J Mol Model (2004) 10:130–138
DOI 10.1007/s00894-003-0176-3

ORIGINAL PAPER

Maria Svensson · Dan Lundh · Mikael Ejdebäck · Abul Mandal

Functional prediction of a T-DNA tagged gene of *Arabidopsis thaliana* by *in silico* analysis

Received: 19 June 2003 / Accepted: 1 December 2003 / Published online: 7 February 2004

Abstract We have employed a gene-knockout approach using T-DNA tagging and in vivo gene fusion in *Arabidopsis thaliana* for identification and isolation of specific plant genes. Screening of about 3,000 T-DNA tagged lines resulted in identification of a mutant line (no. 197) exhibiting a significant delay in flowering. From this line a 600-bp plant DNA fragment downstream of the left T-DNA junction was cloned by inverse PCR. BLAST searching in the *A. thaliana* genomic database indicated a putative gene, *frf* (*f*lowering *r*egulating *f*actor), with unknown function downstream of the T-DNA insert. Bioinformatic tools were used to predict possible protein structure and function. The protein structure predicted by fold recognition indicates that *frf* is a transcriptional regulator, a ligand-binding receptor responsive to steroids and hormones. Analyzing the predicted results and the phenotype of the T-DNA tagged plant we hypothesized that FRF might be involved in hormone response in *A. thaliana*. For verification of this hypothesis we exposed the plants of line no. 197 to gibberellic acid (GA_3), a potential growth regulator in higher plants. This treatment resulted in an earlier onset of flowering, almost similar to that in wild type control plants.

Keywords T-DNA tagging · Fold recognition · Gibberellic acid · Flowering time · *Arabidopsis thaliana*

M. Svensson · M. Ejdebäck · A. Mandal (✉)
Department of Natural Sciences,
University of Skövde,
P.O. Box 408, 541 28 Skövde, Sweden
e-mail: abul.mandal@inv.his.se
Tel.: +46 500 448 608
Fax: +46 500 448 699

D. Lundh
Department of Computer Sciences,
University of Skövde,
P.O. Box 408, 541 28 Skövde, Sweden

Introduction

During the past decade several molecular techniques have been employed for identification and cloning of specific plant genes encoding useful but genetically often not so well defined characteristics. Differential screening of cDNA libraries has been used successfully to isolate inducible plant genes, e.g., genes induced by low temperature, [1] and positional or map-based cloning seems to provide a promising long-term approach for isolation of useful plant genes. [2] Identification of a particular class of plant genes that might be involved in regulating growth and development of plants by the map-based cloning approach requires both a recognizable phenotype and considerable effort. Differential screening is also limited to cloning of genes that are expressed under specific conditions, e.g., in response to low temperature, [1] drought [3] or pathogen infection [4].

A viable alternative for predicting the function of specific plant genes is provided by T-DNA insertion mutagenesis. The use of T-DNA of *Agrobacterium tumefaciens* as an insertion element provides an efficient way of generating both insertion mutants with recognizable phenotypes and reporter gene fusions to plant promoters. [5, 6] In addition, the completion of sequencing of the *Arabidopsis thaliana* genome [7] has opened up new possibilities for identification and cloning of specific classes of plant genes, although the functional annotation of the genes is not yet complete. One approach for predicting the function of the genes of unknown function could be based on determination of the protein's three-dimensional structure.

The bioinformatic tools available today have difficulty predicting the three-dimensional structure directly from the protein's amino acid sequence. With the fold recognition method, a protein fold is predicted by matching a new sequence to an already known fold. [8, 9] This method is therefore limited, as it can only recognize experimentally determined folds but not novel ones. However, the estimated probability of finding a novel gene product that has a genuinely new structure is less

then 10%, of which only 3–4% reveal new folds. [10] Thus, the method of fold recognition is considered to be a powerful tool in obtaining structural information about new genes [11].

Fold recognition and comparative modeling methods for gaining clues of proteins encoded by the genome have been applied successfully to various genomes. This approach involves assembling software that consists of modules for fold assignment, template selection, target–template alignment, model generation and model evaluation [12]. This has been tested widely in different organisms such as *Escherichia coli*, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Mycoplasma genitalium* and *Methanococcus janaschi*. [13]. Fold recognition methods have also been tested on specific proteins for obtaining clues about the protein functionality [14, 15].

In this paper we describe an alternative approach that combines bioinformatics and molecular biology for estimation of the function of one specific gene in the model plant *A. thaliana*. A genomic DNA sequence was cloned from a T-DNA tagged mutant of *A. thaliana* by inverse PCR and used as a template for identification of the corresponding sequence in the *A. thaliana* genome. A candidate gene of unknown function was identified and characterized *in silico* for prediction of protein structure and estimation of gene function.

Materials and methods

Gene tagging

Arabidopsis thaliana (ecotype C24) was used as a source for tagging of structural genes. Production of transgenic plants, by using *Agrobacterium tumefaciens* T-DNA mediated gene transfer, selection and tissue culture procedures were as described previously by Mandal et al. [16]. The vector pMHA2 used for gene tagging was a promoter-probing vector based on *uidA* (*gus*, β-glucuronidase) as a reporter gene placed adjacent to the right end of the T-DNA [17]. The vector also contains a *pnos-nptII* plant-selectable marker gene (kanamycin resistance, Km^R) located at the left end of the T-DNA.

Growth of transgenic plants

All investigations were performed with T_2 progeny of transgenic lines grown on MS medium [18] supplemented with 20 g/l sucrose and 50 mg/l kanamycin sulfate. Growth-chamber conditions were maintained at 22 °C, 70% relative humidity and a 16-h day. Wild-type plants (C24) were treated similarly but without kanamycin selection.

