OPEN ACCESS Potravinarstvo® Scientific Journal for Food Industry

Potravinarstvo, vol. 9, 2015, no. 1, p. 149-153
doi:10.5219/467
Received: 11 March 2015. Accepted: 14 April 2015.
Available online: 25 May 2015 at www.potravinarstvo.com

ISSN 1337-0960 (online)

BIOINFORMATIC APPROACH IN THE IDENTIFICATION OF *ARABIDOPSIS* GENE HOMOLOGOUS IN *AMARANTHUS*

Jana Žiarovská, Michal Záhorský, Zdenka Gálová, Andrea Hricová

ABSTRACT

Bioinfomatics offers an efficient tool for molecular genetics applications and sequence homology search algorithms became an inevitable part for many different research strategies. Appropriate managing of known data that are stored in public available databases can be used in many ways in the research. Here, we report the identification of RmlC-like cupins superfamily protein DNA sequence than is known in *Arabidopsis* genome for the *Amaranthus* – plant specie where this sequence was still not sequenced. A BLAST based approach was used to identify the homologous sequences in the nucleotide database and to find suitable parts of the *Arabidopsis* sequence were primers can be designed. In total, 64 hits were found in nucleotide database for Arabidopsis RmlC-like cupins sequence. A query cover ranged from 10% up to the 100% among RmlC-like cupins nucleotides and its homologues that are actually stored in public nucleotide databases. The most conserved region was identified for matches that posses nucleotides in the range of 1506 up to the 1925 bp of RmlC-like cupins DNA sequence stored in the database. The *in silico* approach was subsequently used in PCR analysis where the specifity of designed primers was approved. A unique, 250 bp long fragment was obtained for *Amaranthus cruentus* and a hybride *Amaranthus hypochondriacus x hybridus* in our analysis. Bioinformatic based analysis of unknown parts of the plant genomes as showed in this study is a very good additional tool in PCR based analysis of plant variability. This approach is suitable in the case for plants, where concrete genomic data are still missing for the appropriate genes, as was demonstrated for *Amaranthus*.

Keywords: BLAST analysis; alignment; Rml-C like cupins; *Amaranthus*; PCR identification

INTRODUCTION

Bioinformatics provides an interdisciplinary tool, that is used to manage and analyse biological data and known sequences of nucleic acids **(Cannataro et al., 2009)**. Many features of nucleic acids can be used in bioinformatic algorithms as motifs for description of their genomic variability and their better understanding. Individual sequence motifs are recognized by their order and nucleotide preference and many motif discovery algorithms have used in different molecular or bioinformatic studies **(Aravind and Koonin, 1999; Hertz and Stormo, 1999; La and Livesay, 2005; Rasouli et al., 2013; Gardner and Slezak, 2014)**.

Here, the bioinformatic algorithms were applied for known cupin DNA sequences. Cupin proteins are reported as to be structurally conserved and in function very divergent superfamily of proteins **(Khuri et al., 2001)** that are germin-related. These proteins were analysed by their EST and microbial GeneBank databases as having representatives in many procaryotic and eukaryotic organisms, moreover, 26 residues of cupins intermotif regions were found in cereal proteins **(Dunwell et al., 2000)**. RmlC-like cupins superfamily protein DNA sequence is available in SeedGeneDB (sgdb.cbi.pku.edu.cn/) database under the accession code AT2G18540, what makes it suitable for applying the bioinformatic tools such as BLAST **(Altschul et al. 1990)** to find homology or conserved regions.

In this study, we identify the conserved nucleotides of available genomic data of Arabidopsis RmlC-like cupins superfamily protein suitable for bioinformatic approach based primer designation and subsequently PCR identification of the presence of this sequence in the genome of *Amaranthus*.

MATERIAL AND METHODOLOGY

Plant material of *Amaranthus cruentus* (Ficha cultivar) and a hybride *Amaranthus hypochondriacus x hybridus* (hybrid K-433) was planted under the in vitro conditions. DNA was extracted following the instructions of GeneJET Plant Genomic DNA Purification Mini Kit (Thermo Scientific). Nanodrop Nanophotometer™ was used for quantity and quality analysis of the extracted DNA.

