Published online 23 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D835–D842***
doi:10.1093/nar/gkp831

SALAD database: a motif-based database of protein annotations for plant comparative genomics

Motohiro Mihara[1], Takeshi Itoh[2] and Takeshi Izawa[1,*]

[1]Plant Genomics Research Unit and [2]Bioinformatics Research Unit, National Institute of Agrobiological Sciences, 2-1-2 Kannondai, Tsukuba, 305-8602, Japan

Received August 10, 2009; Revised September 16, 2009; Accepted September 19, 2009

ABSTRACT

Proteins often have several motifs with distinct evolutionary histories. Proteins with similar motifs have similar biochemical properties and thus related biological functions. We constructed a unique comparative genomics database termed the SALAD database (http://salad.dna.affrc.go.jp/salad/) from plant-genome-based proteome data sets. We extracted evolutionarily conserved motifs by MEME software from 209 529 protein-sequence annotation groups selected by BLASTP from the proteome data sets of 10 species: rice, sorghum, *Arabidopsis thaliana*, grape, a lycophyte, a moss, 3 algae, and yeast. Similarity clustering of each protein group was performed by pairwise scoring of the motif patterns of the sequences. The SALAD database provides a user-friendly graphical viewer that displays a motif pattern diagram linked to the resulting bootstrapped dendrogram for each protein group. Amino-acid-sequence-based and nucleotide-sequence-based phylogenetic trees for motif combination alignment, a logo comparison diagram for each clade in the tree, and a Pfam-domain pattern diagram are also available. We also developed a viewer named 'SALAD on ARRAYs' to view arbitrary microarray data sets of paralogous genes linked to the same dendrogram in a window. The SALAD database is a powerful tool for comparing protein sequences and can provide valuable hints for biological analysis.

INTRODUCTION

Genomic sequence data on several plant species are now available (1–10). This information has enabled bioinformatics researchers to predict non-redundant proteome data sets for those species. Therefore, all possible coding protein sequences have been annotated to create genome-based proteome data sets for those species on the basis of expressed sequence tag and full-length cDNA sequence information, computational gene prediction, and homology with known proteins. Related public databases are open and widely used (1–10).

The biological and biochemical functions of many proteins (or protein-coding genes), however, are not yet well elucidated. Searching for evolutionarily conserved protein domains in protein-sequences is very useful for predicting the biological function of proteins (or protein-coding genes) of interest. Accordingly, functional domain databases such as Pfam (11), SMART (12), PROSITE (13) and InterPro (14) are well known and widely used. Although these databases including manually curated ones, provide high-quality domain information, they focus on the functional domains themselves rather than on the biological functions of proteins. Curated functional domains of plant proteins, however, have not been identified enough in these domain databases yet. For example, when we recently performed a simple Pfam domain search genome-widely we were able to find that only ~150 000 plant proteins contain at least one Pfam domain among ~250 000 plant proteins tested.

Proteins often include several motifs (or domains) of different evolutionary origins. In this situation, typical phylogenetic analysis, in which significantly common multiple alignment is a prerequisite, is sometimes not adequate to predict the biological function of annotated proteins. Therefore, we hypothesized that patterns of evolutionarily conserved peptide sequences (or motifs) in each protein-sequence would reflect the biochemical functions of the annotated proteins and then, constructed a plant comparative genomics database termed SALAD (Surveyed conserved motif ALignment diagram and the Associating Dendrogram) database.

This genome-wide database is based on a similarity clustering by original scoring of distribution patterns of evolutionarily conserved motifs for all possible sequence pairs in a 'high percent similarity' protein group. The 209 529 'high percent similarity' protein groups were formed by BLASTP search (15) using each protein sequence as a query for a proteome data set of 250 687

*To whom correspondence should be addressed. Tel: +81 29 838 7446; Fax: +81 29 838 7446; Email: tizawa@nias.affrc.go.jp

Table 1. Version and species information in the SALAD database

Species	Category	SALAD database version			Source	Reference
		ver. 1.0	ver. 2.0	ver. 3.0		
Oryza sativa	Monocot	✓	✓	✓	RAP-DB (release 2)	(1)
Sorghum bicolor	Monocot		✓	✓	JGI (Sbi1_4)	(2)
Arabidopsis thaliana	Eudicot	✓	✓	✓	MIPS[1], TAIR (TAIR8)[2,3]	(3,4)
Vitis vinifera	Eudicot			✓	French-Italian Public Consortium	(5)
Selaginella moellendorffii	Lycophyte			✓	JGI (v1.0)	–
Physcomitrella patens	Moss		✓	✓	JGI (v1.1)	(6)
Ostreococcus tauri	Green alga		✓	✓	JGI (v2.0)	(7)
Chlamydomonas reinhardtii	Green alga			✓	JGI (v4.0)	(8)
Cyanidioschyzon merolae	Red alga	✓	✓	✓	*Cyanidioschyzon merolae* Genome Project	(9)
Saccharomyces cerevisiae	Yeast		✓	✓	Saccharomyces Genome Database (SGD1.01.50[2], SGD1.01.54[3])	(10)

