Contents lists available at ScienceDirect

Phytochemistry

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

Update in Bioinformatics

IT3F: A web-based tool for functional analysis of transcription factors in plants

Paul C. Bailey [a,*], Jo Dicks [a], Trevor L. Wang [b], Cathie Martin [b]

[a] Department of Computational and Systems Biology, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK
[b] Department of Metabolic Biology, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK

ARTICLE INFO

Article history:
Received 30 July 2008
Accepted 7 August 2008
Available online 2 September 2008

Keywords:
Arabidopsis thaliana
Oryza sativa
Phylogenetics
Gene expression
Website
PHYLIP
Transcription factor
myb
bHLH

ABSTRACT

A web-based tool, the Interspecies Transcription Factor Function Finder (IT3F), has been developed to display both evolutionary gene relationships and expression data for plant transcription factors, focussing primarily on the R2R3MYB gene subfamily for proof of concept. The graphical display of information allows users to make direct comparisons between structurally related genes and to identify those genes that are potentially orthologous, thereby assisting with their understanding of gene function.

A key feature of the website is the provision of an interrogative phylogenetic tree that allows submission of new sequences corresponding to a transcription factor family or subfamily and maps their relative positions to the products of other genes on an 'existing' tree containing proteins encoded by Arabidopsis and rice genes, along with key proteins encoded by genes from other species that have been characterised functionally. In addition, a feature to select clusters of related sequences has been developed so that more detailed phylogenetic analysis can be performed to highlight potential orthologous and paralogous genes within related clusters. Arabidopsis genes that reside on duplicated regions of the genome are indicated on the tree, providing further information for interpreting gene function.

An additional feature of the website allows a selected number of key Arabidopsis and rice microarray experiments to be visualised alongside the tree as a tabulated heat map of expression intensity values. Through this display, it is possible to observe relative expression levels across a whole gene family and the extent to which the expression of closely related genes within subgroups has altered since their ancestral divergence.

The website is available at http://jicbio.bbsrc.ac.uk/IT3F/.

1. Introduction

Transcription factors (TFs) are an important class of proteins involved in regulating the production of mRNA transcripts from genes by binding to DNA *cis*-acting elements in the promoters that they target. TFs can be part of the basal transcription machinery or regulatory in nature, whereby they control specific groups of genes in particular cell types, time periods or environmental conditions. The website described here concerns families of the regulatory type of TF in higher plants, in particular those families whose gene members have been shown to be involved in the regulation of plant secondary metabolism.

There is considerable interest in understanding the different functions that TFs perform in plants due to their potential contribution to plant improvement; many TFs have been shown to turn on whole metabolic pathways or pathway branches by activating or repressing the transcription of sets of genes coordinately (for example, Grotewold et al., 1994) and over-expression of a TF can produce large changes in metabolism, overcoming the problem of flux control that has been encountered when the amount of only one enzyme of a pathway is increased by transgenesis (for example, Niggeweg et al., 2004; Luo et al., 2007).

It has been recognised for over 20 years that TFs from within and between species can be grouped into families based on shared characteristic amino acid signatures in their DNA binding domains (Liu et al., 1999). Following the complete sequencing of the *Arabidopsis thaliana* genome (Arabidopsis Genome Initiative, 2000), it is apparent that many of these TF families consist of a large number of proteins, termed 'superfamilies', some of which are specific to plants (Riechmann et al., 2000). The existence of large families of TFs in plants shows that the individual genes encoding proteins within each family have emerged, then evolved, from a common ancestor. There is now strong evidence that multiple gene duplication events over millions of years – the major affect being through whole genome duplications (see Section 2.3) – have played an important evolutionary role in shaping the functional diversity within each family (for example, Martin and Paz-Ares, 1997; Shiu et al., 2005).

A useful way to visualise the structural and functional diversity between the protein members of each TF family within a single species is to draw a phylogenetic tree using a protein alignment

[*] Corresponding author. Tel.: +44 1603 450805; fax: +44 1603 450021.
E-mail address: Paul.Bailey@bbsrc.ac.uk (P.C. Bailey).

0031-9422/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2008.08.003

based on the shared DNA binding domain alone. Given the large size of many TF families, a distance matrix method is most appropriate for this task because the calculations are fast even for large data sets; this type of method has also been shown to give results of sufficient quality in simulation studies (for example, Kuhner and Felsenstein, 1994). The resulting tree topologies for TF families show very clearly that many proteins in the family are encoded by structurally related genes that fall into discrete clades. Detailed studies of closely related genes from different species provide strong support for the conclusion that genes within a clade encode proteins which share similar biochemical functions. For example, AmMYB308 and AmMYB330 (Tamagnone et al., 1998) and AtMYB4 (Jin et al., 2000) belong to the same clade (subgroup 4) of the R2R3MYB family. All these proteins repress production of hydroxycinnamic acid derivatives. In some cases the biological function of proteins within a clade may diverge, depending on the cellular context in which the TFs are active (for example, Lee and Schiefelbein, 2001), a phenomenon explained by the 'cocktail party model' (Sieweke and Graf, 1998). Nevertheless, if proteins with a known biochemical function are added to the tree data set their presence often helps to place other related but as yet unstudied genes into context, indicating their likely function. Furthermore, the addition of a full complement of proteins in a family from another species can identify potential orthologues and inparalogues (homologues that have arisen via duplication after species divergence – Remm et al., 2001).

Studies on the analysis of whole TF families from Arabidopsis have focussed on gene structure, phylogenetic analysis, gene expression, gene redundancy and functional diversity (for example, the AP2/ERF family (Riechmann and Meyerowitz, 1998), the R2R3 MYB family (Stracke et al., 2001) and the bHLH family (Heim et al., 2003)). More recent studies have compared Arabidopsis and rice TF families (for example, bHLH (Li et al., 2006), MYB (Yanhui et al., 2006) and AP2/ERF (Nakano et al., 2006)). Unfortunately, there is no easy way to integrate new data with the existing knowledge available in these earlier reviews and there are few bioinformatics tools that provide an integrated view of the different data types. With other plant genome sequencing projects completed or in draft form for *Populus balsamifera* ssp. *trichocarpa* (black cottonwood [poplar], Tuskan et al., 2006), *Vitis vinifera* (grape, Jaillon et al., 2007), *Amborella trichopoda* (Soltis et al., 2008), *Carica papaya* (papaya, Ming et al., 2008), *Physcomitrella patens* (moss, Rensing et al., 2008) and with others on the way, there will be an increasing need to map genes from these new data sets to existing information requiring easy-to-use tools that facilitate comparisons between different species.

With the need for simple and rapid analysis of new data sets, we have developed a comparative analysis website tool to aid the functional characterisation of newly discovered gene family members. In particular, we wanted to create a website with the following features: (1) an interrogative tree; users would be able to submit protein 'query' sequence(s) from any species and instantly have them mapped onto a phylogenetic tree comprising all members of the corresponding TF family from Arabidopsis and rice. (2) A web page display that uses evolutionary relationships as an intuitive way to guide the display of other data types, for example, gene expression profiles; these data types for closely related members of a TF family, possibly orthologues, could then be compared directly to reveal similarities or differences between them.

