Plant Gene 6 (2016) 18–29

Contents lists available at ScienceDirect

Plant Gene

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

Genome-wide analysis of superoxide dismutase gene family in *Gossypium raimondii* and *G. arboreum*

Wei Wang [1], Minxuan Xia [1], 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 22 October 2015
Received in revised form 21 January 2016
Accepted 25 February 2016
Available online 2 March 2016

Keywords:
Cotton
Superoxide dismutase
Genome-wide analysis

ABSTRACT

Superoxide dismutases (SODs) convert highly reactive superoxide radicals to hydrogen peroxide and molecular oxygen, and belong to a class of proteins with important roles in plant responses to stress. Genome-wide analysis was performed in cotton species *Gossypium raimondii* and *Gossypium arboreum* to characterize *SOD* genes and proteins. From the two genomes, 18 *SOD* genes were identified with several bioinformatics tools, and classified into two subfamilies: Cu/Zn-SODs (ten genes) and Mn/Fe-SODs (eight genes). The highest number of *SOD* genes was on chromosome 13 with two genes in *G. raimondii*, and on chromosomes 9, 10 and 13 with two genes on each in *G. arboreum*. Four (50%) *SOD* genes from *Arabidopsis thaliana* had one putative ortholog in *G. raimondii*, and three (37.5%) had one putative ortholog in *G. arboreum*; and eight (88.89%) from *G. arboreum* had one putative ortholog in *G. raimondii*. There were 4–8 introns in *SOD* genes of *G. raimondii* and 5–8 of *G. arboreum*. Phylogenetic analysis revealed that Cu/Zn-SODs (92%) and Mn/Fe-SODs (100%) were separated by high bootstrap value. Tissue-specific expressions of cotton *SOD* genes showed that 9, 10, 9, 8 and 18 of a total of 18 putative *SOD* genes were expressed in root, stem, leaf, flower and ovule, respectively. Stage-specific expression patterns in ovule showed that expressions of *GaFSD1*, *GaMSD2*, *GrMSD1* and *GrMSD2* peaked during the elongation stage and declined coincident with the initiation of secondary cell wall synthesis, and had similar patterns to genes expressed primarily during cell elongation in fiber development. Three-dimensional structures were determined and compared within each cotton SOD protein. These results will improve understanding of *SOD* genes and proteins in these cotton species, and especially in *Gossypium hirsutum*.

1. Introduction

Oxidative stress is the adverse effect of oxidants on physiological function and occurs when abnormally high levels of reactive oxygen species (ROS) are generated during oxygen metabolism (Halliwell and Gutteridge, 1984). Aberrant production of ROS, such as superoxide anion ($O_2{\bullet}^-$), hydroxyl radical ($\cdot$OH), hydrogen peroxide (H_2O_2) and singlet oxygen ($1O_2$) cause irreparable damage to cellular components such as DNA, proteins and lipids that require additional defense mechanisms. Plants have developed efficient mechanisms to cope with ROS toxicity with the enzymatic arsenal of ROS scavengers and non-enzymatic antioxidant defense systems (Gopavajhula et al., 2013). SOD is the first line of enzymatic defenses against oxidative damage in cells, and catalyzes the conversion or dismutation of toxic $O_2{\bullet}^-$ radicals to H_2O_2 and molecular oxygen (O_2) (Gill et al., 2015).

In plants, SODs have been classified into three groups according to the type of prosthetic metals: copper and zinc-containing (Cu/Zn-SODs), manganese-containing (Mn-SODs) and iron-containing (Fe-SODs). Subcellular fractionation studies of plants showed that plants generally contain mitochondrial matrix-localized Mn-SOD, cytosolic Cu/Zn-SOD, Fe-SOD and/or chloroplastic Cu/Zn-SOD and Fe-SOD (Bowler et al., 1994). Plant SODs are known to contribute to defense against toxic oxygen species and are consequently important for stress tolerance. For example, Transgenic tobacco plants that overexpressed a tomato chloroplast SOD exhibited greater resistance to virus-induced hypersensitive necrosis (Viczián et al., 2014). Transgenic *Arabidopsis*, overexpressing a novel *Cu/Zn-SOD* gene (*p35S:JcCu/Zn-SOD*) cloned from *Jatropha curcas*

Abbreviations: Aa, amino acid; A, Ala, alanine; At, *Arabidopsis thaliana*; C, cytoplasm; CDS, coding sequences; CP, chloroplast; CSD, Cu/Zn-SOD, copper and zinc-containing SOD; Cys, cysteine; CZ, Cu/Zn-superoxide dismutase domain; D, Asp, aspartate; DPA, day post anthesis; EST, expressed sequence tags; F, Phe, phenylalanine; FSD, Fe-SOD, iron-containing SOD; G, Gly, glycine; Gr, *Gossypium raimondii*; Ga, *Gossypium arboreum*; H_2O_2, hydrogen peroxide; His, histidine; IMA, Fe/Mn-SODs alpha-hairpin domain; IMC, Fe/Mn-SODs C-terminal domain; MAPK, mitogen-activated protein kinase; ML, the maximum-likelihood method; MSD, Mn-SOD, manganese-containing SOD; MT, mitochondrion; MW, theoretical molecular weight of proteins; NAC, named from NAM, ATAF1/2 and CUC2; $O_2{\bullet}^-$, superoxide anion; $1O_2$, singlet oxygen; $\cdot$OH, hydroxyl radical; PCW, primary cell wall; *pI*, theoretical isoelectric point of proteins; P, peroxisome; Q, Gln, glutamine; R, Arg, arginine; ROS, reactive oxygen species; S, Ser, serine; SCW, secondary cell wall; SOD, superoxide dismutase; V, vacuole; W, Trp, tryptophane; Y, Tyr, tyrosine; ZF, C2H2-type zinc finger.

* Corresponding author.
E-mail address: ffshen@sdau.edu.cn (F. Shen).

[1] The authors who contribute equally.

http://dx.doi.org/10.1016/j.plgene.2016.02.002

L., had enhanced tolerance to salt stress during germination, seedling establishment and growth in terms of longer root and larger rosette area compared with wild type (Liu et al., 2014). These studies added more evidence of the role of *SOD* genes in development and response to abiotic/biotic stresses in different species, but related genome-wide resources are little known.

SODs are encoded by a small multigene family, and the first plant *SOD* gene cloned was from *Zea mays* (Cannon et al., 1987). Extensive studies have been carried out on the *SOD* multigene family in many plant species, including *Z. mays* (Scandalios, 1997), *Arabidopsis thaliana* (Kliebenstein et al., 1998), *Oryza sativa* (Feng et al., 2006; Kaminaka et al., 1999), *Lotus japonicus* (Rubio et al., 2007), *Gossypium hirsutum* (Kim et al., 2008), *Picea sitchensis* (Ralph et al., 2008b), *Populus trichocarpa* (Ralph et al., 2008a), *Nelumbo nucifera* (Dong et al., 2011), *Haberlea rhodopensis* (Apostolova et al., 2012), *Dimocarpus longan* (Lin & Lai, 2013), *Musa acuminata* (Feng et al., 2015) and *Sorghum bicolor* (Filiz and Tombuloğlu, 2015). The most detailed studies have been conducted in *Z. mays*, *Arabidopsis*, *D. longan* and *S. bicolor*. In *Z. mays*, the *SOD* multigene family consists of at least nine different isoforms, of which six are Cu/Zn-SODs (*Sod1*, *Sod2*, *Sod4*, *Sod4A*, *Sod5*, and *Sod9*), one is a Fe-SOD (*SodB*) and four are Mn-SODs (*Sod-3.1*, *Sod-3.2*, *Sod-3.3*, and *Sod-3.4*) (Scandalios, 1997). Seven *SOD* genes were identified in *Arabidopsis*, including one Mn-SOD (*MSD1*), three Fe-SOD isoforms (*FSD1-FSD3*) and three Cu/Zn-SODs (*CSD1-CSD3*), and *CSD2* and Fe-SOD proteins were also observed in *Arabidopsis* chloroplast (Kliebenstein et al., 1998). In *D. longan*, the *SOD* multigene family consists of cytoplasmic *CSD1a* and *DlCSD1b*, chloroplastic *DlCSD2a* and *DlFSD1a*, and plastidic *DlFSD1b* (Lin and Lai, 2013). In *S. bicolor*, the *SbSOD* multigene family included three members of Cu/Zn-SODs (*SbSOD1*, *SbSOD2* and *SbSOD5*) localized in the cytoplasm, and two Cu/Zn-SODs (*SbSOD3* and *SbSOD4*) localized in chloroplast. Furthermore, two members (*SbSOD6* and *SbSOD8*) and one (*SbSOD7*) member of Mn/Fe-SODs were localized in mitochondria and chloroplast, respectively (Filiz and Tombuloğlu, 2015).

Cotton is an economically important crop grown worldwide as a source of fiber and edible oil (Wendel et al., 2009). SODs are ubiquitous proteins that are believed to play important roles in cotton development and stress tolerance. However, no genome-wide information on the *SOD* gene family is currently available in cotton. *Gossypium raimondii* and *Gossypium arboreum* are diploid cotton species and release of their genomes enabled us to identify and analyze the family of *SOD* genes in these species. There are a few reports on studies of cotton transcriptomics (Hovav et al., 2008). Here, we looked at expression of the different SODs in transcriptome data obtained at the National Center for Biotechnology Information (NCBI) (Altenhoff and Dessimoz, 2009). In this study, we characterized a comprehensive and non-redundant set of *SOD* genes at a genome-wide scale in *G. raimondii* and *G. arboreum* with some bioinformatics tools. Furthermore, predicted three-dimensional (3D) structure modeling, subcellular localization and physiochemical properties of the SOD proteins of the two cotton species were evaluated.

2. Materials and methods

2.1. Identification of SOD genes

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). We then filtered gene annotation results based on the following criteria (Ma et al., 2013): (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 ≥ 80%) to RepBase sequences (http://www.girinst.org/repbase/index.html) were filtered out. To identify members of the SOD gene family in *Arabidopsis*, *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) (Camacho et al., 2009). 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) (Quevillon et al., 2005) was further used to confirm the inclusion of the SOD domain in each candidate sequence using the Pfam database.

Physicochemical characteristics of SOD proteins were calculated using the ProtParam tool (http://www.expasy.org/tools/protparam.html), containing the number of amino acids, molecular weight, and theoretical isoelectric point (*pI*). Subcellular localization was analyzed using the CELLO version 2.5 (http://cello.life.nctu.edu.tw/) server. *SOD* genes data, including accession number, chromosomal location and ORF length were collected from the JGI database.

2.2. Phylogenetic analysis

The full-length *SOD* genes encoding sequences of *Arabidopsis*, *G. raimondii* and *G. arboreum* were aligned with the query CDS previously mentioned using PRANK (version 140,603) (Löytynoja and Goldman, 2008) with default settings under the codon model. These multiple sequence alignments were used for subsequent molecular evolutionary analyses.

Phylogenetic trees were constructed using the maximum-likelihood (ML) method of the PhyML (version 20,120,412) (Guindon et al., 2010) with GTR + I + gamma substitution model and a bootstrap value of 1000. The reliability of the obtained trees was tested using the Bayesian analysis method. Bayesian trees were constructed using MrBayes (version 3.2.4) (Ronquist and Huelsenbeck, 2003) 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.

