Modelling differential phenotypic expression

Fred A. van Eeuwijk[1], Marcos Malosetti[1], Xinyou Yin[2], Paul C. Struik[2] and Piet Stam[1]

[1] Laboratory of Plant Breeding, Wageningen University, P.O. Box 386, 6700 AJ Wageningen, The Netherlands.
[2] Crop and Weed Ecology Group, Wageningen University, P.O. Box 430, 6700 AK Wageningen, The Netherlands
E-mail: fred.vaneeuwijk@wur.nl

Abstract

To study the performance of genotypes under different growing conditions, plant breeders evaluate their germplasm in multi-environment trials. These trials produce genotype by environment data. We present various statistical models for the analysis of such data that differ in the extent to which additional genetic, physiological and environmental information is incorporated into the model formulation. The simplest model in our exposition is the additive two-way analysis of variance model, without genotype by environment interaction and with parameters whose interpretation depends strongly on the set of included genotypes and environments. The most complicated model is a synthesis of a multiple quantitative trait locus model and an eco-physiological model to describe a collection of genotypic response curves. Between those extremes, we discuss linear-bilinear models, whose parameters can only indirectly be related to genetic and physiological information, and factorial regression models that allow direct incorporation of explicit genetic, physiological and environmental covariables on the levels of the genotypic and environmental factor. Factorial regression models are also very suitable for the modeling of QTL main effects and QTL by environment interaction. Our conclusion is that statistical and physiological models can fruitfully be combined for the study of genotype by environment interaction.

Media Summary

To study the performance of genotypes under different growing conditions, germplasm is evaluated in multi-environment trials. In this paper, we demonstrate that statistical and physiological models can fruitfully be combined for the study of genotype by environment interaction.

Key words

Crop growth model, Eco-physiological model, Factorial regression, Genotype by environment interaction, QTL by environment interaction, Response curves

Introduction

A major objective in many advanced plant breeding programs is to assess the suitability of individual crop genotypes for agricultural purposes across a range of agro-ecological conditions. To this purpose breeders perform so-called multi-environment trials. In a multi-environment trial, a set of genotypes is evaluated across a number of environments that hopefully represent the environmental range across which the genotypes should partially (specific adaptation) or wholly (wide adaptation) perform well. The performance of genotypes in multi-environment trials is analyzed by statistical models developed to describe and interpret genotype by environment data. The statistical analysis should provide estimates for parameters that indicate both how well genotypes perform on average across the environmental range and how well they perform in specific environmental conditions. Traditionally, the statistical parameters used by breeders to characterize genotypic responses across environments were largely devoid of physiological meaning. More recently, it has become popular to use statistical models whose parameters relate better to physiological knowledge and that permit varying degrees of integration between statistical and physiological approaches to description and prediction of genotypic responses across environments.

This paper discusses various classes of statistical models for the analysis of genotype by environment data. All the models can be interpreted in terms of response functions for individual genotypes to environmental variables. The differences between the models reside in the amount of genetic and physiological characterization of the genotypes, the amount of physical and meteorological characterization of the environments, and the complexity of the response curves. We intend to show that statistical and physiological models for the description of genotype by environment data can be reconciled and combined in a fruitful way.

The additive model in quantitative genetics and plant breeding

Within plant breeding, a tradition exists to describe phenotypic responses across environments in terms of statistical parameters that have well defined statistical properties, but that are hard to interpret in physiological terms. The dominant quantitative genetic paradigm in plant breeding dictates models for phenotypic expression to consist of sums of terms that are indexed by either genotypes, environments, or combinations of both. The simplest model for the description of phenotypic responses across environments, the additive model, contains only single indexed terms. For the expected phenotypic response for genotype i (i = 1,…,I) in environment j (j = 1,…,J), $?_{ij}$, we write $?_{ij} = ? + G_i + E_j$, where ? stands for the general mean, G_i is the genotypic main effect expressed as a deviation from the general mean, and E_j symbolizes the environmental main effect, again expressed as a deviation from the mean.

Although the statistical description of the additive model suggests already some complexity, the above model merely states that we might try to describe the phenotypic responses for a set of genotypes as a set of parallel straight lines, where the differences between the responses are given by the differences between the genotypic main effects. To illustrate this, we consider the increase in the mean response for a genotype i, when going from environment j to j*, where we assume that j* represents the better environment, or, $E_{j^*} > E_j : ?_{ij^*} - ?_{ij} = E_{j^*} - E_j$. It is obvious that all genotypes will show the same increase in phenotypic response, when going from the inferior environment j to the superior environment j*. When the environmental main effect is interpreted as an indicator of environmental quality, we might say that all genotypes exhibit the same sensitivity to the environment. To emphasize the parallel response character of the additive model, we can write $?_{ij} = G'_i + \beta'_i E_j$, where $G'_i = ? + G_i$, the predicted mean performance for genotype i across environments, and the slope β'_i is equal to 1 for all genotypes.