Analysis of transgenic plants

Identification of mutant phenotype

Four-week-old axenically grown plants of wild-type, vector-transformed control plants and transgenic line no. 197 were transferred to soil and kept at room temperature (22–24 °C) using a 16-h day. Flowering time was measured by counting the number of days from sowing until the first flower bud was visible.

Treatment with gibberellic acid

Two-week-old axenically grown seedlings of wild-type control plants and transgenic line no. 197 were transferred to soil pots and kept in growth chambers at 22 °C, 70% relative humidity and with a 16-h day. The plants were then sprayed once a week with 100 μM GA_3 (gibberellic acid, Sigma). Control plants were sprayed similarly with tap water. Flowering time was measured as described earlier.

Statistical analysis

Data obtained from measuring of the flowering time were analyzed statistically by two-sample *t*-test [19] assuming equal variance using MINITAB Statistical software, release 13.32 [20]. *T*-value (*T*), *P*-value (*P*) and degree of freedom (df) were used for explanation of the results.

Cloning of T-DNA flanking plant DNA

For cloning of the plant DNA sequences adjacent to the left end of the T-DNA, genomic DNA from line 197 was digested with SspI. The DNA was then self-ligated using T4 DNA ligase, 2 U for a 50 μl reaction. Inverse PCR was performed using PCR kit AmpliTaq Gold from Applied Biosystems. Upper LB Primer (5'-ATTTGTCGTTTTATCAAAATGTAC-3') and Lower LB Primer (5'-CATTCCCAGATACCCATTTCA-3') were used for cloning of the left T-DNA-plant DNA junction fragment. The IPCR reaction was carried out in a total volume of 50 μl containing 5 ng self-ligated plant DNA, 3 mM $MgCl_2$, 1X PCR buffer, 0.8 mM dNTP, 0.5 μM primers Upper LB and Lower LB and 1.25 U AmpliTaq Gold DNA polymerase. The PCR was performed with an initial denaturation at 92 °C for 15 min followed by 35 cycles of 94 °C for 20 s, 51 °C for 1 min and 72 °C for 2 min. The final step of elongation was maintained at 72 °C for 10 min. The IPCR fragment was cleaned with QIAquick PCR Purification kit (Qiagen) and cloned into a TOPO TA cloning vector for sequencing (Invitrogen). Synthetic oligonucleotides were bought from MedProbe.

Analysis of plant DNA

DNA sequencing was performed by using ABI PRISM BigDye Terminator Cycle Sequencing Ready Reaction Kit and the DNA sequencer ABI PRISM 310 Genetic Analyzer from Applied Biosystems. For cycle sequencing, 400 ng plasmid DNA and 3.2 pmol primer were added. Two primers T3 (5'-ATTAACCC-TCACTAAAGGGA-3') and T7 (5'-AATACGACTCACTATA-GGG-3') were used for sequencing the IPCR fragment cloned previously into a TOPO TA vector.

For identification of the location of the T-DNA insertion in the plant genome, the IPCR-cloned plant DNA sequence was searched against the *A. thaliana* GenBank [21] by using BLAST at NCBI.

Similarity search and homology determination

The predicted protein sequence of FRF obtained from GenBank was run against standard databases such as the Non Redundant Data Base (NRDB), and the Protein Data Bank (PDB) [22] to ensure that any functionally related sequence was collected. For this purpose BLAST [23] and PSI-BLAST [24] were used. The PSI-BLAST search was iterated until no new sequences above the PSI-BLAST threshold were indicated. In searching for neighbors of the putative protein FRF, four iterations were required. For identification of known motifs, the predicted protein sequence of FRF was run against secondary databases using Interpro [25] as a working tool. Based on the BLAST search an alignment was constructed and plotted using the default parameters of the Clustal W algorithm [26] and OMIGA 2.0 [27].