Superscript numbers in *A. thaliana* genome source indicate SALAD database versions 1, 2 and 3, respectively.

protein sequences of 10 species (1–10). The SALAD database can provide valuable information for plant researchers for the design of molecular biology experiments and for elucidating the biological function of the proteins. In particular, it will be very useful for translating the knowledge from model plants such as *Arabidopsis thaliana* into possible biological functions of related sequences. As an example of data in SALAD database, we here present data that was unintentionally consistent with data in recently published articles on structural analyses of the phytohormone gibberellin receptor GID1. Furthermore, we report on our development of a viewer termed 'SALAD on ARRAYs' (16), which enables users to compare any microarray data of paralogous genes in a window of the database.

MAKING OF THE SALAD DATABASE

Selection of proteome data sets

To extract evolutionarily conserved motifs from homologous protein groups, the use of non-redundant proteome data sets is a prerequisite. Therefore, we originally evaluated some public genome-sequence data sets against the following criteria: (i) an assembled, non-redundant sequence for most of the genome was registered to public databases; (ii) one representative amino acid sequence with a certain locus (or annotation) ID code for one locus was assigned and (iii) the ratio of sequences containing apparent premature stop codons was low. The frequency of such proteins with premature stop codons can be used as an indicator for annotation quality. With these criteria, we selected three proteome data sets (rice, *A. thaliana* and red alga) in 2006 to start the construction of version 1 of the database (Table 1), and we released this version in October 2008. We next selected seven proteome data sets (rice, sorghum, *A. thaliana*, moss, green alga, red alga and yeast) in 2008 and released that version in March 2009 (Table 1). This year, with the same criteria, we selected three more data sets for version 3, for a total of 10 proteome data sets: rice (*Oryza sativa*), sorghum (*Sorghum bicolor*), *A. thaliana*, grape (*Vitis vinifera*), lycophyte (*Selaginella moellendorffii*), moss (*Physcomitrella patens*), green algae (*Ostreococcus tauri*, *Chlamydomonas reinhardtii*), a red alga (*Cyanidioschyzon merolae*) and a yeast (*Saccharomyces cerevisiae*) (as an outgroup). We released version 3 in August 2009 (Table 1). All processed data sets of versions 1, 2 and 3 are now available online (http://salad.dna.affrc.go.jp/salad/). Users can select appropriate dataset according to their purposes.

Making of 'high percent similarity' protein groups

We first performed BLASTP search for each predicted coding sequence in the dataset as a query. Based on the BLASTP results, the corresponding annotations were selected in ascending order to make a 'high percent similarity' protein group for each protein annotation. The threshold was less than 1.0*e*–5 of P value and the number limit in a 'high percent similarity' protein group was at maximum 70. This number, '70', was mainly determined by two reasons. A bunch of MEME analysis of data with many sequences requires a big PC power. Therefore, considering the efficiency of data analysis, '70' was in a sort of limitation at the moment. In addition, when users view the SALAD data, 70 proteins in a dendrogram is good enough size to overview the entire data at a glance in a typical PC display. In the latest version (version 3) (Table 1), the total number of 'high percent similarity' protein groups was 209 529, derived from the 250 687 BLASTP results for all sequences of the 10 proteome data sets.

The clustering method in SALAD database

We next extracted evolutionarily conserved motifs (8–50 amino acids) from each 'high percent similarity' protein group by using MEME software (EM algorithm) (17–19). Note that we randomized the order of amino acid sequences in the group every time before the use of MEME to increase the entire efficiency of motif extraction. Then we got a set of motifs [i.e. a set of PSSM (Position Specific Score Matrix)] found by MEME for each 'high percent similarity' protein group. Here, we considered both the presence/absence of motifs in the group and the similarity between amino acid sequences in