The most curious property of the additive model is that its parameters suggest a reference to genotypic and environmental entities outside the model, i.e., there appear to exist things or processes that might be called genetic (genotypic) in nature as well as environmental. However, the genotypes in the additive model are nothing more than the levels of a nominal variable, where the idea is that the major differences between the levels of that factor reside ultimately in differences in DNA composition. In the additive model, the environment is a collection of discrete sets of conditions under which the plants pertaining to particular genotypes have been grown. The parameters G_i and E_j are estimated by averaging over phenotypic observations, and at no point in this process does something evidently genetic or environmental enter the calculations. For balanced data (for example, all genotype by environment combinations were observed equally often without missing values), the estimate for the main effect of genotype i follows from the average across environments of the phenotypic observations indexed by i. Likewise, the estimate for the main effect of environment j follows from the average across genotypes of observations indexed by j. Thus, genotypic main effects depend on the collection of environments that were included in the experiments, while environmental main effects depend on the genotypes that were included. Suppose we evaluate yield for a set of genotypes that consists of two subsets: one subset of genotypes that are tolerant against a major stress factor and another subset of genotypes that are susceptible to the same stress factor. Genotypic main effects in the tolerant and susceptible subset will, other things being equal, be of similar magnitude as long as the particular environmental stress factor does not occur in the sample of environments included in the trials. In contrast, when in at least some of the environments the pertinent stress factor does occur, the susceptible genotypes will rank lower than the tolerant ones.

An equivalent argument can be constructed for the environments. Environments may differ in nutrient and water availability, but without genotypic variation in sensitivity to the quality of the environment, the better environments will not be recognizable for their higher yield, i.e., higher environmental main effects. Therefore, strictly speaking, neither the genotypic main effects nor the environmental main effects represent entities that exist outside of the collection of genotypes and environments that were included in the trials and the model for which they have been estimated. The main purpose of the additive model is to interpret phenotypic differences in terms of differences between the levels of the genotypic factor on the one hand and between levels of the environmental factor on the other hand for the included sets of genotypes and environments. Of course, the genotypes and environments in the trials may be chosen to be representative of some population of interest. For the environments, we then speak of the target population of environments. For the latter case, the environmental main effects are often assumed to follow a normal distribution. Whatever the statistical details, it may be clear that it will be difficult to encounter in an individual plant the physiological counterpart of its genotypic main effect.

Models for interaction using phenotypic characterizations of the environment

The additive model is an elementary model that is more important as a didactical tool to introduce statistical models for genotype by environment data than as a serious description of such data. The additive model provides a null model against which to test models that are more complicated with terms for genotype by environment interaction. Genotype by environment interaction occurs whenever genotypes react differently to environmental changes. So, whenever the difference in phenotypic performance between two environments j and j* varies between two genotypes i and i*, i.e, $?_{ij^*} - ?_{ij} \neq ?_{i^*j^*} - ?_{i^*j}$, the additive model will be inadequate and a more elaborate model should be formulated. Traditionally, the additive model is extended to a full interaction model with double indexed genotype by environment terms for each combination of genotype and environment: $?_{ij} = ? + G_i + E_j + (GE)_{ij}$. In the full interaction model there are as many independent parameters as genotype by environment combinations and from the point of view of parsimony little has been accomplished by fitting this model to the data. Predictions of phenotypic responses for environments that were not in the set of trial environments are impossible, as there will be no estimates for the particular $(GE)_{ij}$ terms. Compare this with the situations for which the additive model provides a good fit. In those cases rough predictions are possible as long as the quality of the new environment can be ranked as being in between two environments that were part of the multi-environment trial.

An alternative, more attractive extension of the additive model, which, like the additive model, describes phenotypic responses as straight lines, but allows for differential environmental sensitivity between genotypes, is the regression on the mean model, popularized by Finlay and Wilkinson (1963). The philosophy behind this model is that in the absence of explicit physical or meteorological characterizations of an environment, a good approximation to the general biological quality of the environment is given by the average phenotypic performance across the genotypes. The phenotypic responses of individual genotypes are then regressed on the average performance, and the genotype by environment interaction (GEI) expresses itself by differences in the slopes between the genotypes.

