Available online at www.sciencedirect.com
SCIENCE DIRECT
PHYTOCHEMISTRY
Phytochemistry 67 (2006) 534–539
www.elsevier.com/locate/phytochem

Update in Bioinformatics

PeroxiBase: A class III plant peroxidase database

Nenad Bakalovic [a,1], Filippo Passardi [a,1], Vassilios Ioannidis [b], Claudia Cosio [a], Claude Penel [a], Laurent Falquet [b], Christophe Dunand [a,*]

[a] *Laboratory of Plant Physiology, University of Geneva, Quai Ernest-Ansermet 30, CH-1211 Geneva 4, Switzerland*
[b] *Swiss Institute of Bioinformatics, CH-1066 Epalinges/Lausanne, Switzerland*

Received 13 September 2005; received in revised form 9 December 2005
Available online 26 January 2006

Abstract

Class III plant peroxidases (EC 1.11.1.7), which are encoded by multigenic families in land plants, are involved in several important physiological and developmental processes. Their varied functions are not yet clearly determined, but their characterization will certainly lead to a better understanding of plant growth, differentiation and interaction with the environment, and hence to many exciting applications. Since there is currently no central database for plant peroxidase sequences and many plant sequences are not deposited in the EMBL/GenBank/DDBJ repository or the UniProt KnowledgeBase, this prevents researchers from easily accessing all peroxidase sequences. Furthermore, gene expression data are poorly covered and annotations are inconsistent. In this rapidly moving field, there is a need for continual updating and correction of the peroxidase superfamily in plants. Moreover, consolidating information about peroxidases will allow for comparison of peroxidases between species and thus significantly help making correlations of function, structure or phylogeny. We report a new database (PeroxiBase) accessible through a web server (http://peroxidase.isb-sib.ch) with specific tools dedicated to facilitate query, classification and submission of peroxidase sequences. Recent developments in the field of plant peroxidase are also mentioned.

Keywords: Database; Multigenic family; Evolution; Phylogeny; Peroxidases

1. Introduction

Class III plant peroxidases (EC 1.11.1.7, donor:hydrogen-peroxide oxidoreductase) are present in all land plants (Table 1). Genes encoding this enzyme family are particularly numerous in Angiosperms. The high number of isoenzymes and their remarkable catalytic versatility allow them to be involved in a broad range of physiological and developmental processes all along the plant life cycle (Passardi et al., 2005). Plant peroxidases have been shown to be involved in the cross-linking of cell wall constituents, lignin polymerization, the catabolism of auxin – a hormone having a critical role in plant growth and development – and the formation of reactive oxygen species (superoxide, hydroxyl radical). They also play a prominent role in defence reactions against many pathogenic organisms. However, until now the in vivo functions of a particular peroxidase have not been reported. This knowledge is however crucial to understand the evolution, the roles and the regulations of this key multifunctional enzyme. Plant peroxidases are an example of a multigenic family whose number of members increased since the conquering of land by plants due to constant evolution. The *Arabidopsis* genome contains 73 genes encoding a peroxidase (Tognolli et al., 2002) and rice contains 138 (Passardi et al., 2004). The homology between paralogs in a plant ranges from 30% to 100%, but very close orthologs exist, even between evolutionarily distant plants. All plant peroxidases contain invariant amino acids essential for their catalytic properties and for their proper folding (Welinder et al., 2002). They are structurally related to other heme-containing

* Corresponding author. Tel.: +41 223793012; fax: +41 223793017.
E-mail address: christophe.dunand@bota.unige.ch (C. Dunand).
[1] The two first authors have contributed equally to this work.