corresponding motifs to get the pairwise score for all possible pair of proteins in the group. The pairwise score was calculated in the following. When a certain motif found by MEME in a 'high percent similarity' protein group existed in both proteins, the score (Score1) was calculated using amino acid sequences in the corresponding motif by an amino acid substitution matrix. When the certain motif existed in either of proteins, the similarity score (Score2) for all possible amino acid sequences with the same length with the motif was calculated using the same amino acid substitution matrix. When the certain motif did not exist in either proteins, we empirically decided to give an average score (Score3) between Score1 and Score2 as a similarity score for the corresponding motif. We summed such scores (Score1, Score2 or Score3) for all the motifs found by MEME in the 'high percent similarity' group as the pairwise score. This pairwise score of similarity between proteins was used as the distance for clustering. In this way, all proteins in a given 'high percent similarity' protein group were clustered into a bootstrapped dendrogram by the pvclust routine in R software (http://www.r-project.org/). For each 'high percent similarity' protein group, this bootstrapped dendrogram linked to the corresponding motif-pattern diagram is presented for users in the SALAD database viewer (Figure 1).

WEB APPLICATION

Search system

Browsing of the SALAD database (version 3) starts with the selection of one 'high percent similarity' protein group from the 209 529 groups. Users can retrieve any protein group by a keyword search for gene ID, gene name, or gene function or by a BLAST search of the 250 687 sequences (amino acid or nucleotide) of the 10 species. The descriptive retrieval information for this keyword search is derived from annotation information from the web-site of National Center for Biotechnology Information (NCBI) and original databases such as TAIR (3) and RAP-DB (1).

SALAD data viewer

The SALAD database provides a user-friendly graphical viewer that displays SVG-formatted output, which contains a motif pattern diagram linked to a bootstrapped similarity dendrogram (Figure 1). In this viewer, the output can be manipulated on the display of the personal computer (e.g. zooming in or out and moving the output) (Figure 1A). The viewer also contains various functions such as graphical alignment to arbitrary motifs with some highlighted coloring; drawing of a neighbor-joining (NJ) tree for alignment of multiple motif sequences, which users can freely select (see below); a link to a Pfam-pattern diagram in the same dendrogram order; and a link to a description-list in the same dendrogram order for each homologous group. Each gene annotation has references to external databases [e.g. RAP-DB, TAIR, ATTED-II (20), etc.] and related internal data.

As an example, the data of a protein group selected by use of Os05g0407500 as a key word are presented in Figure 1. The Os05g0407500 gene (or *GIBBERELLIN INSENSITIVE DWARF1, GID1*) encodes a rice gibberellin (GA) receptor. Motif 12 located at the N terminal of GID1 related proteins, shown in dark green, was observed only in sequences of rice, sorghum, *A. thaliana*, grape and lycophyte, and not in the moss sequences. This N-terminal motif is functionally very important for interactions between the GA receptor (GID1) and its interacting target protein (DELLA) as mediated by GA, and for controlling the stability of this GID1–GA–DELLA complex under the regulation of ubiquitin and the 26S proteasome (21,22), and the moss GID1-like proteins have been shown not to bind to GA (23), as mosses do not have that signaling pathway (21,23).

Motif phylogenetic tree viewer

As shown in Figure 2, the motif phylogenetic tree viewer in the SALAD database provides both amino acid–sequence-based and nucleotide-sequence-based NJ trees (with bootstrapped values) for each motif or for arbitrarily combined MEME motifs. Users can select any motifs of interest by inputting motif ID numbers (Figure 1A-2) to make the NJ tree. The motif sequence alignment corresponding to the NJ tree is displayed beside the tree. In addition, a logo-comparison diagram between the user-selected clades of NJ trees can be created to compare conservation of sequences (or sites) among each clade by use of WebLogo software (24). Examples are shown for the phylogenetic tree of the GID1 group and the corresponding logo comparison diagram for the angiosperm clade and the lycophyte clade (Figure 2A and B). The amino acid locus of no. 52 is fixed to isoleucine (Ile, I) in the angiosperm clade (Figure 2B). Recently, it was shown experimentally that this isoleucine plays an important role in more specific and sensitive recognition of GA_4, an active endogenous GA in higher plants, than other amino acids such as leucine (Leu, L) and valine (Val, V) (21). Another example of a logo diagram is shown to demonstrate a clear conservation only in nucleotide sequences of an miRNA target, *miR156*, in a clade of an NJ tree for the SBP transcription factor family (Figure 2C).