Bioinformatic Mega-BLAST **(Zhang et al. 2000)** alignment of the 2672 bp RmlC-like cupins superfamily protein DNA sequence (SeedGeneDB accession code AT2G18540) was performed. BLAST searches were used in nonredundant, nonmouse and nonhuman nucleotide databases by BLATtn against plants (taxid:3193) nucleotide sequences in the NCBI database to alight existing accessions. To analyse the returned alignments for the purposes of primer designations, only those nucleotide

sequences were chosen that posses the query cover more than 75% and E-value 0.0.

Primers were designed in Primer-BLAST **(Ye et al, 2012)** in a manner to get RmlC-like cupins amplification only based on the conservative part of this gene as predicted bioinformatically. Following primers were returned as specific and used in the study: forward 5´ccgaagtttcatccgatggc 3´and reverse 5´ctttgaaagctccccctccg 3´. PCR amplifications were performed in a Bio-Rad C1000™ Thermocycler with the following program: an initial denaturation step at 95 °C for 5 min followed by 40 cycles at 95 °C for 30 s, 58 °C for 40 s, and 72 °C for 40 sec with a final cycle at 72 °C for 10 min. The amplified products were inspected by electrophoresis in 1.5% agarose in a 1×TBE buffer, visualized after GelRed™ staining and photographed under UV light.

RESULTS AND DISCUSSION

First, alignment of *Arabidopsis* RmlC-like cupins sequence was done using megaBLAST. Here, query cover from 10% up to the 100% was found among RmlC-like cupins nucleotides and its homologues that are actually known for taxid 3193 and stored in databases (Figure 1). In total, 64 hits were found in nucleotide database for *Arabidopsis* RmlC-like cupins sequence. Subsequently, conserved region was identified for matches that posses the query cover more than 75% and E-value 0.0 and was returned for the nucleotides in the range of 1506 up to the 1925 bp of RmlC-like cupins DNA. Variable regions of the most conserved part of the RmlC-like cupins sequence of *Arabidopsis* are listed in the Table 1.

Comparing the obtained data for possibility to design primers for RmlC-like cupins PCR identification in plant species, candidates sites were found as displayed in figure 2 for the nucleotides 1533-1552 and 1765-1784. For primer design, Primer-BLAST **(Ye et al, 2012)** software was used. In literature, not only Primer-BLAST, but also similar softwares like FastPCR **(Kalendar et al., 2011)**, or csPCR **(Dasu et al., 2010)** are reported by **(Gardner et al., 2014)** as to be optimal for low throughput analyses for the purposes of manual inspection or a graphical user interface. They design the primers together with primer Tm, secondary structure and primer-dimer prediction.

Bäumlein et al. (1995) reported for cupin superfamily sharing conserved residues with vicilin and legumin storage proteins. This was confirmed by performed BLAST again, as vicilin-like and provicilin-like alignments were found in nucleotide database for *Camelina sativa*, *Morus notabilis*, *Elaeis guineensis*, *Citrus sinensis*, *Vitis vinifera*, *Brassica rapa* and *Tarenaya hassleriana*. None all of these returned alignment share the query cover more than 75% and E-value 0.0 and were not used in comparison for primer design purposes.

Table 1 Characterization of variable nucleotide motifs in the most conserved region of RmlC-like cupins superfamily protein.

