BioTechnologia vol. 93(1) • pp. 41-47 • 2012
Journal of Biotechnology, Computational Biology and Bionanotechnology RESEARCH PAPER

Computational analysis of alternative spliced genes on maize chromosome 1

JOANNA GRACZ [1*], ALEKSANDRA ŚWIERCZ [1, 2], TOMASZ TWARDOWSKI[1]

[1] Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

[2] Institute of Computing Science, Poznan University of Technology, Poznań, Poland

* Corresponding author: jgracz@ibch.poznan.pl

Abstract

Alternative splicing is an important part of mRNA processing which results in protein diversity in all eukaryotes. In plants, this process is still poorly understood, however recent computational analysis has shown that alternative splicing is far more prevalent than previously thought. For better characterization of alternative splicing in maize, one of the most important crop species, we used AUGUSTUS (web-application deployed on local or remote host) to predict multiple transcripts and alternative splicing events from maize chromosome 1 sequence. From over 300 million bp of chromosome 1, AUGUSTUS software predicted *ab initio* 46 400 genes and 20% of the estimated genes have at least two transcripts. For 412 genes with three transcripts we performed additional analysis including EST's alignment, protein identification and comparative evaluation with genes predicted using F genesh and Gramene software. In consequence we have identified alternative splicing events for 42 genes from maize chromosome 1.

Key words: *Zea mays*, alternative splicing, AUGUSTUS, expressed sequence tags

Introduction

In 1978, a year after the discovery that the gene structure consists of coding regions (exons) interrupted by noncoding sequences (introns), it was proposed that in RNA transcribed from a single gene different exons could be joined together to produce various mRNA transcripts (Reddy, 2007). This particular process, called alternative splicing (AS), was initially described for individual genes as isolated events (in 1994, the level of alternative spliced genes in human genome was estimated at less than 5% (Sharp, 1994)). Recent studies indicate that 95% of human genes and around 20-30% of plant genes undergo AS (Severing et al., 2009). This mechanism generates transcriptome diversity and partially explains the discrepancy between an estimated number of genes compared to the number of proteins encoded by the same genome. Most of AS events occur within exons, which affects the binding properties, intercellular localization, stability and activity of translated proteins. Furthermore, some splice isoforms contain premature termination codons (PTC) which are targeted for nonsense-mediated mRNA decay (NMD) (Maquat, 2004). This pathway is one of the post-transcriptional mechanisms for regulating gene expression.

Although most of AS studies have been performed on humans and other vertebrates, so far several thousand plant genes are known to produce multiple transcripts (Reddy, 2007; Wang and Brendel, 2006). Genes that show AS in plants are often part of fundamental processes such as photosynthesis, flowering or defense responses. One of the first AS events studied in *Arabidopsis* and spinach was ribulosebisphosphate carboxylase/oxygenase (rubisco) activase – nuclear-encoded chloroplast enzyme required for rapid formation of the carbamate on the active site of rubisco (Werneke et al., 1989). Other well-characterized examples of alternative spliced genes in plants include: RNA polymerase II (Dietrich et al., 1990), chorismate synthase (Gorlach et al., 1995), *FCA* gene which encodes a nuclear ABA receptor (Macknight et al. 2002; Quesada et al., 2003) or splicing factors like serine/arginine-rich (SR) proteins (Lazar and Goodman, 2000; Reddy, 2004). SR proteins are highly conserved non-small nuclear ribonucleoprotein (non-snRNP) that play an important role in regular and alternative splicing.

The majority of studies on AS or SR proteins in plants were performed on *Arabidopsis*, however it has been established that SR proteins produce multiple transcripts and utilize non-canonical splice sites in maize as well (Gupta et al., 2005). A few other studies were conducted on maize and revealed, for instance, AS of maize regulatory MuDR transposable element (Hershberger et al., 1995), glutathione S-transferase (*bronze2*) (Marrset al., 1997), *waxy* (Mareillonnet et al., 1997) or *Knox7* (Morere-Le et al., 2007) genes. Despite these reports, the abundance of alternatively spliced genes is still underestimated in maize and other less-characterized plants. Most of AS events in plants are estimated on the basis of ESTs or cDNA data, which are generally biased and have diverse quality. Moreover, EST collections represent only a small part of different environmental conditions, whereas plant responses to distinct stresses differ a lot and might affect AS efficiency or splicing patterns. It is likely that better EST coverage in less-studied plants will increase the number of AS events discoveries (Barbazuk et al., 2008). Progress in understanding the process of AS still requires answers to many questions, such as differences of AS patterns between plants and animals, monocots and dicots and between different plant species. To achieve this goal, first of all one has to elucidate constitutive and regulated splicing in specific plant species. An analysis of important bioenergetic and crop plants such as *Zea mays* will be very helpful in further comparative genomic research, studies on domestication and trait selection for breeding programs. In this paper we have described AS events on maize largest chromosome using *ab initio* gene prediction and aligning results with EST collection. This approach allows us to identify novel AS genes,determinethe significance of AS phenomenon and can be used to analyze the entire genomes as well.