SALAD on ARRAYs viewer

Recently a new viewer, called 'SALAD on ARRAYs' was incorporated into our SALAD database (16). This viewer provides gene expression data for paralogous genes from microarray data sets linked to the dendrogram of the SALAD database in a window (Figure 3). The gene expression level is shown as a gray-scale gradient in the colored boxes (Figure 3). Here, any public microarray data set can be put into this SALAD on ARRAYs viewer upon user's request. Therefore, users can easily find which paralogous genes are highly expressed (or not) in the microarray data of their interest. So far, the laser microdissection (LM) microarray data on rice pollen development are available for viewing through this SALAD on ARRAYs viewer (16,25–27) (http://

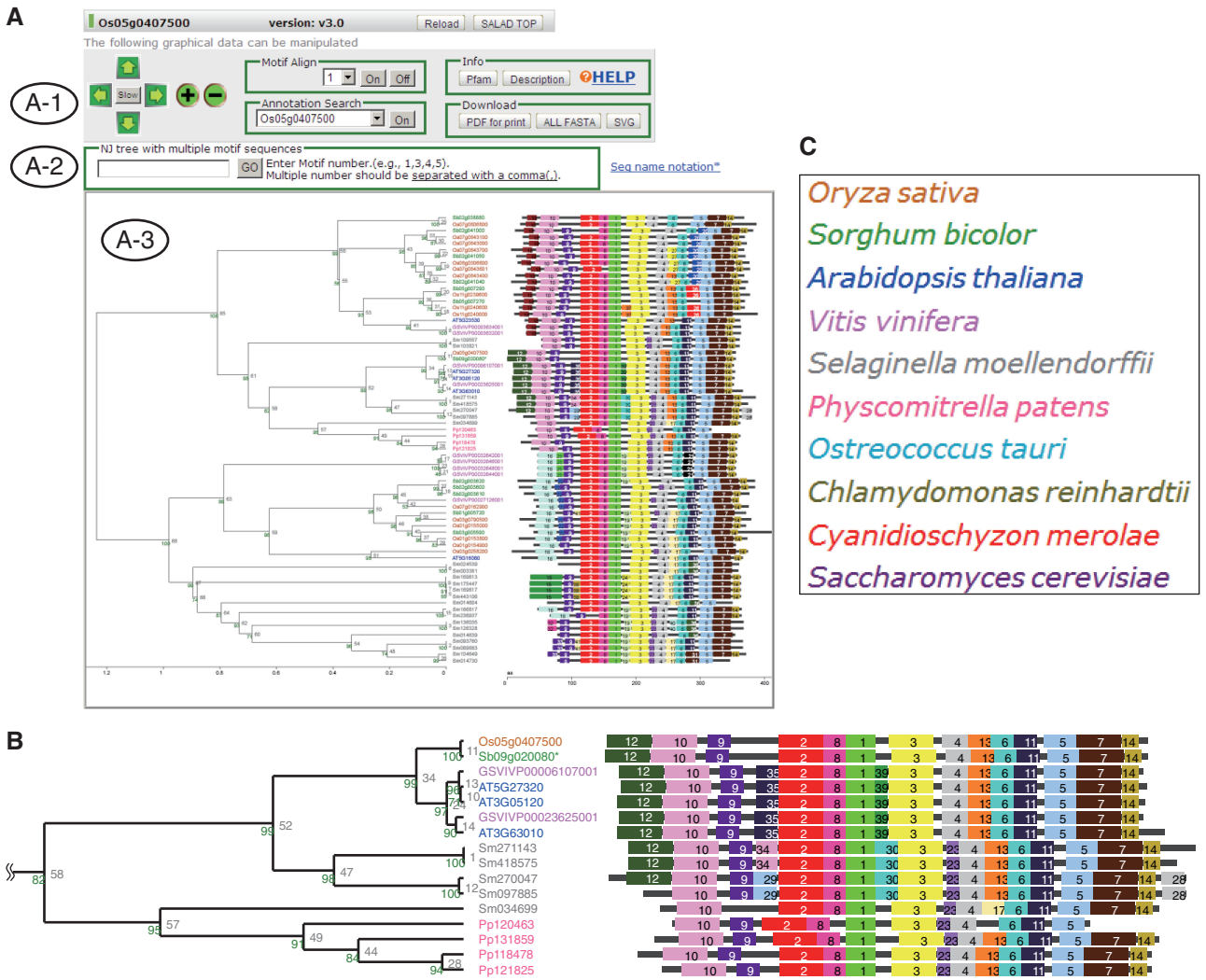

Figure 1. SALAD database viewer. (**A**) Data of the GID1 (Os05g0407500) group. (A-1) Operation panel for manipulating the output, and buttons for 'Motif Align', 'Annotation Search', and 'Download'. The 'Pfam' button brings up the pattern diagram of Pfam domains of each sequence. The 'Description' button brings up the annotation list derived from NCBI or the original database (such as TAIR, RAP-DB, etc.). (A-2) Toolbox for constructing phylogenetic trees based on sequence alignments of selected multiple motifs. Users can input the motif ID numbers of interest to make an NJ tree for motif alignment. (A-3) Typical SALAD analysis results come in two parts: a dendrogram of sequences clustered according to the presence and similarity of extracted conserved motifs, and a diagram that displays positional information of the extracted motifs in each sequence. (**B**) Expansion section around GID1. Each motif is assigned to a sequence number and color in the 'high percent similarity' protein group, and the same color box indicates the same extracted motif. (**C**) Color key of species.

salad.dna.affrc.go.jp/CGViewer/MicroArrayPollen/). A SALAD on ARRAYs window for the LM microarray data of GA3-oxidase (*OsGA3ox*), which is related to GA biosynthesis, is shown in Figure 3. One of the *GA3ox* genes is highly expressed in the tricellular pollen of rice (stage TC).

Interactive analysis

The SALAD database also provides an interactive analysis page, where users can submit a query sequence set (in multi-FASTA format) to the SALAD analysis output on the Web (http://salad.dna.affrc.go.jp/CGViewer/en/cgv_upload.html). Using this function, users can view SALAD database-style data for any protein group other than the 209 529 groups of 10 species.

DISCUSSION AND CONCLUSIONS

The SALAD database is a genome-wide protein-comparison database that was developed to connect the biological information of well-characterized proteins (or protein-coding genes) of a plant species with other related (but uncharacterized) proteins according to similarity of amino acid sequences. In the latest version (ver. 3.0), the database contains data on 250 687 protein sequences of 10 plant species, and the extracted motifs including those conserved only in land plants, higher plants, monocots, dicots and other categories. Therefore, users can compare any proteins with motif sequences conserved among the 10 species. As shown as an example of SALAD clustering (Figure 1), land plants share a characteristic N-terminal motif (assigned to