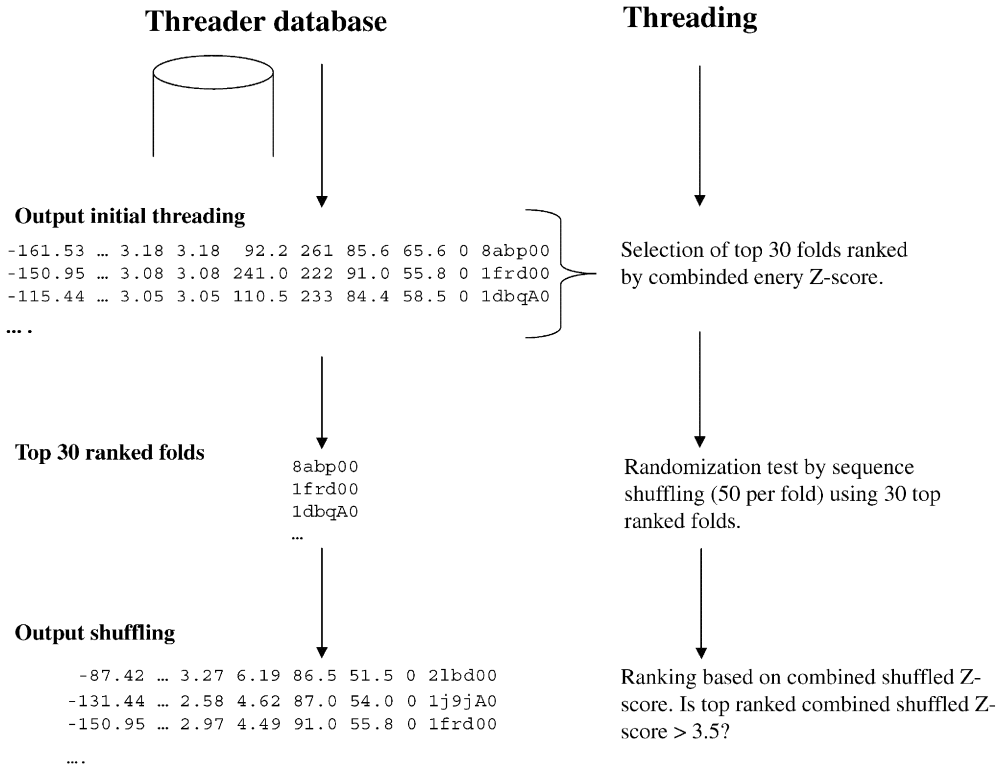

Fig. 1 Schematic presentation of threading. The threading procedures are performed as outlined in the THREADER manual

Fold recognition

For threading the predicted protein sequence FRF, we employed a fold recognition method and for identification of the possible template folds we used the software THREADER 3.3 [9, 28]. Threading was performed in a two-step procedure as described in Fig. 1. Step one was an initial threading for capturing interesting folds, i.e. running of the THREADER against all folds in the database (in total 5,257 folds and domains). From this initial threading, the top 30 folds were selected based on combined energy *Z*-score values (the higher the *Z*-score value the higher is the ranking). Classification and interpretation of the *Z*-score values according to the THREADER manual can be summarized as follows: very significant—possibility for a correct prediction is very high ($Z>4.0$); significant—good chance of being correct ($Z>3.5$); borderline significant—possibly correct ($2.7<Z<3.5$); poor score—could be right, but needs other conformation ($2.0<Z<2.7$); very poor—probably there are no suitable folds in the library ($Z<2.0$). The second step was the shuffling of the top 30 captured folds for determination of their significance. Each captured fold was shuffled 50 times for detection of the false positives and for obtaining a more accurate ranking. Ranking was made based on the combined shuffled *Z*-scores, i.e. the *Z*-score values obtained after shuffling. As indicated in the THREADER manual significant folds were determined by the combined shuffled *Z*-scores higher than 3.5. Furthermore, for verification of the data obtained by THREADER we employed an alternative threading approach by using 3D-PSSM [29].

Additional threadings were performed similarly by using both THREADER and 3D-PSSM for investigation of the hypnotized domains (QLQ and WRC) of FRF.

Structure prediction

The predicted protein sequence FRF was matched (superimposed) to the template fold identified by fold recognition. The alignment of structurally related regions obtained from THREADER and FRF sequences was used as an input to superimpose the sequence on the structure. To increase the quality of the alignment between the template and the FRF sequence, all long unrelated regions (regions longer than nine amino acids) were scanned separately in the PDB database. When no matches were identified, the sequences were then threaded against the THREADER database. A template match for the unrelated region was determined based on the combined energy *Z*-score value ($Z>3.0$). For prediction of the protein structure, the fold 2LBD exhibiting the highest level of significance after shuffling was used as a template for the main fold. For unrelated regions missing in this thread, the following regions and folds were used as templates: 1GNF for residues 175–184 and 1KQ1 for residues 217–239.

In order to output a structure similar to the native fold we used the homologous modeling tool MODELLER [30]. As the quality of the structures may vary depending on the fold and alignment, we ensured the highest possible probability for a correct structure by using a validation procedure. The output from MODELLER may result in a number of structures with slightly different conformations. The most suitable structure, i.e. closest to the native state, was identified based on manual inspection of the energy profile together with solvent accessibility. The structures with the lowest energy (stereochemical clashes and total energy) as well as unsuspicious structures (without knots) were chosen for further investigation, i.e. energy minimization by molecular dynamics. Further validation of the putative structure was done by using the software PROCHECK [31].

Results

Analysis of transgenic plants

During screening of about 3,000 T-DNA tagged *A. thaliana* individuals, we identified a line (no. 197) that exhibits delayed flowering in comparison with control plants (Fig. 2). Flowering time in line no. 197, wild-type and vector-transformed control plants was estimated

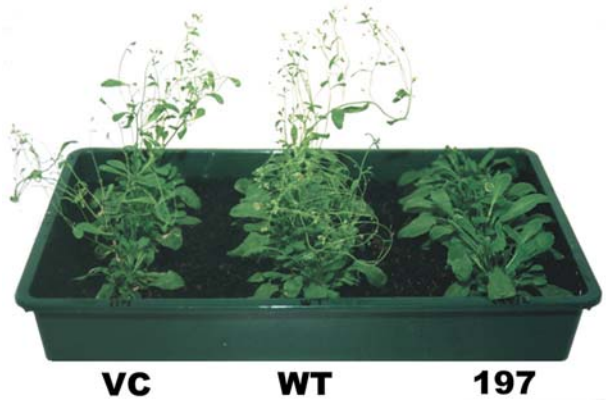

Fig. 2 Identification of a mutant phenotype in plants of line 197. Transgenic plants VC (vector-transformed control plants) and 197 were germinated axenically (22 °C, 70% relative humidity and a 16-h day) on MS-media supplemented with 50 mg/l kanamycin sulfate. Wild-type plants (WT) were treated similarly but without kanamycin selection and used as control plants. Four-week-old plants were transferred to soil and kept at room temperature (22–24 °C) using a 16-h day. Photograph taken after 8 weeks of growth

based on the number of days from sowing to appearance of the first flower bud. Plants of line no. 197 flowered on average 67 days after sowing, while the wild-type and vector-transformed control plants flowered in average after 44 and 47 days, respectively. This result suggests that the flowering time in plants of line 197 was delayed for about 20 days. In order to verify whether the observed difference in flowering time between the tagged line and the control plants was significant, we performed a two-sample *t*-test (data not shown). Results of this test indicated that the difference in flowering time between plants of line no. 197 and control plants was significant ($T_{\text{WT-197}}$=8.44, $P \ll 0.01$, df=165; $T_{\text{VC-197}}$=7.50, $P \ll 0.01$, df=173), whereas the difference between wild-type (WT) and vector-transformed (VC) control plants was not ($T_{\text{WT-VC}}$=1.89, P>0.01, df=108).