0031-9422/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2005.12.020

Table 1
Representation of the major plant lineages found in the PeroxiBase

Order (number of species/number of sequences): Genus
Angiosperms
Dicotyledons
Rosids
Cucurbitales (3/7): Cucumis (2x), Cucurbita
Fabales (14/314): Arachis, Cicer, Glycine, Lotus, Lupinus (2x), Medicago (2x), Phaseolus (2x), Pisum, Stylosanthes, Trifolium, Vigna
Rosales (3/8): Ficus, Malus, Urtica
Fagales (1/2): Quercus
Malpighiales (10/92): Euphorbia, Linum, Manihot, Mercurialis, Populus (6x)
Malvales (3/58): Gossypium, Theobroma
Sapindales (3/20): Citrus (2x), Poncirus
Brassicales (7/103): Arabidopsis, Brassica (2x), Armoracia, Raphanus, Thellungiella (2x)
Vitaceae (1/21): Vitis
Asterids
Gentianales (3/10): Coffea, Hedyotis (2x)
Lamiales (6/10): Avicennia, Eucommia, Orobanche (2x), Scutellaria, Striga
Solanales (11/199): Capsicum (2x), Ipomoea (2x), Lycopersicon (2x), Nicotiana (3x), Petunia, Solanum
Asterales (8/73): Artemisia, Cichorium, Helianthus (2x), Gerbera, Lactuca, Stevia, Zinnia
Apiales (1/1): Petroselinum
Ericales (1/1): Vaccinium
Saxifragales (1/6): Ribes
Caryophyllales (4/56): Beta, Mesembryanthemum, Mirabilis, Spinacia
Ranunculales (2/37): Eschscholzia, Aquilegia
Monocotyledons
Magnolids
Laurales (1/10): Persea
Magnoliales (1/11): Liriodendron
Piperales (1/0): Saruma
Liliopsida
Poales (15/655): Aegilops, Ananas, Avena, Cenchrus, Hordeum, Lolium, Oryza, Saccharum (2x), Secale, Setoria, Sorghum, Triticum (2x), Zea
Zingiberales (0/0) EST project in run
Liliales (2/2): Lilium, Alstroemeria
Acorales (1/5): Acorus
Alismatales (1/1): Spirodela
Arecales (1/1): Elaeis
Asparagales (3/36): Allium, Asparagus, Hyacinthus
Basal magnoliophyta
Austrobaileyales (1/0): Illicium
Nymphaeales (1/3): Nuphar
Amborellaceae (1/13): Amborella
Gymnosperms
Gnetopsida
Gnetales
Welwitschiales (1/1): Welwitschia
Coniferales (6/24): Cryptomeria, Pinus (3x), Picea (2x)
Ginkgoales (1/8): Ginkgo
Cycadales (3/12): Cycas, Zamia (2x)
Cryptogames (4/21)
Marchantia, Physcomitrella, Selaginella, Ceratopteris
Algae
Charales (0/0) no sequence available/activity found
Chlamydomonadales (0/0) no sequence available/no activity found

Number of species and of peroxidases found in each lineage are represented in brackets.

proteins, like peroxidases from prokaryotes, fungi. The key amino acids that interact with heme are also found in hemoglobins and cytochromes. The broad molecular diversification of plant peroxidases mainly results from gene duplication events. Newly duplicated genes were likely conserved because they acquired new modes of expression, regulation (subfunctionalization) or novel functions (neofunctionalization).

The automated annotation of the whole genomes of *Arabidopsis* (Arabidopsis Genome Initiative, 2000) and *Oryza sativa* (Goff et al., 2002; Yu et al., 2002), the automated clustering and assembling of EST sequences, and numerous EST projects led to the identification of a large number of sequences coding for class III plant peroxidases. We decided to construct a database devoted to this large, multigenic family because in our experience automated processing sometimes yields sequences of poor quality. Specificity is compromised and BLAST searching often requires manual sorting. Using the highly conserved motifs of the class III peroxidases (Welinder, 1992), manual annotation and editing can retrieve whole peroxidase sequences, that are unrecognized in automation due to poor quality sequences. *Arabidopsis* and rice are completely sequenced and are considered to be plant models, but they are not representative of plant diversity. The large number of EST projects developed with more diverse plants will provide a better overview of peroxidase evolution throughout green plants. The first goal of the PeroxiBase is to centralize most of the annotated and non-annotated class III peroxidase-encoding sequences and to make them publicly available, so that the research community has a unique tool for discovery, comparison, and exchange of peroxidase sequences. The second goal is to compile information concerning putative function and transcription regulation in order to facilitate cross-checking between close paralogs and orthologs. The final goal is to confirm the hypothesis that the number of class III isoforms increased after the emergence of the land plants.