2.3. Gene structure and domain analysis

All candidate SOD protein sequences (see Table 1 in Ref [Wang et al., 2016]) were analyzed using InterProScan (version 4.8) with the Pfam database (Finn et al., 2014). Exon–intron structure information of these *SOD* genes was parsed from the Generic Feature Format (GFF) file downloaded along with the genomic data. Gene structures of the *SOD* genes were generated using home-made PERL scripts.

2.4. Syntenic blocks and genome duplication identification

Synteny analysis was conducted locally using a method similar to that developed for the Plant Genome Duplication Database (http://chibba.pgml.uga.edu/duplication/) (Tang et al., 2008). We used program BLAST version 2.2.9 (Altschul et al., 1990) for the pairwise comparison of the filtered SOD protein sets of *Arabidopsis*, *G. raimondii* and *G. arboreum*. Then, MCscanX (Y. Wang et al., 2012) was employed to identify homologous regions, and syntenic blocks were evaluated using Circos-0.64 (Krzywinski et al., 2009). Default parameters were

Table 1
The details of *SOD* genes and proteins, containing physiochemical, structural and sequence properties.

Gene name	Sequence ID	Genomic position	ORF length (bp)	Length (aa)	MW (kDa)	*pI*	Subcellular prediction by CELLO	Predicted Pfam domain
AtCSD1	AT1G08830.1	Chr1:2,827,061.2829315 (+)	1354	152	15.10	5.38	C	CZ
AtCSD2	AT2G28190.1	Chr2:12,014,498.12016569 (+)	1756	216	22.24	7.02	CP	CZ
AtCSD3	AT5G18100.1	Chr5:5,987,187.5988885 (+)	1486	164	16.94	7.73	P, V	CZ
AtMSD1	AT3G10920.1	Chr3:3,417,954.3419853 (+)	1567	231	25.44	8.84	MT	IMA, IMC
AtMSD2	AT3G56350.1	Chr3:20,893,946.20895625 (−)	1471	241	26.89	6.76	MT	IMA, IMC
AtFSD1	AT4G25100.1	Chr4:12,884,300.12886695 (−)	1853	212	23.79	6.51	CP	IMA, IMC
AtFSD2	AT5G51100.1	Chr5:20,773,296.20775701 (−)	2279	305	34.66	4.60	CP	IMA, IMC
AtFSD3	AT5G23310.1	Chr5:7,850,523.7852532 (+)	1618	263	30.36	8.70	CP	IMA, IMC
GrCSD1	Cotton_D_gene_10024020	scaffold149:2,141,278.2144241 (−)	459	152	15.11	5.92	C	CZ
GrCSD2	Cotton_D_gene_10006229	Chr13:13,535,426.13536512 (−)	462	153	15.34	5.65	C	CZ
GrCSD3	Cotton_D_gene_10039927	Chr7:13,570,863.13572956 (−)	468	155	15.95	6.82	C	CZ
GrCSD4	Cotton_D_gene_10006544	Chr6:24,009,719.24015344 (−)	1404	467	49.57	6.23	CP	CZ, ZF
GrCSD5	Cotton_D_gene_10032111	Chr9:11,503,418.11505248 (−)	601	211	22.39	6.40	CP	CZ
GrMSD1	Cotton_D_gene_10016246	Chr1:22,165,427.22168170 (+)	683	230	25.70	7.14	MT	IMA, IMC
GrMSD2	Cotton_D_gene_10018648	Chr11:35,302,138.35304979 (−)	686	231	26.01	8.81	MT	IMA, IMC
GrFSD1	Cotton_D_gene_10009356	scaffold141:134,027.136904 (−)	930	309	35.53	4.85	CP	IMA, IMC, IMC
GrFSD2	Cotton_D_gene_10030063	Chr13:27,683,350.27685860 (−)	756	251	28.82	5.64	CP	IMA, IMC
GaCSD1	Cotton_A_21978	chr8:126,578,106–126,579,386 (−)	609	202	20.89	5.73	C	CZ
GaCSD2	Cotton_A_24238	chr13:34,195,003–34,196,511 (+)	459	152	15.34	5.30	C	CZ
GaCSD3	Cotton_A_30467	chr4:58,314,910–58,317,769 (+)	456	151	15.48	6.82	C	CZ
GaCSD4	Cotton_A_32487	chr10:66,118,430–66,120,830 (−)	645	214	22.10	6.02	CP	CZ
GaCSD5	Cotton_A_36793	chr9:1,110,059–1,112,675 (−)	675	224	23.23	6.48	CP	CZ
GaMSD1	Cotton_A_04050	Chr10:81,012,722.81015489 (−)	693	230	25.70	7.14	MT	IMA, IMC
GaMSD2	Cotton_A_21263	Chr9:3,364,759.3366909 (−)	696	231	25.94	8.50	MT	IMA, IMC
GaFSD1	Cotton_A_03623	Chr1:66,936,792.66939676 (+)	930	300	35.65	4.84	CP	IMA, IMC
GaFSD2	Cotton_A_26478	Chr13:2,099,667.2102408 (−)	771	256	29.33	6.17	CP	IMA, IMC

At: *Arabidopsis thaliana*; Gr: *Gossypium raimondii*; Ga: *Gossypium arboreum*; CSD: Cu/Zn-SOD; FSD: Fe-SOD; MSD: Mn-SOD; aa: amino acid; *pI*: theoretical isoelectric point of proteins; MW: theoretical molecular weight of proteins; C: cytoplasm; CP: chloroplast; MT: mitochondrion; P: peroxisome; V: vacuole; CZ: Cu/Zn-superoxide dismutase (SOD), IMA: Fe/Mn-SODs, alpha-hairpin domain, IMC: Fe/Mn-SODs, C-terminal domain; ZF: C2H2-type zinc finger.

used in all steps. Tandem duplication was characterized as multiple genes of one family located within the same or neighboring intergenic region (Du et al., 2013).

2.5. Gene expression analyses

The tissue-specific expression patterns of cotton *SOD* genes were analyzed in the NCBI-EST database (http://www.ncbi.nlm.nih.gov/dbEST/) using the MEGABLAST tool. Searching parameters were maximum identity > 95%, length > 200 bp and E value < 10^{-10}.

The stage-specific expression pattern of the *SOD* genes was analyzed using whole transcriptome sequencing data from cotton ovules at various stages (i.e. 10, 20, 30 and 40 DPA; DPA: day post anthesis) of *G. raimondii* and *G. arboreum*, and every stage had three respective replicates. These data were obtained from the NCBI Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/bioproject/PRJNA179447). Accession numbers for *G. raimondii* were: SRX204399, SRX204400, SRX204401, SRX204405, SRX204406, SRX204407, SRX2044529, SRX204530, SRX204531, SRX204532, SRX204533, and SRX204534; and for *G. arboreum* were SRX204558, SRX204557, SRX204556 and SRX204555.

We then used fastq-dump from SRAToolkit.2.4.5–2-centos_linux64 (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=software) to convert the aforementioned SRA data into fastq format. These data were then filtered by Trimmomatic-0.32 (Bolger, Lohse, & Usadel, 2014), including the removal of the adapter and the interception and removal of some low quality bases in leading and trailing sequences. Parameters of filtration were adopted as: SE-phred33 ILLUMINACLIP:TruSeq3-SE.fa:2:30:10 HEADCROP:15 CROP:80 LEADING:5 TRAILING:5 SLIDINGWINDOW:4:15 MINLEN:36. We found that the data size of each duplicate was small after filtration, so the three duplicates of each stage were pooled into one dataset. TopHat2 (D. Kim et al., 2013) (http://ccb.jhu.edu/software/tophat) pipeline was used to respectively align clean reads generated from the above steps to their reference genomes. These alignments were then subject to cufflinks pipeline to assemble the transcripts and perform rigorous statistical analyses. Cufflinks (Trapnell et al., 2012) (http://cufflinks.cbcb.umd.edu/) was used to calculate the expression level of each transcript, namely the FPKM value (fragments per kilobase per million reads). Finally, the FPKM values of these SOD candidates were extracted and plotted using the R programming language.

2.6. Predicted 3D structure of SODs

The predicted 3D structures of SODs were generated using the SWISS-MODEL server (http://swissmodel.expasy.org/) (Biasini et al., 2014). The reliability of the predicted 3D structures details including predicted binding sites and conserved residues were tested using the online COACH server (http://zhanglab.ccmb.med.umich.edu/COACH/) (Yang et al., 2013a, 2013b). Structural evaluation and stereochemical analyses were assessed using RAMPAGE Ramachandran plot analysis (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php) (Lovell et al., 2003).

3. Results and discussion

3.1. Identification of SOD genes in Arabidopsis, G. raimondii and G. arboreum

To characterize *SOD* genes, all known *Arabidopsis*, *G. hirsutum*, *Po. trichocarpa*, *Z. mays* and *O. sativa* SOD protein sequences were used as query sequences to perform genome-wide searches using BLASTP.

In *Arabidopsis*, we detected eight *SOD* genes: three Cu/Zn-SODs, two Mn-SODs and three Fe-SODs. Physicochemical analysis of SOD proteins revealed that the length, molecular weight and *pI* values of SOD proteins were within the ranges of 152–305 amino acids, 15.10–34.66 kDa and 4.60–8.84, respectively. The Cu/Zn-SODs had variable *pI* values: two members (*AtCSD2* and *AtCSD3*) were slightly basic and one (*AtCSD1*) was acidic in character; of Mn-SODs, one member (*AtMSD1*) was basic and one (*AtMSD2*) was acidic in character; of Fe-SODs, one member (*AtFSD3*) was basic and two members (*AtFSD1* and *AtFSD2*) were acidic. The subcellular localizations of SOD proteins showed that three members of Cu/Zn-SODs were localized in the cytoplasm, chloroplast and peroxisome/vacuole, respectively. Furthermore, two members of Mn-SODs and three of Fe-SODs were localized in mitochondria and chloroplast, respectively (Table 1).

In *G. raimondii*, we detected nine *SOD* genes including five Cu/Zn-SODs, two Fe-SODs and two Mn-SODs. The length, molecular weight and *pI* values of SOD proteins were within the ranges of 152–467 amino acids, 15.11–49.57 kDa, and 4.85–8.81, respectively. All Cu/Zn-SODs and Fe-SODs were acidic, and all Mn-SODs were basic in character. The prediction of subcellular localizations showed that three members of Cu/Zn-SODs (*GrCSD1*–*GrCSD3*) were localized in the cytoplasm, and two (*GrCSD4* and *GrCSD5*) in chloroplasts. Furthermore, two members of Mn-SODs and two of Fe-SODs were localized in mitochondria and chloroplasts, respectively (Table 1).

In *G. arboreum*, nine *SOD* genes were determined five Cu/Zn-SODs, two Fe-SODs and two Mn-SODs. The length, molecular weight and *pI* values of SOD proteins were within the ranges of 151–300 amino acids, 15.34–35.65 kDa and 4.84–8.50, respectively (Table 1). All Cu/Zn-SODs and Fe-SODs were acidic, and all Mn-SODs were basic in character. Prediction of subcellular localizations showed that three members (*GaCSD1*-*GaCSD3*) of Cu/Zn-SODs were localized in the cytoplasm, and two (*GaCSD4* and *GaCSD5*) in chloroplasts. Furthermore, two members of Mn-SODs and two of Fe-SODs were localized in mitochondria and chloroplasts, respectively (Table 1).