Material and methods

As the object of our research we selected maize for its high economic value and in correlation with the studies on this plant previously conducted in our laboratory (Twardowski et al., 2010; Korbin et al., 2011). For *ab initio* gene prediction, we used AUGUSTUS program available at WWW server (http://augustus. gobics.de) and maize chromosome 1 sequence from GenBank database (accession number GK000031.1). We decided to choose AUGUSTUS program, because it is one of the first gene finder programs that can predict the formation of multiple transcripts for predicted genes. It is also freely available from the web server and has pre-calculated sets of parameters for *Zea mays* (and also for 48 other species). This software is based on generalized hidden Markov model (GHMM) – complex probabilistic model which combines information from a variety of different sources – signal and sequence content. Exons, introns, the sequence around a splice site etc., correspond to states in the model and each state is connected to a DNA sequence with certain pre-defined emission probabilities. AUGUSTUS, like any other HMM-based gene prediction software, finds an optimal parse, that is the segmentation of the input DNA sequence which is consistent with all the given constrains (Stenke and Morgenstern, 2005; Stenke et al., 2005). For input at AUGUSTUS server, a user can upload sequences of maximum total length of 3 mln bp, therefore it was decided to divide maize chromosome 1 (total length 300 239 041 bp) into one hundred fragments of 3 mln bp each and one of 239 041 bp. To avoid omitting genes that can be located near the ends of those fragments, we created one hundred of 40 000-bp long sequences which overlapped at neighboring fragments. AUGUSTUS web server has four different options for the number of reported alternative transcripts depending primarily on posterior probability of exons and introns. To perform the scheduled analysis, we chose options with few reported transcripts. The output data in General Feature Format (GFF), consisting of a coding sequence and an amino acids sequence of predicted gene, were converted for further analysis to FASTA file format.

In the next step, we selected 412 genes with 3 or more transcripts predicted by AUGUSTUS and searched NCBI EST database with MEGABLAST algorithm. Amino acid sequences translated from selected genes were used to search UniProtKB Database with BLASTP algorithm. The search was performed with Geneious Basic 5.0.2. software (Drummond et al., 2010), which is a single desktop application to store, visualize and analyze complex biological data (free version for non-commercial use is available at http://www.geneious.com).

We also compared the selected genes with predictions obtained from other platforms – Gramene and F genesh using the maize genome browser (http://maizesequence.org) and BLASTN algorithm. Gramene is a comparative genome mapping database specializing in

Table 1. List of proteins[1] identified from BLASTP alignment to UniProtKB with obtained parameters and UniProtKB accession

