Theoretical basis of identification of genotypes by their phenotypes in process of selection in segregating generations

Accepted 5th November, 2013

ABSTRACT

The formalized theory of the identification of genotypes by their phenotypes in modern breeding technologies is offer. In our first paper has presented the main principles of the simulation system for phenomenon "genotype-environment" interaction, and shows the possibility of the proposed models for a number of the solution of basic problems of modern genetics and breeding (Mikhailenko and Dragavtsev, 2012). Of all the problems solved with the help of such models, we selected the most important (Figure 1) assessment of the mechanisms of transgressions and the selection of parental pairs for a desirable result mating, 2) evaluation of the contributions (in productivity) of genetic-physiological systems of parental pairs, 3) prediction of transgressions of breeding traits in the F2 generation, and obtaining a population mating F2, 4) identification and selection of best genotypes by their phenotypes. These problems in their indissoluble unity represent stages in the same general problem - the strict control of genetic-breeding process. Development of the theory for solving this problem is one of the key perspectives of the modern genetics and plant breeding. Figure 1 shows the scheme of relations between tasks.

Key words: Genotypes, phenotypes, mathematical models, identification, breeding technologies, genetic-physiological systems, evaluation.

INTRODUCTION

As can be seen from Figure 1, the problem forms a closed loop control by the breeding process. Here, the initial step is the selection of the parent pairs for a desirable result mating. To optimize this selection used, the prediction of the results of parental mating pairs and actual results are analyzed on the crossing point for the identification of genotypes by their phenotypes. They are also used to correct the models predicting the results of crossing.

THE GENERAL SCHEME OF THE PROBLEM OF IDENTIFICATION OF GENOTYPES BY PHENOTYPES

In content, the selection of genotypes by phenotypes is quite complicated in terms of scientific classification of the information problem. Its aim is to find (or form) genotype, which includes the maximum number of positive shifts of given breeding character (BC). Therefore, the algorithm is based on the principle of background character (Dragavtsev, 1966) and the principle of differently-directional shifts of quantitative trait of individual genotype under the influence of genetic and environmental causes in the two-dimensional character's coordinates (Dyakov and Dragavtsev, 1994). The ideal background character has zero genetic variation, so that it "writes" on a variation that is only the ecological fluctuation of limiting factor (Dragavtsev, 1966). Respectively, the individual, which have a deviation of background character from an average population; this is a plus-modification, which got the better micro-ecological niche. At the same time, if the breeding
character of this individual is shifted in the positive direction from population's average, then this is a common modification and it does not make sense to select. If the other individual's background character is expressed at the level of average population value and breeding character shifted in the positive direction from average in the population, this recombination (or mutation), and it is necessary to select for productive breeding. The principle of differently-directional shifts propounds studying not characters, but behavior of genetic-physiological systems in the two-dimensional characters coordinates. For example in coordinates “mass of stem – mass of wheat-ear” effects of modifications and effects of genetic adaptively must “walk” along positive regression line, and effects of attraction – along negative regression line.

THE INFORMATIONAL SENSE OF THE PROBLEM

Here is more information situation in which we need to solve this problem. In view of the concepts discussed, problem of identification of genotype by phenotype is essentially reduced to the identification of breeding character on which this individual can be selected or not for further breeding. We have a mathematical model of the “genotype-environment” interaction that allows us to predict quantitative traits of individuals or populations (Mikhailenko and Dragavtsev, 2012). In addition, we have data for monitoring of all environmental factors affecting, as controllable and uncontrollable throughout the growing season of plants, as well as data on the actual growth and development of individuals and populations, ranging from planting until the end of quantitative results. That's the end results we have to classify individuals. The result of this classification will be the division of the whole set of phenotypes new generation of fissile subset of genotypes that have different sets of quantitative traits, some of which are economically valuable. The number of individuals in certain subsets may be very small or even be a few units. Given that the formation of these subsets will require modeling of states of each individual, without which it is impossible to correct classification, after the formation of subsets of genotypes is expedient to find the boundary separating them, which will further simplify and speed up the classification of the individual of other generations, not resorting to their modelling. With this approach, the first stage of classification is the training, where as a "non-ideal" (real) teachers' use mathematical models of individuals. The result of training will determine the number of possible classes and identification of the boundaries of subsets of classes, and the second phase of the task itself is the operative classification of the genotypes of individuals according to their phenotypic characteristics. Figure 2 shows the structural scheme of the problem (identification of genotype by phenotype).