Cloning of T-DNA flanking plant DNA

Inverse PCR (IPCR) was employed to clone and sequence the plant DNA flanking the left end of the T-DNA. The sequencing results showed that a 635-bp plant DNA fragment adjacent to the T-DNA had been cloned. The fragment was then found to belong to *A. thaliana* chromosome II. In the *Arabidopsis* genomic database two putative genes of unknown function, At2g36410 and At2g36400, were identified in the vicinity of the cloned sequence. At2g36410 was found in the upstream, whereas At2g36400 was in the downstream region of this sequence.

Homology search

The putative gene At2g36400 (from now on called *frf, f*lowering *r*egulating *f*actor) identified downstream of the T-DNA insert was analyzed for prediction of protein sequence. The predicted protein sequence FRF, as obtained from the GenBank, was then used for similarity searching by using PSI-BLAST in the non-redundant database until no new sequences above the PSI-BLAST threshold were indicated. Hypothetical proteins of unknown function were omitted from further analysis. Of the remaining sequences, the highest scoring sequence (Score: 123; *E*-value: 3e−27) was found to be a growth-regulating factor, Os-GRF1 from *Oryza sativa*. [32] Os-GRF1 is encoded by a novel gibberellin (GA) induced gene and is characterized by the domains QLQ, WRC and TQL [32]. The predicted amino-acid sequence of FRF was aligned with that of Os-GRF1. A part of this alignment indicating essential regions of the protein is shown in Fig. 3. The alignment shows high sequence

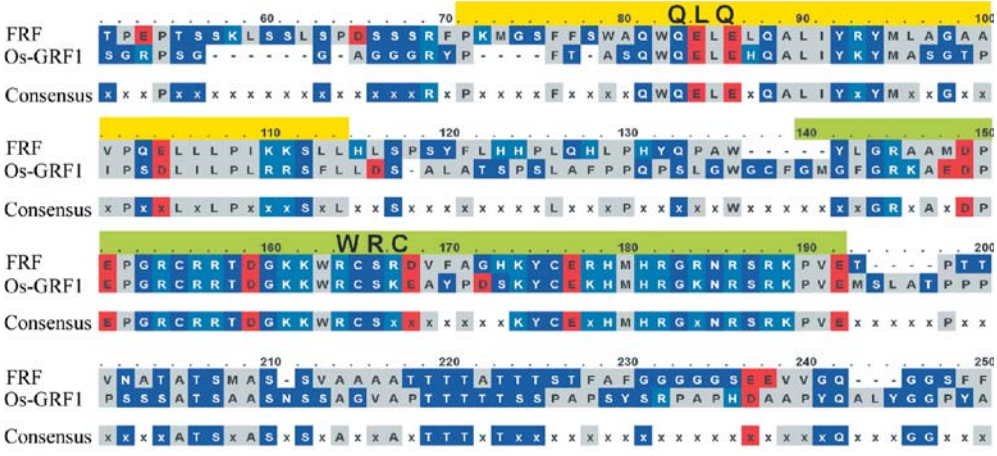

Fig. 3 Partial alignment of predicted protein sequence of FRF and Os-GRF1. The protein sequence alignment was constructed and plotted by using the Clustal W algorithm and the program OMIGA 2.0. The putative domains QLQ and WRC are marked in *yellow* and *green* colors, respectively. The residue color codes showing functionality are as follows: *blue* (basic), *red* (acidic), *gray* (hydrophobic) and *dark blue* (hydrophilic amino acid)

Table 1 Initial threading of FRF by THREADER. Ranking was determined based on the combined energy *Z*-score values

Ranking number	THREADER folds (PDB id.)	Combined energy *Z*-scores	Description and putative function of the identified folds
1	8ABP	3.18	Arabinose-binding protein mutant (binding protein)
2	1FDR	3.08	Flavodoxin reductase from *E. coli* (flavoprotein)
3	1DBQ	3.05	DNA-binding regulatory protein
4	1A80	2.68	Native 2,5-diketo-D-gluconic acid reductase (oxido-reductase)
5	2LBD	2.51	Ligand-binding domain of the human retinoic acid receptor 2 gamma bound to all-*trans* retinoic acids
6	1RYP	2.38	20S proteasome from yeast (multicatalytic proteinase)
7	1HYQ	2.37	Mind bacterial cell division regulator from *A. fulgidus*
8	1J9J	2.33	Sure protein from *T. maritima* (unknown function)

Table 2 Randomization test of the captured folds. The top 30 folds from the initial threading were shuffled by THREADER. Ranking was determined based on the combined shuffled *Z*-score values. Combined shuffled *Z*-scores were based on 50 shufflings for each fold

Ranking number	THREADER folds (PDB id.)	Combined shuffled Z-scores	Description and putative function of the identified folds
1	2LBD	6.19	Ligand-binding domain of the human retinoic acid receptor 2 gamma bound to all-*trans* retinoic acids
2	1J9J	4.62	Sure protein from *T. maritima* (unknown function)
3	1FDR	4.49	Flavodoxin reductase from *E. coli* (flavoprotein)
4	1A80	4.37	Native 2,5-diketo-D-gluconic acid reductase (oxido-reductase)
5	1DBQ	3.25	DNA-binding regulatory protein
6	8ABP	3.18	Arabinose-binding protein mutant (binding protein)
7	1CG2	3.05	Carboxypeptidase G2 (metallocarboxypeptidase)
8	1PEA	2.88	Amide receptor/negative regulator of the amidase operon of *Pseudomonas aeruginosa* (binding protein)

identities in the QLQ and WRC [32] domains resulting 50% and 72% similarity, respectively.

