D1118–D1122 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 8 November 2010*
doi:10.1093/nar/gkq1120

AGRIS: the Arabidopsis Gene Regulatory Information Server, an update

Alper Yilmaz[1,*], Maria Katherine Mejia-Guerra[1], Kyle Kurz[2], Xiaoyu Liang[2], Lonnie Welch[2] and Erich Grotewold[1]

[1]Department of Plant Cellular and Molecular Biology and Plant Biotechnology Center, The Ohio State University, Columbus, OH and [2]School of Electrical Engineering and Computer Science, Ohio University, Athens, OH, USA

Received September 3, 2010; Revised October 5, 2010; Accepted October 19, 2010

ABSTRACT

The Arabidopsis Gene Regulatory Information Server (AGRIS; http://arabidopsis.med.ohio-state.edu/) provides a comprehensive resource for gene regulatory studies in the model plant *Arabidopsis thaliana*. Three interlinked databases, AtTFDB, AtcisDB and AtRegNet, furnish comprehensive and updated information on transcription factors (TFs), predicted and experimentally verified *cis*-regulatory elements (CREs) and their interactions, respectively. In addition to significant contributions in the identification of the entire set of TF–DNA interactions, which are the key to understand the gene regulatory networks that govern *Arabidopsis* gene expression, tools recently incorporated into AGRIS include the complete set of words length 5–15 present in the *Arabidopsis* genome and the integration of AtRegNet with visualization tools, such as the recently developed ReIN application. All the information in AGRIS is publicly available and downloadable upon registration.

INTRODUCTION

A first step toward elucidating the dynamic behavior of gene regulatory networks involves compiling the parts lists, formed by the transcription factors (TFs), all the gene regulatory regions and the corresponding interactions (1). Over the past few years, *Arabidopsis* has emerged as a model plant (2), and system approaches that rely on the easy access to genomic and gene regulation information have begun to identify the architecture of the underlying networks for this organism (3). AGRIS is dedicated to storing information on TFs and *cis*-regulatory elements (CREs), and helping reveal gene regulatory networks in *Arabidopsis*. Different from other community resources such as ATTED-II, which focuses primarily on co-expression analysis (4), Athena that focuses on promoter analysis (5) or PlnTFDB (6), which harbors computationally curated information on TF from many different plant species, AGRIS is *Arabidopsis*–centered, primarily hand-curated, integrates TF and promoter information primarily from experimental sources and combines them into gene regulatory networks. AGRIS is currently composed of databases of TFs (AtTFDB), predicted and experimentally demonstrated CREs (AtcisDB), and the experimentally verified physical interactions between TFs and gene regulatory regions, represented in AtRegNet and visualized by the ReIN tool.

AtTFDB contains information on regulatory proteins, where TFs are grouped into families based on the presence of conserved domains and following prior classification criteria (7). The current version of AtTFDB contains 50 families and 1773 TFs. The identification of TFs involved a combination of BLAST and motif searches based on the literature available on known TFs, and on continuous manual literature curation.

AtcisDB holds information on sequences for the upstream regions of genes with CRE annotations. The current version of AGRIS (August 2010) contains 33 239 nuclear promoter sequences (not restricted to RNA polymerase II genes) with descriptions of putative and experimentally confirmed CREs. AtcisDB maps CREs to their respective locations in gene promoters, and displays them in a graphical form, clearly distinguishing predicted CREs from their experimentally validated counterparts.

DATABASE CONTENT

In the past year, AGRIS underwent major updates through the design of an updated frontend user interface and a major Arabidopsis genome release version. All

*To whom correspondence should be addressed. Tel: +1 614 688 4954; Fax: +1 614 292 5379; Email: yilmaz.11@osu.edu

(M)IPS <> (S)ALK <> T(A)IR
(C)onfirmed direct targets (U)nconfirmed direct targets

TF Family Name	TF Locus Id	Protein Name	Sub Family	Gene Name , Synonym	Links	Sequences	BindingSites/NFM	Direct Targets	Total Direct Interactions
bHLH	At1g09530	AtbHLH8	NA	PAP3, PIF3, POC1	M, S, A	Nucl-Prot		C(6)	VIEW (8)
bHLH	At1g32640	AtbHLH6	NA	ATMYC2, JAI1, JIN1, MYC2, RAP-1, RD22BP1, ZBF1	M, S, A	Nucl-Prot	bindingsites	C(2) - U(2)	VIEW (5)
bHLH	At1g51140	AtbHLH122	NA		M, S, A	Nucl-Prot			VIEW (2)
bHLH	At1g69010	AtbHLH102	NA	BIM2	M, S, A	Nucl-Prot			VIEW (2)
bHLH	At2g20180	AtbHLH15	NA	PIF1, PIL5	M, S, A	Nucl-Prot		C(189) - U(561)	VIEW (751)
bHLH	At2g43060	AtbHLH158	NA		M, S, A	Nucl-Prot			VIEW (3)
bHLH	At3g19860	AtbHLH121	NA		M, S, A	Nucl-Prot			
bHLH	At4g09820	AtbHLH42	NA	TT8	M, S, A	Nucl-Prot		C(1)	VIEW (1)
bHLH	At5g41315	AtbHLH1	NA	GL3, MYC6.2	M, S, A	Nucl-Prot		C(25) - U(696)	VIEW (722)

