J. Amer. Soc. Hort. Sci. 133(3):459–461. 2008.

DGTF: A Database of Grape Transcription Factors

Bin Cai
Biotechnology Research Institute, Shanghai Academy of Agricultural Sciences, 2901 Beidi Road, Shanghai, 201106 China; and College of Horticulture, Nanjing Agricultural University, 1 Weigang, Nanjing, 210095, China

Cheng-Hui Li
Suzhou Polytechnic Institute of Agriculture, 279 Xiyuan Road, Suzhou, 215008, China

Ai-Sheng Xiong, Ri-He Peng, Jun Zhou, and Feng Gao
Biotechnology Research Institute, Shanghai Academy of Agricultural Sciences, 2901 Beidi Road, Shanghai, 201106 China

Zhen Zhang
College of Horticulture, Nanjing Agricultural University, 1 Weigang, Nanjing, 210095, China

Quan-Hong Yao[1]
Biotechnology Research Institute, Shanghai Academy of Agricultural Sciences, 2901 Beidi Road, Shanghai, 201106 China

Additional index words. *Vitis vinifera* L., family, domain, transcriptional regulator

Abstract. The database of grape transcription factors (DGTF) is a plant transcription factor (TF) database comprehensively collecting and annotating grape (*Vitis* L.) TF. The DGTF contains 1423 putative grape TF in 57 families. These TF were identified from the predicted wine grape (*Vitis vinifera* L.) proteins from the grape genome sequencing project by means of a domain search. The DGTF provides detailed annotations for individual members of each TF family, including sequence feature, domain architecture, expression information, and orthologs in other plants. Cross-links to other public databases make its annotations more extensive. In addition, some other transcriptional regulators were also included in the DGTF. It contains 202 transcriptional regulators in 10 families.

Transcription factors (TF) are identified by their affinity for specific motifs in promoters, upstream regulatory elements, or enhancer regions of target genes (Riechmann et al., 2000). These factors bind specifically to their DNA-binding sites near target genes and then activate or repress gene transcription (Zhang, 2003). It is essential to identify and characterize TF on a genome-wide level to understand their biological function and to explore the mechanisms of transcriptional regulation. Recently, some databases of eukaryotic transcription factors have become available on the web. TRANSFAC (Matys et al., 2003), DBD (Wilson et al., 2008), PlnTFDB (Riaño-Pachón et al., 2007), AGRIS (Palaniswamy et al., 2006), DATF (Guo et al., 2005), DRTF (Guo et al., 2006), and DPTF (Zhu et al., 2007) have provided some comprehensive information on TF.

Grape is one of the most important horticultural crops in the world, being used for the production of wine and juice, and as fresh and dried fruit (Tinlot and Rousseau, 1993). Therefore, it is necessary to develop genomic tools to accelerate the acquisition of knowledge about its important agronomic characteristics such as resistance to diseases, tolerance to abiotic stress, and maturation and quality of the fruit. Given the importance of TF in the life cycle of plants, identification and annotation of the TF in grape will improve the understanding of these agronomic characteristics at the level of gene expression and regulation. With the completion of the grape genome sequence (Jaillon et al., 2007), the entire complement of genes coding for TF can be identified and described. A database of grape transcription factors (DGTF) will provide a resource for researchers to explore the expression and function of TF of grape.

Materials and Methods

Source datasets. We downloaded 30,434 wine grape mRNAs (Vitis_vinifera_mRNA _v1.fa), the corresponding protein sequences (Vitis_vinifera_peptide_v1.fa), and general feature formats (GFF; Wellcome Trust Sanger Institute, 2007) of genes (Vitis_vinifera_annotation_v1.gff) from Genoscope (Center National de Séquençage, Evry, France).

Data processing. TF can be identified and grouped into different families based upon their DNA-binding domains (Riaño-Pachón et al., 2007; Riechmann et al., 2000). For some families, a TF contains a single domain, which is sufficient to assign its membership. However, in other families, a TF may contain more than one DNA-binding domain; likewise, some domains are shared by different TF families. To correctly identify and classify TF into different families, we constructed a rule for identification and classification of each TF based on the literature (Riaño-Pachón et al., 2007; Riechmann et al., 2000) and the grape-specific combination of domains in each of the grape TF (data not shown). The rule was depicted as a graph on the DGTF web site (for details, see the DGTF Help page).