The Interpro [25] scan indicated an O-Glycosyl hydrolase (PRODOM PD203330, region 56–76) with a putative function as a growth-regulating factor (PRODOM PD025033, region 77–189). These results remained in complete agreement with those we obtained from BLAST searching. Matching of a growth-regulating factor was inherent from research on Os-GRF1. [32]

Fold recognition

As the BLAST search against the PDB structural database revealed no matches, fold recognition by THREADER was used to predict the protein structure. In the initial step of threading 30 folds were first captured based on combined energy *Z*-score values. These results are presented in Table 1. Table 1 shows that the eight top ranked folds had combined energy *Z*-scores either within a borderline significant fold (which is possibly correct) or within a poor score (which could be right, but needs other confirmation). To identify the false positive matches and to highlight the true positive ones, the top 30 captured folds were then shuffled by using THREADER. The output of this shuffling was ranked based on the folds' combined shuffled *Z*-scores as mentioned in the THREADER manual. These results are shown in Table 2. As indicated in Table 2, a ligand-binding domain of the human retinoic acid receptor gamma (PDB id: 2LBD) was ranked in the first place (combined shuffled *Z*-score=6.19, which is significant for just 50 shuffles). For this reason 2LBD was selected as the first candidate for a possible fold of the predicted FRF sequence.

For further analysis of the function of FRF, we threaded the suggested QLQ and WRC domains separately against THREADER and 3D-PSSM. For the QLQ domain it was not possible to deduce any function, no matches were found above the significance level. However, for the WRC domain some borderline significant matches were found within the DNA binding domains (transcription factors such as GATA-1; 1GNF, data not shown).

Structure prediction of FRF

Structural models of FRF matching the template 2LDB were built. Two large insertions were introduced in the 2LBD sequence for optimal alignment. One insertion was 10 amino acids long and corresponded to a region of FRF from amino acid 175 to 184. The second insertion was, however, 23 amino acids long and corresponded to a region of FRF from amino acid 217 to 239. The C-terminal region of FRF (117 amino acids) did not correspond to 2LBD and was therefore omitted from the modeling experiment. The best-derived model had an energy profile similar to that of the template and no high-energy regions could be detected. Verification of the best

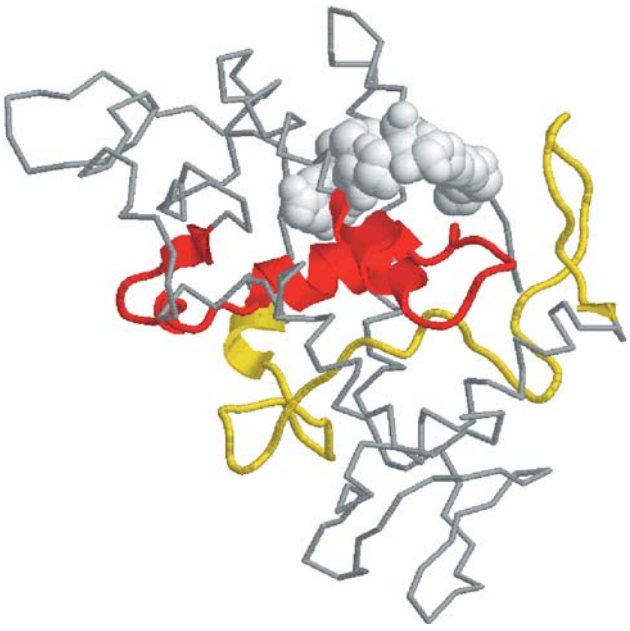

Fig. 4 Prediction of three-dimensional structure of FRF. The tools THREADER and MODELLER were used for predicting the three-dimensional structure of FRF; the C-terminal was not included in the model. The region marked in *red* illustrates the QLQ domain, *gray* denotes the substrate retinoic acid, and *yellow* illustrates the positioning of the WRC domain

Fig. 5 Treatment of plants with gibberellic acid (GA_3). Flowering time and stem elongation in plants of line 197 and wild-type (WT) *A. thaliana* plants with (+) or without (−) GA-treatment were compared. Transgenic plants of line 197 were germinated axenically (22 °C, 70% relative humidity and a 16-h day) on MS-media supplemented with 50 mg/l kanamycin sulfate. Wild-type plants (WT) were treated similarly but without kanamycin selection and used as control plants. Two-week-old plants were transferred to soil pots and kept in growth chambers at 22 °C with a 16-h day. Plants were sprayed once a week with 100-μM GA_3. Control plants were sprayed similarly with tap water

structure for FRF showed that in the Ramachandran plot [33] 91.9% of the residues were in the allowed areas, 4.1% in the generously allowed regions and 4.0% in the disallowed. For further validation of the putative structure the software PROCHECK [31] was used. This software was developed to assess how the normal or unusual geometry of the residues in a given protein structure was built as compared with the stereochemical parameters derived from well-refined and high-resolution structures. Deviations from the mean values were classified into quality classes of the standard PDB structure. The score of the quality classification for the best FRF structure resulted 1–1–3 whereas the score of the template 2LBD resulted 1–2–2 (the first digit refers to Phi-psi distribution, the second to Chi-1 standard deviation and the third to H-bond energy standard deviation; 1=one standard deviation below average, 2=average, 3=one standard deviation above average and 4=more than one standard deviations above average). In the Ramachandran plot the template structure 2LBD had 99.5% of the residues in the allowed areas while the remaining 0.5% in the generously allowed regions. Based on these data the structure of FRF was predicted (Fig. 4).