Since the inception of the work described in this paper, a plethora of new websites has emerged that provide a list of gene members for all TF families (AGRIS, PlantTFDB that includes DATF and DRTF, RARTF and PlnTFDB), with some websites also providing information on phylogenetic relationships (PlanTAPDB and DATFAP) while the GreenPhyl website presents this information on the entire proteome (Table 1). However, to date these websites lack the features proposed above for our website. Our focus has been on developing tools for the R2R3 MYB family of plant specific TFs, but other TF families are included within the website to demonstrate that it can be used to analyse any TF family and other protein classes that possess discrete protein domains.

2. The creation of an interrogative gene tree

A Perl-CGI program (IT3F.pl) was written to chain together distinct methods into an analytical pipeline, in order to; add query sequence(s) to an existing and potentially large alignment of other gene family members comprising only the DNA binding domain (see Section 1), deal with file format issues, pass data into distance matrix programs and, finally, process a Newick format tree file output to generate a graphical depiction of the tree (Fig. 1). For the pipeline to work effectively, three critical components were developed, one to guarantee that the query sequences were aligned perfectly to the existing sequence alignment, one to ensure rapid completion of the distance matrix method and one component to display the tree.

2.1. Aligning the query sequence to an existing large alignment

The first critical component recognises and extracts the DNA binding domain from the query sequence and aligns it perfectly against a large 'existing' alignment of proteins encoded by gene family members that is based on alignment of the DNA binding domain only. A perfect alignment is one in which all amino acids of the query DNA binding domain align to their homologous counterparts in the existing alignment. The new alignment is passed to modified PHYLIP programs (Version 3.66). This software (Felsenstein, 2004) is used because it is one of the most widely cited packages in publications, is freely available and permission is granted for the source code to be modified, provided the copyright notice is not removed.

In early versions of the website, the alignment procedure used CLUSTALW, but a more robust alignment was subsequently found to be obtained with the HMMALIGN program from the HMMER

Table 1
Websites that provide gene and phylogeny data for plant TFs

Website	Acronym	Reference	Website address
Arabidopsis gene regulatory information server	AGRIS	Davuluri et al. (2003)	http://arabidopsis.med.ohio-state.edu/AtTFDB/
Plant transcription factor databases:	PlantTFDB	Guo et al. (2008)	http://planttfdb.cbi.pku.edu.cn/
Database of Arabidopsis transcription factors	DATF	Guo et al. (2005)	http://datf.cbi.pku.edu.cn/
Database of rice transcription factors	DRTF	Gao et al. (2006)	http://drtf.cbi.pku.edu.cn/
RIKEN Arabidopsis transcription factor database	RARTF	Iida et al. (2005)	http://rarge.gsc.riken.jp/rartf/
Plant transcription factor database	PlnTFDB	Riano-Pachon et al. (2007)	http://plntfdb.bio.uni-potsdam.de/v2.0/
Plant transcription associated protein database	PlanTAPDB	Richardt et al. (2007)	http://www.cosmoss.org/bm/plantapdb
GreenPhyl		Conte et al. (2008)	http://greenphyl.cirad.fr/cgi-bin/greenphyl.cgi
Database of Arabidopsis transcription factors with alignments and primers	DATFAP	Fredslund (2008)	http://cgi-www.daimi.au.dk/cgi-chili/datfap/frontdoor.py