Gene no.	E-value	% of identities	Identified protein	Accession (UniProtKB)
1	2.92e-107	44.70%	DNA (cytosine-5)-methyltransferase DRM1 from *Arabidopsis thaliana*	9LXE5
2	7.11e-114	45.70%	DNA (cytosine-5)-methyltransferase DRM2 from *Arabidopsis thaliana*	Q9M548
3	8.92e-169	65.80%	Protein FIZZY-RELATED 2 from *Arabidopsis thaliana*	Q8L3Z8
4	5.23e-175	62.10%	Probable serine/threonine-protein kinase from *Arabidopsis thaliana*	Q3EDL4
5	0	68.50%	Phytoene dehydrogenase, chloroplastic/chromoplastic from *Solanum lycopersicum*	P28554
6	1.66e-131	63.40%	26S proteasome non-ATPase regulatory subunit 4 from *Arabidopsis thaliana*	P55034
7	2.01e-156	79.90%	Asparagine synthetase [glutamine-hydrolyzing] from *Sandersonia urantiaca*	O24338
8	1.27e-134	96.90%	Myb-related protein P from *Zea mays*	P27898
9	1.73e-151	81.70%	Homocysteine S-methyltransferase 2 from *Zea mays*	Q9FUM9
10	0	78.10%	Glutamate decarboxylase 2 from *Arabidopsis thaliana*	Q42472
11	5.48e-106	69.50%	Probable 125 kDakinesin-related protein from *Arabidopsis thaliana*	P82266
12	1.33e-108	76.20%	12-oxophytodienoate reductase 3 from *Arabidopsis thaliana*	Q9FUP0
13	0	83.30%	Alpha-galactosidase from *Oryza sativa subsp. japonica*	Q9FXT4
14	0	55.00%	Seed lipoxygenase-3 from *Glycine max*	P09186
15	0	77.80%	Lipoxygenase 2 from *Oryza sativa subsp. japonica*	P29250
16	2.39e-148	46.70%	65-kDa microtubule-associated protein 6 from *Arabidopsis thaliana*	Q9SIS3
17	5.44e-168	76.90%	Ferredoxin-NADPreductase, root isozyme, chloroplastic from *Oryza sativa subsp. japonica*	P41345
18	7.01e-141	86.8%	Glutamate dehydrogenase *Zea mays*	Q43260

[1]Amino acid sequences from predicted by AUGUSTUS genes and exon coordinates of respective genes are available in Supplementary materials 1 Table 2

monocots. It brings together ESTs collections, genetic maps, map relations, genes, proteins and publications providing connection between different sources of information (Ware et al., 2002;Youens-Clark et al., 2011). F genesh (Find GENES using HMM); on the other hand, is based on a statistical approach and uses HMM trained on monocots, but does not consider alternative splicing (Solovyev et al., 2006). It was previously used for gene prediction in maize and in comparison with four other programs (GeneMark.hmm, GENESCAN, GlimmerR and Grail) it yielded the most accurate genes predictions (Yao et al., 2005).

For the selected gene pool we performed additional search at NCBI Reference RNA Sequences (refseq_rna) database using BLASTN algorithm. The obtained results for mRNA and ESTs have been presented in a graphical form using Integrative Genomic Viewer (free available at http://www.broadinstitute.org/software/igv).

Results and discussion

In the first step of the analysis from AUGUSTUS *ab initio* prediction based on the information in the genome sequence, we located 46 400 genes on a maize chromosome 1. This result suggests the existence of a high number of false-positive predictions, because the estimated number of genes in the whole maize genome is about 32 000 (Schnable et al., 2009). One fifth (9410) of the genes predicted by AUGUSTUS have more than one transcript and 412 genes out of them have three or more. For the following evaluation, we chose these 412 genes with the highest number of predicted transcripts. We assumed that it is more likely that they match real transcripts