In developing the algorithm for the general classification of genotypes based on the earlier principles will give a brief description of the evolution of the models used in modern genetics. So in 1984, the two existing models that describe the relationships genes - characters (the first model of Mendel (Mendel, 1865), the second – R. Fisher, C. Mather, S. Wright (Wright, 1950)), added a third - a model of eco-genetic organization of quantitative characters (MEGOQC) (Dragavtsev et al, 1984). The period from 1984 to 2011 has been theoretically predicted and experimentally confirmed 23 issues of this model. The most important are interpreting of nature and prediction: transgression, environmentally dependent heterosis, changes in the signs and levels of genotypic and genetic (additive) correlations, effects of the interaction "genotype-environment", changes in the numbers of genes and genetic variation of the amplitude characteristics of productivity, genetic homeostasis, and others (Kocherina and Dragavtsev, 2008).

Phenotypical meaning of individual plant's quantitative character in concrete point of growth is describe by classical model of R. Fisher:

\[ \Psi_i = \mu + \gamma_i + \pi_i \]  

(1)

Where \( \Psi_i \) is the phenotypic value of a quantitative trait in the \( i \)-th individual, \( \mu \) is the average trait in the population, \( \gamma_i \) is the genotypic deviation from the mean trait values of individuals, \( \pi_i \) is the \( i \)-individual environmental deviation from the mean of population.

The new model proposed in Dragavtsev (1998), describes the efficiency integral property of the \( i \)-th individual:

\[ \Psi_i = \mu \text{ (plant's productivity)} = \gamma_{\text{atr}} + \gamma_{\text{mic}} + \gamma_{\text{ad}} + \gamma_{\text{imm}} + \gamma_{\text{st}} + \gamma_{\text{soc}} + \gamma_{\text{com}} + \gamma_{\text{ont}} + \pi_{\text{com}} + \pi_{\text{ont}} + \pi_i \]  

(2)

where \( \Psi_i \) is the phenotypic trait value of productivity character at the \( i \)-th individual; \( \mu \) is the average productivity of the population; \( \gamma_{\text{atr}} \) is the deviation of attraction products of photosynthesis from the stems and leaves in the ear; \( \gamma_{\text{mic}} \) is the deviation of the distribution of products of attraction between the grains and chaff in the ear; \( \gamma_{\text{ad}} \) is the influence of the deviation system adaptability to the products effectiveness, measured by total dry biomass of plants, \( \gamma_{\text{imm}} \) is the impact of horizontal resistance in productivity; \( \gamma_{\text{st}} \) is the payment by biomass of limiting factors of soil nutrition; \( \gamma_{\text{com}} \) is a deviation tolerance to density; \( \gamma_{\text{ont}} \) is the deviation of the genetic variability in the duration of the phases of ontogenesis; \( \gamma_{\text{com}} \) is the
1. Evaluation of mechanisms transgressions and selection of parent’s pairs from a collection of varieties

2. Estimation of quantities of the contributions of genetic-physiological systems of the parents in breeding character

3. Prediction of breeding evidence of transgressive genotypes

4. Crossing and obtaining F\textsubscript{2} population

5. Identification of individual genotypes by phenotypes and comparison of selected transgressions with forecasts.

**Figure 1.** Block diagram of the relationship management tasks of genetic-breeding process.

**Figure 2.** Block diagram of a general algorithm for the identification of genotypes by phenotypes.
deviation of the genetic competition of plants for moisture, food, light, etc.; $\pi_{\text{com}}$ deviation is not the genetic competition caused by the unequal growth of the initial conditions, $\pi_{\text{cont}}$ is the deviation caused by the change of limiting factors in ontogenesis between bookmarks and development mark, $\pi_i$ is the deviation caused by the influence of the environmental fluctuations. Decipher each component of the model in the form of specific states of genetic-physiological systems and the modular structure of the model:

