed by Lamboni et al. (2009) that allows the application of global sensitivity analysis to models having a multivariate (eg. dynamic) output. The idea is to perform a principal components analysis on the outputs, and then compute the sensitivity indexes for each principal component. The results are summarized by the generalized sensitivity indices (GSI) that provide a unique ranking of the parameters over the whole output.

This method was applied to the 23 parameters of the original model and to the measured enzymatic activities v. Each parameter was studied at three levels, corresponding to 0.05, 0.5 and 0.95 quantiles of the previously estimated 14 parameters values (Desnoues et al. 2018) and to a variation of −20% and +20% of the fixed values for the other parameters. For time-dependent enzyme activities, the same −20% and +20% variation was applied on their average values over the whole dynamics.
In order to evaluate the impact of the genotype choice on the results of the sensitivity analysis, simulations were performed according to a factorial design, following the ANOVA model $genotypes \times (p_1 + \ldots p_{23} + v_1 + \ldots + v_7)^2$. The package "Planor" in **R** (R Development Core Team 2015) was used. The minimum resolution of the plan was fixed by using the tool MinT (Schürer et al. 2006) to test all main effects and interactions. The factorial design resulted in $10 \times 3^9 = 196\,830$ simulations.

Multivariate sensitivity analysis was performed independently on the dynamics of the four output sugars (*i.e* sucrose, glucose, fructose, and sorbitol) that compose peach fruit. In order to determine the least sensitive parameters, the whole sugar phenotype has to be taken into account, with respect to the relative proportions of each sugar. For this aim, an aggregate generalized sensitivity index ($aGSI$) was constructed for each parameter as

$$aGSI = \sum_{i=1}^{4} GSI_i \, \beta_i \tag{2.3}$$

where GSI is the generalized sensitivity indice computed for the sugar i and β_i the relative proportion of sugar i in the fruit. $\beta = (0.72, 0.13, 0.09, 0.05)$ for sucrose, glucose, fructose, and sorbitol, respectively.

2.2.3.2 Structural simplification methods

This section aims to simplify the structure of the model in terms of network and reaction rates while preserving its predictive ability. The structural simplification includes the three following strategies :

Simplifying the description of enzymatic capacities

Seven enzymatic capacities V_{max} are represented in the original model. Some of these capacities were assumed to vary over time (temporal effect) and/or to depend

on the phenotypic group (phenotype effect), according to experimental evidences (Desnoues et al. 2014). The characteristics of enzyme capacities are summarized in Table 2.1. In order to simplify the model, we systematically tested the impact of the suppression of the phenotype and/or the temporal effect on each single capacity. Depending on the characteristics of the considered enzyme (Table 2.1), the procedure is slightly different :

$$\text{Phenotype effect :} \quad \begin{cases} V^1_{max} \\ V^2_{max} \end{cases} \rightarrow \frac{V^1_{max} + V^2_{max}}{2} \tag{2.4}$$

$$\text{Temporal effect :} \quad V_{max}(t) \rightarrow < V_{max}(t) >_t \tag{2.5}$$

$$\text{Double effect :} \quad (2.4) \text{ then } (2.5) \text{ applied} \tag{2.6}$$

where $< . >_t$ stands for temporal average over the whole dynamics.

TABLE 2.1 – Characteristics of enzymatic activities in Desnoues et al. (2018)

V_{max}	Phenotype effect	Temporal effect
v_1	No	No
v_2	No	Yes
v_3	Yes	No
v_4	No	Yes
v_5	No	Yes
v_6	Yes	Yes
v_7	Yes	Yes

Rate simplification

In the original model, enzymatic reactions were represented by an irreversible Michaelis-Menten (MM) equation :

$$u(x,t) = V_{max} \frac{x(t)}{K_m + x(t)} \tag{2.7}$$

where V_{max} is the enzymatic capacity. K_m is the affinity of the enzyme for the substrate, $x(t)$ is the concentration of the substrate at time t.

The objective here is to simplify Eq.(2.7) in order to improve the efficiency of the numerical simulation. Depending on the relative levels of the substrate concentration and the MM equation affinity, two simplifications of the flows' equations can be made :

Case 1: if $x(t) << K_m$
Substrate concentration is small compared to the affinity of the enzyme for the substrate then we can write : $u(x,t) = \frac{V_{max}}{K_m} x(t)$.

Case 2: if $x(t) >> K_m$
Substrate concentration exceeds the affinity of the enzyme for the substrate, so that the enzyme can be supposed close to saturation : $u(x,t) = V_{max}$.

Futile cycle removal

The presence of internal cycles within a metabolic network can lead to the appearance of thermodynamically unfeasible loops i.e. reactions that run simultaneously in opposite directions (for example Fig. 2.2) and have no overall effect on the exchange fluxes of the system. This is an undesirable situation that causes numerical issues and makes the estimation of the corresponding parameter values an ill-posed problem.

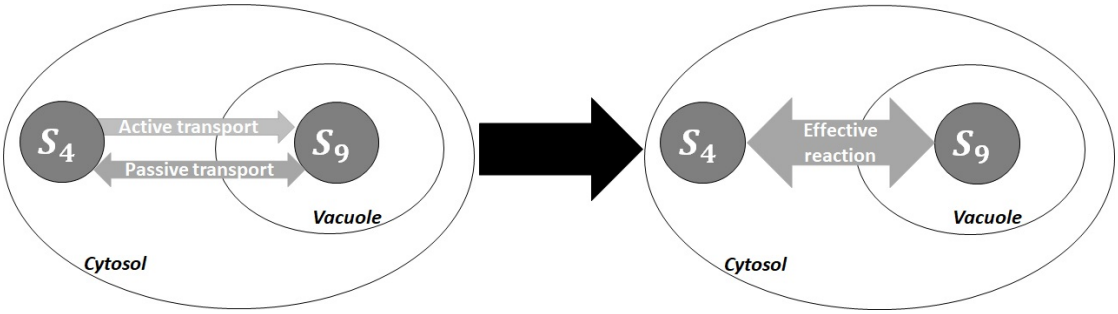


FIGURE 2.2 – S_4 is the glucose in the cytosol transported to the vacuole as S_9 via an active (unidirectional transport) and passive (reversible transport).

In this context, our strategy was to remove each futile cycle by replacing the antagonist reactions by a single effective reaction preserving the net exchange flux of the system. Different kinetics can be tested for the effective reaction, as alternative reduction approaches. Consistently with the previous reduction method, we decided to test two linear reaction forms, namely

$$u(x,t) = k_i x_i - k_j x_j \tag{2.8}$$

and

$$u(x,t) = k_i (x_i - x_j) \tag{2.9}$$

where x_i, x_j are the variables involved in the futile cycle and k_i, k_j are the coefficients to be estimated.

2.2.3.3 Time-scale analysis and QSS approximation

Biological systems are often characterized by the presence of different time scales (seconds, hours, days). Following Heinrich et al. (1996), an appropriate measure of the time scales involved is given by

$$\tau_i(t) = -\frac{1}{Re(\lambda_i(t))} \tag{2.10}$$

where $Re(\lambda_i)$ are real parts of the eigenvalues λ_i of the Jacobian matrix of the system, along a given trajectory. The presence of fast modes in the system allows the reduction of the number of variables based on a quasi-steady-state assumption.

Based on the above information and on the analysis of time-series of the full model, variables can be divided into two groups $x = (x^{(1)}, x^{(2)})$, where $x^{(1)}$ and $x^{(2)}$ correspond respectively to the slow and fast variables of the system (Heinrich et al. 1996; López Zazueta et al. 2018).

Application of the QSS approximation states that

$$\frac{dx^{(2)}}{dt} = f_2(x^{(1)}, x^{(2)}, I(t), v(t), p) = 0 \quad \rightarrow \quad x^{(2)}_{ss} = g(x^{(1)}) \tag{2.11}$$

It follows that, after a relaxation period, the system can be approximated by the reduced model :

$$\frac{dx^{(1)}}{dt} = f_1((x^{(1)}, g(x^{(1)}), I(t), v(t), p) \tag{2.12}$$

of lower dimension.

2.2.4 Experimental and artificial data

2.2.4.1 Experimental data

The 106 peach genotypes used in this study come from an inter-specific progeny obtained by two subsequent back-crosses between *Prunus davidiana* (Carr.) P1908 and *Prunus persica* (L.) Batsch 'Summergrand' and then 'Zephyr' (Quilot et al. 2004a). They were planted in 2001 in a completely randomized design in the orchard of the INRAE Research Centre of Avignon (southern France). Experimental monitoring of peach fruit growth and quality has been conducted in 2012, as described in (Desnoues et al. 2014). The concentration of different metabolites, namely sucrose, glucose, fructose, sorbitol, and hexoses phosphates, the fruit flesh fresh weight and dry matter content were measured at different time points during fruit development, for all genotypes. In addition, the temporal evolution of enzymatic capacities (maximal activity) of the twelve enzymes involved in sugar metabolism was measured over the whole population (Desnoues et al. 2014). The resulting dynamic patterns were analyzed and compared by means of a generalized mixed linear-effect model (GLMM). Accordingly, some enzyme activities were shown to vary over time and/or depend on the phenotypic group (Desnoues et al. 2014).

Training set

The 10 genotypes already used by Desnoues et al. (2018) were selected as the training set for our reduction strategies. They include five genotypes having a 'standard phenotype', namely a balanced fructose-to-glucose ratio at maturity between 0.6 and 0.9, and five considered to have a 'low fructose phenotype' due to the lower proportion

of fructose compared with glucose based on their sugar composition at maturity (Desnoues et al. 2018). For these 10 genotypes, 3 biological measurements are available at 6 dates after bloom.

The training set was used to test each reduction method individually as well as their combination, based on the AIC_C value and the calibration time (see section 2.2.5.3).

Validation set

The quality of the final reduced model was evaluated by calibration on a validation set for which fewer data points were available (one single biological measurement at 6 dates). The idea was to select 30 additional genotypes of the inter-specific peach progeny, which in complement to the training set, represented the greatest diversity in terms of growth rate and duration. For this aim, experimentally measured growth curves were interpolated with a smoothing spline algorithm (Chambers et al. 1992) with 16.4 degrees of freedom in **R** (R Development Core Team 2015) and the maximum and average growth rate quantified as the maximum and the average of the growth curve's derivative over fruit development. A principal component analysis (PCA) was performed on growth rate and growth duration for the whole progeny of 106 genotypes using the **R ADE4** library. The first two principal components accounted for more than 90% of the genetic diversity. The first axis was mainly related to the growth rate whereas the second one reflected the duration of growth. As shown in Fig. 2.3, the ten genotypes of the original study provided a good representation of the observed diversity in growth rate. However, their growth duration was relatively short, compared to the existing variability. As a consequence, most of the new genotypes have been selected in the upper-left panel of the plan, in order to capture the greatest genetic diversity in terms of fruit development. An equal proportion of the two phenotypic groups was maintained.

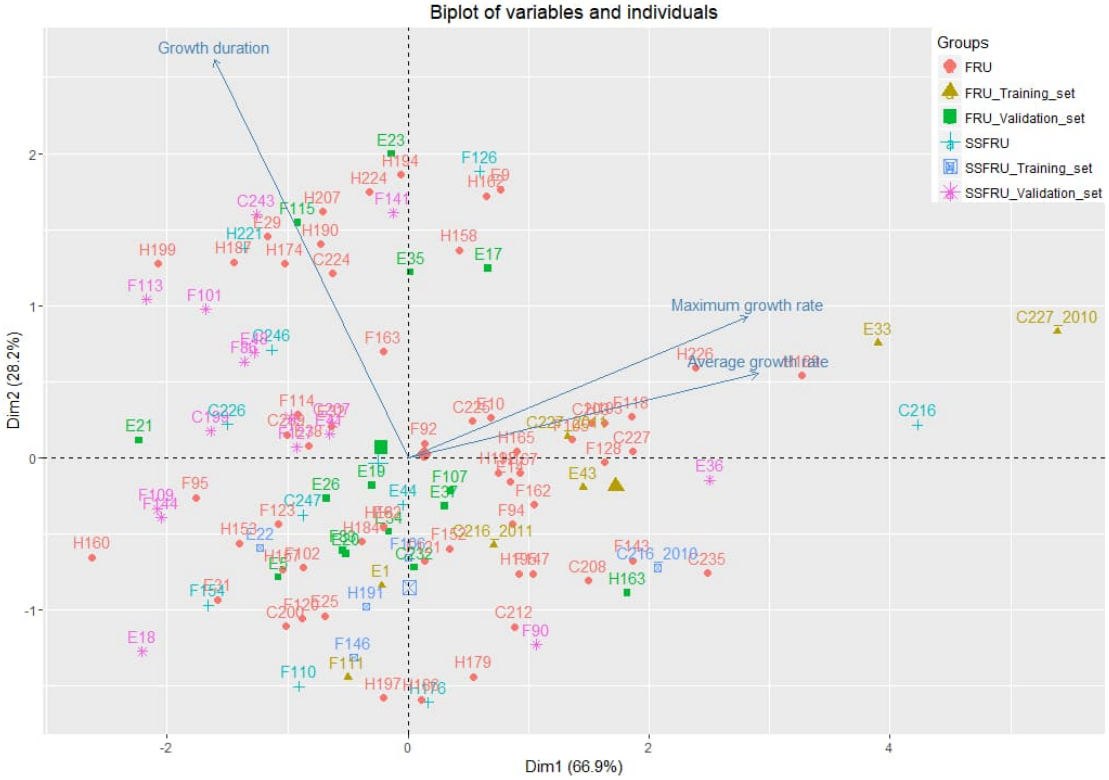


FIGURE 2.3 – Principal component analysis (PCA) for the whole progeny of 106 genotypes. It represents the projection on the Dim1 and Dim2 of the growth duration and growth rate obtained with growth curves.

2.2.4.2 Virtual genotypes

In addition to the training set, a virtual experiment was performed to evaluate the reliability of the reduction methods to variations in parameter values, initial conditions, and input functions, expected in large genetic populations. For this aim, 20 000 virtual genotypes were generated by randomly assigning model parameters and inputs, based on data from the 10 profiles used in (Desnoues et al. 2018).

The values of the parameters p were taken randomly using a uniform distribution between the minimum and the maximum of the previously estimated values over the set of 10 genotypes (Desnoues et al. 2018). Initial conditions, such as $initial\ fruit\ weight$, and $initial\ sugar\ concentration$ were assigned randomly using a uniform distribution within the range of observed values plus a variation of 40%.
Given the high correlation among parameters describing fruit growth curves (Barrasso et al. 2019), model inputs, such as $fruit\ weight$, were randomly assigned using a uniform distribution picking one of the observed growth dynamics and adding an overall random variation between zero and 10% on fruit weight. Finally, shifts in the duration of fruit development among genotypes were also considered. The maturity date was chosen randomly using a uniform distribution within the range of observed

dates broaden of 40%.

2.2.5 Numerical methods

2.2.5.1 Mathematical notations

- $x(t, p^{(k)})$: original model associated to parameters $p^{(k)}$ (i.e. genotype k)
- $\tilde{x}(t, \tilde{p}^{(k)})$: reduced model for the genotype k.

Note that the notation $\tilde{x}(t, \tilde{p}^{(k)})$ can apply to different versions of the reduced model, depending on the considered reduction method.

- $\mathcal{T}_S^{(k)}$: set of the N_S simulation times for the genotype k
- $\mathcal{T}_M^{(k)}$: set of the N_M measurement times for the genotype k
- $X^{(k)}(t_j)$: N experimental observations for the genotype k, with $t_j \in \mathcal{T}_M^{(k)}$. Note that $N = 4 \times N_M \times r$, where r is the number of replicates at time t_j, for the 4 different sugars (sucrose, glucose, fructose and sorbitol). $r = 3$ for the training set and $r = 1$ for the validation set.

2.2.5.2 Parameter estimation

In this section, we aim to estimate the parameters of the models to fit our observations i.e. our measured sugars concentrations. For this purpose, we note $X^{(k)} = (X_1^{(k)}, \ldots, X_N^{(k)})$ the vector of the experimental observations at several times for the genotype k and suppose that :

$$\mathbb{E}(X_i^{(k)}) = \mathcal{M}_{p^{(k)}}(x_i^{(k)})$$

where $x_i^{(k)} = (x^{(k)}(t_i))$ is the set of system variables at $(t_i)_{i\in[1,N]}$, $p^{(k)}$ is the vector of parameters to be estimated and $\mathcal{M}_{p^{(k)}}$ is the mathematical function relying the considered model to the data (see Appendix A for more information). Here, the observations $X^{(k)}$ are assumed to follow a Gaussian law $\mathcal{N}(\mathcal{M}_{p^{(k)}}(x^{(k)}), \sigma_k^2)$ with constant variance σ_k^2.

The estimation of our parameters can be performed through the maximization of the likelihood. We note $\ell(p^{(k)}, \sigma_k^2)$ the log-likelihood function for the genotype k.

Under the assumption of observation independence, the log-likelihood can be defined as follows :

$$\ell(p^{(k)}, \sigma_k^2) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_k^2) - \frac{1}{2\sigma_k^2}\sum_{i=1}^{N}(X_i^{(k)} - \mathcal{M}_{p^{(k)}}(x_i^{(k)}))^2 \tag{2.13}$$

A maximum log-likelihood estimator $(\hat{p}^{(k)}, \hat{\sigma_k}^2)$ of $(p^{(k)}, \sigma_k^2)$ is a solution to the maximization problem :

$$(\hat{p}^{(k)}, \hat{\sigma_k}^2) = \underset{p^{(k)}, \sigma_k^2}{\arg\max}\, \ell(p^{(k)}, \sigma_k^2) \tag{2.14}$$

In this Gaussian case, the maximum log-likelihood estimator is thus equivalent to the ordinary least-square estimator :

$$\hat{p}^{(k)} = \underset{p^{(k)}}{\arg\max} \left\{ -\sum_{i=1}^{N} (X_i^{(k)} - \mathscr{M}_{p^{(k)}}(x_i^{(k)}))^2 \right\} \tag{2.15}$$

$$\hat{\sigma}_k^2 = \frac{1}{N} \sum_{i=1}^{N} (X_i^{(k)} - \mathscr{M}_{\hat{p}^{(k)}}(x_i^{(k)}))^2 \tag{2.16}$$

Matlab software (MATLAB R2018a, The MathWorks Inc., Natick, MA) was used for model integration (solver ode23tb Hosea et al. (1996)) and calibration. A genetic algorithm (function ga Goldberg (1989) of Global Optimisation Toolbox) was used for maximization of Eq. (2.15). The population size, the maximum number of generations, and the crossover probability have been respectively set at 200, 300, and 0.7. For each reduced version of the model (individual or combined reduction methods), free parameters were numerically re-estimated. The fitting process was considered at convergence when the average relative change in the best-cost function, i.e. the sum of squared errors, value over generations was less than 10^{-6}. For each genotype k and reduced model, estimations procedure has been repeated ten times to take into account the stochastic nature of the genetic algorithm and to ensure the good exploration of the parameters' space. The solution having the best score was kept for subsequent analyses.

2.2.5.3 Model selection

Individual and combined reduction methods were evaluated according to three criteria of major importance for our application : the corrected Akaike Information Criterion (AIC_C), the gain in calibration time (%) and the expected error (%) between the original and reduced models.

Akaike Information Criterion

The AIC gives information on the likelihood of the proposed model based on available experimental data and weighted by the number of free parameters : (Burnham et al. 2002) :

$$AIC(\mathscr{M}_p) = -2\,\ell(p,\sigma^2) + 2n_p \tag{2.17}$$

where n_p is the number of estimated parameters p and $\ell(p,\sigma^2)$ is the maximum log-likelihood. In this paper, we used the corrected AIC as we deal with a small set of observations and a considerable number of parameters.

$$AIC_C(\mathscr{M}_p) = AIC(\mathscr{M}_p) + \frac{2n_p(n_p+1)}{N-n_p-1} \tag{2.18}$$

where N is the number of observations. For genotype k and for each reduction method, we defined

$$\Delta^{(k)}_{AIC_C}(\mathscr{M}_{\tilde{p}^{(k)}}, \mathscr{M}_{p^{(k)}}) = AIC_{C\,reduced}(\mathscr{M}_{\tilde{p}^{(k)}}) - AIC_{C\,original}(\mathscr{M}_{p^{(k)}}) \tag{2.19}$$

as the AIC_C difference between the reduced and the original model. Note that Δ_{AIC_C} is always computed using the best estimated parameter solution for the considered model. Whenever the average over the 10 genotypes ($< \Delta_{AIC_C} >_G$) was negative, the reduction method was validated.

Gain in calibration time

We used the calibration of a specific genotype ($E43$) as a proxy of the maximum expected calibration time on the population. Genotype $E43$ was selected because it required a long calibration time on the original model proposed by Desnoues et al. (2018) (approximately 11 hours on average on a 3.1GHz Intel(R) Xeon(R) processor) but it did not suffer from numerical instabilities, that could complicate the calibration process. Note that the overall calibration time of a model depends both on the integration time of each evaluation step and on the convergence of the cost function that sets the actual number of generations performed by the algorithm. Both aspects may be affected by the model reduction.

To evaluate the gain in calibration time due to model reduction, parameter estimation was performed for each reduction method, following the general procedure (see section 2.2.5.2), and compared to the calibration time obtained for the original model. An initial population $\mathscr{P}_0$ was randomly selected assuming a uniform distribution in the parameter range and then kept fixed for all calibration processes (both original and reduced models). For models having a reduced number of parameters, the initial population was directly derived from $\mathscr{P}_0$.

The gain (G_t) was defined as the gain (in %) in calibration time T between the original and the reduced model :

$$G_T = \frac{T_{original} - T_{reduced}}{T_{original}} \times 100$$

Expected error

Simulations of the original and reduced models were compared by the Normalized Root Mean Square Error over the 10 model variables :

$$J_i(p^{(k)}, \tilde{p}^{(k)}) = \frac{\sqrt{\frac{1}{N_S}\sum_{j=1}^{N_S}(x_i(t_j, p^{(k)}) - \tilde{x}_i(t_j, \tilde{p}^{(k)}))^2}}{max_j(x_i(t_j, p^{(k)})) - min_j(x_i(t_j, p^{(k)}))} \quad \forall i \in \{1, \ldots, 10\} \tag{2.20}$$

where $x(t, p_k)$ and $\tilde{x}(t, \tilde{p}_k)$ are the concentration predicted by the original and reduced model, respectively. Parameters for the reduced model were derived from the values

of the corresponding parameters in the original model.

The quality of the QSS approximation was assessed by computing J_i for each variable in the model, over the whole dynamics.

In the context of the virtual experiment, the Expected Error (%) of the reduced model was defined as the average distance J over the virtual population :

$$\text{Expected Error} = \frac{1}{N_{VG}} \sum_{k=1}^{N_{VG}} < J_i(p^{(k)}, \tilde{p}^{(k)}) > \times 100 \tag{2.21}$$

with

$$< J_i(p^{(k)}, \tilde{p}^{(k)}) >= \frac{1}{10} \sum_{i=1}^{10} J_i(p^{(k)}, \tilde{p}^{(k)})$$

where N_{VG} is the number of virtual genotypes and 10 is the number of variables. In our case, $N_{VG} = 20\,000$. The Expected Error was used to quantify the reliability of the reduction.

2.2.6 Results

2.2.6.1 Strategy 1 : Identification of low sensitive parameters

The objective of the sensitivity analysis was to identify parameters having a significant influence on the outputs of the model, over the whole dynamics and for all tested genotypes. A multivariate sensitivity analysis (Lamboni et al. 2009) was used for this purpose. The aggregate generalized sensitivity indices (aGSI) (see section 2.2.3.1) shown in Fig. 2.4 give a common ranking of model parameters according to their influence on the whole sugar phenotype, as it is made up by the four output sugars (sucrose, sorbitol, glucose, and fructose).

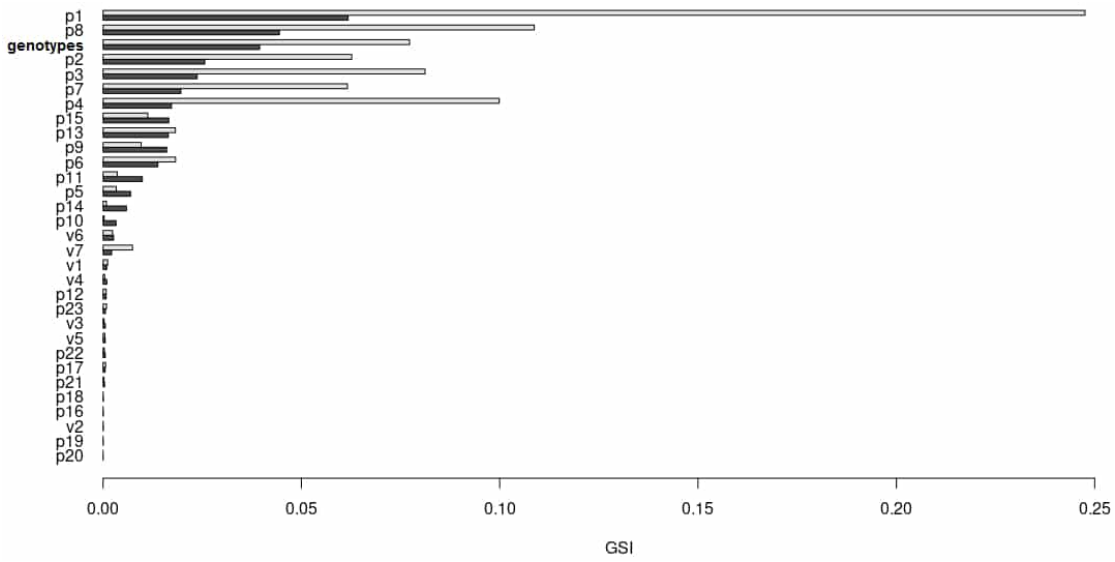


FIGURE 2.4 – Aggregate Generalized sensitivity indices (aGSI) for the parameters of the model and genotypes (the training set) on four outputs (Sucrose, Sorbitol, Glucose and Fructose) of the sugar model. The main sensitivity indices are in dark bars and interaction ones are in grey bars.

Parameter (p_1) related to the action of cell-wall invertase in fruit apoplasm and the coefficient of sucrose import (p_8) are the most important parameters, followed by the activities of acid invertase (p_2), the activities of Fructokinase (p_3), Hexokinase (p_4) and the resynthesis rate of sucrose from hexose phosphate (p_7). Indeed, p_1, p_3, and p_4 parameters are the most sensitive parameters for sucrose, fructose and glucose concentrations respectively (see Fig. 2.13).

Interestingly, the `genotype` factor is ranked third, meaning that it does not affect parameters' sensitivity as much as expected. A closer look at the results shows that the choice of the genotype essentially affects the second principal component, via the definition of the initial conditions of the model (see the supplemental information Fig. 2.12).

Among the 14 parameters estimated ($p_1, \ldots, p_{14}$) in the original model, four parameters, namely p_5, p_{10}, p_{12} and p_{14}, have a negligible effect on the four outputs, independently of the peach genotype. Accordingly, these parameters can be fixed to their nominal values i.e. their average value over the ten genotypes, without affecting the quality of predictions. The validity of such a reduction strategy was tested on the ten genotypes of the training set. The difference in Akaike criterion (Δ_{AIC_C}) between the reduced and the original models was computed for each genotype. Results presented in Table 2.2 show that such a reduction in the number of parameters is strongly beneficial for nine out of the ten genotypes with largely negative Δ_{AIC_C} values, and roughly neutral for one genotype ($\Delta_{AIC_C} \sim 0$). The gain in calibration time, however, is important (25%) and the expected error over the progeny of virtual genotypes is low, demonstrating a good reliability of the proposed simplification. For these reasons, the

model with 10 parameters to be estimated was selected.

TABLE 2.2 – Δ_{AIC_C} calculated between reduced and original models for the training set and the gain in calibration time (%) for E43. The Expected error ± standard deviation (Std) between original and different reduced models for 20 000 virtual genotypes.

Simplification method		Δ_{AIC_C}											Calibration	Expected Error
		E1	E33	E43	F111	E22	F106	F146	H191	C216	C227	$<\Delta_{AIC_C}>_G$	Time gain %	Virtual genotypes
Low sensitive parameters fixed		-11.5	-6.4	-0.9	-14.04	-13.2	-28.3	-13.5	-14.3	-18.7	-87.7	-20.8	25.8	4.9 ± 6.5
V_{max} **Type**	v_3	-1.01	-5.9	-4.15	-4.2	1.1	-2.3	-0.3	-6.1	-6.1	-72.02	-7.9	22.4	0.5 ± 1.3
effect	v_6	-0.1	-4.3	0.06	- 3.9	0.7	-5.5	-0.3	-5.4	-6.1	-87.7	-11.3	26.6	1.7 ± 1.6
removed	v_7	-0.7	-36.4	0.2	-6.02	1.4	-5.6	1.9	-5.2	-4.9	-94.5	-14.9	33.9	2.9 ± 4.5
V_{max}	v_2	-0.1	-3.1	0.06	-3.7	0.7	-5.1	-0.3	-0.3	-6.3	-83.4	-10.1	31.6	0.3 ± 0.7
Temporal	v_4	-0.8	-8.2	0.7	-5.6	-5.03	-2.5	-2.5	-5.1	- 6.1	-90.3	-12.3	19.4	2.9 ± 2.5
effect	v_5	0.2	-6.8	0.5	-4.8	1.8	-5.8	2.03	-2.9	-6.1	-91.1	-11.3	20.3	5.5 ± 5.7
removed	v_6	-0.3	-0.4	-0.1	-27.04	1.7	-5.5	-0.2	-5.3	-5.7	-84.9	-12.7	30.5	4.1 ± 3.1
	v_7	8.6	-25.1	21.1	11.02	19.6	20.01	29.05	12.4	15.6	-97.5	1.5	24.2	6.8 ± 4.5
Rate simplification		-17.2	-53.4	8.9	-35.4	-2.9	2.7	-14.7	-22.9	-5.7	- 71.04	-21.1	6.7	18.6 ± 9.7
Futile cycle	Eq.(2.8)	2.5	-0.9	15.6	-1.6	-0.01	-2.3	-0.6	-1.5	-5.9	-43.23	-3.8	23.6	12.7 ± 14.7
removal	Eq.(2.9)	0.7	- 56.7	-5.6	-37.1	- 9.02	- 10.5	-6.7	-35.5	-12.2	- 70.7	-24.3	24.1	11.5 ± 9.9
Intermediate reduced model		-32.7	-18.6	-3.7	- 24.5	-11.8	-24.04	-16.5	-20.3	-18.8	-43.1	-21.4	30.5	22.5 ± 8.4
Final reduced model		-32.5	-19.1	-4.3	-25.1	-12.7	-1.01-	- 16.4	-20.4	-18.8	-43.3	-18.5	43.3	22.5 ± 8.5

2.2.6.2 Strategy 2 : Structural simplification of the model

Structural simplification methods are another way to reduce the complexity of dynamic systems by improving the generality of the model and the numerical integration of the ordinary differential equations.

Firstly, we tried to remove the temporal and the phenotype effects in the enzyme activities, $v_2, \ldots, v_7$ (v_1 has neither phenotype nor temporal effects). The results of this simplification are shown in Table 2.2. The elimination of the phenotype effect for v_3, v_6 and v_7 resulted in a decrease of the AIC_C value for nine genotypes, neutral for one genotype, and was thus selected for the final reduction. The elimination of the temporal effect for v_2, v_4, v_5, v_7 was also advantageous on the corrected AIC results for all ten genotypes. Nevertheless, when we tried to eliminate the temporal effect of v_7, the resulting Δ_{AIC_C} was positive for most genotypes. This is in line with the results of multi-variate sensitivity analysis according to which $v_2, \ldots, v_6$ have a low sensitivity on the four outputs of the model, whereas v_7 has a non-negligible effect on the dynamics of glucose concentration. According to these results, the elimination of the temporal effect was validated only for v_2, v_4, v_5, v_6. In support of this choice, the test with the virtual genotypes shows that the expected error between the reduced and the original model is small (Table 2.2).

In the second phase, we tested the possibility of simplifying the enzymatic reaction rates (Eq.(4.8)). For each reaction in the model, Fig. 2.5 compares the order of magnitude of the substrate $x(t)$ to the corresponding affinity K_m. The boxplots show that (**Case 2**, see section 2.2.3.2) simplification strategy can be applied only for the reaction rates u_5 and u_7. Therefore, their reaction rates can be written as $u = V_{max}$. All other flows verify the (**Case 1**, see section 2.2.3.2) and can therefore be expressed as $u = \frac{V_{max}}{K_m} x(t)$. The rates simplification improves the corrected AIC for eight genotypes

and yields a substantial gain in the calibration time. The expected error over the virtual progeny is higher than in the previous reduction steps, but still in the range of accuracy of the original model Desnoues et al. 2018. According to these observations, the enzymatic reaction rates simplification strategy was validated.

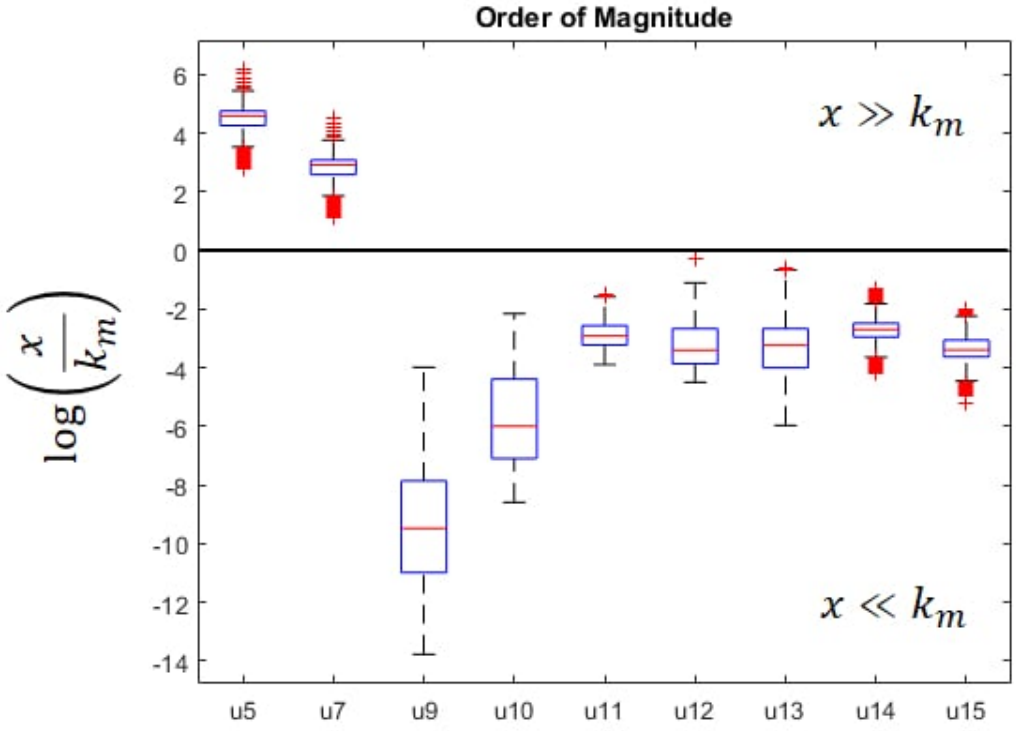


FIGURE 2.5 – Differences in order of magnitude between enzyme affinity (K_m) and substrate concentration (x) calculated over the whole dynamics and the training set for each reaction rate u_i, $i \in \{5, 7, 9 \dots 15\}$.

Eventually, futile cycles were detected to reduce the full system. In the original model, glucose, and fructose sugars can be transported to the vacuole via two possible mechanisms : an active, unidirectional transport (u_5, u_7) and passive reversible transport (u_6, u_8). Simulations showed that, whenever the genotype, the net flux mostly pointed in the direction of an export for both fructose and glucose from the vacuole to the cytosol (Desnoues et al. 2018). However, futile cycles occurred due to the presence of the active transport mechanism, that continually brings glucose and fructose back into the vacuole. Indeed, u_5 and u_6 (respectively u_7 and u_8) had the same evolution over the whole dynamics for all ten genotypes (Fig. 2.6) : the active and passive transport ran simultaneously in two opposite directions.

According to our strategy (section 2.2.3.2), we tried to remove futile cycles by replacing reactions (u_5, u_6) (respectively (u_7, u_8)) with an effective reaction rate of the form $p_{10}\, x_9 - p_{11}\, x_4$ (respectively $p_9\, x_8 - p_{12}\, x_3$) preserving the net export flux from vacuole

to the cytosol. We compared the performance of the reduced model with respect to the original one (Table 2.2). The corrected AIC values were generally slightly negative, with the exception of genotypes E_1 and E_{43}, suggesting an overall improvement of the model structure. Notice that the present strategy did not reduce the total parameters number but decreased model complexity and improved the calibration time.
As a further simplification, we then tried to use a special case of the above mentioned reaction rate with $p_{10} = p_{11}$ (respectively $p_9 = p_{12}$). This time, the simplification was fully validated by the corrected AIC on all genotypes (Table 2.2, Eq.(2.9)). The expected error over the virtual genotypes was estimated to 13% and the calibration time was lowered by 24% with respect to the original model, thanks to structural simplification and the reduction of the number of parameters to be estimated. Accordingly to these results, the simplification by Eq.(2.9) was validated.

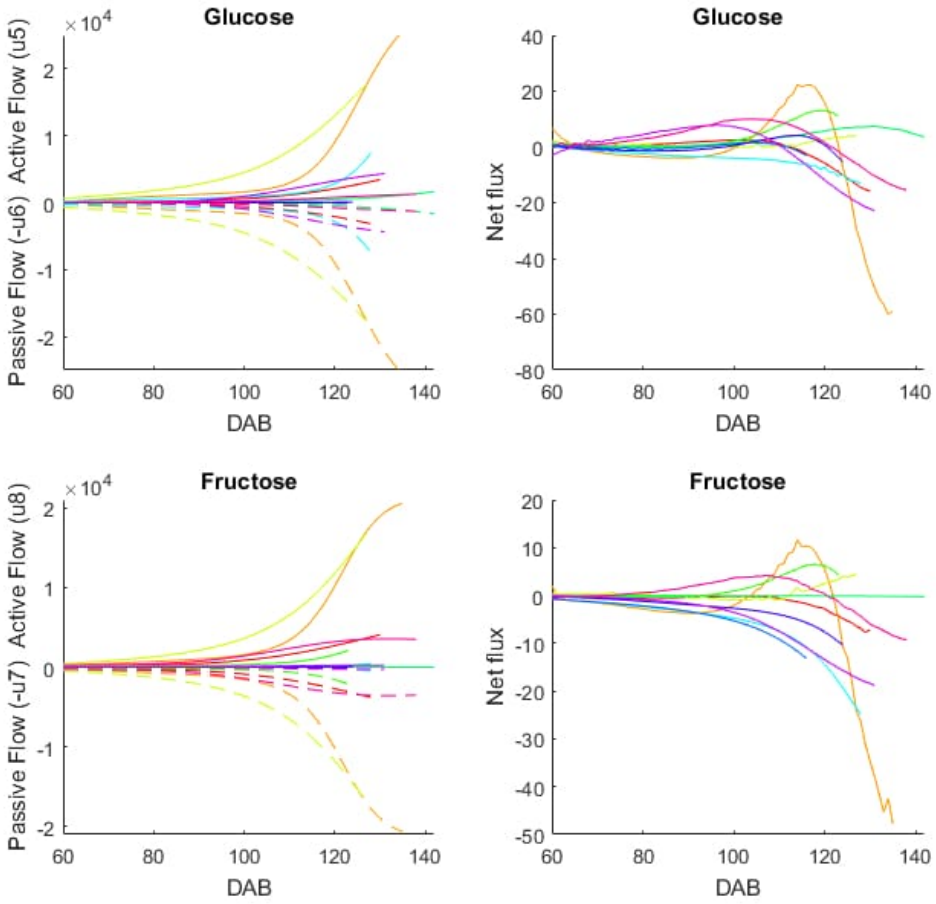


FIGURE 2.6 – Evolution of the active flux (solid lines) and passive transport (dashed lines) for glucose (respectively fructose) and net flux during fruit development (DAB, day after bloom) for the ten genotypes of the training set (different colors).

2.2.6.3 Strategy 3 : Time-scale analysis and QSSA

Results from the reduction strategies 1 and 2 were combined into an intermediate reduced model. This model had only 9 parameters to be estimated, linear flows and only one temporal enzymatic capacity, common to all genotypes. Improvement in AIC_C with respect to the original model confirmed a strong benefice for all ten genotypes (Table 2.2). The expected error over a large progeny was estimated around 20%, close to the performance of the original model.

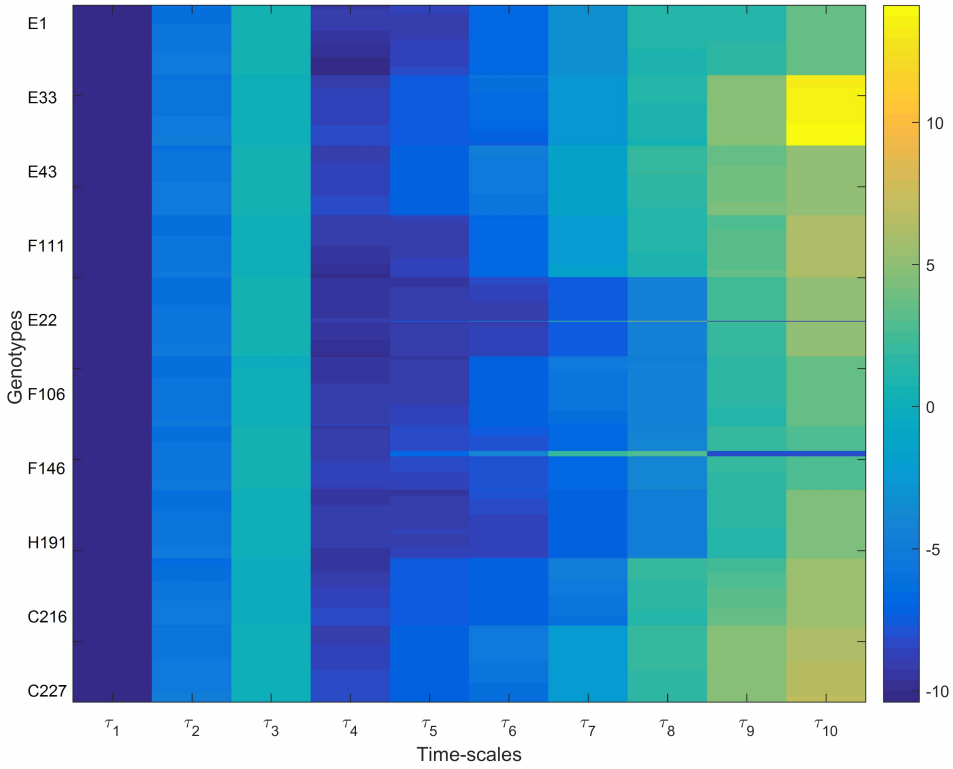


FIGURE 2.7 – Order of magnitude of time scales τ_i along fruit development (DAB, days after bloom) for the 10 genotypes of the training set.

On the basis of this intermediate reduced model, time scale analysis was performed to detect the possible presence of fast modes in the system. The analysis of the Jacobian matrix, indeed, confirmed the presence of different modes, with typical time scales spanning a few seconds up to days, for all tested genotypes (Fig. 2.7).

A fast transient dynamics, followed by a slow one, was observable in the numerical simulations of the original and intermediate reduced models for the hexose phosphates concentration (variable x_5, see supplemental information, Fig. 2.15). In addition, following the method proposed in (López Zazueta et al. 2018; Heinrich et al. 1996), we analyzed the predicted concentration of sugars in both intracellular compartments,

for all genotypes. The concentration of the hexose phosphate (x_5) was systematically lower than the concentrations of the other variables in the system, as expected for the fast components of the system (Fig. 2.8). Accordingly, x_5 was assumed to be at quasi-steady-state and its equation was replaced by an algebraic function of the slow variables.

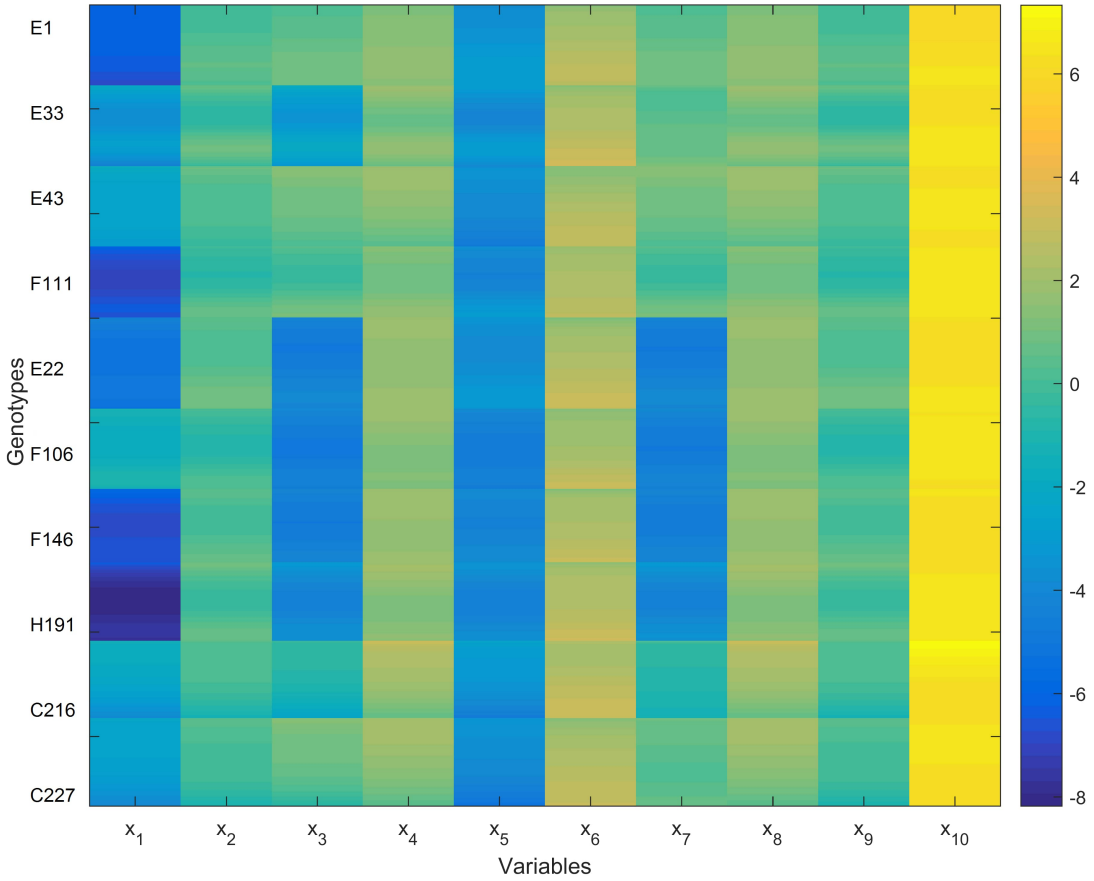


FIGURE 2.8 – Order of magnitude of the predicted sugars concentrations (mg gFW^{-1}) in the cytosol (x_1 : Sucrose, x_2 : Sorbitol, x_3 : Fructose, x_4 : Glucose, x_5 : Hexose Phosphate, x_{10} : Other compounds) and vacuole (x_6 : Sucrose, x_7 : Fructose, x_8 : Glucose, x_9 : Sorbitol), along fruit development (DAB, days after bloom) for the ten genotypes of the training set.

We compared the intermediate reduced model with its QSS approximation by calculating J_i (Eq.(2.20)) as explained previously. J_i was very low, less than 1%, over the whole dynamics for all variables (Fig. 2.9). This result was validated also on the virtual genotypes simulated with QQS approximation (see the supplemental information Fig. 2.16). In addition the QSS assumption, further increased the performance of the model, leading to a gain in the calibration time of 40% with respect to the original model.

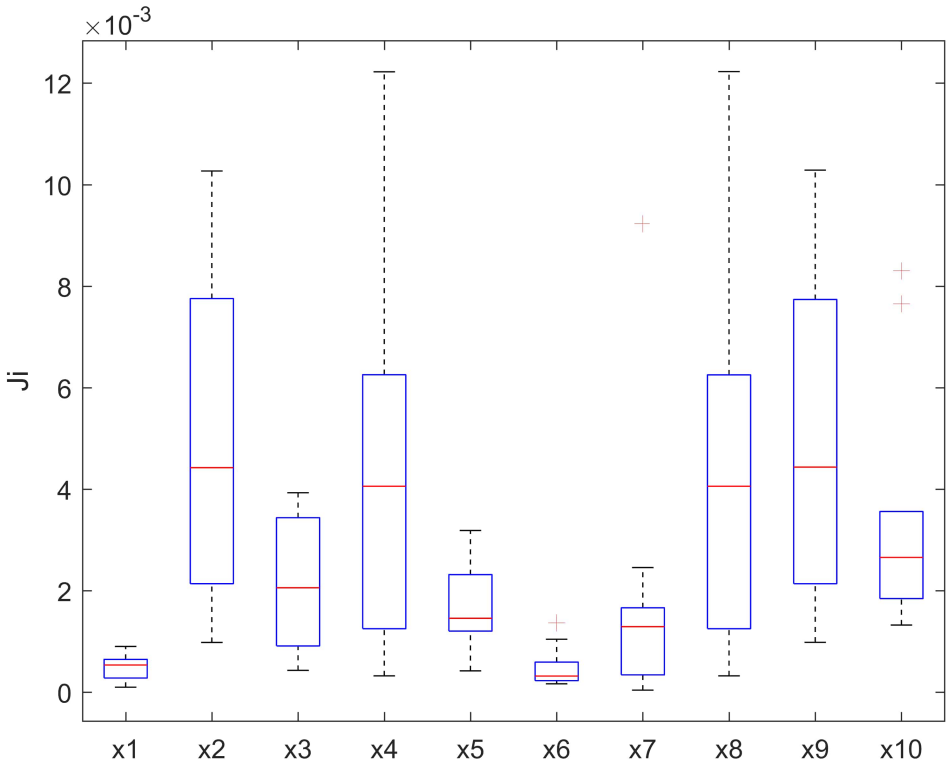


FIGURE 2.9 – Normalized Root Mean Square Errors $J_i, i \in \{1,\ldots,10\}$ between the intermediate and reduced models after application of the QSSA to x_5. The boxplot shows the variability of J_i over the training set

2.2.6.4 Evaluation of the reduced model

The validity of the reduced model was verified on some new genotypes of the interspecific peach progeny, for which few data were available.

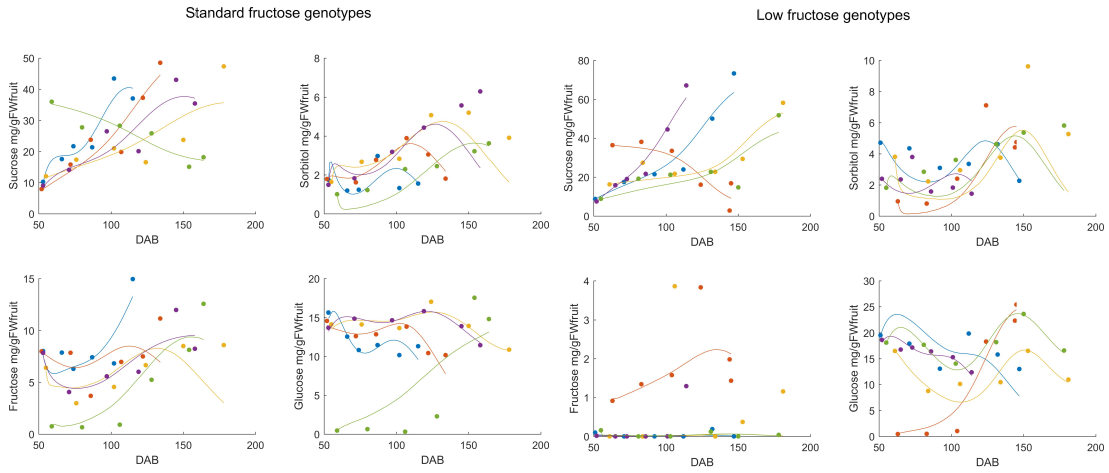


FIGURE 2.10 – Evolution of the concentration ($mggFW^{-1}$) of sugars during fruit development (DAB, days after bloom) for ten representative genotypes of the validation set with standard (left) and low fructose (right) phenotypes. Dots represent experimental data and lines are model simulations.