2. Construction of the database

The database was constructed following two parallel procedures: one exhaustive and another more specific (Fig. 1). Firstly, the plant/fungal/bacterial heme peroxidase proteins are characterized by the motif PEROXIDASE_4 (http://ca.expasy.org/cgi-bin/nicesite.pl? PS50873). Using this signature, systematic data mining with MyHits (Pagni et al., 2004) from different predicted protein databases (TrEST, TrGEN) (Pagni et al., 2001) provides a global view of the peroxidase encoding sequences. The resulting hits are already treated data (assembled and translated sequences) with the risk of automatic compilation and translation.

We have then used AtPrx42 and OsPrx73, two sequences potentially related to an ancestral sequence (Passardi et al., 2004), for a second, more specific approach by scanning numerous public sources of plant ESTs and genome sequences in order to obtain a large collection of peroxidase-encoding sequences. For rare species, a tBLASTn was first performed against the NCBI whole database using limited queries for the date and for the organism. The Plant Genome Project (http://www.pgn.cornell.edu, 2004), Plant GDB (Dong et al., 2004) and Sputnik (Rudd, 2005) were used to complete the short sequences obtained from TIGR and to find new ones.

Each sequence obtained (assembly or singleton sequences) was individually translated and the presence of characteristic peroxidase motifs was verified using FingerPRINTScan and InterPro Scan softwares. Low quality

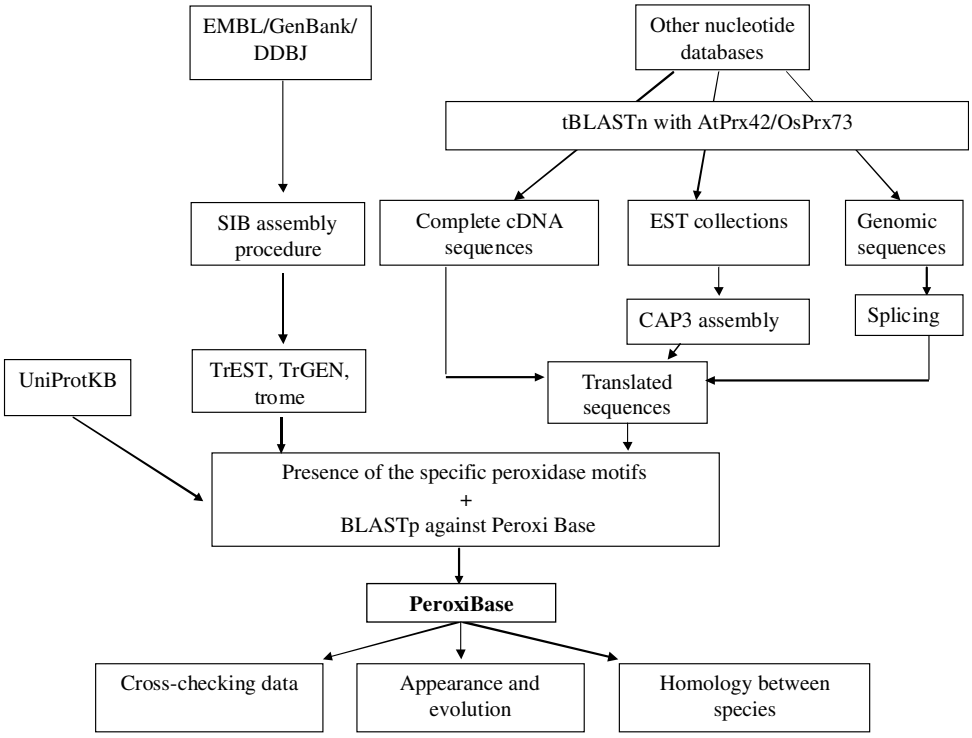

Fig. 1. Procedure of data analysis for generation of the PeroxiBase. Various EST and genomic databases have been used as sources of peroxidase encoding sequence.

sequences are not included in TIGR consensus sequences due to the high sequence stringency TIGR uses. Using peroxidase motifs as a guide, manual inspection of these poor quality sequences allows for increasing their length and assembling new sequences. All distinct sequences, even short ones, are kept in the database. We have also compared the UTR regions to confirm that sequences with high homology are truly distinct.