An elaborate way to write the regression on the mean model, that shows the relation with the full interaction analysis of variance model, is $?_{ij} = ? + G_i + E_j + \beta_i E_j$. The GEI is modeled as differential genotypic sensitivity, represented by the parameters β_i, to the environmental characterization E_j, with the average sensitivity being zero. A reformulation of the model makes evident the non parallel straight lines nature of the regression on the mean model $?_{ij} = ? + G_i + E_j ?+?\beta_i E_j = (? + G_i) + (1 +\beta_i)E_j =$ $G_i' + \beta_i' E_j$, where the average sensitivity now will be unity; $\overline{\beta_i'} = 1$. When all β_i are zero, or all β_i' are one, the regression on the mean model reduces to the additive model. Alternatively, the regression on the mean model will be equivalent to the full interaction model when $(GE)_{ij} = \beta_i E_j$ for all genotype by environment combinations.

The estimate for the sensitivity, or responsiveness, to the environment of individual genotypes depends on the average potential of the genotypes to change in relation to the environmental conditions. Therefore, the interpretation of the magnitudes of individual genotypic sensitivities should take into account the composition of the genotype and environment sets included in the multi-environment trial. For example, one susceptible genotype in a collection of otherwise tolerant genotypes evaluated under environmental conditions that include at least one instance of the pertinent stress, will have a far higher estimated environmental sensitivity than the same susceptible genotype evaluated within a predominantly susceptible set of genotypes. Like the additive model, the regression on the mean model can be used for prediction to the extent that new environments can be ranked with respect to the environments included in the trial set.

The regression on the mean model partitions the genotype by environment interaction term, $(GE)_{ij}$, in the full interaction model in a part due to regression on the environmental main effect (environmental index), $\beta_i E_j$, and a residual, $(GE)'_{ij}$. (This residual is usually interpreted as a random variable with zero mean, in that it does not appear in the expectation.) The statistical success of the regression on the mean model depends on the proportion of genotype by environment interaction that is described by the differential environmental sensitivity of the genotypes, or, equivalently, by the quality of the environmental effect as a reflection of the environmental forces that cause phenotypic differences between genotypes. The regression on the mean model provides only limited flexibility for describing GEI, because of its rather specific, one-dimensional incorporation of the environmental factors affecting the phenotypic responses. However, other models from the model class of which the regression on the mean model is a member, the class of linear-bilinear models (Gabriel, 1978, 1998; van Eeuwijk, 1995a; Denis and Gower, 1996; Crossa and Cornelius, 2002), allow considerably more flexible environmental characterizations of the environment. All these models describe GEI by differential genotypic sensitivities to environmental characterizations that are derived from the phenotypic data themselves.

Linear-bilinear models consist of sums of single indexed additive and multiplicative terms. In the regression on the mean model, using the regression formulation, $G'_i + \beta'_i E_j$, the linear part of the model is given by G'_i, while the bilinear part is equal to $\beta'_i E_j$. The term $\beta'_i E_j$ contains genotypic and environmental parameters that need to be estimated simultaneously. The name bilinear models stems from the observation that these models become standard linear models in the genotypic parameters upon fixation of the environmental parameters and vice versa. This property also forms the basis of a general estimating procedure for the parameters (Gabriel and Zamir, 1979, Gabriel, 1998; van Eeuwijk, 1995b).

In comparison with the regression on the mean model more flexible linear-bilinear models for the modeling of GEI can be constructed by the inclusion of additional bilinear terms. A popular example of a linear-bilinear model with a varying number of bilinear terms for the description of GEI is the additive main effects and multiplicative interaction effects model (Gollob, 1968; Mandel 1969; Gabriel 1978; Gauch, 1988). The model can be formulated as $?_{ij} = ? + G_i + E_j + \sum_{k=1}^{K} a_{ki} b_{kj}$, where a_{ki} and b_{kj} are genotypic and environmental parameters (scores) for the bilinear term k, and where K indicates the number of multiplicative terms necessary for an adequate description of the genotype by environment interaction. Following the same logic as for the regression on the mean model, the genotypic scores, a_{ki}, can be interpreted as sensitivities or responsivenesses, while the environmental scores, b_{kj}, are environmental characterizations. The environmental scores for the first bilinear term represent the best environmental characterization possible for the description of the genotype by environment interaction in terms of differences in genotypic sensitivity. The second bilinear term represents the second best environmental characterization, etcetera. The environmental characterizations in bilinear terms are acquired by minimization of a least squares criterion, and may not always have an immediate physiological interpretation. Still, regressing the environmental scores on explicit environmental measurements usually allows the genotype by environment interaction to be related to physiological processes (Vargas et al., 1999).