The reduced model was then calibrated on the dynamics of sugar concentration of these selected genotypes, as described in section 2.2.5.2. The results presented in (Fig. 2.10) showed a satisfactory agreement between model and data, all over fruit development, for most genotypes. The average $NRMSE$ (Table 2.8) ranged from 10% to 30% for the main sugars, in good agreement with estimations over the virtual progeny. These results confirmed that the reduced model offered a quality of prediction close to the original one with fewer parameters to be estimated and shorter integration time.

From a biological perspective, an important prediction of the model developed by Desnoues et al. (2018) was that a difference in fructokinase affinity could be at the origin of the phenotypic difference observed between standard and low fructose genotypes.

We checked if the estimations obtained with the reduced model still supported this hypothesis. Fig. 2.11 shows a significant difference of estimated fructokinase affinity between the two phenotypic groups, in agreement with the original model based on the Student t-test (p-value $<2.0187e^{-9}$)

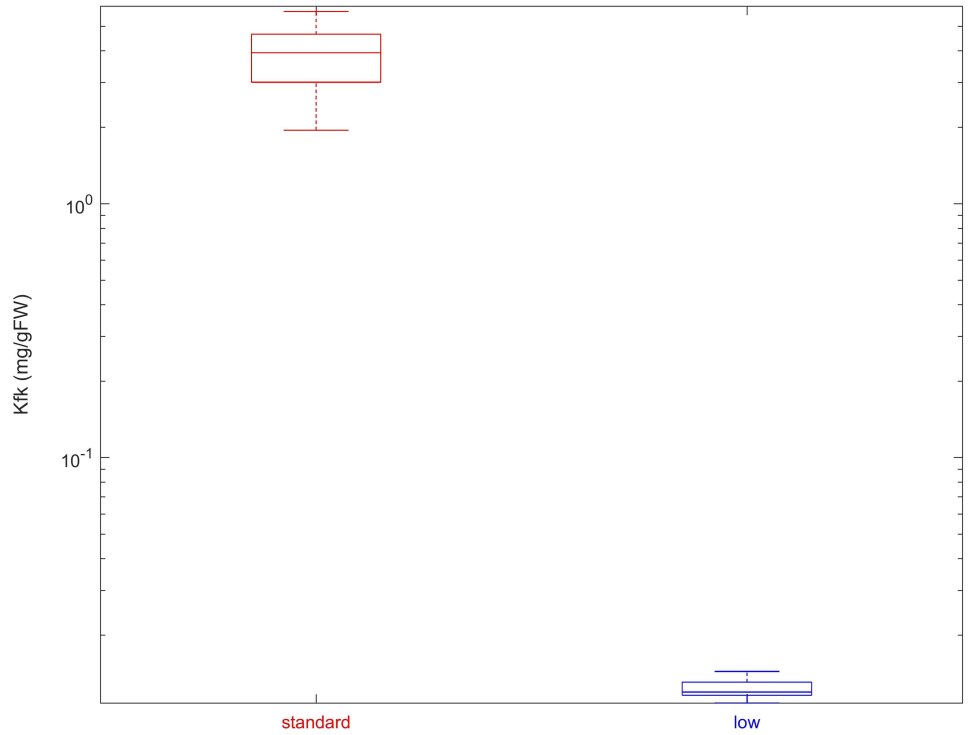


FIGURE 2.11 – Difference in the estimated fructokinase affinity between standard and low fructose phenotypes, for forty genotypes (training and validation sets). The difference is significant with a p-value < $2.0187e^{-9}$.

2.2.7 Discussion

Models of metabolic systems are usually very complex. Complexity stems from the number of components and the high degree of non-linearity included in both the network structure and the individual reaction rates. As a consequence, metabolic models usually suffer from numerical and identifiability issues that seriously hamper their application in the context of genetic studies, especially when they have to be calibrated for hundreds of genotypes. In this paper, we present a reduction scheme that explicitly accounts for genomic diversity. Our approach is based on the systematic evaluation of different reduction methods, that, if successful, are then combined together to yield the final reduced model. When applied to the model of sugar metabolism developed by Desnoues et al. (2018) our approach led to a reduced model that could be efficiently calibrated on a large diversity of genotypes, for which few data are available. The reduced model showed comparable predictions and biological interpretation as the original model, with only a limited number of estimated parameters. Indeed, calibration time was reduced by 40%, a considerable improvement when considering that

the calibration of the original model could span up to 30 hours for a single genotype. Moreover, mitigation of model non-linearities can help limiting numerical issues and increase the reliability of estimated parameters, an important aspect in the context of genetic studies, where large genetic populations have to be calibrated.

The proposed reduction scheme is especially suitable for dynamical models of metabolic and biochemical networks, in which a large number of chemical reactions interact with similar non-linear kinetics. In these systems, indeed, the connectivity properties of the network usually prime over the precise description of the individual rate laws (Barabási et al. 2004). The presence of saturating kinetic functions (like the classical Michaelis-Menten), in particular, allows the simplification of the rate function depending on the substrate range whereas the presence of redundant or opposite reactions opens the way to structural simplification of the system. The extension of these reduction steps to another kind of models is less straightforward. Crop models for instance can involve a large variety of process kinetics, one for each described physiological process. The complexity of the cellular network is replaced by the interaction of a comparatively small number of processes but described by complicated, ad-hoc kinetic functions that can involve several model components as well as external environmental variables (temperature, humidity, light). The simplification of individual rate laws is still possible but it involves case-by-case study.

Although the application of specific reduction methods is tailored to model structure, the proposed evaluation strategy is pretty generic and easily adaptable to a large range of biological models. The main objective of this work was to provide a method to build a reduced model that is adapted to the application to a large panel of genotypes. In this sense, we do not look for the best model for a given genotype but rather for the best *compromise* in terms of accuracy and efficiency over a large genetic diversity. The question recalls the one of "model validation domain" i.e. the ability for a given model to describe data obtained in conditions different from those in which the model itself was calibrated (Mairet et al. 2019). Here it is about selecting for a reduced model having a large validation domain and able to cope with changes in model's inputs, parameter values, and initial conditions.

For this aim, we proposed a criterium based on the simulation of a large number of virtual genotypes and the systematic comparison of the expected distance between the original and the reduced models. Virtual genotypes are built based on the variability observed in a sub-sample of the population, plus a basal variability, expressed as a random effect, to limit the bias due to the choice of the initial sample and to assure a minimal diversity across the virtual population. A few remarks are needed. First, the above method tests the reliability of the reduction, assuming that the original model is valid. In this sense, the amplitude of the basal random effect should be subject to an expert knowledge so to avoid biologically unreasonable situations, that fall outside the conditions of applicability of the model. Second, it is worth to notice that, given the virtual nature of our comparison, the reduced model is parameterized using parameter values that are directly derived from the parameters of the original model, to which it is compared. In this sense, the 'expected NRMSE error' of the reduced model

represents an upper bound of its actual accuracy over an experimental dataset, as parameter re-calibration can significantly improve the performances of the reduced model on real genetic populations.

Ultimately, the existence of a reduced model will considerably speed up the integration of genetic control into ecophysiological models. Currently, most genetic-improved ecophysiological models make use of Quantitative Trait loci (QTL) to describe the genetic architecture of specific model parameters. Basically, each parameter has a specific distribution in the population of genotypes and QTL analyses can be performed for each parameter to decipher the architecture of its genetic control (QTL number and effects, linkage). However, a major drawback of this approach is the difficulty in the calibration of the models for a large number of genotypes (due to a large number of parameters along with restricted number of observations) (Martre et al. 2015; Bertin et al. 2010). Indeed, the statistical power of QTL analyses strongly depends on the size of the population and on the QTL effects i.e. their contribution to the variation of the trait they are associated with (Mackay et al. 2009). So, in order to be of interest, genetic parameters have to vary among genotypes and be quantifiable with relevant accuracy either experimentally or through numerical optimization.

In this perspective, a reduced model with a simpler structure will allow for a better exploration of the parameter space and a more accurate estimation of parameter values. Moreover, the improved calibration time opens the possibility of exploring larger genetic populations so to get more robust QTLs estimation. Finally, it will allow to do simulations over a large number of environmental conditions and/or climatic scenarios.

This is an important step towards dealing with complex Genotype x Environment x Management interactions issues expected in the near future. The development of reliable gene-to-phenotype models will be an important lever to optimize farming in the future climatic conditions.

Acknowledgments

HK was founded by a scholarship of the Lebanese government. We would like to thank V. Signoret for her help in maintaining the inter-specific peach progeny. We are grateful to the IE-EMMAH UMR1114 and IE-GAFL UR1052 teams for taking care of the experimental orchard, and to Dr Olivier Martin for his help in statistics.

2.2.8 Appendices

2.2.8.1 Model description

Model equations

The original model (Desnoues et al. 2018) was written in terms of species *carbon* quantities $C(t)$. Here, we decided to rewrite the system as a function of species concen-

tration $x_i(t)$, for a better readability. The quantity of carbon as a sugar i (C_i) depends on the concentration of i (x_i) according to the following equation :

$$C_i = \sigma_i \, x_i \, V_j \tag{2.22}$$

where σ_i is the carbon concentration of sugar i and V_j is the volume of the intracellular compartment (cytosol or vacuol) in which species i is located. The carbon content σ_i for the different sugar molecules is reported in Table 2.3. Table 2.4 specifies variable location within the cell's compartments.
Differentiation of Equation (Eq. (2.22)) leads to :

$$\frac{dx_i}{dt} = \frac{1}{\sigma_i V_j}\frac{dC_i}{dt} - \frac{1}{V_j}x_i\frac{dV_j}{dt} \tag{2.23}$$

Accordingly, for variables $1,\ldots,5,10$, $C_i = \sigma_i x_i V_1$ whereas $C_i = \sigma_i x_i V_2$ for $i \in [6,9]$. For simplicity,we assume $\frac{V_1}{V_2} = \alpha$. This leads to $\mu(t) = \frac{1}{V_1}\frac{dV_1}{dt} = \frac{1}{V_2}\frac{dV_2}{dt}$.

TABLE 2.3 – Carbon content of each sugar

σ	Sugar	Value
σ_1, σ_6	Sucrose	0.421
σ_3, σ_8	Fructose	0.4
σ_2, σ_7	Sorbitol	0.39
σ_4, σ_9	Glucose	0.4
σ_5	Hexose phosphate	0.27
σ_{10}	Other compounds	0.44

TABLE 2.4 – Model variables and location

x_1	Sucrose	Cytosol
x_2	Sorbitol	Cytosol
x_3	Fructose	Cytosol
x_4	Glucose	Cytosol
x_5	Hexose phosphate	Cytosol
x_6	Sucrose	Vacuole
x_7	Sorbitol	Vacuole
x_8	Fructose	Vacuole
x_9	Glucose	Vacuole
x_{10}	Other compounds	Cytosol

TABLE 2.5 – Reaction rates of the original and reduced models

Equations	Original Model	Reduced Model
Input flows	$I(t) = \sigma_f \frac{dDW}{dt} + R(t) = (\sigma_f + q_g)\frac{dDW}{dt} + q_m DW Q_{10}^{\frac{(T-20)}{10}}$ $R(t) = q_m DW Q_{10}^{\frac{(T-20)}{10}} + q_g \frac{dDW}{dt}$ $DW = DW(t_0) + w_1(1 - e^{-w_2 t}) + \frac{w_3}{1+e^{-w_4(t-w_5)}}$ $u_1(I) = \frac{1}{\sigma_1 V_1}\lambda\lambda_{suc}(t)I(t)$ $\lambda_{suc}(t) = \frac{p_1 t}{t_{max}}$ where t_{max} corresponds to the maturation time $u_2(I) = \frac{1}{\sigma_2 V_1}(1-\lambda)I(t)$ $u_3(I) = \frac{1}{\sigma_3 V_1}\frac{\lambda}{2}(1-\lambda_{suc}(t))I(t)$	
Metabolism	$u_9(v_2, x_1) = \frac{v_2(t)}{p_5+x_1}x_1(t)$ $u_{10}(x_1) = \frac{v_3}{p_{21}+x_1}x_1(t)$ $u_{11}(v_4, x_2) = \frac{v_4(t)}{p_{22}+x_2}x_2(t)$ $u_{12}(v_5, x_2) = \frac{v_5(t)}{p_{13}+x_2}x_2(t)$ $u_{13}(x_6, x_8, x_9) = \frac{v_1}{(1+\frac{x_8+x_9}{p_2})p_{23}+x_6}x_6(t)$ $u_{14}(v_6, x_3) = \frac{v_6(t)}{p_3+x_3}x_3(t)$ $u_{15}(v_7, x_4) = \frac{v_7(t)}{p_4+x_4}x_4(t)$ $u_{16}(x_5) = p_7 x_5(t)$ $u_{17}(x_5) = p_6 x_5(t)$ $u_{18}(R) = R(t)$	$u_9(x_1) = \frac{v_2}{p_5}x_1(t) = r_1 x_1(t)$ $u_{10}(x_1) = \frac{v_3}{p_{21}}x_1(t) = r_2 x_1(t)$ $u_{11}(x_2) = \frac{v_4}{p_{22}}x_2 = r_3 x_2(t)$ $u_{12}(x_2) = \frac{v_5}{p_{13}}x_2(t) = r_4 x_2(t)$ $u_{13}(x_6) = r_5 x_6(t)$ $u_{14}(x_3) = \frac{v_6}{p_3}x_3(t) = r_6 x_3(t)$ $u_{15}(v_7, x_4) = \frac{v_7(t)}{p_4}x_4(t)$ $u_{16}(x_5) = p_7 x_5(t)$ $u_{17}(x_5) = p_6 x_5(t)$ $u_{18}(R) = R(t)$
Transport processes	$u_4(S, x_1) = p_8 x_1(t)S(t)$ $u_5(S, x_3, x_4) = \frac{p_{11}}{p_{19}+x_3+x_4}x_4(t)S(t)$ $u_6(S, x_4, x_9) = (x_9 - x_4)p_{10}S(t)$ $u_7(S, x_3, x_4) = \frac{p_{12}}{p_{20}+x_3+x_4}x_3(t)S(t)$ $u_8(S, x_3, x_8) = (x_8 - x_3)p_9 S(t)$ $u_{19}(S, x_2, x_7) = p_{14}(x_7 - x_2)S(t)$	$u_4(S, x_1) = p_8 x_1(t)S(t)$ $u_5 = 0$ $u_6(S, x_4, x_9) = (x_9 - x_4)p_{10}S(t)$ $u_7 = 0$ $u_8(S, x_3, x_8) = (x_8 - x_3)p_9 S(t)$ $u_{19}(S, x_2, x_7) = p_{14}(x_7 - x_2)S(t)$

The original model by Desnoues et al. (2018) was composed by a network of 19

reactions and one input function $I(t)$. The latter described the carbon supply from the mother plant to the fruit and it was estimated as the sum of the carbon used for fruit dry mass (DW) increase and the carbon lost by respiration ($R(t)$). Two parameters λ and λ_{suc} described the fraction of the input flow that is converted into the different forms of sugars. Fruit respiration was computed following the growth-maintenance paradigm, as described in (Desnoues et al. 2018).
Reaction rates are reported in Table 2.5. Enzymatic reactions were generally described using an irreversible Michaelis-Menten kinetics, with experimentally-measured capacities $v_i(t)$. Transport processes between cytosol and vacuole were assumed proportional to the vacuole surface (hypothesis of constant density of transporters) computed from vacuole fresh mass (proxy of the volume) supposing the vacuole as a sphere of surface $S(t) = (4\pi)^{\frac{1}{3}}(V_2)^{\frac{2}{3}}$ (see Desnoues et al. (2018) for more information). Both active and passive transport mechanisms were considered for fructose and glucose.
Model equations are reported in Table 2.6, for both the original and the reduced model.

TABLE 2.6 – System of original and reduced models

System of original model	**System of reduced model**
$\frac{dx_1}{dt} = u_1 + \frac{\sigma_5}{\sigma_1} u_{16} - u_{10} - u_4 - \mu(t)x_1$	
$\frac{dx_2}{dt} = u_2 - u_{11} - u_{12} + \frac{1}{\sigma_2 V_1} u_{19} - \mu(t)x_2$	
$\frac{dx_3}{dt} = u_3 + \frac{1}{\sigma_3 V_1} u_8 + \frac{1}{2}\frac{\sigma_1}{\sigma_3} u_9 + \frac{1}{2}\frac{\sigma_1}{\sigma_3} u_{10} + \frac{\sigma_2}{\sigma_3} u_{11} - u_7 - u_{14} - \mu(t)x_3$	$\frac{dx_3}{dt} = u_3 + \frac{1}{\sigma_3 V_1} u_8 + \frac{1}{2}\frac{\sigma_1}{\sigma_3} u_9 + \frac{1}{2}\frac{\sigma_1}{\sigma_3} u_{10} + \frac{\sigma_2}{\sigma_3} u_{11} - u_{14} - \mu(t)x_3$
$\frac{dx_4}{dt} = u_3 + \frac{1}{\sigma_4 V_1} u_6 + \frac{1}{2}\frac{\sigma_1}{\sigma_4} u_{10} + \frac{\sigma_2}{\sigma_4} u_{12} - u_5 - u_{15} - \mu(t)x_4$	$\frac{dx_4}{dt} = u_3 + \frac{1}{\sigma_4 V_1} u_6 + \frac{1}{2}\frac{\sigma_1}{\sigma_4} u_{10} + \frac{\sigma_2}{\sigma_4} u_{12} - u_{15} - \mu(t)x_4$
$\frac{dx_5}{dt} = \frac{1}{2}\frac{\sigma_1}{\sigma_5} u_9 + \frac{\sigma_3}{\sigma_5} u_{14} + \frac{\sigma_4}{\sigma_5} u_{15} - u_{17} - u_{16} - \frac{1}{\sigma_5 V_1} u_{18} - \mu(t)x_5$	$x_5 = \frac{1}{p_6+p_7+\mu(t)}(\frac{1}{2}\frac{\sigma_1}{\sigma_5} u_9 + \frac{\sigma_3}{\sigma_5} u_{14} + \frac{\sigma_4}{\sigma_5} u_{15} - \frac{1}{\sigma_5 V_1} u_{18})$
$\frac{dx_6}{dt} = \alpha u_4 - u_{13} - \mu(t)x_6$	
$\frac{dx_7}{dt} = -\frac{1}{\sigma_7 V_2} u_{19} - \mu(t)x_7$	
$\frac{dx_8}{dt} = \alpha u_7 + \frac{1}{2}\frac{\sigma_6}{\sigma_8} u_{13} - \frac{1}{\sigma_8 V_2} u_8 - \mu(t)x_8$	$\frac{dx_8}{dt} = \frac{1}{2}\frac{\sigma_6}{\sigma_8} u_{13} - \frac{1}{\sigma_8 V_2} u_8 - \mu(t)x_8$
$\frac{dx_9}{dt} = \alpha u_5 + \frac{1}{2}\frac{\sigma_6}{\sigma_9} u_{13} - \frac{1}{\sigma_9 V_2} u_6 - \mu(t)x_9$	$\frac{dx_9}{dt} = \frac{1}{2}\frac{\sigma_6}{\sigma_9} u_{13} - \frac{1}{\sigma_9 V_2} u_6 - \mu(t)x_9$
$\frac{dx_{10}}{dt} = \frac{\sigma_5}{\sigma_{10}} u_{17} - \mu(t)x_{10}$	

Model parameterization and initialization

A total of 23 parameters are needed to fully define the reaction rates of Table 2.5. Following (Desnoues et al. 2018), 9 of these parameters were fixed based on published data, which were obtained from research studies of peach or fruit. The remaining 14 parameters were estimated numerically, as described in section 2.2.5.2. In order to compare model and data, sugar concentrations at the fruit level have to be computed from model variables, describing the metabolite concentration within intra-cellular compartments. Assuming a constant proportion of vacuolar space in fruit cells, the concentration of each sugar j (sucrose, glucose, fructose, and sorbitol) at the fruit level is given by :

$$\mathbb{E}(X_j) = \mathscr{M}_p(x_i) = x_i^{vac}\frac{1}{\alpha+1} + x_i^{cyt}\frac{\alpha}{\alpha+1} \tag{2.24}$$

where x_i^{vac} and x_i^{cyt} are respectively the variables located in the vacuole ($i \in [6,9]$) and cytosol ($i \in [1,5]$) (see Table 2.4) and $\alpha = \dfrac{V_1}{V_2}$ is the intra-cellular volume ratio. The value of α was estimated by cytological analysis to 0.08 (see (Desnoues et al. 2018) for more information). Fruit fresh mass was assumed as a proxy for total volume $V_1 + V_2$.

For each genotype k, initial conditions $x_0^{(k)}$ were set using the concentrations $X^{(k)}(t_0)$ of sucrose, glucose, fructose, sorbitol, and hexose phosphates, measured at the fruit level at the first stage of development. The conversion between total and intra-cellular metabolite concentrations was performed based on metabolite localization at maturity. Accordingly, 98% of fructose, glucose, sucrose content and 90% of sorbitol content were assumed to be located in the vacuole, whereas the hexose phosphates were restricted to the cytosol. Accordingly, for sucrose, fructose, and glucose :

$$\begin{aligned}
\text{cytosol}: &\quad x_i^{(k)}(t_0) = 0.02\, X^{(k)}(t_0)\frac{(1+\alpha)}{\alpha} && i \in \{1,3,4\}\\
\text{vacuole}: &\quad x_i^{(k)}(t_0) = 0.98\, X^{(k)}(t_0)\,(1+\alpha) && i \in \{6,8,9\}
\end{aligned}$$

for sorbitol,

$$\begin{aligned}
\text{cytosol}: &\quad x_i^{(k)}(t_0) = 0.10\, X^{(k)}(t_0)\frac{(1+\alpha)}{\alpha} && i = 2\\
\text{vacuole}: &\quad x_i^{(k)}(t_0) = 0.90\, X^{(k)}(t_0)\,(1+\alpha) && i = 7
\end{aligned}$$

and for the hexoses phosphates

$$\text{cytosol}: \quad x_i^{(k)} = X^{(k)}(t_0)\frac{(1+\alpha)}{\alpha} \qquad i = 10$$

TABLE 2.7 – Table of original and reduced models parameter description

p_i	Parameter	Corresponding model	Description	Reference	Value	Unit
p_1	λ_{Suc}	original and reduced	sucrose proportion hydrolyzed in the apoplasm	Estimated	0-1	
p_8	$TactifSuc$	original and reduced	coefficicent of sucrose transport (active import) from cytosol to vacuole	Estimated	0-400	$mgFW^{-1}day^{-1}$
p_{10}	$TpassifGlu$	reduced	coefficient of glucose passive transport between cytosol and vacuole and in the opposite direction	Section. 2.2.3.1	112.1559	$mgFW^{-1}day^{-1}$
		original		Estimated	0-150	
p_9	$TpassifFru$	original and reduced	coefficient of fructose passive transport between cytosol and vacuole and in the opposite direction	Estimated	0-150	$mgFW^{-1}day^{-1}$
$r_1 = \frac{v_2}{p_5}$	$R_{susy} = \frac{V_{susy}}{K_{susy}}$	reduced	coefficient of the transfer function between sucrose and (fructose + hexoses phosphate) under action of sucrose synthase (susy) enzyme	Section. 2.2.3.2	1.8809	day^{-1}
$r_2 = \frac{v_3}{p_{21}}$	$R_{ni} = \frac{V_{ni}}{K_{ni}}$	reduced	coefficient of the transfer function between sucrose and (glucose +fructose) under action of neutral invertase (ni) enzyme	Section. 2.2.3.2	95.5875	day^{-1}
$r_3 = \frac{v_4}{p_{22}}$	$R_{sdh} = \frac{V_{sdh}}{K_{sdh}}$	reduced	coefficient of the transfer function between sorbitol and fructose under action of sorbitol dehydrogenase (sdh) enzyme	Section. 2.2.3.2	7.1592	day^{-1}
$r_4 = \frac{v_5}{p_{13}}$	$R_{so} = \frac{V_{so}}{K_{so}}$	reduced	coefficient of the transfer function between sorbitol and glucose under action of sorbitol oxydase (so) enzyme	Estimated	0-10	day^{-1}
$r_5 = \frac{v_1}{p_{23}}$	$R_{ai} = \frac{V_{ai}}{K_{ai}}$	reduced	coefficient of the transfer function between sucrose and (glucose +fructose) under action of acid invertase (ai) enzyme	Estimated	0-1	day^{-1}
p_2	Ki_{AI}	original	inhibitor constant of acid invertase	Estimated	0-10	$mggFW^{-1}$
$r_6 = \frac{v_6}{p_3}$	$R_{fk} = \frac{V_{fk}}{K_{fk}}$	reduced	coefficient of the transfer function between fructose and hexoses phosphate under action of fructokinase (fk) enzyme	Estimated	0-5000	day^{-1}
v_7	$Vhk(t)$	reduced	hexokinase activity (hk) to transfer glucose to hexoses phosphate	Section. 2.2.3.2	$86.2 - 2.3t + 2e^{-2}t^2 - 8.3e^{-5}t$	$mgFW^{-1}day^{-1}$
p_4	Khk	original and reduced	hexokinase affinity	Estimated	1-300	$mggFW^{-1}$
p_7	$ReSyntSuc$	original and reduced	coefficient of the transfer function between hexoses phosphate and sucrose	Estimated	0-300	day^{-1}
p_6	$OthComp$	original and reduced	coefficient of the transfer function between hexoses phosphate and other compounds	Estimated	450-1500	day^{-1}
p_{14}	$TpassifSor$	reduced	coefficient of sorbitol passive transport between cytosol and vacuole	Section. 2.2.3.1	4.1305	$mgFW^{-1}day^{-1}$
		original		Estimated	0-150	
σ_f	$PropCdw$	original and reduced	carbon concentration of the mesocarp	Génard et al. (2010)	0.44	$gCgDW^{-1}$
p_{17}	q_g	original and reduced	growth respiration coefficient	Génard et al. (2010)	0.084	$gCgDW^{-1}$
p_{18}	q_m	original and reduced	maintenance respiration coefficient	Génard et al. (2010)	2.76e-4	$gCgDW^{-1}day^{-1}$
p_{16}	Q_{10}	original and reduced	temperature ratio of maintenance respiration	Génard et al. (2010)	1.9	
p_{15}	λ	original and reduced	sucrose sap proportion	Desnoues et al. (2018)	0.65	
p_{11}	$VmtactifFru$	original	fructose active import (activity)	Estimated	0-150	$mgFW^{-1}day^{-1}$
p_{12}	$VmtactifGlu$	original	Glucose active import (activity)	Estimated	0-150	$mgFW^{-1}day^{-1}$
p_{19}	$KmtactifGlu$	original	Glucose active import (affinity)	Shiratake et al. (1997)	0.054	$mggFW^{-1}$
p_{20}	$KmtactifFru$	original	fructose active import (affinity)	Shiratake et al. (1997)	0.288	$mggFW^{-1}$

2.2.8.2 Multi-variate sensitivity analysis

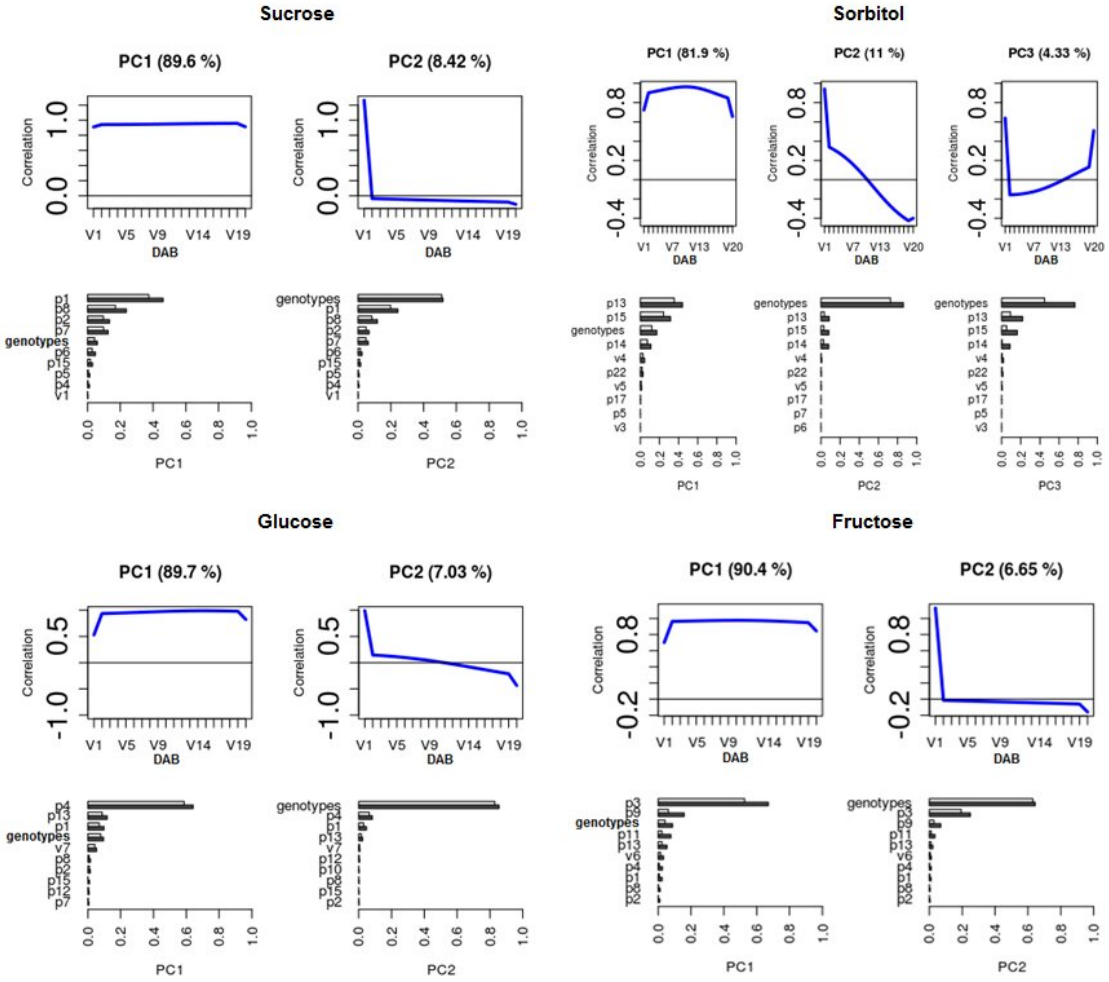


FIGURE 2.12 – PCA-based sensitivity analysis of the sugar model. Columns : principal components. Top row : correlation coefficients (y-axis) between the principal component and the output of each sugar during fruit development (DAB, days after bloom on the x-axis). Bottom row : first order sensitivity indices (dark bars) and total sensitivity indices (pale bars).

Multivariate sensitivity (Lamboni et al. 2009) was used to identify the influence of each parameter on the dynamic output $x(t)$ during fruit development. Where $x(t)$ is the sugar concentration (sucrose, glucose, fructose and glucose) and t is the independent time variable for 20 days after bloom ($t = (V1 = min(DAB), V2 = max(DAB)/2 + 2, \ldots, V19 = max(DAB)/2 + 19, V20 = max(DAB))$). Results of the principal components and sensitivity principal indices are presented in Fig. 2.12. For sucrose, glucose and fructose, the first two components explained 96% of the total inertia of the simulated sugar dynamics. For sorbitol, the first three components

explained 97%. The first component was positively correlated with all time-points. Correlation values in Fig. 2.12 show that the first principal component corresponds to the average concentration of sugars (sucrose, glucose, fructose and sorbitol) produced during the whole fruit development. The second principal component was positively correlated with sugar concentration at stage 1 and poorly or slightly negatively correlated with simulated sugar during the second half of fruit development. Thus, the second principal component corresponds to the difference in sugar initialization values, that strongly depends on the `genotype` factor. For sorbitol, the third principal component accounts for a much smaller part of inertia, associated with the difference between the sorbitol produced in the middle of fruit development and the sorbitol produced both very early and late. It was sensitive to the set of genotypes.

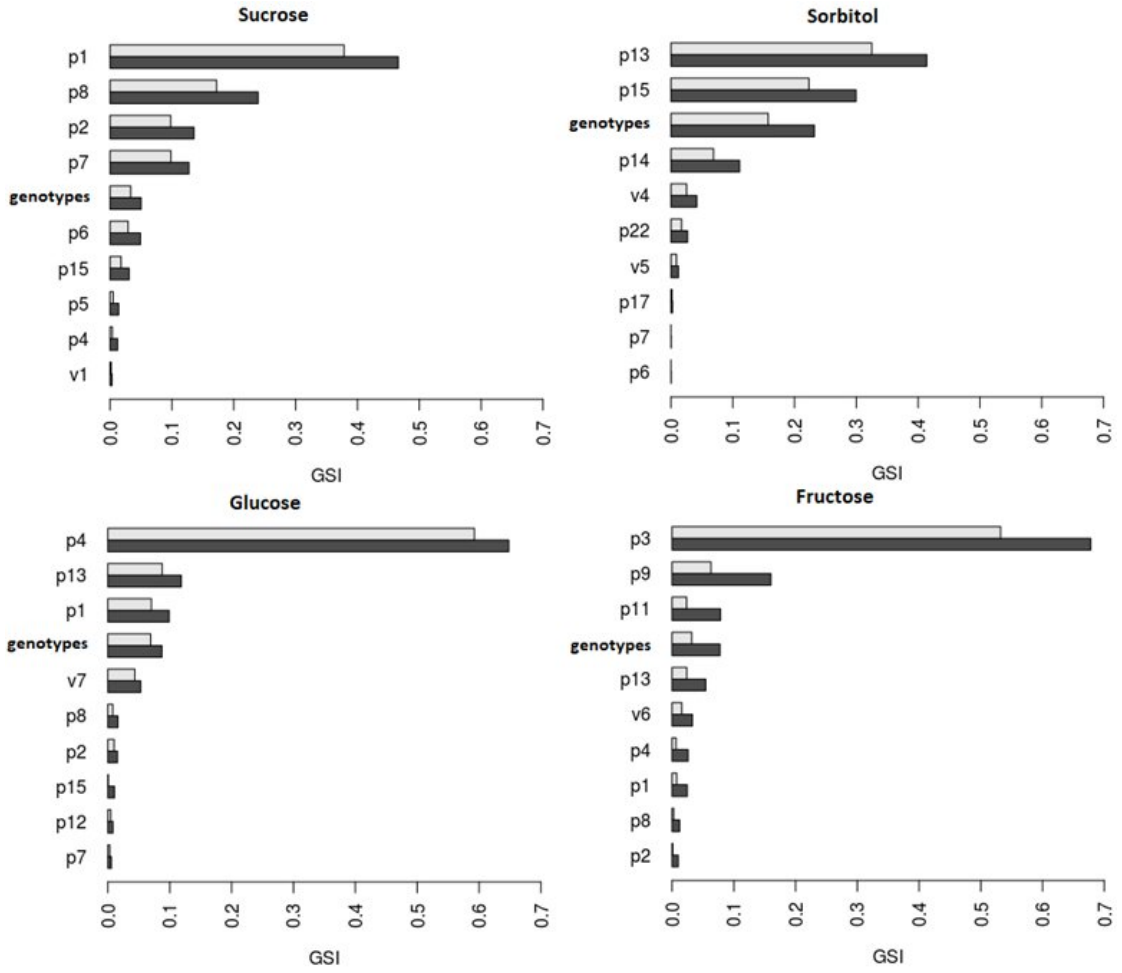


FIGURE 2.13 – Generalized sensitivity indices (GSI) for the first ten sensitive parameters (p_i) and ten genotypes (the training set) on four outputs (Sucrose, Sorbitol, Glucose, and Fructose) of the sugar model. The main sensitivity indices are in dark bars and interaction ones are in grey bars.

The generalized sensitivity indices (GSI) shown in Fig. 2.13 gives a common ranking of model parameters according to their influence on the four output sugars (Sucrose, Sorbitol, Glucose and Fructose), for all tested genotypes. Parameter p_1 related to the action of cell-wall invertase in fruit apoplasm is the most important parameter, for its effect on both sucrose (rank 1) and glucose (rank 3) dynamics. The activities of Fructokinase (p_3) and Hexokinase (p_4) are the most sensitive parameters for fructose and glucose concentrations, respectively, whereas the sorbitol oxydase affinity (p_{13}) and the proportion of sucrose in the plant sap (p_{15}) affect sorbitol content in the fruit. Interestingly, the `genotype` factor is ranked only third to fifth, depending on the sugar, meaning that it does not affect parameters' sensitivity as much as expected. A closer look at the results shows that the choice of the genotype essentially affects the second principal component, via the definition of the initial conditions of the model (see the supplemental information Fig. 2.12).

2.2.8.3 Virtual experiment

In order to evaluate the reliability of the proposed simplifications over a larger diversity, a progeny of virtual genotypes was randomly created based on a careful recombination, with noise, of the original 10 dynamics. This included changes in parameters values, initial conditions and input functions.

We used the results from the principal component analysis (PCA) performed on growth rate and growth duration for the whole progeny of 106 genotypes to verify the distribution of virtual genotypes. To this aim, growth rates and durations of the 20 000 virtual genotypes were projected on the PCA plan defined by the previous analysis. As shown in Fig. 2.14, the virtual genotypes provide a good representation of the diversity in growth rate and growth duration observed in the real progeny.

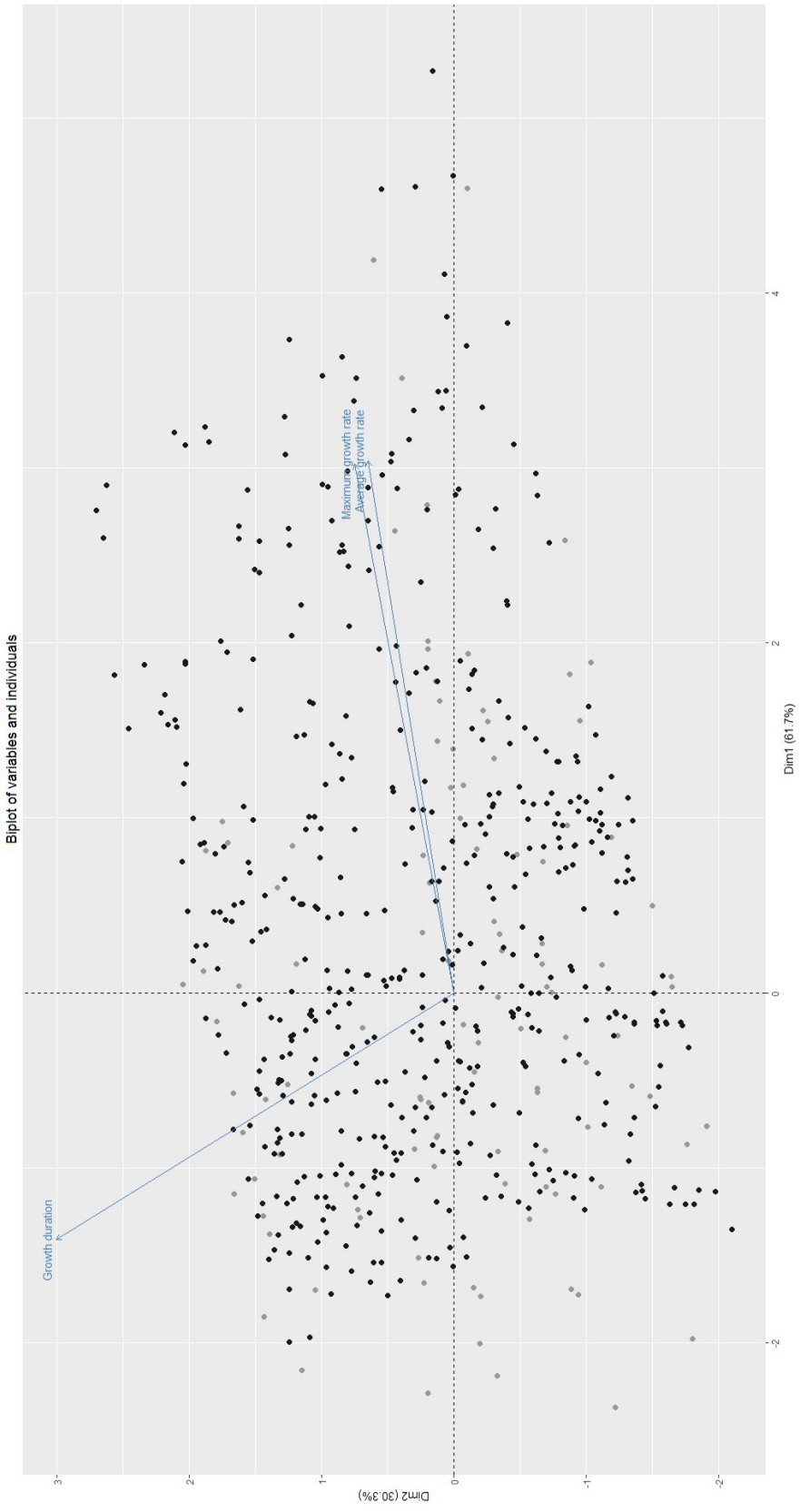


FIGURE 2.14 – Principal component analysis (PCA) for the whole progeny of 106 genotypes (grey) and 500, out of 20000, virtual genotypes (black). Represents the projection on the Dim1 and Dim2 of the growth duration and growth rate obtained with curves growth.

2.2.8.4 Time-scale analysis and QSSA

Timescale-based approaches and quasi-steady-state approximation (Heinrich et al. 1996) were applied to reduce the number of ODEs of the model and to obtain the final reduced model. The predicted concentrations of sugars in both intracellular compartments were analyzed. A fast transient dynamics of the concentration of the hexose phosphate (x_5), followed by a slow one, was observable in the numerical simulations of the original and intermediate reduced model (Fig. 2.15. Together with the analysis of the Jacobian matrix, this observation led to the assumption of x_5 as a fast variable of the system. Quasi-steady-state approximation on x_5, indeed, strongly reduced the fast transient dynamics in the final reduced model, for most genotypes. Notice that a few fast modes (already pointed out by the analysis of the Jacobian matrix) may nonetheless remain in the system. Their elimination would require a linear combination of the original variables, which is incompatible with our objective to preserve the biological interpretation of the model. We therefore decided not to push the simplification of the model further.

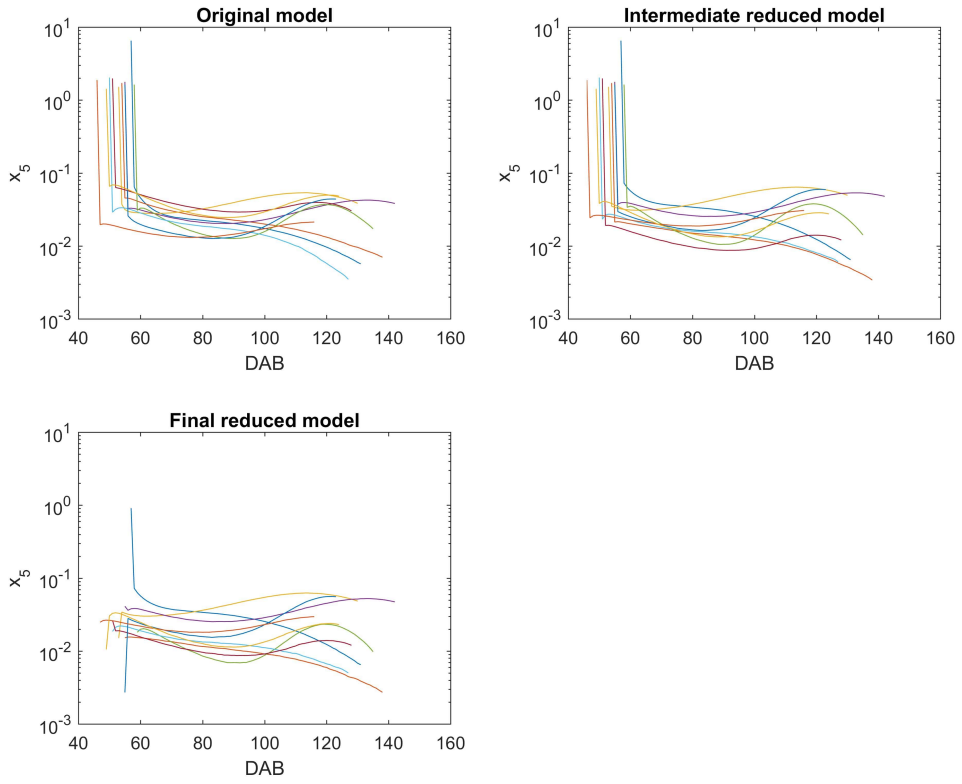


FIGURE 2.15 – Evolution of the concentration ($mggFW^{-1}$) of x_5 : $Hexose\ Phosphate$ during fruit development (DAB, days after bloom) for ten genotypes for the original, intermediate reduced and final models.

Results of quasi-steady-state approximation applied on the intermediate reduced model for the 20 000 virtual genotypes

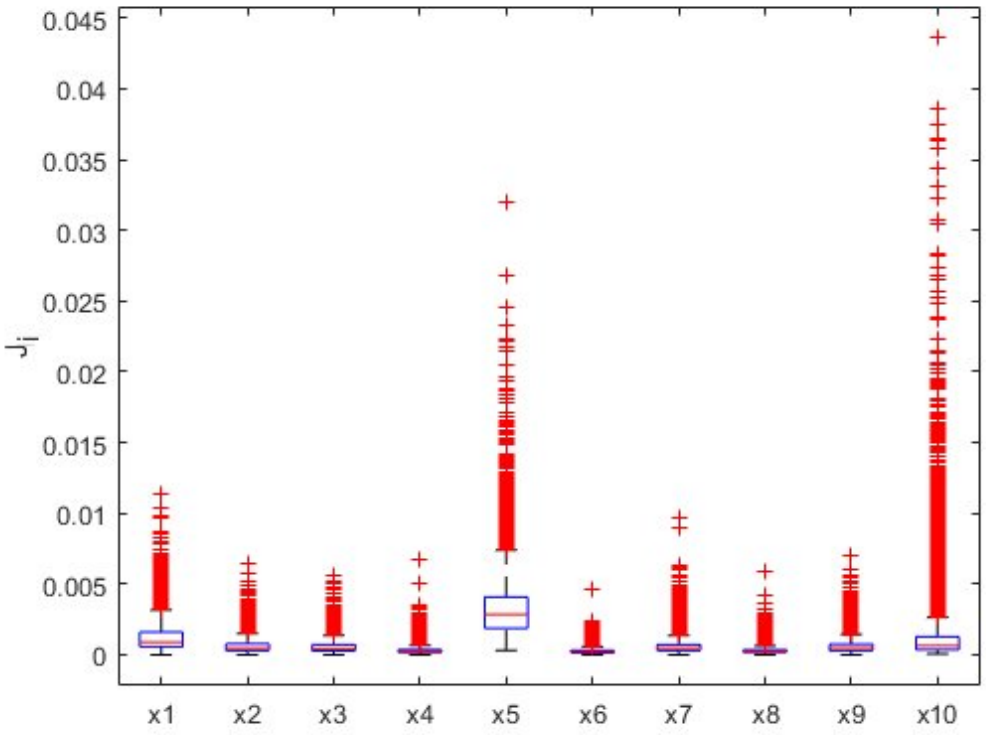


FIGURE 2.16 – Normalized Root Mean Square Errors $J_i, i \in \{1,\ldots,10\}$ between the intermediate and reduced model after application of the QSSA to x_5. The boxplot shows the variability of J_i over the virtual genotypes

We compared the intermediate reduced model with its QSS approximation by calculating the Normalized Root Mean Square Error (J_i) on the 20 000 virtual genotypes. All J_i was very low, less than 0.045, over the whole dynamics for all variables (Fig. 2.16).

TABLE 2.8 – NRMSE between model simulation and experimental data. Calculated values of the normalized root mean squared error (NRMSE) are presented for each genotype, the four sugars separately.

	Genotype	Phenotype	Year	Sucrose	Sorbitol	Fructose	Glucose	Mean
Trainig set	E1	Standard	2012	0.09	0.14	0.16	0.26	0.16
	E33	Standard	2012	0.04	0.19	0.27	0.17	0.16
	E43	Standard	2012	0.07	0.11	0.24	0.16	0.14
	F111	Standard	2012	0.13	0.13	0.21	0.23	0.17
	C227	Standard	2011	0.07	0.28	0.16	0.13	0.16
	E22	Low	2012	0.11	0.09	0.21	0.18	0.14
	F106	Low	2012	0.07	0.7	0.14	0.15	0.26
	F146	Low	2012	0.05	0.14	0.02	0.17	0.10
	H191	Low	2012	0.07	0.15	0.18	0.16	0.14
	C216	Low	2011	0.08	0.26	0.25	0.11	0.17
Validation set	H163	Standard	2012	0.10	0.35	0.18	0.19	0.20
	F107	Standard	2012	0.11	0.11	0.24	0.25	0.17
	E23	Standard	2012	0.15	0.30	0.10	0.17	0.18
	E17	Standard	2012	0.14	0.37	0.19	0.05	0.18
	E21	Standard	2012	0.13	0.21	0.16	0.24	0.18
	E41	Low	2012	0.08	0.35	0.35	0.24	0.25
	E18	Low	2012	0.13	0.26	0.26	0.09	0.18
	F113	Low	2012	0.12	0.32	0.35	0.20	0.24
	F90	Low	2012	0.06	0.47	0.26	0.13	0.23
	C243	Low	2012	0.18	0.49	0.24	0.09	0.25
	C199	Low	2012	0.22	0.46	0.25	0.19	0.28
	C207	Low	2012	0.19	0.28	0.23	0.24	0.23
	E36	Low	2012	0.09	0.13	0.13	0.16	0.12
	E48	Low	2012	0.07	0.41	0.14	0.14	0.19
	F101	Low	2012	0.15	0.43	0.20	0.05	0.20
	F109	Low	2012	0.07	0.31	0.27	0.24	0.22
	F127	Low	2012	0.09	0.17	0.22	0.10	0.14
	F144	Low	2012	0.19	0.11	0.24	0.22	0.19
	F141	Low	2012	0.14	0.22	0.36	0.16	0.22
	F86	Low	2012	0.06	0.13	0.30	0.32	0.20
	C232	Standard	2012	0.05	0.30	0.25	0.13	0.18
	E5	Standard	2012	0.22	0.07	0.12	0.24	0.16
	E19	Standard	2012	0.05	0.14	0.24	0.03	0.11
	E20	Standard	2012	0.19	0.17	0.07	0.18	0.11
	E26	Standard	2012	0.05	0.20	0.18	0.38	0.20
	E34	Standard	2012	0.22	0.43	0.20	0.15	0.25
	E35	Standard	2012	0.11	0.27	0.19	0.07	0.16
	E37	Standard	2012	0.24	0.24	0.13	0.26	0.21
	F83	Standard	2012	0.09	0.19	0.06	0.27	0.15
	F115	Standard	2012	0.03	0.20	0.10	0.07	0.10

The NRMSE can defined as follows :

$$NRMSE(\tilde{p}^{(k)}) = \sum_{i=1}^{4} J_i(\tilde{p}^{(k)})$$

with

$$J_i(\tilde{p}^{(k)}) = \frac{\sqrt{\frac{1}{N_M}\sum_{j=1}^{N_M}(\tilde{x}_i(t_j,\tilde{p}^{(k)}) - X_i^{(k)}(t_j))^2}}{max_j(X_i^{(k)}(t_j)) - min_j(X_i^{(k)}(t_j))} \tag{2.25}$$

where N_M is the number of observations, $\tilde{x}(t,\tilde{p}^{(k)})$ are the concentrations predicted by the model and $X^{(k)}t(t)$ are the experimental data and i is the sugar index.