Treatment with gibberellic acid

Considering the mutant phenotype of plants of line no. 197 as well as the fact that gibberellin is a growth regulator and is involved in many developmental processes in plants such as stem elongation and flowering time, [34] we hypothesized that GA synthesis in this line could be incomplete. To verify this hypothesis, plants of line 197 were treated with gibberellic acid, GA_3. These results are shown in Fig. 5. When exposed to GA, plants of line no. 197 flowered earlier (41 days after sowing) than the untreated control plants (52 days after sowing). A two-sample *t*-test (data not shown) indicated that this difference in the flowering time between treated and untreated plants was significant (T=3.83, $P \ll 0.001$, df=18). In these experiments wild-type (WT) control plants flowered only 31 days after sowing. Thus, there was still a significant difference (T=9.31, $P \ll 0.01$, df=18) in the flowering time between the untreated WT control plants and the GA-treated plants of line no. 197. We did not observe any significant difference in flowering time between the treated and untreated WT control plants (T=1.90, P>0.01, df=18).

Discussion

During screening of the T-DNA tagged lines of *A. thaliana* one line, no. 197, attracted our attention as this line showed a significant delay in flowering. The delayed flowering time in this line was hypothesized to be due to a

gene mutation caused by T-DNA insertion. In order to complement the mutant phenotype we backcrossed the plants of line 197 with wild-type *A. thaliana*. Segregation analysis of the F_2 hybrid offspring based on the activity of the promoterless *gus* reporter gene (data not shown) indicates that plants of line no. 197 harbor more than one copy of the integrated T-DNA. Within the F_2 offspring, kanamycin-resistance plants exhibit either GUS-positive or GUS-negative phenotype. These results are preliminary and at present we are verifying these by Southern blot hybridization using T-DNA sequences as hybridization probes. For further verification of the contribution of this gene to the observed phenotype, we are now analyzing several lines of SIGnAL (Salk Institute Genomic Analysis Laboratory) mutants of *A. thaliana* harboring a mutation in *frf* or in its vicinity. As indicated in the results, by employing inverse PCR (IPCR) we could successfully amplify only a 635-bp fragment of plant DNA flanking the left end of the T-DNA. For IPCR amplification, the template DNA was isolated from the plants of line 197 exhibiting both GUS activity and delayed flowering. This IPCR-cloned genomic sequence was then used in BLAST searching for identification of the hypothetical target gene. Two putative genes were subsequently identified in the genomic database upstream (At2g36410) and downstream (At2g36400) of the T-DNA insert. In the initial stage of our investigations both genes were characterized by *in silico* analysis. However, later we continued with only one gene At2g36400 (*frf*) found downstream of the T-DNA insert. There were mainly three reasons behind this decision. The first one was that the plants of line 197 show *in vivo* gene fusion with the promoterless *gus* reporter gene placed adjacent to the right end of the T-DNA. The activation of the *gus* reporter gene indicates that the promoter of the target gene might be located upstream of the right junction, whereas the coding sequences might be downstream of the integrated T-DNA. This is a transcriptional gene fusion and the direction of transcription of the reporter gene when fused with a plant promoter will be from the right to the left junction. The second reason was the hypothetical function of this gene predicted later by *in silico* analysis. The predicted function of *frf* was the hormone response, and one could easily correlate this hypothetical function with the phenotype observed in the plants of line no. 197. The third reason for selecting *frf* was based on the results we obtained from reverse transcriptase PCR (RT-PCR). Results of RT-PCR indicated that the level of *frf* transcript especially in the shoot apex of the mutant plant was severely reduced in comparison with that of the wild-type control plants (data not shown).

Searching for protein-sequence similarity resulted in the identification of a growth-regulating factor (Os-GRF1) of *Oryza sativa*, [32] which shows a high protein-sequence similarity with FRF, as demonstrated in Fig. 3. In *Oryza sativa* this protein is encoded by a novel gibberellin-induced gene and has a potential regulatory role in stem growth. [32] Os-GRF1 contains three domains QLQ, WRC and TQL with similarity to other sequences in the database. According to van der Knapp et al. [32] the QLQ domain may be involved in protein-to-protein interactions, whereas the WRC domain is most likely to function in DNA binding. The TQL domain, however, had not been recognized previously in plant proteins and its function could therefore not be determined [32]. Os-GRF1 thus displays general features of a transcription factor and may play a regulatory role in GA-induced stem elongation [32]. Van der Knapp et al. [32] expressed the gene Os-GRF1 in *A. thaliana* to study its role in plant growth. However, rather than promoting plant growth, expression of this gene led to a severe reduction in stem elongation and the normal growth of the plant could not be recovered by application of GA. This dwarf phenotype of the transgenic *Arabidopsis* plants could be a result of a gain-of-function with respect to one component of growth and thereby, to an imbalance between growth-related processes [32]. Van der Knapp et al. [32] also suggested that the expression of Os-GRF1 in *Arabidopsis* disrupts the function of the shoot apical meristem. In the protein sequence alignment shown in Fig. 3, it could be seen that in the QLQ and WRC [32] domains of Os-GRF1 the sequence identity was high, resulting in 50% and 72% similarity, respectively. Considering the results of similarity search and homology determination as well as the phenotype of the mutant plants (line no. 197) it could be hypothesized that the function of FRF is similar to that of Os-GRF1.