Alignment	Variable nucleotides within the motif 1506-1925 nt
Arabidopsis thaliana cupin family protein mRNA	none
Camelina sativa provicilin-like (predicted)	1547, 1655, 1581, 1587, 1598, 1601, 1604, 1606, 1641, 1644, 1647, 1651, 1653, 1660, 1670, 1673-4, 1679, 1696, 1706, 1718, 1722, 1749, 1752, 1757, 1761, 1767, 1773, 1780, 1789, 1809, 1820, 1838, 1840, 1842, 1847, 1851-2, 1858-60, 1876, 1878-9, 1886, 1890, 1894-97, 1903-4, 1906
Camelina sativa globulin-1 S allele-like (predicted)	1547, 1556, 1565, 1580, 1586, 1601, 1604, 1606-7, 1613, 1646, 1650, 1652, 1670, 1671, 1674, 1696, 1706, 1709, 1716, 1718, 1722, 1730, 1748, 1751, 1756, 1760, 1772, 1779, 1789, 1790, 1809, 1820, 1840, 1842, 1852, 1853, 1859, 1860, 1861, 1874-88, 1885-88, 1898, 1899, 1906, 1910
Camelina sativa globulin-1 S allele (predicted)	1547, 1565, 1580, 1586, 1601, 1604, 1606-7, 1611, 1613, 1634, 1646, 1650, 1652, 1659, 1670, 1674, 1696, 1706, 1726, 1728, 1730, 1748, 1751, 1756, 1760, 1772, 1779, 1789, 1790, 1820, 1840, 1842, 1852, 1853, 1859, 1860, 1871, 1875-80, 1888-9, 1892, 1898, 1903, 1904, 1906, 1910
Arabidopsis lyrata subsp. lyrata hypothetical protein*	1547, 1556, 1567, 1594, 1601, 1604-5, 1607, 1613, 1628, 1637, 1640, 1646, 1650, 1652, 1666, 1670, 1647, 1694, 1700, 1706, 1709, 1714, 1716, 1718, 1730, 1772, 1776, 1779, 1786, 1801, 1840, 1845
Brassica rapa uncharacterized LOC103874069 (predicted)	1514, 1517, 1536, 1542, 1547, 1556, 1557, 1567, 1577, 1580-1, 1583, 1592, 1601, 1607, 1613, 1629, 1640, 1656, 1642, 1646, 1650, 1652, 1655, 1661-2, 1647, 1679-81, 1692, 1697-8, 1702, 1707, 1709-10, 1713, 1793, 1810, 1811, 1838, 1842, 1843, 1853-70, 1888, 1889, 1890, 1900-4, 1906, 1916, 1919

* alignment found here only for the nucleotides 1506-1869

In Arabidopsis, analysed RmlC-like cupins superfamily protein is described to have nutrient reservoir activity function (http://sgdb.cbi.pku.edu.cn/gene_info.php?id=AT2G18540). For other alignments found by BLAST in this study, beside above mentioned vicilin and provicilin like characteristics, three types of other characteristics were returned: globulin 1-S, hypothetical protein and uncharacterized one. For most of germin-like proteins isolated from plants are in literature reported unknown function or bifunctionality (Woo et al., 2000).

After the bioinformatic analysis of Arabidopsis RmlC-like cupins sequence, primers were designed for the purposes of its PCR identification in those plants, for which no homologous nucleotides exist in databases. *Amaranthus cruentus* (Ficha cultivar) and a hybride