Models for interaction using explicit environmental characterizations

Bilinear models for interaction are very useful for a first round of exploratory analyses in which differences between genotypes are modeled by sensitivities to hypothetical environmental characterizations that describe a maximum amount of the genotype by environment interaction. Whether the results of analyses by bilinear models contain any physiological interest depends on the relation that the environmental main effects and scores bear with a description of the environment in terms of external, physical and meteorological variables. For example, suppose that it is concluded from the analysis of a particular data set that the regression on the mean model gives an adequate description of the genotype by environment interaction and that the environmental main effect is mainly driven by average daily temperature, T_i. We then think of E_j in the regression on the mean model as a function of T_j, $E_j = f(T_j)$, and write: $?_{ij} = G'_i + \beta'_i E_j = G'_i + \beta'_i f(T_j)$. The latter model would definitely be a lot closer to the kind of models physiologists are used to work with than the purely phenotypic regression on the mean model. In addition, the latter kind of model would allow the phenotypic responses to be non-linear, i.e., to become response curves as exponential, logistic, Gompertz, Gaussian, etcetera, as long as the curve parameters are genotype-independent.

The simplest way of replacing the environmental effect by a function of an explicit environmental variable, is by using the identity function for f(.). For example, describing the interaction as driven by temperature would lead us to $?_{ij} = G'_i + \beta'_i z_j$, with z_j the temperature in environment j. The extension to more than one environmental variable is straightforward. Suppose that the genotype by environment interaction is driven by both the average temperature, z_{1j}, and the amount of rainfall, z_{2j}, then the following model might be appropriate: $?_{ij} = G'_i + \beta'_{1i} z_{1j} + \beta'_{2i} z_{2j}$, where β'_{1i} and β'_{2i} are the sensitivities to temperature and rainfall, respectively.

Contrary to what physiologists would do, plant breeders customarily want to correct the phenotypic data for the environmental main effect, and thereby concentrate on that part of the phenotypic differences that is caused by genotype-related sources of variation (G_i and $(GE)_{ij}$). Plant breeders are not particularly interested in a (physiological) model for the trial mean, they especially want to understand the differences between genotypes. When we follow that convention and include the environmental main effect, the above model becomes $?_{ij} = ? + G_i + E_j ?+ \beta_{1i} z_{1j} + \beta_{2i} z_{2j}$. The resemblance of the latter regression-like model with a linear-bilinear model with two bilinear terms for the interaction, $?_{ij} = ? + G_i + E_j + a_{1i} b_{1j} + a_{2i} b_{2j}$, is evident. The environmental scores b_{1j} and b_{2j} are, theoretically, the best environmental covariables for explaining GEI, but for a physiological understanding of the GEI, these scores should be interpreted in terms of measured or simulated environmental characterizations. One way to do so would be by regressing the environmental scores on a set of environmental covariables. In exceptional cases, the environmental scores of the linear-bilinear model can be replaced by environmental covariables without loss of descriptive adequacy for the GEI. In such cases, a physiology-inspired description of the GEI will coincide with the best statistical description for the particular data.

Statistical models for phenotypic responses across environments that describe genotype by environment interaction by differential sensitivity to explicit environmental variables belong to the class of factorial regression models (Denis, 1988; van Eeuwijk et al., 1996). The name is derived from the inclusion of covariables on the levels of the classifying factors in analysis of variance models. The critical issue for factorial regression models is the choice of covariables. In former days, in the absence of explicit information about the environment, the regression on the mean model, or another linear-bilinear model, was an obligatory choice. As a continuous registration of the environment has come within reach of many plant breeding trials, the question nowadays has become how to summarize the most relevant features of the environment from the point of view of genotype by environment interaction. Exclusively statistical approaches as variable subset-selection procedures are not very satisfactory, because they result mostly in physiologically difficult-to-interpret models. The most promising way forward seems to be the use of physiological knowledge to delimit the vast amount of potentially useful sets of environmental covariables.

Examples of the use of factorial regression guided by physiological knowledge to analyze adaptation and genotype by environment interaction in barley can be found in Voltas et al. (1999a,b).

Instead of physical measurements of the environment, one could also use simulated characterizations of the environments in a multi-environment trial. Crop growth models can be used to integrate environmental information over the growing season, which may result in a characterization of the environments in terms of different stress classes (Chapman et al., 2000). This type of environmental characterization can then be introduced as a categorical variable in a factorial regression model. Of course, when a crop growth model produces a quantitative stress index, this index could also be included in a factorial regression to model GEI. Note that although environmental covariables enter the factorial regression models linearly, there is no restriction to linear responses as quadratic and high order terms can be included. Furthermore, response surfaces based on multiple environmental covariables are equally feasible, provided the data contain enough information for the estimation of all the parameters.