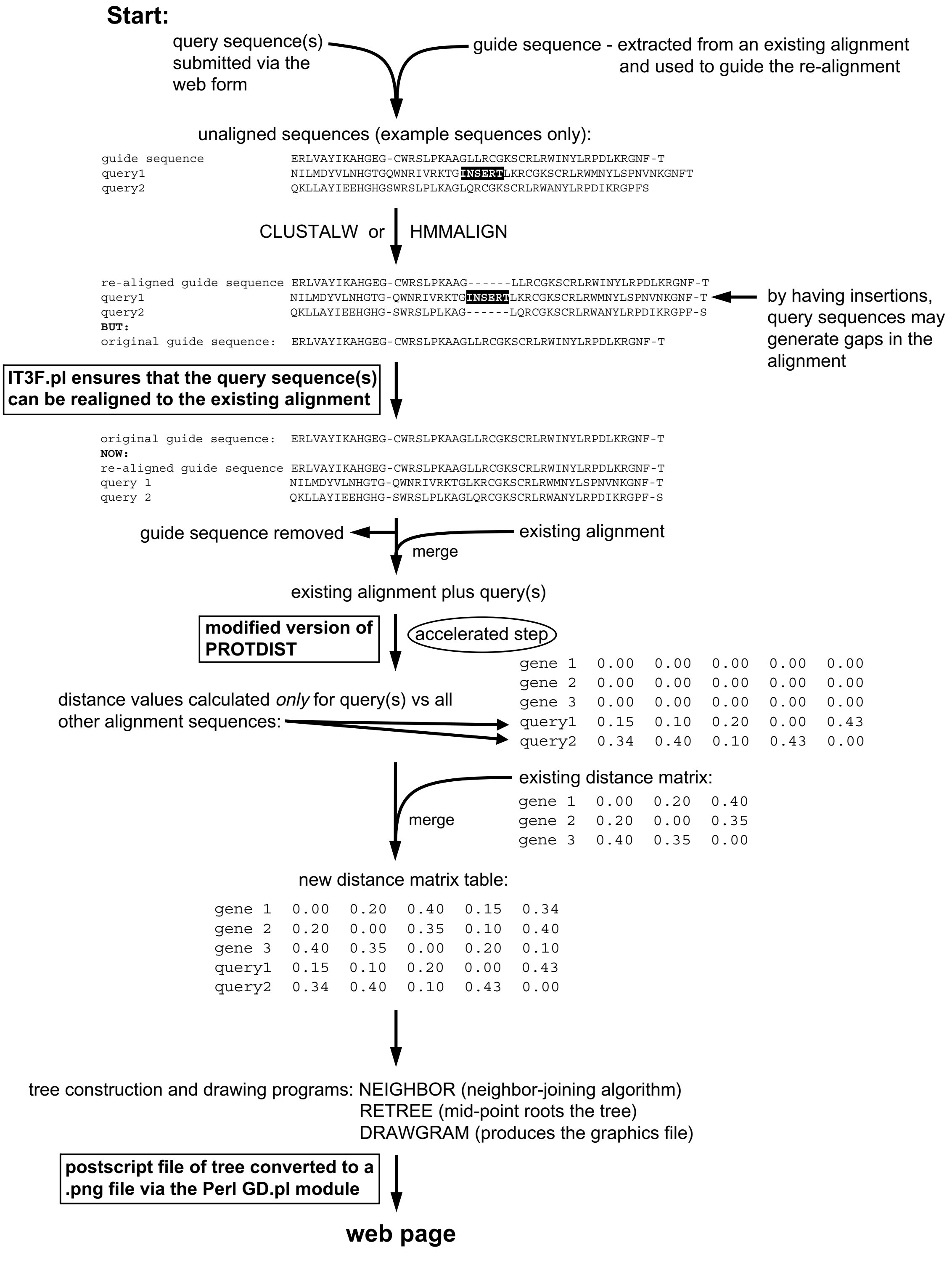
Start:
query sequence(s) submitted via the web form
guide sequence - extracted from an existing alignment and used to guide the re-alignment
unaligned sequences (example sequences only):
guide sequence ERLVAYIKAHGEG-CWRSLPKAAGLLRCGKSCRLRWINYLRPDLKRGNF-T
query1 NILMDYVLNHGTGQWNRIVRKTGINSERTLKRCGKSCRLRWMNYLSPNVNKGNFT
query2 QKLLAYIEEHGHGSWRSLPLKAGLQRCGKSCRLRWANYLRPDIKRGPFS
CLUSTALW or HMMALIGN
re-aligned guide sequence ERLVAYIKAHGEG-CWRSLPKAAG------LLRCGKSCRLRWINYLRPDLKRGNF-T
query1 NILMDYVLNHGTG-QWNRIVRKTGINSERTLKRCGKSCRLRWMNYLSPNVNKGNF-T
query2 QKLLAYIEEHGHG-SWRSLPLKAG------LQRCGKSCRLRWANYLRPDIKRGPF-S
BUT:
original guide sequence: ERLVAYIKAHGEG-CWRSLPKAAGLLRCGKSCRLRWINYLRPDLKRGNF-T
by having insertions, query sequences may generate gaps in the alignment
IT3F.pl ensures that the query sequence(s) can be realigned to the existing alignment
original guide sequence: ERLVAYIKAHGEG-CWRSLPKAAGLLRCGKSCRLRWINYLRPDLKRGNF-T
NOW:
re-aligned guide sequence ERLVAYIKAHGEG-CWRSLPKAAGLLRCGKSCRLRWINYLRPDLKRGNF-T
query 1 NILMDYVLNHGTG-QWNRIVRKTGLKRCGKSCRLRWMNYLSPNVNKGNF-T
query 2 QKLLAYIEEHGHG-SWRSLPLKAGLQRCGKSCRLRWANYLRPDIKRGPF-S
guide sequence removed
existing alignment
merge
existing alignment plus query(s)
modified version of PROTDIST
accelerated step
gene 1 0.00 0.00 0.00 0.00 0.00
gene 2 0.00 0.00 0.00 0.00 0.00
gene 3 0.00 0.00 0.00 0.00 0.00
query1 0.15 0.10 0.20 0.00 0.43
query2 0.34 0.40 0.10 0.43 0.00
distance values calculated *only* for query(s) vs all other alignment sequences:
existing distance matrix:
gene 1 0.00 0.20 0.40
gene 2 0.20 0.00 0.35
gene 3 0.40 0.35 0.00
merge
new distance matrix table:
gene 1 0.00 0.20 0.40 0.15 0.34
gene 2 0.20 0.00 0.35 0.10 0.40
gene 3 0.40 0.35 0.00 0.20 0.10
query1 0.15 0.10 0.20 0.00 0.43
query2 0.34 0.40 0.10 0.43 0.00
tree construction and drawing programs: NEIGHBOR (neighbor-joining algorithm)
RETREE (mid-point roots the tree)
DRAWGRAM (produces the graphics file)
postscript file of tree converted to a .png file via the Perl GD.pl module
web page

Fig. 1. Overview of the interrogative gene tree pipeline controlled by the IT3F.pl Perl-CGI program. The key steps developed in this work are contained in boxes. Use of HMMALIGN over CLUSTALW was considered to be the best alignment strategy for query sequences.

suite (Eddy, 1998) which can align query sequences to an HMM profile whilst guided by a corresponding alignment. However, an outstanding issue remains with the HMM-based procedure. Normally, if a query sequence is identical to one of the sequences in the existing alignment, the user would expect to observe these sequences appearing together in the tree next to a vertical line with no terminal branches. Sometimes this does not occur and the query sequence diverges from its matching sequence, being separated by a small branch. This artefact may be confusing to users who might infer that the two sequences represent different genes. The cause of this divergence is due to an alignment ‘wobble’, a difference between the alignment of the sequence in the pre-existing alignment and that of its identical query sequence, aligned slightly differently by the HMMALIGN program, particularly adjacent to gaps in the alignment. The alignment pipeline will be improved in the near future to overcome this minor problem.

2.2. Speeding up the distance matrix step

The second critical component accelerates the distance matrix method for large data sets so that a tree is returned to the web page within a few seconds. The rate-limiting step for a distance matrix method is the first step which is the calculation of the distance estimate for all sequence pairs in an alignment, with each pairwise comparison being calculated independently of all others in the data set. Therefore those comparisons for the existing Arabidopsis and rice sequences can be pre-calculated leaving only the pairwise comparisons involving the query sequence against all other sequences to be calculated dynamically while the user waits (Fig. 1). This shortcut saves a considerable amount of time (several minutes for large datasets). The only caveat to this time-saving approach is that it requires the query sequence to align to the existing alignment perfectly which the previous step described in Section 2.1 ensures.

To implement this idea, it was necessary only to modify the C source code of the PHYLIP PROTDIST program in one line of the `makedists()` function which controls an iterative loop to calculate the distance estimates for each sequence or species, `i`, against all others in turn:

```
for (i = 1; i ⩽ spp; i ++)  {...
```

This line states that for each species starting from species 1 (`i = 1`) at the top of the alignment, distance estimates will be calculated and the loop will end when all species (`i = spp`) in the alignment have been processed. In the modified PROTDIST version, the loop counter was set to the number corresponding to the first query sequence located at the bottom of the alignment, using a variable called `noNewSpp`:

```
for (i = noNewSpp; i ⩽ spp; i ++)  {...
```

The modified PROTDIST program was set to use the JTT model of protein evolution (Jones et al., 1992) but not to account for the variation in amino acid substitution rate between sites in the alignment using the gamma distribution.

To our knowledge this is the first report of manipulating distance matrices in this way to speed up a sequence-to-phylogeny pipeline for a web page significantly, although Zmasek and Eddy (2002) employed precalculation of pairwise distances to provide increased time efficiency in their Resampled Inference of Orthologs (RIO) procedure.

2.3. The tree display

For the tree display, the IT3F.pl program calls three other PHYLIP programs: NEIGHBOR to estimate a neighbor-joining tree (Saitou and Nei, 1987), RETREE to root the tree by the midpoint rooting method, and DRAWGRAM to convert the resulting Newick format tree file to postscript format. In the third critical component of the pipeline, the latter file is converted to a .png file using the Perl GD.pl wrapper module to the GD Graphics Library, assisted by IT3F.pl. The resulting .png file is displayed in the web page with query sequences highlighted in red, necessary for easy identification in large trees (Fig. 2). Users can also see the locus ID numbers for each gene alongside the tree. The alignment that gives rise to the tree is displayed at the bottom of the web page (not shown in Fig. 2).