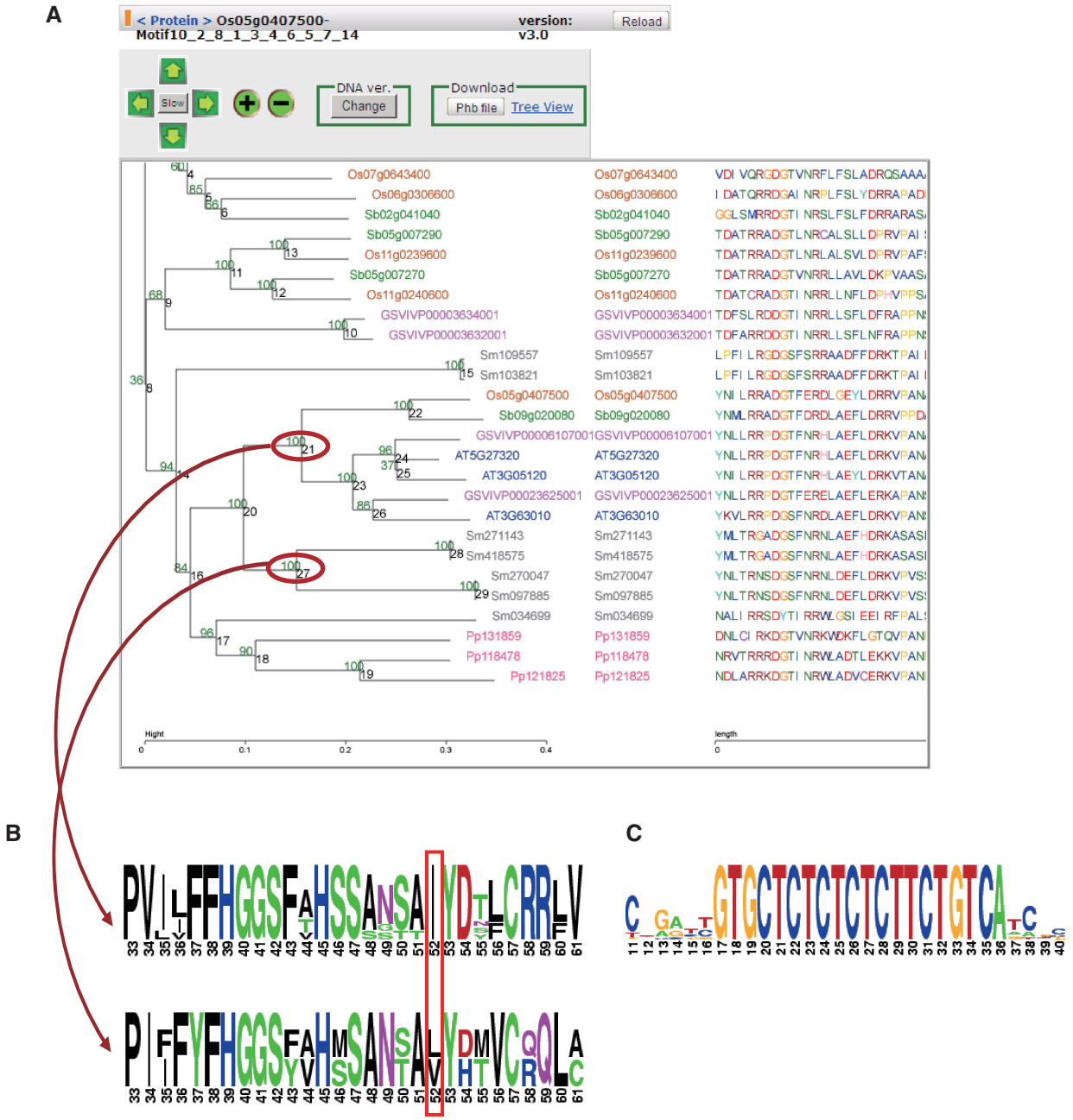

Figure 2. Motif phylogenetic tree view in SALAD database. (**A**) Motif phylogenetic tree view of the GID1 (Os05g0407500) group. A phylogenetic tree based on selected motifs was constructed. The 'Change' button enables one to switch the phylogenetic tree back and forth between nucleotide and amino acid alignments. Motif nucleotide or amino acid sequences are displayed to the right of the tree. (**B**) Logo comparison diagrams of two conserved clades in the phylogenetic tree; the upper one shows GID1-related sequences of flowering plants and the lower one shows those of the lycophyte *S. moellendorffii*. The red box highlights an amino acid site with a functionally important change (21). (**C**) A nucleotide-sequence logo diagram shows the binding site of miR156 for motif alignment in an protein family of an SBP transcription factor (http://salad.dna.affrc.go.jp/CGViewer/en/v3.0/cgv_motif_view.jsp?pfamid = AT3G60030:17).

no. 12) but moss does not (Figure 1B). This N-terminal motif has not yet been registered in any other public domain databases. This example clearly indicates that our SALAD database can provide biologically relevant information.

The SALAD database can provide other useful information as well. For example, the presence of Leu-133 or Val-133 in place of Ile-133 makes recognition of GA_4 (specificity and sensitivity) by the fern GID1 significantly lower than that by GID1s of flowering plants (21). This type of amino acid substitution can easily be found by creating a logo comparison diagram in the SALAD database (Figure 2B). The logo diagram can be also created for clades in a nucleotide sequence NJ tree. For

