Data in Brief 9 (2016) 345–348

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Data set for phylogenetic tree and RAMPAGE Ramachandran plot analysis of SODs in *Gossypium raimondii* and *G. arboreum*

Wei Wang, Minxuan Xia, Jie Chen, Fenni Deng, Rui Yuan, Xiaopei Zhang, Fafu Shen *

State Key Laboratory of Crop Biology, College of Agronomy, Shandong Agricultural University, Tai'an 271018, Shandong, PR China

ARTICLE INFO

Article history:
Received 2 March 2016
Received in revised form
3 May 2016
Accepted 14 May 2016
Available online 18 June 2016

Keywords:
Phylogenetic tree
RAMPAGE Ramachandran plot analysis
SOD
Cotton

ABSTRACT

The data presented in this paper is supporting the research article "Genome-Wide Analysis of Superoxide Dismutase Gene Family in *Gossypium raimondii* and *G. arboreum*" [1]. In this data article, we present phylogenetic tree showing dichotomy with two different clusters of SODs inferred by the Bayesian method of MrBayes (version 3.2.4), "Bayesian phylogenetic inference under mixed models" [2], Ramachandran plots of *G. raimondii* and *G. arboreum* SODs, the protein sequence used to generate 3D sructure of proteins and the template accession via SWISS-MODEL server, "SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information." [3] and motif sequences of SODs identified by InterProScan (version 4.8) with the Pfam database, "Pfam: the protein families database" [4].

DOI of original article: http://dx.doi.org/10.1016/j.plgene.2016.02.002
* Corresponding author.
E-mail address: ffshen@sdau.edu.cn (F. Shen).

http://dx.doi.org/10.1016/j.dib.2016.05.025

Specifications Table

Subject area	Biology
More specific subject area	Genetics and Molecular Biology
Type of data	Figure
How data was acquired	Database analysis
Data format	Analyzed
Experimental factors	Amino acid sequences were retrieved from NCBI, TAIR10, Joint Genome Institute (JGI) and/or CottonGen.
Experimental features	Sequences were aligned using BLAST for Proteins (BLASTP), Structural evaluation and stereochemical analyses were assessed using RAMPAGE Ramachandran plot analysis
Data accessibility	With this article

Value of the data

- Data on phylogenies separately estimated using Bayesian method of MrBayes enable researchers to examine how the topologies differ from each other.
- Data on phylogenies of Gossypium SOD proteins enable researchers to infer the possible ranges of time frames in the divergence events of Gossypium *SOD* genes and its molecular evolution in general.
- Data on RAMPAGE Ramachandran plot analysis of Gossypium SOD proteins enable researchers to evaluate the accuracy of the predicted models.

1. Data

The phylogenetic tree obtained using Maximum-Likelihood (ML) method of PhyML (version 20120412) [5] and the 3D structure of SODs generated by using SWISS-MODEL server (http://swissmodel.expasy.org/) [3] and using the online COACH server (http://zhanglab.ccmb.med.umich.edu/COACH/) [6]. were presented in [1]. The data shown here represent the showing dichotomy with two different clusters of SODs (I: Cu/Zn; II: Mn/Fe-SODs) inferred by the Bayesian method of MrBayes (version 3.2.4) [2] and Cu/Zn-SOD cluster had three subgroups (Ia–Ic), whereas the Mn/Fe-SOD cluster had two subgroups (IId and IIe) (Figure. 1). We analysed the accuracy of the predicted models evaluated by Ramachandran plot using the RAMPAGE server (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) [3]. The refined SOD models showed good proportions of residues in favoured, allowed and outlier regions (Figs. 2 and 3). In-depth analyses of the data is presented in the associated research article [1].

2. Experimental design, materials and methods

2.1. Information access

The latest versions of the *G. raimondii* (V1.0) and *G. arboreum* (V2.0) genomes and annotation files were downloaded from CottonGen (https://www.cottongen.org/data/genome). The latest version of the Arabidopsis (TAIR10) genome and annotation files were downloaded from the Joint Genome Institute (JGI) (http://www.phytozome.net).

2.2. Data filtering

We then filtered gene annotation results based on the following criteria [7]: (1) the longest transcript in each gene loci was chosen to represent that locus; (2) coding sequences (CDS) with length <150 base pair bp were filtered out; (3) CDS with the percentage of ambiguous nucleotides ('N') $>50\%$ were filtered out; (4) CDS with internal termination codon were filtered out; and (5) the CDS with hits(Basic Local Alignment Search Tool (BLAST) identity $\geq 80\%$) to RepBase sequences (http://www.girinst.org/repbase/index.html) were filtered out.