Genes that are thought to have undergone duplication through large-scale genome duplication events, as inferred for Arabidopsis by Blanc et al. (2003), are indicated by dots to the right of the gene names in the tree. These genes in particular are likely to have closely related functions. Such information in the tree will help users to identify whether functional gene redundancy is likely and will indicate, for example in a reverse genetics experiment, whether a double gene knockout mutant is likely to be necessary to produce a phenotype for duplicated genes.

To our knowledge this is the first website that provides an interrogative tree in which multiple query sequences can be placed into established protein datasets. Although the GOST tool (Conte et al., 2008) at the GreenPhyl website is similar, identifying orthologous sequences to user query sequences, the query sequences are not placed on a returning tree.

3. Analysis of MYB and other TF sequence data sets

In addition to the MYB family, sequence data were prepared for the following plant TF families: B3 (VP1/ABI3-like), AP2 (AP2/ERF), bHLH, bZIP, HSF, TCP and WRKY (of these the B3, AP2, TCP and WRKY families are specific to plants). These data sets were obtained for Arabidopsis from DATF plus any additional genes present in PlnTFDB (Table 1). For rice, DRTF data sets were used in preference to PlnTFDB data sets because version 1.0 of the latter database contained no locus IDs which were critical to the success of the pipeline. At present, only the first version of each gene model is included in the trees (e.g., At1g01010.1, not the alternative splice forms, At1g01010.2 or At1g01010.3, if present). These latter gene models are likely to be identical across the region encoding the DNA binding domain so a large number of them would clutter the tree. The number of genes encoding proteins present in each tree is indicated in the pulldown menu of TF families on the home page and should contain most members of each family.

To obtain as accurate a phylogeny as possible, it is desirable to maximise the sequence length (i.e., the number of amino acid residues) over which the phylogeny can be inferred. It was possible to use longer sequences in situations where TF superfamilies contained subfamilies. The MYB superfamily has six subfamilies (R1R2R3, R2R3, SHAQKY, GARP, CAPRICE-like, telomere binding protein-like) and each has inherited a particular number of MYB motifs; the AP2/ERF family has two subfamilies, one inheriting two AP2 DNA binding motifs, the other inheriting just one motif. Clearly, if a subfamily contains two motifs, whereas another subfamily contains one, the conserved alignment of the DNA binding domain will be approximately twice as long and therefore contain twice as much evolutionary information. In these cases, subfamily trees have been provided rather than a supertree for the entire TF family. The separation of these subfamilies into different trees provides further natural boundaries for defining functional categories and augments the work by Riano-Pachon et al. (2007) to classify groups of TF-related genes into functional categories based on their shared evolutionary past.

Proteins encoded by thirty-nine genes from other species for which functions are known were included in the R2R3MYB data sets. Including these proteins helps to provide a clearer tree topology for inferring the putative functions of uncharacterised, but closely related proteins. Although these proteins must be added to the data sets ‘by hand’, users of the website can also use known proteins as ‘queries’. They can keep a list of favourite proteins from the literature or new proteins emerging from sequencing projects and periodically interrogate the tree at a click of a button to refine their knowledge of protein relationships.

The main purpose of the large trees representing whole TF families is to define subgroups of related sequences. The distance matrix method targeted against the DNA binding domain achieves this successfully; subgroups are separated clearly from other subgroups by relatively long branches indicating an old divergence event. However, the relationships between the subgroups in the tree are often poorly defined with the nodes towards the root of the tree having low bootstrap support values (below 70% – data