Models for interaction using explicit genotypic and environmental characterizations

The inclusion of covariables on factor levels in analysis of variance models for the description of GEI is not only useful for the environmental factor(s), but is equally recommendable for the genotypic factor(s). For example, a laboratory test may have been developed to assess the tolerance of a set of genotypes against a particular stress factor and one wants to include the results of such a test in an analysis of variance model for a multi-environment trial on yield to describe genotypic differences dependent on the environment. Assume the values of the laboratory tests are expressed by the genotypic covariable x, with values x_i, then we can incorporate this covariable in the two-way analysis of variance model as follows, $?_{ij} = ? + G_i + E_j ?+ x_i\rho_j$. The parameters ρ_j then relate to the severity of the particular stress in environment j.

Genotypic covariables can also be used for the description of differences in genotypic means across environments: $?_{ij} = ? + x_i\rho +$ G^*_i $+ E_j$, where G^*_i represents a residual genotypic main effect that should be smaller when the description by x_i is more successful. An interesting application of this type of factorial regression model is in the detection and localization of quantitative trait loci (QTLs). The regression based approaches to QTL mapping, as initiated by Haley and Knott (1992), can be seen as a form of factorial regression with genotypic covariables that are functions of marker genotypes and the type of QTL effect (additive, dominance, epistasis). Consider a co-dominant marker that can assume the genotypes MM, Mm, and mm. A genetic covariable, or genetic predictor, for estimating a possible QTL at the position of this marker can be constructed by giving the predictor the value 2 for genotypes of the MM type, 1 for Mm and 0 for mm. Equivalently, the values 1, 0, -1 may be preferable for estimating the additive QTL effect when a genetic predictor for QTL dominance is going to be included with value 1 for Mm and 0 for MM and mm. By constructing genetic predictors at all marker positions, a genome scan for QTLs by marker regression can be performed. Genetic predictors in between marker positions, necessary for simple interval mapping, can be constructed as functions of the probabilities of QTL genotypes given flanking markers. Lynch and Walsh (1998) provide an introduction to procedures for constructing genetic predictors, while Jiang and Zeng (1997) present a very general algorithm for all kinds of biparental segregating populations. Composite interval mapping requires the inclusion of so-called co-factors, markers that correct for QTLs elsewhere on the genome. These co-factors can be chosen to be the genetic predictors corresponding to the QTLs identified during a genome scan by simple interval mapping. In model form we write, $?_{ij} = ? +$ $\sum_{c \in C} x_c \rho_c$ $+ x_i\rho +$ G^*_i $+ E_j$, where the terms $x_c\rho_c$ correct for putative QTLs elsewhere on the genome, and C represents the full set of such putative QTLs, while $x_i\rho$ is the QTL under test.

In the framework of factorial regression, modeling of QTL by environment interaction is a natural extension of modeling main effect QTLs, i.e., QTLs that are supposed to have constant expression across environments. A model with a QTL main effect and QTL by environment interaction at the same location in the genome can be written as $?_{ij} = ? + x_i\rho +$ G^*_i $+ E_j + x_i\rho_j +$ $(GE)^*_{ij}$. The $(GE)_{ij}$ from the analysis of

variance model is partitioned in a part due to differential QTL expression, $x_i\rho_j$, and a residual, $(GE)^*_{ij}$, that is usually taken random and for that reason then disappears from the expression for the expectation. In the light of QTL by environment interaction, the parameter ρ_j adjusts the average QTL expression across environments, ρ, to a more appropriate level for the individual environment j. The QTL by environment interaction parameters, ρ_j, can themselves be regressed on an environmental covariable, z, in an attempt to link differential QTL expression directly to key environmental factors. The QTL by environment interaction term $x_i\rho_j$ is replaced by a regression term $x_i(\lambda z_j)$ and a residual term $x_i \rho^*_j$, that again disappears from the expectation when ρ^*_j is assumed to be random. The parameter λ is a proportionality constant that determines the extent to which a unit change in the environmental covariable z, influences the effect of a QTL allele substitution.

