Bioinformatic and molecular identification of wheat genes playing role in grain development

Ph. D. thesis

ATTILA SZŰCS

MTA Szegedi Biológiai Központ,

Növénybiológiai Intézet

Supervisor: Dr. Fehér Attila

Szeged

2007

Background

Wheat (Triticum sp.) is one of the most important crops in the world. It is cultivated from Helsinki (Finland) to Cape Town (South Africa). The quantity and quality of the wheat yield depends on local conditions (daily oscillation of temperature, the length of daytime, the type of soil, the distribution of rainfall, etc.) In Hungary, wheat is cultivated in 1,000,000 ha/year and a typical yield is 4 ton/ha. The value of the product is usually more than 100 billion HUF/year.

From the last century, the average temperature of the world continuously increases. It is not clear yet that human activity or some natural processes are the main reasons of this phenomenon. The "global warming" does not mean that every regions of the world become warmer, but means the aggregate of local changes in weather. The most probable prediction for Hungary is that the summer will be hotter and droughtier. After the problem was recognized, several research projects were established to better understand wheat genetics underlying stress tolerance in order to promote the development of new cultivars with improved yield safety.

My Ph.D. thesis describes our genomic research approaches which can be used for identifying those genes that are responsible for the quantity and quality of wheat harvest and play role in yield safety. We primarily concentrated our efforts on the drought and heat tolerance of the early phases of grain development in wheat. We hope that in a long term our results will help to accelerate wheat breeding and improve the efficiency of the selection of new cultivars.

Transcription profiling of the wheat egg cell

The life cycle of plants can be divided into a diploid sporophyte and a haploid gametophyte phase. Most of our knowledge comes from the dominant sporophyte cycle, because the gametophyte consists of only few cells which are encapsulated deeply in the tissues of the sporophyte generation. The new modern methods enable us to investigate this hidden life cycle in plants. With T-DNA insertion mutagenesis more than a hundred mutations were identified which affect the developing gametophyte (Pagnussat et al.,

2005). Using DNA micro array, 225 genes were identified as gametophyte specific (Yu et all., 2005).

The egg cell has a special role in the gametophyte, because after the fertilisation the new sporotophyte generation is developing from it. In most of the species studied (mainly animals), maternally stored mRNAs of the egg were identified as involved in the establishment of embryonic axes, diversification of cell types and morphological changes during early embryogenesis. Compared with animals, little is known about transcripts stored in egg cells of flowering seed plants (angiosperms) and activation of the zygotic genome after fertilization. There is only limited information about the biological activity of the egg cell and its biological processes. Investigation of the egg cell transcriptosome can give us a comprehensive view about these processes.

The model organism, Arabidopsis, is inadequate for the direct studying of the earliest steps in embryogenesis, because of it's small size and the difficulties in the isolation of egg cells. On the contrary, in the case of wheat and maize, there are routine methodologies for egg cell isolation. These techniques were partly (wheat) developed in the laboratory of Beáta Barnabás (Agricultural Research Institute, HAS, Martonvásár, Hungary). On the basis of their experiences and technical possibilities, we have made a wheat egg cell cDNA library and investigated that what kind of genes are expressed in this specific cell type. Meanwhile it cleared up that other research groups had also choosen this approach. Sprunck and her colleagues, in 2005, published the identification of 404 egg cell-expressed genes and their primary characterization. The sequences of these EST are avaible from international databases, so we could compare them with those 251 cDNA clones, which were identified by ourselves. The number of common sequences were small, so we identified 237 new egg cell specific genes. 73 of these are unknown wheat sequences, while 188 sequences were already described from other cDNA libraries. The analysis of both sequence pools suggests that there are high metabolic activities and intensive changes at the protein leveli n the wheat egg cells, so this cell type is not so quiescent, as it was postulated earlier (Russell 1993).