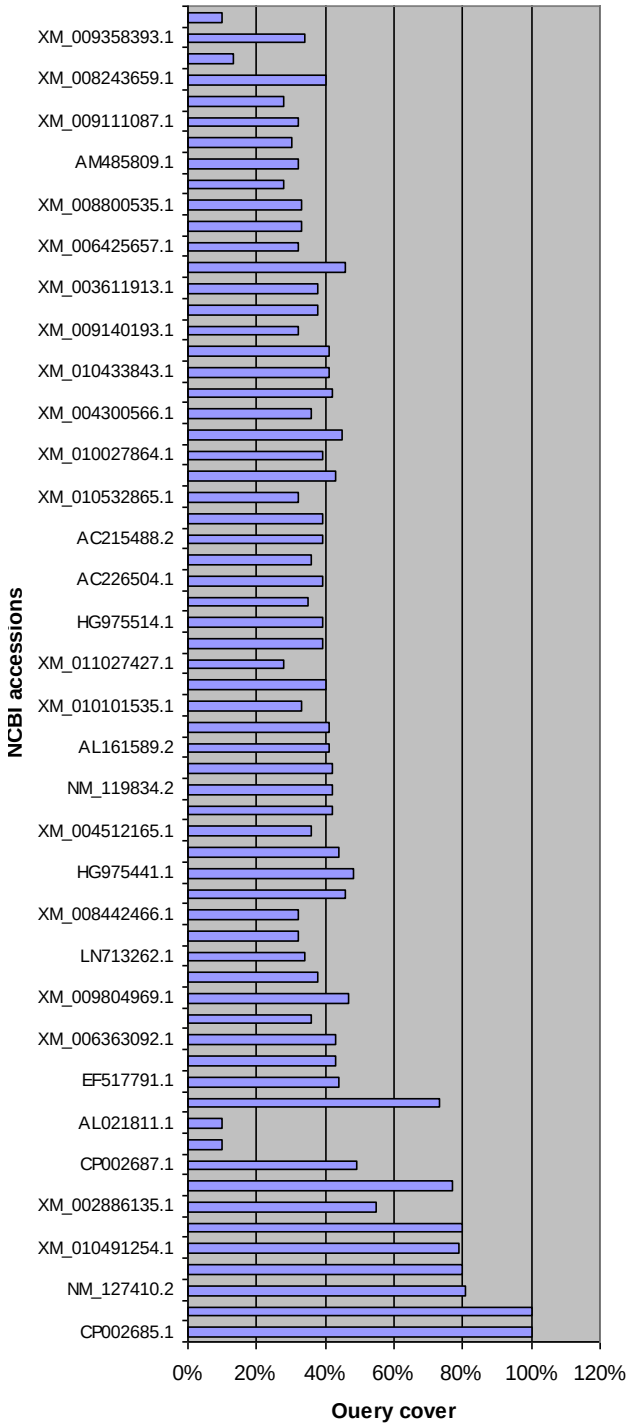

Figure 1 Differences in query covers of returned sequences in the alignment of *Arabidopsis* RmlC-like cupins sequence against taxid taxid:3193.

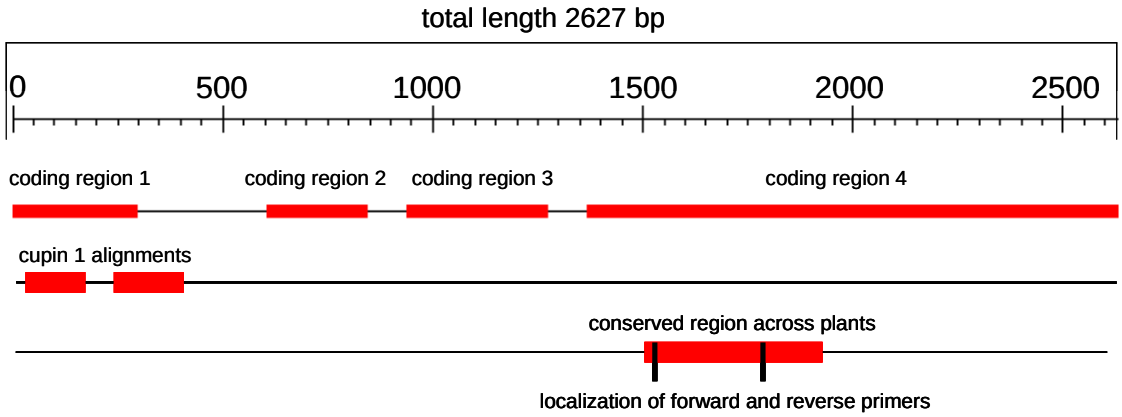

Figure 2 Characteristics of *Arabidopsis* RmlC-like cupins sequence with the position of primers for its identification designed in this study.

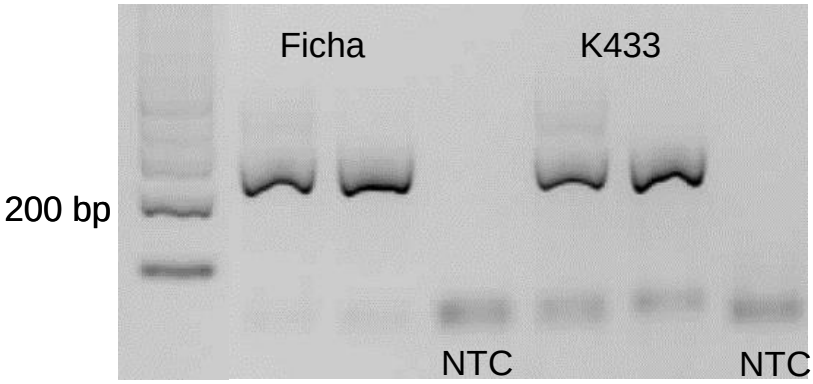

Figure 3 PCR amlification of *Arabidopsis* RmlC-like cupins sequence in *Amaranthus cruentus* and a hybride *Amaranthus hypochondriacus* x *hybridus* based on bioinformatically predicted data.