From a breeding and physiological point of view, the above model is an interesting option, because it allows the prediction of differential genotypic responses to environmental changes from marker information characterizing the genotypes and environmental covariables characterizing the environment. Van Eeuwijk et al. (2001, 2002) give an example of differential QTL expression in relation to the minimum temperature during flowering for yield in maize data from the CIMMYT program on drought stress. Malosetti et al. (2004) analyzed yield data from the North American Barley Genome Project with added environmental information. QTL by environment interaction at chromosome 2H was found to depend on the temperature range during heading. A QTL allele substitution increased/decreased yield with 0.112 ton ha^{-1} for every degree Celsius that the temperature range increased.

Models for GEI based on QTLs whose expression is a function of environmental covariables may help to solve the recurrent point of dispute on the extent to which input parameters for crop growth models are ‘genetic’, where the implicit assumption is that input parameters that exhibit GEI are not ‘genetic’. On the realization that the discussion point here is one of predictability based on genotypic information, it follows that GEI for input parameters creates no problems as long as it can be modeled as QTL expression in relation to environmental covariables, because all parameters will then still be single indexed.

Models for response curves

A drawback of the QTL models described in the previous paragraph may be that they are linear in the parameters, while most physiological and developmental processes behave essentially in a non-linear manner in relation to the environment. Although polynomial expansions can provide good approximations to those non-linear functions, an intrinsically non-linear approach will usually be preferable. Wu et al. (2002) formulated a two-step approach that acknowledges the non-linearity of response curves for physiological traits, but they still use linear QTL models for the parameters of those curves. First, they fit non-linear functions to growth data for each of the genotypes separately and then analyze the estimated parameter vectors jointly in a multivariate composite interval mapping procedure. A fully non-linear approach to physiological response curves is presented by Ma et al. (2002). Their methodology is based on mixture models and an EM algorithm for estimation. The paper contains an example for logistic growth curves in poplar. The authors claim that their intrinsically non-linear approach to the unraveling of the genetic basis of growth curves has higher power than alternative approaches. In Malosetti and van Eeuwijk (2004), the philosophy proclaimed by Ma et al. (2002) has been translated into the slightly less demanding non-linear mixed model framework. The process of senescence in potato was modeled by a logistic curve for individual diploid potato genotypes stemming from a biparental cross. The model for the expectation of the state of senescence for genotype i at time point j was $?_{ij} = A + \frac{C}{1 + e^{-b_i(z_j - m_i)}}$, with z_j the time from planting to observation. The lower asymptote, A, and the difference between lower and upper asymptote, C, were the same for all genotypes, while the inflection point, m_i, and the slope parameter at this point, b_i, were genotype specific. The inflection point is the time at which the process of senescence reaches the point half way between the upper and lower asymptote. The values for slope and inflection points were modeled on genetic predictors inside the non-linear mixed model, i.e., different

QTL alleles had different average slopes and inflection points, while the genotype-specific deviations from those averages were given a bivariate normal distribution. Various QTLs were detected for both slopes and inflection points. As the QTLs for slopes and inflection points were largely uncorrelated, speed and onset of the senescence process seemed amenable to independent genetic improvement.

The non-linear QTL models of Ma et al. (2002) and Malosetti and van Eeuwijk (2004) provide powerful methods for eco-physiologically inspired genetic models for differential phenotypic expression in relation to environmental variables and (developmental) time. However, it cannot be denied that these models require a considerable amount of statistical skill for successful application. Therefore, it is reassuring that the simpler two-step approach that first estimates the parameters in non-linear response curves for individual genotypes by standard non-linear regression methodology and next searches for the genetic basis of those estimated parameters by applying standard QTL mapping methods, before feeding the QTL based parameters back into the eco-physiological crop model, also produces satisfactory results. Yin et al. (2004) studied days to flowering in barley in this way. They modeled daily rate of progress towards flowering as a non-linear function of temperature and photoperiod and estimated four genotypic parameters from a photoperiod-controlled greenhouse experiment. The four genotypic input parameters of the eco-physiological model for days to flowering were subjected to a QTL analysis, with each physiological input parameter treated as a classical phenotypic response. From the QTL models fitted to the estimated physiological parameters, genotypic predictions were calculated and fed back into the non-linear eco-physiological model in the hope that these QTL-based input parameter values would lead to better predictions of the time to flowering than the 'phenotypic' parameters from the non-linear regressions. In formula form, the expectation for days to flowering for genotype i in environment j was $?_{ij} = f(x_{1i}(\boldsymbol{\rho}_{1i}), x_{2i}(\boldsymbol{\rho}_{2i}), x_{3i}(\boldsymbol{\rho}_{3i}), x_{4i}(\boldsymbol{\rho}_{4i}), z_{1j}, z_{2j})$, with f(.) the non-linear function producing the days to flowering from four genotypic input traits, x_1 to x_4, and two environmental covariables, daily temperature and photoperiod, z_1 and z_2. The parameter vectors $\boldsymbol{\rho}_{1i}$ to $\boldsymbol{\rho}_{4i}$ represent the QTL basis of the four genotypic input traits. A promising result of this study was that days to flowering in barley could indeed well be predicted from the QTL based eco-physiological model, so that the combination of marker profile and environmental characterization (daily temperature and photoperiod) sufficed for prediction of days to flowering for new genotypes in new environments.