2.3 Conclusions et perspectives

Conclusions

- Le modèle réduit ne comporte que 9 paramètres à estimer, des flux linéaires, 9 EDOs et une seule capacité enzymatique temporelle, commune à tous les génotypes
- De nouveaux génotypes de la descendance interspécifique du pêcher, pour lesquels peu de données sont disponibles, ont pu être calibrés
- Les résultats ont montré un accord satisfaisant entre le modèle et les données expérimentales, ouvrant de nouvelles perspectives prometteuses pour les études génétiques et la sélection virtuelle

Perspectives

Le modèle réduit avec une structure plus simple permettra

- une meilleure exploration de l'espace des paramètres et une estimation plus précise des valeurs des paramètres
- d'ouvrir la possibilité d'explorer de grandes populations génétiques
- d'effectuer de nombreuses simulations sur un grand nombre de conditions environnementales et/ou de scénarios climatiques

3 Estimation des paramètres du modèle cinétique réduit pour une population génétique de pêchers

Sommaire

DANS LE CHAPITRE précédent, nous avons présenté une stratégie pour simplifier le modèle développé par Desnoues et al. (2018), permettant d'obtenir un modèle réduit tout en conservant les principales informations biologiques. Dans ce chapitre, nous nous intéressons à la calibration de ce modèle réduit (Kanso et al. 2020) sur l'ensemble d'une population de 106 génotypes.

La première partie du chapitre décrit des travaux complémentaires effectués autour du modèle réduit visant à : i) vérifier son identifiabilité structurelle, ii) améliorer la description de la fonction d'entrée sur l'ensemble de la population, et iii) étudier la sensibilité de différentes fonctions objectif aux paramètres à estimer.

La seconde partie constitue le coeur du chapitre. Elle est rédigée sous forme d'un article en préparation et décrit la stratégie utilisée pour calibrer le modèle et estimer les paramètres génotype-dépendants. Deux approches ont été testées : i) la calibration du modèle pour chaque génotype indépendamment, ii) la calibration du modèle pour tous les génotypes simultanément en utilisant une approche basée sur la population proposée par Baey et al. (2018). Les performances de ces deux stratégies de calibration ont été comparées afin d'identifier la stratégie adaptée pour une estimation fiable et robuste des paramètres qui feront l'objet d'une analyse du contrôle génétique.

3.1 Travaux complémentaires autour du modèle réduit du métabolisme des sucres chez la pêche

La stratégie développée au Chapitre 2 a permis d'obtenir un modèle cinétique à la fois plus économe en nombre de variables et de paramètres mais aussi mathématiquement moins complexe que le modèle original. En particulier, les flux sont désormais décrits par des fonctions linéaires du substrat et les activités enzymatiques sont supposées constantes dans le temps, à l'exception de la capacité enzymatique de l'hexokinase (HK). Ces simplifications réduisent considérablement la complexité du système EDO et ouvrent la voie à son application à un large nombre de génotypes,

une étape essentielle à l'analyse du contrôle génétique du métabolisme des sucres chez la pêche.

En raison du caractère plus appliqué des chapitres suivants, nous avons opté pour une nouvelle notation des paramètres, moins abstraite et plus lisible pour un public biologiste. Le tableau 3.1 illustre la correspondance entre la notation du Chapitre 2 et celle des Chapitres 3 et 4.

TABLE 3.1 – Tableau de correspondance avec les paramètres du chapitre 2

Paramètre Chapitre 2	Nouveau Paramètre	Description
p_1	LHx	fraction de saccharose hydrolysé dans l'apoplasme
p_3	KFk	affinité de la fructokinase (FK)
p_4	KHk	affinité de l'hexokinase (HK)
p_6	OCp	taux de conversion des hexoses phosphates en d'autres composés (parois, acides..)
p_7	RSS	taux de conversion des hexoses phosphates en saccharose
p_8	TAS	coefficient de transport du saccharose (actif) du cytosol à la vacuole
p_9	TPF	coefficient de transport du fructose (passif) entre cytosol et vacuole
r_4	RSO	taux de conversion du sorbitol en glucose sous l'action de la sorbitol oxydase (SO)
r_5	RAI	taux de conversion du saccharose en glucose et fructose sous l'action de l'invertase acide (AI)

3.1.1 Identifiabilité structurelle du modèle réduit

Avant de procéder à la calibration du modèle, il est important de vérifier si celui-ci est structurellement identifiable *i.e.* s'il est en principe possible de déterminer les valeurs des paramètres à partir de l'observation (supposée infiniment précise) de ses sorties et de la connaissance de ses équations. Cette propriété est donc préliminaire et nécessaire avant toute estimation sur des données réelles, pour lesquelles des problèmes d'identifiabilité pratique, liés à la qualité et à la quantité des observations, peuvent aussi se poser.

Plusieurs méthodes existent pour analyser l'identifiablité structurelle d'un modèle (voir Section 1.4.1). Dans ce travail, nous avons utilisé l'outil logiciel STRIKE-GOLDD (STRuctural Identifiability taKen as Extended-Generalized Observability with Lie Derivatives et la décomposition) développé par Villaverde et al. (2016).

Cet outil met en oeuvre une méthodologie d'analyse d'identifiabilité structurelle locale dédiée aux systèmes non linéaires au sens large, y compris les systèmes non rationnels, et basée sur le concept d'observabilité d'un système (Hermann et al. 1977) *i.e.* la possibilité de définir son état interne à partir de la mesure de ses sorties, en

un temps fini. Pour l'analyse d'identifiablité, le concept d'état interne est élargi aux paramètres, considérés comme des états à dynamique nulle. L'analyse d'observabilité-identifiabilité qui en découle permet notamment de savoir si le modèle est structurellement identifiable, et, le cas échéant, d'identifier les paramètres n'étant pas identifiables individuellement. Sous certaines conditions, la méthode permet aussi de suggérer des combiner des paramètres non identifiables en combinaisons qui pourraient être identifiables.

D'un point de vue pratique, l'application de la méthode STRIKE-GOLDD passe par la définition du modèle sous la forme suivante :

$$M : \begin{cases} \dot{x}(t) & = f[x(t), u(t), p], \\ y(t) & = g[x(t), p], \\ x_0 & = x(t_0) \end{cases}$$

où $x(t)$ est le vecteur de variables du système, $u(t)$ le vecteur des fonctions d'entrée, et p le vecteur des paramètres du modèle. $y(t)$ définit les sorties du modèle pouvant être mesurées, x_0 est la condition initiale. Dans notre cas, le modèle réduit contient 9 variables, 9 paramètres à estimer et 3 fonctions d'entrée correspondant aux masses fraîche (FW) et sèche (DW) du fruit, et à la dérivée de cette dernière (dDW), nécessaire au calcul de l'entrée de carbone dans le fruit. Les observations $y(t)$ sont les concentrations des 4 sucres à l'échelle du fruit entier, moyenne des concentrations dans le cytosol et dans la vacuole pondérées par les volumes intracellulaires, auxquelles s'ajoutent, en vertu du bilan de masse, la somme de deux autres variables, les hexoses phosphates (variable x_5, à l'état quasi-stationnaire) et les autres composés carbonés (parois, acides organiques etc, variable x_{10}). La sortie du logiciel résume l'identifiabilité du modèle. Il s'avère que toutes les variables sont observables et que les 9 paramètres de notre modèle réduit sont structurellement identifiables.

La conclusion de cette étude nous permet donc d'aborder l'étape d'identifiabilité pratique des paramètres, qui fera l'objet de la Section 3.2 de ce chapitre. Mais avant nous devons améliorer la description de la fonction d'entrée sur l'ensemble de la population.

3.1.2 Modélisation de la dynamique de croissance de la chair du fruit

Comme évoqué ci-dessus, la fonction de croissance du fruit est utilisée comme entrée du modèle cinétique du métabolisme des sucres. Spécifique à chaque génotype, elle détermine la quantité de carbone qui rentre dans le fruit en provenance de l'arbre (croissance en matière sèche) et elle est utilisée pour l'estimation des volumes et de la croissance des compartiments intracellulaires (croissance en matière fraîche). Le choix de la fonction de croissance joue donc un rôle important dans la caractérisation de la variabilité phénotypique entre génotypes et peut avoir un impact fort sur les prédictions du modèle dynamique.

Dans l'article original (Desnoues et al. 2018), la dynamique de croissance de la chair a été décrite à l'aide de fonctions double-sigmoïde ce qui impose l'estimation de 6 paramètres pour la masse fraîche et 6 pour la masse sèche. L'estimation de ces paramètres était possible car pour chaque génotype (10 génotypes considérés dans l'étude initiale), 3 réplicats de mesures étaient disponibles pour chaque stade de développement. Or, cette stratégie n'est pas adaptée aux autres génotypes de la population pour lesquels on dispose d'un nombre réduit de mesures (1 seul réplicat par stade de développement). L'objectif de cette section est donc de trouver une représentation de la dynamique de croissance du fruit qui soit à la fois moins coûteuse en nombre de paramètres et suffisamment souple pour décrire l'ensemble des dynamiques observées. Dans la suite, deux stratégies sont proposées et comparées sur un échantillon de 10 génotypes de la population, pour lesquels on dispose d'un nombre réduit de mesures.

Données Expérimentales

La teneur en matière sèche de la chair et sa masse fraîche ont été mesurées expérimentalement pour tous les génotypes de la population, à six stades de développement du fruit. A partir de ces données, la masse sèche de la chair a ensuite été calculée.

La Figure 3.1 montre la variabilité de dynamiques de croissance observées sur les 106 génotypes (courbes grises). La croissance des fruits varient significativement à la fois en vitesse et en durée de croissance, résultant en un large éventail de masses fraîches finales.

Malgré une tendance à l'augmentation pour certains génotypes, la teneur en masse sèche (voir Figure 3.2) reste comprise entre 0.10 et 0.15 pour la plupart des génotypes. Sur la base de ces données, 10 génotypes (courbes noires) ont été sélectionnés pour représenter la variabilité de dynamiques observées et utilisés dans cette étude. Il s'agit des génotypes E17, E21, E23, F107 et H163 et E21 ayant un phénotype standard, et les génotypes C243, E18, E41, F90 et F113 caractérisés par un phénotype pauvre en fructose.

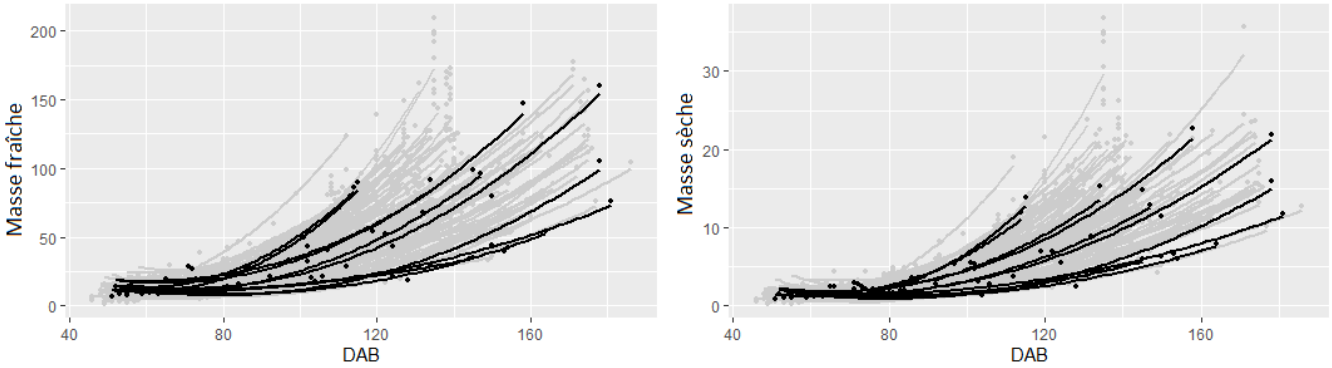


FIGURE 3.1 – Évolution des masses fraîche (gauche) et sèche (droite) de la chair dans le temps (Jours après floraison, DAB) pour les 106 génotypes de la population. En noir, les 10 génotypes utilisés dans cette étude (C243, E17, E18, E21, E23, E41, F90, F107, F113 et H163).

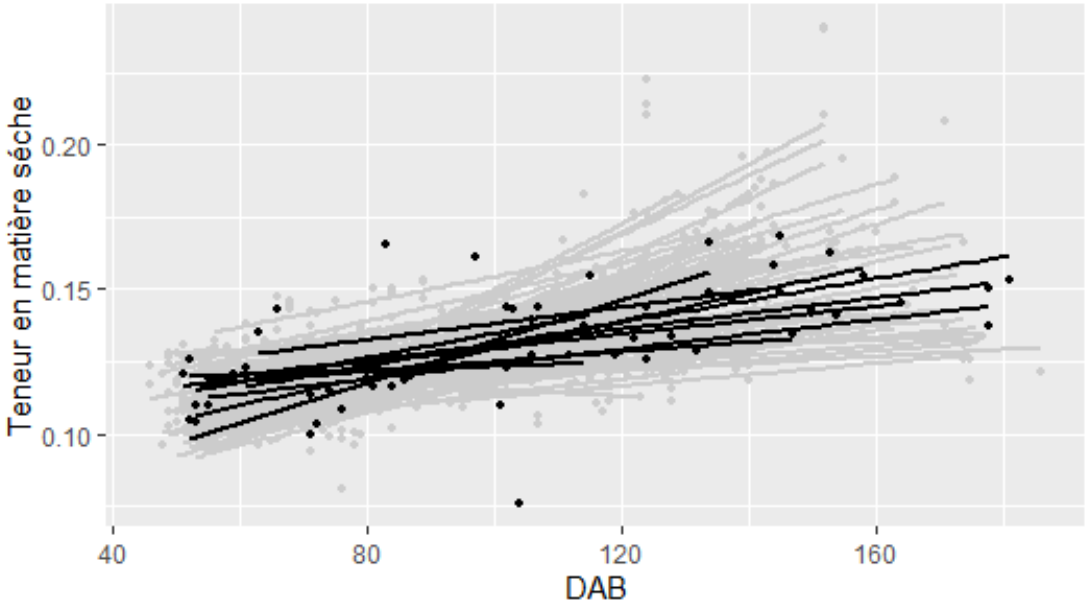


FIGURE 3.2 – Teneur en matière sèche de la chair mesurée pour 106 génotypes durant la croissance du fruit en jours après floraison (DAB). En noir, les 10 génotypes étudiés dans la suite.

Choix de la fonction de croissance

Stratégie 1

La première stratégie consiste à décrire les masses fraîches et sèches de la chair en utilisant des fonctions logistiques simples, chacune impliquant 3 paramètres à

estimer.
La courbe pour la masse fraîche a été définie comme (3.1) :

$$FW(t) = FW_0 + \frac{a_1}{1 + e^{-a_2(t-a_3)}} \tag{3.1}$$

où $FW(t)$ représente la masse fraîche du fruit au temps t, FW_0 la masse fraîche à la date initiale et t le temps en jours après la floraison.
La masse sèche $DW(t)$ de la chair a été ajustée en utilisant la fonction (3.2) :

$$DW(t) = DW_0 + \frac{a_4}{1 + e^{-a_5(t-a_6)}} \tag{3.2}$$

où DW_0 est la masse sèche à la date initiale.

TABLE 3.2 – Paramètres des courbes de croissance pour la stratégie 1

a_i	Value	Description
a_1	$0-300$	masse fraîche finale de la chair
a_2	$0-1$	taux de croissance de la chair en masse fraîche
a_3	$0-150$	date d'occurrence de la vitesse maximale de croissance en masse fraîche
a_4	$0-150$	masse sèche finale de la chair
a_5	$0-1$	taux de croissance de la chair en masse sèche
a_6	$0-50$	date d'occurrence de la vitesse maximale en masse sèche

Stratégie 2

Dans la deuxième stratégie, l'évolution de la masse fraîche de la chair a été modélisée comme dans l'article original, en utilisant une double sigmoïde (3.3 proposée par Génard et al. (1991)) :

$$FW(t) = FW_0 + a_1(1 - e^{-a_2 t}) + \frac{a_3}{1 + e^{-a_4(t-a_5)}} \tag{3.3}$$

où $FW(t)$, FW_0, et t ont la même signification que dans la fonction 3.1.

La teneur en matière sèche, en revanche, est supposée constante et considérée comme un paramètre à estimer (a_6). La masse sèche de la chair est donc proportionnelle à la masse fraîche :

$$DW(t) = a_6.FW(t) \tag{3.4}$$

TABLE 3.3 – Paramètres de la deuxième courbe de croissance

b_i	Value	Description
a_1	$0-50$	masse fraîche cumulée de la chair pour la première phase de croissance
a_2	$0-0.1$	taux de croissance relatif initial de la chair
a_3	$0-300$	masse fraîche finale de la chair
a_4	$0-1$	taux de croissance relatif de la chair dans la deuxième phase de croissance
a_5	$0-150$	date d'occurrence de la vitesse maximale de croissance dans la deuxième phase
a_6	$0-1$	teneur en matière sèche de la chair

Estimation des paramètres et sélection de la fonction de croissance

La qualité de l'ajustement aux données expérimentales obtenue selon les 2 stratégies ci-dessus a été comparée sur un panel de 10 génotypes.

La calibration a été effectuée en utilisant simultanément les données de masses fraîche et sèche de la chair au cours du temps. Nous notons $W^{(k)} = (W_1^{(k)}, \ldots, W_N^{(k)})$ le vecteur d'observations expérimentales des masses à plusieurs dates pour le génotype k, où $W^{(k)} = [FW^{(k)}, DW^{(k)}]$.
Supposons que :

$$W^{(k)} = w^{(k)}(a^{(k)}) + e_k \tag{3.5}$$

où $w^{(k)}(a^{(k)})$ est la masse prédite (fraîche et sèche), $a^{(k)}$ l'ensemble des paramètres à estimer et e_k est l'erreur résiduelle pour le génotype k qui suit une distribution normale avec une moyenne 0 et une variance σ_k^2. L'estimation des paramètres peut être réalisée par la maximisation de la vraisemblance, comme dans le chapitre précédent, en supposant σ_k constante et égale entre masses fraîche et sèche de la chair.

Dans ce cas, l'estimateur du maximum de log-vraisemblance est donc équivalent à l'estimateur ordinaire des moindres carrés :

$$\hat{a}^{(k)} = \underset{a^{(k)}}{\arg\min} \left\{ \sum_{i=1}^{N} (W_i^{(k)} - w_i^{(k)}(a^{(k)}))^2 \right\} \tag{3.6}$$

$$\hat{\sigma}_k^2 = \frac{1}{N} \sum_{i=1}^{N} (W_i^{(k)} - w_i^{(k)}(\hat{a}^{(k)}))^2 \tag{3.7}$$

La fonction ga (Goldberg 1989 de Global Optimisation Toolbox) du logiciel Matlab (MATLAB R2018a, The MathWorks Inc., Natick, MA) a été utilisée pour l'estimation des paramètres, avec le même paramétrage que dans le Chapitre 2.

Le critère d'information d'Akaike (AIC_c) (voir Section 1.4.3) et l'erreur quadratique moyenne normalisée (NRMSE) Eq. 2.25 sont calculés comme des indicateurs de sélection de la meilleure stratégie .

Résultats

La différence du critère d'Akaike $\Delta_{AIC_c} = {AIC_c}^{S2} - {AIC_c}^{S1}$ entre les deux stratégies $S1$ et $S2$ a été calculée pour chacun des 10 génotypes. Les résultats présentés dans la Figure 3.3 montrent que $S2$ est bénéfique pour huit des dix génotypes, avec des valeurs Δ_{AIC_c} largement négatives, approximativement neutre pour un génotype ($\Delta_{AIC_c} \approx 0$) et $S1$ est bénéfique pour un seul génotype.

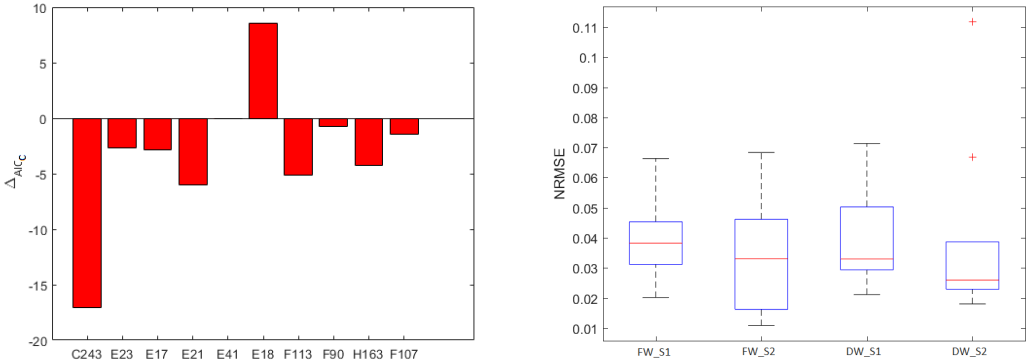


FIGURE 3.3 – Gauche : Δ_{AIC_c} calculé entre les deux stratégies pour chacun des dix génotypes. Droite : NRMSE : Erreur quadratique moyenne normalisée entre la courbe ajustée et l'observation pour FW and DW pour dix génotypes.

La moyenne des NRMSE pour les deux stratégies est proche et se situe entre 2% et 4%, respectivement pour les masses fraîche FW et sèche DW de chair. L' accord entre les prédictions et les observations est donc satisfaisant pour la plupart des génotypes, indépendamment de la stratégie.

A la lumière de ces deux indicateurs et des profils simulés (voir Figure 3.4), $S2$ (Eq. 3.3 et Eq. 3.4) est sélectionnée pour la suite de la thèse.

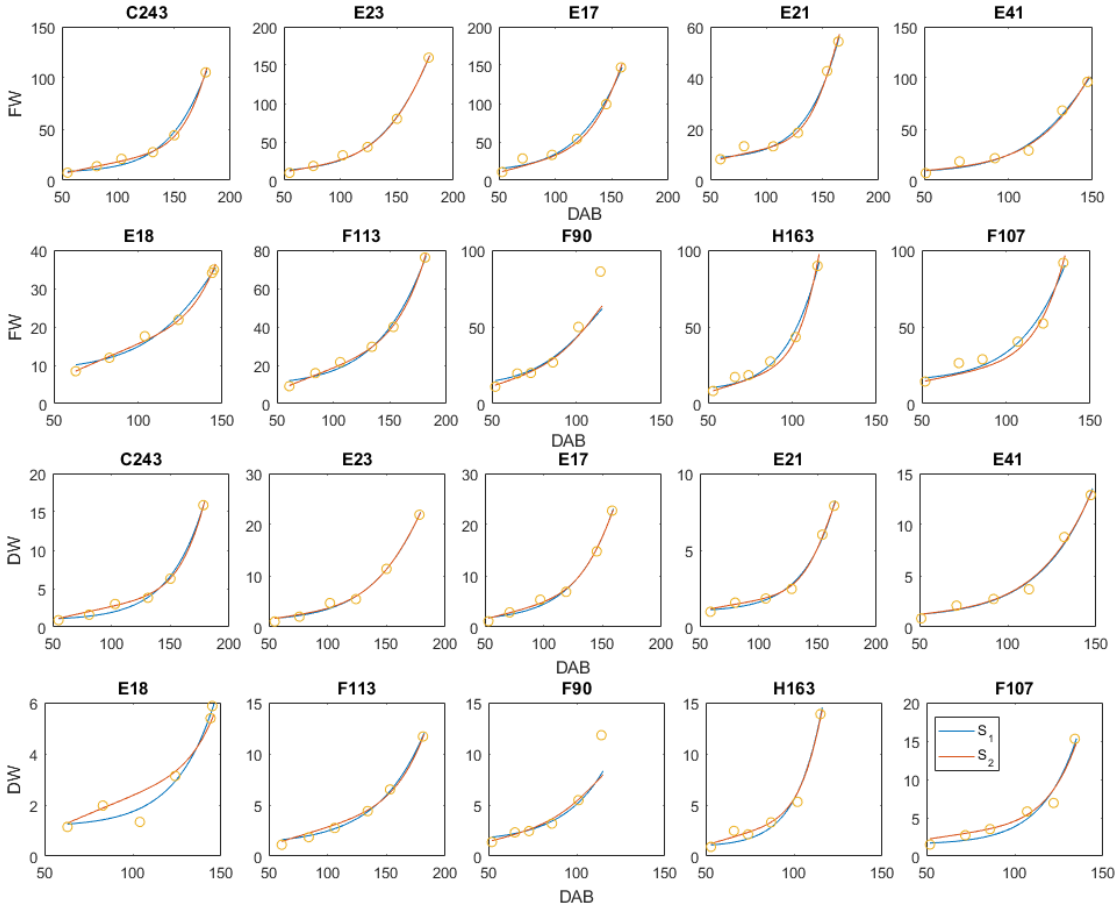


FIGURE 3.4 – Évolution des masses fraîche (FW) et sèche (DW) de la chair au cours de la croissance du fruit, en jours après floraison (DAB), pour les dix génotypes étudiés. Les cercles représentent les données expérimentales et les lignes sont les simulations de FW et DW en utilisant les stratégies 1 (bleu) et 2 (rouge).

3.1.3 Analyse de sensibilité sur la fonction objectif

Dans les chapitres précédents, comme dans l'article original (Desnoues et al. 2018), nous avions fait le choix d'estimer les paramètres du modèle génotype par génotype, en combinant les données des quatre sucres (saccharose, sorbitol, glucose et fructose) dans une seule fonction objectif, du type :

$$f(y^{(k)}, p^{(k)}) = \sum_{j=1}^{N_{SG}} \sum_{i=1}^{N_j^{(k)}} \left(y_{ij}^{(k)} - \hat{y}_{ij}^{(k)}(p^{(k)}) \right)^2 \tag{3.8}$$

où $y_{ij}^{(k)}$ et $\hat{y}_{ij}^{(k)}$ représentent respectivement la mesure et la prédiction de la concentration en sucre j ($j = 1,\ldots,N_{SG}$) avec $N_{SG} = 4$, au stade i ($i = 1,\ldots,N_j^{(k)}$) pour le génotype k.

Dans cette section, nous essayons de prendre un peu de recul par rapport à ce choix en considérant d'autres définitions de fonction objectif et en étudiant leur sensibilité aux paramètres du modèle. En particulier nous considérons la possibilité de travailler sur chaque sucre indépendamment, en définissant quatre objectif séparés :

$$f_j(y_j^{(k)}, p^{(k)}) = \sum_{i=1}^{N_j^{(k)}} \left(y_{ij}^{(k)} - \hat{y}_{ij}^{(k)}(p^{(k)}) \right)^2 \tag{3.9}$$

Une analyse de sensibilité globale basée sur la méthode de Sobol (Saltelli et al. 2008) a été réalisée sur la fonction 3.8 et sur les quatre fonctions 3.9 afin de déterminer les paramètres qui ont le plus d'influence sur la fonction objectif et donc sur le processus d'estimation (Mathieu et al. 2018).

Plan d'expérience

L'analyse de sensibilité a été appliquée aux 9 paramètres du modèle réduit (Kanso et al. 2020). Chaque paramètre a été étudié à trois niveaux, correspondant aux quantiles 0.05, 0.5 et 0.95 des valeurs estimées au chapitre 2. Comme précédemment, afin d'évaluer l'impact du choix du génotype sur les résultats de l'analyse de sensibilité, des simulations ont été effectuées selon un plan factoriel, suivant le modèle ANOVA $genotypes \times (LHx + \ldots + RSO)^2$. Le paquet "Planor" de **R** (R Development Core Team 2015) a été utilisé, avec une résolution maximale de 5, pour un total de $106 \times 3^5 = 25\ 758$ simulations. Matlab software (MATLAB R2018a, The MathWorks Inc., Natick, MA) a été utilisé pour l'intégration du modèle (solver ode23tb Hosea et al. (1996)).

Résultats

La Figure 3.5 montre les indices de sensibilité des 9 paramètres du modèle, pour les fonctions 3.9 pour les quatre sucres pris individuellement.

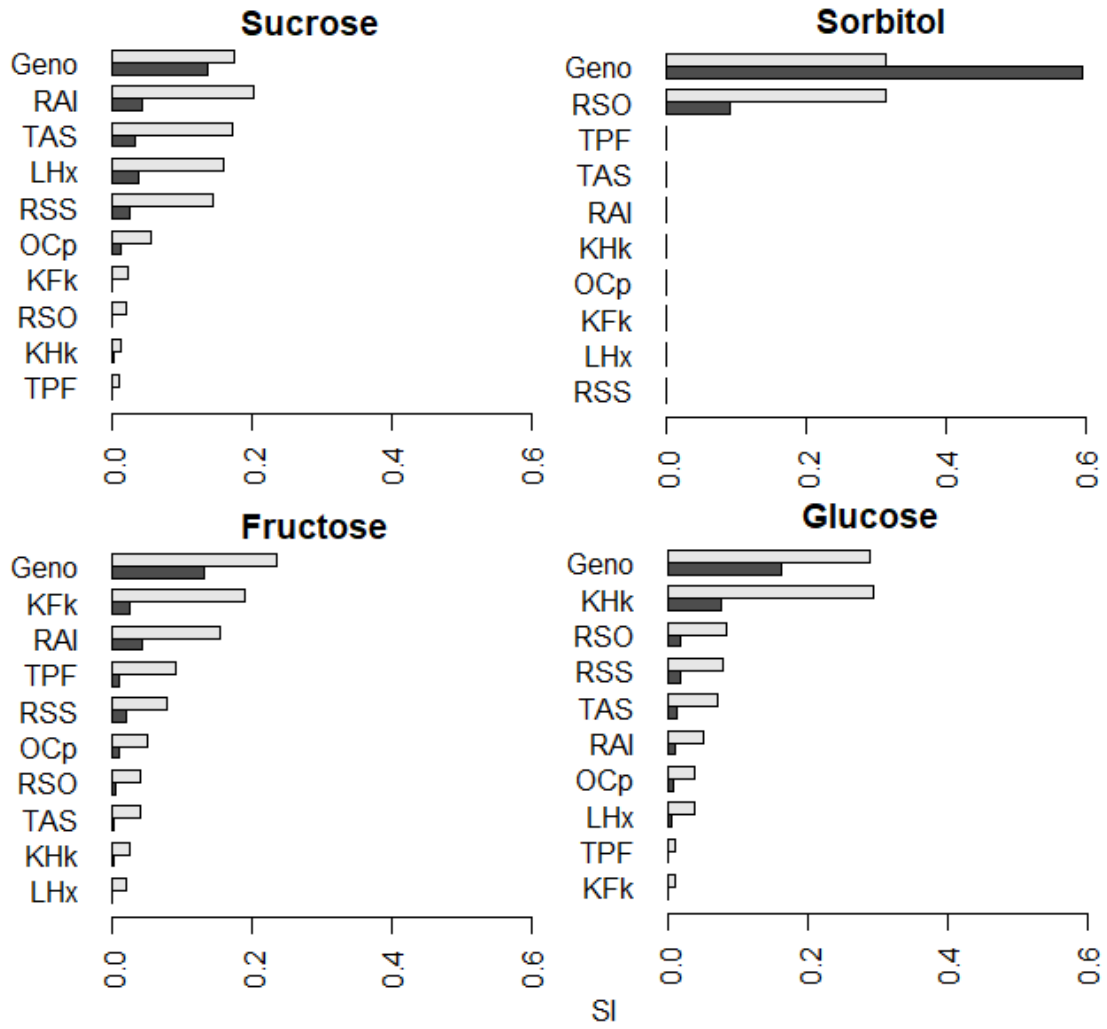


FIGURE 3.5 – Indices de sensibilité (SI) de la fonction objectif de chaque sucre (saccharose, sorbitol, glucose et fructose) pour chacun des 9 paramètres du modèle. Les indices principaux de sensibilité sont représentés par les barres noires et les indices d'interaction sont représentés par les barres grises.

On observe que les paramètres les plus influents dépendent du sucre considéré. Pour la fonction objectif du sorbitol, il était prévisible que seul le paramètre RSO ait une influence. Cependant, pour les 3 autres sucres, des différences notables sont observées. La fonction objectif du glucose dépend essentiellement du paramètre KHk. Celle du saccharose dépend principalement de quatre paramètres RAI, TAS, LHx et RSS ayant tous un fort effet d'interaction. Enfin, pour le fructose la sensibilité de la fonction objectif est répartie d'une façon plus uniforme entre les paramètres, avec une prédominance pour les paramètres KFk et RAI. Enfin, le génotype (via les fonctions d'entrée au modèle) a un impact très important sur la sensibilité, quelque soit le sucre considéré.

La Figure 3.6 montre les résultats de l'analyse de sensibilité pour la fonction objectif agrégée définie comme la somme quadratique des écarts pour les quatre sucres (Eq. 3.8). En comparant les deux figures, on constate que la sensibilité de la fonction agrégée est dominée par les paramètres du saccharose, le sucre majoritaire dans le fruit de pêche.

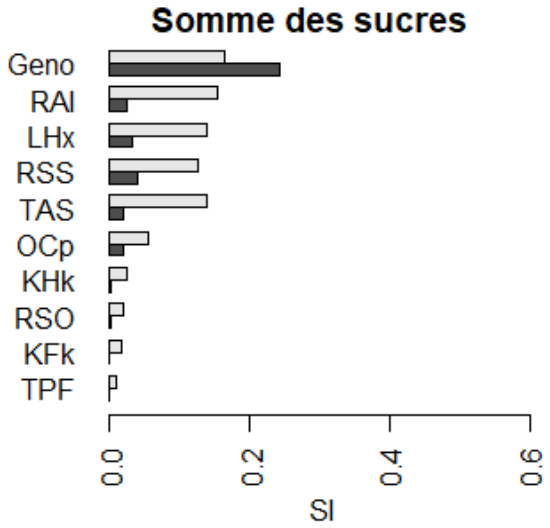


FIGURE 3.6 – Indices de sensibilité (SI) de la fonction objectif agrégée (somme d'erreur quadratique des quatre sucres) pour chacun des 9 paramètres du modèle. Les indices principaux de sensibilité sont représentés par les barres noires et les indices d'interaction sont représentés par les barres grises.

À la vue de ces résultats, il est légitime d'imaginer qu'une optimisation multi-objectif, prenant en compte l'accord données-prédictions pour les 4 sucres séparément (Eq. 3.9), pourrait permettre une meilleure calibration du modèle. En effet, la sensibilité variable des quatre objectif aux paramètres du modèle devrait aboutir sur une estimation plus précise des valeurs de KFk, KHk TPF et RSO qui sont très peu influents sur la fonction agrégée (Eq. 3.8).

Aussi, nous allons tester cette hypothèse dans la suite du chapitre.

3.2 Robust parameters estimation of kinetic model of sugar metabolism in peach to assess the inter-individual genetic variability

Abstract

Integrating genetic information into dynamical models is a key to understand variations among genotypes. It could provide information on how to improve quality and response to genetic, physiological and environmental controls. Being able to calibrate dynamical models on a large progeny of genotypes is a first and necessary step towards genotype-phenotype models. An ODE kinetic model of sugar metabolism has been proposed by Kanso et al. (2020) to simulate the accumulation of different sugars during peach fruit development. We compared here two different calibration strategies. First, the model was calibrated for each genotype independently. Two formulations of the problem have been tested, either as a Single-Objective Optimization (SOO) problem or as a Multi-Objective Optimization (MOO) problem. Second, the model was calibrated for all genotypes simultaneously using the population-based model proposed by Baey et al. (2018). The two strategies were applied on a set of simulated data and then on a real dataset derived from an interspecific population of 106 peach genotypes. Results showed that the independent calibration of each genotype allowed for a high goodness of fit for most genotypes, especially in the SOO formulation. However, the estimated parameters suffered from a lack of practical identifiability as independent repetitions of the estimation algorithm did not always converge to a same value for most genotypes. This issue was resolved using the population-based calibration strategy. Hence, it showed a good identifiability of the population parameter values, a goodness of fit comparable to the one obtained with the first strategy and a good characterisation of parameter variations within the progeny, which is a key to assess the inter-individual genetic variability. These results are an important step towards the development of reliable gene-to-phenotype models.

keyword

Peach fruit, Kinetic model, Inter-individual variability, Parameter estimation, Nonlinear mixed model, Population model, Optimisation, MCMC-SAEM, genetic algorithms, NSGA-II.

3.2.1 Introduction

Predicting genotype-to-phenotype relationships under contrasting environments is a big challenge for plant biology and breeding. Process-based ecophysiological models are considered as powerful tools to deal with such a challenge (Martre et al. 2015) as they are able to simulate plant phenotypic traits as the result of developmental,

architectural and environmental factors. Within this framework, differences among genotypic responses are usually encoded into a set of parameters values, considered as the phenotypic fingerprint of the underlying genetic information (Martre et al. 2015). The aim of gene-to-phenotype modelling is to make this link explicit, by expressing the value of genetic parameters as a function of specific genetic loci (QTL) or allelic combinations of genes, depending on the available information. Ultimately, gene-to-phenotype models can be used to explore trait correlations in a population and to analyse Genotype (G) x Environment (E) x Management (M) interactions, in the perspective of virtual breeding. The construction of gene-to-phenotype models is challenging as it relies on the calibration of a large number of genotypes, usually with few data available. Moreover, process-based models are generally non-linear and involve a large number of parameters among which correlations may exist (Chou et al. 2009; Gutenkunst et al. 2007). This may lead to a *non-identifiable* problem i.e. parameter values cannot be unambiguously inferred from available data and so multiple parameter sets may provide equally good fits of the data. This situation is particularly troublesome in the perspective of studying the genetic control of the parameters since they are expected to represent the genotypic fingerprint of the system (Hass et al. 2017; Becker et al. 2010). In general, a reliable estimate of the parameters depends on the structure of the model, the quality of the observations and the choice of the calibration strategy. Although model reduction approaches can be effectively used to improve model identifiability (Maiwald et al. 2016; Chou et al. 2009; Snowden et al. 2017), we focus here on the last point, namely the formalization and implementation of the calibration process.

The fitting process usually starts with the definition of a likelihood function. A likelihood is a function that measures the fit of a statistical model to data (experimental observations) according to the value of the parameters of the model. Different hypotheses can be made at this stage, concerning the relationship and the error associated to experimental data, leading to different possible formulations of the likelihood function. In the case of complex phenotypes, a multi-objective formulation can be adopted, allowing for a deeper analysis of possible trade-offs among physiological traits. Then, a numerical method has to be used to maximise the likelihood and provide estimates for the model parameters. Several optimization algorithms are available, either local or global, divided into deterministic and stochastic methods (Floudas et al. 2013). Local search algorithms often start from an initial solution of the problem and iteratively try to improve the current solution by searching for better solutions in its local neighbourhood. In general, local minima or maxima are not guaranteed to be global minima or maxima. Conversely, global search algorithms try to find the best solution by exploring the whole space of feasible solutions using different strategies. The global optimisation algorithms are more suitable for the calibration of complex models but they are often more difficult to implement. Moreover, in some cases, local strategies may do the job while being simpler and/or faster to implement.

When dealing with large genetic populations, the issues of estimating individual parameters are even more complex, due to the presence of possible correlations among

parameters, due, for instance, to a shared genetic control. In some recent studies, Baey et al. (2016) et Baey et al. (2018) proposed a population-based approach using a Non Linear Mixed Model (NLMM). This approach is adapted to cases where repeated measurements are available for several individuals from the same population. This approach takes into account two sources of variability deriving from observations : intra-individual variability, i.e. how the measurements of the same individual vary between repetitions and inter-individual variability i.e how the measurements between individuals vary. The main idea of this strategy is to catch these sources of variability in the whole progeny. It assumes that individual parameters, defined for each genotype, are independent and identically distributed random variables following a common probability distribution.

The aim of this paper is to compare calibration strategies using a model of sugar metabolism (Desnoues et al. 2018 ; Kanso et al. 2020) and 106 genotypes which belong to an inter-specific peach cross, as an example. More specifically, we looked at the impact of two calibration strategies on the resulting parameter estimation, in terms of quality of fit as well as robustness and uniqueness of the solution. These comparisons were performed through a simulation study on artificial data and on real experimental data.

The first strategy aimed to calibrate the model for each genotype independently. The calibration problem was formulated as a Single-Objective Optimisation problem (SOO) and as a Multi-Objective Optimisation problem (MOO) looking for the maximum value of likelihood functions. In the second strategy, the model was calibrated for all genotypes simultaneously using the Population-Based (PB) model proposed by Baey et al. (2018). In this strategy, two covariance structures for the parameters vector were considered (one where parameters are assumed to be correlated and another one where parameters are assumed independent) and compared using AIC criterion to choose the best formulated model. Then to assess the interest of each strategy, the modelling efficiency (EF) (Mayer et al. 1993) was used as an indicator of goodness of fit and the intra-genotype variation between estimations and the corresponding correlations were computed as a measure of the reliability of the parameters estimation.

In the sections 3.2.2 and 3.2.4 of this paper, the mathematical model is presented as well as the simulated and experimental data. Sections 3.2.5 and 3.2.6 introduces the problem formulations, optimisation methods and indicators used for the assessment of the two strategies. Afterwards, the results of the two strategies to calibrate the model of sugar metabolism are reported, first for the virtual simulation and then for the experimental data. Finally, a general discussion on the advantages and limitations of each strategy closes the paper.

3.2.2 Mathematical model

The model used in this paper is the one proposed by Kanso et al. (2020) as a reduced version of the model previously developed by Desnoues et al. (2018). The model

describes the accumulation of four different sugars (sucrose, glucose, fructose, and sorbitol) in peach fruit during its development, as a system of ordinary differential equations (ODE). Carbon enters the fruit from the plant sap which is transformed by a metabolic network, including enzymatic reactions and transport mechanisms between the cytosol and the vacuole (see Figure 3.7). A total of 17 reactions are included in the model and represented by a linear flows. The model involves 30 parameters, among which 21 are considered as known (measured or taken from literature) and identical for all genotypes. The remaining 9 parameters are considered as unknown and genotype-dependent.

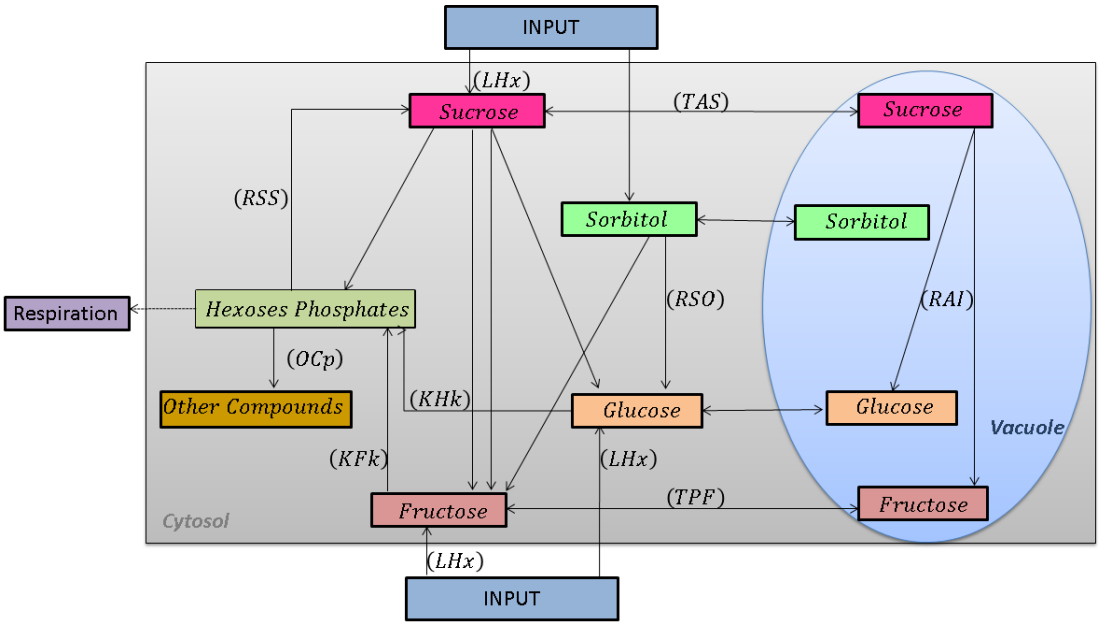


FIGURE 3.7 – Schematic network of the reduced model of sugar accumulation in the peach fruit. Arrows represent carbon flows. The corresponding kinetic equations are presented in Kanso et al. (2020)

In this work, we focused on the estimation of the 9 unknown parameters described in the following table :

TABLE 3.4 – Description of the 9 unknown genotype-dependent parameters of the sugar model

Parameter	Description	Range	Unit
LHx	sucrose proportion hydrolyzed in the apoplasm	]0, 1]	
RAI	coefficient of the transfer function between sucrose and (glucose +fructose) under action of acid invertase (AI) enzyme	]0, 0.1]	day^{-1}
KFk	fructokinase (FK) affinity	]0, 100]	$mggFW^{-1}$
KHk	hexokinase (HK) affinity	]0-500]	$mggFW^{-1}$
OCp	coefficient of the transfer function between hexoses phosphate and other compounds	[250, 1500]	day^{-1}
RSS	coefficient of the transfer function between hexoses phosphate and sucrose	]0, 1000]	day^{-1}
TAS	coefficicent of sucrose transport (active import) from cytosol to vacuole	]0, 400]	$mggFW^{-1}day^{-1}$
TPF	coefficient of fructose passive transport between cytosol and vacuole and in the opposite direction	]0, 150]	$mggFW^{-1}day^{-1}$
RSO	coefficient of the transfer function between sorbitol and glucose under action of sorbitol oxydase (SO) enzyme	]0, 10]	day^{-1}

3.2.3 Problem formulations and calibration strategies

To estimate the 9 genotype-dependent parameters, two calibration strategies were used to fit the data i.e. virtual or measured fruit sugar concentrations, at six developmental stages, on the whole progeny : i) parameters were estimated independently for each genotype and ii) parameters were simultaneously estimated for the whole progeny. Figure 3.8 gives a schematic representation of the calibration steps, detailed below.

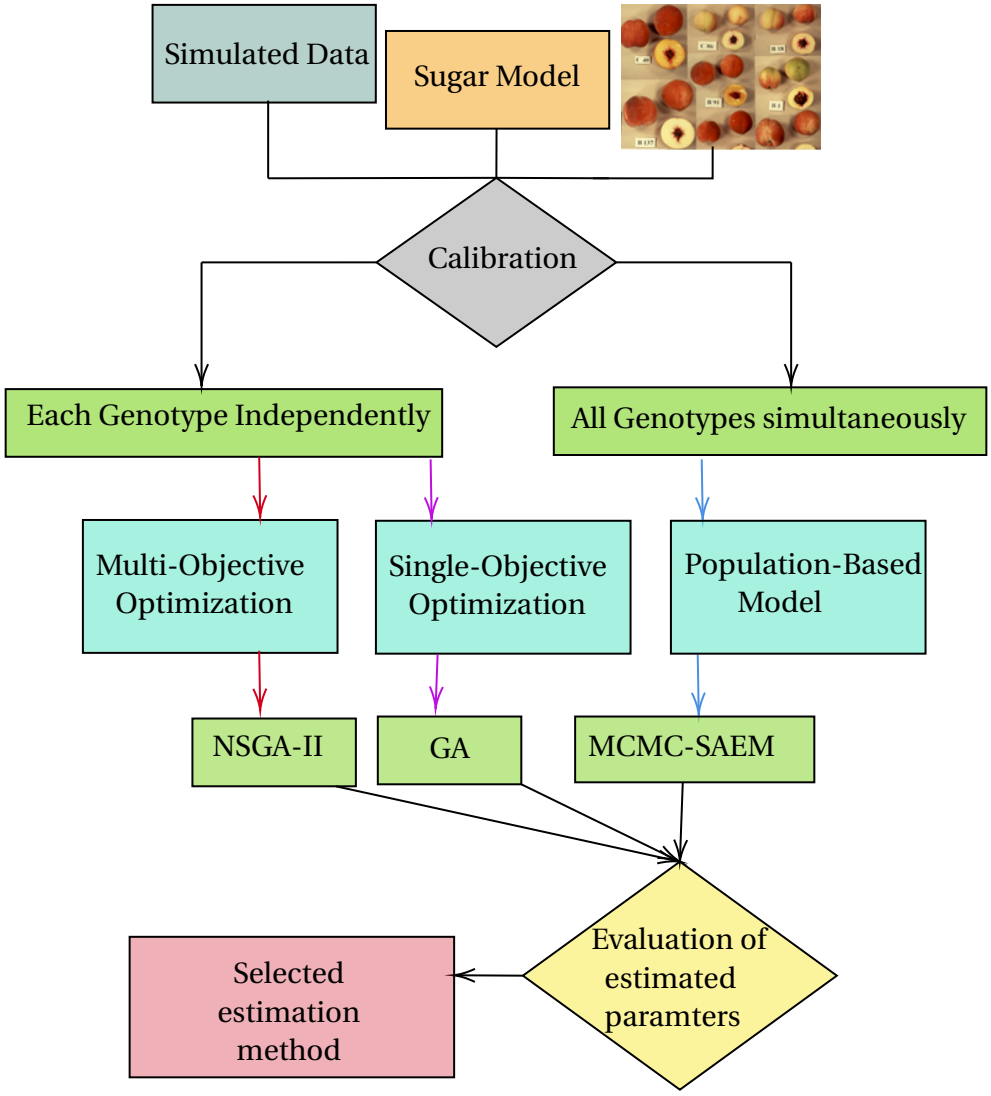


FIGURE 3.8 – Graphical representation of the proposed steps to calibrate the sugar model. Two strategies were used to calibrate the model : 1) for each genotype independently ; two formulations were considered : Single-Objective Optimisation (SOO) and Multi-Objective Optimisation (MOO), 2) for all genotypes simultaneously using Population Based (PB) model. To estimate the unknown 9 genotype-dependent parameters, 3 algorithms were used : GA for (SOO), NSGA-II for (MOO) and MCMC-SAEM for (PB).

3.2.3.1 Mathematical notations

The following mathematical notations were used :

- G : number of genotypes in the population i.e. 106 for the peach progeny, 100 for the virtual populations

- n_{SG} : number of sugars i.e. 4.
- $\mathcal{T}_M^{(k)}$: set of N_M measurement times (i.e at 6 different time points during fruit development) for the genotype k
- $N^{(k)}$: number of experimental observations for the genotype k
- $N_j^{(k)}$: number of experimental observations for the genotype k and for the sugar j. Note that : $N^{(k)} = \sum_{j=1}^{n_{SG}} N_j^{(k)} = n_{SG} \times r \times N_M$, where r is the number of replicates at time t_i, for the four sugars (sucrose, sorbitol, fructose and glucose). $r = 3$ for 10 genotypes and $r = 1$ for 96 genotypes
- $\phi^{(k)}$: vector of 9 parameters to be estimated for each genotype k
- $\varphi^{(k)}$: log transformation of $\phi^{(k)}$
- $\mathcal{M}_{ij}(\phi^{(k)})$: i-th *prediction* of sugar j concentrations of genotype k at time $i \in \mathcal{T}_M^{(k)}$, obtained with model $\mathcal{M}$ and parameters $\phi^{(k)}$.
- $\widetilde{\mathcal{M}_{ij}}(\phi^{(k)})$: log transformation of $\mathcal{M}_{ij}(\phi^{(k)})$
- $y_{ij}^{(k)}$: i-th *observation* of sugar j concentrations of genotype k, with $i \in \mathcal{T}_M^{(k)}$
- $\widetilde{y}_{ij}^{(k)}$: log transformation of $y_{ij}^{(k)}$

3.2.3.2 Model's calibration for each genotype independently

In the framework of this first strategy, we used two alternative formulations of the model calibration as a Single-Objective Optimization (SOO), and as a Multi-Objective Optimization (MOO) problem.