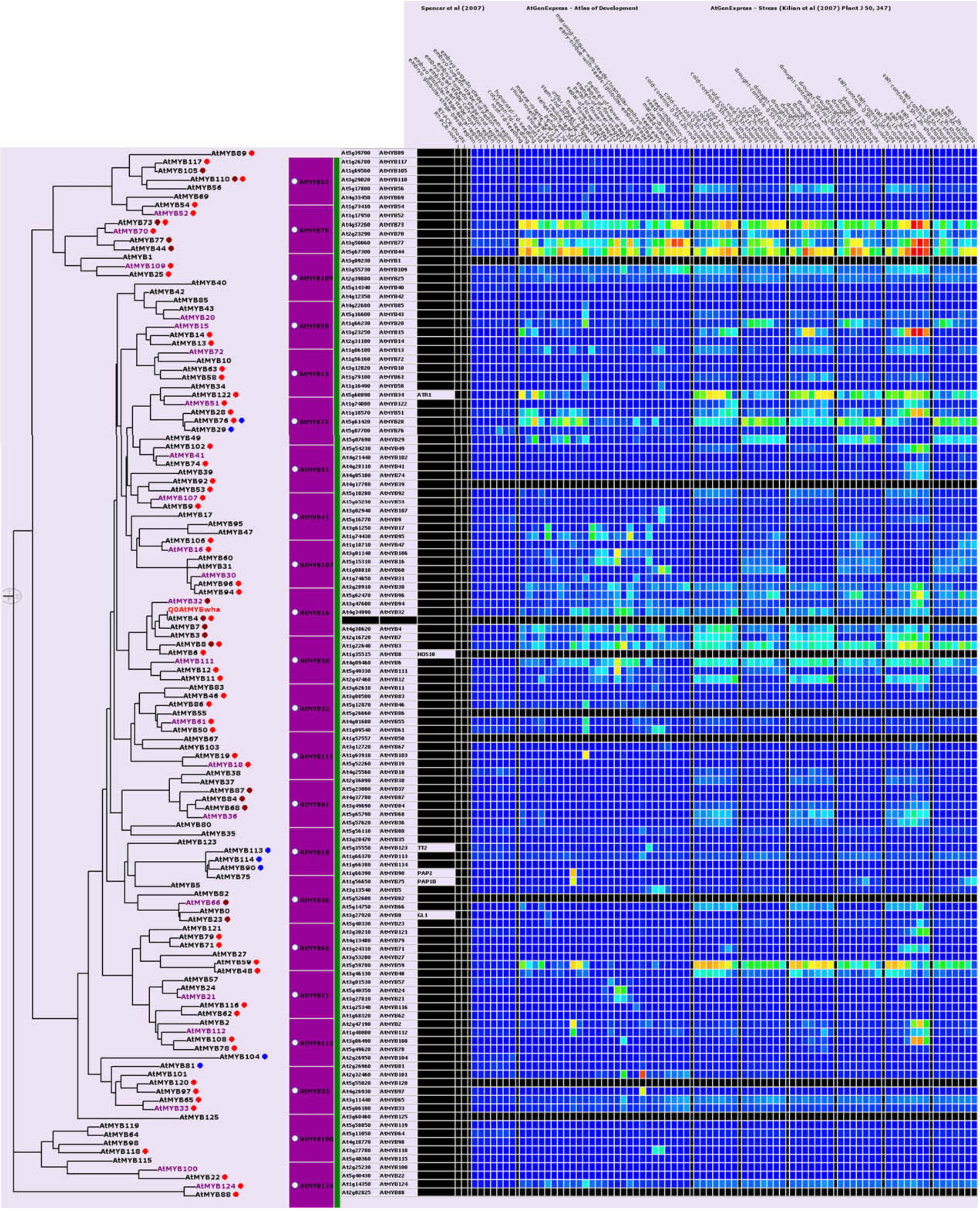

Fig. 2. A screen view of the main phylogenetic tree and expression display showing the Arabidopsis R2R3 MYB family with a single query sequence added (name in red). Genes residing on duplicated regions of the genome are indicated: red or dark red for recent or old duplication events, respectively. Genes present along the genome in tandem are indicated by blue dots. Subgroups can be selected in a web form via the purple boxes to obtain a tree for each subgroup (for example, see Fig. 3b). Expression values are displayed as a tabulated heat map (blue – low expression, through green then yellow to orange – higher expression). These values are the normalised raw signal intensity values so that the full dynamic range of expression strength can be visualised for genes within and between species. Performing a 'mouseover' on the table cells shows the numerical expression value for each data point. Gaps ('black holes') in the expression data appear for several genes. They occur when the sequence and expression data sets contain conflicting locus identifiers for the same gene or when expression data are unavailable.

not shown). This lack of definition in the tree is probably due to erosion of the phylogenetic 'signal' in the DNA binding domain during evolution coupled with the relatively small number of amino acid columns available for the analysis.

4. The gene expression display

Alongside the returning tree containing the query sequences is a table showing the pattern of gene expression across a range of tissues and environmental conditions for genes present in the existing phylogenetic data set. The numerical expression values are displayed as a heat map enabling expression patterns for closely related genes to be visualised side by side in a phylogenetic setting (Fig. 2). Expression profiles for the TFs were extracted from the expression datasets (described below) and the IT3F.pl program was extended to include them alongside the tree in the web display.

Arabidopsis Affymetrix microarray experiments were taken from three experimental series: the AtGenExpress global stress expression data set (Kilian et al., 2007), the AtGenExpress 'Expression Atlas of Arabidopsis Development' data set (Schmid et al., 2005), and an embryo development data set (Spencer et al., 2007). The AtGenExpress experiment contains 47 individual experiments comprising 1387 individual arrays. To display these data effectively in a limited amount of space on the web page, a subset of the data was examined: cold, drought and salt stress data sets, 27 tissue types from the Expression Atlas and seven embryo data sets. The raw Affymetrix .cel files were normalised using RMAExpress (http://rmaexpress.bmbolstad.com/). Per gene, an average value is presented for each redundant set of samples; in the stress experiments an average value is presented to represent all control samples except the 0 h control sample which is displayed separately.

Affymetrix experiments for rice that were equivalent to those available for Arabidopsis were analysed to demonstrate that the expression patterns of closely related genes from different species could also be compared side by side with their phylogenetic profiles. There are only a few Affymetrix experiments available for rice to date, but recently Jain et al. (2007) performed a survey of F-box proteins in rice which included Affymetrix data sets. Of the tissues they investigated, eight of these were considered to be comparable to the samples analysed for Arabidopsis. Three experiments investigating the response of 7 d old seedlings to cold, drought or salt stress after 3 h were also considered to be comparable to the 3 h Arabidopsis whole shoot or root stress samples. The raw Affymetrix .cel files were normalised as for the Arabidopsis data.

From the data sets presented in the web page, one might expect to be able to visualise the processes of neo- and subfunctionalisation in which duplicated genes diverge with respect to sequence and expression following a genome duplication event (reviewed by Briggs et al. (2006)). For the R2R3MYB family, it appears that many recently duplicated genes do have significantly different expression patterns across the range of tissues presented. Arabidopsis and rice expression patterns for closely related genes, perhaps orthologous genes, can be extremely different. This observation may mean that apparently similar dicot and monocot tissues, for example young or mature leaves, cannot be assumed to be physiologically equivalent. An alternative explanation is that mRNA expression patterns for these genes is not a good indicator of function, having migrated away from those patterns present in the common ancestor of Arabidopsis and rice without affecting the biochemical roles that the genes perform in the plant. The prediction might be that comparing more closely related species, i.e., those within the dicot or monocot lineages – for example, Arabidopsis and a legume species – would reveal more similar expression patterns; Frickey et al. (2008) have reported that their AffyTrees tool finds similar expression patterns for closely related Medicago and Arabidopsis genes. Nevertheless, conservation in gene expression patterns can still be observed between Arabidopsis and rice genes and between genes that are known to interact with each other in the transcriptional initiation complex. For example, similar expression profiles are apparent for two clearly orthologous sequences, AtMYB5 and Os01g50110 in developing seed tissue; co-expression of AtMYB5 and AtbHLH042 (TT8), two genes known to interact (Zimmermann et al., 2004), can also be seen in the corresponding trees; stress-induced gene expression of AtMYB15 is mirrored in two closely related rice genes, Os04g43680 (OsMYB4) and Os02g41510. (Refer to the R2R3MYB and bHLH data sets containing Arabidopsis and rice sequences.) The identification of TF genes whose expression patterns have been conserved over large evolutionary distances may be the more significant attribute of the IT3F website, since expression patterns are likely to be most closely related to function for such genes.

5. Creating a subgroup tree

An alignment of protein sequences belonging to a TF family subgroup can reveal further zones of sequence similarity in addition to the DNA binding domain. Re-estimation of the phylogenetic tree based on an extended alignment to include these additional zones, should render a more fine-grained tree topology enabling potential orthologous genes or functional clusters to be pinpointed with greater accuracy.

A second Perl-CGI program was written to select a subgroup of sequences specified by the user from the whole TF family tree and then perform phylogenetic analysis on this subset of sequences. The subgroup boundary is established using pre-defined subgroup identifiers which typify each group. These identifiers are shown inside purple boxes (the Inner Tree Form) on the web page and are also highlighted in purple in the tree (Fig. 2). Subgroups can be identified easily from the whole TF family tree or are known from existing studies (for example, for the R2R3 MYB family, refer to the 25 subfamilies described by Stracke et al. (2001)). When an identifier is selected from the Inner Tree Form on the web page, the Perl-CGI program analyses the contents of a Newick tree file between the parentheses (green) and continues outwards from the selected subgroup identifier, in this example AtMYB16, until a bracket pair (red) contains another subgroup identifier (for example AtMYB51) (Fig. 3a). The genes contained within the previous bracket pair then represent the subgroup and the corresponding names are used to make the inner tree. The second subgroup identifier found, in this case AtMYB51, is used as an outgroup to stabilize the base of the tree.