Reymond et al. (2003) used a similar combination of eco-physiological modeling and QTL mapping for the prediction of GEI for leaf elongation rate in maize. The QTL analysis was performed on parameters of a linear model for predicting the leaf elongation rate as affected by meristem temperature, water vapor pressure difference, and soil water status. Combined QTL- and eco-physiological model successfully predicted leaf elongation rates for environments characterized by different climatic scenarios.

Concluding remarks

The success of the approximate two-step approaches to combined QTL-eco-physiological modeling of Reymond et al. (2003) and Yin et al. (2004), and the one-step approaches discussed by Malosetti and van Eeuwijk (2004), Malosetti et al. (2004), and Ma et al. (2002), show that the methodology for the prediction of complex physiological responses in relation to genetic and environmental information has become sufficiently reliable to try its practical implementation in real life breeding programs.

Two points of concern are the following. Firstly, so far, the examples concerned simple, closed form, physiological processes. Should not more complicated physiological responses be the object of study? The answer comes from Bidinger et al. (1996) in that such simple expressions can provide already substantial insight into the physiological processes underlying GEI.

Another point of concern may be that the present examples of combined QTL-eco-physiological modeling were all done on classical segregating populations from biparental crosses. The question arises whether the results of such studies can be extrapolated to other crosses, i.e., other genetic backgrounds. An alternative to classical QTL studies with offspring from biparental crosses are association studies that look at marker-trait associations in collections of selected genotypes. A strong advantage of association studies is that they can be done on germplasm that represents a far wider genetic range than biparental offspring populations; a disadvantage is that marker-trait associations do not necessarily follow from

genetic linkage between marker and QTL. However, there are various ways to cope with that disadvantage. The work by Kraakman et al. (2005) illustrates the potential of association mapping for QTL-eco-physiological modeling. Danish variety trial yield data, spanning the period 1993-2000, on 146 modern two-row spring barley cultivars, representing the current commercial germplasm in Europe, was used to estimate mean performance, adaptability (slopes of the regression on the mean model), and stability (variance around the regression on the mean line). The cultivars were genotyped with 236 AFLP-markers, of which 123 were identified on an integrated map. Regression of the traits on individual marker data disclosed marker-trait associations for mean yield and yield stability. Many of the associated markers were located in regions where earlier QTLs were found for yield and yield components. To study the oligogenic base of the traits, multiple linear regression of the traits on markers was carried out using stepwise selection. By this procedure, 18 to 20 markers were selected to account for 40 to 58% of the variation in the studied complex traits. It was concluded that association mapping approaches constitute a viable alternative to classical QTL approaches, especially for complex traits with costly measurements. As statistical models for association mapping are very comparable to the models for classical QTL studies, the way forward to the integration of QTL-modeling and eco-physiological modeling would seem to be the application of one-step QTL-eco-physiological models in an association mapping context, using selections of genotypes that are known to exhibit interesting physiological contrasts on a phenotypic level.

References

Bidinger, F.R., Hammer, G.L., Muchow, R.C., 1996. The physiological basis of genotype by environment interaction in crop adaptation. In: Plant adaptation and crop improvement. Eds. Cooper, M., Hammer, G.L, 636 pp., CAB International, Oxon, UK.

Chapman, S.C., Cooper, M., Hammer, G.L., Butler, D., 2000. Genotype by environment interactions affecting grain sorghum. II. Frequencies of different seasonal patterns of drought stress are related to location effects on hybrid yields. Austr J Agric Res 51: 209-221.

Crossa, J., Cornelius, P., 2002. Linear–bilinear models for the analysis of genotype–environment interaction. In: Kang, M.S. (Ed.), Quantitative genetics, genomics and plant breeding, pp. 305-322. CAB International, Wallingford.

Denis, J.-B., 1988. Two-way analysis using covariates. Statistics 19: 123-132.

Denis, J.-B., Gower, J.C., 1996. Asymptotic confidence regions for biadditive models: interpreting genotype-environment interactions. Applied Statistics 45: 479-492.

Finlay, K. W., Wilkinson, G.N., 1963. The analysis of adaptation in a plant breeding programme. Australian Journal of Agricultural Research 14: 742-754.