3.2. Phylogenetic analysis

In the present study, the phylogenetic relationships of *G. raimondii* and *G. arboreum SOD* genes were evaluated with respect to *Arabidopsis SOD* genes using the ML approach implemented in PhyML (Fig. 3A). There were two data on each branch node, the upper (50–100) is the ML bootstrap support rate (percentage), the lower (0–1) are the probability values after testing using the Bayesian analysis method. Phylogenetic analysis revealed that two major groups were obtained with Cu/Zn and Mn/Fe-SODs. The Cu/Zn-SOD cluster had three subgroups (a–c), whereas the Mn/Fe-SOD cluster had two subgroups (d and e). In this tree, cotton Cu/Zn-SODs (*GaCSD1* and *GrCSD1*, and *GaCSD2* and *GrCSD2*) and *Arabidopsis* cytoplasmic Cu/Zn-SOD (*AtCSD1*) were clustered in subgroup-a with 97% bootstrap values. Subgroup-b contained cotton Cu/Zn-SODs (*GaCSD3* and *GrCSD3*) and *Arabidopsis* peroxisomal Cu/Zn-SOD (*AtCSD3*) with 100% bootstrap values. Subgroup-c contained cotton Cu/Zn-SODs (*GaCSD4* and *GrCSD4*, and *GaCSD5* and *GrCSD5*) and *Arabidopsis* chloroplast Cu/Zn-SOD (*AtCSD2*) with 98% bootstrap values. Cotton Mn-SODs (*GaMSD1* and *GrMSD1*, and *GaMSD2* and *GrMSD2*) and *Arabidopsis* mitochondrial Mn-SOD (*AtMSD1* and *AtMSD2*) were clustered in subgroup-d with 52% bootstrap values. Subgroup-e contained cotton Fe-SODs (*GaFSD1* and *GrFSD1*, and *GaFSD2* and *GrFSD2*) and *Arabidopsis* chloroplast Fe-SODs (*AtFSD1* and *AtFSD2*, *AtCSD3*) with 96 and 99% bootstrap values, respectively (Fig. 3A). This result shows that these SODs were closely related to each other and that the sequences retrieved were accurate.

Interestingly, predicted subcellular localization data did not support the phylogenetic data regarding *GaCSD3* and *GrCSD3*. We expected to find them in subgroup-a or -c, but they were located in peroxisomal Cu/Zn-SOD in subgroup-b. Additionally, the bootstrap value of subgroup-d was 52%, which was lower than other values. Various reports have demonstrated that Mn- and Fe-SODs are separated in plants, and plant Mn-SODs have 70% homology with plant Fe-SODs, suggesting different ancestral gene origination (Fink and Scandalios, 2002; Miller, 2012). The ML tree showed that the *SOD* genes identified here were important and deserve further investigation.

For phylogeny reconstruction, a total of 50 sequences of Cu/Zn-, Mn- and Fe-SODs from different plant species were selected and aligned in PRANK. The consensus tree generated by ML method showed dichotomy with two distinct clusters: I and II (Fig. 1). Cu/Zn-SOD sequences fell in cluster I whereas Mn- and Fe-SODs were in cluster II. Three sub-clusters (Ia–Ic) existed within cluster I and there were two sub-clusters (IIa and IIb) in cluster II, indicating evolution of the enzyme in different plants. In near relatives, cotton and cacao, both belonging to the Malvales, were clustered together (*TcCSD1*, *GrCSD1* and *GaCSD1*; and *TcFSD1*, *GrFSD1* and *GaFSD1*); *Z. mays*, *O. sativa* and *Arabidopsis* belonging to the Poaceae and Brassicaceae were clustered together (*ZmCSD1*, *OsCSD1* and *AtCSD1*; *ZmCSD2*, *OsCSD2* and *AtCSD2*; *ZmMSD1*, *OsMSD1* and *AtMSD1*; and *ZmFSD2*, *OsFSD1* and *AtFSD3*). The results showed that the phylogenetic relationship of SOD in different plants had distinct characteristics for particular species or genera. The reliability of the obtained ML trees was tested using the Bayesian analysis method (see Fig. 1 in Ref [Wang et al., 2016]). Results of the phylogeny analysis indicated separate evolution of Cu/Zn-SODs from that of Mn- and Fe-SODs, which may have evolved from the same ancestral enzyme.

3.3. Chromosomal distributions and duplications of SOD genes

Chromosomal distributions of *SOD* genes were determined. The eight *SOD* genes of *Arabidopsis* were distributed across all five chromosomes: one, one, two, one and three genes on chr1, chr2, chr3, chr4 and chr5, respectively (Table 1). The highest number of *SOD* genes was three on chromosome 5. In *G. raimondii*, the genes were segregated as follows: one, one, one, one, one, two, one and one gene on chr1, chr6, chr7, chr9, chr11 and chr13 and scaffold141 and scaffold149, respectively (Table 1). The highest number of *SOD* genes was two on chr13. Additionally, chromosomes 2, 3, 4, 5, 8 and 12 did not contain any *SOD* genes. In *G. arboreum*, the genes were segregated as follows: one, one, one, two, two and two genes on chr1, chr4, chr8, chr9, chr10 and chr13, respectively (Table 1). The highest number of *SOD* genes was on chr9, chr10 and chr13 with two genes each, while all others had one gene member. Additionally, chromosomes 2, 3, 5, 6, 7, 11 and 12 did not contain any *SOD* genes.

Gene duplication can be a crucial factor for diversification. Duplicated genes appear by regional genomic events or genome-wide events (polyploidization) (Lawton-Rauh, 2003). Functional divergences in duplicated genes also contribute to molecular innovation in higher organisms (Ganko et al., 2007). In addition, recent studies have shown that *G. raimondii* and *G. arboreum* experienced the hexaploidization event (γ-WGD) shared by the eudicots, as well as a cotton-specific whole-genome duplication (Li et al., 2014; K. Wang et al., 2012). To analyze the relationship between the *SOD* genes and genome-wide duplications, we mapped the *SOD* genes to the duplicated blocks. Based on gene duplication analysis, one (*GrMSD1* and *GrMSD2*) and two (*AtMSD1* and *AtMSD2*, and *AtFSD1* and *AtFSD2*) segmental duplication events were identified in *G. raimondii* and *Arabidopsis*, respectively (Fig. 2). Segmental duplications may have played an important role in the expansion of *SOD* genes in *Arabidopsis* and *G. raimondii* genomes. The segmental duplication events may provide support for finer regulation of SOD activities by functional divergences in various stress conditions and for expression in different plant structure by protein separation during various growth and developmental stages (Bindschedler et al., 2008; Rajjou et al., 2008).

We also examined the orthologous relationships between *SOD* genes from *Arabidopsis*, *G. raimondii* and *G. arboreum*, given that orthologs often retain equivalent functions during the course of evolution (Altenhoff and Dessimoz, 2009). We found that four (50%) *SOD* genes from *Arabidopsis* had one putative ortholog in *G. raimondii*: *AtCSD3* and *GrCSD3*, *AtFSD1* and *GrFSD1*, *AtFSD2* and *GrFSD1*, and *AtFSD3* and *GrFSD2*. Three (37.5%) from *Arabidopsis* had one putative ortholog in *G. arboreum*: *AtFSD1* and *GaFSD1*, *AtFSD2* and *GaFSD1*, and *AtFSD3* and *GaFSD2*. Eight (88.89%) from *G. arboreum* had one putative ortholog in *G. raimondii*: *GaCSD1* and *GrCSD1*, *GaCSD2* and *GrCSD2*, *GaMSD1* and *GrMSD1(2)*, *GaMSD2* and *GrMSD2*, *GaFSD1* and *GrFSD1*, and *GaFSD2* and *GrFSD2* (Fig. 2). Of these, *GrFSD1* was an ortholog of *GaFSD1* in *G. arboreum* and *AtFSD1(2)* in *Arabidopsis*. *GrFSD2* was an ortholog of *GaFSD2* in *G. arboreum* and *AtFSD3* in *Arabidopsis* (Fig. 2).

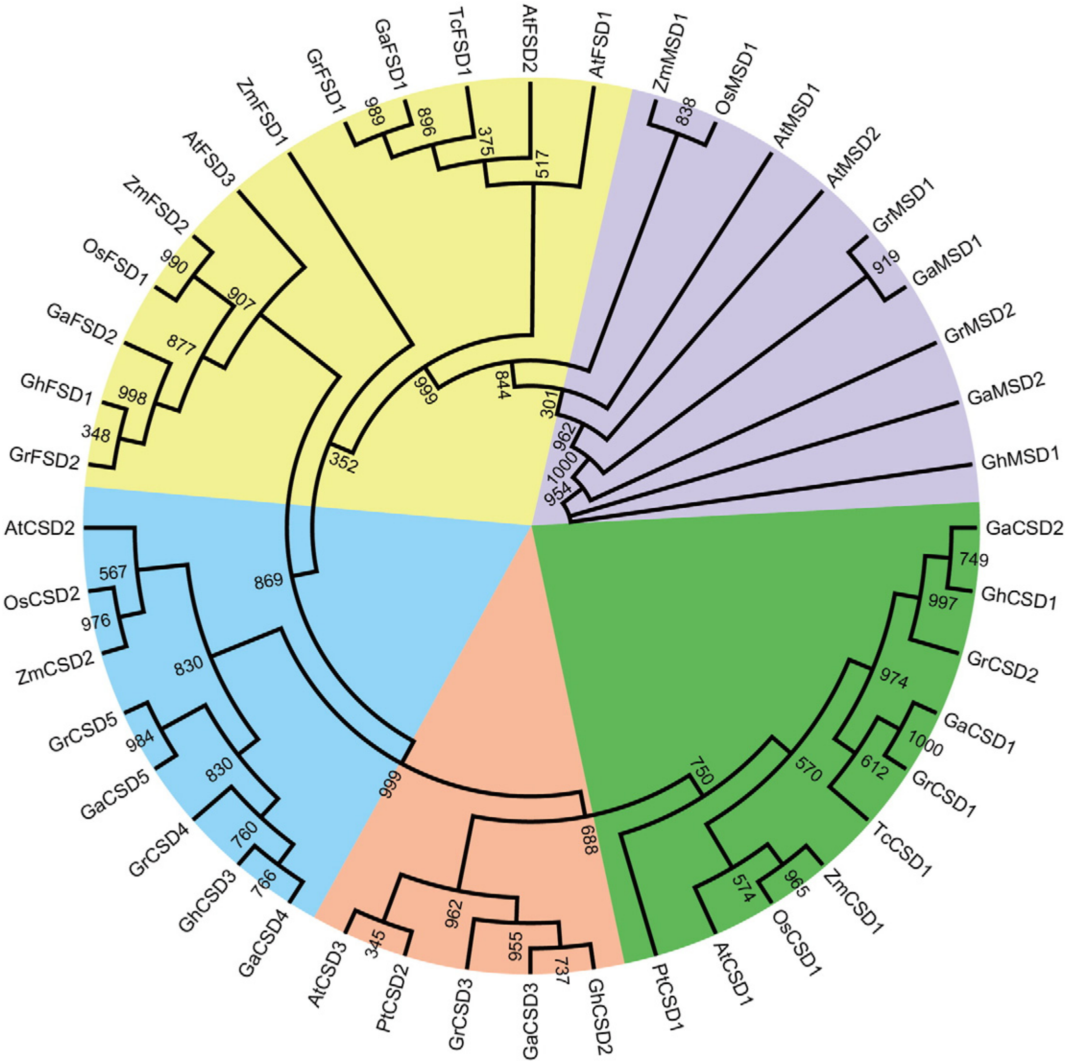

Fig. 1. Phylogenetic tree showing dichotomy with two different clusters of superoxide dismutases (SODs) inferred by Maximum-Likelihood method of PhyML and using 1000 bootstrap value. Clusters Ia, Ib, Ic, IIa and IIb are indicated in green, orange, blue, purple and yellow, respectively.