Figure 3. SALAD on ARRAYs viewer. Gene expression profiles of paralogous genes in the SALAD on ARRAYs viewer corresponding to the OsGA3ox1 (Os05g0178100) group. The figure shows the LM microarray data on rice pollen development (16,25–27). The gene expression level is displayed as a gray-scale gradient in the colored boxes. Spaces beside the *A. thaliana* gene annotation are filled by gene description retrieved from TAIR and NCBI. MEI, meiotic phase; TET, tetrad phase; UN, unicellular microspore; BC, bicellular; TC, mature tricellular pollen.

example, in the binding site of an miRNA, *miR156*, the nucleotide sequence was highly conserved in the logo diagram of the SALAD database (Figure 2C). In this way, users can compare not only motif patterns, but also sequences of motifs to speculate on the biological function of any proteins of interest.

The SALAD database is focused on the entire protein sequences rather than on evolutionarily conserved motifs, differentiating it from other domain databases such as Pfam and Interpro. In the SALAD database, the pairwise similarity between protein sequences is scored by all-or-none information on motifs and an amino acid substitution matrix. The clustering based on this scoring distinguishes the SALAD clustering from those in other comparative genomics databases such as GreenPhyl, in which distinct motifs are not well defined for each protein sequence (28). Although GreenPhyl is a specialized database for searching evolutionary orthologs between *A. thaliana* and rice, the SALAD database is specialized for inferring the biological function of uncharacterized related proteins in plant kingdoms. For example, as shown in the case of N terminal motif in the GID family (Figure 2), if a user may find a unique motif specific to some phylogenetic clade in SALAD database, the motif may reflect an important biological function.