Calibration as Single-Objective Optimisation (SOO)

The general main goal of SOO is to find the best solution, which corresponds to the minimum or maximum value of a single-valued objective function. With this aim, the observations of the four sugars were combined together into a single objective defined as the sum of the squared errors of each sugar. We assumed that the observations $\{y_{ij}^{(k)}\}_{j=1,\dots,n_{SG}}$ follow a Gaussian distribution $\mathcal{N}(\mathcal{M}_{ij}(\phi_{SOO}^{(k)}), \sigma_k^2)$ with constant variance σ_k^2 independent of the sugar type. Under these assumptions, the calibration problem can be formulated as follows :

$$\hat{\phi}_{SOO}^{(k)} = \underset{\phi_{SOO}^{(k)}}{\arg\min} \left\{ \sum_{j=1}^{n_{SG}} \sum_{i=1}^{N_j^{(k)}} \left(y_{ij}^{(k)} - \mathcal{M}_{ij}(\phi_{SOO}^{(k)}) \right)^2 \right\}. \tag{3.10}$$

Calibration as Multi-Objective Optimisation (MOO)

As a second case, model calibration was defined as a Multi-Objective Optimisation (MOO). Here, the optimisation was performed on a vector of objective functions. Contrary to SOO, each sugar was considered independently. Thus, the problem can be decomposed into 4 objective functions, each of which addressing one sugar. We used the sum of squared errors of each sugar as objective function. These functions

are considered as separate objectives to be optimised independently. The idea was to estimate the parameters and find a compromise solution between all objective functions. In this context, the observations of each sugar $y_{ij}^{(k)}$ were assumed to follow a Gaussian law $\mathcal{N}(\mathcal{M}_{ij}(\phi_{MOO}^{(k)}), \sigma_{kj}^2)$ with specific constant variance σ_{kj}^2 for each sugar.

The calibration problem was defined as :

$$\left(\tilde{\phi}_{MOO}^{(k)}, (\tilde{\sigma}_{kj}^2)_{j=1,\dots,n_{SG}}^{T}\right) = \underset{(\phi_{MOO}^{(k)}, \sigma_{kj}^2)}{\operatorname{argmin}} \left\{F_j^{(k)}(\phi_{MOO}^{(k)}, \sigma_{kj}^2)\right\}_{j=1,\dots,n_{SG}}^{T} \tag{3.11}$$

where

$$F_j^{(k)}(\phi_{MOO}^{(k)}, \sigma_{kj}^2) = \frac{N_j^{(k)}}{2} log\left(\sigma_{kj}^2\right) + \frac{1}{2\sigma_{kj}^2} \sum_{i=1}^{N_j^{(k)}} \left(y_{ij}^{(k)} - \mathcal{M}_{ij}(\phi_{MOO}^{(k)})\right)^2$$

3.2.3.3 Model's calibration for all genotypes simultaneously : Population-based Model

We used the statistical approach developed by Baey et al. (2018) for simultaneous estimation of all genotypes in the population. The statistical model was defined as a two-stage hierarchical formulation. The first stage concerned the intra-individual variability. The observations for each genotype were modeled assuming a multiplicative error (log-additive) as :

$$\tilde{y}_{ij}^{(k)} = \widetilde{\mathcal{M}}_{ij}(\phi^{(k)}) + \varepsilon_{ij}^{(k)}, \quad \varepsilon_{ij}^{(k)} \sim \mathcal{N}(0, \sigma^2) \tag{3.12}$$

where $\tilde{y}_{ij}^{(k)} = log(y_{ij}^{(k)})$ and $\widetilde{\mathcal{M}}_{ij}(\phi^{(k)}) = log(\mathcal{M}_{ij}(\phi^{(k)}))$, $\varepsilon_{ij}^{(k)}$ is the residual error assumed to follow a Gaussian distribution with mean 0 and constant variance σ^2. The vectors $(\varepsilon_{ij}^{(k)})$ (with $1 \leq k \leq G, 1 \leq j \leq n_{SG}, 1 \leq i \leq N_j^{(k)}$) are assumed independent, and the sequences $(\varepsilon_{ij}^{(k)})$ and $(\varphi^{(k)})$ are assumed mutually independent. In the second stage, we took into account the inter-individual variability and the set of nine parameters for each genotype $\phi^{(k)}$ was considered as a random effect in the whole progeny. In our case, log transformations were applied to these parameters as they are all strictly positive by definition :

$$\varphi^{(k)} = \log\left(\phi^{(k)}\right) = \beta + \xi^{(k)}, \quad \xi^{(k)} \sim \mathcal{N}(0, \Gamma), \tag{3.13}$$

where $\phi^{(k)}$ is the initial individual parameters of the model, β is the population mean and Γ is the covariance matrix. In the following, the term 'random effects' will refer equivalently to $\phi^{(k)}$ or $\varphi^{(k)}$.

We considered two types of covariance : first, a full matrix Γ_1 corresponding to the case where all random effects are assumed to be correlated, and second, a diagonal matrix Γ_2 corresponding to the case where random effects are assumed to be independent.

The two resulting models were compared using the AIC criterion (see section 3.2.6.6).

Under the above assumptions, and noting $\widetilde{y} = (\widetilde{y}_{ij}^{(k)}, 1 \leq k \leq G, 1 \leq j \leq n_{SG}, 1 \leq i \leq N_j^{(k)})$, $\varphi = (\varphi^{(k)}, 1 \leq k \leq G)^T$ and $\theta = (\beta, \Gamma, \sigma^2)$, we have

$$f(\widetilde{y}|\varphi;\theta) = \prod_{k=1}^{G} \prod_{j=1}^{n_{SG}} \prod_{i=1}^{N_j^{(k)}} f(\widetilde{y}_{ij}^{(k)}|\varphi^{(k)};\theta), \tag{3.14}$$

and

$$f(\varphi;\theta) = \prod_{k=1}^{G} f(\varphi^{(k)};\theta), \tag{3.15}$$

where $f(\widetilde{y}_{ij}^{(k)}|\varphi^{(k)};\theta)$ is the conditional density of the observations $\widetilde{y}_{ij}^{(k)}$ given the individual parameters, and $f(\varphi^{(k)};\theta)$ is the marginal density of $\varphi^{(k)}$.

The vector of parameters is $\theta = (\beta, \Gamma, \sigma^2)$, and can be estimated by maximizing of the likelihood defined as

$$L(\theta) = \int f(\widetilde{y}|\varphi;\theta) f(\varphi;\theta) d\varphi \tag{3.16}$$

In the framework of the model defined in Eq. 3.12 and Eq. 3.13, the non-linearity of the mathematical model makes in general the calculation of this integral (3.16) analytically impossible. Mixed models can be seen as a special case of incomplete data models, with $\widetilde{y}$, the observed data, and the random effects φ being the unobserved data. In this case, maximum likelihood estimation of $\theta = (\beta, \Gamma, \sigma^2)$ can be done using different approaches that can be found in Davidian et al. (1995). In our case, the MCMC-SAEM algorithm as described in Baey et al. (2018) was used and adapted to our mathematical model. In order to get an estimate of parameters of individual genotypes, the individual parameters were computed as $\hat{\varphi}^{(k)} = \mathbb{E}(\varphi^{(k)})|\widetilde{y}^{(k)};\hat{\theta})$ for each genotype k, where $\hat{\theta}$ is the value of θ that maximizes the likelihood.

3.2.4 Experimental and simulated data

3.2.4.1 Simulation study

The proposed calibration strategies were first applied to a set of simulated datasets, in order to check the quality and stability of the corresponding estimator. Given the intrinsic difference between the genotype and population-centered approaches, distinct procedures were used to built the corresponding datasets, that are detailed in the following sections.

Datasets for the independent calibration of genotypes

To test the quality of the genotype-centered strategies (SOO and MOO formulations), we built a set of simulated datasets corresponding to repeated measurements of a *same* virtual genotype. Accordingly, a single set of parameters and the same model

inputs, i.e. the genotype growth curve, were used to simulate the dynamics of the four sugars by means of our ODE model. Parameter values were selected randomly using a uniform distribution within the range of previous estimates (Desnoues et al. 2018; Kanso et al. 2020) and the input signal was picked among the observed growth dynamics of the 106 real genotypes. The exact values are reported in Table 3.5.

In order to obtain 100 different datasets for parameter estimation, six dates were selected, evenly distributed along the fruit developmental period, and a random observation noise added to the simulation values.

TABLE 3.5 – Parameters used to simulate the dynamics of sugar concentrations and the growth curve

Sugar model	LHx	RAI	KFk	KHk	OCp	RSS	TAS	TPF	RSO	σ^2
	0.03	0.01	12	20	665	33	20	12	1	0.05
Growth curve	a_1	a_2	a_3	a_4	a_5	a_6				
	188.7	0.08	102.6	21.9	0.008	0.17				

Datasets for population-based calibration of the genotypes

This time, one hundred populations were generated, each one including 100 different genotypes. All effects were considered random, i.e.

$$\phi^{(k)} = (LHx, RAI, KFk, KHk, OCp, RSS, TAS, TPF, RSO)^T$$

with log-mean population β and a diagonal variance $diag(\Gamma)$ (see Table 3.6).

TABLE 3.6 – Parameter values used to simulate 100 populations considering all effects are random

Parameters	LHx	RAI	KFk	KHk	OCp	RSS	TAS	TPF	RSO
β	-3.5	-4.5	2.5	3	6.5	3.5	3	2.5	0.01
$diag(\Gamma)$	3.5	1.5	1.5	2.5	0.2	3.5	2.5	2	1.5

For each simulated population, a set of 100 individual parameters was generated according to Eq. 3.13. Observations were then obtained by running the model according to Eq. 3.12, with $\sigma^2 = 0.05$. Six dates were selected, evenly distributed along the fruit developmental period. For each genotype, parameters describing model inputs (fruit dry and fresh weights) were randomly assigned by picking one of the observed growth dynamics and adding an overall random variation between zero and 10% on fruit weight. Moreover, in order to further increase variability among genotypes, the duration of fruit development was defined randomly using a uniform distribution within the range of observed timespans plus 40%.

3.2.4.2 Experimental data

A progeny of 106 genotypes has been used in this study. As described by Quilot et al. (2004a), these genotypes are issued from an inter-specific progeny obtained by two subsequent back-crosses between *Prunus davidiana* (Carr.) P1908 and *Prunus persica* (L.) Batsch 'Summergrand' and then 'Zephyr'. The concentrations of different metabolites, namely sucrose, glucose, fructose, sorbitol, and hexoses phosphates, the fruit flesh fresh mass and the flesh dry matter content, as well as enzymatic capacities (maximal activity) of twelve enzymes were measured at six time points during fruit development, for all genotypes, as described in Desnoues et al. (2014). At each time point, available data consists in 3 biological measurements for 10 genotypes and 1 biological measurement for the remaining 96 genotypes. All these observations were measured at 6 dates after full blooming for all genotypes. Genotypes can be sorted into two classes : those having a 'standard phenotype', namely a balanced fructose-to-glucose ratio at maturity, and those having a 'low fructose phenotype' due to the lower proportion of fructose compared with glucose at maturity (Desnoues et al. 2018).

3.2.5 Parameter estimation

3.2.5.1 Optimisation algorithms and their settings

For the first calibration strategy i.e. each genotype independently, we used two genetic algorithms : GA (Goldberg 1989 ; Conn et al. 1991 ; Conn et al. 1997) for the SOO formulation and NSGA-II (Deb et al. 2002) for MOO formulation. Population size was set to 200 and the maximum number of generations was 300.

For the second calibration strategy i.e. all genotypes simultaneously, a MCMC-SAEM algorithm (Kuhn et al. 2004 ; Baey et al. 2018), coupling a Markov chain Monte Carlo (Hastings 1970) with a Stochastic Approximation Expectation-Maximization (Robbins et al. 1951), was used for the estimation of $\theta = (\beta, \Gamma, \sigma^2)$. The details of the algorithm can be found in Baey et al. (2018). The code was implemented in MATLAB software by the authors. For the Markov chains, a burn-in period of 1000 iterations was performed and we kept the last 10 iterations (for a chain of size equal to ten) as estimates for individual parameters. The burn-in period was used to give the Markov Chain time to reach its equilibrium distribution (Johansen et al. 2010). The choice of the step size at l iteration of the algorithm was fixed according to Kuhn et al. (2005). During the first stage of K_1 iterations, it was fixed to 1, and then decrease it as $1/(l - K_1)$ for K_2 more iterations. The values of K_1, K_2 were stated respectively to 180, 120 in the context of simulation study and 300, 200 when the experimental data was used.

Settings are described in Table 3.7. Matlab software (MATLAB R2018a, The MathWorks Inc., Natick, MA) was used for all calibrations.

TABLE 3.7 – Settings of three optimization algorithms. C.P : Crossover probability

Algorithm	Parameters			Stopping criterion
GA	Population size 200	Number of generations 300	C.P 0.7	Stopped when the best-cost function value over generations was less than 10^{-6}
NSGA-II	200	300	0.7	Stopped when the number of generations was reached
MCMC-SAEM	Iterations 300 for simulation study 500 for experimental data	Length of Markov chain 10		Stopped when total number of iterations was reached

3.2.5.2 Selection of the reference solution

For GA and NSGA-II, the estimation procedure was repeated 10 (real data) to 20 (simulated data) times for each genotype k, to take into account the stochastic nature of algorithms and to ensure the good exploration of the parameters' space and ultimately to escape local convergence. In the case of SOO/GA the solution having the minimum likelihood was selected as the reference estimation, for each genotype k. In addition, the solutions which did not exceed 5% deviation from the best score were kept to analyse the variability of the estimated parameter values.

Unlike in single-objective optimisation problems, multi-objectives problems like MOO/NSGA-II do not yield a single optimal solution but a set of trade-off solutions among our four criteria. All solutions $\left(\tilde{\phi}^{(k)}_{MOO}, (\tilde{\sigma}^2_{kj})^T_{j=1,\dots,n_{SG}}\right)$, resulting from the independent repetitions of the calibration process, were first pooled together and then filtered thanks to the $is.dominated$ function of the "emoa" package (developed for R) in order to identify the Pareto-optimal set i.e. solutions allowing the best trade-offs between calibration objectives. Let's call $\left(\hat{\phi}^{(k)}_{MOO}, (\hat{\sigma}^2_{kj})^T_{j=1,\dots,n_{SG}}\right)$ the resulting set of Pareto-optimal solutions. Let $|\hat{\phi}^{(k)}_{MOO}|$ be the cardinality of this Pareto set for the k-th genotype. The issue is then how to select a single reference solution among all these options. We tested two alternative strategies, already used in other related works (Constantinescu et al. 2016).

First, for each solution in the set $\left(\hat{\phi}^{(k)}_{MOO}, (\hat{\sigma}^2_{kj})^T_{j=1,\dots,n_{SG}}\right)$, we summed the four predictions associated to this solution. Then, we selected the solution(s) associated to the minimal value of errors sum. The decision function was defined as :

$$(\phi^{(k)}_{MOO}, (\sigma^2_{kj})^T_{j=1,\dots,n_{SG}}) = \underset{(\hat{\phi}^{(k)}_{MOO}, \hat{\sigma}^2_{kj})}{\arg\min} \sum_{j=1}^{n_{SG}} F^{(k)}_j(\hat{\phi}^{(k)}_{MOO}, \hat{\sigma}^2_{kj}) \tag{3.17}$$

Note that there was no need for any optimisation at this step but only a simple calculus on the basis of the optimisation results.

Second, we looked for the solution(s) that avoid the worst prediction for each error (sugar). To achieve this, for each solution in the set $\left(\hat{\phi}^{(k)}_{MOO}, \hat{\sigma}^2_{kj}\right)$, we associated its maximal prediction error corresponding to the worst prediction of one or more sugars. Then, we selected the solution(s) associated with the lowest value(s) among all the

worst values. The decision function in this case can be formulated as follows :

$$(\phi_{MOO}^{(k)}, (\sigma_{kj}^2)_{j=1,\dots,n_{SG}}^T) = \underset{(\hat{\phi}_{MOO}^{(k)}, \hat{\sigma}_{kj}^2)}{\operatorname{argmin}} \max_j (F_j^{(k)}(\hat{\phi}_{MOO}^{(k)}, \hat{\sigma}_{kj}^2))^T \quad \text{with } j = 1, \dots, n_{SG} \tag{3.18}$$

For MCMC-SAEM, the estimation procedure was repeated 5 (simulated data) to 10 (real data) times, for all genotypes simultaneously. Two situations were possible : i) either the repetitions converged to the same value (within the Monte Carlo error) and in this case the average of the all estimated values was chosen as the reference estimation or ii) the repetitions did not converge to the same value and in this case we chose the one with the highest likelihood value.

3.2.5.3 Confidence intervals

Confidence intervals for θ can be obtained via parametric bootstrap (Efron 1982 ; « Chapman & Hall ; London : 1993 »). More precisely, given the value of $\hat{\theta}$, 1000 datasets were simulated. For each dataset d, a set of 106 individual parameters was generated according to Eq. 3.13. Observations were then obtained by model simulation for each individual according to Eq. 3.12. Then, for each of these bootstrap samples, we rerun the estimation procedure. This yielded 1000 estimators $\hat{\theta}_\rho$, from which we estimated the distribution of the estimator $\hat{\theta}$. The bootstrap confidence interval of level $1-\alpha$ ($\alpha = 0.05$) was then given by :

$$IC_{1-\alpha}(\theta) = [\hat{\theta}_d^{(\frac{\alpha}{2}1000)}; \hat{\theta}_d^{(1-\frac{\alpha}{2}1000)}] \tag{3.19}$$

where $\hat{\theta}_d^{(\frac{\alpha}{2}1000)}$ is the statistic of order $\frac{\alpha}{2}1000$ of bootstrap estimators.

3.2.6 Strategy selection

The different estimation methods were evaluated using different criteria, including the quality of phenotypic predictions, the accuracy and variations between the parameter estimates, and the intra-genotype correlations. Specific indicators were defined and described in the following sections. Moreover, the Aikake criterion was used to compare alternative formulations of the parameter estimation procedures, in the case of population-based approaches.

3.2.6.1 Modelling efficiency

The modelling efficiency (EF) proposed by Mayer et al. (1993), is a goodness of fit indicator that can be computed as a measure of the accuracy of phenotypic predictions.

It is defined as follows :

$$EF_j = 1 - \frac{\sum_{k=1}^{G}\sum_{i=1}^{N_j^{(k)}}(y_{ij}^{(k)} - \mathscr{M}_{ij}(\phi^{(k)}))^2}{\sum_{k=1}^{G}\sum_{i=1}^{N_j^{(k)}}(y_{ij}^{(k)} - \overline{y}_j^{(k)})^2}, \quad EF_j \in]-\infty\ 1], \tag{3.20}$$

where $\overline{y}_j^{(k)}$ is the mean of observations of sugar j for genotype k over all dates i. $EF = 1$ means a perfect equality between the predictions and the observations, and if $EF < 0$ means that the model predictions are worse than the mean of the observations.

3.2.6.2 Mean squared Error

The mean squared error (MSE) of an estimator $\hat{\phi}$ is a measure of its accuracy. It quantifies the risk of using this estimator as a proxy of the real parameter value ϕ, in the framework of parameter estimation.
The MSE is defined as :

$$MSE(\hat{\phi}) = \text{Bias}^2(\hat{\phi}) + \text{Var}(\hat{\phi}), \tag{3.21}$$

where

$$\text{Bias}(\hat{\phi}) = \mathbf{E}(\hat{\phi}) - \phi \tag{3.22}$$

and

$$\text{Var}(\hat{\phi}) = E(\hat{\phi}^2) - (E(\hat{\phi}))^2. \tag{3.23}$$

In practice, the bias was estimated as the difference between the average prediction of our reference estimates (as a proxy of the expected value) and the correct value which we are trying to predict. The variance was estimated by

$$\text{Var}(\hat{\phi}) = \frac{1}{n}\sum_{d=1}^{n}(\hat{\phi}_d - \overline{\phi})^2 \tag{3.24}$$

where $\overline{\phi}$ is the mean value of our reference estimates and n=100 the number of simulated data.

3.2.6.3 Expected Error (%)

In the context of the simulation study for the population-based approach, the quality of the *individual* parameters was assessed by computing the relative distance between the true values $\phi_{r,k}$ and the estimated values $\hat{\phi}_{r,k}$, for the r-th individual parameter and the genotype k. For each simulated population, the Expected Error E_r (%) of the r-th parameter was defined as :

$$\text{E}_r = \frac{1}{G}\sum_{k=1}^{n}\left|\frac{\phi_{r,k} - \hat{\phi}_{r,k}}{\phi_{r,k}}\right| \times 100 \tag{3.25}$$

where $G = 100$ is the number of virtual genotypes included in the population. Overall our simulation study yielded 100 values of E_r, for each of the 9 parameters.

3.2.6.4 Intra-genotype parameter correlation

In the framework of the genotype-by-genotype strategy, we checked for correlation among parameter values obtained via independent estimations (repetitions) of the algorithm, on a same genotype. To this aim, Person linear correlation was used. For each genotype k, a matrix $\rho^{(k)}$ containing the pairwise correlation coefficients between each pair of parameters was built using the Matlab function `corr`. The average correlation between two parameters i and j was defined as :

$$\rho(i,j) = \frac{1}{G}\sum_{k=1}^{G}\rho^{(k)}(i,j), \tag{3.26}$$

where G is the number of genotypes of the population. The procedure was repeated for both SOO/GA and MOO/NSGA-II formulations.

3.2.6.5 Normalized estimate distance between calibration strategies

In order to quantitatively compare the estimation obtained with different calibration strategies, we define a pair-to-pair distance between the reference estimates $\phi_{r,a}^{(k)}$ obtained with algorithm a ($a = 1,2,3$) for each r-th parameter ($r = 1\ldots9$) and genotype k as

$$D_r^{(k)}(a1,a2) = \frac{|\phi_{r,a1}^{(k)} - \phi_{r,a2}^{(k)}|}{med_r^{(k)}} \tag{3.27}$$

where $med_r^{(k)}$ is the median estimated value of the parameter $\phi_r^{(k)}$ over the three methods :

$$med_r^{(k)} = median(\phi_r^{(k,a)}) \quad a = 1\ldots3 \tag{3.28}$$

This allows to rescale the difference among estimates with respect to the expected parameter value. We repeated the procedure for each couple $(a1,a2) \in \{(1,2),(1,3),(2,3)\}$.

3.2.6.6 Akaike information criterion (AIC)

Different formulations of the PB optimisation problem were compared using the AIC criterion, which gives information on the likelihood of the proposed model based on available experimental data and weighted by the number of free parameters (Burnham et al. 2002) :

$$AIC(\theta) = -2\,log\,L(\theta) + 2n_\theta \tag{3.29}$$

where n_θ is the number of estimated parameters θ, N the number of observations and L the likelihood function. In the context of nonlinear mixed-effect models, the

likelihood function is not directly computable. However, it is possible to approximate the likelihood at final point estimates $\hat{\theta}$, by Monte Carlo simulations (see Section 3.2.9.1).

3.2.7 Results

In this section, results from the simulation and experimental studies are presented. In both cases, we first considered the genotype-by-genotype strategy consisting in the calibration on each genotype taken independently, with the two optimisation formulations (SOO and MOO). Then, we examined results from the second strategy, when the model was calibrated simultaneously on all genotypes of the population.

3.2.7.1 Simulation study

Model calibration for each genotype independently

Twenty independent repetitions of the GA and NSGA-II algorithms were performed for each of the 100 simulated datasets to assess the stability of these algorithms. Results showed that the estimations of the parameter values were highly unstable, resulting in large variations of the estimates, both with the SOO/GA and MOO/NSGA-II formulations. An illustration of these variations is presented in Figure 3.9. Indeed, the different repetitions did not converge to the true value or even to an identical value, suggesting that several local maxima were found. A large variability was generally observed on the estimations of the parameters OCp, TPF, RSO and TAS with both SOO and MOO formulations (see Figure 3.13). However, estimations obtained for LHx, RAI, KFk, KHk and RSS using the SOO formulation were noticeably less variable than those obtained using the MOO formulation.

TABLE 3.8 – Results of mean squared error (MSE) obtained on simulated data with the GA and NSGA-II algorithms. MSE$(NSGA-II)^{(1)}$ was computed using the solutions selected from 3.17. MSE$(NSGA-II)^{(2)}$ was computed using the solutions selected from 3.18. For each parameter, the smallest MSE value is indicated in bold.

Parameter	MSE(GA)	MSE$(NSGA-II)^{(1)}$	MSE$(NSGA-II)^{(2)}$
LHx	**0.015**	0.181	0.809
RAI	**0.004**	0.213	0.577
KFk	**0.0007**	0.001	0.108
KHk	**0.008**	**0.008**	0.352
OCp	0.064	**0.044**	0.088
RSS	**0.0401**	0.102	0.761
TAS	**0.219**	0.249	0.856
TPF	1.551	**0.226**	0.566
RSO	0.253	**0.204**	0.801
σ^2	**8.9e-06**	-	-
σ_1^2	-	**0.047**	0.126
σ_2^2	-	**0.125**	1.266
σ_3^2	-	0.043	**0.038**
σ_4^2	-	**0.038**	0.106

The mean squared error (MSE) was computed to draw conclusions on the performance of each algorithm to estimate the model parameters (see Table 3.8). For each simulated dataset, the repetition corresponding to the highest likelihood value was selected for the computation of the MSE of the estimator. In addition, for MOO/NSGA-II, two procedures of selection were tested, i) selection of the solution minimizing the sum of the 4 objectives (Eq. 3.17) and ii) selection of the solution minimizing the worst error (Eq. 3.18). Results showed that the first procedure outperformed the second one, systematically resulting in lower MSE values. When compared to the estimations obtained by means of the single objective formulation (SOO/GA), the MOO/NSGA-II strategy was generally less accurate, yielding larger or equivalent MSE values. In particular, small MSE were obtained on the parameters LHx, RAI, KFk, RSS and TAS when SOO/GA was used, whereas estimations of the parameters OCp, RSO and KHk were comparable among the two strategies. One parameter, TPF, however, stood out as poorly predicted by the SOO/GA approach. It is interesting to notice that this parameter corresponded to the least sensitive to SOO objective function and this could explain that the algorithm wandered around a wide range of values without any impact on the prediction quality. On the other hand, this parameter ranked third for the objective function of fructose and is involved in the interactions for the remaining sugars. Therefore, the MOO/NSGA-II better accounted for these multiple involvements and thus this could partly explain the difference between the SOO and MOO results in favour of the last one.

TABLE 3.9 – Average values of the modelling efficiency (EF) for the four sugars over the simulated datasets. Values between parentheses represent the min-max range of the corresponding EF values, over the 100 replicate data for SOO and MOO, and the min-max range over the 100 virtual populations for Population-Based represent

Method	Sucrose	Sorbitol	Fructose	Glucose
SOO	0.999(0.993/0.999)	0.999(0.998/0.999)	0.999(0.998/0.999)	0.998(0.993/0.999)
MOO	0.979(0.949/0.999)	0.985(0.979/0.999)	0.918(0.837/0.999)	0.978(0.862/0.999)
Population-Based	0.996(0.991/0.998)	0.997(0.974/0.999)	0.991(0.978/0.994)	0.991(0.979/0.995)

Additionally, the quality of fit between SOO and MOO was compared by computing the Modelling efficiency (EF) on the dynamics of the four sugars, on the 100 simulated datasets (see Table 3.9). Results on sucrose, glucose and fructose were better using SOO (first strategy), and with an equivalent performance for sorbitol.

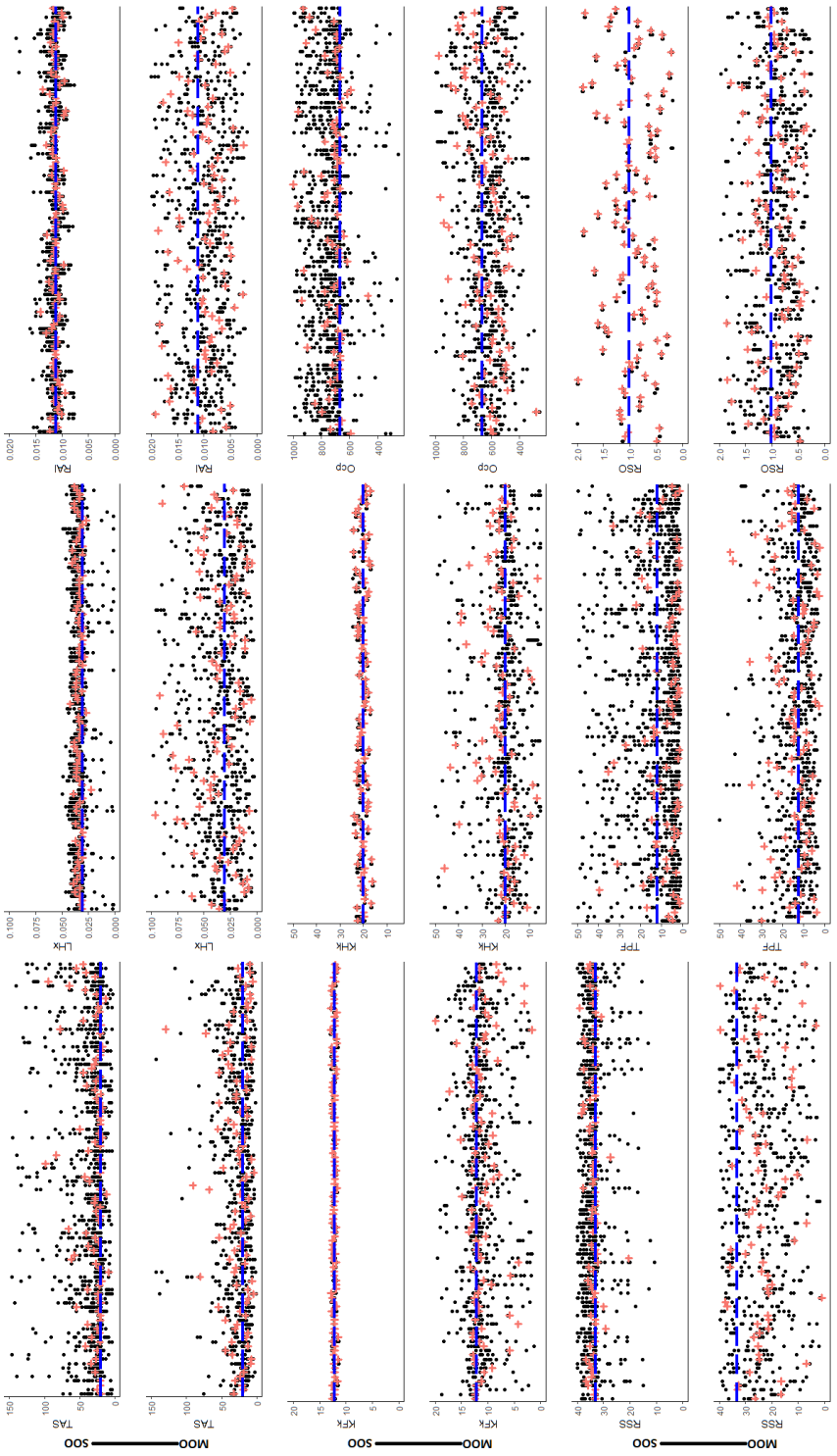


FIGURE 3.9 – Stripcharts of individual estimations of the nine parameters using the genotype independent calibration strategy for each replicate data. On the top panels : results for SOO formulation/ GA algorithm. On the bottom panels MOO formulation/NSGA-II algorithm. Blue dashed line represents the true value of the parameter. Red crosses represent the selected reference solutions. Only solutions having a maximum likelihood ±5% were considered.

Model Calibration for all genotypes simultaneously

Results of the 5×100 independent estimations of the parameters β showed a very stable behaviour of the MCMC-SAEM algorithm (see Figure 3.10). Whenever the simulated dataset and the starting value, the algorithm converged to the same final estimates, close to the true mean value of the nine parameters over the population.

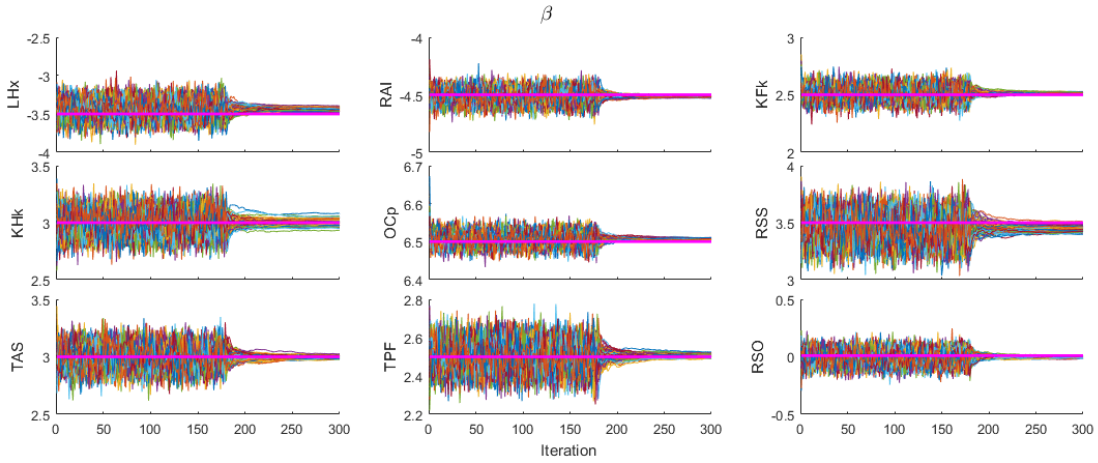


FIGURE 3.10 – Independent estimations along iterations (5×100) of β associated with (LHx, RAI, KFk, KHk, OCp, RSS, TAS, TPF, RSO) with the population-based strategy using MCMC-SAEM algorithm.

The mean squared error(MSE) of the estimator of $\theta = (\beta, \Gamma, \sigma^2)$ was computed for the population-based approach. Given the good convergence of the algorithm, the mean of the five repetitions was used as reference estimate. The estimator was reliable and accurate, with low MSE values for most parameters (see Table 3.10).

TABLE 3.10 – Results of mean squared error (MSE) obtained on simulated data with the MCMC-SAEM algorithm.

β	MSE	Γ	MSE
β_{LHx}	0.011	Γ_{LHx}	0.049
β_{RAI}	0.021	Γ_{RAI}	0.002
β_{KFk}	0.073	Γ_{KFk}	0.002
β_{KHk}	0.045	Γ_{KHk}	0.074
β_{OCp}	0.013	Γ_{OCp}	0.006
β_{RSS}	0.031	Γ_{RSS}	0.061
β_{TAS}	0.0004	Γ_{TAS}	0.003
β_{TPF}	6e-05	Γ_{TPF}	0.078
β_{RSO}	6e-05	Γ_{RSO}	0.002
σ^2	0.012		

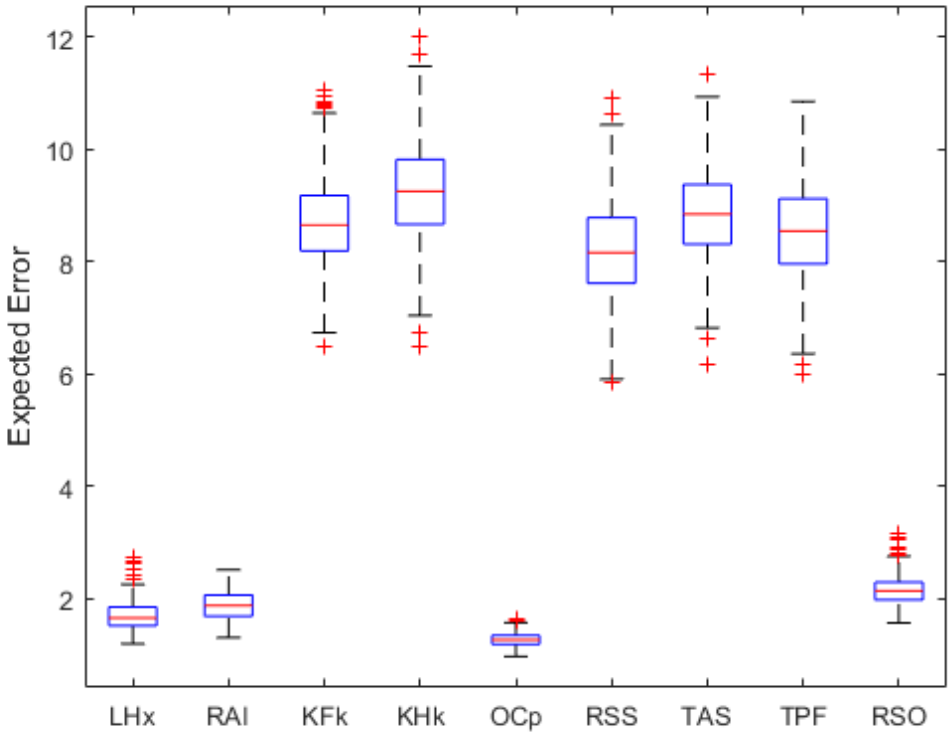


FIGURE 3.11 – Expected error (%) on the individual estimates of PB/MCMC-SAEM approach over 100 simulated populations and 5 repetitions of algorithm.

The good performances observed for the estimation of mean parameter values, also extend to the estimation of individual parameters. Figure 3.11 shows a comparison between estimated and true individual parameter values out of the 100 virtual populations. The expected error was estimated around 2% for LHx, RAI, OCp and RSO, and around 9% for KFk, KHk RSS, TAS and TPF. In addition, the results presented in Table. 3.9 showed a satisfactory agreement between model predictions and simulated data for the four sugars with an average of the modelling efficiency close to 0.9.

3.2.7.2 Experimental study

Model calibration for each genotype independently

SOO/GA and MOO/NSGA-II approaches were used for the independent calibration of the model on 106 peach genotypes from a peach population. Following the results of the simulation study, results from the first selection strategy only, for MOO/NSGA-II, are presented here for the real data.

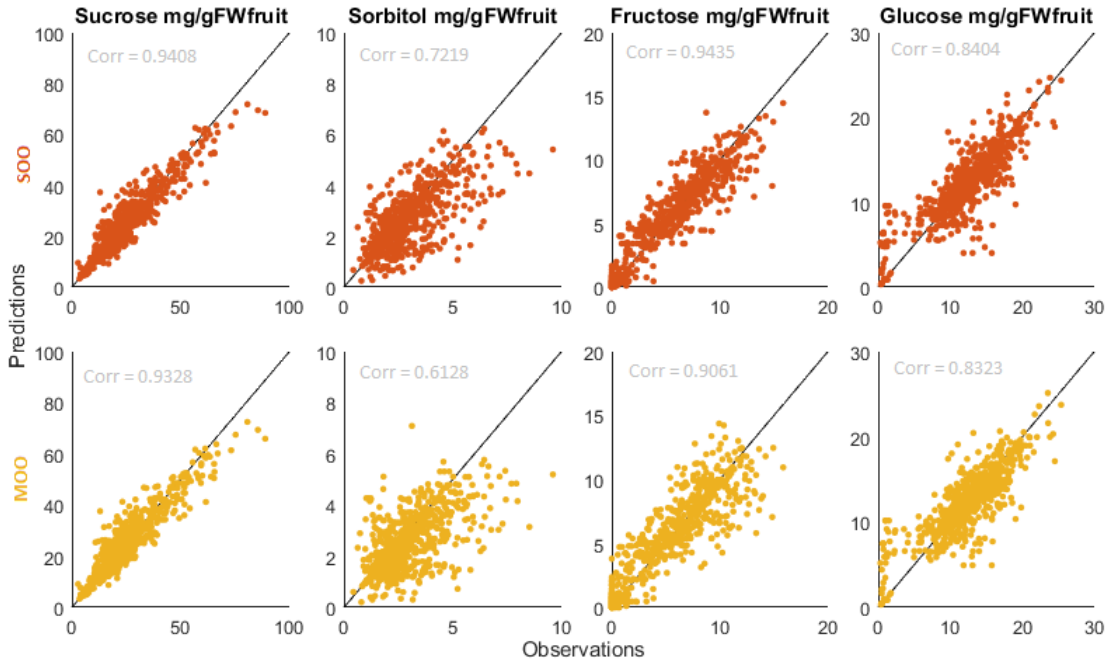


FIGURE 3.12 – Predictions vs observations using the reference estimates over the peach population. Top : results from (SOO/GA) Bottom : results from (MOO/NSGA-II). Corr : Correlations were computed between predictions and observations for each sugar and each calibration formulation/used algorithm.

Figure 3.12 shows the predictions-observations plot obtained for the four sugars using genotype-by-genotype calibration approaches. In general, the genotype-by-genotype estimation of the model provided acceptable predictions for the four sugars. Zooming on prediction from each formulation, predictions from SOO/GA were slightly better than those obtained with MOO/NSGA-II on the four outputs of the model (sucrose, sorbitol, fructose and glucose). Results were confirmed by computing correlation values between predictions and observations for each problem, underlying that a good fit was obtained for sucrose and fructose on the whole progeny of 106 genotypes with a correlation close to one followed by glucose and sorbitol respectively. Weak correlations were obtained for sorbitol compared to the others, that can be partially explained by a significant error on the experimental measurements due to the low concentration of sorbitol in the fruit.

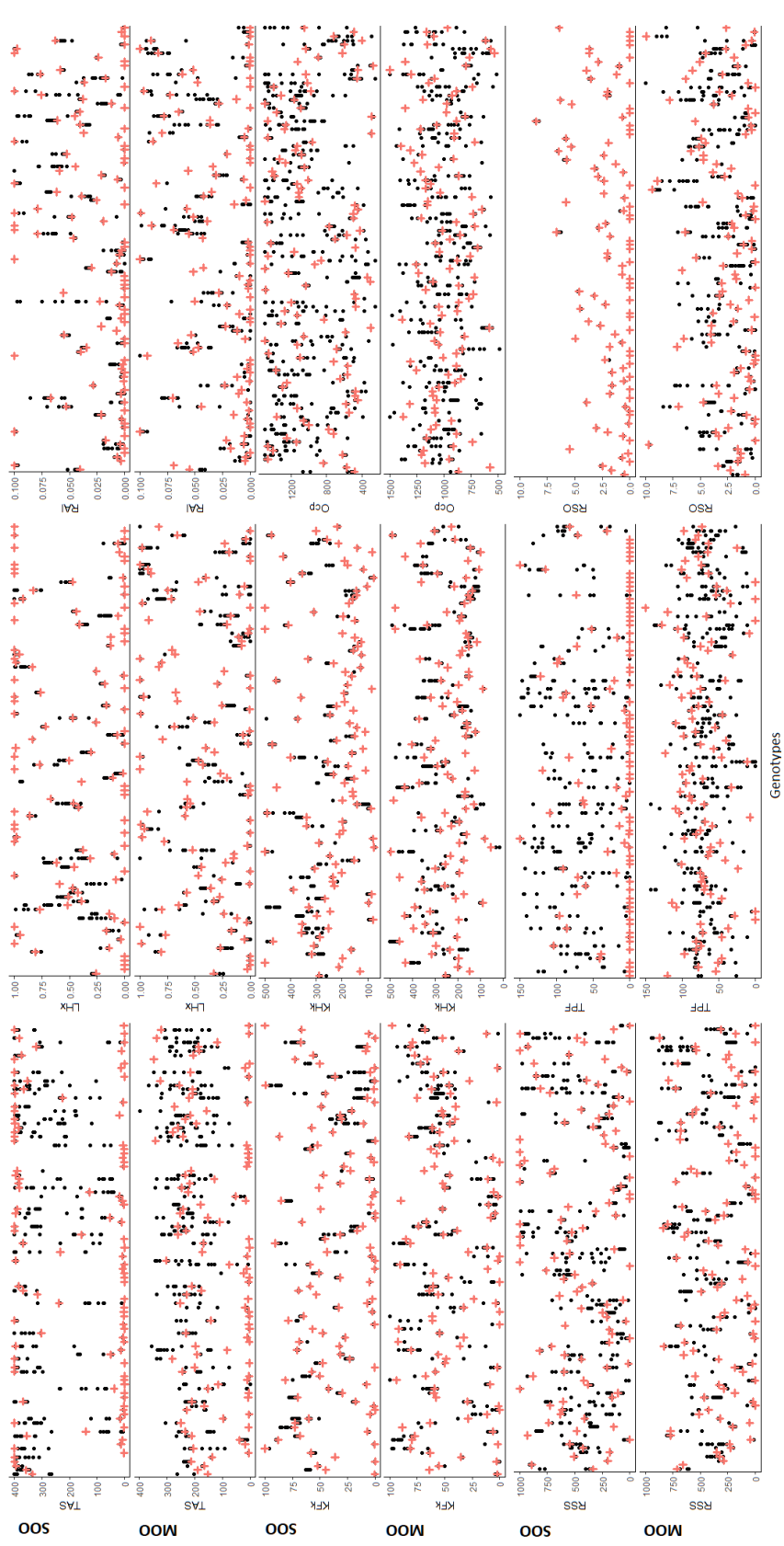


FIGURE 3.13 – Stripcharts of individual estimations of the nine parameters using genotype independent calibration strategy for the 106 genotypes of the peach progeny. On the odd lines : results for SOO formulation/GA algorithm. On the even lines : MOO formulation/NSGA-II algorithm. Only solutions having a maximum likelihood ±5% were considered. Red crosses represent the reference solutions selected for each genotype

When looking at the estimation of the underlying parameter values, however, results on the peach progeny confirmed the poor accuracy of genotype-by-genotype strategies, as already observed in the simulation study. Figure 3.13 shows the set of solutions having a maximum likelihood ±5%, for the nine estimated parameters with the two considered formulations, SOO and MOO. Results show that the degree of variation strongly depended on the genotype, the parameter and the considered optimisation formulation/algorithm. Overall, large variations in the estimated values were observed, especially for parameters TAS, TPF, OCp and RSS. It is interesting to notice that these parameters do not always correspond to high expected MSE values (see Table 3.8). Even when focusing on estimates having a large likelihood, the values of estimated parameters (see Figure 3.13, red crosses), were markedly different between the two formulation/optimisation schemes. This is particularly striking for a few parameters, as TAS or TPF for which most SOO estimates, but not MOO's, lie close to the boundary of the parameter range. A more quantitative comparison of estimated parameter values for the three calibration methods can be found at the end of the Result section.

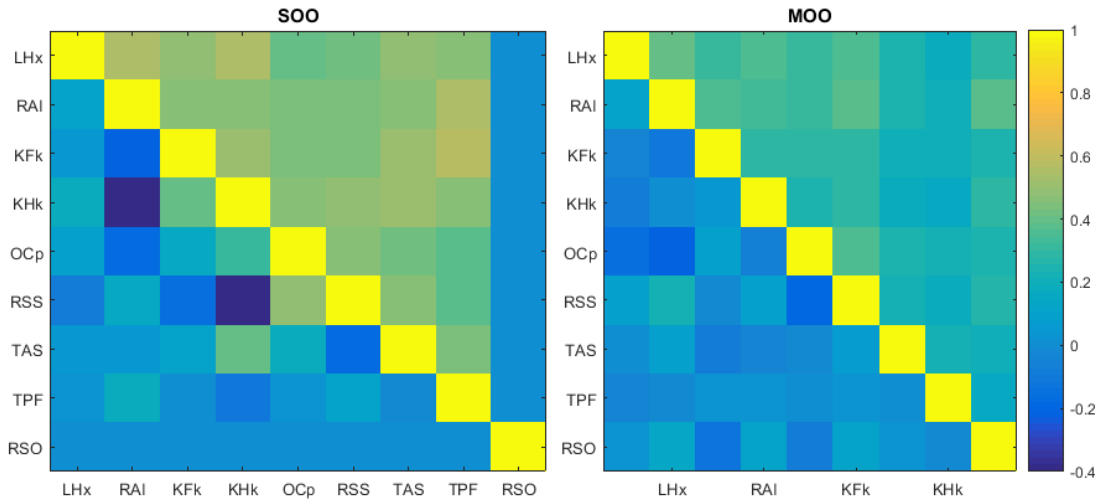


FIGURE 3.14 – Intra-genotype correlations i.e correlations between repeated estimations obtained using the genotype-by-genotype strategy, for the SOO and MOO formulations. Lower triangular matrix : mean intra-genotype correlations over the 106 genotypes of the peach population, upper triangular matrix : standard deviation of intra-genotype correlations on 106 genotypes

To better understand this point, we computed the average pair-to-pair correlation between parameters estimates of a same genotype (see Figure 3.14, low triangle). Results revealed a high degree of correlation among certain parameter estimates, especially for those obtained with the SOO/GA formulation. As an example, Figure 3.15 shows the ten estimations obtained for RAI (sensitive for sucrose concentration)

and KHk (sensitive for glucose concentration) parameters for a selection of genotypes. Plots clearly show that equivalent solutions (green dots) could be obtained by increasing RAI value and decreasing KHk, making the choice of the "true" estimation difficult on the basis of the objective function only. Indeed, the aggregation of the four sugars into a single fitting criterion (SOO) allows for offsets among the quality of individual sugar predictions, leading to the emergence of equivalent solutions with different underlying phenotype. In this perspective, the use of a multi-objective scheme, by individualising the prediction of the four sugars, reduced the correlation between parameter estimates for a same genotype compared to the SOO approach (see Figure 3.14, Right).

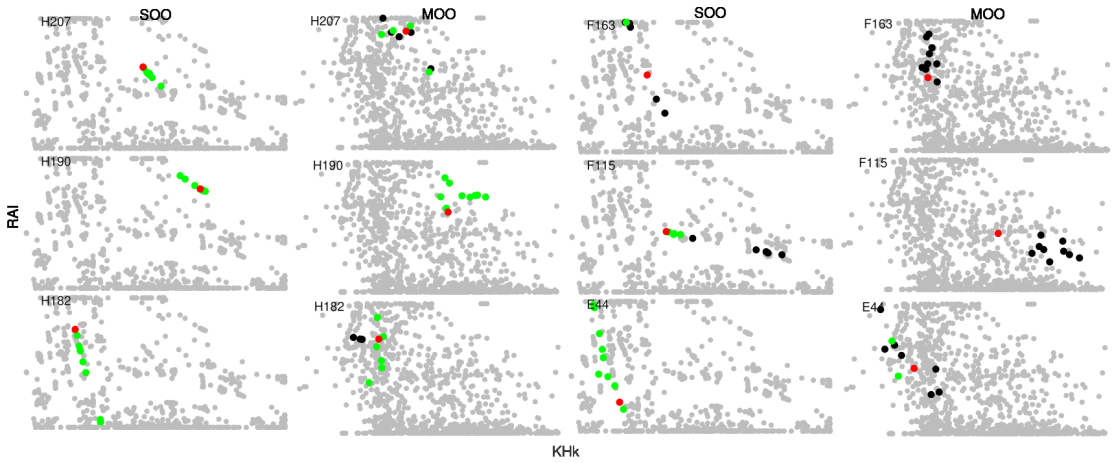


FIGURE 3.15 – Intra-genotype correlations between parameters RAI and KHk for SOO and MOO approaches, in the framework of the genotype-independent strategy. Red dots stand for the reference solution, green and black dots for solution(s) having respectively less or more than 5% deviation from the reference one. Grey dots represent the solutions obtained on the whole population of 106 genotypes

Model Calibration for all genotypes simultaneously

The population-based (PB) approach proposed by Baey et al. (2018) was applied to the population of 106 peach genotypes. Two structures for the covariance matrix were compared : i) a full covariance matrix (Γ_1) corresponding to the most general hypothesis of a correlation among all parameters in the model, and ii) a diagonal covariance matrix (Γ_2), *i.e.* assuming mutually independent parameters. Both models were calibrated on the whole population using the MCMC-SAEM algorithm. Results of the estimation are shown in tables 3.11 and 3.12, for Γ_1 and Γ_2 respectively. The inter-individual variability was high on the nine parameters of the model, for both Γ_1 and Γ_2. These results suggest the presence of a variability for each effect. Instead, low covariance values were observed among parameters for Γ_1 (Matrix 3.11), suggesting that the parameters could be only weakly dependent.