Amaranthus hypochondriacus x *hybridus* (hybrid K-433) became the biological object for the test of the designed primer pair suitability and amplification efficiency. After optimization of PCR conditions a specific monomorphic fragment with the size of 250 bp it was obtained (Figure 3). This fragment lenght corresponded to the size of the conservative region that is flanked by designed primers in the bioinformatic part of the study. Amplicons were inspected for specifity on 1.5% agarose gel.

Sequence homology search algorithms became commonly used and efficient tools in molecular genetics **(Karpov and Bloom, 2010)**. Nowadays, a number of different motifs finding algorithms are available and **(Lin, http://biochem218.stanford.edu/Projects%202012/Lin.pdf)** reported them to be impossible to provide a comprehensive report of all of them. Each algorithm has its own advantages and disadvantages. One of the aims of different patterns discovery is finding of specific motifs in nucleotide or protein sequences for the purpose of better understanding of their structure and function **(Bailey, 2008)** or for their identification **(Khuri et al., 2001)**.

CONCLUSION

We presented a bioinformatic approach for identification of specific parts of the plant genomes that are known for some species, but not for all. Using nucleotide comparisons based on BLAST analysis offer a tool that can be used for selecting the most conservative sites of known sequences. Based on comparison such as those, universal primers can be designed and used for species, where concrete genomic data are still missing for the appropriate gene.

REFERENCES

Brown, D. R., Southern, L. L. 1985. Effect of citric acid and ascorbic acids on performance and intestinal pH of chicks. *Poultry Science*, vol. 64, no. 7, p. 1390-1401. http://dx.doi.org/10.3382/ps.0641399 PMid:4022914

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, vol. 25, no. 17, p. 3389-3402. http://dx.doi.org/10.1093/nar/25.17.3389 PMid:9254694

Aravind, L., Koonin, E. V. 1999. Gleanin non-trivial structural, functional and evolutionary information about proteins by interactive database searches. *J. Mol. Biology*, vol. 287, no. 5, p. 1023-1040. http://dx.doi.org/10.1006/jmbi.1999.2653 PMid:10222208

Bailey, T. L. 2008. Discovering sequence motifs. *Methods Mol. Biology*. vol. 452, p. 231-251. http://dx.doi.org/10.1007/978-1-60327-159-2_12 PMid:18566768

Bäumlein, H., Braun, H., Kakhovskaya, I. A., Shutov. A. D. 1995. Seed storage proteins of spermatophytes share a

common ancestor with desiccation proteins of fungi. Journal of Molecular Evolution, vol. 41, no. 6, p. 1070-1075. http://dx.doi.org/10.1007/BF00173188 PMid:8587105

Dasu, S., Williams, A., Fofanov, Y., Putonti, C. 2010. csPCR: A computational tool for the simulation of the Polymerase Chain Reaction. *Online Journal of Bioinformatics*, vol. 11, no. 1, p. 34-37. [cit. 2015-03-03] Available at: http://onljvetres.com/cspcrabs2010.htm

Dunwell, J. M. 1998. Cupins: a new superfamily of functionally diverse proteins that include germins and plant storage proteins. *Biotechnol. Genet. Eng. Rev.*, vol. 15, no. 1, p.1-32. http://dx.doi.org/10.1080/02648725.1998.10647950 PMid:9573603

Hertz, G. Z., Stormo, G. D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*, vol. 15, no. 7-8, p. 563-577. http://dx.doi.org/10.1093/bioinformatics/15.7.563 PMid:10487864