In addition, as in numerous other genera, ESTs from *Gossypium* (cotton), *Picea* (spruce), *Populus* (poplar) have been assembled by TIGR. However, the sequences used for the construction of the contigs are treated as if derived from a single species although they originate from different species. For example, cotton is derived from *Gossypium arboreum* and *hirsutum*, spruce from a collection of *Picea abies*, *Picea glauca* and *Picea sitchensis* and poplar from a mix of *Populus alba*, *Populus balsamifera*, *Populus euphratica*, *Populus kitakamiensis*, *Populus nigra*, *Populus tremula*, *Populus tremuloides*, and *Populus trichocarpa*. The Sputnik database (http://sputnik.btk.fi/) helped us unscramble this mixture of sequences for *Gossypium* and *Populus* and *Picea* species. The other species have been assembled directly from the NCBI entries.

After analysis of sequence alignment within each species (ClustalW and BioEdit), the protein sequences were individually entered in the database with their corresponding accession numbers as well as various information concerning the putative functions and transcription regulation (localization, induction and repression).

3. Web interface

The PeroxiBase web interface includes four main modules. (i) *Search*: this module enables a text query from the entire dataset with keywords such as tissue type, accession number, inducer/repressor, and name of sequences and organisms. (ii) *Organism*: each organism possesses a link with its taxonomic identity. Information for the peroxidases present in each organism can be viewed independently. Each file contains a direct link to the corresponding database (NCBI, TIGR, PGN, Sputnik) and to Swiss-Prot and DNA sequences when these entries exist. In addition, numbers of ESTs, cellular localization and tissue type are all included. The three closest homologous sequences, with their corresponding score and *E*-value are also described in the files. (iii) *BLAST*: two BLAST searches can be performed against the whole peroxidase database (Altschul et al., 1997), BLASTp for protein sequence and BLASTx for nucleotide sequences. The alignments are visualized and linked to the entry of each peroxidase. (iv) *FingerPrintscan*: this tool helps to find out which peroxidase family the sequence belongs to.

Three minor modules *Tissue type*, *Inducers/repressors* and *Cellular localization* are used as alternative ways of viewing sequences.

4. Modelization of the number of class III isoforms evolution

Class III peroxidase encoding sequences and peroxidase activity are both absent from the green alga *Chlamydomonas reinhardtii* (Passardi et al., 2004). On the contrary, in various *Chara* species, guaiacol oxidation in the presence of H_2O_2, specific to class III peroxidase activity (Greppin et al., 1986), can be detected (data not shown). The exhaustive data mining performed for the setup of the PeroxiBase confirms that the class III peroxidases are present in all land plants. For some species the total EST count is low (less than 1000), yet several independent isoforms were identified. In the case of large EST libraries (over 10,000), multigenic families with numerous putative class III peroxidases sequences can be found confirming previous results obtained with *Arabidopsis* and *Oryza* (Tognolli et al., 2002; Duroux and Welinder, 2003). In addition, the case of *Physcomitrella patens* seems to further validate the

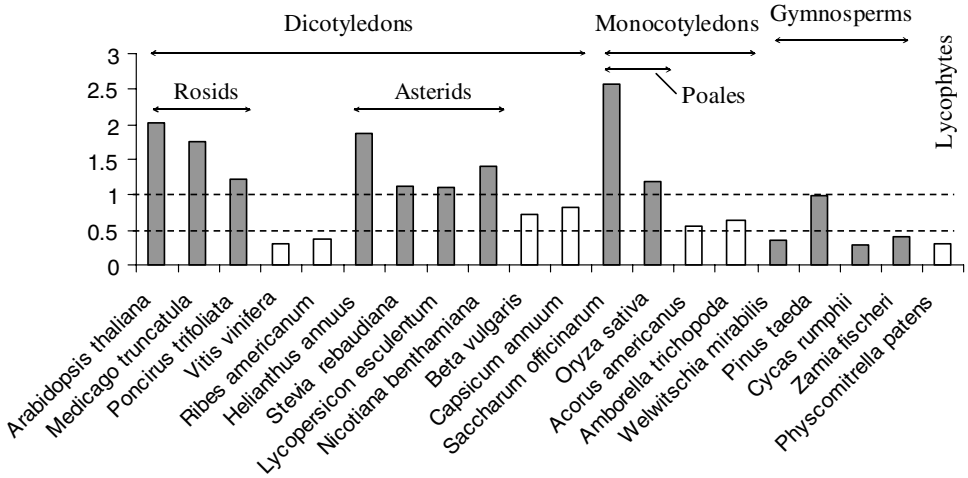

Fig. 2. Evolution of the peroxidase encoding genes versus the total number of genes in various key organisms. The *y*-axis represents a ratio obtained as following: (number of peroxidase encoding EST/number of independent peroxidase encoding genes)/(number of total EST/number of unigenes).