TABLE 3.11 – Estimated covariance matrix Γ_1.

Parameter	LHx	RAI	KFk	KHk	OCp	RSS	TAS	TPF	RSO
LHx	1.7								
RAI	5e-3	0.9							
KFk	3e-3	7e-3	1.9						
KHk	-6e-3	-5e-3	-2e-2	0.7					
OCp	-4e-3	-3e-3	-1e-3	5e-3	0.1				
RSS	1e-3	7e-3	-2e-3	-8e-3	3e-3	0.7			
TAS	8e-4	2e-3	-3e-3	-3e-3	1e-3	5e-4	0.7		
TPF	-7e-3	6e-4	4e-3	-3e-4	4e-3	-8e-5	-1e-3	0.7	
RSO	-6e-3	2e-3	5e-3	-1e-2	-1e-3	3e-3	1e-2	6e-3	0.8

TABLE 3.12 – Estimated covariance matrix Γ_2, assuming mutually independent parameters.

Parameter	LHx	RAI	KFk	KHk	OCp	RSS	TAS	TPF	RSO
diag(Γ_2)	0.7	1	1.9	0.7	0.2	0.7	0.8	0.8	0.9

The AIC criterion (see Section 3.2.9.1) was computed to select the best structure for the covariance matrix. The results of the AIC confirmed that the model with diagonal covariance matrix had to be preferred (AIC_{Γ_1} = 2396.1 vs AIC_{Γ_2} = 2299.7). For this reason, the model with diagonal covariance ($\Gamma = \Gamma_2$) was selected for the subsequent analyses.

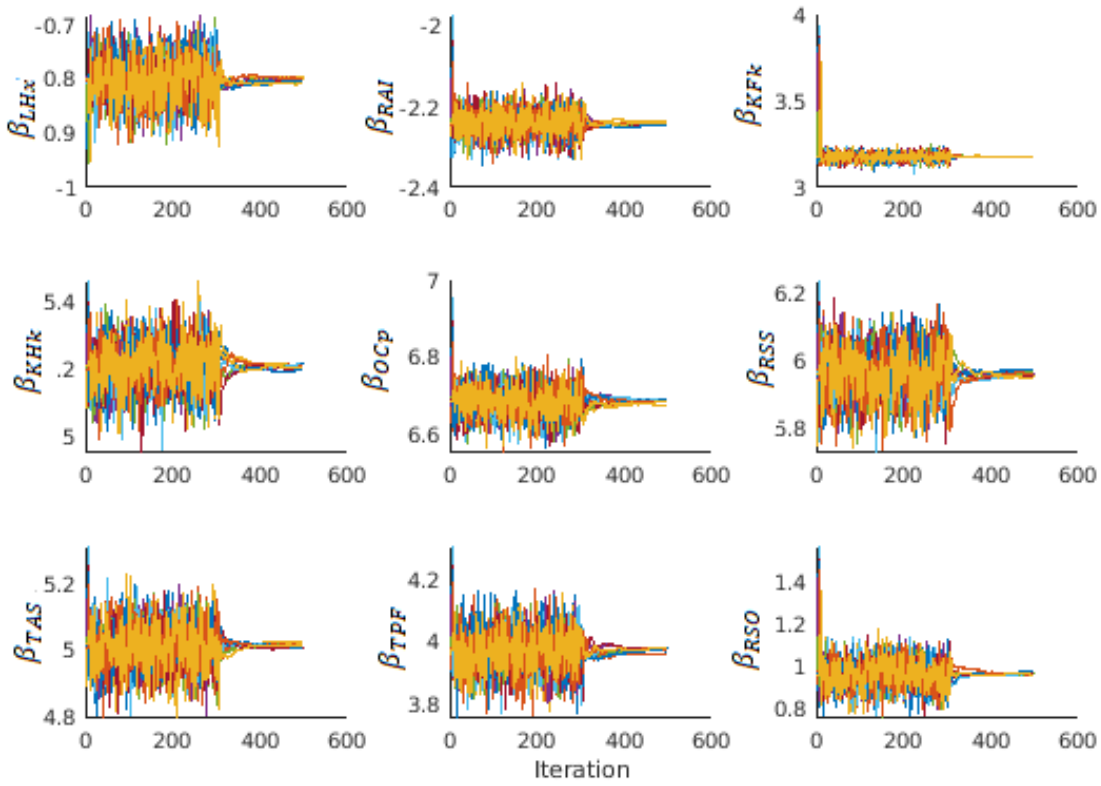


FIGURE 3.16 – Evolution along iterations of ten independent estimations of β associated with LHx, RAI, KFk, KHk, OCp, RSS, TAS, TPF, RSO, using MCMC-SAEM algorithm, on data from the peach progeny.

Using this model, we thus evaluated the quality of parameter estimation with PB approach, both in terms of fit quality and robustness of parameter estimates. Satisfactory agreement between predictions and observations was obtained on the four sugars using individual parameters (see Figure 3.17). Some outliers were observable though, especially for sorbitol and glucose.

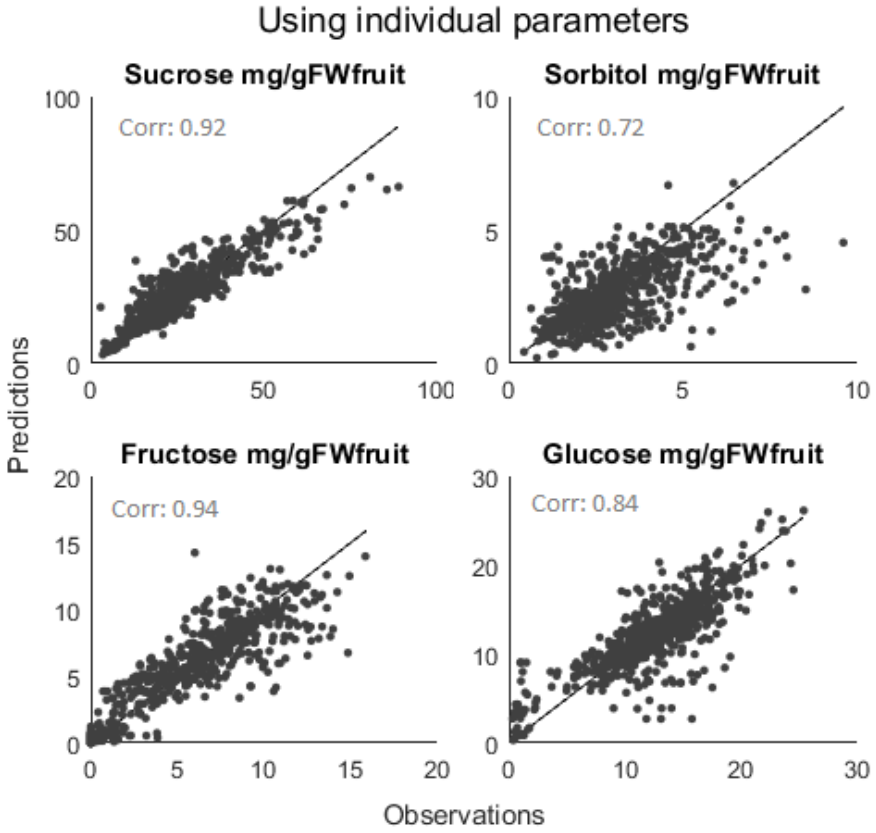


FIGURE 3.17 – Predictions vs observations using individual MCMC-SAEM estimates over the 106 genotypes of the peach progeny. Corr : correlations between predictions and observations

Results of ten independent estimations of the parameter β are shown in Figure 3.16. The corresponding figures for σ^2 and Γ can be found in appendix 3.2.9 (see Figure 3.23). The different runs showed a very stable behaviour and converged close to the same final estimates for the mean population value of the nine parameters, independently of the starting value. In agreement with this observation, Figure 3.18 shows that the range of the estimates of individual parameters was small for most of the genotypes, meaning that the algorithm allowed a robust estimation of both means and individual parameter values of the population. Table 3.13 reports the confidence intervals of the vector of parameters $\theta = (\beta, \Gamma, \sigma^2)$ (see Section 3.2.5.3 for details on their computation).

TABLE 3.13 – Population-Based calibration : results of the estimation of mean population parameters θ for the solution corresponding to the highest likelihood

Parameter	Estimate	SD	95% CI
β_{LHx}	-0.810	0.019	[-0.843 ; -0.782]
β_{RAI}	-2.240	0.005	[-2.284 ; -2.197]
β_{KFk}	3.176	0.023	[3.077 ; 3.280]
β_{KHk}	5.202	0.014	[5.189 ; 5.208]
β_{OCp}	6.682	0.006	[6.584 ; 6.720]
β_{RSS}	5.944	0.010	[5.877 ; 6.115]
β_{TAS}	5.014	0.010	[4.885 ; 5.135]
β_{TPF}	3.978	0.015	[3.809 ; 4.049]
β_{RSO}	0.961	0.036	[0.895 ; 1.039]
Γ_{LHx}	0.867	0.018	[0.791 ; 0.937]
Γ_{RAI}	1.001	0.007	[0.992 ; 1.021]
Γ_{KFk}	1.937	0.012	[1.887 ; 1.987]
Γ_{KHk}	0.776	0.016	[0.751 ; 0.803]
Γ_{OCp}	0.199	0.001	[0.198 ; 0.204]
Γ_{RSS}	0.727	0.011	[0.707 ; 0.752]
Γ_{TAS}	0.802	0.011	[0.795 ; 0.839]
Γ_{TPF}	0.883	0.014	[0.855 ; 0.913]
Γ_{RSO}	0.961	0.011	[0.971 ; 1.019]
σ^2	0.563	0.034	[0.502 ; 0.571]

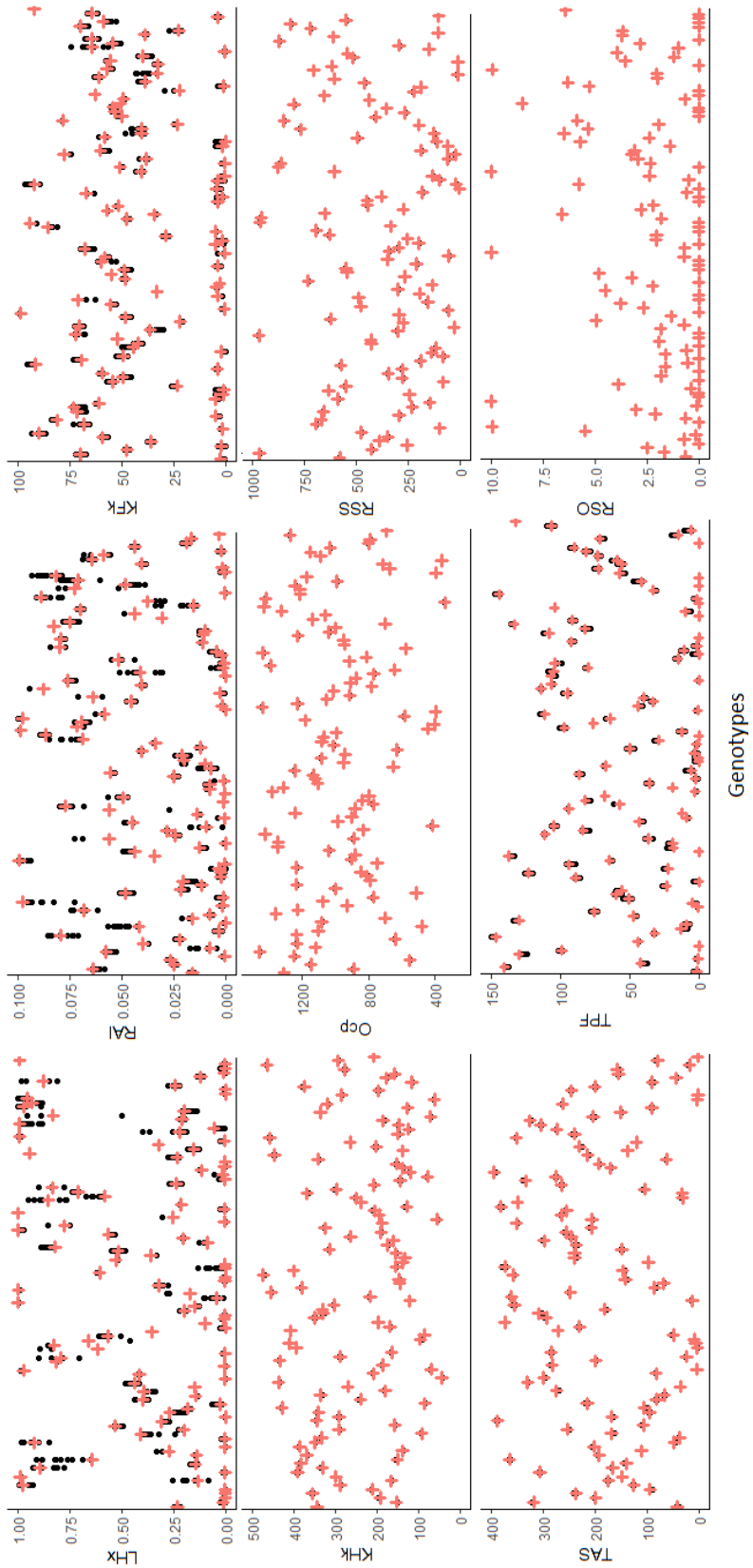


FIGURE 3.18 – Variability across individual parameters obtained over ten independent runs (black points) of the MCMC-SAEM algorithm on the whole progeny of 106 genotypes. Red crosses represent the solutions corresponding to the highest likelihood.

Comparison of the two calibration strategies

The Modelling efficiency (EF) was computed to compare numerically the predictions from the two calibration strategies on the dynamics of the four sugars, on the 106 genotypes of the population (see Table 3.14). Results on sucrose and glucose were better using SOO (genotype-by-genotype calibration), followed by PB and then MOO. PB and SOO approaches outperformed MOO in the prediction of fructose concentration, whereas equivalent results were obtained for sorbitol.

TABLE 3.14 – Average values of the modelling efficiency (EF) for the three calibration strategies and for the four sugars. Values between parentheses represent the min-max range over the 106 genotypes.

Method	Sucrose	Sorbitol	Fructose	Glucose
SOO	0.88(0.75/0.98)	0.45(0.30/0.98)	0.88(0.41/0.98)	0.70(0.42/0.98)
MOO	0.87(0.71/0.96)	0.25(0.22/0.96)	0.82(0.32/0.98)	0.69(0.38/0.97)
Population-Based	0.84(0.73/0.97)	0.45(0.29/0.99)	0.88(0.41/0.98)	0.68(0.38/0.98)

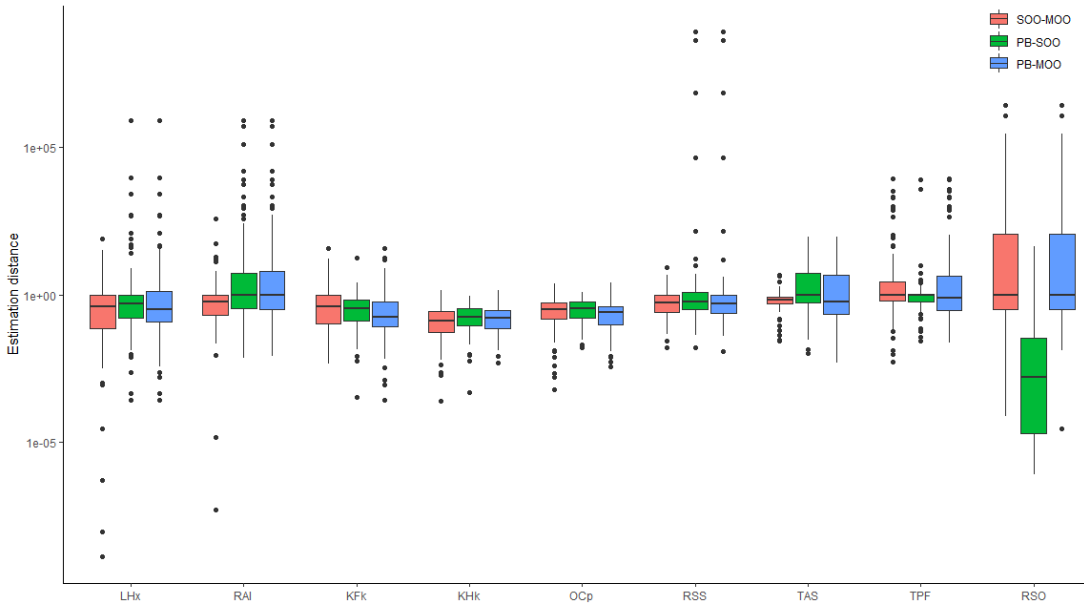


FIGURE 3.19 – Normalized pair-to-pair distance between the reference parameter estimates obtained with the three calibration methods, over the 106 genotypes of the peach progeny. Distances were normalized with respect to the median estimated value for each parameter and genotype in the population.

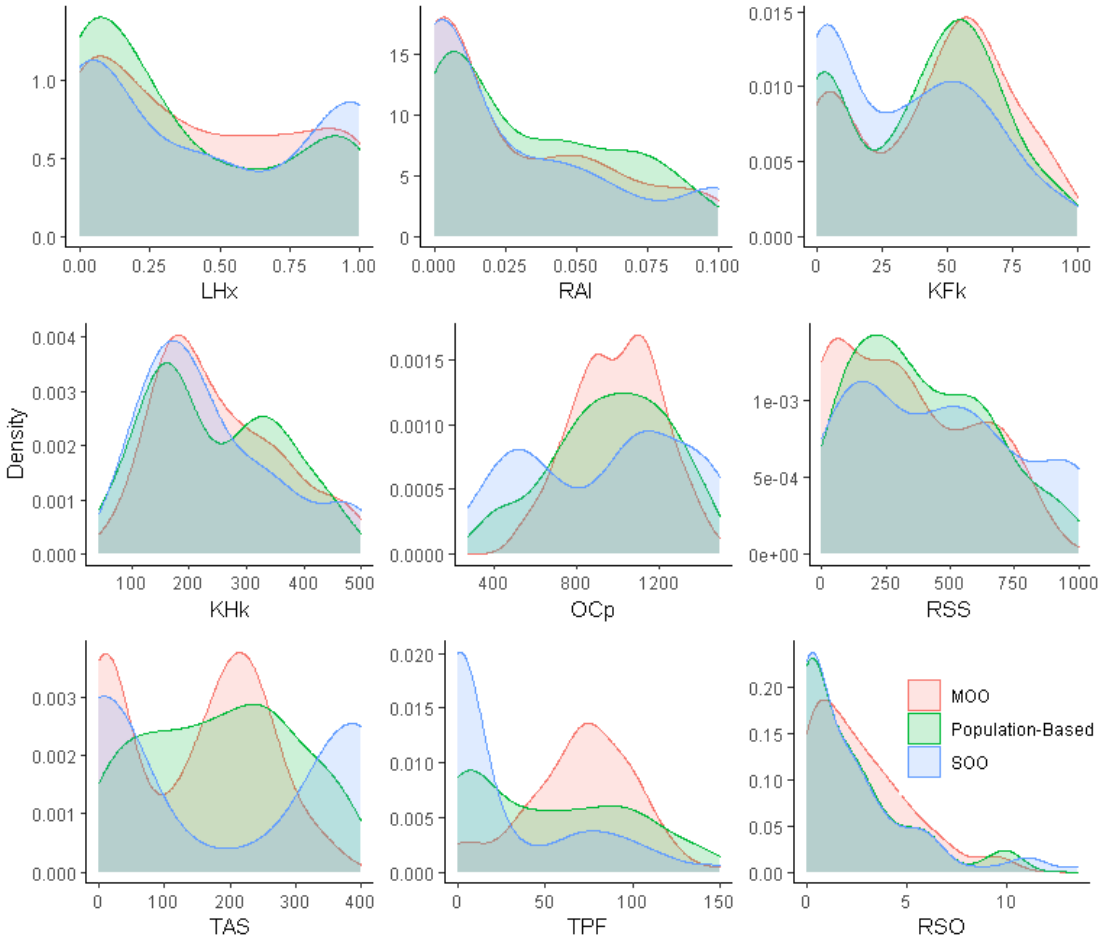


FIGURE 3.20 – Distribution of parameter values over the peach population obtained with the three considered methods : single-objectif (SOO), multi-objectives (MOO) and population-based (PB) approach.

When looking to estimated parameter values, the differences between the three approaches became more pronounced. Figure 3.19 compares the reference estimates obtained with the three calibration methods for the 106 genotypes of the peach population. On average, estimated values were quite close, but huge quantitative differences (up to five order of magnitude) appeared for a handful of genotypes. This is particularly true for the estimated values of RSO and RSS and, to a lesser extent, of LHx, RAI, TPF parameters. Figure 3.20 shows the distribution of individual parameters obtained for the whole progeny. Despite the above-mentioned discrepancies in the estimates, the overall range of estimated values as well as the shape of the distribution for parameters LHx, RAI, KFk, KHk, RSS et RSO were quite similar across the three methods. Stronger differences emerged instead for the distributions of OCp, TAS and TPF, that may be linked to the bad identifiability of these parameters with the genotype-by-genotype strategy.

Interestingly, the distribution of KFk was markly bimodal, suggesting the presence

of two sub-populations, with contrasted affinity values. This result is in agreement with the prediction of the model (Desnoues et al. 2018) according to which the fructokinase affinity may be responsible for the appearance of two distinct fructose phenotypes, as observed in the peach population. In order to further investigate the relationships between the parameter values and the resulting fructose concentration, we systematically compared the estimations obtained over the two phenotypic populations, with the three calibration methods (see Figure 3.21). Results confirmed a significative and robust difference in the estimated values of KFk for the two phenotypic groups, corroborating the involvement of the fructokinase affinity in the specification of the fructose content of the fruit. Other parameters emerged different between the two phenotypic groups when considering the results from the genotype-independent calibration methods, but these results were not supported by the population-based estimations. This was notably the case of KHk and TPF parameters that were also pointed out as putative mechanisms in the original modeling work by Desnoues et al. (2018) (using an SOO approach). These results underline the importance of the choice of the calibration method which may impact the biological interpretation of the results.

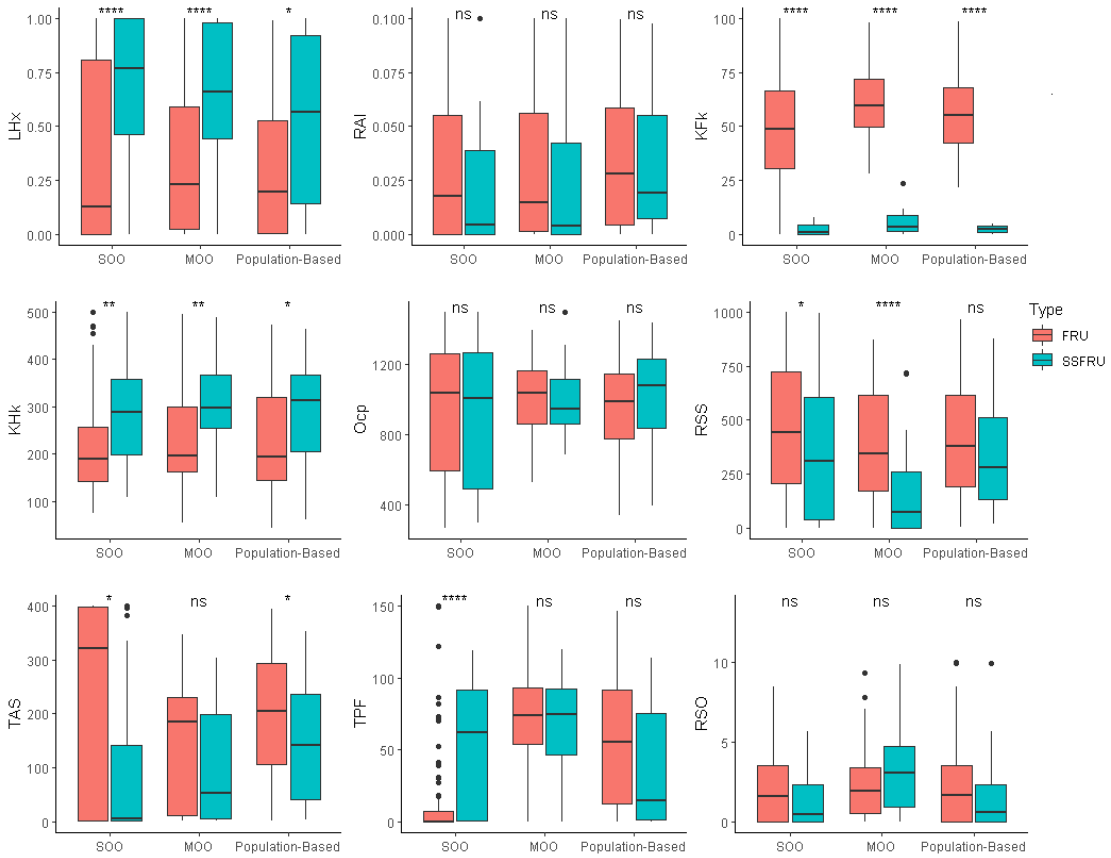


FIGURE 3.21 – Difference in the estimated model parameters between standard ('FRU') and low fructose phenotypes ('SSFRU'), over 106 genotypes obtained with the three considered methods : single-objectif (SOO),multi-objectives (MOO) and population-based (PB) approach.

Last but not least, the two optimisation strategies were compared for their calibration time (see Table 3.15). The population-based strategy turned out to be 250 times faster than the genotype-by-genotype strategy, over a single iteration (generation) and over the whole population. This is more valuable and appreciated if we take into account the dimensions of explored spaces in each strategy. Indeed, the SOO and MOO (first strategy) estimated, respectively, 9 and 13 (9 model parameters plus 4 standard deviations) unknown parameters whereas the PB (second strategy) estimated 19 unknown parameters (assuming a diagonal covariance matrix).

TABLE 3.15 – Average calibration time for the three considered methods over the peach population.

Method	Time per iteration (s)	Time for the whole progeny
SOO	≈ 6.25	≈ 2 days
MOO	≈ 6.27	≈ 2 days
Population-Based	≈ 0.025	≈ 13 minutes

3.2.8 Discussion

In the context of agronomy, models are increasingly used to describe plant/organ development and metabolism, testing their performances in different conditions. Being able to account for intra-specific variability is a key challenge in the field as it can provide useful information for varietal selection and plant breeding.

In this paper, a mathematical model of sugar metabolism in peach fruit (Kanso et al. 2020) was first calibrated on a set of simulated data and then on a progeny of 106 genotypes using two contrasted strategies that strongly differ in the way data are *perceived* and analysed. The first strategy corresponds to the classical way to proceed : parameter estimation is addressed on each genotype independently, and the obtained estimates compared in order to identify the genotypic parameters of the system (Bertin et al. 2010). Examples of this approach can be found in the study of many plant processes, including leaf elongation (Reymond et al. 2003), plant development (Yin et al. 2000; Yin et al. 2005; Messina et al. 2006), phenology (Nakagawa et al. 2005), nitrogen adaptation laperche2006, fruit growth (Constantinescu et al. 2016) and quality (Quilot et al. 2005; Prudent et al. 2011). In particular, we considered two options to describe the agreement between model and simulated and experimental data, either as a single objective optimisation (SOO) or as a multi-objectives optimisation (MOO) problem. In the first case i.e. SOO/GA formulation, the estimation process results in an unique value for the objective function (goodness of fit). Hopefully, this value of objective function could correspond to a unique set of parameter values for each genotype of the progeny. Whereas using the MOO formulation, an entire set of trade-off between antagonist objective functions (Pareto front) is identified for each genotype, among which the desired parameter set has to be selected based on some other criteria. This choice could be done using graphical tools, decision-making methods, or some particular heuristics designed for a particular classification problem (Coello et al. 2007; Akdemir 2019).

The second calibration strategy relies on the use of nonlinear mixed-effect model (Davidian et al. (2003) et Baey et al. (2018)) and applies to all genotypes simultaneously, following a population-based (PB) perspective. Accordingly, parameters from individual genotypes are considered as random variables and supposed to follow the same probability distribution. The aim of the estimation process shifts from the individuals to the population, the result of the calibration being the parameters of such a distribution, namely the average population values and the covariance matrix, in addition to

individual genotypic parameters.

When applied to the model proposed by Kanso et al. (2020), all approaches showed a satisfactory agreement between predictions and data, for the whole progeny of 106 genotypes. In term of individual (phenotypic) predictions, the SOO/GA formulation of the first strategy generally provided better results, especially concerning the dynamics of sucrose and glucose, the prevailing sugars in peach fruit. However, when looking at the corresponding parameter values, the first strategy suffered from a lack of reproducibility/variability, with several parameters sets giving an equivalent agreement with data. This is a well-known issue of collective fit in system biology : in spite of a good predictive power, parameter estimation may show large uncertainty, in correspondence to sloppy direction of the parameter space (Gutenkunst et al. 2007). Although in many biological applications, useful insights may be derived from model predictions rather than parameter values, uncertainty in parameter estimates poses a problem for genetic analysis, in which the parameters values are expected to be the genetic fingerprint of the system. Moreover, spurious correlations among parameter estimates along with sloppy directions, may prevent the identification of true genetic correlations due to epistastic effects or common regulatory mechanisms.

The quality and the nature of available data are important determinants of the resulting accuracy of parameter estimation. Indeed, a reasoned experimental design can considerably reduce the uncertainty on parameter values, even in the case of collective fit (Raue et al. 2009; Apgar et al. 2010). In this paper, we showed that the choice of the calibration strategy may have an impact too. The use of a multi-objective formulation in the first strategy, that separates the optimisation of individual sugars, helped to reduce correlation among parameter estimates (see Figure 3.14) and the possibility of offset among the quality of prediction of the different sugars, an issue of the single-objective formulation. However, the selection of a pertinent solution on the Pareto front remains a major concern. Our simulation study showed indeed that the criterion used for selection can strongly affect the quality of the estimator, as measured by the mean-squared error (see Table 3.8). Moreover, even within a same selection criteria, "equivalent" trade-off solutions can correspond to very different parameters estimates, making difficult to identify an unique parameter set for each genotype. The observed variability and lack of reproducibility observed with the first strategy (genotype-by-genotype calibration) could be linked to the results (not shown) of the sensitivity analysis conducted separately on the objective function of SOO and those of MOO. This sensitivity analysis has shown a largely dominant effect of genotype over the considered parameters on the model outputs (sugars). It shows also some interactions (reflecting in somewhat correlations) between parameters in both cases (SOO and MOO). This is particularly true in the case of experimental data where the inputs of our model depend on the genotype and thus it is unsurprising to see accordingly the variability of parameter estimations between genotypes. Regarding the poor performances of MOO/NSGA-II compared to SOO/GA in terms of phenotype predictions, it could be worth here to mention this is well known that NSGA-II is particularly effective for optimisation problems involving 2 or 3 objective functions

but its performances decrease for more than three (many-objective optimisation) and become poor as the dimension increases (Wang et al. 2013).

In light of our results, the use of genotype-by-genotype calibration approaches in the context of genetic studies is not recommended. Better results were undoubtedly obtained with the population-based approach. The advantage of this approach is that it takes into account the observation of the whole progeny and cast them in a same general framework (common probability distribution). The presence of multiple genotypes with a similar sugar concentration thus helped to further constraint the estimation of parameters, resulting in a strong reproducibility and high accuracy of the estimates (see Table 3.13).

In addition, the use of population-based strategy allowed for the direct analysis of genetically-relevant correlations among parameters, via the selection of the covariance matrix. Results on our dataset showed that a diagonal covariance matrix, corresponding to the parameter independence, was the best choice according to the AIC criterion. This suggests that the genetic control of the biochemical reactions involved in peach sugar metabolism may be shared among many different mechanisms, in complex regulatory network (Desnoues et al. 2016; Carreno-Quintero et al. 2013). On the other hand, an important inter-genotypic variability was observed for most estimated parameters. The parameter KFk corresponding to the affinity of the fructokinase showed the larger variations. A closer look to estimation results revealed that the corresponding probability distribution was bimodal *i.e.* the observed variability corresponds indeed to the presence of two sub-population of genotypes, with different distributions. This result agrees remarkably well with the existence of two distinct phenotypes in the population, with contrasted fructose contents, and with previous analysis (Desnoues et al. 2018) suggesting KFk as the main determinant of such a phenotypic trait. The fact that PB estimations could distinguish among different populations is extremely promising and corroborates the suitability of this approach for genetic studies. The advantages of PB strategy with respect to a genotype-by-genotype approach, however, may partly depend on the population size. In Baey et al. (2018) parameters estimations over a population of 34 genotypes were less stable, due to convergence problems. Moreover, this result was sensitive to the parametrisation of the algorithm, further complicating the estimation process. Last but not least, the fundamental hypothesis of the approach, namely that individual genetic parameters are realization of a same probability distribution, might not be valid when multi-specific or natural (not inbred) populations are considered.

In perspective, a reliable estimation of model parameters will offer the possibility to study their genetic architecture, looking for genomic regions (QTL) linked to the parameter modification (Cooper et al. 2009; Yin et al. 2016). At term, the integration of the genetic control into ecophysiological models will permit to simulate different environmental scenarii , allowing for in-depth analysis of genotype x environment interactions. This will open the door towards new opportunities of virtual breeding i.e. the design of new genotypes that better meet the challenges of modern agriculture

(e.g. increased production, new cultivation techniques, future climates).

3.2.9 Appendices

3.2.9.1 Estimation of the likelihood

To compare different models using AIC criterion, we need to compute the likelihood at $\hat{\theta}$, where $\hat{\theta}$ is the value of θ that maximizes the likelihood. It is possible to approximate the likelihood at a given point, using the method of importance sampling (Robert et al. 2013). Let us denote by $\widetilde{y} = (\widetilde{y}_{ij}^{(k)}, 1 \leq k \leq G, 1 \leq j \leq n_{SG}, 1 \leq i \leq N_j^{(k)})$, $f(\widetilde{y}^{(k)};\theta)$ the marginal likelihood of the k-th individual, and by $\ell(\theta)$ the joint log-likelihood

$$\ell(\theta) = \log\left(\prod_{k=1}^{G} f(\widetilde{y}^{(k)};\theta)\right) = \sum_{k=1}^{G} \log\left(\int f(\widetilde{y}^{(k)} \mid \varphi^{(k)};\theta)\, p(\varphi^{(k)};\theta)\, d\varphi^{(k)}\right)$$

where $f(\cdot \mid \varphi^{(k)};\theta)$ is the conditional probability density function of $y^{(k)}$ given the random effect $\varphi^{(k)}$ (i.e. a Gaussian distribution with mean given by the model and variance σ^2), and $p(\cdot;\theta)$ is the probability density function of the random effect $\varphi^{(k)}$ (i.e. a multivariate Gaussian distribution with mean β and covariance matrix Γ).

Importance sampling can be used to approximate an integral, using an instrumental distribution q, and noticing that :

$$\int f(\widetilde{y}^{(k)} \mid \varphi^{(k)};\theta)\, p(\varphi^{(k)};\theta)\, d\varphi^{(k)} = \int f(\widetilde{y}^{(k)} \mid \varphi^{(k)};\theta) \frac{p(\varphi^{(k)};\theta)}{q(\varphi^{(k)};\theta)} q(\varphi^{(k)};\theta)\, d\varphi^{(k)}.$$

If q is well chosen, the second integral can be easier to approximate using Monte-Carlo method. In our case, we can use a Gaussian distribution, with mean $\hat{\beta}$ and covariance matrix $\hat{\Gamma}$. In particular, in this case we have $p = q$, the instrumental distribution q is equal to the distribution p of the random effects. More precisely, to compute the likelihood at the final estimate point $\hat{\theta}$, we can generate for k-th individual, N_{sp} random variables $\varphi_{sp}^{(k)} \sim \mathcal{N}(\hat{\beta}, \hat{\Gamma})$. The marginal likelihood of the k-th individual $f(\widetilde{y}^{(k)};\theta)$ can

be estimated by

$$\begin{aligned}
\hat{f}(\tilde{y}^{(k)};\hat{\theta}) &= \frac{1}{N_{sp}} \sum_{sp=1}^{N_{sp}} f(\tilde{y}^{(k)} \mid \varphi_{sp}^{(k)};\hat{\theta}) \, \frac{p(\varphi_{sp}^{(k)};\hat{\theta})}{q(\varphi_{sp}^{(k)};\hat{\theta})} \\
&= \frac{1}{N_{sp}} \sum_{sp=1}^{N_{sp}} f(\tilde{y}^{(k)} \mid \varphi_{sp}^{(k)};\hat{\theta}) \\
&= \frac{1}{N_{sp}} \sum_{sp=1}^{N_{sp}} \prod_{j=1}^{n_{SG}} \prod_{i=1}^{N_j^{(k)}} f(\tilde{y}_{ij}^{(k)} \mid \varphi_{sp}^{(k)};\hat{\theta}) \\
&= \frac{1}{N_{sp}} \sum_{sp=1}^{N_{sp}} \prod_{j=1}^{n_{SG}} \prod_{i=1}^{N_j^{(k)}} \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{\left(\tilde{y}_{ij}^{(k)} - \widetilde{\mathscr{M}}_{ij}(\varphi_{sp}^{(k)})\right)^2}{2\hat{\sigma}^2} \right) \\
&= \frac{1}{N_{sp}} \sum_{sp=1}^{N_{sp}} \frac{1}{\hat{\sigma}^{N^{(k)}}(2\pi)^{N^{(k)}/2}} \exp\left(-\frac{1}{2\hat{\sigma}^2} \sum_{j=1}^{n_{SG}} \sum_{i=1}^{N_j^{(k)}} \left(\tilde{y}_{ij}^{(k)} - \widetilde{\mathscr{M}}_{ij}(\varphi_{sp}^{(k)})\right)^2 \right)
\end{aligned}$$

with $N^{(k)} = \sum_{j=1}^{n_{SG}} N_j^{(k)}$. We can obtain an unbiased estimate of the likelihood at $\hat{\theta}$ by taking :

$$\hat{L}(\hat{\theta}) = \prod_{k=1}^{G} \hat{f}(\tilde{y}^{(k)};\hat{\theta}).$$

and then, we can compute the log-likelihood :

$$\hat{\ell}(\hat{\theta}) = log\hat{L}(\hat{\theta}) = \sum_{k=1}^{G} \ln \hat{f}(\tilde{y}^{(k)};\hat{\theta}).$$

3.2.9.2 Γ and σ^2 estimated on the artificial data

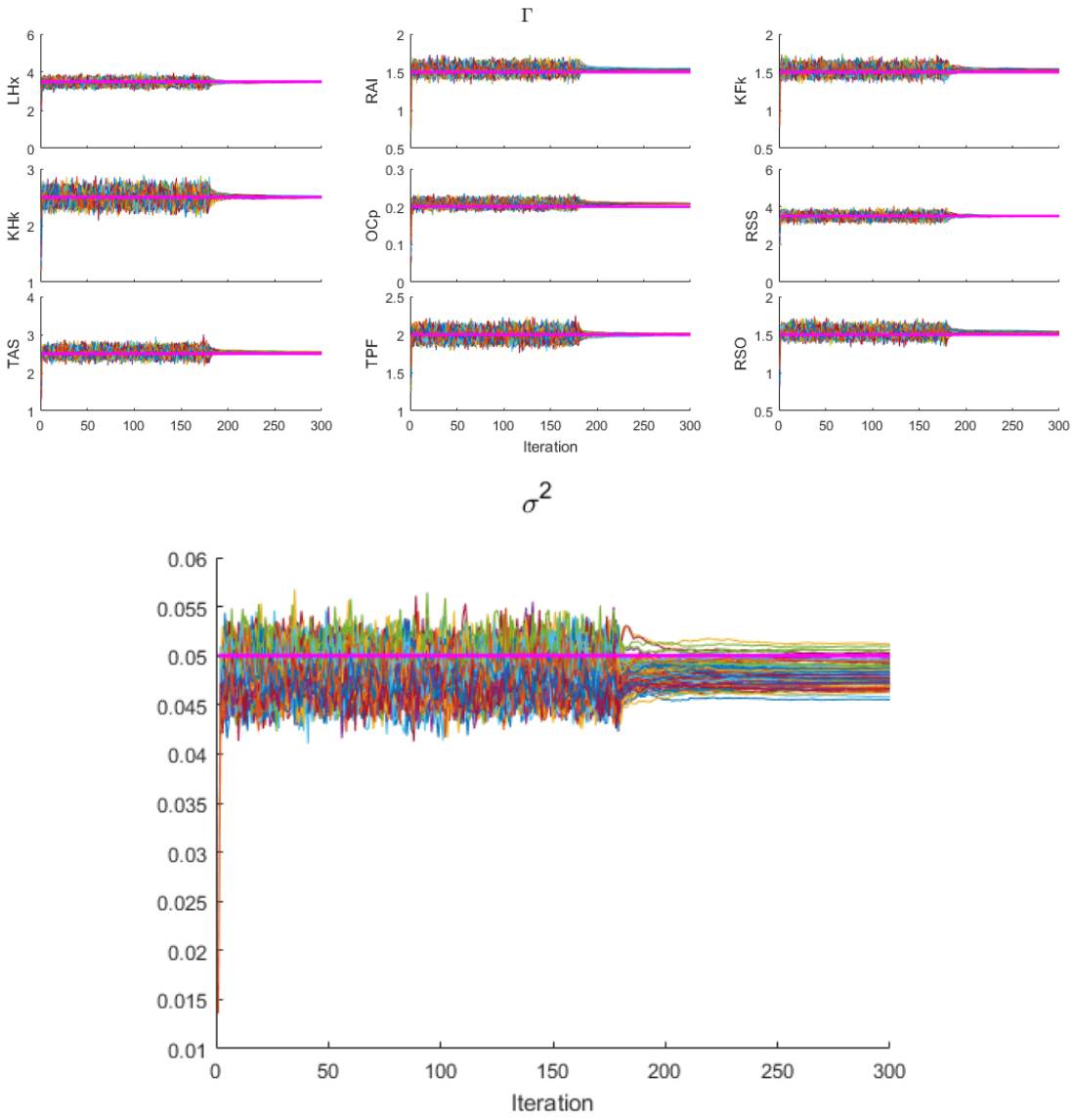


FIGURE 3.22 – 5×100 independent estimations of (Γ, σ^2) associated with (LHx, RAI, KFk, KHk, OCp, RSS, TAS, TPF, RSO) with the second strategy using MCMC-SAEM algorithm on the artificial data

3.2.9.3 Γ and σ^2 estimated on the experimental data

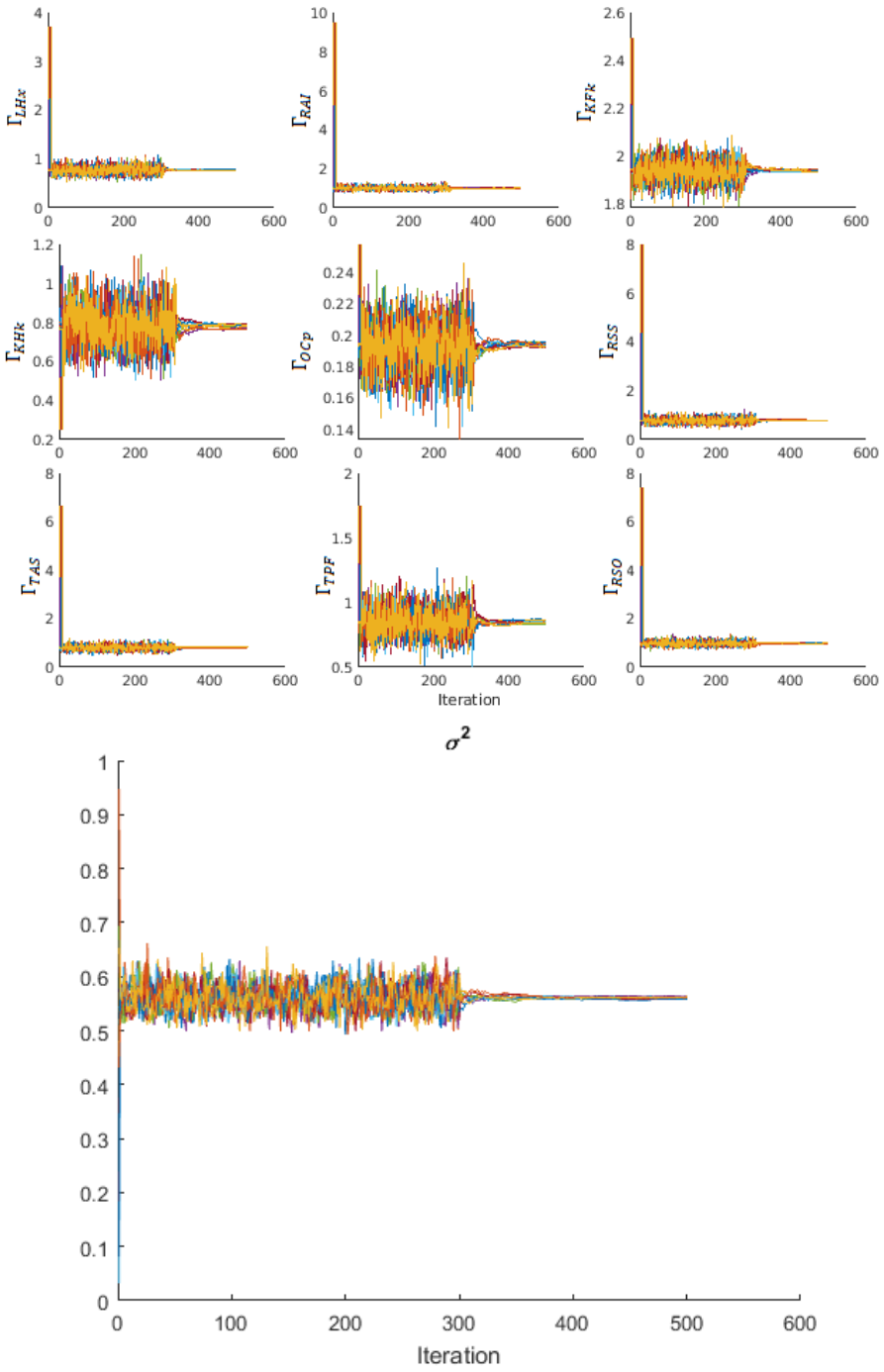


FIGURE 3.23 – Ten independent estimations of (Γ, σ^2) associated with (LHx, RAI, KFk, KHk, OCp, RSS, TAS, TPF, RSO) with the second strategy using MCMC-SAEM algorithm on the experimental data

3.3 Conclusions et perspectives

Conclusion

La calibration du modèle réduit en utilisant deux stratégies (calibration du modèle pour chaque génotype indépendamment et calibration du modèle pour tous les génotypes simultanément) a permis

- de calibrer le modèle avec succès pour 106 genotypes
- d'estimer 9 paramètres génotype-dépendant pour toute la population
- d'augmenter la fiabilité des estimations en utilisant la deuxième stratégie
- d'obtenir un bon ajustement avec les deux stratégies et une bonne caractérisation de la variabilité de la population

Perspectives

Une estimation fiable des paramètres du modèle offrira la possibilité

- de réaliser une analyse de la variabilité inter-genotypes et des interactions entre eux
- d'étudier leur variabilité génétique
- de rechercher des régions génomiques (QTL) liées à la modification des paramètres

4 Analyse du contrôle génétique du métabolisme des sucres chez la pêche assistée par le modèle cinétique réduit

Sommaire

DANS LE CHAPITRE précédent, le modèle réduit (Kanso et al. 2020) a été calibré avec succès pour les 106 génotype de la population BC2 en appliquant plusieurs approches d'optimisation.

Dans ce chapitre, nous nous sommes intéressés à l'analyse du contrôle génétique de ces paramètres calibrés du modèle cinétique. Nous avons utilisé les cartes génétiques disponibles pour la population BC2. Les paramètres génotype-dépendants du modèle ont été traités comme des variables quantitatives et une recherche de QTL a été réalisée en considérant deux approches. La première est l'approche séquentielle à deux niveaux ('Independent analysis') et la seconde est l'approche intégrée ou 'joint-analysis' développée par Onogi (2020). Dans le premier cas, une recherche de QTL a été effectuée sur la base des données issues du chapitre précédent (Chapitre 3), i.e. les valeurs des 9 paramètres estimés pour tous les génotypes simultanément (stratégie dite 'population-based'). Nous avons comparé ces résultats avec ceux obtenus avec la seconde approche qui consiste à estimer les effets des marqueurs et les paramètres du modèle simultanément. Au préalable, une étude de simulation a été menée pour comparer l'efficacité de ces deux approches sur des données virtuelles. Ce chapitre, présenté sous la forme d'un article en préparation, décrit l'ensemble des résultats obtenus à partir de ces différentes analyses.

4.1 Impact of the estimation methods on the detection of QTLs controlling the parameters of a peach sugar model

Abstract

Systems biology aims at understanding complex interactions within biological systems by putting its pieces together, thanks to computational and mathematical analysis and modeling. In particular, the effect of genetic diversity and/or environmental modifications on the elaboration of complex traits can be explored this way. Integrating information related to genetic control into biological functioning models is a fraction of this interdisciplinary field and is a primary step towards virtual breeding. An improved ODE kinetic model of sugar metabolism proposed by Kanso et al. (2020) to predict the sugar concentration during peach fruit development was used to analyse the genetic control of this process. We proposed here two strategies for QTL detection. First, a 'population-based' two-step approach, called 'Independent analysis' (IA) : the 9 genetic parameters were estimated and then considered as quantitative traits to perform an independent QTL analysis. Second, a 'joint-analysis' approach (JA) :

the calibration of the 9 genetic parameters and the detection of the corresponding QTL were performed simultaneously. The two approaches were compared both on simulated and real data. Comparing the two approaches on the former data showed : i) their ability to infer the model parameters was good, ii) their capacity to identify the responsible genomic regions was comparable and satisfactory for major QTL, iii) the marker effects were, on average, underestimated and the phenotypic variance explained was always underestimated. On real data, i) the quality of fit of the model was very similar between the two approaches, ii) the parameter values were quite close and iii) the number of QTL detected was significantly increased using JA approach. QTL of parameters very often colocolized with metabolic QTL and candidate genes. Thus, the JA approach appeared slightly better to calibrate parameters and decipher their genetic control, especially since it is faster. These results are an important step towards developing of reliable gene-to-phenotype models and design ideotypes.

4.1.1 Introduction

Mathematical models have been considered as efficient tools to assess biological processes. In the context of plant breeding, predicting phenotypes with process-based models interacting with climate would help to test the performances of new genotypes under different conditions and thus orientate selection.