For further estimation of the function of FRF, a secondary database search was performed. The tool Interpro [25] was used to run the predicted protein sequence of FRF against secondary databases. The derived information suggested that FRF is involved in regulation of plants growth. Furthermore, the matches against a motif for O-Glycosyl hydrolase also supported the predicted function of FRF as a growth-regulating factor since many of the growth-regulating processes are initiated by glycosyl hydrolyase [35].

Although FRF and Os-GRF1 are similar to some extent in the amino-acid sequence level, not much could be assumed about their common function. Sequence comparison could fail to identify many of the relationships that appear once the protein's structure was known. For further continuation of this analysis we employed a fold-recognition method to predict the three-dimensional structure of FRF.

By using THREADER, a possible fold 2LBD was predicted for the FRF sequence based on a very high combined shuffled *Z*-score value of 6.19. Since threading using THREADER might be misguided because of a small fold database (i.e. the recently discovered folds might not be included in the database), and the fact that the threading algorithms underlying these tools are different (i.e. they might capture different aspects) we performed an alternative threading using 3D-PSSM as described in the methods. However, no significant results were obtained (data not shown). In our experience, 3D-PSSM performs well when sequence similarity is fairly high. However, when sequence similarity to known

structures decreases, as it was in our case, THREADER, which accounts and evaluates also the atomic features of the structure, is generally preferable. The reason for using multiple tools is, apart from the sequence similarity, that the THREADER does not perform well with all kind of proteins such as the transmembrane proteins. However, the results we obtained (data not shown) from the cellular prediction using PSORTb [36] indicated that FRF does not belong to transmembrane proteins. The probability that FRF belongs to nuclear proteins is high (60.9%), to mitochondrial proteins is medium (30.4%) and to cytoplasmic proteins is very low (8.7%). These data indicate that the THREADER results are trustworthy.

As shown in Table 1, the initial threading did not give any significant fold (i.e. combined energy *Z*-score >3.5). The top ranked folds were either borderline significant ($2.7<Z<3.5$) which means that the match is possibly correct, or had a poor score ($2.0<Z<2.7$) which means that the match could be right, but needs other confirmation. For this confirmation the top 30 folds captured in the initial threading were shuffled in order to eliminate the false positive matches and to highlight the true positives ones. The result, shown in Table 2, indicates that the best match obtained by THREADER is the fold 2LBD. It stands out quite clearly with a combined shuffled *Z*-score of 6.19, indicating that the match is correct. Table 2 also indicates that 2LBD is not the only fold with a high combined shuffled *Z*-score. Three other folds with a *Z*-score higher than 3.5 were also suggested (1J9 J, 1FDR and 1A80), but with a lower confidence (Table 2). Here we followed the THREADER manual that describes combined shuffled *Z*-scores of 3.5 as significant for 50 shuffles. These 50 shufflings are assumed to give rise to a normal distribution.

For obtaining an increased quality of structural models, all unrelated regions of the template fold 2LBD (longer than nine amino acids) were scanned separately in the PDB database. As this scanning resulted in no matches, the sequences were then threaded against the THREADER database. Our approach was based on insertion of two additional templates (1KQ1 and 1GNF) to supplement the model based on the major template 2LBD. A similar type of structure prediction has been described previously [37, 38]. These authors predicted the structures of the long unrelated regions based on the proteins' conformational energy. In our case, we performed the structure prediction of the unrelated regions by providing templates 1KQ1 and GNF1 obtained from THREADER. The final model was then adjusted based on the proteins' conformational energy.

The C-terminal region of FRF was not included in the prediction of structural models, as this region did not correspond to the template fold 2LBD. Deletion of a particular region of a protein sequence from modeling experiments has been described previously [39]. To investigate the C-terminal domain of FRF, we employed a similar procedure as for the entire sequence, but for threading only amino acids 272 to 398 were used. This investigation resulted matching of the sequence with two significant folds, 1FDR and 1HYH with *Z*-scores of 4.33 and 3.94, respectively. These analyses also revealed that both of these folds are involved in binding variants of adenine-dinucleotides (flavin and nicotinamide).

2LBD is a ligand-binding domain of the human retinoic acid receptor gamma-2 protein (RRG2). RRG2 is a receptor for retinoic acid, which, in higher eukaryotes, is involved in the regulation of numerous essential physiological processes [40]. RRG2 belongs to the superfamily of nuclear hormone receptors, which being transcription regulators are involved in diverse physiological functions such as the control of embryonic development, cell division and differentiation [40]. The family of nuclear hormone receptors contains a conserved domain that is involved in specific DNA binding of the receptor to its target DNA sequence. This family also contains a ligand-binding domain that is responsible for binding of hormones. The ligand-binding domain is characterized by its ligand-dependent activation, which is critical for the regulation of transcription. So, in the absence of ligands the receptors are weakly associated with nuclear components. The nuclear hormone receptors are ligand-activated transcription factors, which interact with specific DNA elements (hormone response elements) to regulate transcription [41].

The results obtained from both the homology search and the fold recognition indicate that FRF is involved in transcriptional regulation. The fold recognition and the structure prediction results suggest that this transcriptional regulation is controlled by binding of a hormone or steroid to the ligand-binding domain of FRF. However, this does not exclude other possible substrates with similar composition and structure to bind to FRF.