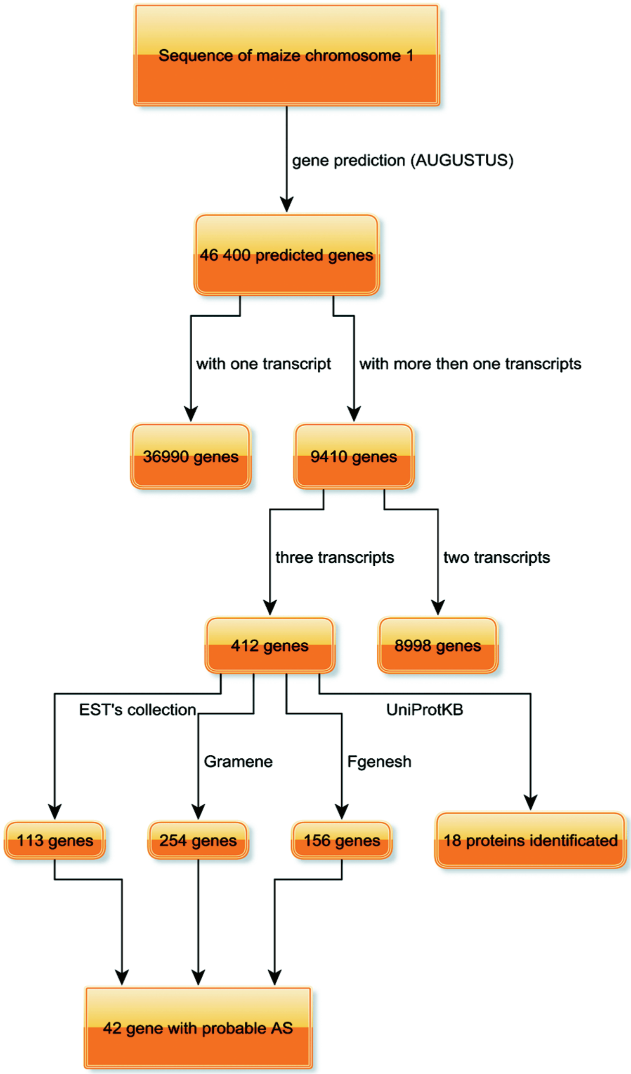

Fig. 1. Outline of AS analysis performed on maize chromosome 1. Flow of data analysis and gene pool restriction with exact number of considered genes. For EST's collection searching we used MEGABLAST algorithm, for F genesh and Gramene – BLASTN and for UniProtKB – BLASTP

expressed in plants (ESTs) or at least one of them does. For aligning transcripts with EST collection using MEGABLAST algorithm, we established E-value threshold on 1e-100 and the level of percentage of identities at 85% for meaningful result. After aligning transcripts (over 1200 sequences) to EST collection for 272 out of 412 (66%) we confirmed the existence of:

- all three transcripts for 113 genes (27.4%),
- two transcripts for 129 genes (31.3%),
- only one transcript for 30 genes (7.3%).

In the next step, we compared the results obtained from AUGUSTUS with predictions from Gramene and F genesh. We confirmed AS events for 254 and 156 genes, respectively (for meaningful results we used the same parameters as for MEGABLAST analysis). From Gramene predictions only 23 genes had more than one transcripts, which points towards high sensitivity of AUGUSTUS program. Forty two gens from the pool of 113 genes with three transcripts confirmed by alignment with ESTs were also validated by Gramene and F genesh predictions (Supplementary materials 1 Table 1). Thus we considered these 42 genes as the most probable AS events (see Figure 1). An example of AUGUSTUS outcome (for gene 1) is presented in Figure 2 and in GFF format for every gene in Supplementary material 2. The distribution and location of those 42 genes pool on chromosome 1 is presented in Figure 3.

To better depict the relative position of the predicted transcripts, for those 42 gene pool and its confirming data (ESTs and mRNAs), we used Integrative Genomics Viewer to present them in correlation with chromosome 1 sequence (Supplementary materials 1 Figure 1). For 16 genes, we found a very accurate exon coordinate prediction made by AUGUSTUS in comparison to the corresponding mRNA sequence (genes number: 1, 2, 3, 4, 5, 8, 18, 20, 23, 24, 27, 29, 31, 34, 38 and 41). For the remaining transcripts, there was only a rather small coverage of mRNA sequence, which can be partially explained by incomplete collection of the reference maize mRNA sequence.

Parallel to query ESTs collection, an analysis at the protein level was also performed. It included alignment of amino acids sequences (translated from AUGUSTUS-predicted genes) to UniProtKB using BLASTP algorithm. We set up E-value threshold at 1e-100 and the percent of identities at least 40%. In this way we identified 18 proteins – most of them were from *A. thaliana* and 3 were from maize (Table 1). For one of the proteins identified from *Zea mays*, Myb-related protein P, alternative splicing had been confirmed earlier (Grotewold et al., 1991). Similarly, other 3 proteins (– probable serine/threonine-protein kinase (Yamada et al., 2003), glutamate decarboxylase 2 (Zik et al., 1998) and 65-kDa microtubule associated (Iida et al., 2009) from *A. thaliana*, have two isoforms arising from alternatively spliced genes.

The growing availability of genome sequences has produced enormous amount of data which requires appropriate interpretation to obtain useful information. Therefore, there is still a great need for bioinformatics