The apomictic „Salmon" genotype of wheat is capable of parthenogenetic development. Kumlehn and their colleges (2001) identified more than 500 ESTs from these egg cells, but their detailed description has not been submitted into any freely available databases.

We can conclude that up to now 1000-1500 wheat genes have been identified which are expressed in the egg cell. Only our cDNA library was prepared in phagemids and it is the most representative (app. $1.5 \cdot 10^6$ genes) that may allow a more detailed analysis. The size of the two other published libraries are much smaller as they consist only few thousands of ESTs, which are cloned into plasmids.

The transcription profiles of one- and two-celled zygotes

Sprunck and her colleagues (2005) made also a cDNA library from two-celled wheat embryos. They determined 789 EST sequences and based on these data they compared the expression profile of the egg cell and the two-celled zygote. In the dividing zygote, the transcription and the translation were more intensive and, naturally, cell cycle specific genes appeared. From two-celled maize embryos, Okamoto and his colleagues (2005) identified cDNA sequences that were expressed in the basal and/or the apical cell. They demonstrated that there is a major difference between the expression profile of this two cell type. These experiments did not answer, however, the question: At what time does the transition of the transcriptosome occur? Our cDNA library, which was made from zygotes 8 hours following the fertilization may help to solve this problem. Up to now, we have identified 369 EST sequences from this one-celled zygote cDNA library. Several transcription factors and proteins participating in signal transduction appeared, which may indicate that the zygote's own gene set is already active at this time. The explicit verification of the early activation of a zygote specific gene set needs additional experiments. 81 of the identified 369 zygote ESTs are new wheat genes or alleles. 23 of them also exist in the egg cell, 22 of them were identified in the 2-days-old zygote cDNA library (Sprunck et al. 2005). The amount of the available sequences are not sufficient for a comprehensive analysis, so the identification of new transcripts should be continued in order to get a better picture on fertilization induced changes in trancription.

Genomic approaches in wheat research

As it was mentioned earlier, wheat (Triticum astivum L.) is one of our most important agricultural crops, therefore the investigation of this species has big importance. The

application of modern genetic methods, however, are not so easy, because the genome size of wheat is huge, the chromosomes are colossal and the level of polyploidy is high. The allohexaploid wheat genome consist of 16 Gbase of nucleotides (Bennett and Smith, 1976), while the genome of the model organism rice consists of only 450 Mbase (Sasaki and Sederoff 2003). The main difference between the two genomes comes from the repetitive, non-coding, DNA sequences . The gene sets of these organisms are quite similar to each other (Sorells et al. 2003). The rice genome consists of about 40000 genes (Bennetzen et al. 2004). The estimations of the number of genes in wheat are in quite a big range, but surely they are far more than in rice (e.g. considering polyploidy).

The rice and barley as model organisms of cereal research can be useful for wheat research and breeding (Appels et al. 2003; Feuillet and Keller 2002; Ware et al. 2002). Because of the size of the wheat genome and the high number of repetitive sequences, the sequencing of the whole wheat genome will not be completed in the close future. Therefore, the EST sequencing projects have high priority. For this purpose several, cDNA libraries has been constructed representing various developmental stages and responses for different environmental signals. Our laboratory contributes to this effort by the production of egg cell- and zygote-specific cDNA libraries. 154 out of 620 of our EST sequences show no high local similarity to any other already identified wheat EST sequences (less then 90 % percent of identity), so probably they are new, up to now not identified wheat genes or alleles. In several other instances we could extend the already known sequences by our ESTs, while in those cases where the cDNA sequences represented by our ESTs were already fully known, the new information was that these genes are expressed in the egg cell and/or zygote.