hypothesis that the number of peroxidase encoding genes increases over evolution. With more than 100,000 EST and only 12 peroxidase encoding sequences, the number of peroxidases sequences for this moss is far below those from other organisms such as *Arabidopsis*.

A potential evolution of peroxidase gene numbers can be drawn up based on information concerning each species such as total EST count and number of unigenes with the following formula: (number of peroxidase encoding EST/number of independent peroxidase encoding genes)/(number of total EST/number of unigenes). The value is proportional to the number of peroxidase genes in each species and gives information regarding the putative evolution of the peroxidase isoforms. Other completed genome sequencing and increasing EST sequences should confirm this hypothesis. Independently of the EST number, the size of the family seems also to follow a gradual increase from *Charales* (few isoforms) to the higher plants (numerous isoforms) confirming the previous hypothesis. The species can be classified in two major groups following the value of this ratio (Fig. 2). Rosids, Asterids and Poales considered to be higher plants, show a high diversification rate value over 1. On the other hand, species issued from basal Gymnosperms, and from small Mono- and Dicotyledons orders such as Vitaceae, Saxifragales, Caryophyllales, and Ranunculales have values around 0.5 or smaller.

5. Current status and future developments

The first goal of the PeroxiBase was to develop an efficient tool for the study of the evolution of a plant multigenic family. We tried not to be exclusive and to include as many sequences as possible from different organisms. The base currently consists of a core dataset containing over 2000 complete or partial peroxidase-encoding sequences from 125 organisms (Table 1), and it is still in constant evolution.

New peroxidase encoding sequences can be easily and directly added to the database by external people with individual user name and password. Continuous data mining will be performed until a putative complete analysis of the available sequences is achieved (EST and genomic sequences). At this point, a semi-automatic update will be set up to collect the peroxidase encoding sequences newly submitted to general databases (NCBI, Swiss-Prot). Information concerning the expression profile will also be updated and new features such as results of knock-out, knock-down or overexpression studies will be added when available.

The superfamily of plant, fungal and bacterial heme peroxidases contains *class I* (Cytochrome *C* peroxidase (EC 1.11.1.5), catalase peroxidase (EC 1.11.1.6) and ascorbate peroxidase (EC 1.11.1.11)), *class II* (lignin peroxidases (EC 1.11.1.14) and manganese peroxidases (EC 1.11.1.13)) and *class III*. The next update will include class I and class II peroxidases in the database. To our knowledge, the PeroxiBase will become the first repository devoted exclusively to a superfamily composed of multigenic families. The database could help to confirm the hypothesis that the three classes evolved from a single ancestral sequence (Zamocky, 2004). Another major addition will be to relate the major lineages containing peroxidase-encoding sequences to a schematic evolutionary tree. Peroxidases may become key markers for the evolution of plants, from as early as the first moments of land colonization to the human impact on genetics of cultivated plants today. The varied functions of peroxidases will be characterized and lead to a better understanding of plant growth, differentiation and interaction with the environment, and eventually to many exciting applications.

6. Useful web links

BioEdit: http://www.mbio.ncsu.edu/BioEdit/bioedit.html
ClustalW: http://www.ebi.ac.uk/clustalw/
Expasy translate: http://us.expasy.org/tools/dna.html
FingerPRINTScan: http://www.bioinf.man.ac.uk/fingerPRINTScan/
InterPro Scan: http://www.ebi.ac.uk/InterProScan/
MyHits: http://myhits.isb-sib.ch/cgi-bin/motif_query
NCBI: http://www.ncbi.nlm.nih.gov/
PlantGDB: http://zmdb.iastate.edu/PlantGDB/
Plant Genome Network: http://pgn.cornell.edu/
Reverse-Complement: http://bioinformatics.org/sms/rev_comp.html
SoftBerry- FGENESH: http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind
Sputnik: http://sputnik.btk.fi/ests
TIGR: http://www.tigr.org/tdb/tgi/plant.shtml

Acknowledgments

We thank Sonia Guimil for her critical reading. The financial support of the Swiss National Science Foundation (Grant 31-068003.02) to C.P. and C.D. is gratefully acknowledged, N.B. is paid by the Office Cantonal de l'Emploi.