3.4. Gene structure and domain analysis

Exon–intron structure information of these *SOD* genes was parsed from the GFF file downloaded along with the genomic data. Schematic structures of *SOD* genes were generated using the home-made PERL scripts (Fig. 3B). Gene structure analysis revealed that the ORF length of *SOD* genes in *Arabidopsis* was within the range of 456–2279 bp. Intron numbers were 5–8, with the highest number of introns in *AtFSD2* and the lowest number in *AtMSD1* and *AtMSD2*. In *G. raimondii*, the ORF length was 459–1404 bp and intron numbers were 4–8, with the highest number of introns in *GrFSD1* and the lowest in *GrCSD1*. In *G. arboreum*, the ORF length was 456–930 bp and intron numbers were 5–8, with the highest number of introns in *GaFSD1* and the lowest number in *GaMSD1* and *GaMSD2*. Fink and Scandalios (Fink and Scandalios, 2002) reported that plant SODs have highly conserved intron patterns, with different numbers of introns between the genes encoding cytosolic and chloroplastic Cu/Zn-SODs. They found that chloroplastic *SODs* had an extra intron located in the corresponding second exon of the cytosolic genes. In our analysis, not only was there one more introns in chloroplastic *SODs* in *G. raimondii* than in cytosolic genes, but the location was not in the corresponding second exon of the cytosolic genes. Thus, our findings did not corroborate the findings of Fink and Scandalios (2002). Divergences of exon–intron structures are shaped by three main mechanisms: exon/intron gain/loss, exonization/pseudoexonization, and insertion/deletion (Xu et al., 2011). The structural divergences of *SOD* genes were affected by these mechanisms in the *SOD* evolution of *Arabidopsis*, *G. raimondii* and *G. arboreum*. These structural divergences may be related to enzyme function that responds to various biotic and abiotic stress conditions with expression pattern divergences. In addition, we found that the same types of *SOD* genes in *G. raimondii* and *G. arboreum* had the same or similar gene structures (Fig. 3B).

The candidate protein sequences (see Table 1 in Ref [Wang et al., 2016]) were analyzed for the presence of a SOD domain and 4 types of motifs were identified as shown in the combined block diagram (Fig. 3C). Furthermore, motif sequences and their domain patterns were shown in Table 2 in Ref (Wang et al., 2016). Based on the domain analysis, among all SOD sequences of *G. raimondii* and *G. arboreum*, Cu/Zn-SODs had a copper/zinc superoxide dismutase domain (Pfam: 00,080) and conserved Cu ion and Zn ion binding sites (Table 3, see Table 2 in Ref [Wang et al., 2016]), and Mn- and Fe-SODs both had an iron/manganese superoxide dismutase alpha-hairpin domain (Pfam: 00,081) and an iron/manganese superoxide dismutase C-terminal domain (Pfam: 02,777). The Mn- and Fe-SODs domain included the conserved metal-binding domain “DVWEHAYY” (see Table 2 in Ref

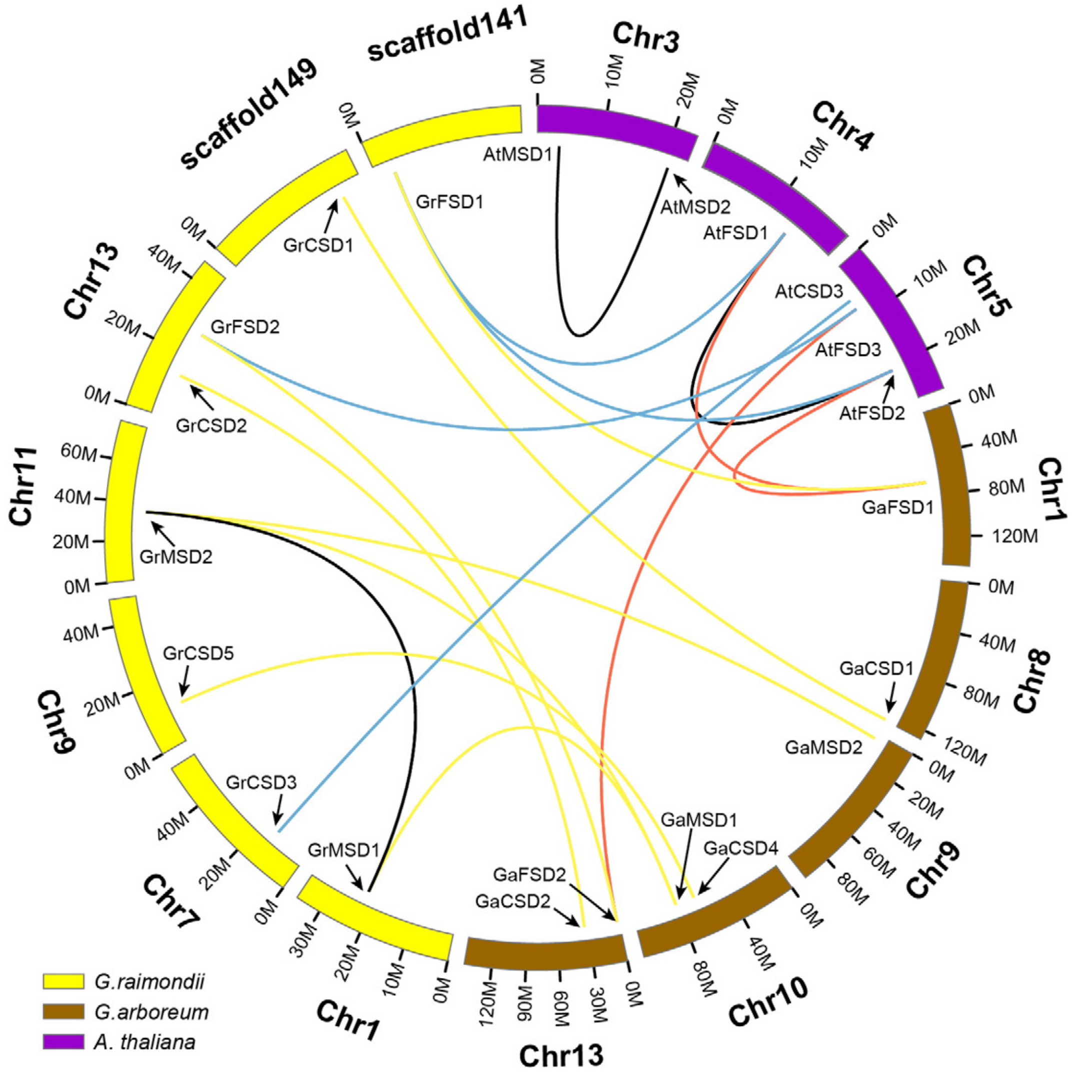

Fig. 2. Syntenic relationships between *SOD* genes from *Arabidopsis*, *G. raimondii* and *G. arboreum*. *Arabidopsis*, *G. raimondii* and *G. arboreum* chromosomes are indicated in purple, yellow, and brown, respectively. The putative orthologous *SOD* genes between *Arabidopsis* and *G. raimondii*, *Arabidopsis* and *G. arboreum*, and *G. raimondii* and *G. arboreum* are connected by blue, red and yellow lines, respectively.

[Wang et al., 2016]), and conserved Mn ion or Fe ion binding sites were showed in Table 3. Five residues were specific for Mn-SODs (Gly, Gly, Phe, Gln and Asp) and Fe-SODs (Ala, Gln, Trp, Phe and Ser), respectively (see Table 2 in Ref [Wang et al., 2016]). Additionally, we found a C2H2-type zinc finger domain (Pfam: 13,912) downstream of *GrCSD4*. The result showed that *GrCSD4* and *GrCSD4* had similar gene structures and close evolutionary relationships but differed in ORF length, protein length, molecular weight (MW) and *pI*. We also found that *GrFSD1* had two iron/manganese SOD C-terminal domains (Pfam: 02,777). Compared with *GaFSD1*, the length and position of the domain was no different; the amount and the mechanism were not satisfactorily explained.

3.5. Express analysis

Tissue-specific expressions of cotton *SOD* genes were evaluated in five different tissues and organs: roots, stems, leaves, flowers and ovules. Digital expression analyses of cotton *SOD* genes were performed using the NCBI-EST database and mixed expressed sequence tags (EST) data were not taken into account (Table 2). Nine *SOD* genes were found expressed in roots, ten in stems, nine in leaves, eight in flowers and 18 in ovules. Exceptionally, *GaCSD1* expression was only found in ovule compared with other tissues and organs. Duplicated gene pairs (i.e. *GrMSD1* and *GrMSD2*, *GaMSD1* and *GaMSD2*, *GrFSD1* and *GrFSD2*, and *GaFSD1* and *GaFSD2*) shared the same expression patterns and could indicate tissue-specific expression patterns. *GhCSD1* located in cytosol was cloned and expressed in *G. hirsutum*, and the results showed that Cu/Zn-SOD mRNA was expressed in different organs: roots, stems, leaves and flowers (Hu, Yu, Fan, & Song, 2007) — similar findings to those of the present study. *GrCSD2* and *GaCSD2* were both predicted to be located in cytosol and clustered together with *GhCSD1* — they were all expressed in the five tissues and organs. *SOD* genes that were predominantly expressed in root and/or leaf tissue were involved in scavenging ROS, and played a pivotal role as triggers of gene expression during abiotic stresses (Ranjan & Sawant, 2015). In the present study, ten of all *SOD* genes were found to be expressed in roots and/or leaves tissue, but determining their role(s) during abiotic stresses requires further experiments. All of the cotton *SOD* genes were found to be expressed in ovules. The results indicate stage-specific expression patterns.