Programs to align the subgroup sequences (CLUSTALW) and perform the phylogeny (PROTDIST, NEIGHBOR, and DRAWGRAM) are then called by the Perl-CGI program and the resulting neighbour-joining tree is returned to the webpage (Fig. 3b).

Below the tree in the web page is an alignment that shows how sequences in the subgroup are related across their full length. Conserved columns of the alignment are indicated using a colour scheme that picks out shared amino acid motifs existing outside the DNA binding domain as well as other more extensive zones of sequence similarity such as protein domains (Fig. 3c). These highlighted columns were all present in a common ancestor, so each one contains homologous amino acids. All these columns are used together to derive the subgroup phylogenetic tree. The number of evolutionarily constrained columns within the DNA binding domain of closely related proteins is often extremely high. Therefore, the additional columns in bold, blue and red that lie outside the DNA binding domain are likely to be important for providing an improved estimate of the tree.

6. Future developments

The subgroup tool is still under development. The tree that is displayed currently is a neighbor-joining tree but it would be interesting to test more sophisticated but time-consuming phylogenetic methods such as maximum likelihood, at least for small subgroups of proteins (about 10 members). This should be possible as computer speeds increase, together with the adoption of maximum

a

(((AtMYB34 , (AtMYB122 , **AtMYB51**)) , ((AtMYB28 , AtMYB76) , AtMYB29)) , (^{4}(AmML1 , AmMYBML3) , (^{3}(AmMIXTA , (GhMYB25 , ((Os03g33660 , AmML2) , ((Q2LeTC1896 , LjMYB8) , (Q1SdMYBML2 , PhMYB1))))) , (^{2}(Os02g36890 , (TaMYB38 , Os04g38740)) , (^{1}AtMYB106 , **AtMYB16**^{1})^{2})^{3})^{4})) , ((AtMYB95_PA,AtMYB47) , (Os03g26130 , ((Os09g24800 , (Os07g43580 , Os08g33940)) , (LjMYB48 , ((AtMYB31 , (((Os11g35390 , AtMYB60) , (Os12g03150 , Os11g03440)) , LjMYB51)) , ((**AtMYB30** , AmMYB306) , (AtMYB96 , AtMYB94)))))))

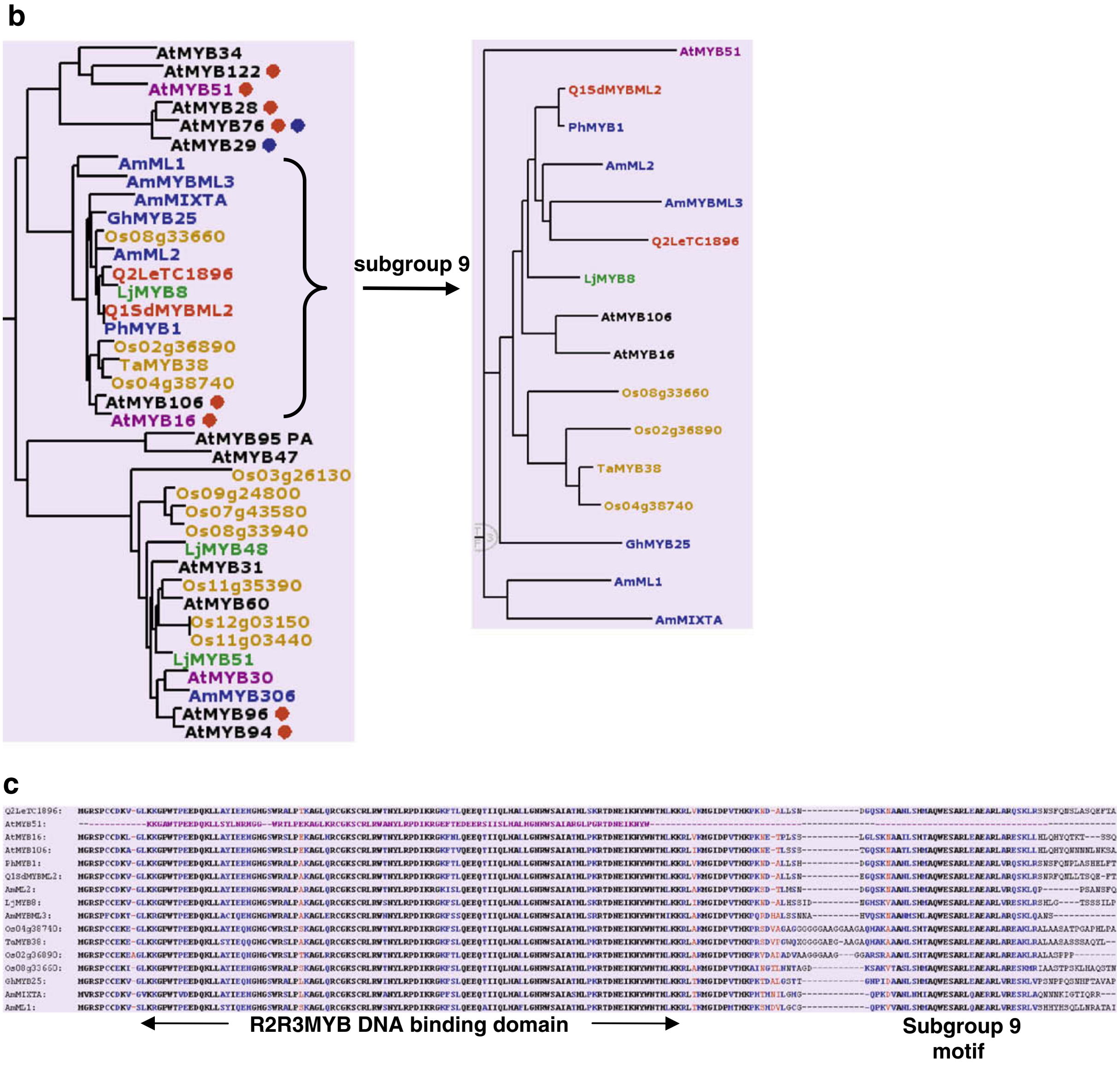

Fig. 3. (a) An inner tree traversal to identify members of a subgroup in a protein family tree, in this example, subgroup 9 of the R2R3MYB family. The tree shown is in Newick tree format (without branch lengths), annotated with the steps of the traversal described in the main text. Pairs of green parentheses (numbered) contain subgroup 9 members and a pair of red parentheses contains neighbouring subgroup members, including a second subgroup identifier (the outgroup). Subgroup identifiers are drawn in purple. (b) A screen view showing three major clades from the full R2R3MYB tree (left) and the equivalent tree after the sequence data for subgroup 9 have been processed by the second Perl-CGI program (right). Arabidopsis, rice and legume genes are drawn in black, yellow and green, respectively. Query sequences are drawn in red. (c) The N-terminal region of the full alignment for subgroup 9 proteins which has highlighted a conserved motif that is common to all members of this subgroup known to date. The columns highlighted in bold, blue and red contain 100%, greater than 50% and less than 50% amino acid identity, respectively. The outgroup sequence is drawn in purple.

likelihood methods for parallelisation (for example, fastDNAml (Olsen et al., 1994), see http://rac.uits.iu.edu/hpc/fastDNAml/index.shtml). However, due to the relatively small amount of data available, i.e., the small number of amino acid residues available for TF phylogeny, maximum likelihood methods may not prove to be significantly more accurate than a neighbor-joining tree. Tests are underway to compare the two strategies. While this comparison is ongoing, it should be possible to modify the tool to perform bootstrap analysis for the current neighbor-joining trees and to display the results on the tree.