1. The system of attraction: the mass of stem and ear $\varphi_{11}$, $\varphi_{12}$ (commodity and non-market part of the plant);
2. Micro distribution: weight of the grain $\varphi_{21}$ and $\varphi_{22}$ of non-grain spike (chaff, awns, etc.);
3. Adaptation (resistance to environmental stressors and chemical environment): the degree of deceleration of growth processes under the influence of adverse environmental factors (stressors), speed and recovery time course of the normal growth of the processes;
5. The "paying" of doses of soil nutrition elements by dry biomass: parameters of the sensitivity characteristics of the productive to the doses of nutrients;
6. The tolerance to density: the parameters of sensitivity productive of indicators to the density of phytocenosis;
7. The variability of the periods of ontogenesis: is used in the selection for the "withdrawing" of the critical phases of the ontogenesis from the stressor, "beating" in a critical phase.

**THE BASIC MATHEMATICAL MODEL**

Reflect the given characteristics of the genetic-physiological systems of the first on the modular structure of the model of "genotype-environment" for the crops (Figure 3). Figure 3 corresponds to the mathematical model of the main (output) module [4].

where the following notation: $x_{ij}$ is the mass of grain in the ear $i$-th individual, $x_{2i}$ is the mass of chaff in the ear, $x_{3i}$ is the mass of straw in the ear, $u$—provision (control) of nitrogen nutrition; $f_1$ is the luminous efficiency factor, $f_2$ is the thermal efficiency factor products, $f_3$ is the moisture as a factor of productivity; $\Delta \varphi_1, \Delta \varphi_2$ is the influence of genetic-physiological systems; $\xi_1, \xi_2, \xi_3$ is the random perturbations, reflecting the uncertainty in the information model; $a_{kj}, b_k$, $c_{kj}, d_k$ is the dynamic parameters of the model.

We represent the model (2) in a more compact vector-matrix form:

$$\dot{X} = A(j_{1i}, j_{2i}, j_{3i})X(t) + b(j_{3i})u(t) + C(j_{3i})F(t) + D^* [j_{4i}(t) j_{6i}(t)] + x(t),$$

$$t \int [t_{0}(j_{\gamma}); T(j_{\gamma})]$$

in which all variables and parameters are combined into the corresponding vectors and matrices.

Model (4) determines the state of the $i$-th individual, and the effect of limiting factors, the differential for all individuals, as the action of genetic-physiological systems,
leading to perturbations of the states of individuals and the emergence of environmental and genetic variance. These perturbations can be represented as follows:

$$DX_i = X - X_i = U_E \Delta E + U_j \Delta j,$$

$$U_E = \frac{\|X_i\|}{\|E\|}, \quad U_j = \frac{\|X_i\|}{\|j\|},$$

where the $U_E, U_\phi$ are state vectors of the sensitivity functions of the module, respectively, to environmental and genetic perturbations; $\Delta E, \Delta \phi_i$ are vectors of the observed variations in environmental factors and unobservable genetic effects.

Equation (5) reflects the modeled contributions of environmental and genetic factors. However, the breeder usually has to do with the observed variations in the recognition of magnitude, which we denote as $\Delta Y_i$. In this case, the meaning of the identification of genotypes is to establish the causes of observed variations in characteristics of individuals compared with average of the population values. In the event that such causes are environmental factors, then we are dealing with modifications of the same genotype, and in establishing the genetic basis - with a new genotype.