Kalendar, R., Lee, D., Schulman, A. H. 2011. Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. *Genomics*, vol. 98, no. 2, p. 137-144. http://dx.doi.org/10.1016/j.ygeno.2011.04.009 PMid:21569836

Karpov, P. A., Nadezhdina, E. S., Yemets, A. I., Matusov, V. G., Nyporko, A. Y., Shashina, N. Y., Blume, Y. B. 2010. Bioinformatic search of plant microtubule-and cell cycle related serine-threonine protein kinases. *BMC Genomics*, vol. 11, Suppl. 1, S14; http://dx.doi.org/10.1186/1471-2164-11-S1-S14

Khuri, S., Bakker, F. T., Dunwell, J. M. 2001. Phylogeny, Function, and Evolution of the Cupins, a Structurally Conserved, Functionally Diverse Superfamily of Proteins. *Molecular Biology and Evolution*, vol. 18, no. 4, p. 593-605. http://dx.doi.org/10.1093/oxfordjournals.molbev.a003840 PMid:11264412

La, D., Livesay, D. R. 2005. Predicting functional sites with an automated algorithm suitable for heterogeneous datasets. *BMC Bioinformatics*, vol. 6, p. 116. http://dx.doi.org/10.1186/1471-2105-6-116 PMid:15890082

Lin, I. 2012. Discovering Transcription Factor Binding Motif Sequences. Bioc218 Final Report. [cit. 2015-03-03] Available at: http://biochem218.stanford.edu/Projects%202012/Lin.pdf

Rasouli, H., Kahrizi, D., Ghadernia, P. 2013. Identification of conserved domains and motifs for *TaWdhn13* gene in *Triticum aestivum* by in silico analysis. *Advances in Environmental Biology*, vol. 7, p. 586-590.

Shea, N., Gardner, S. H., Slezak, T. 2014. Simulate_PCR for amplicon prediction and annotation from multiplex, degenerate primers and probes. *BMC Bioinformatics*, vol. 15, p. 237. http://dx.doi.org/10.1186/1471-2105-15-237 PMid:25005023

Woo, E.-J., Dunwell, J. M, Goodenough, P. W., Marvier, A. C., Pickersgill, R. W. 2000. Germin is a manganese containing homohexamer with oxalate oxidase and superoxide dismutase activities. *Nat. Struct. Biol*. vol. 7, no. 11, p. 1036-1040. http://dx.doi.org/10.1038/80954 PMid:11062559

Ye, J., Coulouris, G., Zaretskaya, I., Cutcutache, I., Rozen, S., Madden, T. L. 2012. Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, vol. 13, p. 134. http://dx.doi.org/10.1186/1471-2105-13-134 PMid:22708584

Zhang, Z., Schwartz, S., Wagner, L., Miller, W. A. 2000. A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, vol. 7, no. 1-2, p. 203-214. http://dx.doi.org/10.1089/10665270050081478 PMid:10890397

Acknowledgment:

This research was co-funded by VEGA project no. 2/0066/13: Exploitation of modern biotechnologies in amaranth breeding programme and European Community project ITMS: 26220220180: Building Research Centre "Agrobiotech".

Contact address:

doc. Ing. PaedDr. Jana Žiarovská, PhD., Slovak University of Agriculture, Faculty of Agrobiology and Food Resources, Department of Genetics and Plant Breeding, Tr. A. Hlinku 2, 949 76 Nitra, Slovakia, E-mail: jana.ziarovska@uniag.sk.

Ing. Michal Záhorský, Institute of Plant Genetics and Biotechnology SAS, Akademická 2, 949 01 Nitra, Slovakia, E-mail: nrgrzahm@savba.sk.

prof. RNDr. Zdenka Gálová CSc., Slovak University of Agriculture in Nitra, Faculty of Biotechnology and Food Sciences, Department of Biochemistry and Biotechnology, Tr. A. Hlinku 2, 949 76 Nitra, Slovakia, E-mail: zdenka.galova@uniag.sk.

Ing. Andrea Hricová, PhD. Slovak Academy of Sciences, Institute of Plant Genetics and Biotechnology, Akademická 2, 949 01 Nitra, Slovakia, E-mail: andrea.hricova@savba.sk.