To investigate the stage-specific expression patterns of the cotton *SOD* genes in ovules, we analyzed the whole-transcriptome sequencing data from cotton ovules at various stages (i.e. 10, 20, 30 and 40 DPA) of *G. raimondii* and *G. arboreum*, with three replicates for every stage. Cotton fibers are single-cell trichomes arising from the epidermis of developing cotton ovules. Cotton fiber development occurs in four

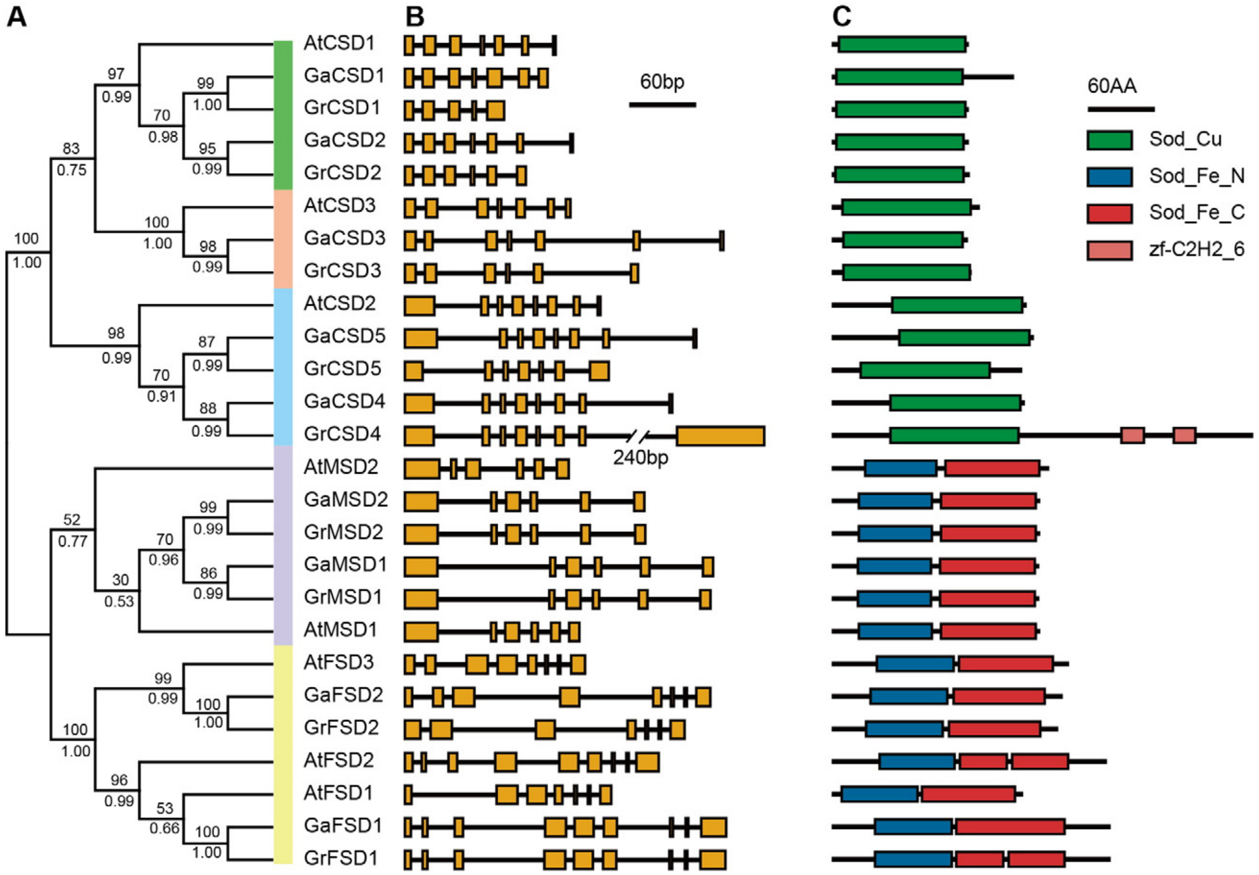

Fig. 3. Phylogenetic tree, gene structure and domain analysis of superoxide dismutase (SOD) in *Arabidopsis*, *G. raimondii* and *G. arboreum*. (A) Phylogenetic tree of *Arabidopsis*, *G. raimondii* and *G. arboreum* SODs constructed with ML method of the PhyML and using 1000 bootstrap value. Subgroup-a, -b, -c, -d and -e colored in green, orange, blue, purple and yellow, respectively. (B) Exon–intron structures of *SOD* genes. (C) Conserved domains annotated by Pfam database.

overlapping stages: initiation, elongation, secondary cell wall (SCW) deposition, and maturation (Naithani et al., 1981). At 10 DPA, fibers are in the elongation stage and have only a thin primary cell wall (PCW); at 20 DPA, fibers are in the transition from elongation to SCW deposition and have a thin PCW and some SCW; at 30 DPA, fibers are in the SCW deposition stage and have a thin PCW and a thicker SCW; and at 40 DPA, fibers are in the maturation stage (Ding et al., 2015; Ralph et al., 2008b). The expression patterns of 18 *SOD* genes are shown in Fig. 4. All predicted cotton *SOD* genes were expressed in each stage of ovule development and exhibited developmentally regulated expression patterns. Exceptionally, *GrCSD5* was expressed at low levels in all four stages, and was lowest at 10 DPA. Of the 18 predicted *SOD* genes, four were stably expressed in all four development stages, with two *SOD* genes (*GrCSD2* and *GaCSD2*) almost identical to each other and expressed at high levels in all stages. Many genes involved in cell elongation or SCW synthesis in cotton fibers are transcriptionally regulated, and there are multiple developmental programs controlling gene expression throughout cotton fiber development (Haigler et al., 2005). The expression levels of *GaFSD1*, *GaMSD2*, *GrMSD1* and *GrMSD2* peaked during the elongation stage (10 DPA) and declined coincident with the initiation of SCW synthesis (20 DPA) — these were similar to the patterns of genes expressed primarily during the cell elongation stage of fiber development (Kim and Triplett, 2004). In contrast, four genes (*GaFSD1*, *GaMSD2*, *GrMSD1* and *GrMSD2*) were most abundant in 10-DPA ovules, five (*GaCSD3*, *GaCSD4*, *GaCSD5*, *GaFSD2* and *GrCSD5*) were most abundant in 20-DPA ovules, one (*GrFSD2*) was most abundant in 40-DPA ovules, two (*GaMSD1* and *GrCSD1*) was most abundant in 10- and 20-DPA ovules, one (*GrCSD4*) was most abundant in 20- and 40-DPA ovules, and one (*GrFSD1*) was most abundant in 30- and 40-DPA ovules. Moreover, the profile similarities of transcripts encoded by duplicated genes, such as *GrCSD2* and *GaCSD2*, suggested functional redundancy between these family members, the different expression patterns of members of the *GrCSD4,5* and *GaCSD4,5* indicated that other cotton *SOD* genes have been preserved by sub-functionalization.

Table 2
Digital expression analysis of *SOD* genes in *G. raimondii* and *G. arboreum*.

Group	Gene name	Tissue and organ type					
		Root	Stem	Leaf	Flower	Ovule	Mixed
Cu/Zn-SOD	*GrCSD1*				+	+	
	GrCSD2	+	+	+	+	+	
	GrCSD3			+	+	+	+
	GrCSD4		+	+	+	+	+
	GrCSD5	+	+			+	
	GaCSD1					+	
	GaCSD2	+	+	+	+	+	
	GaCSD3			+	+	+	+
	GaCSD4	+	+			+	
	GaCSD5	+	+			+	
Mn-SOD	*GrMSD1*	+	+			+	+
	GrMSD2	+	+			+	
	GaMSD1	+	+			+	+
	GaMSD2	+	+			+	
Fe-SOD	*GrFSD1*			+		+	
	GrFSD2			+		+	
	GaFSD1			+	+	+	+
	GaFSD2			+	+	+	+

+: Expressed; blank: not expressed.

Table 3
The details of predicted SOD 3D structures, containing predicted binding sites, conserved residues, and Ramachandran plot statistics.

Gene name	Sequence ID	Predicted binding sites		Conserved residues		Ramachandran plot		
		Cu	Zn	Disulfide bond	Enzyme activity	Favored regions	Allowed regions	Outlier regions
		(H,H,H,H)	(H,H,H,D)	(C,C)	(R)			
GrCSD1	Cotton_D_gene_10024020	45,47,62,119	62,70,79,82	56,145	142	95.3%	4.7%	0.0%
GrCSD2	Cotton_D_gene_10006229	45,47,62,119	62,70,79,82	56,145	142	95.3%	4.7%	0.0%
GrCSD3	Cotton_D_gene_10039927	52,54,69,126	69,77,86,89	63,152	149	91.7%	6.2%	2.1%
GrCSD4	Cotton_D_gene_10006544	106,108,123,180	123,131,140,143	117,206	203	91.0%	7.6%	1.4%
GrCSD5	Cotton_D_gene_10032111	73,75,90,147	90,98,107,110	84,173	170	93.0%	5.7%	1.3%
GaCSD1	Cotton_A_21978	45,47,62,119	62,70,79,82	56,145	142	95.2%	4.8%	0.0%
GaCSD2	Cotton_A_24238	45,47,62,119	62,70,79,82	56,145	142	95.1%	4.9%	0.0%
GaCSD3	Cotton_A_30467	52,54,59,116	59,67,76,79	N,142	139	91.4%	7.2%	1.4%
GaCSD4	Cotton_A_32487	106,108,123,180	123,131,140,143	117,206	203	92.3%	6.7%	1.0%
GaCSD5	Cotton_A_36793	116,118,133,190	133,141,150,153	127,216	213	93.0%	6.0%	1.0%

		Mn	Fe	Hydrogen bond			
		(H,H,D,H)	(H,H,D,H)	(Q,Y,H,Y)			
GrMSD1	Cotton_D_gene_10016246	54,102,191,195	–	175,62,58,198	97.6%	2.4%	0.0%
GrMSD2	Cotton_D_gene_10018648	55,103,192,196	–	176,63,59,199	97.2%	2.8%	0.0%
GaMSD1	Cotton_A_04050	54,102,191,195	–	174,62,58,198	97.6%	2.4%	0.0%
GaMSD2	Cotton_A_21263	55,103,192,196	–	175,63,59,199	97.2%	2.8%	0.0%
GrFSD1	Cotton_D_gene_10009356	–	73,125,224,228	121,81,77,231	93.4%	3.8%	2.8%
GrFSD2	Cotton_D_gene_10030063	–	62,115,199,203	N,70,66,206	92.7%	6.3%	1.0%
GaFSD1	Cotton_A_03623	–	73,125,224,228	121,81,77,231	93.4%	3.8%	2.8%
GaFSD2	Cotton_A_26478	–	67,120,204,208	N,75,71,211	92.7%	5.8%	1.5%

Gr: *G. raimondii*; Ga: *G. arboreum*; CSD: Cu/Zn-SOD; FSD: Fe-SOD; MSD: Mn-SOD; H: His, histidine; D: Asp, aspartate; C: Cys, cysteine; R: Arg, arginine; Q: Gln, glutamine; Y: Tyr, tyrosine; N: not found residues site.