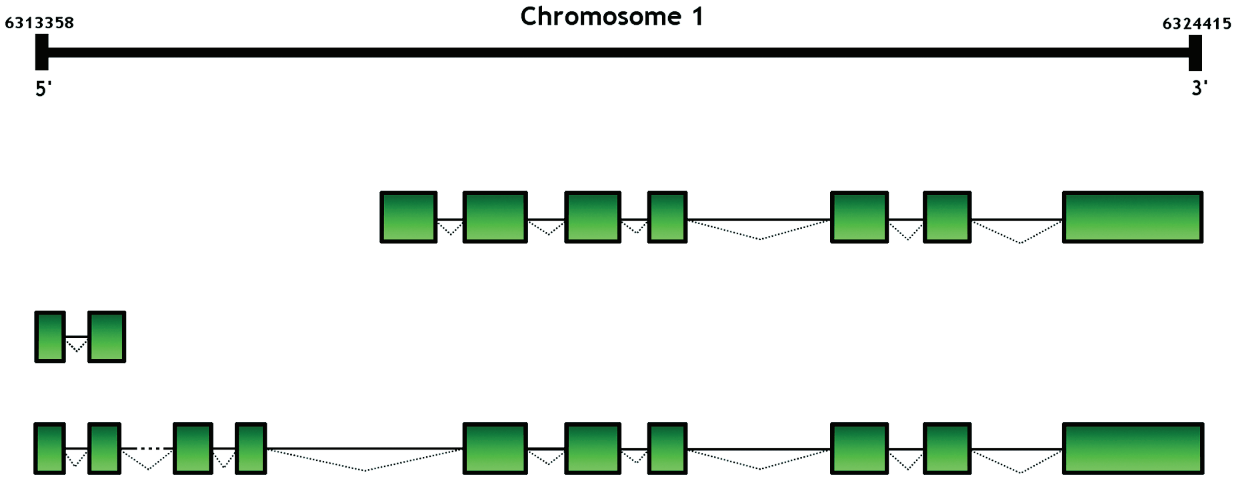

Fig. 2. Scheme of intron and exon organization of 3 transcript (1.1, 1.2, 1.3) from gene 1 predicted by AUGUSTUS. Each of presented transcripts match different ESTs and mRNA. Longest transcript (1.3) consist of 10 exons and one very large intron – 6782 nt (between exon 8 and 9), transcript 1.2 consist of only 2 exons and raise from alternative 3′ splice site acceptor of exon 9. In transcript 1.1 there is an additional exon, which is located in intron 6 of transcript 1.3

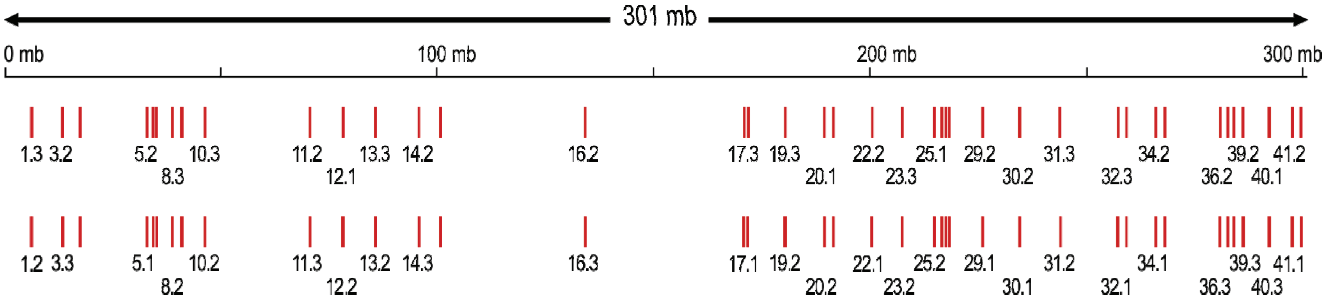

Fig. 3. Distribution and localization of restricted pool of genes predicted by AUGUSTUS

tools for gene prediction and specialized applications e.g. searching for miRNA, AS events or specific sequences motifs. The purpose of this study was the identification of AS events in transcripts arising from genes located on the maize's largest chromosome. The results of our analysis allowed us to verify the significance of AS, helped us diminish the total number of chromosome 1 genes to those undergoing AS and thus focus on specific ones. Within this *in silico* experiment we had a chance to evaluate the accuracy of AUGUSTUS predictions, compare it to the other currently available programs and to build a simple pipeline to restrict the pool of candidate genes for further experiments (Fig. 1).