Gabriel, K.R., 1978. Least squares approximation of matrices by additive and multiplicative models. Journal of the Royal Statistical Society, Series B 40:186-196.

Gabriel, K. R., 1998. Generalised bilinear regression. Biometrika 85: 689-700.

Gabriel, K.R., Zamir, S., 1979. Lower rank approximations of matrices by least squares with any choice of weights. Technometrics 21: 489-498.

Gauch, H. G., 1988. Model selection and validation for yield trials with interaction. Biometrics 44: 705-715.

Gollob, H.F., 1968. A statistical model which combines features of factor analysis and analysis of variance techniques. Psychometrika 33:73-115.

Haley, C.S., Knott, S.A., 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315-324.

Jiang, C., Zeng, Z.-B., 1997. Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica 101: 47-58.

Kraakman, A.T.W., Niks, R.E., van den Berg, P. M. M. M., Stam, P., van Eeuwijk, F.A., 2005. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics: accepted for publication.

Lynch, M., Walsh, J.B., 1998. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, Massachusetts.

Ma, C.X., Casella, G., Wu, R., 2002. Functional Mapping of Quantitative Trait Loci Underlying the Character Process: A Theoretical Framework. Genetics 161: 1751-1762.

Malosetti, M., van Eeuwijk, F.A., 2004. QTL methodology for response curves on the basis of non-linear mixed models, with an illustration to senescence in potato. Submitted.

Malosetti, M., Voltas, J., Romagosa, I., Ullrich, SE, van Eeuwijk, F.A., 2004. Mixed models including environmental variables for studying QTL by environment interaction. Euphytica: in press.

Mandel, J., 1969. The partitioning of interaction in analysis of variance. J. Res. NBS., 73B, 309-328.

Reymond, M., Muller, B., Leonardi, A., Charcosset, A., Tardieu, F. 2003. Combining QTL analysis and an ecophysiological model to analyse the genetic variability of the responses of maize leaf growth to temperature and water deficit. Plant Physiology 131:664-675.

van Eeuwijk, F.A., 1995a. Linear and bilinear models for the analysis of multi-environment trials: I. An inventory of models. Euphytica 84: 1-7.

van Eeuwijk, F.A., 1995b. Multiplicative interaction in generalized linear models. Biometrics 51: 1017-1032.

Van Eeuwijk, F.A., Crossa, J., Vargas, M., Ribaut, J.M. 2001. Variants of factorial regression for analysing QTL by environment interaction In: Eucarpia, Quantitative genetics and breeding methods: the way ahead. Gallais, A., Dillmann, C., Goldringer, I. (Eds.), INRA Editions, Versailles. Les colloques 96: 107-116

van Eeuwijk F.A., Crossa, J., Vargas, M., Ribaut, J.-M., 2002. Analysing QTL by environment interaction by factorial regression, with an application to the CIMMYT drought and low nitrogen stress programme in maize. In: Kang, M.S. (Ed.), Quantitative genetics, genomics and plant breeding, pp. 245-256. CAB International, Wallingford.

van Eeuwijk, F. A., Denis, J.-B., Kang, M.S., 1996. Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables. In: Kang, M.S., Gauch, H.G. (Eds.), Genotype-by-environment interaction, pp. 15-50. CRC Press, Boca Raton.

Vargas, M., Crossa, J., van Eeuwijk, F.A., Ram?rez, M.E., Sayre, K., 1999. Using AMMI, factorial regression, and partial least squares regression models for interpreting genotype x environment interaction. Crop Science 39: 955-967.

Voltas, J., van Eeuwijk, F.A., Sombrero, A., Lafarga, A., Igartua, E., Romagosa, I., 1999a. Integrating statistical and ecophysiological analysis of genotype by environment interaction for grain filling of barley in Mediterranean areas. I. Individual grain weight. Field Crops Research 62: 63-74.

Voltas, J., van Eeuwijk, F.A., Araus, J.L., Romagosa, I., 1999b. Integrating statistical and ecophysiological analysis of genotype by environment interaction for grain filling of barley in Mediterranean areas. II. Grain growth. Field Crops Research 62: 75-84.

Wu, W., Zhou, Y., Li, W., Mao, D., Chen, Q., 2002. Mapping of quantitative trait loci based on growth models. Theoretical and Applied Genetics 105:1043-1049.

Yin, X., Struik, P.C., van Eeuwijk, F.A., Stam, P., Tang, J., 2004. QTL analysis and QTL-based prediction of flowering phenology in recombinant inbred lines of barley.