We introduce the quadratic functional identification quality:

$$J_1 = \int_0^t [(DX_i(t) - DY_i(t))^T (DX_i(t) - DY_i(t))] dt,$$

whose meaning is "balancing" of the simulated and observed shifts of breeding traits by unobserved effects of seven genetic-physiological systems.

Required to minimize the criterion (6) of unobservable influences of genetic factors for the known variations $\Delta \phi_i$ environmental factors $\Delta E_i(t)$, which will assess the "contributions" of each of the genetic-physiological systems.

Now, having regions of admissible values of the impacts of genetic-physiological systems for individual genotypes $\Omega_k, k = 1, 2, 3 ... K$-indices of the genotypes (classes), we can determine the decision rule of identification:

$$k_i = k^*, \text{ if } Df_i^j \hat{W}_k.$$

(7)

With background information on deviations of environmental factors $\Delta E_i(t)$ and the observed deviations of quantitative traits i-th individual $\Delta Y_i(t)$ we have by the procedure (2) - (7) may determine that it belongs to the
genotype with the specified tag-mi. Fixing at the time of this procedure for each class of sets \((\Delta E_k (t), \Delta Y_k (t))\), we can construct a boundary between the genotypes in the space of environmental deviations and variation of quantitative traits:

\[
L_k = F_{k,k+1}(DE, DY),
\]

where \(F(.)\) - special function approximation bounds.

In this case, the decision rule is as follows:

\[
k_i = k^*, \text{ if } F_{k,k+1}(DE, DY) - c \leq 0,
\]

\[
k_i = k^* + 1, \text{ if } F_{k,k+1}(DE, DY) - c > 0,
\]

Where, \(c\) is the threshold value, which is one of parameters of the decision rule.

Thus, we examined the whole scheme of identification of genotypes for the observed phenotypes shown in Figure 1. This entire procedure is prior separation of the genotypes of individuals and modifications to the stage of teaching "teacher" more than a simple decision rule (8) (9). Due to the fact that this algorithm can make mistakes, then that teacher is "imperfect," or, more accurately, a "real teacher".

**ALGORITHM OF IDENTIFICATION**

Consider the Hamiltonian of the system:

\[
H = (DX_i(t)[DE] - DY_i(t))^T(DX_i(t)[DE] - DY_i(t)) +
+ l^T[A_(j) \mu(t) + b(t) + C(t)D_j(t) + D*(J(t)D_j(t) + (t)J(t)D_j(t))],
\]

(10)

Where: \(\lambda\) is the vector of conjugate variables, which is a solution of the system in reverse time.

\[
p_k \rightarrow - \frac{\partial H}{\partial DX} = - 2(DX_i(t)[DE] - DY_i(t)) + A^T \mu(t) + C(t)D_j(t) + D*(J(t)D_j(t) + (t)J(t)D_j(t)),
\]

(11)

Given the introduction of new auxiliary variables identification procedure is to minimize the criterion (5) on no observed effect of genetic-physiological systems will look like the following multi-step procedure:

\[
D_j^T_{i,j+1} = D_j^T_{i,j} - g \frac{\partial H}{\partial D_j}_{i,j},
\]

(12)

Where: \(j\) is the number of iterations of the process of working to minimize the criterion (5).

Upon reaching the iterations (12) breakpoint conditions the estimates of impacts genetic-physiological systems in the future will be denoted by \(-\Delta \phi_i^*\). With the obtained values of the separation vectors in a subset of the classes according to the rule (7), their boundaries are convenient to specify the system of inequalities:

\[
W_k^* j \in \Delta E_k \min \ell \in j \in I_k, l = \frac{1}{2}, \ldots, n,
\]

Where: \(l\) - the indices of genetic-physiological systems.

The system of inequalities (13), whereby the space of seven influences genetic-physiological systems by separation of the individuals in the future we will be called "eco-genetic portrait" of the genotype, bearing in mind that in the light of developing theory of eco-genetic organization of quantitative characteristics (TEGRQ), it is the only possible representation of the differences in genotype.