References

Altschul, S.F., Madden, T.L., Schaffer, 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 Res. 25, 3389–3402.

Arabidopsis Genome Initiative, 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408, 796–815.

Dong, Q., Schlueter, S.D., Brendel, V., 2004. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 32, D354–D359.

Duroux, L., Welinder, K.G., 2003. The peroxidase gene family in plants: a phylogenetic overview. J. Mol. Evol. 57, 397–407.

Goff, S.A., Ricke, D., Lan, T.H., Presting, G., Wang, R., Dunn, M., Glazebrook, J., Sessions, A., Oeller, P., Varma, H., Hadley, D., Hutchison, D., Martin, C., Katagiri, F., Lange, B.M., Moughamer, T., Xia, Y., Budworth, P., Zhong, J., Miguel, T., Paszkowski, U., Zhang, S., Colbert, M., Sun, W.L., Chen, L., Cooper, B., Park, S., Wood, T.C., Mao, L., Quail, P., Wing, R., Dean, R., Yu, Y., Zharkikh, A., Shen, R., Sahasrabudhe, S., Thomas, A., Cannings, R., Gutin, A., Pruss, D., Reid, J., Tavtigian, S., Mitchell, J., Eldredge, G., Scholl, T., Miller, R.M., Bhatnagar, S., Adey, N., Rubano, T., Tusneem, N., Robinson, R., Feldhaus, J., Macalma, T., Oliphant, A., Briggs, S., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica). Science 296, 92–100.

Greppin, H., Penel, C., Gaspar, T., 1986. Molecular and Physiological Aspects of Plant Peroxidases. University of Geneva, Switzerland.

Pagni, M., Iseli, C., Junier, T., Falquet, L., Jongeneel, V., Bucher, P., 2001. trEST, trGEN and Hits: access to databases of predicted protein sequences. Nucleic Acids Res. 29, 148–151.

Pagni, M., Ioannidis, V., Cerutti, L., Zahn-Zabal, M., Jongeneel, C.V., Falquet, L., 2004. MyHits: a new interactive resource for protein annotation and domain identification. Nucleic Acids Res. 32, W332–W335.

Passardi, F., Longet, D., Penel, C., Dunand, C., 2004. The class III peroxidase multigenic family in rice and its evolution in land plants. Phytochemistry 65, 1879–1893.

Passardi, F., Cosio, C., Penel, C., Dunand, C., 2005. Peroxidases have more functions than a Swiss army knife. Plant Cell Rep. 24, 255–265.

Rudd, S., 2005. openSputnik: a database to establish comparative plant genomics using unsaturated sequence collections. Nucleic Acids Res. 33, D622–D627.

Tognolli, M., Penel, C., Greppin, H., Simon, P., 2002. Analysis and expression of the class III peroxidase large gene family in *Arabidopsis thaliana*. Gene 288, 129–138.

Welinder, K.G., 1992. Plant peroxidases: structure–function relationships. In: Penel, C., Gaspar, T., Greppin, H. (Eds.), Plant Peroxidases. University of Geneva, Switzerland.

Welinder, K.G., Justesen, A.F., Kjaersgard, I.V., Jensen, R.B., Rasmussen, S.K., Jespersen, H.M., Duroux, L., 2002. Structural diversity and transcription of class III peroxidases from *Arabidopsis thaliana*. Eur. J. Biochem. 269, 6063–6081.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Li, J., Liu, Z., Qi, Q., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Zhao, W., Li, P., Chen, W., Zhang, Y., Hu, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Tao, M., Zhu, L., Yuan, L., Yang, H., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica). Science 296, 79–92.

Zamocky, M., 2004. Phylogenetic relationships in class I of the superfamily of bacterial, fungal, and plant peroxidases. Eur. J. Biochem. 271, 3297–3309.