To infer the biological function of uncharacterized proteins (or protein-coding genes), SALAD users would want information on the analyzed protein sequences, such as literature. To connect the motif-based information with such linguistic information, we adopted the Pfam domain diagram in the SALAD clustering dendrogram. The Pfam domain information is easily obtained through the link in the Pfam diagram page of the SALAD database. We also provide description lists linked to the SALAD clustering dendrogram for each protein group to support the inference of biological functions by users. For users who are interested in proteins in a species other than the 10 species, we provide an interactive analysis page to submit a list of any protein sequences for SALAD analysis.

The SALAD on ARRAYs viewer allows users to compare various microarray gene expressions of paralogous (or related) genes in a SALAD clustering window. As in Figure 3, one can easily compare expression patterns among paralogous genes. In this case, data of *OsGA3ox1* (Os05g0178100) and *OsGA3ox2* (Os01g0177400) (29), which are related to GA biosynthesis, are shown in the viewer. Using this viewer, it is easy to compare their expression patterns and to find differences among the microarrays registered: *OsGA3ox1* was well expressed in mature pollen (stage TC) but *OsGA3ox2* was not expressed at any stages (Figure 3). Upon request, we are ready to register any public microarray data into the SALAD database. In addition, we are now incorporating more than 1000 publicly available microarray data of *A. thaliana* from AtGenExpress in NCBI GEO repository (30–32) into SALAD on ARRAYs.

We believe that these functions of the SALAD database will provide researchers with many hints for designing molecular biology studies and will help to elucidate the biological functions of proteins (or protein-coding genes).

ACKNOWLEDGEMENTS

The authors thank the members of the LM-Microarray of Rice Pollen Project for providing microarray data. Some sequence data were produced by the US Department of Energy's Joint Genome Institute (http://www.jgi.doe.gov/) in collaboration with the user community. We also thank IRGSP for rice genome sequencing, the JGI, and other consortiums for providing genome data sets.

FUNDING

Ministry of Agriculture, Forestry and Fisheries of Japan (Agrobiological Genomics, GD1003; Genomics for Agricultural Innovation, GIR1002, RTR0004 to T. Izawa, and GIR1001 to T. Itoh). Funding for open access charge: National Institute of Agrobiological Sciences (Japan).

Conflict of interest statement. None declared.