The plant hormones that are frequently involved in regulation of flowering time in higher plants are the gibberellins [42]. Considering the phenotype of the mutant plant and the hypothetical function of the predicted protein, we believe that FRF could be involved in regulating GA biosynthesis in *A. thaliana*. To test this hypothesis, we exposed the plants of line no. 197 to GA. Results of this experiment demonstrate that the delay in flowering time observed in the mutant plants of line 197 was significantly reduced, exhibiting an earlier onset of flowering almost similar to that in wild-type control plants (Fig. 5).

Our approach is based on the hypothesis that bioinformatic tools can be applied for prediction of gene functions. However, a successful outcome requires accurate tools and techniques as well as correct information, i.e. databases. Unfortunately, the accuracy of the tools and the database information are not 100% correct. To minimize this problem we have combined information from different levels, e.g. primary structures and 3D folds of the proteins, and employed different alternative tools for the same task. By integrating multiple tools and information sources we believe that we have eliminated the possible imperfections that could occur in our investigations. Molecular and biochemical experiments

can be designed for providing further evidences supporting the results we have predicted by *in silico* analyses.

References

1. Pearce RS, Houlston CE, Atherton KM, Rixon JE, Harrison P, Hughes MA, Dunn MA (1998) Plant Physiol 117:787–795
2. Martin GB (1998) Curr Opin Biotech 9:220–226
3. Urao T, Katagiri T, Mizoguchi T, Yamaguchi-Shinozaki K, Hayashida N, Shinozaki K (1994) Mol Gen Genet 244:331–340
4. Nawrath C, Métraux J-P (1999) Plant Cell 11:1393–1404
5. Feldmann KA (1991) Plant J 1:71–82
6. Krysan PJ, Young JC, Sussman MR (1999) Plant Cell 11:2283–2290
7. The Arabidopsis Genome Initiative (2000) Nature 408:796–815
8. Bowie JU, Luethy R, Eisenberg D (1991) Science 253:164–170
9. Jones DT (1999) J Mol Biol 287:797–815
10. Brenner SE, Chothia C, Hubbard TJP (1997) Curr Opin Struct Biol 7:369–376
11. Jones DT (2000) Curr Opin Struct Biol 10:371–379
12. Sánches R, Peiper U, Melo F, Eswar N, MartŢ-Renom MA, Madhusudhan MS, Mirković N, Šali A (2000) Nat Struct Biol 7:986–990
13. Marti-Renom M, Stuart AC, Fiser A, Sanches R, Melo F, Sali A (2000) Annu Rev Biophys Biomol Struct 29:291–325
14. Fetrow JS, Godzik A, Skolnick J (1998) J Mol Biol 282:703–711
15. Sali A, Matsumoto R, McNeil HP, Karplus M, Stevens RL (1993) J Biol Chem 268:9023-9034
16. Mandal A, Lång V, Orczyk W, Palva ET (1993) Theor Appl Genet 86:621–628
17. Mandal A, Sandgren M, Holmström K-O, Gallois P, Palva ET (1995) Plant Mol Biol Rep 13:243–254
18. Murashige T, Skoog F (1962) Phys Plantarum 15:473–497
19. Bailey NTJ (1993) The use of t-tests for small samples. In: Bailey NTJ (ed) Statistical methods in biology. Cambridge University Press, Cambridge
20. MINITAB Reference manual (1996) Minitab Inc
21. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL (2000) Nucleic Acids Res 1:15–18
22. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) J Mol Biol 215:403–410
24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Nucleic Acids Res 25:3389-3402
25. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJA, Zdobnov EM (2001) Nucleic Acids Res 29:37–40
26. Thompson JD, Higgins DG, Gibson TJ (1994) Nucleic Acids Res 22:4673–4680
27. Omiga™ 2.0 (1999) User Guide. Oxford Molecular Ltd CS1090-02
28. Jones DT (1997) Curr Opin Struct Biol 7:377–387
29. Kelley LA, MacCallum RM, Sternberg MJE (2000) J Mol Biol 299:499–520
30. Sanchez R, Sali A (1997) Curr Opin Struct Biol 7:206–214
31. Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) J Appl Crystallogr 26:283–291
32. van der Knaap E, Kim JH, Kende H (2000) Plant Physiol 122:695–704
33. Ramachandran GN, Ramakrishnan C, Sasisekharan V (1963) J Mol Biol 7:95–99
34. Levy YY, Dean C (1998) Curr Opin Plant Biol 1:49–54
35. Thomas BR, Inouhe M, Simmons CR, Nevins DJ (2000) Int J Biol Macromol 27:145–149
36. Nakai K, Horto P (1999) Trends Biochem Sci 24:34–35
37. Rehm BHA, Antonio RV, Spiekermann P, Amara AA, Steinbüchel (2002) Biochim Biophys Acta 1594:178–190
38. Ettrich R, Melichercik M, Teisinger J, Ettrichova O, Krumscheid K, Hofbauerova K, Kvasnicka P, Schoner W, Amler E (2001) J Mol Model 7:184–192
39. Léonard N, Lambert C, Depiereux E, Wouters J (2003) NeuroToxicology (in press)
40. Egea PF, Rochel N, Birck C, Vachette P, Timmins PA, Moras D (2001) J Mol Biol 307:557–576
41. Lehmann JM, Hoffmann B, Pfahl M (1991) Nucleic Acids Res 19:573–578
42. Mouradov A, Cremer F, Coupland G (2002) Plant Cell S111-S130