3.6. Predicted 3D structures of SODs

We predicted the 3D structures of SODs with the protein sequences (see Table 1 in Ref [Wang et al., 2016]) using the SWISS-MODEL serve, and downloaded the built models files (see Supporting Information) which were used to view the 3D structures. According to the SWISS-MODEL homology modeling report, we collected homology templates whose accession were showed in Table 1 in Ref (Wang et al., 2016). In *G. raimondii* and *G. arboreum* Cu/Zn-SODs, there was a highly conserved advanced structure with two identical subunits, including an antiparallel β-barrel consisting of eight antiparallel β-strands in a Greek key topology (Figs. 6A–E and 7A–E,). In each subunit, the active site contained one Cu ion ligated by four histidines when in the reduced state and one Zn ion ligated by one aspartic acid and three histidines, whose side-chains all resided outside of the β-barrel. One of the histidine ligands of the Zn ion ligands also ligated the Cu ion when in the oxidized state and thus has been termed the bridging histidine. The template was fully conserved as for the known Cu/Zn-SODs, with the Cu and Zn metal ion-interacting residues shown in Table 3. The conserved disulfide bond residues were determined in the active site channel of each Cu/Zn-SOD subunit except for *GaCSD3*, including Cys56/Cys145, Cys56/Cys145, Cys63/Cys152, Cys117/Cys206 and Cys84/Cys173 in *GrCSD1*, *GrCSD2*, *GrCSD3*, *GrCSD4* and *GrCSD5*, respectively; and Cys56/Cys145, Cys56/Cys145, Cys117/Cys206 and Cys127/Cys216 in *GaSOD1*, *GaSOD2*, *GaCSD4* and *GaCSD5*, respectively (Table 3). The alignment between the *GaCSD3* amino acid sequence and the sequence of template protein showed that *GaCSD3* had lost ten aa, which might contain one cysteine (Fig. 5). Thus, our findings mostly corroborated the findings of Ding and Dokholyan (2008) and Perry et al. (2010). Both the binding of the active site metal ions and the formation of the conserved disulfide bond in each subunit would contribute to the framework stability and specificity of the protein fold and dimer assembly. Also conserved was an arginine residue, which is considered important for enzymatic activity (Table 3) (Fisher et al., 1994).

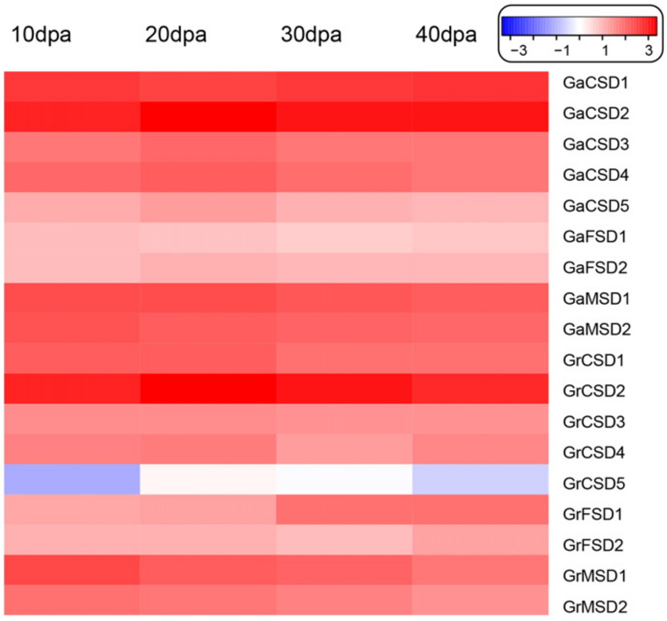

Fig. 4. Expression patterns of *SOD* genes in *G. raimondii* and *G. arboreum*. Heatmap showing the *SOD* genes across four stages (ovules at 10, 20, 30 and 40 DPA; mentioned at the top of each lane). The color scale at the top of the gene names indicates the relative expression levels. Color scale represents reads per kilobase per million normalized $\log_{10}$ transformed counts, where blue indicates low level and red indicates high level.

The analysis of Mn/Fe-SODs of *G. raimondii* and *G. arboreum* revealed three antiparallel β-sheets dominated by α-helices as also found in *O. sativa* (Dehury et al., 2013), *Glycine max* (Gopavajhula et al., 2013) and *So. bicolor* (Filiz and Tombuloğlu, 2015) (Figs. 6F–I and 7F–I). There was a slight difference in the number of α-helices, with both enzymes lacking the disulfide bond, which is normally a unique characteristic of Cu/Zn-SODs. For example, the active site of Mn/Fe-SOD was located between the N- and C-terminal domains, and it differed from that of Cu/Zn-SODs by containing a single metal ion. The metal ion is coordinated in a strained trigonal bipyramidal geometry by four amino acid side-chains (His-His-Asp-His) (Table 3), and by one solvent

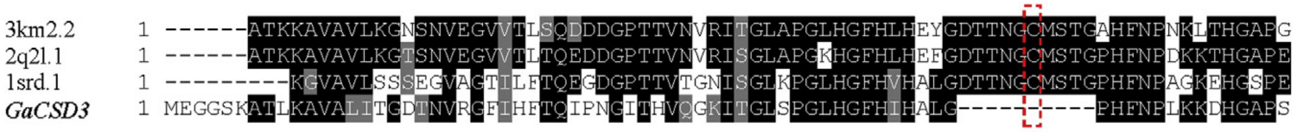
3km2.2 1 ------ATKKAVAVLKGNSNVEGVVTLSQDDDGPTTVNVRITGLAPGLHGFHLHEYGDTTNGCMSTGAHFNPNKLTHGAPG
2q2l.1 1 ------ATKKAVAVLKGTSNVEGVVTLTQEDDGPTTVNVRISGLAPGKHGFHLHEFGDTTNGCMSTGPHFNPDKKTHGAPE
1srd.1 1 ---------KGVAVLSSSEGVAGTILFTQEGDGPTTVTGNISGLKPGLHGFHVHALGDTTNGCMSTGPHFNPAGKEHGSPE
GaCSD3 1 MEGGSKATLKAVALITGDTNVRGFIHFTQIPNGITHVQGKITGLSPGLHGFHIHALG--------PHFNPLKKDHGAPS
3km2.2 76 DEIRHAGDLGNIVANADGVAEVTLVDNQIPLTGPNSVVGRALVVHELEDDLGKGGHELSLTTGNAGGRLACGVVGL----
2q2l.1 91 DEVRHAGDLGNIVANTDGVAEATIVDNQIPLTGPNSVVGRALVVHELEDDLGKGGHELSPTTGNAGGRLACGVVGL----
1srd.1 88 DETRHAGDLGNITVGDDGTACFTIVDKQIPLTGPHSIIGRAVVVHADPDDLGKGGHELSKSTGNAGGRIACGIIGLQ---
GaCSD3 81 DGERHAGDLGNIIAGPDGVAEVSIKDWQIPLSGQHSILGRAVVVHADPDDLGKGGHELSKTTGNAGARVGCGIIGLQSSV

Fig. 5. Multiple alignment of the *GaCSD3* amino acid sequence with template protein. The multiple alignment was obtained using ClustalW and conserved amino acids were shaded using Boxshade (v.3.21). Dashes (—) indicate gaps in the alignment; tomato Cu/Zn-SOD (PDB ID: 3km2.2); potentilla Cu/Zn-SOD (PDB ID: 2q2l.1) (Yogavel et al., 2007); and spinach Cu/Zn-SOD (PDB ID: 1srd.1) (Kitagawa et al., 1991).

molecule. The active site of Mn/Fe-SOD also contains a hydrogen bond network (Gln-Tyr-His-Tyr) that extends from the metal-bound solvent molecule to solvent-exposed residues at the interface between subunits (Table 3). A water molecule situated between Tyr and the His side-chains mediates the hydrogen bonding between these two residues. The hydrogen bond network supports proton transfer in catalysis

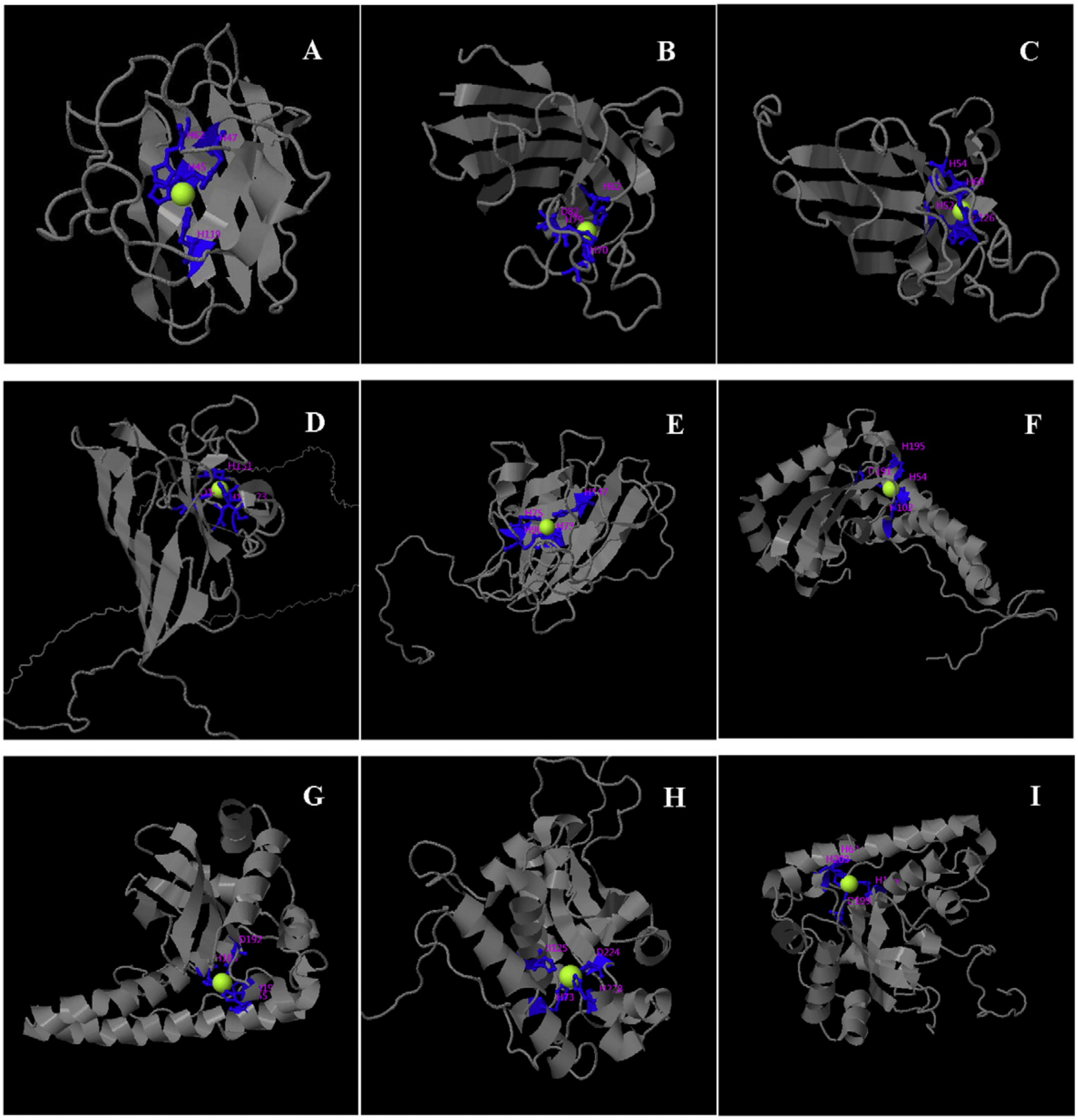

Fig. 6. Predicted 3D structures and binding sites of *G. raimondii* superoxide dismutases using the SWISS-MODEL server. A: *GrCSD1*; B: *GrCSD2*; C: *GrCSD3*; D: *GrCSD4*; E: *GrCSD5*; F: *GrMSD1*; G: *GrMSD2*; H: *GrFSD1*; I: *GrFSD2*.