Addition of further TF family data sets will be important to make the website relevant to a wider audience. Updating existing data sets will also be important but will require the development of automated tools. In particular, the locus IDs unique to each gene need to be updated because they are essential for retrieving information from each data resource. At the moment, not all locus IDs are identical in each data set (sequence, gene expression and the evidence for recent gene duplication events – see Fig. 2).

The aim of this work in the future will be to focus on TF subgroups that regulate secondary metabolism in plants. They are likely to be from the MYB, AP2, WRKY and bHLH families. Categorising functionally similar regulators into discrete phylogenetic subgroups and being able to observe other experimental data, should assist with understanding their associations with corresponding downstream target gene(s) encoding the enzyme(s) of specific secondary metabolic pathways for any plant species. One way that associations between regulator and target genes can be inferred is through the identification of co-regulated genes (for example, from Arabidopsis gene expression studies (Manfield et al., 2006)). The methods developed in this work could be adapted for the analysis of the target genes which are common in secondary metabolism (for example, the cytochrome P450, glucosyl transferase and BAHD acyl transferase families). We propose that it would then be possible for users to interrogate phylogenetic trees for both regulator and target genes and for the web page to display the evidence for any associations between them in the area of the subgroups of interest.

Acknowledgements

P.C.B. gratefully acknowledges financial support from the John Innes Centre for a distance-learning M.Sc. course in Bioinformatics (University of Manchester) with supervision of the dissertation from J.D. and from Dr. David Robertson (University of Manchester).

References

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

Blanc, G., Hokamp, K., Wolfe, K.H., 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genome Research 13, 137–144.

Briggs, G.C., Osmont, K.S., Shindo, C., Sibout, R., Hardtke, C.S., 2006. Unequal genetic redundancies in Arabidopsis—a neglected phenomenon? Trends in Plant Science 11, 492–498.

Conte, M.G., Gaillard, S., Lanau, N., Rouard, M., Perin, C., 2008. GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Research 36, D991–D998.

Davuluri, R.V., Sun, H., Palaniswamy, S.K., Matthews, N., Molina, C., Kurtz, M., Grotewold, E., 2003. AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis *cis*-regulatory elements and transcription factors. BMC Bioinformatics 4, 25.

Eddy, S.R., 1998. Profile hidden Markov models. Bioinformatics (Oxford, England) 14, 755–763.

Felsenstein, J., 2004. PHYLIP (Phylogeny Inference Package) Version 3.6 Distributed by the Author. Department of Genome Sciences, University of Washington, Seattle.

Fredslund, J., 2008. DATFAP: a database of primers and homology alignments for transcription factors from 13 plant species. BMC Genomics 9, 140.

Frickey, T., Benedito, V.A., Udvardi, M., Weiller, G., 2008. AffyTrees: facilitating comparative analysis of Affymetrix plant microarray chips. Plant Physiology 146, 377–386.

Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W., Zheng, W., Gu, X., Wei, L., Luo, J., 2006. DRTF: a database of rice transcription factors. Bioinformatics (Oxford, England) 22, 1286–1287.

Grotewold, E., Drummond, B.J., Bowen, B., Peterson, T., 1994. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. Cell 76, 543–553.

Guo, A., He, K., Liu, D., Bai, S., Gu, X., Wei, L., Luo, J., 2005. DATF: a database of Arabidopsis transcription factors. Bioinformatics (Oxford, England) 21, 2568–2569.

Guo, A.Y., Chen, X., Gao, G., Zhang, H., Zhu, Q.H., Liu, X.C., Zhong, Y.F., Gu, X., He, K., Luo, J., 2008. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Research 36, D966–D969.

Heim, M.A., Jakoby, M., Werber, M., Martin, C., Weisshaar, B., Bailey, P.C., 2003. The basic helix-loop-helix transcription factor family in plants: a genome-wide study of protein structure and functional diversity. Molecular Biology and Evolution 20, 735–747.

Iida, K., Seki, M., Sakurai, T., Satou, M., Akiyama, K., Toyoda, T., Konagaya, A., Shinozaki, K., 2005. RARTF: database and tools for complete sets of Arabidopsis transcription factors. DNA Research 12, 247–256.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyere, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pe, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quetier, F., Wincker, P., 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449, 463–467.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A.K., Khurana, J.P., 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiology 143, 1467–1483.

Jin, H., Cominelli, E., Bailey, P., Parr, A., Mehrtens, F., Jones, J., Tonelli, C., Weisshaar, B., Martin, C., 2000. Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in Arabidopsis. The EMBO Journal 19, 6150–6161.

Jones, D.T., Taylor, W.R., Thornton, J.M., 1992. The rapid generation of mutation data matrices from protein sequences. Computer Applications in the Biosciences 8, 275–282.

Kilian, J., Whitehead, D., Horak, J., Wanke, D., Weinl, S., Batistic, O., D'Angelo, C., Bornberg-Bauer, E., Kudla, J., 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 Journal 50, 347–363.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Molecular Biology and Evolution 11, 459–468.

Lee, M.M., Schiefelbein, J., 2001. Developmentally distinct MYB genes encode functionally equivalent proteins in Arabidopsis. Development (Cambridge, England) 128, 1539–1546.

Li, X., Duan, X., Jiang, H., Sun, Y., Tang, Y., Yuan, Z., Guo, J., Liang, W., Chen, L., Yin, J., Ma, H., Wang, J., Zhang, D., 2006. Genome-wide analysis of basic/helix-loop-helix transcription factor family in rice and Arabidopsis. Plant Physiology 141, 1167–1184.

Liu, L., White, M.J., MacRae, T.H., 1999. Transcription factors and their genes in higher plants functional domains, evolution and regulation. European Journal of Biochemistry/FEBS 262, 247–257.

Luo, J., Nishiyama, Y., Fuell, C., Taguchi, G., Elliott, K., Hill, L., Tanaka, Y., Kitayama, M., Yamazaki, M., Bailey, P., Parr, A., Michael, A.J., Saito, K., Martin, C., 2007. Convergent evolution in the BAHD family of acyl transferases: identification and characterization of anthocyanin acyl transferases from *Arabidopsis thaliana*. Plant Journal 50, 678–695.