AUGUSTUS is one of the few programs that have a training set for maize sequences and setup options to control predicted splice variants per gene. This allows users to choose between sensitivity or specificity of the program output. Despite the fact that we have chosen almost the most stringent options with few predicted transcripts, we have obtained many false positive predictions. One reason for this may be an insufficient training set, because the quality of predictions directly depends on the quality of training sets. Compared with *A. thaliana* or human genome, maize genome as well as maize's ESTs collections are still insufficiently studied. Another reason may be the high specificity of the maize genome – it has a lot of repetitive elements, many of which are transposons and retrotransposons, thus genes occupy only a small fraction of about 10-15% of the genomic sequence. The total proportion of genes undergoing AS events (that is 20%) was consistent with the previously estimated number (Barbazuk et al., 2008). When considering genes with only three or more predicted transcripts, AUGUSTUS makes 66% of correct predictions. Moreover, 58.4% of these genes may undergo AS, which was confirmed by aligning the predicted transcripts with different ESTs. More than half (42) of genes with few predicted transcripts were also validated by other programs (Gramene or F genesh). Those 42 genes with 3 alternative transcripts were recognized as novel, previously undescribed, splicing events. Sixteen genes from that pool showed high similarity to mRNA deposited in Refseq_rna database. An analysis of the translated amino acid sequences indicated four documented cases of AS events (from 18 identified proteins). These results confirm that AUGUSTUS is a well-constructed program and may serve as an efficient tool for AS searching, although it will still be necessary to improve its abilities.

The proposed scheme of *in silico* analysis can be used for effective reduction of gene pools, which is important especially when expensive biochemical experiments are going to be carried out in the subsequent step of analysis. In further studies on AS in maize, we plan to use 42 genes which were identified at every level of presented analysis.

Conclusions

46 400 genes (including false-positive predictions) were found by AUGUSTUS program in maize chromosome 1. This gene pool was reduced to 9 410 genes with possible AS events. We further reduced that number to 412 genes with three or more transcripts and have identified 42 genes by aligning their sequences with different ESTs and compared them with predictions obtained using two other gene finder programs (Gramene and F-genesh). We also identified 18 proteins from which 4 were previously demonstrated to have different isoforms that arise from AS events. We recognized AS as a novel event for 42 genes and 14 proteins identified by the described procedure. The foregoing results indicated importance of AS events in maize genome and have demonstrated usefulness of the applied procedure.

References

Barbazuk W.B., Fu Y., McGinnis K.M. (2008) *Genome-wide analyses of alternative splicing in plants: Opportunities and challenges.* Genome Res. 18: 1381-1392.

Dietrich M.A., Prenger J.P., GuilfoyleT.J. (1990) *Analysis of the genes encoding the largest subunit of RNA polymerase II in Arabidopsis and soybean.* Plant Mol. Biol. 15: 207-223.

Drummond A.J., Ashton B., Cheung M., Heled J., Kearse M., Moir R., Stones-Havas S., Thierer T., Wilson A. (2010) Geneious v5.0

Gorlach J., RaeseckeH.R., Abel G., Wehrli R., Amerhein N., Schmid J. (1995) *Organ-specific differences in the ratio of alternatively spliced chorismate synthase (LeCS2) transcripts in tomato.* Plant J. 8: 451-456.

Grotewold E., Athma P., Peterson T. (1991) *Alternatively spliced products of the maize P gene encode proteins with homology to the DNA-binding domain of myb-like transcription factors.* Proc. Natl. Acad. Sci. USA 88: 4587-4591.

Gupta S., Wang B.B., Stryker G.A., Zanetti M.E., Lal S.K. (2005) *Two novel arginine/serine (SR) proteins in maize are differentially spliced and utilize non-canonical splite site.* Biochim. Biophys. Acta 1728: 105-114.

Hershberger R.J., Benito M.I., Hardeman K.J., Warren C., Chandler V.L., Walbot V. (1995) *Characterization of the major transcripts encoded bu the regulatory MuDR transposable element of maize.* Genetics 140: 1087-1098.

Iida K., Fukami-Kobayashi K., Toyoda A., Sakaki Y., Kobayashi M., Seki M., Shinozaki K. (2009) *Analysis of multiple occurences of alternative splicing events in Arabidopsis thaliana using novel sequenced full-lenghtcDNAs.* DNA Res. 16: 155-156.

Korbin M., Kuras A., Keller-Przybyłkowicz S., Adamczyk J., Twardowski T., Kietrys A.M., Szopa A., Adamczewski K. (2011) *Method to precisely differentiate maize genotypes, simplified molecular test and specific identity biomarker.* In Polish Patent Office P393620.

Lazar G., Goodman H.M. (2000) *The Arabidopsis splicing factor SR1 is regulated by alternative splicing.* Plant Mol. Biol. 42: 571-581.

Macknight R., Duroux M., Laurie R., Dijkwel P., Simpson G., Dean C. (2002) *Functional significance of the alternative transcript processing of the Arabidopsis floral promoter FCA.* Plant Cell 14: 877-878.