To develop such tools, it is necessary to add a level of control by the genome of the processes described in the model. In the absence of information on the gene network controlling these processes, one way to proceed is to consider some of the parameters of the model as genetic parameters (Boote et al. 2001 ; Tardieu 2003). The latter behave as quantitative traits and can be subjected to a quantitative trait locus (QTL) analysis to decipher their inheritance in a population (Yin et al. 2000). The resulting QTL-based parameters can be back injected in the process-based model and allow the prediction of phenotypes based on the allelic combination of the molecular markers flanking the detected QTL (Bertin et al. 2010). Promising results have been obtained using physiological components for traits such as plant development (Technow et al. 2015 ; Yin et al. 2000), early plant growth (Brunel et al. 2009), nitrogen uptake and root growth and architecture (Laperche et al. 2006) and peach fruit growth and sweetness(Quilot et al. 2005). Based on estimated QTL effects, phenotypes of untested plants can be predicted for any environments. For example, Reymond et al. (2003) and Bogard et al. (2014) predicted leaf elongation rates in maize and days to heading in wheat, respectively, with reasonable coefficient of determination. Such integrated genetic and process-based models are ideal tools to design ideotypes so as to develop virtual breeding. Promising procedures have been developed to identify parameter combinations that meet breeding objectives (Van Oijen et al. 2016). However, improvements are still needed to get closer to more realistic ideotypes (Quilot-Turion et al. 2016), especially to include more complex and detailed genetic control in process-based models. One direction is to integrate the information from the gene to the whole plant. Among possible approaches, kinetic modelling has been successfully applied to a number of

metabolic networks in plants (Curien et al. 2009 ; Nägele et al. 2014) and fleshy fruit (Beauvoit et al. 2014 ; Desnoues et al. 2014).

Besides the question of the availability of multi-scale process-based models, there are issues on the impact of model formulation and on the approaches used to estimate the parameters and detect QTL of these parameters. Barrasso et al. (2019) showed that different growth model equations, though closely related, led to the detection of different QTL of parameters. They concluded on the poor performance of the two algorithms (a classical nonlinear mixed-effects models algorithm and an evolutionary algorithm) tested for this large-scale global optimization problem showing strong correlations between parameters. Finally, they recommended to limit the number of genotype-dependent parameters to calibrate in order to reduce between parameter correlations. Indeed, these relations of correlations prevent from finding the true parameter value since in this case even a good fit cannot guarantee a unique parameter solution (Li et al. 2013).

Concerning the accurate estimation of the parameters and the genetic mapping, three main approaches are used. Let first mention the traditional straightforward way, known as the two-step approach, in which the 2 steps are performed separately. In this simple approach, the genotype-dependent parameters are first estimated and then, in a second independent step, the values obtained for each individual of the population are used for genetic mapping. The limits of this so-called independent approach are of two types : i) the uncertainty of the parameter values from the optimization step is not taken into account in the mapping step, ii) the genetic relatedness between the individuals of the population is not considered during the optimization step because parameters are often calibrated independently for each individual. A second approach is the functional mapping (Ma et al. 2002) defined by Wei et al. (2018) as a top-down approach because it integrates mathematical model within the genetic mapping. This one-step powerful method uses a maximum-likelihood (ML) based on a finite mixture model and implemented with the Expectation-Maximisation (EM) algorithm. It was successfully applied on different mathematical models involving delay differential equations (DDE) of circadian rhythm (Fu et al. 2011b), stochastic differential equations (SDE) in pharmacogenomics (Wang et al. 2013), ordinary differential equations (ODE) of biological rhythms (Fu et al. 2011a). Despite it can outperform the bottom-up approach in terms of power to detect significant QTL (Wu et al. 2006) and curve estimation precision (Wei et al. 2018), it also displays disadvantages. First, the combination of a complex process-based model together with the QTL mapping theory may not be feasible. Second, the expected values at all time points and for all elements in the covariance matrix need to be estimated, resulting in substantial computational difficulties. The third approach takes up interesting characteristics from functional mapping but combines them with higher feasibility. This one-step integrated approach has been proposed by Sillanpää et al. (2012) to map functional QTL for quadratic and linear curve growths using Bayesian modelling. Subsequently, this approach called JA (Joint Analysis) was also proposed for crop growth models by Technow et al. (2015) and tested by Onogi (2020) et Onogi et al. (2016a) using real

data and different mathematical models, such as daily oscillation of transcriptions of GIGANTEA in rice (Nagano et al. 2012), glucose absorption in humans (Dalla Man et al. 2006) and flowering time of rice (Nakagawa et al. 2005). In these studies, JA (Joint Analysis) proved to be more accurate in prediction than IA (Independent Approach). Indeed, JA offers the advantage of taking into account the whole population of individuals and their relations (shared genetic background between individuals), since it connects the individuals by a covariance matrix.

In this context of developing genetic and process-based models, a kinetic model of sugar metabolism has been designed by Desnoues et al. (2018) and the equation were parameterized from biochemical data (Desnoues et al. 2014; Desnoues et al. 2018). This model was then reduced to enable its use for a large panel of individuals (Kanso et al. 2020) : the number of parameters was scaled down from 14 to 9 parameters and different methods of parameter estimation were tested (Chapter 3 of this thesis) to estimate the 9 parameters for a progeny of 106 individuals derived from an interspecific peach cross. Whatever the methods of parameter estimation used, the reduced model correctly accounted for the variability of accumulation of different sugars during peach fruit development observed between the individuals. Although the fittings were comparable, the different methods of parameters estimation resulted in different sets of parameter values. This illustrated the difficulty to infer the 'true' genetic value of the parameters.

Following this work, the objective of the present study was to explore the genetic control of the 9 parameters of this reduced model, based on a progeny of 106 genotypes. As a one-step approach (Joint Approach, JA) has never been applied to a complex kinetic model with real data, we proposed here to compare the classical two-steps (Independent Analysis, IA) approach and JA. Within IA, QTL detection was conducted on the set of parameter values derived from a Population-Based model using the stochastic approximation expectation maximisation algorithm (Chapter 3 of this thesis). The method of Extended Bayesian Lasso (EBL) (Mutshinda et al. 2010) was used for the genetic analyses. This comparison was performed on simulated and real data. The final results obtained from the two approaches were compared both in terms of i) accuracy of the simulation-observation fit for each sugar and each individual and ii) QTL detected for the parameters compared to the QTL controlling sugars and enzyme activities (Desnoues et al. 2016).

4.1.2 The kinetic model of sugar metabolism in peach fruit

The ODE kinetic model developed by Desnoues et al. (2018) describes the accumulation of four different sugars (sucrose, glucose, fructose and sorbitol) in peach fruit during its development. To be applied to a large panel of individuals, this model counted too many parameters and the calibration time was too long. Thus, different reduction methods have been used and the resulting reduced model, with only 9 genotype-specific parameters to be calibrated, was successfully applied to 30 new individuals (Kanso et al. 2020). This reduced model is used in this study.

The carbon inflow was calculated from measured fruit masses provided as input for each individual and enzyme kinetics measured by Desnoues et al. (2014) served to set genotype-independent parameters. From a mathematical point of view, the model is constructed as a set of parametric ordinary differential equations with 7 reactions represented by linear flows. For the k^{th} genotype, the model can be written as follows :

$$\frac{dx}{dt} = g(x(t), \phi^{(k)}), \tag{4.1}$$

$$x(t_0) = x_0^{(k)}, \tag{4.2}$$

where t is the independent time variable in days after bloom (DAB); $x \in \mathbb{R}^9$ is the concentration vector of metabolites in the corresponding intra-cellular compartment, $x_0^{(k)} \in \mathbb{R}^9$ in Eq.(4.2) is the vector of the corresponding initial values for the genotype k and $\phi^{(k)} \in \mathbb{R}^9$ is the vector of the 9 genotype-dependant parameters to be estimated. $g(x(t), \phi^{(k)})$ of Eq.(4.1) describes the change in compounds concentrations. The full set of equations of the model is presented in Kanso et al. (2020). Here is the description of the 9 genotype-dependent parameters of the kinetic model 4.1 :

TABLE 4.1 – Description of the 9 unknown genotype-dependant parameters of the sugar model

Parameter	Description	Range	Unit
LHx	sucrose proportion hydrolyzed in the apoplasm	]0, 1]	
RAI	coefficient of the transfer function between sucrose and (glucose +fructose) under action of acid invertase (AI) enzyme	]0, 0.1	day^{-1}
KFk	fructokinase (FK) affinity	]0, 100]	$mggFW^{-1}$
KHk	hexokinase (HK) affinity	]0, 500]	$mggFW^{-1}$
OCp	coefficient of the transfer function between hexose phosphates and other compounds	]250, 1500]	day^{-1}
RSS	coefficient of the transfer function between hexose phosphates and sucrose	]0, 1000]	day^{-1}
TAS	coefficicent of sucrose transport (active import) from cytosol to vacuole	]0, 400]	$mggFW^{-1}day^{-1}$
TPF	coefficient of fructose passive transport between cytosol and vacuole and in the opposite direction	]0, 150]	$mggFW^{-1}day^{-1}$
RSO	coefficient of the transfer function between sorbitol and glucose under action of sorbitol oxydase (SO) enzyme	]0, 10]	day^{-1}

4.1.3 Notations

The following mathematical notations are used :

- G : number of genotypes in the population i.e. 106.
- n_{SG} : number of sugars i.e. 4.
- $N_j^{(k)}$: number of experimental observations for the genotype k and for the j-th sugar
- $\mathcal{M}_{ij}(\phi^{(k)})$: i-th *prediction* of sugar j concentrations of genotype k at time $i \in \mathcal{T}_M^{(k)}$, obtained with model $\mathcal{M}$ and parameters $\phi^{(k)}$.
- $\widetilde{\mathcal{M}_{ij}}(\phi^{(k)})$: log transformation of $\mathcal{M}_{ij}(\phi^{(k)})$
- $y_{ij}^{(k)}$: i-th *observation* of j-th sugar concentrations of genotype k

- $\tilde{y}_{ij}^{(k)}$: log transformation of $y_{ij}^{(k)}$
- $\phi^{(k)}$: the vector of 9 parameters (log-value) to be estimated for each genotype k
- B : the marker effects
- $\varkappa_m^{(k)}$: the allelic value at the m^{th} marker coded 1 for QQ and 2 for Qq for genotype k
- M : the number of markers

4.1.4 Real data analysis

The progeny of 106 inter-specific genotypes was obtained by two subsequent back-crosses between *Prunus davidiana* (Carr.) P1908 and two peach cultivars, *Prunus persica* (L.) Batsch 'Summergrand' and then 'Zephyr' (Quilot et al. 2004b). The possible genotypes at any given locus in the progeny are presented in Table 4.2. As described in (Desnoues et al. 2014), one tree per genotype was planted in a randomized design in the orchard of the INRAE Research Centre of Avignon (southern France). Different phenotypic traits were measured at different time points during fruit development, for all genotypes, such as concentration of sucrose, glucose, fructose, sorbitol, and hexoses phosphates, the fruit flesh fresh weight and dry matter content (Desnoues et al. 2014).

TABLE 4.2 – Possible genotypes at a single locus in SD, BC1 and BC2 progenies (from Quilot et al. (2004b)). Cross of P1908 (D) with *P. persica* 'Summergrand' (S) produced an F1 progeny (SD). Then, one F1 hybrid was back-crossed to S to produce a BC1 progeny. Finally, a pool of pollen from BC1 individuals was used to pollinate *P. persica* 'Zephyr' (Z) to derive the progeny studied here (BC2).

$D \times S$			$SD40 \times S$			$BC1 \times Z$			
SD	D_1	D_2	$BC1$	D_1	S_1	$BC2$	D_1	S_1	S_2
S_1	D_1S_1	D_2S_1	S_1	D_1S_1	S_1S_1	Z_1	Z_1D_1	Z_1S_1	Z_1S_2
S_2	D_1S_2	D_2S_2	S_2	D_1S_2	S_2S_1	Z_2	Z_2D_1	Z_2S_1	Z_2S_2
$SD40$ genotype is coded D_1S_1 at one locus			Possible gametes from $BC1$ progeny $D_1(1/4)$, $S_1(1/2)$, $S_2(1/4)$			$D(1/8)$, $S(3/8)$, $Z_1(1/4)$, $Z_2(1/4)$			

Two genetic maps, developed by Desnoues et al. (2016), were used for the QTL detection. The first one displays the polymorphism between the D and the S genomes (DvsS map). It is composed of 340 informative molecular markers. The second one monitors the polymorphism between the 2 alleles of 'Zephyr' (Z_1Z_2) and is composed of 117 markers (Zephyr map). On both maps, the markers were distributed across the 8 autosomal chromosomes of the peach. At any marker, there are two possible genotypes : QQ or Qq. In the following, the two maps have been concatenated one after the other, to take into account all the polymorphism available in the progeny.

Before the analysis, each missing marker genotype was completed (once) by random draws from Bernoulli $\hat{p}_{missing}$, where $\hat{p}_{missing}$ is the conditional expectation

estimated from flanking markers with known genotypes (Haley et al. 1992).

4.1.5 Simulation study

1. *Genetic map*

   The design of the real peach marker data described in Section 4.1.4 was used as a basis for the simulation study. Hence, the map was composed from 340 markers distributed on 8 chromosomes ($DvsSmap$).

2. *Major and Minor QTL selection*

   For each model parameter (total 9 parameters), five major quantitative trait loci (QTL) were selected along the different chromosomes. In addition, 3 minor QTL were selected per chromosome (total 24 minor QTL). The QTL effects of the major QTL were determined as

   $$\sqrt{\frac{1}{2a_f^m(1-a_f^m)}}$$

   where a_f^m is the allele frequency at the marker m linked to the selected QTL obtained by

   $$a_f^m = \frac{\sum_{k=1}^{G} \varkappa^{(k)}(m)}{2G}$$

   where $\varkappa^{(k)}(m)$ is the allelic value at the marker m on the selected QTL and for genotype k.
   Similarly, the QTL effects of the minor QTL were determined as

   $$\sqrt{\frac{0.02}{2a_f^m(1-a_f^m)}}$$

   assuming that each major QTL explains the genetic variance 50 times greater than the minor QTL does. The signs of the QTL are determined randomly.

3. *Genotypic values of the parameters*

   The genotypic values of the parameters were calculated from the QTL effects and allelic values ($DvsSmap$). Assuming that the heritability (h^2) of the model parameters was 0.9, random noises were added to the genotypic values. The model used for this aim was

$$\breve{\phi}^{(k)} = \varkappa^{(k)} B + e_k, \quad e_k \sim \mathcal{N}(0, \vartheta^2 \frac{1-h^2}{h^2}) \tag{4.3}$$

where $\breve{\phi}^{(k)}$ is the matrix of genotypic values of the parameters, $\varkappa^{(k)}$ is the matrix of allelic values at the markers linked to the QTL controlling the parameters, B is the vector of QTL effects and ϑ^2 is the variance of the genotypic values.

The model parameters were then scaled such that the values were settled within the corresponding observed ranges (see Table 4.1). Using this basis, 100 virtual populations of 106 genotypes were generated.

4. *Datasets of four sugars*

For each virtual population, the individual genotypes were simulated from the corresponding generated parameter values and a common variance of the measurement errors was set to $\sigma^2 = 0.05$. For this aim, the following model can be used

$$\widetilde{y}_{ij}^{(k)} = \widetilde{\mathcal{M}}_{ij}(\breve{\phi}^{(k)}) + \varepsilon_{ij}^{(k)}, \quad \varepsilon_{ij}^{(k)} \sim \mathcal{N}(0, \sigma^2) \tag{4.4}$$

where $\widetilde{\mathcal{M}}_{ij}(\breve{\phi}^{(k)}) = log(\mathcal{M}_{ij}(\breve{\phi}^{(k)}))$ is i-th- log-prediction of j-th sugar concentrations of genotype k on the corresponding generated parameter values $\breve{\phi}^{(k)}$. For each population, the input curves of 106 real individual genotypes were used as a input of the model.

4.1.6 Metabolite experimental data

The concentration of the metabolites, namely sucrose, glucose, fructose, sorbitol, and hexoses phosphates, the flesh fresh mass and dry matter content were measured at different time points during fruit development, for all genotypes. In addition, the temporal evolutions of enzymatic capacities (maximal activity) of the twelve enzymes involved in sugar metabolism were measured over the whole progeny (Desnoues et al. 2014). These data served for the calibration of the reduced model (Chapter 3 of this thesis) to produce a set of parameter values, i.e a matrix of 9 parameters x 106 genotypes.

4.1.7 QTL analyses

Two approaches were used to study the genetic control of sugar metabolism. Following the first one, called Independent Analysis **'IA'**, the 9 genetic parameters were first estimated and then considered as quantitative traits to perform an independent QTL analysis. Using the second approach called Joint Analysis **'JA'**, the mathematical model was connected to the genetic maps. In this case, the parameters of the model and markers effects were estimated simultaneously.

4.1.7.1 Regression method

In order to detect the statistical association between markers and phenotypes and estimate the marker effects, the Extended Bayesian Lasso (EBL) proposed by Mutshinda et al. (2010) was used as a Whole-Genome Regression (WGR). For this aim, $\phi = (\phi^{(k)}, 1 \leq k \leq G)^T$ is defined as the set of phenotypic values (parameter estimations in our case), $\varkappa^{(k)} = (\varkappa_m^{(k)}, 1 \leq m \leq M)$ stands for the allele value coded 1 for QQ and 2 for Qq for genotype k at marker m and $B = (B_0, \ldots, B_M)^T$ is the vector of QTL effects. The model can be written for the k^{th} individual as

$$\phi^{(k)} = B_0 + \sum_{m=1}^{M} \varkappa_m^{(k)} B_m + e_k = \varkappa^{(k)} B + e_k \tag{4.5}$$

with B_0 is the intercept, B_m is the effect on the phenotypic value of the QTL at marker m and e_k is the residual error for individual k, following a normal distribution $\mathcal{N}(0, \sigma_0^2)$. In addition, B_m is assumed to follow a normal distribution with mean 0 and constant variance σ_m^2. Details of EBL are summarized in the appendix 4.1.12.1.

4.1.7.2 BLOD score

The approximate Bayesian LOD (BLOD) score proposed by Yi et al. (2008) was used as the test statistic. For marker m, the Bayesian LOD score was computed as follows :

$$BLOD_m = 2log_{10}\left(\frac{\prod_{k=1}^{G} \mathcal{N}(\phi^{(k)} | \varkappa^{(k)} B, \sigma_0^2)}{\prod_{k=1}^{G} \mathcal{N}(\phi^{(k)} | \varkappa^{(k)} B - \varkappa_m^{(k)} B_m, \varrho_m^2)}\right) \tag{4.6}$$

with

$$\varrho_m^2 = \frac{1}{G-2} \sum_{k=1}^{G} (\phi^{(k)} - \varkappa^{(k)} B + \varkappa_m^{(k)} B_m)^2$$

where $\mathcal{N}(\phi^{(k)} | \varkappa^{(k)} B, \sigma_0^2)$ denotes that $\phi^{(k)}$ follow an independent normal distribution with mean $\varkappa^{(k)} B$ and variance σ_0^2. The BLOD score is similar to the LOD score in traditional QTL mapping approach (Lander et al. 1989), but BLOD considers the contribution of a given locus to the LOD after adjusting for the effects of all other loci (Yi et al. 2008). A larger BLOD means that the marker has a higher probability of being associated with the phenotypes.

Permutation tests were performed to confirm a significance threshold for the BLOD score. By re-ordering the phenotypes of all individuals, each individual was randomly reassigned a new phenotypic value and new BLOD scores were computed. The significance threshold at α level for a BLOD score was obtained as follows (Churchill et al. 1994). A permutation test with 1000 repetitions was performed, repeatedly to find the threshold BLOD scores for α =0.05.

4.1.7.3 PVE

To assess the contribution of the estimate of the genetic effects B_m associated with each marker, we can compute the proportion of the phenotypic variance explained (PVE) by m-th genetic effect B_m as :

$$PVE_m = \hat{B}_m^2 \frac{V_m}{V_\phi} \tag{4.7}$$

where $\hat{B}$ is the estimated value of parameter B, V_m is the variance of the marker m and V_ϕ is the variance of the considered traits (Huang et al. 2014).

4.1.7.4 The independent analysis 'IA' approach

The 'IA' approach aimed, in a first step, to fit the model to the data on the whole progeny of 106 genotypes. Then, in a the second step, the estimated parameters were considered as traits and were used for a genetic analysis to estimate the whole-genome marker effects. The estimation of the values of the genetic parameters could be performed using several methods. A former study dedicated to the estimation of the parameters of the reduced model are presented in Chapter 3 of this thesis. In this chapter, the model was calibrated simultaneously on the whole progeny of genotypes using a Population-Based Model (Baey et al. 2016). The dataset of estimated values of model parameters reported in Chapter 3 was used. For the second step, estimated values of the model parameters were regressed on genome-wide markers using EBL (see Section 4.1.7.1), and the marker effects were estimated.

4.1.7.5 The joint analysis 'JA' approach

This approach enabled to estimate simultaneously the values of the parameters of the model and the effects of the markers, in order to fit our observations data and search the links between phenotype and genotype.

First, we denote by $y_{ij}^{(k)} (1 \leq k \leq G,\ 1 \leq j \leq n_{SG},\ 1 \leq i \leq N_j^{(k)})$ the i^{th} observation of j^{th} sugars concentrations of genotype k :

$$\widetilde{y}_{ij}^{(k)} = \widetilde{\mathscr{M}_{ij}}(\phi^{(k)}) + \varepsilon_{ij}^{(k)}, \quad \varepsilon_{ij}^{(k)} \sim \mathscr{N}(0, \sigma_k^2), \tag{4.8}$$

where $\mathscr{M}$ is the mathematical function relying the considered model to the data, $\phi^{(k)}$ are the genetic parameters to be estimated for each genotype k and $\epsilon_{ij}^{(k)}$ is the error term assumed to be normally distributed. Then, ϕ was regressed on genome-wide markers using EBL regression method following the procedure described in Section 4.1.7.1.

With the JA approach proposed by Onogi et al. (2016a) and Onogi (2020), model parameters are inferred using a variational Bayesian approach in which means and

variances of the model parameters are obtained using Markov chain Monte Carlo (MCMC) sampling and are used to update EBL parameters. These two algorithms were iteratively repeated until convergence to maximize the lower boundary of the marginal likelihood of the system. The details of this method are described in the appendix (4.1.12.2).

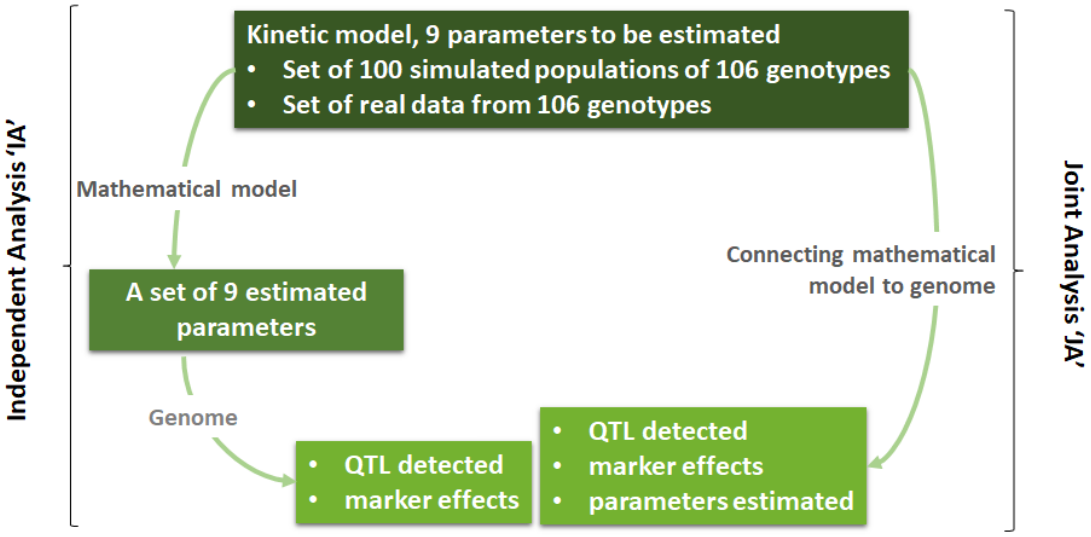


FIGURE 4.1 – Graphical representation of the simulated and experimental studies. Two approaches were used. The first is called 'IA' : genetic parameters were first estimated for all genotypes simultaneously and then used for genetic analyses. The second approach is called 'JA' : the mathematical model and the genome were connected together to estimate simultaneously the genetic model parameters and the marker effects.

4.1.8 Methods to assess the goodness of the calibration process

4.1.8.1 Expected Error (%)

The reliability of *individual* parameters estimated was compared by computing the relative distance between the true values $\check{\phi}_r^{(k)}$ and estimated value $\hat{\phi}_r^{(k)}$, for the individual parameter r and the genotype k. For each simulated population, the Expected Error E_r (%) of the parameter r was defined as :

$$\mathrm{E}_r = \frac{1}{G}\sum_{k=1}^{n}\left|\frac{\check{\phi}_r^{(k)} - \hat{\phi}_r^{(k)}}{\check{\phi}_r^{(k)}}\right| \times 100 \tag{4.9}$$

where $G = 106$ is the number of virtual genotypes included in the population. Overall our simulation study yielded 100 values of E_r, for each of the 9 parameters.

4.1.8.2 Modelling efficiency

Modelling efficiency (EF) proposed by Mayer et al. (1993), is a goodness of fit indicator that can be computed as a measure of the accuracy of phenotypic predictions.

It is defined as follows :

$$EF_j = 1 - \frac{\sum_{k=1}^{G}\sum_{i=1}^{N_j^{(k)}}(y_{ij}^{(k)} - \mathscr{M}_{ij}(\phi^{(k)}))^2}{\sum_{k=1}^{G}\sum_{i=1}^{N_j^{(k)}}(y_{ij}^{(k)} - \overline{y}_j^{(k)})^2}, \quad EF_j \in]-\infty\ 1], \tag{4.10}$$

where $\overline{y}_j^{(k)}$ the mean of observations of sugar j for genotype k. $EF = 1$ means a perfect equality between the predictions and the observations, and if $EF < 0$ means that the model predictions are worse than the mean of the observations.

4.1.9 Results

4.1.9.1 Simulation study

The use of the two approaches on simulated data gave comparable results in terms of detection of the genomic regions responsible for the variations of the parameter values within the population. Five major QTL were simulated for each of the nine parameters and 100 replications datasets built from these and analysed using the two approaches. Both approaches gave accurate results. Indeed, the approaches IA and JA were able to detect respectively 3388 and 3395 significant QTL exactly at the good marker, over the 4500 major QTL (see Table 4.3), which represents 75% of success. In addition, they respectively detected 838 and 833 QTL at the closest neighbouring markers. So the two approaches almost detected 94% of the genomic regions simulated. Their efficiency depended on the parameters, going from 84 to 100% of success. In contrast, very few minor QTL were detected. Counting significant QTL both at the correct marker and closest neighbouring markers, the percentage of success hardly exceeded 5% for both approaches. Finally, no major false QTL was detected by either of the two approaches.

TABLE 4.3 – Performance of the two approaches, IA and JA, to significantly detect the 4500 major and 21600 minor simulated QTL. In the simulation study, 500 and 2400 major and minor QTL respectively were simulated for each of the nine parameters. The number of QTL detected on the correct marker (QTL_C), or on the closest neighbouring markers (QTL_V) (3 markers on either side) is counted. The remaining counts over 500 and 2400 major and minor QTL represent the non detected QTL (QTL_N). QTL_I : incorrect QTL detected.

Count of	Trait	Major QTL		Minor QTL	
		IA	JA	IA	JA
QTL_C	LHx	487	489	12	12
	RAI	409	409	14	11
	KFk	315	325	15	16
	KHk	254	251	36	35
	OCp	325	320	57	49
	RSS	380	384	6	11
	TAS	434	432	44	43
	TPF	359	355	58	65
	RSO	425	430	6	4
Total		3388	3395	248	246
QTL_V	LHx	13	11	69	66
	RAI	44	48	77	81
	KFk	167	164	108	114
	KHk	170	173	140	145
	OCp	135	141	157	169
	RSS	93	86	71	68
	TAS	52	51	66	70
	TPF	105	103	176	190
	RSO	59	56	46	41
Total		838	833	910	944
QTL_N	LHx	0	0	2319	2322
	RAI	47	43	2309	2308
	KFk	18	11	2277	2270
	KHk	76	76	2224	2220
	OCp	40	39	2186	2182
	RSS	27	30	2323	2321
	TAS	14	17	2290	2287
	TPF	36	42	2166	2145
	RSO	16	14	2348	2355
Total		274	245	20442	20410
QTL_I	LHx	0	0	23	23
	RAI	41	36	33	29
	KFk	1	4	21	16
	KHk	1	1	27	25
	OCp	7	9	11	8
	RSS	3	4	12	8
	TAS	15	11	19	21
	TPF	32	31	45	43
	RSO	1	2	3	7
Total		101	98	194	181

Concerning the QTL detected (both exactly at the correct marker and at the neighbouring markers, ie three markers on either side), with a BLOD score over the threshold, their characteristics turned out to be very similar between the two approaches. For both approaches, the BLOD scores were in average lower than the true ones and never higher than them, whatever the replications (see Figure 4.2). In addition, it happened for the two approaches to display low values of BLOD scores compared to the expected ones (factor 40) (see Table 4.4). In a few cases, one of the two approaches performed less well than the other one in detecting a major QTL (QTL1 of RAI, QTL2 and 3 of KFk, QTL1 of OCp, QTL1 of RSO had BLOD scores <10 for one of the two approaches). Finally, the two approaches also detected some false QTL. They represented hardly more than 2% of the total number of major QTL detected.

FIGURE 4.2 – Profiles of BLOD scores over 100 replications (black points) along the genome for nine parameters and for both strategies IA and JA. True BLOD scores for simulated QTL are shown by red asterisks. Dashed lines indicate the 5% genome-wide threshold value for claiming the existence of a QTL at the corresponding marker

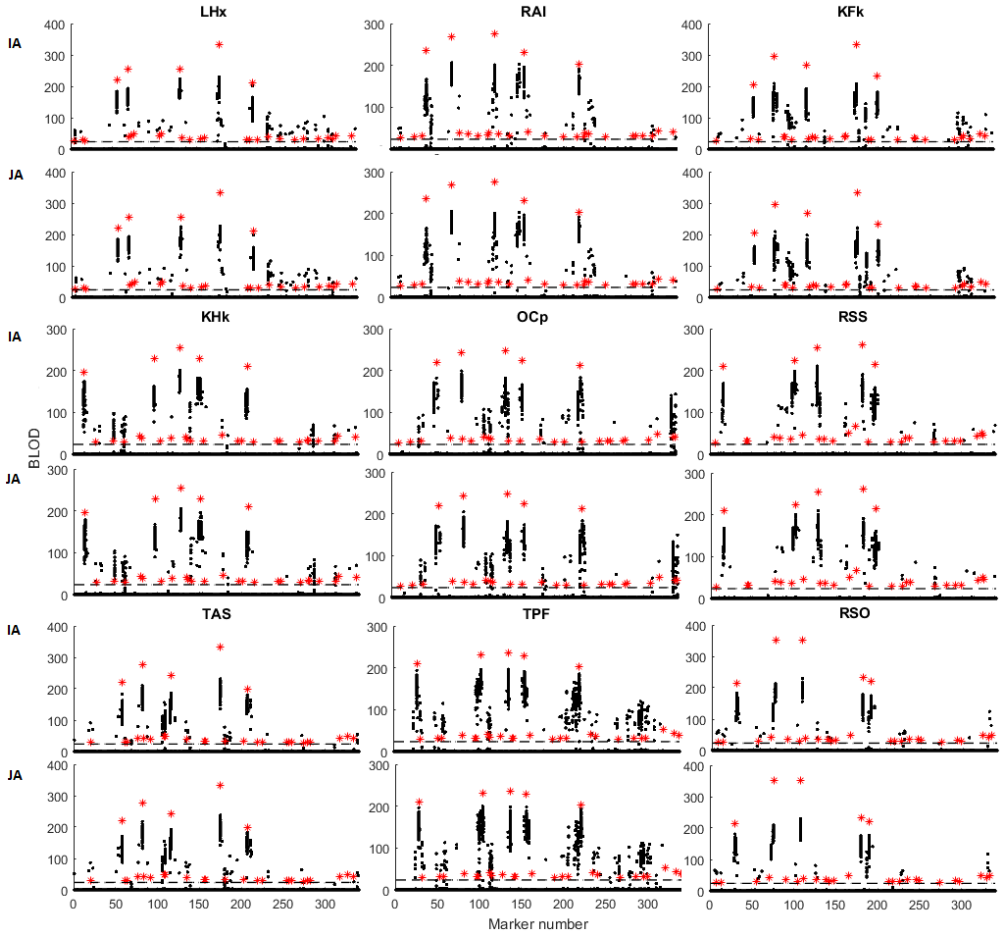
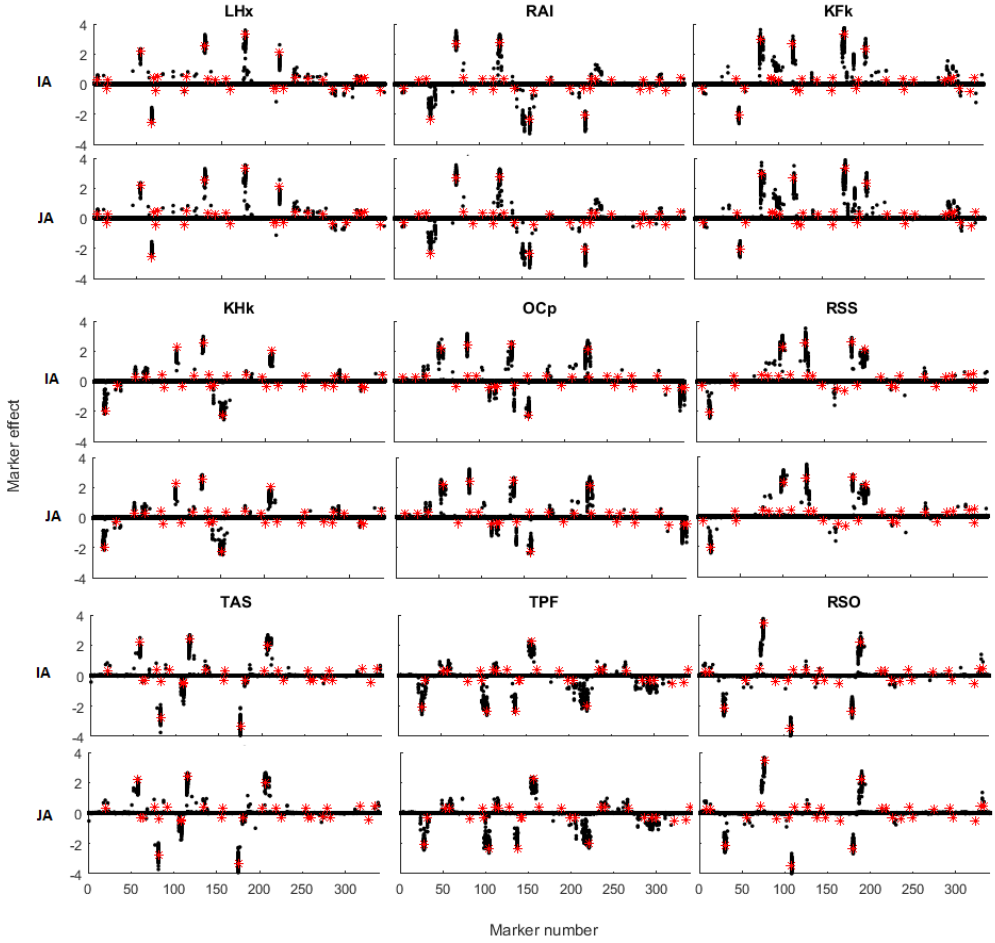


TABLE 4.4 – Characteristics of the 5 major QTL detected for the nine simulated parameters, with the two approaches, IA and JA over 100 replicated data. Average data (minimum / maximum) for significant QTL detected on the correct marker, or on the closest neighbouring markers : BLOD score of the QTL, Marker effects and Phenotypic Variance Explained (%) for the two approaches. True simulated values are also given

Trait	QTL	BLOD			Effects			PVE		
		IA	JA	True	IA	JA	True	IA	JA	True
LHx	1	156 (116/183)	156 (115/184)	231	1.9 (1.3/2.3)	2 (1.3/2.3)	2.2	87.8 (79.2/91.7)	87.8 (78.9/91.8)	95.6
	2	164 (126/193)	164 (125/193)	238	-2.1 (-2.5/-1.5)	-2.1 (-2.5/-1.5)	-2.5	89 (81.9/92.7)	89 (81.6/92.7)	96
	3	187 (162/223)	186 (126/224)	265	2.7 (2.1/3.2)	2.7 (1.3/3.3)	2.5	91.9 (88.9/95.1)	91.8 (81.8/95.2)	97.4
	4	193 (43/229)	196 (118/226)	275	2.8 (0.4/3.5)	2.9 (1.1/3.5)	3.3	92.1 (44.8/95.5)	92.9 (79.8/95.3)	97.6
	5	130 (88/201)	130 (88/199)	207	1.4 (0.9/2.6)	1.5 (0.9/2.5)	2.1	82.4 (69.8/93.4)	82.4 (69.8/93.2)	93.9
RAI	1	120 (8/166)	121 (82/164)	237	-1.4 (-2.3/-0.1)	-1.4 (-2.3/-0.9)	-2.37	79.1 (10.6/89.5)	80 (67.1/89.2)	95.9
	2	180 (152/206)	179 (152/205)	283	2.8 (2.2/3.5)	2.8 (2.2/3.5)	2.68	91.1 (87.3/93.8)	91.1 (87.4/93.8)	97.8
	3	176 (52/201)	174 (10/200)	290	2.7 (0.6/3.3)	2.7 (0.2/3.2)	2.76	90.6 (50.9/93.4)	87.9 (13.2/93.3)	98
	4	168 (77/195)	169 (98/196)	276	-2.5 (-3.2/-0.8)	-2.5 (-3.2/-1)	-2.32	89.5 (64.8/92.9)	89.7 (73.4/93)	97.6
	5	158 (131/191)	158 (132/191)	265	-2.3 (-3.1/-1.8)	-2.4 (-3.1/-1.7)	-2.03	88.1 (83.2/92.5)	88.2 (83.3/92.5)	97.2
KFk	1	136 (102/163)	136 (107/162)	260	-2.1 (-2.6/-1.5)	-2.1 (-2.57/-1.5)	-2.1	84 (75/89.1)	84.1 (76.7/88.8)	97
	2	150 (39/209)	147 (6/209)	298	2.3 (0.5/3.6)	2.2 (0.15/3.7)	2.9	86.1 (41.1/94.1)	84.4 (8.2/94.1)	98.2
	3	144 (6/193)	147 (100/198)	282	2.1 (0.1/3.1)	2.2 (1.31/3.49)	2.6	84.8 (8.6/92.6)	85.9 (74.2/93.1)	97.8
	4	176 (98/208)	175 (94/221)	312	2.9 (1.1/3.7)	2.9 (0.98/3.8)	3.3	90.5 (73.5/94)	90.3 (72.1/95)	98.5
	5	148 (101/181)	145 (11/180)	280	2.2 (1.3/ 3)	2.2 (0.23/3.02)	2.3	86.1 (74.7/91.4)	84.9 (14.1/91.3)	97.7
KHk	1	132 (52/173)	133 (73/178)	215	-1.6 (-2.1/-0.5)	-1.6 (-2.1/-0.8)	-1.9	82.4 (50.9/90.4)	82.9 (62.9/91.1)	94.5
	2	135 (77/162)	135 (89/167)	212	1.5 (0.7/2)	1.5 (0.8/1.9)	2.2	83.7 (64.8/88.8)	83.7 (70.3/89.6)	94.3
	3	176 (147/200)	176 (152/204)	249	2.4 (1.8/2.9)	2.4 (1.8/2.8)	2.5	90.6 (86.3/93.4)	90.7 (87.3/93.7)	96.5
	4	157 (129/181)	156 (66/196)	260	-1.9 (-2.5/-1.4)	-1.9 (-2.4/-0.7)	-2.2	87.9 (82.6//91.4)	87.4 (59.2/92.9)	97.2
	5	125 (85/155)	125 (75/149)	261	1.4 (0.9/1.8)	1.4 (0.7/1.8)	2.1	81.3 (68.5/87.8)	81.3 (64/86.7)	97.3
OCp	1	154 (132/181)	139 (111/154)	232	2.1 (1.6/2.7)	1.9 (1.5/2.1)	2.1	87 (83.4/91.4)	84.6 (77.9/87.6)	95.6
	2	164 (126/198)	165 (121/204)	267	2.4 (1.6/3.1)	2.5 (1.6/3.2)	2.4	89 (81.9/93.2)	89 (80.7/93.7)	97.3
	3	145 (93/181)	144 (106/180)	258	1.9 (1/2.6)	1.9 (1.3/2.6)	2.4	85.6 (71.7/91.4)	85.5 (76.4/91.3)	96.9
	4	137 (83/166)	136 (90/173)	247	-1.8 (-2.4/-1.1)	-1.8 (-2.3/-1.1)	-2.2	83.8 (67.6/89.4)	83.1 (70.6/90.4)	96.5
	5	147 (63/182)	153 (66/183)	253	2.1 (0.7/2.7)	2.2 (0.8/2.7)	2.1	84 (58.9/91.6)	86.7 (59.1/91.6)	96.7
RSS	1	135 (77/169)	136 (79/167)	249	-1.8 (-2.4/-0.9)	-1.8 (-2.4/-0.9)	-2.1	83.6 (65/89.8)	83.7 (66.1/89.6)	96.6
	2	171 (62/198)	174 (145/200)	264	2.2 (0.8/2.9)	2.6 (2.1/3.1)	2.2	89.6 (56.8/93.1)	90.4 (66/93.4)	97.2
	3	170 (117/210)	171 (117/210)	278	2.6 (1.3/3.5)	2.6 (1.5/3.4)	2.5	89.7 (79.5/94.2)	89.9 (79.7/94.1)	97.7
	4	157 (77/190)	158 (59/191)	270	2.2 (0.8/2.9)	2.2 (0.6/2.7)	2.6	87.9 (65.1/92.4)	87.9 (55.2/92.5)	97.4
	5	125 (86/159)	125 (79/159)	235	1.6 (0.9/2.1)	1.6 (0.8/2.2)	2.1	81.1 (68.8/88.3)	81 (66/88.5)	95.8
TAS	1	131 (86/183)	131 (87/169)	266	1.6 (1.2/2.4)	1.6 (1.1/2.2)	2.1	82.6 (69.1/91.6)	82.6 (69.3/89.9)	97.2
	2	177 (132/209)	177 (131/217)	308	-2.7 (-3.7/-1.9)	-2.7 (-3.7/-1.9)	-2.7	90.7 (83.4/94.1)	90.8 (83.2/94.7)	98.4
	3	149 (90/185)	148 (129/192)	284	2.1 (1.1/2.6)	2 (1.6/2.6)	2.4	86.2 (70.6/91.9)	85.7 (82.8/92.5)	97.8
	4	195 (136/231)	195 (142/237)	321	-3.2 (-4.4/-2.1)	-3.2 (-4.2/-2.2)	-3.3	92.7 (84.2/95.6)	92.7 (85.4/96)	98.7
	5	149 (119/180)	148 (114/183)	262	2.1 (1.3/2.6)	2.1 (1.2/2.6)	1.9	86.5 (80.2/91.2)	86.1 (78.8/91.7)	97.1
TPF	1	153 (104/193)	155 (101/195)	257	-1.8 (-2.5/-0.9)	-1.9 (-2.4/-1)	-2.1	86.7 (77.8/92.7)	87.4 (74.5/92.9)	96.9
	2	166 (75/196)	161 (71/199)	255	-2.1 (-2.6/-0.7)	-1.9 (-2.6/-0.6)	-2.3	88.9 (63.9/92.9)	87.6 (62.6/93.3)	96.8
	3	151 (101/196)	149 (92/197)	260	-1.7 (-2.4/-1.2)	-1.7 (-2.3/-1)	-2.3	86.4 (74.5/93)	86.2 (71.4/93.1)	97.1
	4	155 (107/191)	156 (114/197)	255	1.8 (1.2/2.3)	1.8 (1.3/2.3)	2.2	87.3 (76.8/92.5)	87.5 (78.6/93.1)	96.8
	5	145 (103/185)	151 (104/193)	250	-1.7 (-2.3/-1)	-1.7 (-2.3/-1.1)	-2.1	84.6 (72.4/91.8)	86.7 (75.6/92.7)	96.6
RSO	1	141 (95/182)	139 (95/180)	244	-2 (-2.6/-1.2)	-2 (-2.5/-1.1)	-2.1	84.7 (72.4/91.5)	83.8 (69.9/91.3)	96.35
	2	190 (152/212)	189 (152/209)	288	3.1 (2.0/3.7)	3.1 (1.9/3.6)	3.5	92.3 (87.4/94.4)	92.1 (87.3/94.1)	97.99
	3	195 (155/228)	194 (60/229)	288	-3.4 (-4.2/-2.7)	-3.4 (-4.2/-2.6)	-3.5	92.8 (87.9/95.4)	92.7 (88.6/95.5)	97.99
	4	144 (94/178)	144 (92/175)	247	-2.1 (-2.8/-1.4)	-2.1 (-2.6/-1.4)	-2.3	85.3 (72.3/91.1)	85.5 (71.5/90.7)	96.5
	5	145 (105/174)	142 (76/176)	242	2.1 (1.3/2.8)	2.1 (1.1/2.7)	2.1	85.6 (76/90.6)	85 (64.4/90.8)	96.26

Regarding the marker effects, they were also, on average, underestimated by the two approaches, for most of the parameters and QTL, compared to the true effects. However, they were almost always over-estimated by both approaches in the replications showing the maximal effect values. Even the average values happened to be sometimes higher than the true values (6 to 9 times over 45) (see Figure 4.3). The corresponding phenotypic variances explained by the QTL (PVE) were very close to the true values, but always lower in average. Even the maximal values obtained over the replications were lower than the true values. In addition, they sometimes displayed very low values (6 %). The two approaches performed similarly.

FIGURE 4.3 – Estimated marker effects over 100 replications are plotted (black points) along the genome for nine parameters and for both strategies IA and JA. True simulated QTL effects are shown by red asterisks.



The comparison between estimated and true parameter values also gave similar results between the two approaches (see Figure 4.4). The expected error over the 100 virtual populations was estimated below 2% for LHx, OCp, RAI and RSO, and it was estimated around 6% for KFk, KHk, RSS, TAS and TPF.

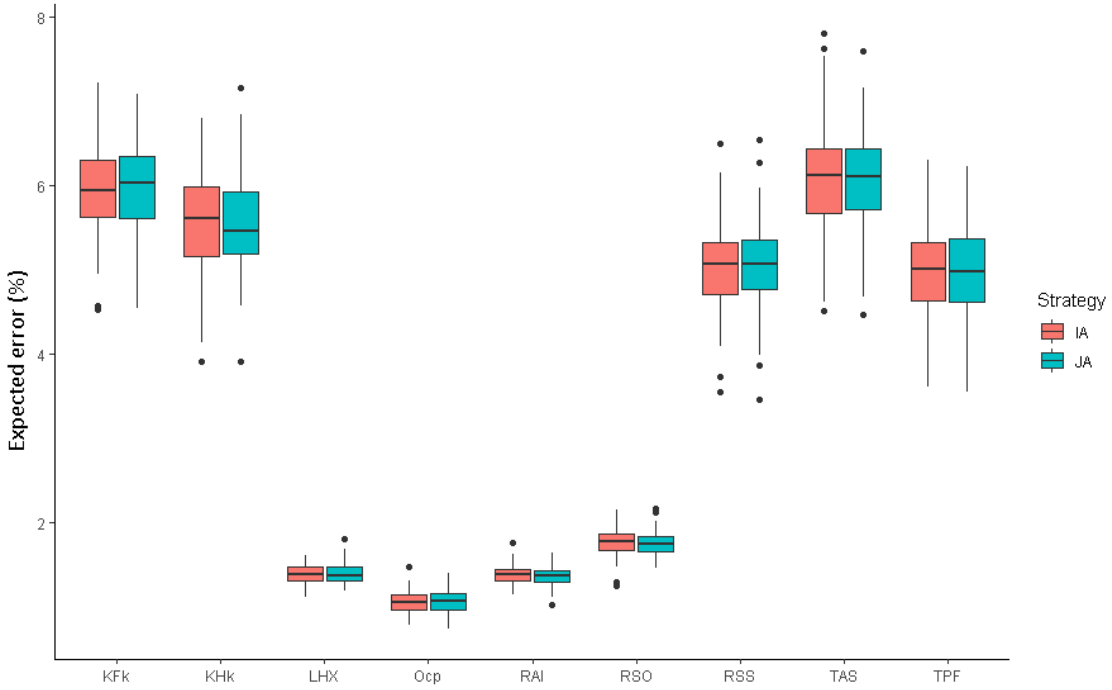


FIGURE 4.4 – Expected error (%) computed on the individual parameters between the true and estimated values obtained with IA and JA over the whole 100 virtual populations.

4.1.9.2 Experimental study

Comparison of the distributions of the sets of parameters

For the IA approach, the set of values of the 9 parameters for the 106 genotypes was taken from Section 3.2. The parameters were calibrated for all genotypes simultaneously by the Population-based model strategy. In addition, the 'JA' approach produced a new dataset of 9 parameters specific of each of the 106 genotypes. In return this dataset was used to generate predictions of concentrations of the four sugars during fruit development for each genotype.
The distributions of the parameter values were quite similar between the two approaches, apart for TAS for which the shift is quite important. Most of the differences between the two distributions lie in the height of the peaks, the length of the distribution queue and in slight shifts. Distributions from JA approach were generally tighter and in most cases, they draw two quite distinct peaks (ie LHx, KFk, KHk, OCp, RSS and TAS) (see Figure 4.5).

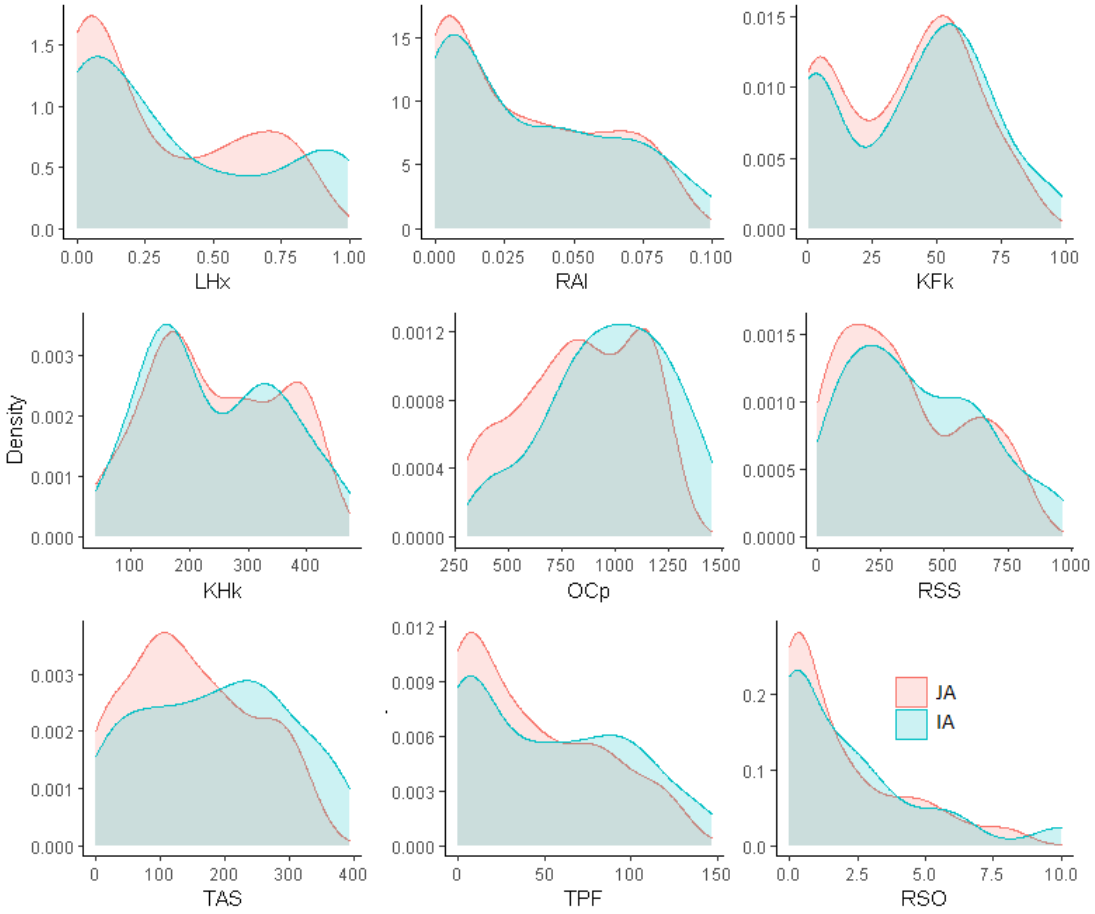


FIGURE 4.5 – Distribution of the estimated values of the nine parameters for the two datasets. The set of values for 'IA' ('Independent analysis') came from Chapter 3. The set of values for JA ('Joint Analysis') were obtained using the procedure described in this Chapter.