Manfield, I.W., Jen, C.H., Pinney, J.W., Michalopoulos, I., Bradford, J.R., Gilmartin, P.M., Westhead, D.R., 2006. Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. Nucleic Acids Research 34, W504–W509.

Martin, C., Paz-Ares, J., 1997. MYB transcription factors in plants. Trends Genetics 13, 67–73.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R.E., Michael, T.P., Wall, K., Rice, D.W., Albert, H., Wang, M.L., Zhu, Y.J., Schatz, M., Nagarajan, N., Acob, R.A., Guan, P., Blas, A., Wai, C.M., Ackerman, C.M., Ren, Y., Liu, C., Wang, J., Wang, J., Na, J.K., Shakirov, E.V., Haas, B., Thimmapuram, J., Nelson, D., Wang, X., Bowers, J.E., Gschwend, A.R., Delcher, A.L., Singh, R., Suzuki, J.Y., Tripathi, S., Neupane, K., Wei, H., Irikura, B., Paidi, M., Jiang, N., Zhang, W., Presting, G., Windsor, A., Navajas-Perez, R., Torres, M.J., Feltus, F.A., Porter, B., Li, Y., Burroughs, A.M., Luo, M.C., Liu, L., Christopher, D.A., Mount, S.M., Moore, P.H., Sugimura, T., Jiang, J., Schuler, M.A., Friedman, V., Mitchell-Olds, T., Shippen, D.E., dePamphilis, C.W., Palmer, J.D., Freeling, M., Paterson, A.H., Gonsalves, D., Wang, L., Alam, M., 2008. The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature 452, 991–996.

Nakano, T., Suzuki, K., Fujimura, T., Shinshi, H., 2006. Genome-wide analysis of the ERF gene family in Arabidopsis and rice. Plant Physiology 140, 411–432.

Niggeweg, R., Michael, A.J., Martin, C., 2004. Engineering plants with increased levels of the antioxidant chlorogenic acid. Nature Biotechnology 22, 746–754.

Olsen, G.J., Matsuda, H., Hagstrom, R., Overbeek, R., 1994. FastDNAmL: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Computer Applications in the Biosciences 10, 41–48.

Remm, M., Storm, C.E., Sonnhammer, E.L., 2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. Journal of Molecular Biology 314, 1041–1052.

Rensing, S.A., Lang, D., Zimmer, A.D., Terry, A., Salamov, A., Shapiro, H., Nishiyama, T., Perroud, P.F., Lindquist, E.A., Kamisugi, Y., Tanahashi, T., Sakakibara, K., Fujita, T., Oishi, K., Shin, I.T., Kuroki, Y., Toyoda, A., Suzuki, Y., Hashimoto, S., Yamaguchi, K., Sugano, S., Kohara, Y., Fujiyama, A., Anterola, A., Aoki, S., Ashton, N., Barbazuk, W.B., Barker, E., Bennetzen, J.L., Blankenship, R., Cho, S.H., Dutcher, S.K., Estelle, M., Fawcett, J.A., Gundlach, H., Hanada, K., Heyl, A., Hicks, K.A., Hughes, J., Lohr, M., Mayer, K., Melkozernov, A., Murata, T., Nelson, D.R., Pils, B., Prigge, M., Reiss, B., Renner, T., Rombauts, S., Rushton, P.J., Sanderfoot, A., Schween, G., Shiu, S.H., Stueber, K., Theodoulou, F.L., Tu, H., Van de Peer, Y., Verrier, P.J., Waters, E., Wood, A., Yang, L., Cove, D., Cuming, A.C., Hasebe, M., Lucas, S., Mishler, B.D., Reski, R., Grigoriev, I.V., Quatrano, R.S., Boore, J.L., 2008. The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. Science (New York, NY) 319, 64–69.

Riano-Pachon, D.M., Ruzicic, S., Dreyer, I., Mueller-Roeber, B., 2007. PlnTFDB: an integrative plant transcription factor database. BMC Bioinformatics 8, 42.

Richardt, S., Lang, D., Reski, R., Frank, W., Rensing, S.A., 2007. PlanTAPDB, a phylogeny-based resource of plant transcription-associated proteins. Plant Physiology 143, 1452–1466.

Riechmann, J.L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., Adam, L., Pineda, O., Ratcliffe, O.J., Samaha, R.R., Creelman, R., Pilgrim, M., Broun, P., Zhang, J.Z., Ghandehari, D., Sherman, B.K., Yu, G., 2000. Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. Science (New York, NY) 290, 2105–2110.

Riechmann, J.L., Meyerowitz, E.M., 1998. The AP2/EREBP family of plant transcription factors. Biological Chemistry 379, 633–646.

Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Molecular Biology and Evolution 4, 406–425.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Scholkopf, B., Weigel, D., Lohmann, J.U., 2005. A gene expression map of *Arabidopsis thaliana* development. Nature Genetics 37, 501–506.

Shiu, S.H., Shih, M.C., Li, W.H., 2005. Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiology 139, 18–26.

Sieweke, M.H., Graf, T., 1998. A transcription factor party during blood cell differentiation. Current Opinion in Genetics & Development 8, 545–551.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Palmer, J.D., Wing, R.A., dePamphilis, C.W., Ma, H., Carlson, J.E., Altman, N., Kim, S., Wall, P.K., Zuccolo, A., Soltis, P.S., 2008. The Amborella genome: an evolutionary reference for plant biology. Genome Biology 9, 402.

Spencer, M.W., Casson, S.A., Lindsey, K., 2007. Transcriptional profiling of the Arabidopsis embryo. Plant Physiology 143, 924–940.

Stracke, R., Werber, M., Weisshaar, B., 2001. The R2R3-MYB gene family in *Arabidopsis thaliana*. Current Opinion in Plant Biology 4, 447–456.

Tamagnone, L., Merida, A., Parr, A., Mackay, S., Culianez-Macia, F.A., Roberts, K., Martin, C., 1998. The AmMYB308 and AmMYB330 transcription factors from antirrhinum regulate phenylpropanoid and lignin biosynthesis in transgenic tobacco. Plant Cell 10, 135–154.

Tuskan, G.A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., Rokhsar, D., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science (New York, NY) 313, 1596–1604.

Yanhui, C., Xiaoyuan, Y., Kun, H., Meihua, L., Jigang, L., Zhaofeng, G., Zhiqiang, L., Yunfei, Z., Xiaoxiao, W., Xiaoming, Q., Yunping, S., Li, Z., Xiaohui, D., Jingchu, L., Xing-Wang, D., Zhangliang, C., Hongya, G., Li-Jia, Q., 2006. The MYB transcription factor superfamily of Arabidopsis: expression analysis and phylogenetic comparison with the rice MYB family. Plant Molecular Biology 60, 107–124.

Zimmermann, I.M., Heim, M.A., Weisshaar, B., Uhrig, J.F., 2004. Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. Plant Journal 40, 22–34.

Zmasek, C.M., Eddy, S.R., 2002. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC Bioinformatics 3, 14.