REFERENCES

1. Tanaka,T., Antonio,B.A., Kikuchi,S., Matsumoto,T., Nagamura,Y., Numa,H., Sakai,H., Wu,J., Itoh,T., Sasaki,T. *et al.* (2008) The rice annotation project database (RAP-DB): 2008 update. *Nucleic Acids Res.*, **36**, D1028–D1033.
2. Paterson,A.H., Bowers,J.E., Bruggmann,R., Dubchak,I., Grimwood,J., Gundlach,H., Haberer,G., Hellsten,U., Mitros,T., Poliakov,A. *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature*, **457**, 551–556.
3. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
4. Schoof,H., Ernst,R., Nazarov,V., Pfeifer,L., Mewes,H.W. and Mayer,K.F.X. (2004) MIPS Arabidopsis thaliana Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.*, **32**, D373–D376.
5. Jaillon,O., Aury,J.M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, U463–U465.
6. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.F., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
7. Palenik,B., Grimwood,J., Aerts,A., Rouze,P., Salamov,A., Putnam,N., Dupont,C., Jorgensen,R., Derelle,E., Rombauts,S. *et al.* (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl Acad. Sci. USA*, **104**, 7705–7710.
8. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Marechal-Drouard,L. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–251.
9. Matsuzaki,M., Misumi,O., Shin-I,T., Maruyama,S., Takahara,M., Miyagishima,S.Y., Mori,T., Nishida,K., Yagisawa,F., Yoshida,Y. *et al.* (2004) Genome sequence of the ultrasmall unicellular red alga *Cyanidioschyzon merolae* 10D. *Nature*, **428**, 653–657.
10. Nash,R., Weng,S., Hitz,B., Balakrishnan,R., Christie,K.R., Costanzo,M.C., Dwight,S.S., Engel,S.R., Fisk,D.G., Hirschman,J.E. *et al.* (2007) Expanded protein information at SGD: new pages and proteome browser. *Nucleic Acids Res.*, **35**, D468–D471.
11. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L.L. *et al.* (2008) The Pfam protein families database. *Nucleic Acids Res.*, **36**, D281–D288.
12. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
13. Hulo,N., Bairoch,A., Bulliard,V., Cerutti,L., Cuche,B.A., de Castro,E., Lachaize,C., Langendijk-Genevaux,P.S. and Sigrist,C.J.A. (2008) The 20 years of PROSITE. *Nucleic Acids Res.*, **36**, D245–D249.
14. Hunter,S., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bork,P., Das,U., Daugherty,L., Duquenne,L. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–D215.
15. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
16. Mihara,M., Itoh,T. and Izawa,T. (2008) *In silico* identification of short nucleotide sequences associated with gene expression of pollen development in rice. *Plant Cell Physiol.*, **49**, 1451–1464.
17. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Second Int. Conf. Intell. Syst. Mol. Biol.*, **2**.
18. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
19. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
20. Obayashi,T., Hayashi,S., Saeki,M., Ohta,H. and Kinoshita,K. (2009) ATTED-II provides coexpressed gene networks for Arabidopsis. *Nucleic Acids Res.*, **37**, D987–D991.
21. Shimada,A., Ueguchi-Tanaka,M., Nakatsu,T., Nakajima,M., Naoe,Y., Ohmiya,H., Kato,H. and Matsuoka,M. (2008) Structural basis for gibberellin recognition by its receptor GID1. *Nature*, **456**, U520–U544.
22. Murase,K., Hirano,Y., Sun,T.P. and Hakoshima,T. (2008) Gibberellin-induced DELLA recognition by the gibberellin receptor GID1. *Nature*, **456**, 459–463.
23. Hirano,K., Nakajima,M., Asano,K., Nishiyama,T., Sakakibara,H., Kojima,M., Katoh,E., Xiang,H., Tanahashi,T., Hasebe,M. *et al.* (2007) The GID1-mediated gibberellin perception mechanism is conserved in the lycophyte Selaginella moellendorffii but not in the bryophyte Physcomitrella patens. *Plant Cell*, **19**, 3058–3079.
24. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: A sequence logo generator. *Genome Res.*, **14**, 1188–1190.
25. Hobo,T., Suwabe,K., Aya,K., Suzuki,G., Yano,K., Ishimizu,T., Fujita,M., Kikuchi,S., Hamada,K., Miyano,M. *et al.* (2008) Various spatiotemporal expression profiles of anther-expressed genes in rice. *Plant Cell Physiol.*, **49**, 1417–1428.
26. Hirano,K., Aya,K., Hobo,T., Sakakibara,H., Kojima,M., Shim,R.A., Hasegawa,Y., Ueguchi-Tanaka,M. and Matsuoka,M. (2008) Comprehensive transcriptome analysis of phytohormone biosynthesis and signaling genes in microspore/pollen and tapetum of rice. *Plant Cell Physiol.*, **49**, 1429–1450.
27. Suwabe,K., Suzuki,G., Takahashi,H., Shiono,K., Endo,M., Yano,K., Fujita,M., Masuko,H., Saito,H., Fujioka,T. *et al.* (2008) Separated transcriptomes of male gametophyte and tapetum in rice: validity of a Laser Microdissection (LM) microarray. *Plant Cell Physiol.*, **49**, 1407–1416.
28. Conte,M.G., Gaillard,S., Lanau,N., Rouard,M. and Perin,C. (2008) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res.*, **36**, D991–D998.
29. Itoh,H., Ueguchi-Tanaka,M., Sentoku,N., Kitano,H., Matsuoka,M. and Kobayashi,M. (2001) Cloning and functional analysis of two gibberellin 3 beta-hydroxylase genes that are differently expressed during the growth of rice. *Proc. Natl Acad. Sci. USA*, **98**, 8909–8914.
30. Goda,H., Sasaki,E., Akiyama,K., Maruyama-Nakashita,A., Nakabayashi,K., Li,W.Q., Ogawa,M., Yamauchi,Y., Preston,J., Aoki,K. *et al.* (2008) The AtGenExpress hormone and chemical treatment data set: experimental design, data

evaluation, model data analysis and data access. *Plant J.*, **55**, 526–542.
31. Kilian,J., Whitehead,D., Horak,J., Wanke,D., Weinl,S., Batistic,O., D'Angelo,C., Bornberg-Bauer,E., Kudla,J. and Harter,K. (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J.*, **50**, 347–363.
32. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Rudnev,D., Evangelista,C., Kim,I.F., Soboleva,A., Tomashevsky,M., Marshall,K.A. *et al.* (2009) NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.