Note that the vectors of the effects of genetic-physiological systems \(\Delta \phi_i^*\) we are only "label" or guidelines for the formation of subsets of causal relationship:

\[
W_k^* (j) : (\Delta E_k, j \in I_k), i = 1, I_k.
\]

(14)

Where the average value of the vector of environmental variations on the final phase between the periods.

That’s it for these sets we construct a simple decision rules. To do so, combine the vector of environmental causes and consequences of the vector \(\Delta Y\) in the unit vector of \(Z^T = [\Delta E_k, \Delta Y]^T\).

Then we define the basic statistical characteristics of the classes on the sets (14) - vectors of expectation and covariance matrix \(M_{I_k} K_{I_k}\) as well as the same probability of occurrence of classes, which represent the ratio of estimates of numbers of individuals who have fallen into many separate classes \(I_k\) to the total number of individuals studied:

\[
p_k = \frac{I_k}{\hat{a} I_k},
\]

(15)

For these characteristics, it is easy to construct a function separating the classes \(F\) (de Groot, 1974):

\[
F_{k,k+1}(Z) = Z^T(\bar{K}_{Z_k} - K_{Z_k})Z + 2(\bar{M}_{Z_k} K_{Z_k} - M_{Z_k} + 1)K_{Z_k} + 1)Z',
\]

(16)

and the threshold number \(c\) of rules (9):

\[
c = 2 \ln \left(\frac{p_{k+1} + 1}{p_k} \frac{K_{Z_k}}{K_{Z_k} + 1}\right) + \ln \left(\frac{M_{Z_k} T_{Z_k} - M_{Z_k}}{K_{Z_k} + 1}\right) + \ln \left(\frac{M_{Z_k} - M_{Z_k} + 1}{K_{Z_k} + 1}\right),
\]

(17)

where: \(|K|\) is the norm of the matrix \(K\).
Obviously, in accordance with rule (9) for each new implementation of the causes and effects, but the signs of \( Z^T = [DF, \Delta \mathbf{Y}] \) we will need to compare the pairs all the possible genotypes.

As mentioned earlier, the vectors of the impacts of genetic-physiological systems \( \Delta \phi \) serve us only "labels" for the formation of subsets of causal relations. However, to solve the following problems of selection, we need a static version of the model of "ecological disturbance-response genetic-physiological systems". To do this we need to form a set of identification, which we estimate the parameters \( W \) of the desired model:

\[
D_j = W^T D_f.
\] (18)

Here, as an illustration for the solution of the problem we have considered only one of the modules of the general model of the identification of genotypes by their phenotypes. In the case of the need to include in the genotypes of characteristic other quantitative trait dimension of the problem can be increased without changing the essence of the approach. In this important feature of the developed theory of identification of genotypes is that it provides a solution to this problem during the whole period of ontogenesis, starting from first stages of ontogenesis, namely, the modules of the lowest level of the hierarchy, completing her final product output modules. This greatly improves the reliability of solving the problem and allows for more productive use of all the genotypic variability in the possession of the geneticist-breeder.

**CONCLUSIONS**

We propose a formalized theory of the identification of genotypes by their phenotypes, including:

Assessment (using the mathematical model and a special algorithm optimizing the unobserved variables seven contributions of genetic-physiological systems in the productivity of the individual); Classification of individuals with a given system of inequalities in the levels of contributions received by the genetic-physiological systems in the productivity of individual; Formation for each of the classes of individual's subsets of variations of environmental factors and variations in quantitative traits with simultaneous estimation of multidimensional statistical characteristics of these subsets combined;

Determination of the statistical characteristics of the variations in environmental factors and variation of quantitative traits boundaries of individual classes of genotypes, which can be implemented in a simplified algorithm for identification of genotype by phenotype.

**REFERENCES**


**Cite this article as:**


Submit your manuscript at http://www.academipublishing.org/journals/ajsr