Exploration of the predictions of the model calibrated using the two approaches

The sugar model enabled to predict the evolution of the four sugars during fruit development and correctly accounted for variations between genotypes. As presented in Figure 4.6, different levels and profiles could be simulated, as for example increasing or decreasing contents of sucrose or very low levels of fructose. The predictions stemmed from the two approaches used to calibrate the model gave very similar profiles. Predictions were close to the observations for both approaches (see Figure 4.7). The correlations between predictions and observations were high and very similar between approaches, with better results for sucrose and fructose than for glucose. Sorbitol was the least well predicted.

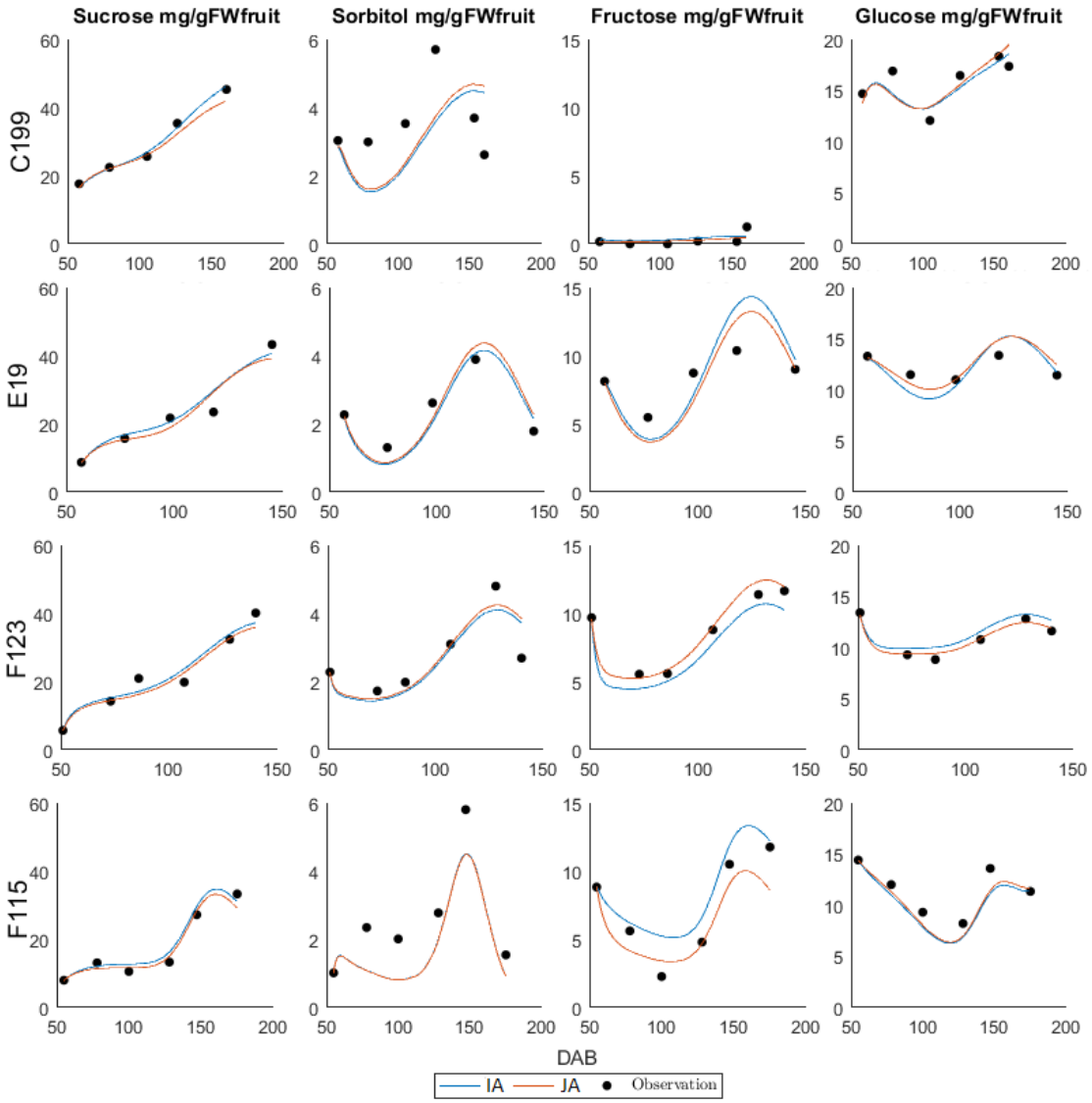


FIGURE 4.6 – Evolution of the concentration ($mggFW^{-1}$) of sugars during fruit development (DAB, days after bloom) for four genotypes. Dots represent experimental data and different colored lines are model simulations using the solution from different approaches : IA (blue) and JA (red).

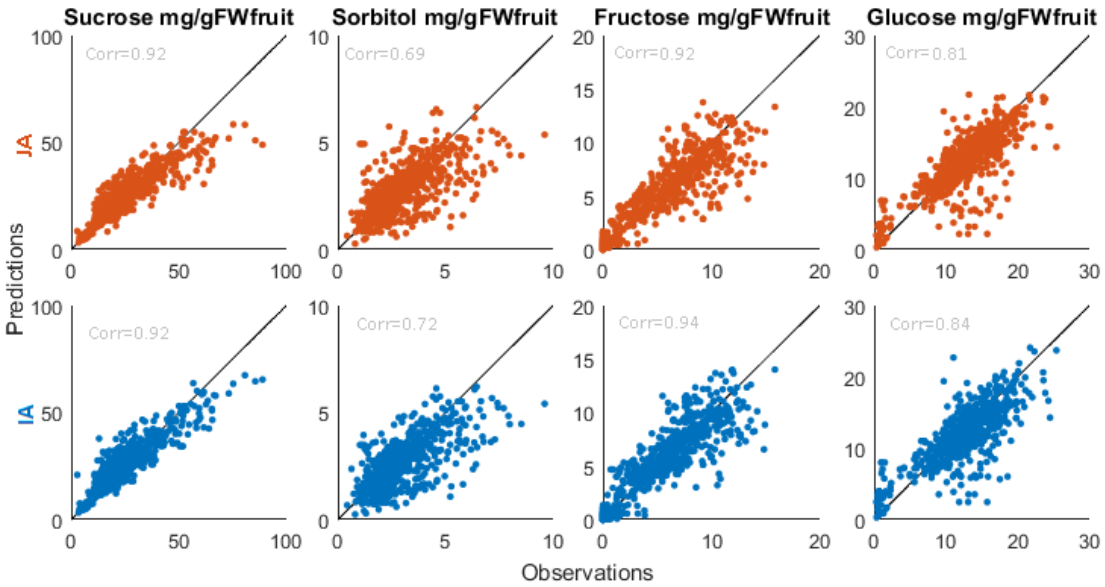


FIGURE 4.7 – Top : Predictions by the model versus observations of sugar contents along fruit development for 106 genotypes. Best individual estimates are plotted for the two optimization methods : 'Independent Analysis' approach (IA) and 'Joint Analysis' approach (JA) ; Corr : correlations between predictions and observations computed for each sugar and for each optimization method.

TABLE 4.5 – The average values of the modelling efficiency (EF) are presented for both IA and JA and for the four sugars.

Method	Sucrose	Sorbitol	Fructose	Glucose
IA	0.84	0.45	0.88	0.68
JA	0.80	0.41	0.83	0.63

Correlations between the values of the parameters calibrated with the two approaches

The correlations between the values of the parameters from the two approaches revealed a very high similarity of the two datasets stemmed from the two approaches (see Figure 4.8). The closeness was very high for KFk, KHk and RSO (correlations > 0.9). It was slightly lower for LHx and then RSS and RAI. Finally, TPF, TAS and OCp displayed smaller correlations but still strong (the smallest is 0.4).
In addition, different parameters proved to be correlated, whatever the approach considered. This is mainly the case of KHk and RSO which were significantly highly negatively correlated. RAI was correlated to RSS and KHk, mostly within the IA dataset,

and to LHx in case of JA dataset. Finally, weaker correlations between other parameters were quite numerous.

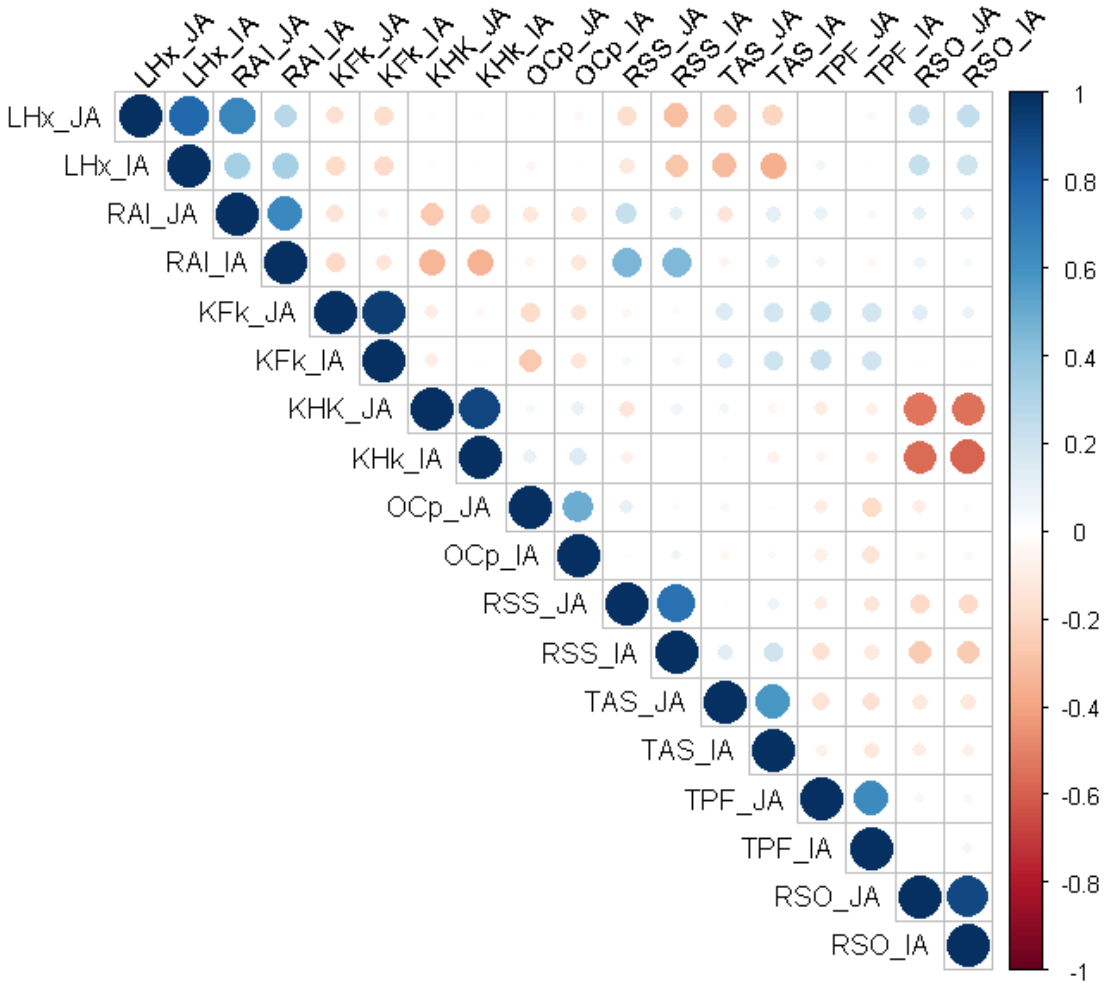


FIGURE 4.8 – Matrix of correlations between parameters, for the nine parameters of the sugar model (LHx, RAI, KFk, KHk, OCp, RSS, TAS and RSO) estimated by 2 different methods. With the 'Independent Analysis' approach (IA), estimations were performed by using Population-Based model and with the 'Joint Analysis' approach (JA) estimations were performed by connecting the model to the genotypes. From blue to red, correlations go from highly positive to highly negative, through no correlation. White shows no significant correlation (Pvalue>0.05). Circle size is proportional to the absolute correlation coefficient.

Results of the QTL analyses performed by the two approaches

With the IA approach, six QTL were detected in total, for four of the nine parameters (RAI, KFk, KHk and RSO) (see Table 4.6). Among them, the major QTL on LG1, with

a very high BLOD (218), explained 88% of the phenotypic variance observed of KFk. Another major QTL (BLOD=44, PVE=36%) was detected on the LG4 of Zephyr map for RSO. The four other QTL detected had small effects. The JA approach enabled to detect nine QTL in total for six of the nine estimated parameters. Among them, five QTL were major ones with BLOD ranging from 44 to 189 and explaining from 36 to 85% of the phenotypic variance observed. They were on four different LG and controlled KFk, KHk, RSO, RAI and TAS parameters. The other four QTL detected had small BLOD. One of the two QTL detected for LHx colocalised with the major QTL of RAI. This is in adequation with the high correlation observed between these two parameters (see Figure 4.8).

TABLE 4.6 – Summary of the QTL detected on the two genetic maps for the 9 parameters (traits), with the two approaches. LG : linkage group ; Marker name and position ; LOD score ; Phenotypic Variance Explained (%)(PVE) : Marker effect ; Map : either DvsS or Zephyr

Method	Trait	LG	Marker	Position	BLOD	PVE(%)	Effect	Map
Independent Analysis	RAI	7	SNP-IGA-783950	60,073	1,282	1.28	-0,139	Zephyr
	KFk	1	MP454-EPPB4232	30,9	218,03	88.8	3,422	DvsS
		2	BPPCT002	18,9	1,125	1.12	0,035	DvsS
	KHk	4	CFF4	56,1	1,825	1.82	0,0349	DvsS
		7	pchms2	35,5	2,225	2.21	0,037	DvsS
	RSO	4	SNP-IGA-409544	42,864	44,48	36.12	2,91	Zephyr
Joint Analysis	LHx	2	SNP-IGA-288054	41	1,031	1.03	0,115	DvsS
		7	pchgms62	41,2	1,022	1.02	0,112	DvsS
	RAI	7	pchgms62	41,2	48,2	38.4	1,198	DvsS
	KFk	1	FRU	31	189,53	85.19	2,385	DvsS
	KHk	7	pchms2	35,5	105,07	65.31	0,608	DvsS
		1	SNP-IGA-104324	48,053	2,06	2.05	0,042	Zephyr
	TAS	3	UDP96008	43,7	44,9	36.39	0,701	DvsS
	RSO	5	EPDCU4658	35,5	1,229	1.23	-0,099	DvsS
		4	SNP-IGA-409544	42,864	67,59	49.39	1,727	Zephyr

The two approaches allowed to detect three similar QTL for KFk, KHk and RSO. They are all three major QTL. Their BLOD scores had same order of magnitude between the two approaches, apart for KHk QTL on LG7 for which the JA approach displayed a BLOD score 50 times higher. Among the three other QTL detected by the IA approach, and not by the JA approach, two were just below the detection threshold (KHK on LG4 and RAI QTL on LG7 of Zephyr map). The same way, five out of the six QTL detected by the JA approach and not the IA approach corresponded to weak but visible peaks and some were not far from being detected by the IA approach.

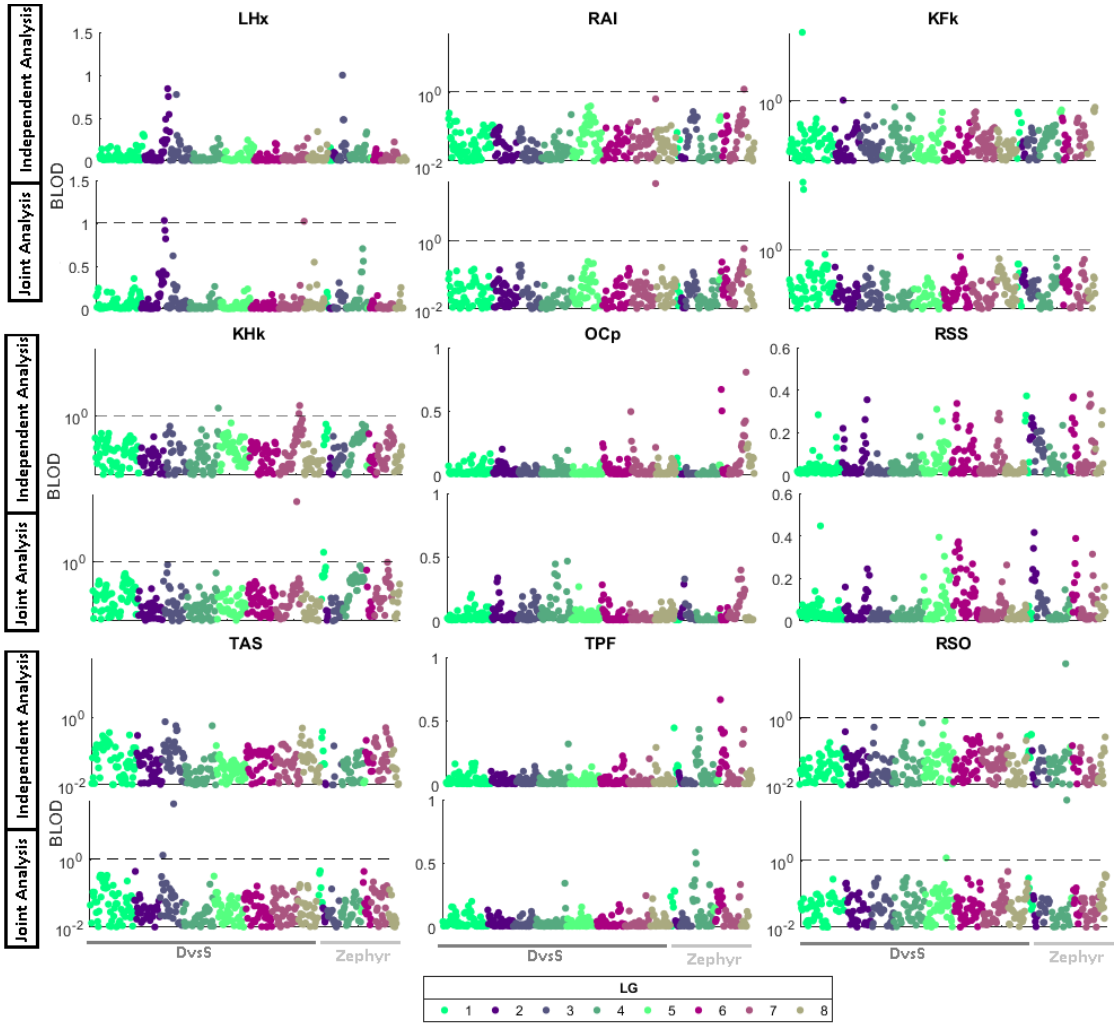


FIGURE 4.9 – Profiles of BLOD scores from the QTL analyses performed on the 9 parameters of the sugar model derived from the 'Independent Analysis' and the 'Joint-Anlaysis' approaches. Different colors indicate different linkage groups (left to right, LG1 to LG8) on the DvsS and Zephyr maps. Dashed lines indicate the 5% genome-wide threshold value for claiming the existence of a QTL at the corresponding marker.

For three out of the nine parameters, OCp, TPF and RSS, no QTL was detected by neither of the two approaches. However, many small peaks could be observed on the profiles obtained with the two approaches, suggesting that these parameters are controlled by many low-effect QTL.

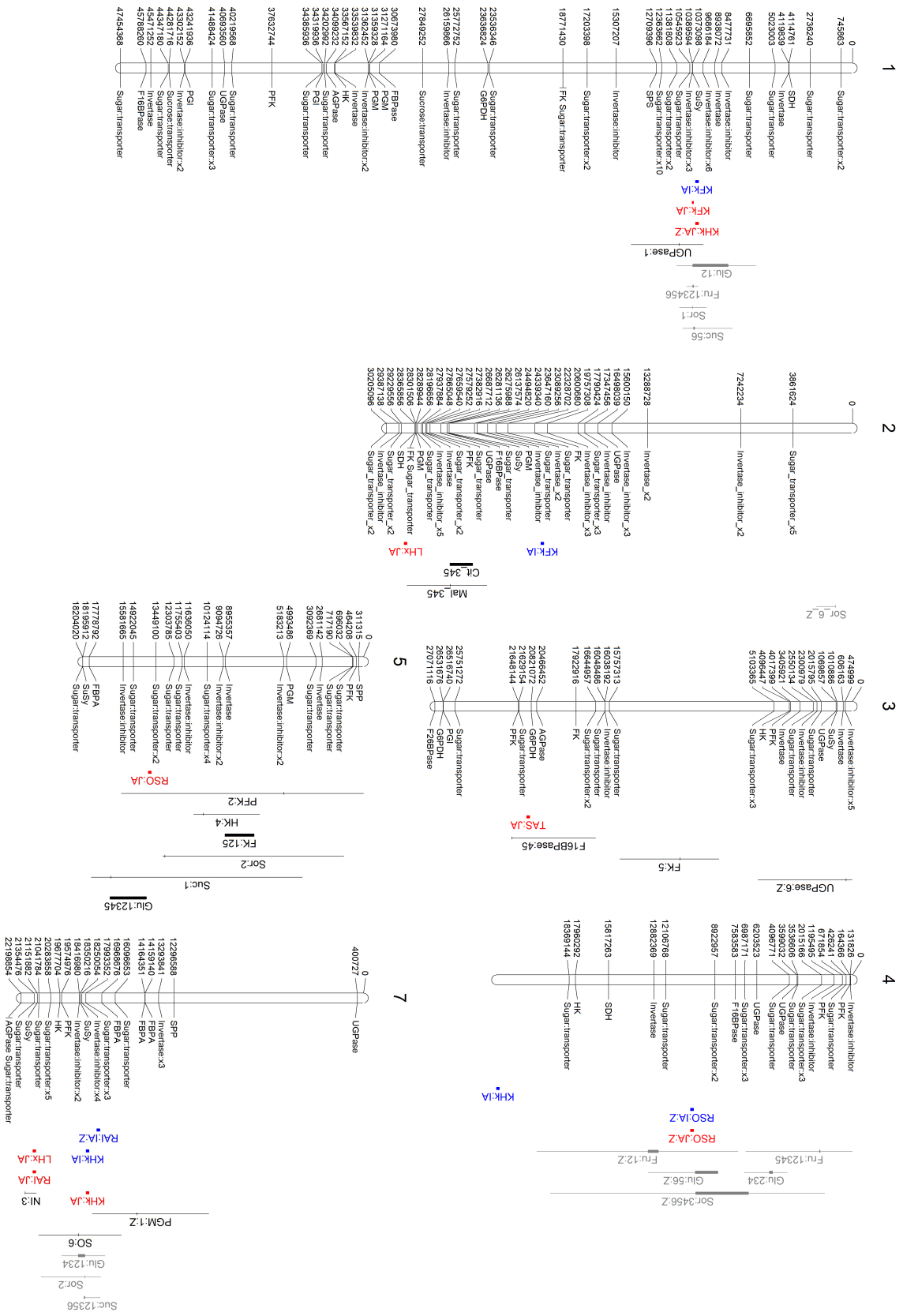


FIGURE 4.10 – (Caption next page.)

FIGURE 4.10 – (Previous page.) Location on the physical map of the QTLs related to the nine parameters of the kinetic model, detected according to two different approaches (IA in blue and JA in red). QTLs detected by Desnoues et al. (2016) on the phenotypic data, namely sugar concentrations (in grey) and enzyme capacities (in black) at six different times of fruit development (numbers) are also positioned. Z at the end of name of the QTLs indicates that the QTL was detected on Zephyr map. Abbreviations : AI, acid invertase ; Cit, citrate ; F16BPase, fructose-1,6-bisphosphatase ; FK, fructokinase ; Fru, fructose ; FW, fresh weight ; Glc, glucose ; HK, hexokinase ; Mal, malate ; NI, neutral invertase ; PFK, ATP-phosphofructokinase ; PGM, phosphoglucomutase ; SDH, sorbitol dehydrogenase ; SO, sorbitol oxidase ; SPS, sucrose phosphate synthase ; Sor, sorbitol ; Suc, sucrose ; SuSy, sucrose synthase ; UGPase, UDP-glucose pyrophosphorylase. Only candidate genes and their locations (pb) are represented in the six of the eight linkage groups where QTL for parameters were detected ; SNP markers have been discarded.

The fifteen QTL detected by the two approaches were positioned on the physical map of the peach genome, together with QTL detected by Desnoues et al. (2016) on the phenotypic data, namely sugar concentrations and enzyme capacities at six different times of fruit development. Candidate genes related to sugar metabolism were represented. A certain number of colocalisations were observable between QTL of parameters and either QTL of metabolites or enzymes and candidate genes of the sugar metabolism. Among the most meaningful colocalisations, the major QTL of KFk, detected with both approaches, colocolised with the major fructose concentration QTL on LG1. On LG2, the QTL of parameter KFk (fructokinase (FK) affinity) colocolised with a HK gene. The QTL of parameter TAS (coefficicent of sucrose transport from cytosol to vacuole) colocolised with a sugar transporter gene on LG3. The QTL of RSO (coefficient of the transfer function between sorbitol and glucose) colocolised with QTL of sorbitol and glucose concentrations on LG4 and LG5. On LG7, one QTL of RAI (coefficient of the transfer function between sucrose and (glucose +fructose) under action of acid invertase enzyme) colocolised with invertase inhibitor genes and with QTL of sucrose and glucose concentrations ; the QTL of KHk (hexokinase (HK) affinity) colocalised with a HK gene and a QTL of glucose concentration.

4.1.10 Discussion

In this work, we used two different approaches, IA and JA, to obtain datasets of parameters for a progeny and to detect genomic regions controlling the parameters. The two approaches tested relied on population-based calibration of the parameters and EBL regression for the association step.

We first developed a simulation study to test the ability of the two approaches to detect major and minor QTL in the case of our experimental framework including a hundred of individuals only and half a thousand of markers. The simulation study gave sufficiently reasonable results to consider the two approaches as appropriate to detect major QTL and to be confident in the results. They displayed very comparable results both in terms of values of parameters calibrated and in terms of QTL detected. This similarity can simply be explained by the involvement of the two same methodologies within the two approaches, namely population-based (MCMC-SAEM) estimation of parameters and EBL regression, ending up to very comparable results. The two approaches proved to be able to detect up to 94% of the major QTL simulated, with quite good estimations of BLOD scores. In return, few false QTL were detected (2% of the major QTL detected). The minor QTL were contrariwise not detected (only 5%), probably because of an experimental design limited by the low number of individuals. The estimations of marker effects for the major QTL were generally underestimated and the phenotypic variance explained by the markers was quite realistic. The results of the experimental study can be then analysed keeping in mind these general characteristics. However, the simulated data were generated considering high heritability of the traits, set to 0.9. Probably, part of the parameters of the sugar model are less heritable than this. So, we can expect the detection power of QTL to be less good in the experimental study. Concerning the differences in results obtained for the nine parameters, the expected error between true and simulated values of the parameters obtained in this simulation study were almost identical to those obtained in Chapter 3 of this thesis. The same trend was obtained, with two groups of parameters more or less easy to calibrate, supporting the impact of the model intrinsic structure on the calibration process. However, it is risky to extrapolate more results, namely the number of QTL detected, to the actual parameters of the model. Indeed, a main part of the differences could come from the random draw of the 5 major QTL and thus corresponding allele frequency and QTL effects deduced from them. To go deeper in the exploration of differences between the parameters, repetitions with different draws of the 5 major QTL should be done in the simulation study.

The two approaches were then applied to the real data. The experimental study resulted in very comparable values of the parameters between the two approaches, although sometimes the distributions were slightly different, especially for the parameters TAS, LHx and OCp. Some correlations between parameters were observed which can be considered as signs of identifiability trouble (Li et al. 2013). Even if the two approaches relied on population-based methods which proved to enhance identifiability of parameters within populations (see Section 3.2), the remaining correlations could impair the estimation of true values of the parameters. Nevertheless, the values of the parameters resulted in accurate predictions of the model by both approaches, with very slightly better results with IA approach in terms of correlations between observations and predictions of sugar concentrations, and in terms of modelling efficiency.

Concerning QTL detection, in most cases, QTL location and peak shapes were very comparable between the two approaches (see Figure 4.9). Indeed, in both cases the calibration using Population-based methods offered an opportunity to use for each genotype the information included in all observations of the progeny and thus to exploit the genetic structure of the progeny. Probably, the small differences in distributions of parameter values obtained by the two approaches led to one part of variations in the QTL profiles. In several cases, BLOD profiles were similar between the two approaches but the BLOD peaks were more often above the threshold for JA approach than for IA approach. This was the case at least for some QTL for LHx and TAS. Consequently, the JA approach allowed to detect 9 QTL whereas only 6 QTL were detected using IA approach. Surprisingly, this superiority of detection power of JA was not foreseeable in light of the results of the simulation study. On the other hand, thanks to this study we may be able to write that the major QTL detected must be reliable and that PVE evaluations were probably underestimated by the two approaches. Unexpectedly, this is contrary to what we are used to writing about the results of QTL analyses by classical linear regression methods which often overestimate QTL effects and PVE. The simulation study also taught us that some major QTL may not have been detected and that undoubtedly QTL with medium or low effects were not detected either. The main reason for this might be limitations of the experimental design. Indeed, in our study, the detection power of QTL is clearly limited by the low number of phenotyped individuals.

Looking more closely at the detection of QTL for each parameter, the differences in success can have several causes. For three of the nine parameters, namely OCp, TPF and RSS, no QTL was detected with neither of the two approaches. Although correlations between RSS and other parameters are quite important, this is not true for the two other parameters. Trouble in identifiability may also intervene. The simulation study performed in Chapter 3 of this thesis showed that the variability of the estimates could be substantial for some parameters. Indeed, important differences between replications were observed in the values of the parameters TPF, RSS, TAS, KHk and KFk (see Figure 3.11). This variability could interfere with the true estimation of the genetic value of the parameters and therefore the detection of QTL. Another matter is that some parameters stand for an integration of many different processes. It is the case of OCp (coefficient of the transfer function between hexose phosphates and other compounds), which could be too integrative of different processes and thus under the control of many loci with low effects.

The model used was built with the concern to give a biological significance to the parameters, in link with the processes involved in sugar accumulation in peach fruit. So it's interesting to compare the QTL detected for the model parameters to those discovered by Desnoues et al. (2016) on the phenotypic data, namely sugar concentrations and enzyme capacities at six different times of fruit development. They found 18

QTL controlling the levels of the four sugars and 21 QTL linked to enzyme capacities and sugar transport. They detected colocations between these two types of QTL. They also observed enzyme QTL co-located with genes coding for the same enzyme and sugar QTL co-located with genes involved in their synthesis, degradation or transport (see Figure 4.10). On LG1, a major locus controls the fructose type of the genotype, either standard or low fructose-to-glucose ratio in fruit at maturity. This locus explained up to 86% of the fructose variability (Desnoues et al. 2016). In this study, a QTL was detected in this area by the two approaches for KFk, directly linked to fructose metabolism and storage. On LG4, RSO (detected with both methods) colocated with a sorbitol and glucose QTL, in the same region as an invertase gene. Finally, a major KHk QTL (detected with both methods and presenting a high PVE value in case of JA approach) colocated on LG7 with a HK gene and glucose and sucrose QTL. Still on LG7, a RAI QTL on LG7 colocated with invertase inhibitor genes and with QTL of sucrose and glucose concentrations. Some QTL were detected with the JA approach only. They were co-located with sugar QTL only (RSO QTL colocated with QTL of sorbitol and glucose concentrations on LG5) or genes only (a TAS QTL colocated with a sugar transporter gene on LG3).

These results militate in favor of the joint approaches. Indeed, the uncertainty usually surrounding parameter calibration (because of identifiability) has been lifted here by the good correspondence of the QTL of parameters with the loci linked to biological data and genome annotation. Finally, the JA approach appeared as slightly better than the IA approach to decipher the genetic control of the parameters of the sugar model we used. However, recent work has shown that the use of non-linear regression methods such as Deep Learning in place of classical linear regressions methods can achieve even better results (Montesinos-López et al. 2018). It would be worth testing such methods as lasso/ridge regressions, random forest, Convolutional Neural Network on a system as complicated as the one studied here.

4.1.11 Conclusion

The functional mapping could not be applied to the kinetic model. On the contrary, the 'Joint-analysis' approach, so called JA, tested by the group of Onogi, appeared through our study as a powerful alternative that may be applied to a greater number of cases. It gave promising results both in terms of estimation of parameters and of QTL detection. Since the parameters and the effects of the markers were estimated simultaneously, it resulted in a significant time saving in computing time. It benefited from the advantages of population-based calibration that proved to be more efficient than the methods performing calibration on each genotype independently (see Chapitre 3). This work made it possible to draw conclusions on the methods available to solve large-scale global optimization problems (LSGO) and calibrate models involving a large number of parameters for numerous linked genotypes.

4.1.12 Appendices

4.1.12.1 Variational Bayesian algorithm for Extended Bayesian Lasso (VB-EBL)

On the basis of variational Bayesian, we can formulate the posterior approximation algorithm for EBL (more details on the framework of variational Bayesian can be found in Tzikas et al. (2008) et Li et al. (2012). The likelihood for the model (4.5) can be specified as

$$p\left(\phi|B,\frac{1}{\sigma_0^2}\right) \propto \prod_{k=1}^{G} \mathcal{N}\left(\phi^{(k)}|\varkappa^{(k)}B,\sigma_0^2\right) \tag{4.11}$$

where $\propto$ stands for "proportional to." (and used to define an unnormalized posterior density).

Note that we assume the same amount of shrinkage for all effects in our study regardless of them being main or interaction effects (cf. Li et al. 2012). Noninformative priors can be used for other unknown parameters in (4.5) and can be represented as

$$p(B_0) \propto 1, \quad p(\frac{1}{\sigma_0^2}) \propto \sigma_0^2 \tag{4.12}$$

These prior hierarchical settings can have a role as a penalty term in the regression, shrinking the marker effects toward zero. In addition the choice of the prior for the variance can be determine the behaviour of the shrinkage. In the following, the remains priors are specified as

$$p(B_m|\frac{1}{\sigma_m^2}) \propto \mathcal{N}\left(B_m|0,\sigma_m^2\right) \tag{4.13}$$

$$p(\frac{1}{\sigma_m^2}|\delta^2,\eta_m^2) \propto Inverse-Gamma\left(\frac{1}{\sigma_m^2}|1,\frac{\delta^2\eta_m^2}{2}\right) \tag{4.14}$$

$$p(\delta^2|\gamma,\nu) \propto Gamma(\delta^2|,\gamma,\nu) \tag{4.15}$$

$$p(\eta_m^2|\Psi,\vartheta) \propto Gamma(\eta_m^2|\Psi,\vartheta) \tag{4.16}$$

where m is the index for m^{th} marker. Thus the logarithm of the joint distribution is

$$\begin{aligned} log\,p(\tilde{y},\phi,\Theta) = &\frac{G}{2}log\sigma_0^2 - \frac{\sigma_0^2}{2}\sum_{k=1}^{G}\left(\phi^{(k)} - B_0 - \sum_{m=1}^{M}\varkappa_m^{(k)}B_m\right)^2 - log\sigma_0^2 + \frac{M}{2}log\sigma_0^2 + \frac{1}{2}\sum_{m=1}^{M}log\sigma_m^2 \\ &- \frac{\sigma_0^2}{2}\sum_{m=1}^{M}\sigma_m^2\varkappa_m^2 - 2\sum_{m=1}^{M}log\sigma_m^2 - \frac{1}{2}\sum_{m=1}^{M}\frac{\delta^2\eta_m^2}{\sigma_m^2} \\ &+ (\gamma-1)log\delta^2 - \nu\delta^2 + (\Psi-1)\sum_{m=1}^{M}log\eta_m^2 - \vartheta\sum_{m=1}^{M}\eta_m^2 + C \end{aligned}$$

where $\widetilde{y} = (\widetilde{y}_{ij}^{(k)}, 1 \leq k \leq G, 1 \leq j \leq n_{SG}, 1 \leq i \leq N_j^{(k)})$, G is the number of genotypes, M is the number of markers, C is a constant, and Θ includes the parameters relating to the genome-wide regression. The regression parameters to be estimated include B_0, B_m, σ_m^2, δ^2, σ_0^2 (residual precision) and η_m^2 and the hyperparameters are γ, ν, Ψ, and ϑ. In the experiments of the presented study, the hyperparameters of γ, ν, Ψ, and ϑ were set to 1.

The factorized posterior distribution of a parameter can be obtained by taking expectations of this density with regard to the other parameters (Li et al. 2012). In practice, parameters of each distribution are initialized and than an iterative algorithm is used to update them successively until convergence. At each iteration the lower bound can be calculated in each iteration and be used as a criterion for stopping. After convergence, we obtain the approximate marginal distributions for each parameter defined in the model.

4.1.12.2 Joint Analysis

In the joint analysis, model parameters (ϕ) and regression parameters (Θ) for EBL model (4.5) were obtained simultaneously. The algorithm was illustrated in the supplementary information of Onogi (2020) and Onogi et al. (2016a). According to the framework, the joint posterior distribution of all parameters is approximated by a factorised distribution to make the posterior distribution tractable.

For model parameters, Markov chain Monte Carlo (MCMC) procedure was used for integration (Hastings 1970). The expectation and the variance of the model parameters are obtained from the MCMC samples and these values are used for the inference of EBL parameters in turn. For parameters relating to the genome-wide regression, we used a variational Bayesian algorithm for the EBL model (see Section 4.1.12.1).

Therefore, the posterior expectations and variances of EBL and model parameters were iteratively updated : First, initial values for model parameters and hyperparameters were set using according to the previous estimated values. σ^2 is arbitrarily set to 0.1. Then, the expectation and the variance of model parameters was computed using MCMC samples for 500 iterations. The sampling interval was 10, yielding 50 MCMC samples. These results were integrated and used for the inference of EBL parameters. In turn, the expectations and variances of the EBL parameters related to model parameters are updated iteratively. This procedure ended when the following criterion was fulfilled. The convergence criterion is

$$\frac{\| \Theta_{new} - \Theta_{old} \|_2}{\| \Theta_{new} \|_2} < 1e-09$$

The algorithm was repeated 100 times and at the last one, the number of iterations (MCMC) is increased to 3000. These samples are output as the posterior distributions of ϕ and used for the posterior distributions of Θ.

4.2 Conclusion et perspectives

Conclusions

La comparaison de méthodes d'optimisation a révélé

- la pertinence d'utiliser des méthodes basées sur la population pour réaliser ensuite une étude génétique
- l'intérêt de l'approche conjointe pour mener ce genre d'étude

Perspectives

L'analyse du contrôle génétique du métabolisme des sucres chez la pêche assistée par le modèle cinétique réduit permettra

- de définir un modèle génétique
- d'intégrer le contrôle génétique dans le modèle cinétique
- de développer une méthodologie pour concevoir des idéotypes afin de progresser vers la sélection virtuelle

5 Discussion générale et perspectives

Sommaire

NOUS nous sommes efforcés de répondre dans cette thèse aux trois problématiques majeures soulevées dans l'introduction, concernant la simplification du modèle dynamique, l'estimation des paramètres pour une population génétique et l'étude de l'architecture du contrôle génétique des paramètres dépendants du génotype. Si les travaux présentés proposent des éléments de réponse à certaines des questions posées, certaines restent encore en suspens, et d'autres émergent également de ces travaux. Nous présentons dans ce dernier chapitre un rappel et une discussion sur les résultats obtenus et proposons quelques perspectives émergeant des résultats.

5.1 Réduction du modèle de métabolisme des sucres chez la pêche

Dans la première partie de la thèse, nous nous sommes intéressés à la réduction d'un modèle du métabolisme des sucres chez la pêche (Desnoues et al. 2018) qui présentait différents inconvénients : d'une part, le nombre de paramètres à estimer, d'autre part le temps d'intégration qui pouvait être conséquent en raison de la non-linéarité du modèle et de fonctions d'entrées dépendant du temps et parfois du génotype. Ainsi, l'ensemble de ces inconvénients empêchaient la calibration du modèle pour la population de 106 génotypes. L'un des objectifs de cette partie était de proposer une stratégie pour simplifier le modèle et faciliter son application à un grand nombre d'individus. Dans ce but, nous avons développé une approche qui combine différentes méthodes de réduction et de simplification en plusieurs étapes parallèles.

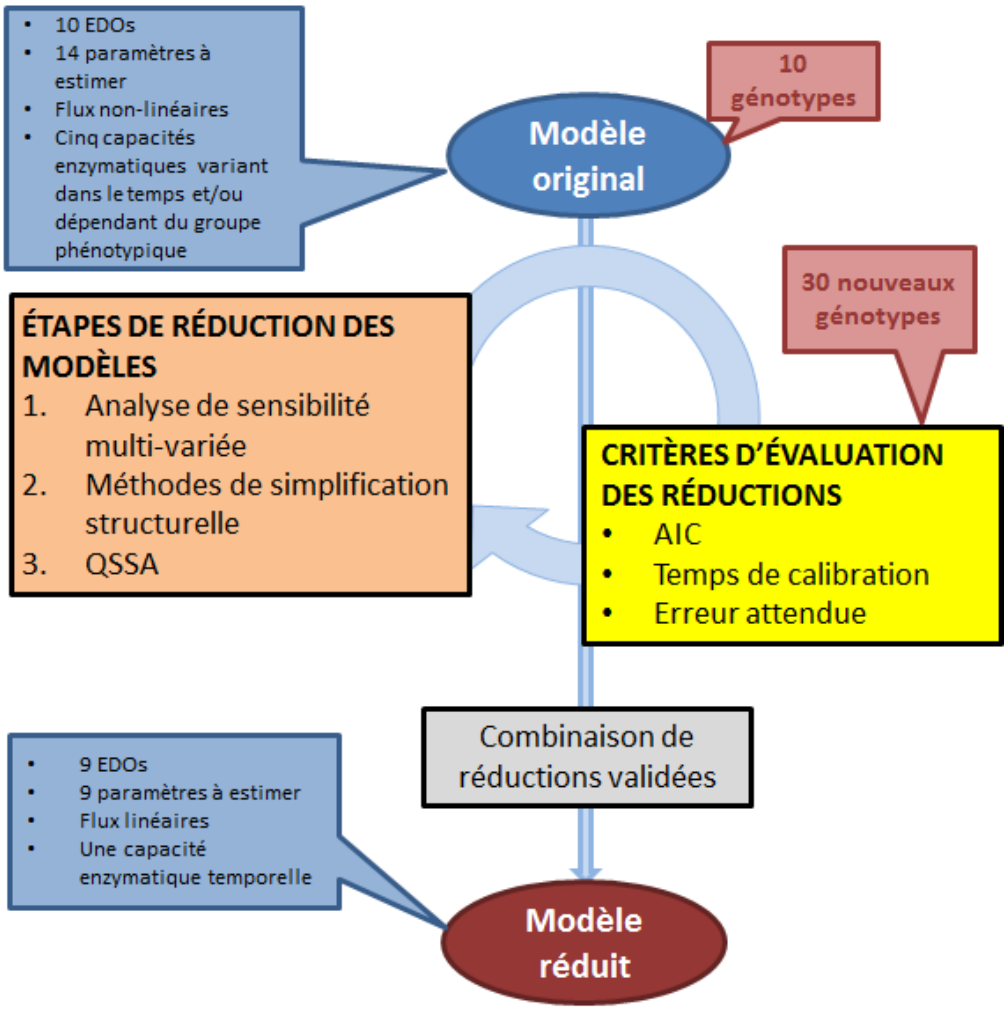


FIGURE 5.1 – Représentation graphique de la stratégie proposée pour simplifier le modèle original (Desnoues et al. 2018). Les spécificités des deux modèles sont indiquées dans les rectangles bleus

Dans un premier temps, une analyse de sensibilité multi-variée (Lamboni et al. 2009) a été appliquée pour identifier les paramètres ayant une influence significative sur les sorties dynamiques du modèle. Nous avons ensuite testé trois simplifications structurelles en termes de réseau et de taux de réaction afin de réduire la complexité du modèle. Enfin, des approches basées sur une échelle de temps et une approximation à l'état quasi-permanent (López Zazueta et al. 2018) ont été appliquées pour réduire le nombre d'EDO du modèle. Les résultats des étapes de réduction individuelles et combinées ont été systématiquement évalués par rapport au modèle d'origine en fonction de trois critères : le critère d'information d'Akaike (AIC), le temps de calibration et l'erreur attendue sur une population de génotypes virtuels.

Les résultats des étapes validées par les différents critères (AIC, temps de calibration et erreur) ont été combinés dans un modèle réduit final (voir Figure 5.1). Ce modèle ne

comporte que 9 paramètres à estimer, des flux linéaires, 9 EDOs et une seule capacité enzymatique temporelle, commune à tous les génotypes.

Les résultats obtenus ont montré un accord satisfaisant entre les prédictions et les données expérimentales. La comparaison entre le modèle réduit et le modèle original a révélé une qualité d'ajustement équivalente et a confirmé un bénéfice important pour la plupart des génotypes. De plus le temps de calibration a été réduit de 40%, ce qui est une amélioration très significative puisque pour certains génotypes la calibration du modèle original pouvait durer jusqu'à 30 heures. Ainsi, la simplification des flux a permis de limiter les problèmes numériques et d'accroître la fiabilité des paramètres estimés. L'ensemble de ces améliorations sont importantes dans un contexte d'études génétiques où un grand nombre d'individus doit être calibré.

Dans cette thèse nous avons proposé une stratégie de réduction qui peut être appliquée à différents modèles dynamiques de réseaux métaboliques et biochimiques. La présence d'une fonction classique comme Michaelis-Menten offre la possibilité de distinguer entre un régime linéaire, pour des faibles concentrations du substrat, et un régime saturé, où la vitesse de la réaction devient quasi-constante. Nous avons exploité cette propriété pour approximer les taux de réaction de notre modèle, en se basant sur la concentration du substrat et sur les valeurs d'affinité précédemment estimées pour le modèle complet (Desnoues et al. 2018). Or cette stratégie n'est pas toujours applicable parce que, d'une part ces valeurs ne sont pas toujours disponibles, et d'autre part, la dynamique du substrat peut couvrir une large gamme de concentration. Dans ce cas, la non-linéarité de la vitesse de réaction doit être considérée dans son intégralité. En alternative, d'autres formes fonctionnelles que Michaelis-Menten peuvent être considérées (Heijnen 2005) qui pourraient présenter une meilleure identifiabilité (Berthoumieux et al. 2011).

Outre la non-linéarité des réactions enzymatiques, la présence au sein d'un réseau métabolique de cycles futiles, c'est-à-dire de réactions qui se produisent simultanément dans des directions opposées et n'ont pas d'effet global sur les flux du système, peut contribuer à des problèmes numériques. Dans notre cas, nous avons détecté deux cycles futiles. Notre stratégie a été de remplacer les réactions antagonistes par une seule réaction efficace préservant l'échange net du flux du système. Cette élimination n'a eu d'impact ni sur la régulation des concentrations, ni sur la détection du groupe phénotypique des génotypes. En revanche, plusieurs études ont constaté que les cycles futiles sont très importants pour la régulation des concentrations de métabolites (Hue et al. 1981 ; Qian et al. 2006 ; Katz et al. 1978). Par exemple, un cycle futile peut permettre de décrire un système oscillant entre deux états, très sensible à de petits changements dans l'activité de l'une des enzymes impliquées (Samoilov et al. 2005).

Les systèmes biologiques sont souvent complexes et impliquent fréquemment un grand nombre de substrats et produits. L'approche que nous avons utilisée pour sim-

plifier le système d'équations repose sur l'idée d'étudier la dynamique des réactions lentes, en supposant que les réactions rapides sont quasi-statiques et capables de suivre quasi-instantanément les variations des variables lentes (Schauer et al. 1983 ; Heinrich et al. 1996). Cependant, la détection des réactions lentes et rapides est souvent difficile et demande un effort supplémentaire en théorie et analyse du modèle. C'est particulièrement compliqué dans le cas de modèles ayant un très grand nombre de variables qui peuvent interagir entre elles.

En conséquences, la stratégie développée ici ne peut pas être généralisée à tous les modèles et le cas de chaque modèle nécessite d'être adressé par une étude spécifique. Cependant, le schéma de réduction construit à partir de différentes méthodes et impliquant des critères d'évaluation, peut servir d'exemple de démarche pour la réduction d'autres modèles dynamiques. Dans ce but, nous avons proposé un critère basé sur la simulation d'un grand nombre de génotypes virtuels et la comparaison systématique de l'erreur entre le modèle original et le modèle réduit.

5.2 Estimation des paramètres pour une population génétique de pêchers

Le deuxième objectif de cette thèse portait sur la calibration du modèle réduit sur l'ensemble de la population de 106 génotypes. Pour arriver à ce but, nous avons comparé deux approches de calibration : i) chaque génotype est calibré indépendamment et ii) l'ensemble de la population est calibré simultanément (Baey et al. 2018). De plus, dans le premier cas, le problème de calibration a été formulé de deux manières : comme un problème d'optimisation mono-objectif ou multi-objectif.

Les résultats des deux approches ont été d'abord évalués sur un jeux de données simulées, afin de determiner la precision et la robustesse de l'estimateur statistique associé, en conditions controlées. Les resultats obtenus sur la population réelle ont ensuite été comparés afin de déterminer la meilleure façon d'estimer les paramètres génétiques sur l'ensemble de la population. Les deux approches ont montré une bonne capacité predictive, avec un accord satisfaisant entre les prédictions du modèle et les données, sur l'ensemble de la population. En revanche, la qualité de l'estimation des paramètres s'est averée peu robuste pour les approches où les génotypes sont calibrés indépendamment, avec de fortes variations dans les valeurs estimées selon les répétitions des algorithmes, que ce soit sur les données simulées ou sur les données réelles. Au contraire, l'approche basée sur la population s'est montrée capable d'estimer les paramètres du modèle de manière sensiblement plus fiable. Une bonne convergence de l'algoritme a été observée pour les paramètres moyens de la population ainsi que pour leur covariance. De plus le fait que cette approche soit basée sur l'ensemble des données de la population peut se révéler important dans le cas d'une étude génétique où les génotypes peuvent avoir des structures génétiques

proches et liées.