One of the potential use of these EST sequences is the production of microarrays and the investigation of comprehensive gene expression patterns. Whole genome microarrays of wheat have not been manufactured yet, due to the unfinished genome sequencing. The most representative microarray for wheat was developed by Affymetrix (www.affymetrix.com). 55052 different transcripts can be detected by this array. It is using short (25 bp) oligonucleotids which guarantees the high level of specificity. The disadvantages of this method: the requirement of a special scanner and the high price (1 slide app. 500 Euro). In order to decrease the cost of the microarray experiments,, several

projects were funded. In the United Kindom, 10000 cDNA sequences from 35 different libraries were amplified using PCR and spotted on microarray slides (Wilson et al. 2004). In this case, the spotted DNA probes were full length cDNA, so this approach is not as specific as the Affymetrix solution. This microarray can be used at a lower price, but the experiments must be done in the UK laboratory. In Japan, a wheat genomic consortium developed an oligonucleotid microarray, which represents 22000 different wheat transcripts (Kawaura et al. 2006) and successfully applied it for identifying genes the expression of which is modified by salt stress.

Our objective was to design and manufacture a relatively cheap, wheat-specific oligonucleotid array (oligo-chip), which can be used even by breeders. We also intended to use this array to investigate the changes of the wheat transcriptosom in the grain, during the grain filling period. We chose about 2000 cDNA sequences from publicly available databases, which exist only or very abundantly in cDNA libraries associated with grain development. Of course, the prepared microarray does not represent the whole gene set which participates in the complex grain filling process. However, it may give a sufficiently detailed picture about the processes underlying this important event. One of our collaborators (BRC, Cell Division Cycle and Stress Adaptation Group) designed additional 1500 oligonucleotides representing stress-responsive genes in wheat. The final oligochip developed in our laborytory therefore contains more than 3000 oligonucleotides.

Handling and processing of large pools of sequence data

The amount of available biological and DNA/protein sequence information is increased so much due to modern high throughput research methods that without the help of complex computer technology the data processing and analysis is impossible. There is also a demand that the new results should be incorporated as a new part into a complex information set. To discover the real relation among several experimental circumstances and the data obtained by various methodologies, we have to establish connections between several databases (3D modeling, transcription profile, interacting partners, protein motifs etc.)

I have developed a software solution which handles several molecular biology specific file formats (BLAST, PICKY, GenBank, FASTA, etc.) and is capable to handle large amount of data (e.g. 2.2 gigabyte input GenBank file). This program also makes it possible for an inexperienced users to create SQL queries on a graphical interface. The software allows the creation of new data tables by importing tab delimited files or the results of SQL queries. It is also possible to cross-link the data tables to hierarchical databases such as KEGG or Gene Ontology. Several specific functions have been introduced in order to allow the genomic analysis of organisms with not fully sequenced genomes based on their EST sequences and the distribution of their EST sequences in the various cDNA libraries ("virtual gene expression").

Publication list:

Dorjgotov D, Szucs A, Otvos K, Szakonyi D, Kelemen Z, Lendvai A, Ponya Z, Barnabas B, Brown S, Dudits D, Feher A.
Specific features of RHO GTPase-dependent signaling in plants.
Cell Biol Int. 2003;27(3),191-2.

Otvos K, Pasternak TP, Miskolczi P, Domoki M, Dorjgotov D, Szucs A, Bottka S, Dudits D, Feher A. Nitric oxide is required for, and promotes auxin-mediated activation of, cell division and embryogenic cell formation but does not influence cell cycle progression in alfalfa cell cultures.
Plant J. 2005 Sep;43(6),849-60.

Szucs A, Dorjgotov D, Otvos K, Fodor C, Domoki M, Gyorgyey J, Kalo P, Kiss GB, Dudits D, Feher A.
Characterization of three Rop GTPase genes of alfalfa (Medicago sativa L.).
Biochim Biophys Acta. 2006 Jan-Feb;1759(1-2),108-15.

Szűcs A, Ignáth I, Jäger K, Barnabás B, Fehér A (2007) Fertilization induces gene expression changes in wheat egg cells within 8 hours. Plant Cell Reports (in preparation)