Received for publication 26 Dec. 2007. Accepted for publication 18 Mar. 2008. The research was supported by the Shanghai Project for ISTC (055407068), by the Shanghai Subject Chief Scientist (06XD14017), by the Project of the key laboratory of Shanghai (05dz223266-07dz22011), by the National Natural Science Foundation (30471258-30670179), and by the 863 Program (2006AA10Z117-06Z358).
[1]Corresponding author. E-mail: yaoquanhong_sh@yahoo.com.cn.

A pipeline we have developed for the identification, classification, and annotation of TF is shown in Fig. 1. In the first step, we collected hidden Markov models (HMM) of domains that occur in TF from the Pfam database (version 22.0; Finn et al., 2006). For the families without DNA-binding domain HMM available in Pfam, new HMM were created in-house based on alignments of TF from Arabidopsis [*Arabidopsis thaliana* (L.) Heynh.]. These HMM were then used to search against 30,434 predicted proteins from Genoscope by using the hmmsearch program (Eddy, 1998). The E-value cut-off 0.01 was used in the search and all significant hits were kept. For the NOZZLE and SAP (STERILE APETALA) families, each of which has only a single member in Arabidopsis, a BLAST (Altschul et al., 1997) search was run for the homolog in grape, and the E-value cut-offs were inspected. The rule was then implemented in a Perl script to classify proteins into different families.

To obtain the UniGene (National Center for Biotechnology Information, Bethesda, MD) entry corresponding to each TF, a BLAST search against expressed sequence tags (EST) of UniGene was performed (E-value < 10^{-10}, identity > 90%). The accession numbers of matched EST and the UniGene ID for each TF were recorded. For each TF, links to various external public sequence databases, Pfam, PlnTFDB, SIMAP (Rattei et al., 2008), and NCBI (National Center for Biotechnology Information, Bethesda, MD; Wheeler et al., 2006) were collected. GFF was used for drawing gene structure. Domain structures were also identified and annotated by hmmpfam (Eddy, 1998). Orthologs of each TF in Arabidopsis, rice (*Oryza sativa* L.), and poplar (*Populus trichocarpa* Torr. & Gray) were detected using best-reciprocal BLAST hits. In addition, the multiple sequence alignment of the DNA-binding domains of TF in each family was created by CLUSTAL W (Thompson et al., 1994). The neighbor-joining phylogenetic tree for each family was also constructed based on the alignment of predicted amino acid sequences.