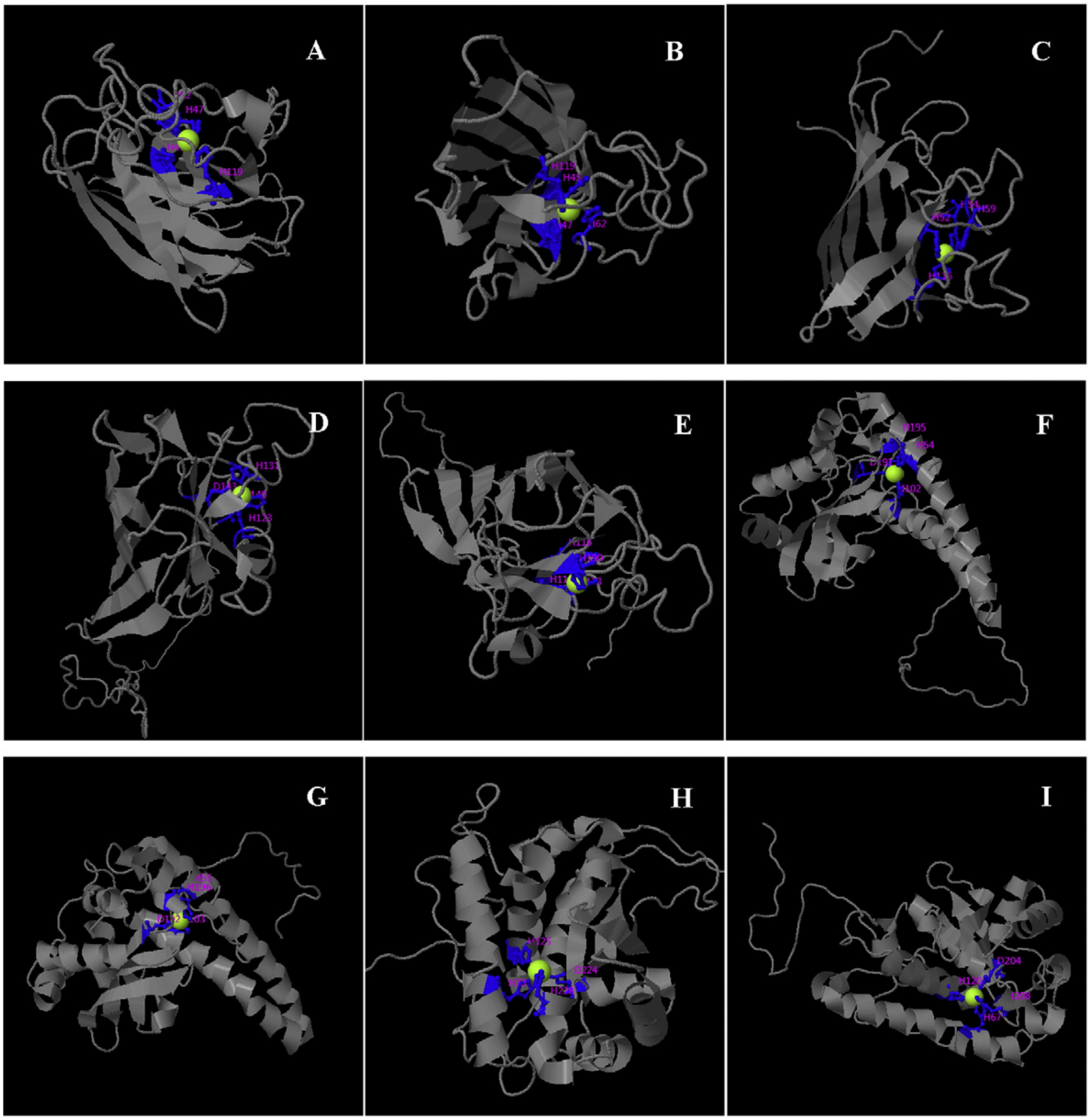

Fig. 7. Predicted 3D structures and binding sites of *G. arboreum* superoxide dismutases using the SWISS-MODEL server. A: *GaCSD1*; B: *GaCSD2*; C: *GaCSD3*; D: *GaCSD4*; E: *GaCSD5*; F: *GaMSD1*; G: *GaMSD2*; H: *GaFSD1*; I: *GaFSD2*.

(Borgstahl et al., 1992). Further structural study of these enzymes is expected to enhance knowledge of the catalytic mechanism and metal ion binding.

The accuracy of the predicted models was evaluated by Ramachandran plot (Table 3, see Figs. 2 and 3 in Ref [Wang et al., 2016]) using the RAMPAGE server. The refined SOD models showed good proportions of residues in favored, allowed and outlier regions (Table 3). The results indicated that the 3D models were of good quality.

4. Conclusions

We reported a genome-wide analysis of the important *SOD* gene family in *G. raimondii* and *G. arboreum*, including genome-wide characterization, phylogenetic analysis, chromosomal location, gene structure, digital expression and 3D structure modeling. Our findings will contribute to the understanding of *SOD* genes and proteins in *G. raimondii* and *G. arboreum*, especially in *G. hirsutum*. To our knowledge, this is the first report of a comparative genome-wide analysis of the *SOD* genes family in *Arabidopsis*, *G. raimondii* and *G. arboreum*. This study should provide a solid foundation for future functional studies, and for guiding future laboratory experimental work on confirming the functional analyses of *SOD* genes in various stress conditions.

Author contributions

Wei Wang, Minxuan Xia and Fafu Shen conceived and designed experiment; Wei Wang, Minxuan Xia, Jie Chen, Fenni Deng, Rui Yuan and Xiaopei Zhang performed experiment; Wei Wang and Minxuan Xia analyzed the data; Jie Chen contributed analysis tools; Wei Wang and Fafu Shen wrote manuscript.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

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).

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.plgene.2016.02.002.

References

Altenhoff, A.M., Dessimoz, C., 2009. Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS Comput. Biol. 5 (1), e1000262. http://dx.doi.org/10.1371/journal.pcbi.1000262.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215 (3), 403–410. http://dx.doi.org/10.1016/S0022-2836(05)80360-2.

Apostolova, E., Rashkova, M., Anachkov, N., Denev, I., Toneva, V., Minkov, I., Yahubyan, G., 2012. Molecular cloning and characterization of cDNAs of the superoxide dismutase gene family in the resurrection plant *Haberlea rhodopensis*. Plant Physiol. Biochem. *55*, 85–92http://dx.doi.org/10.1016/j.plaphy.2012.03.015.

Biasini, M., Bienert, S., Waterhouse, A., Arnold, K., Studer, G., Schmidt, T., ... Schwede, T., 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. *42* (Web Server issue), W252–W258. http://dx.doi.org/10.1093/nar/gku340.

Bindschedler, L.V., Palmblad, M., Cramer, R., 2008. Hydroponic isotope labelling of entire plants (HILEP) for quantitative plant proteomics; an oxidative stress case study. Phytochemistry *69* (10), 1962–1972http://dx.doi.org/10.1016/j.phytochem.2008.04.007.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30 (15), 2114–2120. http://dx.doi.org/10.1093/bioinformatics/btu170.

Borgstahl, G.E.O., Parge, H.E., Hickey, M.J., Beyer Jr., W.F., Hallewell, R.A., Tainer, J.A., 1992. The structure of human mitochondrial manganese superoxide dismutase reveals a novel tetrameric interface of two 4-helix bundles. Cell *71* (1), 107–118http://dx.doi.org/10.1016/0092-8674(92)90270-M.

Bowler, C., Van Camp, W., Van Montagu, M., Inzé, D., Asada, K., 1994. Superoxide dismutase in plants. Crit. Rev. Plant Sci. 13 (3), 199–218.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: architecture and applications. BMC Bioinformatics 10, 421. http://dx.doi.org/10.1186/1471-2105-10-421.

Cannon, R.E., White, J.A., Scandalios, J.G., 1987. Cloning of cDNA for maize superoxide dismutase 2 (SOD2). Proc. Natl. Acad. Sci. U. S. A. 84 (1), 179–183.

Dehury, B., Sarma, K., Sarmah, R., Sahu, J., Sahoo, S., Sahu, M., ... Barooah, M., 2013. In silico analyses of superoxide dismutases (SODs) of rice (*Oryza sativa* L.). J. Plant Biochem. Biotechnol. 22 (1), 150–156. http://dx.doi.org/10.1007/s13562-012-0121-6.

Ding, F., Dokholyan, N.V., 2008. Dynamical roles of metal ions and the disulfide bond in Cu, Zn superoxide dismutase folding and aggregation. Proc. Natl. Acad. Sci. U. S. A. 105 (50), 19696–19701. http://dx.doi.org/10.1073/pnas.0803266105.

Ding, M., Chen, J., Jiang, Y., Lin, L., Cao, Y., Wang, M., ... Ye, W., 2015. Genome-wide investigation and transcriptome analysis of the WRKY gene family in *Gossypium*. Mol. Genet. Genomics 290 (1), 151–171. http://dx.doi.org/10.1007/s00438-014-0904-7.

Dong, C., Zheng, X., Li, G., Zhu, H., Zhou, M., Hu, Z., 2011. Molecular cloning and expression of two cytosolic copper–zinc superoxide dismutases genes from *Nelumbo nucifera*. Appl. Biochem. Biotechnol. 163 (5), 679–691. http://dx.doi.org/10.1007/s12010-010-9074-1.

Du, D., Hao, R., Cheng, T., Pan, H., Yang, W., Wang, J., Zhang, Q., 2013. Genome-wide analysis of the AP2/ERF gene family in *Prunus mume*. Plant Mol. Biol. Report. 31 (3), 741–750. http://dx.doi.org/10.1007/s11105-012-0531-6.

Feng, W., Hongbin, W., Bing, L., Jinfa, W., 2006. Cloning and characterization of a novel splicing isoform of the iron-superoxide dismutase gene in rice (*Oryza sativa* L.). Plant Cell Rep. 24 (12), 734–742. http://dx.doi.org/10.1007/s00299-005-0030-4.

Feng, X., Lai, Z., Lin, Y., Lai, G., Lian, C., 2015. Genome-wide identification and characterization of the superoxide dismutase gene family in *Musa acuminata* cv. tianbaojiao (AAA group). BMC Genomics 16 (1), 1–16. http://dx.doi.org/10.1186/s12864-015-2046-7.

Filiz, E., Tombuloğlu, H., 2015. Genome-wide distribution of superoxide dismutase (SOD) gene families in *Sorghum bicolor*. Turk. J. Biol. 39 (1), 49–59.

Fink, R.C., Scandalios, J.G., 2002. Molecular evolution and structure–function relationships of the superoxide dismutase gene families in angiosperms and their relationship to other eukaryotic and prokaryotic superoxide dismutases. Arch. Biochem. Biophys. *399* (1), 19–36http://dx.doi.org/10.1006/abbi.2001.2739.

Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., ... Punta, M., 2014. Pfam: the protein families database. Nucleic Acids Res. 42 (Database issue), D222–D230. http://dx.doi.org/10.1093/nar/gkt1223.

Fisher, C.L., Cabelli, D.E., Tainer, J.A., Hallewell, R.A., Getzoff, E.D., 1994. The role of arginine 143 in the electrostatics and mechanism of Cu, Zn superoxide dismutase: computational and experimental evaluation by mutational analysis. Proteins: Struct., Funct., Bioinf. 19 (1), 24–34. http://dx.doi.org/10.1002/prot.340190105.