D'un point de vue pratique, l'approche génotype par génotype est l'approche la plus classique et celle la plus souvent utilisée dans la littérature. Aussi, de nombreux algorithmes de calibration sont déjà disponibles pour l'application à un génotype individuel, ce qui rend cette strategie facilement applicable. Quant à l'approche à l'échelle de la population qui prend en compte la variabilité entre les génotypes, elle est plus récente (par exemple Fournier et al. (1999) et Cournède et al. (2008)). Les premiers travaux montrent que leur application s'avère délicate en pratique car il est généralement difficile de décrire et identifier une forme unique du modèle, en partant de la distribution et des variabilités inter- et intra-individuelles sur l'ensemble de la population. De plus, la calibration de ces modèles n'est pas toujours facile puisqu'elle nécessite des informations sur tous les génotypes de la population et repose généralement sur un grand nombre de paramètres. Ainsi, les algorithmes d'estimation à l'échelle de la population sont moins carrossés, peu généralisables et nécessitent une adaptation spécifique pour chaque type de modèle.

Les raisons des grandes differences entre les résultats d'estimation des approches genotype-par-genotype et population peuvent être trouvées dans la structure des algorithms utilisés par les premières approches et dans la quantité (limitée) de données disponibles par rapport au nombre de paramètres à estimer. Plusieurs pistes d'investigation sont evisageables. Une première étape pourrait être de comparer plusieurs algorithmes de calibration. Dans cette thèse, l'algorithme génétique (GA) (Sivanandam et al. 2008) et l'algorithme génétique de tri non dominé II (NSGA-II) (Deb et al. 2002) ont été utilisés. Cependant, il existe dans la littérature d'autres algorithmes qui pourraient également être testés. Par exemple, pour l'optimisation avec une seule fonction objectif, on pourrait utiliser l'algorithme d'évolution différentielle (DE) (Storn et al. 1997) ou Particle Swarm Optimization (PSO) (Kennedy et al. 1995). Il y a aussi des algorithmes de type multi-objectif, tels que l'algorithme d'évolution différentielle multi-objectif (MODE) (Babu et al. 2005) ou l'algorithme Multi-objectif particle swarm optimization (MOPSO) (Coello et al. 2002).

Concernant la formulation du problème d'optimisation, les travaux d'analyse de sensibilité sur la fonction objectif (Chapitre 3) ont montré qu'au moins 3 des 9 paramètres à estimer avaient un indice de sensibilité inferieur à 0.1, à la fois sur la somme des ecarts des 4 sucres et sur chaque sucre individuellement. Pour une amélioration des résultats, une possible option consisterait à reduire le nombre de paramètres à estimer en fixant la valeur de paramètres peu sensibles. Ce-ci permettrait en effet d'avoir un ratio données sur paramètres à estimer plus favorable, contraignant ainsi davantage l'agorithme d'optimisation. Une approche de ce type, couplé à une procédure de sélection basée sur l'AIC, a été utilisée avec succès par Mathieu et al. (2018) pour améliorer la calibration d'un modèle structure-fonction. Une limitation de cette stratégie, par contre, réside dans le choix de la valeur de référence des paramètres à

fixer. Dans Mathieu et al. (2018), les valeurs de référence ont été selectionnées sur la base des calibrations précédentes, sur plusieurs espèces, ainsi que sur des information de la littérature. Dans notre cas, les valeurs moyennes estimées sur la population pourraient être utilisées en première approximation. Dans le cas des paramètres KFk et KHk pour lesquels la distribution est fortement bimodale, on pourrait envisager de selectionner deux valeurs de référence, selon le phénotype en fructose du génotype considéré.

Une autre option consiste à utiliser un approche population-based. Dans cette thèse, nous avons fait l'hypothèse que chaque effet était aléatoire, c'est-à-dire avec une variance non nulle. Ceci nous a notamment permis d'obtenir une vraisemblance complète appartenant à la famille exponentielle, et d'écrire l'étape de maximisation explicitement. Cependant, dans la lignée de la discussion ci-dessus, il pourrait être intéressant de considérer certains effets, et notamment ceux relatifs aux paramètres peu sensibles (Baey et al. 2018), comme fixes, c'est-à-dire avec une variance nulle. Dans ce cas, le modèle n'appartiendrait plus à la famille exponentielle. L'étape de maximisation pourrait ne pas être explicite pour les paramètres considérés, auquel cas des méthodes d'optimisation de type quasi-Newton devraient être mobilisées. L'avantage de cette approche alternative réside dans le fait que la valeur des effets fixes est estimée directement par la procédure d'optimisation, en même temps que les effets aléatoires.

Dans le cadre de l'approche population, il serait également intéressant de tester plusieurs structures de covariance. Jusqu'à présent, seules deux structures ont été testées : la première en imaginant une corrélation entre les effets aléatoires et l'autre en assumant une indépendance entre eux. Aussi pourrait-on également étudier d'autres cas. Mais cela nécessite une connaissance biologique des paramètres pour bien choisir le type de corrélation entre eux. Si, par exemple, on suppose γ comme

$$\begin{pmatrix} \sigma^2_{LHX} & & & & & & & & \\ & \sigma^2_{RAI} & & & & & & & \\ & & \sigma^2_{KFk} & & & & & & \\ & & & \sigma^2_{KHk} & & & & & \\ & & & & \sigma^2_{OCp} & & & & \\ & & & & & \sigma^2_{RSS} & & & \\ & & & & & & \sigma^2_{TAS} & & \\ & & \sigma_{KFk,TPF} & & & & & \sigma^2_{TPF} & \\ & & & & & & & & \sigma^2_{RSO} \end{pmatrix}$$

on considère que seules les paramètres KFk et TPF sont corrélés. La dépendance entre l'affinité de la fructokinase (Fk) et le transport du fructose du cytosol vers la vacuole pourrait s'expliquer par le fait que FK crée un gradient qui permet au transporteur de fructose de fonctionner. Un test dans ce sens serait utile et à comparer avec les autres modèles pour valider ou non cette hypothèse.

Une option supplémentaire pourrait être de définir une distribution bimodale pour le paramètre KFk. En effet, l'estimation de ce paramètre a montré une différence significative entre l'affinité de fructokinase (Fk) des génotypes 'standard' et ceux 'low fructose' (voir Chapitre 2 $KFk_{FRU} >> KFk_{SSFRU}$). Cette option pourrait améliorer la performance de l'algorithme en initialisant la distribution de KFk et la détection de la moyenne pour les deux phénotypes.

Concernant la validation du modèle mixte, les résultats que nous avons présentés sont basés sur un seul jeu de données provenant d'expérimentations réalisées en 2012. Il faudrait étoffer les résultats de l'étude en évaluant le modèle sur d'autres jeux de données, incluant toujours un grand nombre de génotypes différents. De plus, il serait intéressant de tester le modèle sur des sous populations. Comme par exemple, de séparer les deux groupes de phénotypes (Standard et low fructose) et ainsi tester d'une part la performance d'estimation avec une petite population et d'autre part de comparer les estimations de deux populations pour identifier eventuelles differences entre les paramètres estimés en fonction du phenotype.

5.3 Analyse du contrôle génétique du métabolisme des sucres chez la pêche

Le dernier objectif de la thèse était l'analyse du contrôle génétique du métabolisme des sucres chez la pêche basée sur les paramètres génotype-dépendant du modèle cinétique. En effet, chaque paramètre du modèle a une interprétation biologique ; une telle analyse permet donc de mettre en évidence un lien entre un processus décrit dans le modèle mathématique et un polymorphisme génétique responsable des variations des teneurs en sucres dans la pêche.

Dans cette étude, deux approches ont été utilisées pour analyser le contrôle génétique : i) une approche indépendante où, dans un premier temps, les paramètres du modèle sont estimés puis dans un second temps, une cartographie de QTL est réalisée ; ii) une approche jointe ou "one-step". Dans ce cas les paramètres du modèle et les effets des marqueurs sont estimés simultanément. Les deux approches ont été comparées pour leur puissance de détection de QTL. L'approche jointe a présenté des avantages majeurs par rapport à l'approche indépendante. D'une part, les paramètres et les effets des marqueurs sont estimés simultanément ce qui peut conduire à un gain significatif en temps de calcul. D'autre part, le nombre de QTL détectés est légèrement augmenté.

Pour aller plus loin, une première orientation serait de comparer d'autres méthodes de régression pour sélectionner la méthode la plus puissante. La méthode de régression utilisée dans les deux approches considérées était l'"Extended Bayesian Lasso" (EBL, (Mutshinda et al. 2010)). Il existe plusieurs autres méthodes de régression dans la littérature telles que, "Bayesian lasso" (BL : (Park et al. 2008)), "weighted Bayesian shrinkage regression" (wBSR : (Hayashi et al. 2010)) or "Bayesian mixture regression"

(MIX : (Luan et al. 2009))... Ces méthodes sélectionnent les variables importantes (c'est-à-dire les variables liées aux variables de réponse) parmi les variables prédictives définies dans le système de départ. Les différentes méthodes de régression diffèrent par la structure du système, mais dans tous les cas la distribution des effets des marqueurs et l'effet de la variance sont spécifiés ainsi que le nombre d'hyper-paramètres et leur spécification. Cependant, la mise en oeuvre de ces différentes méthodes pour leur comparaison demande des efforts importants pour implémenter et coder les méthodes dans un langage spécifique.

Une approche qu'il pourrait également être intéressant d'implémenter, dans le but d'estimer les effets des marqueurs, est d'ajouter une interaction entre des marqueurs moléculaires. Cette étape pourrait contribuer à mieux décrire la variation phénotypique observée en prenant en compte des effets d'épistasie entre loci, fréquemment décrits dans les réseaux de gènes. Pour cela, le modèle suivant pourrait être proposé :

$$y^{(k)} = B_0 + \sum_{m=1}^{M} \varkappa_m^{(k)} B_m + \sum_{u<v}^{M} \varkappa_u^{(k)} \varkappa_v^{(k)} B_{uv} + e_k, \tag{5.1}$$

avec $y^{(k)}$ représente la valeur phénotypique de l'individu k, B_0 la moyenne de la population, M le nombre de marqueurs du génome entier, B_m représente l'effet du marqueur m, B_{uv} représente l'effet d'interaction entre la paire de marqueurs (u, v) $\varkappa_m^{(k)}$ désigne le génotype au marqueur m pour l'individu k et vaut 1 pour Q_1Q_1 et 2 pour Q_1Q_2 et e_k est l'erreur résiduelle qui est supposée suivre une distribution normale de moyenne zéro et de variance constante.

5.4 Perspectives générales

Les systèmes biologiques permettent d'explorer l'effet de modifications génétiques et/ou environnementales sur des traits complexes (Hammer et al. 2006). Pour cela il est nécessaire d'intégrer des informations relatives au contrôle génétique dans les modèles de fonctionnement biologique ('crop-model', 'process-based' model, modèle cinétique,...).

L'une des perspectives à moyen terme de ce travail de thèse est l'intégration du contrôle génétique dans le modèle métabolique. Cette étape permettra ensuite de développer une méthodologie pour concevoir des idéotypes afin de progresser vers la sélection virtuelle. Les enjeux et méthodes d'intégration du contrôle génétique dans les modèles de plantes ou de culture ont fait l'objet de chapitres de livres récents (Martre et al. 2015 ; Baldazzi et al. 2016) offrant une revue des travaux sur le sujet. Une stratégie assez courante consiste à intégrer les effets des QTL de paramètres dans les modèles de fonctionnement. Le principe est de remplacer les valeurs des paramètres génotype-dépendants du modèle par une combinaison des effets des allèles à chaque QTL détecté pour ces paramètres. C'est la stratégie déjà utilisée par Quilot et al. (2005) pour un modèle de croissance des fruits de la pêche. Les résultats du Chapitre 4 sont

cependant un peu limités pour permettre de définir ces combinaisons. En considérant les résultats issus de l'approche jointe, six paramètres génotype-dépendant du modèle pourraient être remplacés par des combinaisons d'effets d'allèles qui ne comprendraient que 1 ou 2 loci. Ceci n'est pas suffisant pour reconstruire des distributions raisonnablement continues des paramètres, similaires à celles observées. Conscients que la puissance du dispositif expérimental est faible (nombre réduit d'individus et de marqueurs), on pourrait surmonter cette difficulté en relachant le seuil de sélection des QTL pour augmenter le nombre de marqueurs conservés. On se rapprocherait ainsi de l'esprit des méthodologies développées pour la sélection génomique. On pourrait également tester d'autres approches, d'intelligence artificielle, pour tenter de détecter plus de loci impliqués dans le contrôle des paramètres. A partir des combinaisons des effets alléliques, chaque génotype est ensuite simulé par le modèle intégré à partir d'un jeu de paramètres dont les valeurs dépendent des combinaisons alléliques aux QTL. Le modèle intégré peut ainsi simuler les concentrations en sucres pour des génotypes existants ainsi que des génotypes virtuels.

L'étape de conception d'idéotypes proprement dite requiert une étape d'optimisation de façon à proposer des génotypes virtuels aux qualités sucrées améliorées. Classiquement, cette étape d'optimisation portait sur les valeurs des paramètres et ne tenait donc pas compte du contrôle génétique des paramètres. Prendre en compte dans le schéma d'optimisation l'architecture génétique complexe contrôlant les paramètres du modèle est en effet la clé pour espérer créer réellement les solutions apportées par la procédure d'optimisation. En effet, la limitation majeure de la conception d'idéotypes assistée par modèle est le manque de réalisme de ces idéotypes qui ne peuvent jamais être obtenus par un sélectionneur, simplement parce qu'ils enfreignent des contraintes physiologiques et génétiques. De façon à affiner les solutions vers des idéotypes plus réalistes, il est plus judicieux d'optimiser les combinaisons d'allèles qui contrôlent les paramètres génotype-dépendant.

Pour cela, l'algorithme d'optimisation doit être adapté pour explorer un espace de solution discret, binaire, décrivant la présence ou l'absence de l'effet de chaque QTL sur la valeur de chaque paramètre du modèle. Peu de travaux se sont attaqués à ce problème, si ce n'est l'étude menée par Quilot-Turion et al. (2016) visant à comparer une approche classique basée sur l'optimisation des valeurs des paramètres avec l'approche consistant en l'optimisation directe des allèles contrôlant les paramètres. Leurs résultats ont montré que l'utilisation d'un modèle génétique permet de restreindre la dimension de l'espace des paramètres vers des combinaisons plus réalisables de valeurs de paramètres, reproduisant les relations entre les paramètres observés chez une descendance réelle. L'algorithme d'optimisation, adapté pour traiter à la fois des problèmes continus et combinatoires, pourrait être utilisé dans notre cas. Dans un second temps, les contraintes génétiques décrivant la liaison et les interactions entre les gènes (pléiotropie et épistasie) ou QTL pourraient être ajoutées, ce qui réduirait encore l'espace de recherche admissible par l'algorithme.

Bibliographie

AKDEMIR Beavis, Fritsche-Neto Singh Isidro-Sánchez, 2019. Multi-objective optimized genomic breeding strategies for sustainable food improvement. *Heredity*. T. 122, 672—683.

ANDERSON, James et al., 2011. Model decomposition and reduction tools for large-scale networks in systems biology. *Automatica*. T. 47, n° 6, p. 1165 -1174. ISSN 0005-1098. Special Issue on Systems Biology.

APGAR, Joshua F. et al., 2010. Sloppy models, parameter uncertainty, and the role of experimental design. *Molecular BioSystems*. T. 6, n° 10, p. 1890. ISSN 1742-206X. Disp. à l'adr. DOI : 10.1039/b918098b.

APRI, Mochamad et al., 2012. Complexity reduction preserving dynamical behavior of biochemical networks. *Journal of theoretical biology*. T. 304, p. 16-26.

BABU, BV et al., 2005. Multiobjective differential evolution (MODE) for optimization of adiabatic styrene reactor. *Chemical Engineering Science*. T. 60, n° 17, p. 4822-4837.

BAEY, Charlotte et al., 2016. A non linear mixed effects model of plant growth and estimation via stochastic variants of the em algorithm. *Communications in Statistics-Theory and Methods*. T. 45, n° 6, p. 1643-1669.

BAEY, Charlotte et al., 2018. Mixed-effects estimation in dynamic models of plant growth for the assessment of inter-individual variability. *Journal of agricultural, biological and environmental statistics*. T. 23, n° 2, p. 208-232.

BALDAZZI, Valentina et al., 2016. Challenges in integrating genetic control in plant and crop models. In : *Crop Systems Biology*. Springer, p. 1-31.

BANDARA, Samuel et al., 2009. Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput Biol*. T. 5, n° 11, e1000558.

BARABÁSI, A-L et al., 2004. Network biology : understanding the cell's functional organization. *Nat. Rev. Genet*. T. 5, n° 2, p. 101-113.

BARRASSO, Caterina et al., 2019. Model-based QTL detection is sensitive to slight modifications in model formulation. *PloS one*. T. 14, n° 10.

BEAUVOIT, B. P. et al., 2014. Model-assisted analysis of sugar metabolism throughout tomato fruit development reveals enzyme and carrier properties in relation to vacuole expansion. *Plant Cell*. T. 26.

BECKER, Verena et al., 2010. Covering a broad dynamic range : information processing at the erythropoietin receptor. *Science*. T. 328, n° 5984, p. 1404-1408.

BELLMAN, Ror et al., 1970. On structural identifiability. *Mathematical biosciences*. T. 7, n° 3-4, p. 329-339.

BELLU, Giuseppina et al., 2007. DAISY : A new software tool to test global identifiability of biological and physiological systems. *Computer methods and programs in biomedicine*. T. 88, n° 1, p. 52-61.

BERTHOUMIEUX, Sara et al., 2011. Identification of metabolic network models from incomplete high-throughput datasets. *Bioinformatics*. T. 27, n° 13, p. i186-i195.

BERTIN, Nadia et al., 2010. Under what circumstances can process-based simulation models link genotype to phenotype for complex traits ? Case-study of fruit and grain quality traits. *Journal of experimental botany*. T. 61, p. 955-67.

BLUMAN, George W et al., 2013. *Symmetries and differential equations*. Springer Science & Business Media.

BOGARD, Matthieu et al., 2014. Predictions of heading date in bread wheat (Triticum aestivum L.) using QTL-based parameters of an ecophysiological model. *Journal of experimental botany*. T. 65, n° 20, p. 5849-5865.

BOITARD, Simon et al., 2006. Linkage disequilibrium interval mapping of quantitative trait loci. *BMC genomics*. T. 7, n° 1, p. 54.

BOOTE, K.J. et al., 2001. Physiology and modelling of traits in crop plants : implications for genetic improvement. *Agricultural Systems*. T. 70, n° 2, p. 395 -420. ISSN 0308-521X.

BORGONOVO, Emanuele et al., 2004. Sensitivity analysis in investment project evaluation. *International Journal of Production Economics*. T. 90, n° 1, p. 17-25.

BOULIER, François et al., 2009. Towards an automated reduction method for polynomial ODE models of biochemical reaction systems. *Mathematics in Computer Science*. T. 2, n° 3, p. 443-464.

BROCHOT, Céline et al., 2005. Lumping in pharmacokinetics. *Journal of pharmacokinetics and pharmacodynamics*. T. 32, n° 5-6, p. 719-736.

BROMAN, Karl W et al., 2009. *A Guide to QTL Mapping with R/qtl*. Springer.

BRUNEL, S et al., 2009. Using a model-based framework for analysing genetic diversity during germination and heterotrophic growth of Medicago truncatula. *Annals of Botany*. T. 103, n° 7, p. 1103-1117.

BURNHAM, Kenneth P. et al., 2002. *Model selection and multimodel inference : a practical information-theoretic approach*. New York : Springer. ISBN 9780387224565.

CACUCI, Dan G et al., 2005. *Sensitivity and uncertainty analysis, volume II : applications to large-scale systems*. CRC press.

CAMPOLONGO, Francesca et al., 2007. An effective screening design for sensitivity analysis of large models. *Environmental modelling & software*. T. 22, n° 10, p. 1509-1518.

CARIBONI, J. et al., 2007. The role of sensitivity analysis in ecological modelling. *Ecological Modelling*. T. 203, n° 1, p. 167 -182.

CARRENO-QUINTERO, Natalia et al., 2013. Genetic analysis of metabolome–phenotype interactions : from model to crop species. *Trends in Genetics*. T. 29, n° 1, p. 41-50.

CHAMBERS, John M et al., 1992. *Statistical models in S*. Wadsworth & Brooks/Cole Advanced Books & Software Pacific Grove, CA.

CHARDON, Fabien et al., 2013. Leaf fructose content is controlled by the vacuolar transporter SWEET17 in Arabidopsis. *Current Biology*. T. 23, n° 8, p. 697-702.

CHOU, I-Chun et al., 2009. Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical biosciences*. T. 219, n° 2, p. 57-83.

CHURCHILL, Gary A et al., 1994. Empirical threshold values for quantitative trait mapping. *Genetics*. T. 138, n° 3, p. 963-971.

COELLO, CA Coello et al., 2002. MOPSO : A proposal for multiple objective particle swarm optimization. In : *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600)*. T. 2, p. 1051-1056.

COELLO, Carlos A Coello et al., 2007. *Evolutionary algorithms for solving multi-objective problems*. Springer.

CONN, A et al., 1997. A globally convergent Lagrangian barrier algorithm for optimization with general inequality constraints and simple bounds. *Mathematics of Computation*. T. 66, n° 217, p. 261-288.

CONN, Andrew R et al., 1991. A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds. *SIAM Journal on Numerical Analysis*. T. 28, n° 2, p. 545-572.

CONSTANTINESCU, Dario et al., 2016. Model-Assisted Estimation of the Genetic Variability in Physiological Parameters Related to Tomato Fruit Growth under Contrasted Water Conditions. *Frontiers in Plant Science*. T. 7, n° December, p. 1-17. ISSN 1664-462X. Disp. à l'adr. DOI : 10.3389/fpls.2016.01841.

COOPER, Mark et al., 2009. Modeling QTL for complex traits : detection and context for plant breeding. *Current opinion in plant biology*. T. 12, n° 2, p. 231-40. ISSN 1879-0356. Disp. à l'adr. DOI : 10.1016/j.pbi.2009.01.006.

COURNÈDE, Paul-Henry et al., 2008. Computing competition for light in the GREEN-LAB model of plant growth : a contribution to the study of the effects of density on resource acquisition and architectural development. *Annals of Botany*. T. 101, n° 8, p. 1207-1219.

CUKIER, RI et al., 1973. Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. *The Journal of chemical physics*. T. 59, n° 8, p. 3873-3878.

CURIEN, Gilles et al., 2009. Understanding the regulation of aspartate metabolism using a model based on measured kinetic parameters. *Molecular systems biology*. T. 5, n° 1, p. 271.

DA SILVA, David et al., 2014. Influence of the variation of geometrical and topological traits on light interception efficiency of apple trees : sensitivity analysis and metamodelling for ideotype definition. *Annals of botany*. T. 114, n° 4, p. 739-752.

DALLA MAN, Chiara et al., 2006. A system model of oral glucose absorption : validation on gold standard data. *IEEE Transactions on Biomedical Engineering*. T. 53, n° 12, p. 2472-2478.

DAVIDIAN, Marie et al., 1995. *Nonlinear models for repeated measurement data*. CRC press.

DAVIDIAN, Marie et al., 2003. Nonlinear models for repeated measurement data : An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*. T. 8, n° 4, p. 387-419. ISSN 10857117. Disp. à l'adr. DOI : 10.1198/1085711032697.

DEB, Kalyanmoy et al., 2002. A fast and elitist multiobjective genetic algorithm : NSGA-II. *IEEE transactions on evolutionary computation*. T. 6, n° 2, p. 182-197.

DESNOUES, Elsa et al., 2014. Profiling sugar metabolism during fruit development in a peach progeny with different fructose-to-glucose ratios. *BMC Plant Biology*. T. 14, n° 1, p. 336. ISSN 1471-2229.

DESNOUES, Elsa et al., 2016. Dynamic QTLs for sugars and enzyme activities provide an overview of genetic control of sugar metabolism during peach fruit development. *Journal of experimental botany*. T. 67, n° 11, p. 3419-3431.

DESNOUES, Elsa et al., 2018. A kinetic model of sugar metabolism in peach fruit reveals a functional hypothesis of a markedly low fructose-to-glucose ratio phenotype. *The Plant Journal*. T. 94, n° 4, p. 685-698.

DIRLEWANGER, Elisabeth et al., 1999. Mapping QTLs controlling fruit quality in peach (Prunus persica (L.) Batsch). *Theoretical and Applied Genetics*. T. 98, n° 1, p. 18-31.

DOKOUMETZIDIS, Aristides et al., 2009. Proper lumping in systems biology models. *IET systems biology*. T. 3, n° 1, p. 40-51.

EFRON, B et al. Chapman & Hall ; London : 1993. *An introduction to the bootstrap.[Google Scholar]*.

EFRON, Bradley, 1982. *The jackknife, the bootstrap and other resampling plans*. SIAM.

EOM, Joon-Seob et al., 2011. Impaired function of the tonoplast-localized sucrose transporter in rice, OsSUT2, limits the transport of vacuolar reserve sucrose and affects plant growth. *Plant Physiology*. T. 157, n° 1, p. 109-119.

ETIENNE, C et al., 2002. Candidate genes and QTLs for sugar and organic acid content in peach [Prunus persica (L.) Batsch]. *Theoretical and Applied Genetics*. T. 105, n° 1, p. 145-159.

FARNIR, Frédéric et al., 2002. Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees : revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. *Genetics*. T. 161, n° 1, p. 275-287.

FLOUDAS, Christodoulos A et al., 2013. *Handbook of test problems in local and global optimization*. Springer Science & Business Media.

FOURNIER, Christian et al., 1999. ADEL-maize : an L-system based model for the integration of growth processes from the organ to the canopy. Application to regulation of morphogenesis by light availability. *Agronomie*. T. 19, n° 3-4, p. 313-327.

FRIDMAN, Eyal et al., 2000. A recombination hotspot delimits a wild-species quantitative trait locus for tomato sugar content to 484 bp within an invertase gene. *Proceedings of the National Academy of Sciences*. T. 97, n° 9, p. 4718-4723.

FU, Guifang et al., 2011a. A dynamic model for functional mapping of biological rhythms. *Journal of biological dynamics*. T. 5, n° 1, p. 84-101.

FU, Guifang et al., 2011b. A mathematical framework for functional mapping of complex phenotypes using delay differential equations. *Journal of theoretical biology*. T. 289, p. 206-216.

GARCÍA, C López et al., 2010. In-depth modeling of gas oil hydrotreating : From feedstock reconstruction to reactor stability analysis. *Catalysis Today*. T. 150, n° 3-4, p. 279-299.

GÉNARD, Michel et al., 1991. Variabilité de la croissance et de la qualité chez la pêche (Prunus persica L Batsch) et liaison entre croissance et qualité.

GÉNARD, Michel et al., 1996. Modeling the peach sugar contents in relation to fruit growth. *Journal of the American Society for Horticultural Science*. T. 121, n° 6, p. 1122-1131.

GÉNARD, Michel et al., 2010. Virtual profiling : a new way to analyse phenotypes. *The Plant Journal*. T. 62, n° 2, p. 344-355.

GÉNARD, Michel et al., 2014. Metabolic studies in plant organs : don't forget dilution by growth. *Frontiers in plant science*. T. 5, p. 85.

GOLDBERG, David E, 1989. Genetic algorithms in search. *Optimization, and MachineLearning*.

GORBAN, Alexander N et al., 2006. *Model reduction and coarse-graining approaches for multiscale phenomena*. Springer.

GRECHI, Isabelle et al., 2012. Designing integrated management scenarios using simulation-based and multi-objective optimization : Application to the peach tree–Myzus persicae aphid system. *Ecological Modelling*. T. 246, p. 47-59.

GUTENKUNST, Ryan N et al., 2007. Universally sloppy parameter sensitivities in systems biology models. *PLoS computational biology*. T. 3, n° 10, p. 1871-78. ISSN 1553-7358. Disp. à l'adr. DOI : 10.1371/journal.pcbi.0030189.

HALEY, Chris S et al., 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*. T. 69, n° 4, p. 315.

HAMMER, Graeme et al., 2006. Models for navigating biological complexity in breeding improved crop plants. *Trends in plant science*. T. 11, n° 12, p. 587-593.

HASS, Helge et al., 2017. Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ systems biology and applications*. T. 3, n° 1, p. 1-15.

HASTINGS, W Keith, 1970. Monte Carlo sampling methods using Markov chains and their applications.

HAYASHI, Takeshi et al., 2010. EM algorithm for Bayesian estimation of genomic breeding values. *BMC genetics*. T. 11, n° 1, p. 3.

HEIJNEN, Joseph J, 2005. Approximative kinetic formats used in metabolic network modeling. *Biotechnology and bioengineering*. T. 91, n° 5, p. 534-545.

HEINRICH, Reinhart et al., 1996. *The regulation of cellular systems*. Chapman et Hall.

HENRI, Victor, 1903. *Lois générales de l'action des diastases*. Hermann.

HERMANN, Robert et al., 1977. Nonlinear controllability and observability. *IEEE Transactions on automatic control*. T. 22, n° 5, p. 728-740.

HOERL, Arthur E et al., 1970. Ridge regression : Biased estimation for nonorthogonal problems. *Technometrics*. T. 12, n° 1, p. 55-67.

HOLME, Petter et al., 2003. Subnetwork hierarchies of biochemical pathways. *Bioinformatics*. T. 19, n° 4, p. 532-538. ISSN 1367-4803.

HOOGENBOOM, Gerrit et al., 2004. From genome to crop : integration through simulation modeling. *Field Crops Research*. T. 90, n° 1, p. 145 -163. ISSN 0378-4290. Linking Functional Genomics with Physiology for Global Change Research.

HOSEA, ME et al., 1996. Analysis and implementation of TR-BDF2. *Applied Numerical Mathematics*. T. 20, n° 1-2, p. 21-37.

HUANG, Anhui et al., 2014. Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice. *PloS one*. T. 9, n° 1, e87330.

HUE, Louis et al., 1981. The role of futile cycles in the regulation of carbohydrate metabolism in the liver. *Adv Enzymol Relat Areas Mol Biol*. T. 52, p. 247-331.

IOOSS, Bertrand et al., 2015. A review on global sensitivity analysis methods. In : *Uncertainty management in simulation-optimization of complex systems*. Springer, p. 101-122.

JAQAMAN, Khuloud et al., 2006. Linking data to models : data regression. *Nature Reviews Molecular Cell Biology*. T. 7, n° 11, p. 813-819.

JOHANSEN, Adam M et al., 2010. Monte carlo methods. *Lecture notes*. T. 200.

JOHNSON, Kenneth A et al., 2011. The original Michaelis constant : translation of the 1913 Michaelis–Menten paper. *Biochemistry*. T. 50, n° 39, p. 8264-8269.

KADER, Adel A, 2008. Flavor quality of fruits and vegetables. *Journal of the Science of Food and Agriculture*. T. 88, n° 11, p. 1863-1868.

KANAYAMA, Yoshinori et al., 1997. Divergent fructokinase genes are differentially expressed in tomato. *Plant Physiology*. T. 113, n° 4, p. 1379-1384.

KANSO, Hussein et al., 2020. Reducing a model of sugar metabolism in peach to catch different patterns among genotypes. *Mathematical Biosciences*. T. 321, p. 108321.

KAO, Chen-Hung et al., 1999. Multiple interval mapping for quantitative trait loci. *Genetics*. T. 152, n° 3, p. 1203-1216.

KATZ, Joseph et al., 1978. Futile cycling in glucose metabolism. *Trends in Biochemical Sciences*. T. 3, n° 3, p. 171-174.

KENNEDY, James et al., 1995. Particle swarm optimization. In : *Proceedings of ICNN'95-International Conference on Neural Networks*. T. 4, p. 1942-1948.

KNOWLES, Joshua D et al., 2001. Reducing local optima in single-objective problems by multi-objectivization. In : *International conference on evolutionary multi-criterion optimization*, p. 269-283.

KUHN, Estelle et al., 2004. Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM : Probability and Statistics*. T. 8, p. 115-131.

KUHN, Estelle et al., 2005. Maximum likelihood estimation in nonlinear mixed effects models. *Computational statistics & data analysis*. T. 49, n° 4, p. 1020-1038.

KWAK, Il-Youp et al., 2014. A simple regression-based method to map quantitative trait loci underlying function-valued phenotypes. *Genetics*. T. 197, n° 4, p. 1409-1416.

LABATE, Joanne A et al., 2007. Tomato. In : *Vegetables*. Springer, p. 1-125.

LAMBONI, Matieyendou et al., 2009. Multivariate global sensitivity analysis for dynamic crop models. *Field Crops Research*. T. 113, n° 3, p. 312-320.

LAMBONI, Matieyendou et al., 2011. Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*. T. 96, n° 4, p. 450-459.

LANDER, Eric S et al., 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*. T. 121, n° 1, p. 185-199.

LAPERCHE, Anne et al., 2006. A simplified conceptual model of carbon/nitrogen functioning for QTL analysis of winter wheat adaptation to nitrogen deficiency. *Theoretical and Applied Genetics*. T. 113, n° 6, p. 1131-1146.

LEMAIRE, François et al., 2012. Mabsys : Modeling and analysis of biological systems. In : *Algebraic and Numeric Biology*. Springer, p. 57-75.

LERCETEAU-KÖHLER, E et al., 2012. Genetic dissection of fruit quality traits in the octoploid cultivated strawberry highlights the role of homoeo-QTL in their control. *Theoretical and Applied Genetics*. T. 124, n° 6, p. 1059-1077.

LETORT, Veronique et al., 2008. Parametric identification of a functional–structural tree growth model and application to beech trees (Fagus sylvatica). *Functional plant biology*. T. 35, n° 10, p. 951-963.

LI, Genyuan et al., 1989. A general analysis of exact lumping in chemical kinetics. *Chemical engineering science*. T. 44, n° 6, p. 1413-1430.

LI, Pu et al., 2013. Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC systems biology*. T. 7, n° 1, p. 91.

LI, Zitong et al., 2012. Estimation of quantitative trait locus effects with epistasis by variational Bayes algorithms. *Genetics*. T. 190, n° 1, p. 231-249.

LIEBERMEISTER, Wolfram et al., 2005. Biochemical network models simplified by balanced truncation. *The FEBS Journal*. T. 272, n° 16, p. 4034-4043.

LÓPEZ ZAZUETA, Claudia et al., 2018. Analytical Reduction of Nonlinear Metabolic Networks Accounting for Dynamics in Enzymatic Reactions. *Complexity*. T. 2018.

LUAN, Tu et al., 2009. The accuracy of genomic selection in Norwegian red cattle assessed by cross-validation. *Genetics*. T. 183, n° 3, p. 1119-1126.

MA, Chang-Xing et al., 2002. Functional mapping of quantitative trait loci underlying the character process : a theoretical framework. *Genetics*. T. 161, n° 4, p. 1751-1762.

MACKAY, T et al., 2009. The genetics of quantitative traits : challenges and prospects. *Nat. Rev. Genet.* T. 10, n° 8, p. 565-77.

MAIRET, Francis et al., 2019. Twelve quick tips for designing sound dynamical models for bioprocesses. *PLoS Comput Biol*. T. in press, p. 1-15.

MAIWALD, Tim et al., 2016. Driving the model to its limit : profile likelihood based model reduction. *PloS one*. T. 11, n° 9, e0162366.

MARTRE, Pierre et al., 2015. Chapter 14 - Model-assisted phenotyping and ideotype design. In : SADRAS, Victor O. et al. (éd.). *Crop Physiology (Second Edition)*. Second Edition. San Diego : Academic Press, p. 349 -373.

MATHIEU, Amélie et al., 2018. A new methodology based on sensitivity analysis to simplify the recalibration of functional–structural plant models in new conditions. *Annals of botany*. T. 122, n° 3, p. 397-408.

MAYER, DG et al., 1993. Statistical validation. *Ecological modelling*. T. 68, n° 1-2, p. 21-32.

MESSINA, C. D. et al., 2006. A Gene-Based Model to Simulate Soybean Development and Yield Responses to Environment. *Crop Science*. T. 46, n° 1, p. 456. ISSN 1435-0653. Disp. à l'adr. DOI : 10.2135/cropsci2005.04-0372.

MONTESINOS-LÓPEZ, Abelardo et al., 2018. Multi-environment genomic prediction of plant traits using deep learners with dense architecture. *G3 : Genes, Genomes, Genetics*. T. 8, n° 12, p. 3813-3828.

MORIGUCHI, Takaya et al., 1991. Properties of acid invertase purified from peach fruits. *Phytochemistry*. T. 30, n° 1, p. 95-97.

MORRIS, Max D, 1991. Factorial sampling plans for preliminary computational experiments. *Technometrics*. T. 33, n° 2, p. 161-174.

MORTON, Newton E, 1955. Sequential tests for the detection of linkage. *American journal of human genetics*. T. 7, n° 3, p. 277.

MUTSHINDA, Crispin M et al., 2010. Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*. T. 186, n° 3, p. 1067-1075.

NAGANO, Atsushi J et al., 2012. Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. *Cell*. T. 151, n° 6, p. 1358-1369.

NÄGELE, Thomas et al., 2010. Mathematical modeling of the central carbohydrate metabolism in Arabidopsis reveals a substantial regulatory influence of vacuolar invertase on whole plant carbon metabolism. *Plant physiology*. T. 153, n° 1, p. 260-272.

NÄGELE, Thomas et al., 2014. Mathematical modeling reveals that metabolic feedback regulation of SnRK1 and hexokinase is sufficient to control sugar homeostasis from energy depletion to full recovery. *Frontiers in plant science*. T. 5, p. 365.

NAKAGAWA, H et al., 2005. Flowering response of rice to photoperiod and temperature : a QTL analysis using a phenological model. *Theoretical and Applied Genetics*. T. 110, n° 4, p. 778-786.

NIKEREL, I. Emrah et al., 2009. Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics. *Metab. Eng.* T. 11, n° 1, p. 20-30. ISSN 1096-7176.

OKINO, Miles S et al., 1998. Simplification of mathematical models of chemical reaction systems. *Chemical reviews*. T. 98, n° 2, p. 391-408.

ONOGI, Akio, 2020. Connecting mathematical models to genomes : joint estimation of model parameters and genome-wide marker effects on these parameters. *Bioinformatics*. T. 36, n° 10, p. 3169-3176.

ONOGI, Akio et al., 2016a. Toward integration of genomic selection with crop modelling : the development of an integrated approach to predicting rice heading dates. *Theoretical and Applied Genetics*. T. 129, n° 4, p. 805-817.

ONOGI, Akio et al., 2016b. VIGoR : variational Bayesian inference for genome-wide regression. *Journal of Open Research Software*. T. 4, n° 1.

OURA, Yasushi et al., 2000. Purification and characterization of a NAD+-dependent sorbitol dehydrogenase from Japanese pear fruit. *Phytochemistry*. T. 54, n° 6, p. 567-572.

PARK, Trevor et al., 2008. The bayesian lasso. *Journal of the American Statistical Association*. T. 103, n° 482, p. 681-686.

PETREIKOV, Marina et al., 2001. Characterization of native and yeast-expressed tomato fruit fructokinase enzymes. *Phytochemistry*. T. 58, n° 6, p. 841-847.

PIANOSI, Francesca et al., 2016. Sensitivity analysis of environmental models : A systematic review with practical workflow. *Environmental Modelling & Software*. T. 79, p. 214-232.

POHJANPALO, Hannu, 1978. System identifiability based on the power series expansion of the solution. *Mathematical biosciences*. T. 41, n° 1-2, p. 21-33.

PRUDENT, Marion et al., 2011. Combining Ecophysiological Modelling and Quantitative Trait Locus Analysis to Identify key Elementary Processes Underlying Tomato Fruit Sugar Concentration. *J Exp Bot*. T. 62, p. 907-919.

QIAN, Hong et al., 2006. Metabolic futile cycles and their functions : a systems analysis of energy and control. *IEE Proceedings-Systems Biology*. T. 153, n° 4, p. 192-200.

QUILOT, B et al., 2004a. Analysis of genotypic variation in fruit flesh total sugar content via an ecophysiological model applied to peach. *Theoretical and Applied Genetics*. T. 109, n° 2, p. 440-449.

QUILOT, B et al., 2004b. QTL analysis of quality traits in an advanced backcross between Prunus persica cultivars and the wild relative species P. davidiana. *Theoretical and Applied Genetics*. T. 109, n° 4, p. 884-897.

QUILOT, B et al., 2005. Analysing the genetic control of peach fruit quality through an ecophysiological model combined with a QTL approach. *Journal of Experimental Botany*. T. 56, n° 422, p. 3083-3092.

QUILOT-TURION, Bénédicte et al., 2012. Optimization of parameters of the 'Virtual Fruit'model to design peach genotype for sustainable production systems. *European Journal of Agronomy*. T. 42, p. 34-48.

QUILOT-TURION, Bénédicte et al., 2016. Optimization of allelic combinations controlling parameters of a peach quality model. *Frontiers in plant science*. T. 7, p. 1873.

RAUE, Andreas et al., 2009. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*. T. 25, n° 15, p. 1923-1929.

RAUE, Andreas et al., 2014. Comparison of approaches for parameter identifiability analysis of biological systems. *Bioinformatics*. T. 30, n° 10, p. 1440-1448.

REYMOND, Matthieu et al., 2003. Combining quantitative trait loci analysis and an ecophysiological model to analyze the genetic variability of the responses of maize leaf growth to temperature and water deficit. *Plant Physiology*. T. 131, n° 2, p. 664-675.

ROBBINS, Herbert et al., 1951. A stochastic approximation method. *The annals of mathematical statistics*, p. 400-407.

ROBERT, Christian et al., 2013. *Monte Carlo statistical methods*. Springer Science & Business Media.

ROHWER, Johann M, 2012. Kinetic modelling of plant metabolic pathways. *Journal of Experimental Botany*. T. 63, n° 6, p. 2275-2292.

SAEZ-RODRIGUEZ, Julio et al., 2008. Automatic decomposition of kinetic models of signaling networks minimizing the retroactivity among modules. *Bioinformatics*. T. 24, n° 16, p. i213-i219.

SALTELLI, Andrea et al., 2004. *Sensitivity analysis in practice : a guide to assessing scientific models*. Wiley Online Library.

SALTELLI, Andrea et al., 2008. *Global sensitivity analysis : the primer*. John Wiley & Sons.

SAMOILOV, Michael et al., 2005. Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proceedings of the National Academy of Sciences*. T. 102, n° 7, p. 2310-2315.

SCHAUER, M et al., 1983. Quasi-steady-state approximation in the mathematical modeling of biochemical reaction networks. *Mathematical biosciences*. T. 65, n° 2, p. 155-170.

SCHMIDT, Henning et al., 2008. Complexity reduction of biochemical rate expressions. *Bioinformatics*. T. 24, n° 6, p. 848-854.

SCHÜRER, Rudolf et al., 2006. MinT : A Database for Optimal Net Parameters. In : NIEDERREITER, Harald et al. (éd.). *Monte Carlo and Quasi-Monte Carlo Methods 2004*. Berlin, Heidelberg : Springer Berlin Heidelberg, p. 457-469.

SEMENOV, Mikhail A et al., 2013. Designing high-yielding wheat ideotypes for a changing climate. *Food and Energy Security*. T. 2, n° 3, p. 185-196.

SHIRATAKE, Katsuhiro et al., 1997. Characterization of hexose transporter for facilitated diffusion of the tonoplast vesicles from pear fruit. *Plant and cell physiology*. T. 38, n° 8, p. 910-916.

SILLANPÄÄ, MJ et al., 2012. Simultaneous estimation of multiple quantitative trait loci and growth curve parameters through hierarchical Bayesian modeling. *Heredity*. T. 108, n° 2, p. 134-146.

SIVANANDAM, SN et al., 2008. Genetic algorithm optimization problems. In : *Introduction to genetic algorithms*. Springer, p. 165-209.

SLAVIN, Joanne L et al., 2012. Health benefits of fruits and vegetables. *Advances in nutrition*. T. 3, n° 4, p. 506-516.

SNOWDEN, Thomas J et al., 2017. Methods of model reduction for large-scale biological systems : a survey of current methods and trends. *Bulletin of mathematical biology*. T. 79, n° 7, p. 1449-1486.

SOBOL, Ilya M, 1993. Sensitivity analysis for non-linear mathematical models. *Mathematical modelling and computational experiment*. T. 1, p. 407-414.

SOLLER, M et al., 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theoretical and applied genetics*. T. 47, n° 1, p. 35-39.

SOSINSKI, B et al., 1997. Use of AFLP and RFLP markers to create a combined linkage map in peach [Prunus persica (L.) Batsch] for use in marker assisted selection. In : *IV International Peach Symposium 465*, p. 61-68.

SOUNDHARAJAN, Bs et al., 2009. Deficit irrigation management for rice using crop growth simulation model in an optimization framework. *Paddy and Water Environment*. T. 7, n° 2, p. 135-149.

STEPHANI, Hans, 1989. *Differential equations : their solution using symmetries*. Cambridge University Press.

STEVENS, M Allen et al., 1977. Genotypic variation for flavor and composition in fresh market tomatoes. *Journal American Society for Horticultural Science*.

STORN, Rainer et al., 1997. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*. T. 11, n° 4, p. 341-359.

SUN, Xiaodian et al., 2016. Model reduction and parameter estimation of non-linear dynamical biochemical reaction networks. *IET Systems Biology*. T. 10, 10-16(6). ISSN 1751-8849.

SUNNÅKER, Mikael et al., 2010. Zooming of states and parameters using a lumping approach including back-translation. *BMC systems biology*. T. 4, n° 1, p. 28.

SUNNÅKER, Mikael et al., 2011. A method for zooming of nonlinear models of biochemical systems. *BMC systems biology*. T. 5, n° 1, p. 140.

SUROVTSOVA, Irina et al., 2006. Focusing on dynamic dimension reduction for biochemical reaction systems. *Understanding Exploiting Syst Biol Biomed Bioprocesses*. T. 31, p. 31-46.

TARDIEU, Francois, 2003. Virtual plants : modelling as a tool for the genomics of tolerance to water deficit. *Trends in plant Science*. T. 8, n° 1, p. 9-14.

TECHNOW, Frank et al., 2015. Integrating crop growth models with whole genome prediction through approximate Bayesian computation. *PloS one*. T. 10, n° 6, e0130855.

TERWILLIGER, Joseph D, 1995. A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. *American journal of human genetics*. T. 56, n° 3, p. 777.

TIBSHIRANI, Robert, 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society : Series B (Methodological)*. T. 58, n° 1, p. 267-288.

TOMLIN, Alison S et al., 1997. The effect of lumping and expanding on kinetic differential equations. *SIAM Journal on Applied Mathematics*. T. 57, n° 6, p. 1531-1556.

TURÁNYI, Tamás, 1990. Sensitivity analysis of complex kinetic systems. Tools and applications. *Journal of mathematical chemistry*. T. 5, n° 3, p. 203-248.

TZIKAS, Dimitris G et al., 2008. The variational approximation for Bayesian inference. *IEEE Signal Processing Magazine*. T. 25, n° 6, p. 131-146.

UYS, Lafras et al., 2007. Kinetic model of sucrose accumulation in maturing sugarcane culm tissue. *Phytochemistry*. T. 68, n° 16-18, p. 2375-2392.

VALLABHAJOSYULA, Ravishankar R et al., 2006. Complexity reduction of biochemical networks. In : *Proceedings of the 2006 Winter Simulation Conference*, p. 1690-1697.

VAN OIJEN, M et al., 2016. Toward a Bayesian procedure for using process-based models in plant breeding, with application to ideotype design. *Euphytica*. T. 207, n° 3, p. 627-643.

VANGDAL, Eivind, 1985. Quality criteria for fruit for fresh consumption. *Acta Agriculturae Scandinavica*. T. 35, n° 1, p. 41-47.

VANUYTRECHT, Eline et al., 2014. Global sensitivity analysis of yield output from the water productivity model. *Environmental Modelling Software*. T. 51, p. 323 -332. ISSN 1364-8152.

VENTER, Gerhard, 2010. Review of optimization techniques. *Encyclopedia of aerospace engineering*.

VERDE, Ignazio et al., 2013. The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature genetics*. T. 45, n° 5, p. 487-494.

VIENNE, Dominique de, 1998. *Les marqueurs moléculaires en génétique et biotechnologies végétales*. Quae.

VILLAVERDE, Alejandro F et al., 2016. Structural identifiability of dynamic systems biology models. *PLoS computational biology*. T. 12, n° 10, e1005153.

VINCENTE, Ariel R et al., 2014. Nutritional quality of fruits and vegetables. In : *Postharvest handling*. Elsevier, p. 69-122.

VORSTER, Darren J et al., 1998. Partial purification and characterisation of sugarcane neutral invertase. *Phytochemistry*. T. 49, n° 3, p. 651-655.

W. PATRICK, John et al., 2013. Metabolic engineering of sugars and simple sugar derivatives in plants. *Plant biotechnology journal*. T. 11, n° 2, p. 142-156.

WALTER, Eric et al., 1997. Identification of parametric models. *Communications and control engineering*. T. 8.

WANG, Feng-Sheng et al., 2007. Kinetic modeling using S-systems and lin-log approaches. *Biochem. Eng. J.* T. 33, n° 3, p. 238-247.

WANG, Zuoheng et al., 2013. Stochastic modeling of systems mapping in pharmacogenomics. *Advanced drug delivery reviews*. T. 65, n° 7, p. 912-917.

WEI, James et al., 1969. Lumping Analysis in Monomolecular Reaction Systems. Analysis of the Exactly Lumpable System. *Industrial & Engineering Chemistry Fundamentals*. T. 8, n° 1, p. 114-123.

WEI, Kun et al., 2018. An ecophysiologically based mapping model identifies a major pleiotropic QTL for leaf growth trajectories of Phaseolus vulgaris. *The Plant Journal*. T. 95, n° 5, p. 775-784.

WHITE, Jeffrey W. et al., 2003. Gene-Based Approaches to Crop Simulation. *Agron. J.* T. 95, n° 1, p. 52-64.

WU, BH et al., 2012. Application of a SUGAR model to analyse sugar accumulation in peach cultivars that differ in glucose–fructose ratio. *Journal of agricultural science*. T. 150, p. 53-63.

WU, Rongling et al., 2006. Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nature Reviews Genetics*. T. 7, n° 3, p. 229-237.

YI, Nengjun et al., 2008. Bayesian LASSO for quantitative trait loci mapping. *Genetics*. T. 179, n° 2, p. 1045-1055.

YIN, Xinyou et al., 2000. Coupling estimated effects of QTLs for physiological traits to a crop growth model : predicting yield variation among recombinant inbred lines in barley. *Heredity*. T. 85, n° 6, p. 539-549.

YIN, Xinyou et al., 2005. QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley. *J Exp Bot*. T. 56, n° 413, p. 967-976. ISSN 0022-0957. Disp. à l'adr. DOI : 10.1093/jxb/eri090.

YIN, Xinyou et al., 2016. Modelling QTL-Trait-Crop Relationships : Past Experiences and Future Prospects. In : *Crop Systems Biology*. Cham : Springer International Publishing, p. 193-218. Disp. à l'adr. DOI : 10.1007/978-3-319-20562-5_9.

ZHANG, Chunhua et al., 2013. Cloning and expression of genes related to the sucrose-metabolizing enzymes and carbohydrate changes in peach. *Acta physiologiae plantarum*. T. 35, n° 2, p. 589-602.

ZHOU, Yong et al., 2014. Overexpression of OsSWEET5 in rice causes growth retardation and precocious senescence. *PLoS One*. T. 9, n° 4, e94210.

ZOBELEY, Jürgen et al., 2005. A new time-dependent complexity reduction method for biochemical systems. In : *Transactions on Computational Systems Biology I*. Springer, p. 90-110.