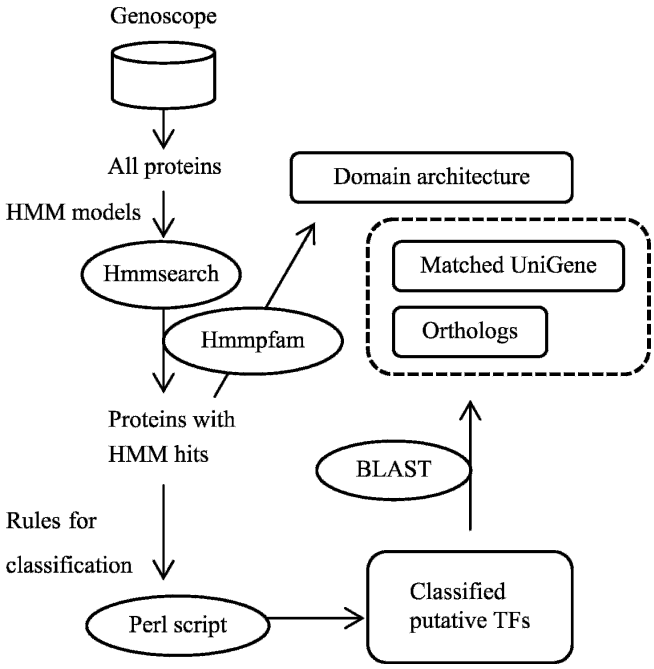

Fig. 1. Pipeline for the identification, classification, and annotation of transcription factors (TF). The pipeline starts with the complete collection of predicted proteins from Genoscope (Center National de Séquençage, Evry, France). Hidden Markov models (HMM) of DNA-binding domains are collected or created. These HMM are then used to search against all of the predicted proteins by using the hmmsearch program, and all significant hits are kept. A Perl script produces a list of putative TF grouped into different families according to a rule we developed for family classification. A BLAST search against expressed sequence tags (EST) of UniGene (National Center for Biotechnology Information, Bethesda, MD) is performed to obtain the UniGene entry corresponding to each TF. Finally, domain structures are also identified and annotated by hmmpfam, and orthologs of each TF in other plants are detected using best-reciprocal BLAST hits.

In addition to TF, some other types of transcriptional regulators were also included in the DGTF. These transcriptional regulators contain binding domains for: ARID (A-T rich interaction domain), HMG (high mobility group), MBF1 (multiprotein bridging factor 1), SNF2 (named after the *Saccharomyces cerevisiae* Hansen protein SNF2), Aux/IAA, Jumonji, PHD (plant homeodomain), DDT (DNA-binding homeobox and different TF), LUG (LEUNIG), and SET [named after three *Drosophila melanogaster* Meigen genes involved in epigenetic processes, *Su(var)*, *E(z)* and *trithorax*].

Results and Discussion

Identification and classification of grape putative TF. Using the pipeline we developed, 1423 putative TF in grape were identified and classified into 57 families. These putative TF were identified from 30,434 proteins that were predicted by the grape genome sequencing project. Several resources were used to build the gene models automatically with GAZE (Jaillon et al., 2007). Although our pipeline can improve the reliability and accuracy of gene prediction, some proteins may be annotated incorrectly. Therefore, users should validate any TF from our database that are going to be used in further research. In addition, 202 other transcriptional regulators were identified and classified into 10 families.

Annotation of grape putative TF. A UniGene entry is a set of transcript sequences that appear to come from the same transcription locus, together with information on protein similarities, gene expression, cDNA clone reagents, and genomic location (Wheeler et al., 2006). In this study, 55% of the TF matched UniGene clusters. The corresponding UniGene entry for each TF was stored as annotation information, which may provide valuable expression information for further analysis of TF. Using a similarity search, of 1423 grape TF, 98% were found to have orthologs in Arabidopsis, rice, and poplar.

Implementation and user interface. The DGTF allows users to start their data-mining by browsing a list of families (Cai, 2008). Clicking on one of the family names will show users a summary page of the family, including the family description extracted from the literature, the list of protein names, the multiple sequence alignment of the DNA-binding domains, and the neighbor-joining phylogenetic tree. Detailed information for each of the TF can be accessed by clicking the protein name or by entering the protein name into a search form on the top of the page. Alternatively, users can search the DGTF by running a BLAST search against the sequences in the DGTF.

The DGTF provides detailed information on individual TF, including chromosomal location, predicted molecular weight and pI (isoelectric point), gene structure, gene feature, corresponding UniGene, and expression profile. The detailed information on EST in the UniGene is also available. For each TF, the information provided in the DGTF is linked to various external public sequence databases: Pfam, plnTFDB, SIMAP,

NCBI, and the grape genome browser in Genoscope. All the coding sequences (CDS), protein sequences, and GFF can be downloaded through the DGTF website for further analysis.

Future plans. DGTF is the first database that comprehensively collects and annotates grape TF based on genome-wide data. The DGTF will be a useful resource for research on grape transcription regulation. As more data and information on the genome sequence becomes available, we will maintain and update the DGTF regularly. Furthermore, as soon as new plant genomes become available, the methods described in this study will be applied to them.

Literature Cited

Altschul, S., T. Madden, A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Cai, B. 2008. The database of grape transcription factors. List of transcription factor families. 15 Mar. 2008. <http://www.yaolab.sh.cn/dgtf.html>.