Ganko, E.W., Meyers, B.C., Vision, T.J., 2007. Divergence in expression between duplicated genes in Arabidopsis. Mol. Biol. Evol. 24 (10), 2298–2309. http://dx.doi.org/10.1093/molbev/msm158.

Gill, S., Anjum, N., Gill, R., Yadav, S., Hasanuzzaman, M., Fujita, M., ... Tuteja, N., 2015. Superoxide dismutase — mentor of abiotic stress tolerance in crop plants. Environ. Sci. Pollut. Res. 22 (14), 10375–10394. http://dx.doi.org/10.1007/s11356-015-4532-5.

Gopavajhula, V.R., Chaitanya, K.V., Akbar Ali Khan, P., Shaik, J.P., Reddy, P.N., Alanazi, M., 2013. Modeling and analysis of soybean (*Glycine max*. L) Cu/Zn, Mn and Fe superoxide dismutases. Genet. Mol. Biol. 36 (2), 225–236. http://dx.doi.org/10.1590/S1415-47572013005000023.

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., Gascuel, O., 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59 (3), 307–321. http://dx.doi.org/10.1093/sysbio/syq010.

Haigler, C.H., Zhang, D., Wilkerson, C.G., 2005. Biotechnological improvement of cotton fibre maturity. Physiol. Plant. 124 (3), 285–294. http://dx.doi.org/10.1111/j.1399-3054.2005.00480.x.

Halliwell, B., Gutteridge, J.M., 1984. Oxygen toxicity, oxygen radicals, transition metals and disease. Biochem. J. 219 (1), 1–14.

Hovav, R., Udall, J. A., Chaudhary, B., Hovav, E., Flagel, L., Hu, G., & Wendel, J. F. (2008). The evolution of spinnable Cotton fiber entailed prolonged development and a novel metabolism. PLoS Genet., 4(2), e25. doi: http://dx.doi.org/10.1371/journal.pgen.0040025

Hu, G.-H., Yu, S.-X., Fan, S.-L., Song, M.-Z., 2007. Cloning and expressing of a gene encoding cytosolic coppereinc superoxide dismutase in the upland cotton. Agric. Sci. China *6* (5), 536–544http://dx.doi.org/10.1016/S1671-2927(07)60080-7.

Kaminaka, H., Morita, S., Tokumoto, M., Yokoyama, H., Masumura, T., Tanaka, K., 1999. Molecular cloning and characterization of a cDNA for an iron-superoxide dismutase in rice (*Oryza sativa* L.). Biosci. Biotechnol. Biochem. 63 (2), 302–308. http://dx.doi.org/10.1271/bbb.63.302.

Kim, H., Triplett, B., 2004. Cotton fiber germin-like protein. I. Molecular cloning and gene expression. Planta 218 (4), 516–524. http://dx.doi.org/10.1007/s00425-003-1133-1.

Kim, H., Kato, N., Kim, S., Triplett, B., 2008a. Cu/Zn superoxide dismutases in developing cotton fibers: evidence for an extracellular form. Planta 228 (2), 281–292. http://dx.doi.org/10.1007/s00425-008-0734-0.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., Salzberg, S.L., 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14 (4), R36. http://dx.doi.org/10.1186/gb-2013-14-4-r36.

Kitagawa, Y., Tanaka, N., Hata, Y., Kusunoki, M., Lee, G.-P., Katsube, Y., Morita, Y., 1991. Three-dimensional structure of Cu, Zn-superoxide dismutase from spinach at 2.0 &aring; resolution. J. Biochem. 109 (3), 477–485.

Kliebenstein, D.J., Monde, R.-A., Last, R.L., 1998. Superoxide dismutase in *Arabidopsis*: an eclectic enzyme family with disparate regulation and protein localization. Plant Physiol. 118 (2), 637–650.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., ... Marra, M.A., 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19 (9), 1639–1645. http://dx.doi.org/10.1101/gr.092759.109.

Lawton-Rauh, A., 2003. Evolutionary dynamics of duplicated genes in plants. Mol. Phylogenet. Evol. *29* (3), 396–409 http://dx.doi.org/10.1016/j.ympev.2003.07.004.

Li, F., Fan, G., Wang, K., Sun, F., Yuan, Y., Song, G., ... Yu, S., 2014. Genome sequence of the cultivated cotton *Gossypium arboreum*. Nat. Genet. 46 (6), 567–572. http://dx.doi.org/10.1038/ng.2987.

Lin, Y.-L., Lai, Z.-X., 2013. Superoxide dismutase multigene family in longan somatic embryos: a comparison of CuZn-SOD, Fe-SOD, and Mn-SOD gene structure, splicing, phylogeny, and expression. Mol. Breed. 32 (3), 595–615. http://dx.doi.org/10.1007/s11032-013-9892-2.

Liu, Z., Zhang, W., Gong, X., Zhang, Q., Zhou, L., 2014. A Cu/Zn superoxide dismutase from *Jatropha curcas* enhances salt tolerance of *Arabidopsis thaliana*. Genet. Mol. Res. 14 (1), 2086–2098.

Lovell, S.C., Davis, I.W., Arendall, W.B., de Bakker, P.I.W., Word, J.M., Prisant, M.G., ... Richardson, D.C., 2003. Structure validation by Cα geometry: ϕ, ψ and Cβ deviation. Proteins: Struct., Funct., Bioinf. 50 (3), 437–450. http://dx.doi.org/10.1002/prot.10286.

Löytynoja, A., Goldman, N., 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320 (5883), 1632–1635. http://dx.doi.org/10.1126/science.1158395.

Ma, T., Wang, J., Zhou, G., Yue, Z., Hu, Q., Chen, Y., ... Liu, J., 2013. Genomic insights into salt adaptation in a desert poplar. Nat. Commun. *4*. http://dx.doi.org/10.1038/ncomms3797.

Miller, A.-F., 2012. Superoxide dismutases: ancient enzymes and new insights. FEBS Lett. 586 (5), 585–595.

Naithani, S.C., Rama-Rao, N., Krishnan, P.N., Singh, Y.D., 1981. Changes *o*-diphenol oxidase during fibre development in cotton. Ann. Bot. (No.3), 379–385.

Perry, J.J., Shin, D.S., Getzoff, E.D., Tainer, J.A., 2010. The structural biochemistry of the superoxide dismutases. Biochim. Biophys. Acta 1804 (2), 245–262. http://dx.doi.org/10.1016/j.bbapap.2009.11.004.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., Lopez, R., 2005. InterProScan: protein domains identifier. Nucleic Acids Res. *33* (Web Server issue), W116–W120. http://dx.doi.org/10.1093/nar/gki442.

Rajjou, L., Lovigny, Y., Groot, S.P.C., Belghazi, M., Job, C., Job, D., 2008. Proteome-wide characterization of seed aging in Arabidopsis: a comparison between artificial and natural aging protocols. Plant Physiol. 148 (1), 620–641. http://dx.doi.org/10.1104/pp.108.123141.

Ralph, S.G., Chun, H.J.E., Cooper, D., Kirkpatrick, R., Kolosova, N., Gunter, L., ... Bohlmann, J., 2008a. Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. BMC Genomics 9, 57. http://dx.doi.org/10.1186/1471-2164-9-57.

Ralph, S.G., Chun, H.J.E., Kolosova, N., Cooper, D., Oddy, C., Ritland, C.E., ... Bohlmann, J., 2008b. A conifer genomics resource of 200,000 spruce (*Picea* spp.) ESTs and 6,464 high-quality, sequence-finished full-length cDNAs for Sitka spruce (*Picea sitchensis*). BMC Genomics 9, 484. http://dx.doi.org/10.1186/1471-2164-9-484.

Ranjan, A., Sawant, S., 2015. Genome-wide transcriptomic comparison of cotton (*Gossypium herbaceum*) leaf and root under drought stress. *3*. Biotech 5 (4), 585–596. http://dx.doi.org/10.1007/s13205-014-0257-2.

Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19 (12), 1572–1574.

Rubio, M.C., Becana, M., Sato, S., James, E.K., Tabata, S., Spaink, H.P., 2007. Characterization of genomic clones and expression analysis of the three types of superoxide dismutases during nodule development in lotus japonicus. Mol. Plant-Microbe Interact. 20 (3), 262–275. http://dx.doi.org/10.1094/MPMI-20-3-0262.

Scandalios, J.G., 1997. Molecular genetics of superoxide dismutase in plants. In: Scandalios, J.G. (Ed.), Oxidative Stress and Molecular Biology of Antioxidant Defenses. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, pp. 527e–568e.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., Paterson, A.H., 2008. Synteny and collinearity in plant genomes. Science 320 (5875), 486–488. http://dx.doi.org/10.1126/science.1153917.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., ... Pachter, L., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat. Protoc. 7 (3), 562–578. http://dx.doi.org/10.1038/nprot.2012.016.

Viczián, O., Künstler, A., Hafez, Y., Király, L., 2014. Catalases may play different roles in influencing resistance to virus-induced hypersensitive necrosis. Acta Phytopathol. Entomol. Hung. 49 (2), 189–200.

Wang, Y., Tang, H., Debarry, J.D., Tan, X., Li, J., Wang, X., ... Paterson, A.H., 2012a. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 40 (7), e49. http://dx.doi.org/10.1093/nar/gkr1293.

Wang, K., Wang, Z., Li, F., Ye, W., Wang, J., Song, G., ... Yu, S., 2012b. The draft genome of a diploid cotton *Gossypium raimondii*. Nat. Genet. *44* (10), 1098–1103 (http://www.nature.com/ng/journal/v44/n10/abs/ng.2371.html#supplementary-information).

Wang, W., Xia, M., Chen, J., Deng, F., Yuan, R., Zhang, X., Shen, F., 2016. Data set for genome-wide analysis of superoxide dismutase (SOD) gene family in *Gossypium raimondii* and *G. arboreum*. Data Brief (submitted for publication).

Wendel, J., Brubaker, C., Alvarez, I., Cronn, R., Stewart, J., 2009. Evolution and Natural History of the Cotton Genus. In: Paterson, A. (Ed.)Genetics and Genomics of Cotton Vol. 3. Springer, US, pp. 3–22.

Xu, L., Zhu, L., Tu, L., Liu, L., Yuan, D., Jin, L., ... Zhang, X., 2011. Lignin metabolism has a central role in the resistance of cotton to the wilt fungus verticillium dahliae as revealed by RNA-seq-dependent transcriptional analysis and histochemistry. J. Exp. Bot. 62 (15), 5607–5621.

Yang, J., Roy, A., Zhang, Y., 2013a. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic Acids Res. 41 (Database issue), D1096–D1103. http://dx.doi.org/10.1093/nar/gks966.

Yang, J., Roy, A., Zhang, Y., 2013b. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 29 (20), 2588–2595. http://dx.doi.org/10.1093/bioinformatics/btt447.

Yogavel, M., Gill, J., Mishra, P. C., & Sharma, A. (2007). SAD phasing of a structure based on cocrystallized iodides using an in-house Cu Kα X-ray source: effects of data redundancy and completeness on structure solution. Acta Crystallogr. D Biol. Crystallogr., 63(8), 931–934. http://dx.doi.org/10.1107/S0907444907029174