Marrillonnet S., Wessler S.R. (1997) *Retrotransposon insertion into maize waxy genes results in tissue-specific RNA processing.* Plant Cell 9: 967-978.

Marrs K.A., Walbot V. (1997) *Expression and RNA splicing of the maize glutathione S-transferase Bronze2 gene is regulated by cadmium and other stresses.* Plant Physiol.113: 93-102.

Maquat L.E. (2004) *Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics.* Nat. Rev. Mol. Cell Biol. 5: 89-99.

Morere-Le Paven M.C., Anzala F., Recton A., Limami A.M. (2007) *Differential transcription initiation and alternative RNA splicing of Konox7 a class 2 homeobox gene of maize.* Gene 401: 71-79.

Quesada V., Macknight R., Dean C., Simpson G.G. (2003) *Autoregulation of FCA pre-mRNA processing controls Arabidopsis flowering time.* EMBO J. 22: 3142-3152.

ReddyA.S.N. (2004) *Plant serine/arginine-rich proteins and their role in pre-mRNA splicing.* Trends Plant Sci. 9: 541-547.

Reddy A.S.N. (2007) *Alternative Splicing of Pre-Messenger RNAs in Plants in the Genomic Era.* Annu. Rev. Plant Biol. 58: 267-294.

Schnable P.S., Ware D., Fulton R.S., Stein J.C., Wei F., Pasternak S. et al. (2009) *The B73 maize genome: complexity, diversity and dynamics.* Science 326: 1112-1115.

Severing E.I., Van Dijk A.D.J., Stiekema W.J., Van Ham R.C.H.J. (2009) *Comperative analysis indicates that alternative splicing in plants has a limited role in functional expansion of the proteome.* BMC Genomics 10: 154.

Sharp P.A. (1994) *Split genes and RNA splicing.* Cell 77: 805-815.

Solovyev V., Kosarev P., Seledsov I., Vorobyev D. (2006) *Automatic annotation of eukaryotic genes, pseudogenes and promoters.* Genome Biol. 7: S10.1-12

Stanke M., Morgenstern B. (2005) *AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constrains.* Nucl. Acid Res. 33: 465-467.

Stanke M., Keller O., Gunduz I., Hayes A., Waack S., Morgenstern B. (2006) *AUGUSTUS: ab initio prediction of alternative transcripts.* Nucl. Acids Res. 34: 435-439.

Twardowski T., Kietrys A.M., Szopa A., Adamczewski K., Adamczyk J., Korbin M., Kuchta P. (2010) *Method to determine the herbicide stress resistance of maize varietes, a diagnostics kit to determine the resistance of maize lines to herbicide stress and use of RNA aptamers to detect resistant and sensitive maize lines.* In: Polish Patent Office, P392284.

Wang B.B., Brendel V. (2006) *Genomewidecomperative analysis of alternative splicing in plants.* Proc. Natl. Acad. Sci. USA 103: 7175-7180.

Ware D.H., Jaiswal P., Junjian N., Yap I.V., Pan X., Clark K.Y. at al. (2002) *Gramene, a tool for grass genomics.* Plant Physiol. 130: 1606-1613.

Werneke J.M., Chatfield J.M., Ogren W.L. (1989) *Alternative mRNA splicing generates the two ribulosebisphosphate carboxylase/oxygenaseactivase polypeptides in spinach and Arabidopsis.* Plant Cell 1: 815-825.

Yamada K., Lim J., Dale J.M., Chen H., Shinn P., Palm C.J. et al. (2003) *Empirical analysis of transcriptional activity in the Arabidopsis genome.* Science 302: 842-846.

Yao H., Guo L., Fu Y., Borsuk L.A., Tsui-Jung W., Skibbe D.S. et al. (2005) *Evaluation of five ab initio gene prediction programs for the discovery of maize genes.* Plant Mol. Biol. 57: 445-460.

Youens-Clark K., Buckler E., Casstevens T., Chen C., DeClerck G., Derwent P. et al. (2011) *Gramene database in 2010: updates and extensions.* Nucl. Acids Res. 39: D1085-1094.

Zik M., Arazi T., Snedden W.A., Fromm H. (1998) *Two isoforms of glutamate decarboxylase in Arabidopsis are regulated by calcium/calmodulin and differ in organ distribution.* Plant Mol. Biol. 37: 967-975.