Eddy, S.R. 1998. Profile hidden Markov models. Bioinformatics 14:755–763.

Finn, R.D., J. Mistry, B. Schuster-Böckler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S.R. Eddy, E.L.L. Sonnhammer, and A. Bateman. 2006. Pfam: Clans, web tools and services. Nucleic Acids Res. 34:247–251.

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

Guo, A., K. He, D. Liu, S. Bai, X. Gu, L. Wei, and J. Luo. 2006. DRTF: A database of rice transcription factors. Bioinformatics 22:1286–1287.

Jaillon, O., J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, A. Vezzi, F. Legeai, P. Hugueney, C. Dasilva, D. Horner, E. Mica, D. Jublot, J. Poulain, C. Bruyère, A. Billault, B. Segurens, M. Gouyvenoux, E. Ugarte, F. Cattonaro, V. Anthouard, V. Vico, C. Del Fabbro, M. Alaux, G. Di Gaspero, V. Dumas, N. Felice, S. Paillard, I. Juman, M. Moroldo, S. Scalabrin, A. Canaguier, I. Le Clainche, G. Malacrida, E. Durand, G. Pesole, V. Laucou, P. Chatelet, D. Merdinoglu, M. Delledonne, M. Pezzotti, A. Lecharny, C. Scarpelli, F. Artiguenave, M.E. Pè, G. Valle, M. Morgante, M. Caboche, A.F. Adam-Blondon, J. Weissenbach, F. Quétier, and P. Wincker, and French-Italian Public Consortium for Grapevine Genome Characterization. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463–468.

Matys, V., E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A.E. Kel, O.V. Kel-Margoulis, D.U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31:374–378.

Palaniswamy, S.K., S. James, H. Sun, R.S. Lamb, R.V. Davuluri, and E. Grotewold. 2006. AGRIS and AtRegNet: A platform to link cis-regulatory elements and transcription factors into regulatory networks. Plant Physiol. 140(3):818–829.

Rattei, T., P. Tischler, R. Arnold, F. Hamberger, J. Krebs, J. Krumsiek, B. Wachinger, V. Stümpflen, and W. Mewes. 2008. SIMAP: Structuring the network of protein similarities. Nucleic Acids Res. 36:289–292.

Riaño-Pachón, D.M., S. Ruzicic, I. Dreyer, and B. Mueller-Roeber. 2007. PlnTFDB: An integrative plant transcription factor database. BMC Bioinformatics 8:1–10.

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

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.

Tinlot, R. and M. Rousseau. 1993. The state of viticulture in the world and the statistical information in 1992. Bulletin de l'O.I.V. 66:861–946.

Wellcome Trust Sanger Institute. 2007. GFF: An exchange format for feature description. 15 Mar. 2008. <http://www.sanger.ac.uk/Software/formats/GFF/>.

Wheeler, D.L., T. Barrett, D.A. Benson, S.H. Bryant, K. Canese, V. Chetvernin, D.M. Church, M. DiCuccio, R. Edgar, S. Federhen, L.Y. Geer, W. Helmberg, Y. Kapustin, D.L. Kenton, O. Khovayko, D.J. Lipman, T.L. Madden, D.R. Maglott, J. Ostell, K.D. Pruitt, G.D. Schuler, L.M. Schriml, E. Sequeira, S.T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, T.O. Suzek, R. Tatusov, T.A. Tatusova, L. Wagner, and E. Yaschenko. 2006. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 34:173–180.

Wilson, D., V. Charoensawan, S.K. Kummerfeld, and S.A. Teichmann. 2008. DBD: Taxonomically broad transcription factor predictions: New content and functionality. Nucleic Acids Res. 36:88–92.

Zhang, J.Z. 2003. Overexpression analysis of plant transcription factors. Curr. Opin. Plant Biol. 6(5):430–440.

Zhu, Q.H., A.Y. Guo, G. Gao, Y.F. Zhong, M. Xu, M. Huang, and J. Luo. 2007. DPTF: A database of poplar transcription factors. Bioinformatics 23:1307–1308.