2.3. Identification of SOD protein

To identify members of the SOD protein in *G. raimondii* and *G. arboreum*, we retrieved SOD protein sequences from the NCBI protein database (http://www.ncbi.nlm.nih.gov/protein/). These protein sequences from six species, including Arabidopsis (accession nos. NP_172360.1, NP_565666.1, NP_197311.1, NP_199923.1, NP_197722.1 and NP_187703.1), Theobroma cacao (XP_007030135.1 and XP_007038205.1), G. hirsutum (ABA00453.1, ACC93639.1, ABA00454.1, ABA00456.1 and ABA00455.1), Po. trichocarpa (XP_002319589.1 and XP_002325843.1), Z. mays (NP_001105704.1, BAI50563.1, ACG41865.1, ACG32380.1 and NP_001105742.1) and O. sativa (AAA33917.1, BAD09607.1, BAA37131.1 and NP_001055195.1), were used as query sequences to perform multiple database searches using BLAST for Proteins (BLASTP) [8]. After removing alignments with identity $<50\%$, the resultant candidate SOD proteins were aligned to each other to ensure that no gene was represented multiple times. InterProScan (version 4.8) [9]was further used to confirm the inclusion of the SOD domain in each candidate sequence using the Pfam database. Furthermore, we gathered the SOD protein sequences, the template accession and motif sequences.

2.4. Construct phylogenetic trees

Phylogenetic trees were constructed using the Bayesian analysis method. Bayesian trees were constructed using MrBayes (version 3.2.4) [2] with GTR+I+gamma substitution model. The Markov chain Monte Carlo process performed 5,000,000 iterations with sampling every 500 iterations resulting in 10,000 samples and a burn-in of 25% samples. Other parameters were the default settings.

2.5. Structural evaluation and stereochemical analysis

Structural evaluation and stereochemical analyses were assessed using RAMPAGE Ramachandran plot analysis (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) [10].

Acknowledgements

This research was mainly supported by the China Major Projects for Transgenic Breeding (Grant Nos. 2011ZX08005-004 and 2011ZX08005-002) and the China Key Development Project for Basic Research (973) (Grant No. 2010CB12606).

Transparency document. Supporting information

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.05.025.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.dib.2016.05.025.

References

[1] W. Wang, M. Xia, J. Chen, F. Deng, R. Yuan, X. Zhang, F. Shen, Genome-wide analysis of superoxide dismutase gene family in *Gossypium raimondii* and *G. arboretum*, Plant Gene 6 (2016) 18–29.
[2] F. Ronquist, J.P. Huelsenbeck, MrBayes 3: Bayesian phylogenetic inference under mixed models, Bioinformatics 19 (2003) 1572–1574.
[3] M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T.G. Cassarino, M. Bertoni, L. Bordoli, T. Schwede, SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information, Nucleic Acids Res. 42 (2014) W252–W258.
[4] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E. L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, Nucleic Acids Res. 42 (2014) D222–D230.
[5] S. Guindon, J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, O. Gascuel, New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0, Syst. Biol. 59 (2010) 307–321.
[6] J. Yang, A. Roy, Y. Zhang, BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions, Nucleic Acids Res. 41 (2013) D1096–D1103.
[7] T. Ma, J. Wang, G. Zhou, Z. Yue, Q. Hu, Y. Chen, B. Liu, Q. Qiu, Z. Wang, J. Zhang, K. Wang, D. Jiang, C. Gou, L. Yu, D. Zhan, R. Zhou, W. Luo, H. Ma, Y. Yang, S. Pan, D. Fang, Y. Luo, X. Wang, G. Wang, J. Wang, Q. Wang, X. Lu, Z. Chen, J. Liu, Y. Lu, Y. Yin, H. Yang, R.J. Abbott, Y. Wu, D. Wan, J. Li, T. Yin, M. Lascoux, S.P. DiFazio, G.A. Tuskan, J. Wang, J. Liu, Genomic insights into salt adaptation in a desert poplar, Nat. Commun. 4 (2013).
[8] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, T.L. Madden, BLAST+: architecture and applications, BMC Bioinforma (2009), http://dx.doi.org/10.1186/1471-2105-10-421.
[9] E. Quevillon, V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, R. Lopez, InterProScan: protein domains identifier, Nucleic Acids Res. 33 (2005) W116–W120.
[10] S.C. Lovell, I.W. Davis, W.B. Arendall, P.I.W. de Bakker, J.M. Word, M.G. Prisant, J.S. Richardson, D.C. Richardson, Structure validation by Cα geometry: ϕ, ψ and Cβ deviation., Proteins: Struct., Funct., Bioinforma. 50 (2003) 437–450.