Figure 1. AtTFDB query result table. Either TF family browsing or search query results indicate Locus ID, synonyms, links, sequences and AtRegNet related data. The updated AtTFDB provides improved integration with AtRegNet. Confirmed direct target counts are indicated with 'C', and unconfirmed one are indicated with 'U'.

gene models in AtTFDB, AtcisDB and AtRegNet were upgraded from TAIR4 to TAIR9 (http://www.arabidopsis.org). This had a minimal impact on AtTFDB but substantially improved AtcisDB. Earlier versions of the AGRIS page-code were written in Java, which has now been replaced with a Perl-based graphical user interface. The HTML::Mason (http://www.masonhq.com/) module was used to embed Perl code within the HTML format. Additional functionality was added via Javascript to make the website more interactive.

AtTFDB

The *Arabidopsis* Transcription Factor Database (AtTFDB) contains information on 1773 TFs, grouped into 50 families, based on the presence of conserved domains, often involved in DNA-binding and/or dimerization. The current version of AGRIS contains an additional 83 TFs since the last major release (8) and these TFs were identified following the same criteria as previously described (7).

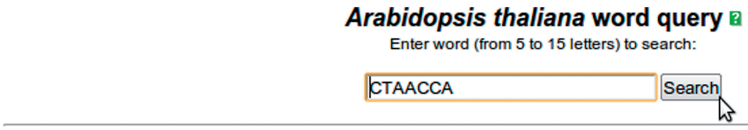
Users can access information on Individual TFs by browsing or searching with unique gene locus identifiers (AGI ID, Atxgxxxxx) or common gene names. The resulting table contains the AGI ID, gene synonyms, links to MIPS (9) (http://mips.helmholtz-muenchen.de/plant/athal/), SALK (10) (http://signal.salk.edu/cgi-bin/tdnaexpress) and TAIR (11) (http://arabidopsis.org/), nucleotide and protein sequences and information on the participation of the respective TF in a regulatory sub-network. The locus ID is linked to a summary page for each particular TF, which displays the most prominent information for it, including whether clones or other resources are available upon request from the Arabidopsis Biological Resource Center (ABRC, http://abrc.osu.edu/).

The updated version of AGRIS includes two additional columns ('Direct Targets' and 'Total Direct Interactions') in order to provide integrated regulatory information for each TF (Figure 1), which is extracted from the contents of AtRegNet (see below). The 'Direct Targets' column corresponds to the number of confirmed and unconfirmed targets; confirmed targets corresponding to those for which two or more experimental approaches confirmed them as directly regulated by a particular TF, and unconfirmed corresponding to those identified by just one experiment. The 'Total Direct Interactions' column corresponds to the total number of interactions in which that particular TF is involved, including the cases where a particular TF is targeted by other regulatory factors. In graph theory language, the 'Total Direct Interactions' number represents the degree or valency of that particular node.

AtcisDB

The *Arabidopsis cis*-regulatory element Database (AtcisDB) consists of a searchable relational database, which includes multiple data types, including TF-binding sites and complete upstream promoter sequence information. In the updated version of AGRIS, the upstream 3000 bp promoter sequence region of genes were obtained from TAIR9, resulting in a total of 33 239 sequences. To increase the scalability and extensibility, the Genome Data Visualization Toolkit (GDVTK) (12) used

A



B

Summary for word CTAACCA per genomic segment

The queried word matches the following experimentally verified binding sites
Mouse over binding site names for publication reference.

Motif	Binding Site Name
CTAACCA	AtMYB2 binding site in RD22

C

Segment Name	Unique sequence occurences	Total occurences	Expected unique sequences	Expected occurences	Rank	Score
5'UTR	135	138	138.1770	145.4980	11474	-3.1405
3'UTR	145	147	140.4380	144.9910	6672	4.6356
Intron	1299	1318	1242.2000	1303.0400	4249	58.0813
Core Promoter	180	180	173.4370	185.0920	6968	6.6861
Proximal Promoter	1612	1680	1653.6500	1717.5100	10997	-41.1195
Distal Promoter	3194	3448	3403.3900	3647.0400	13394	-202.8150
Genome-wide	5	8610	5.0000	8839.4500	9198	0.0000

D

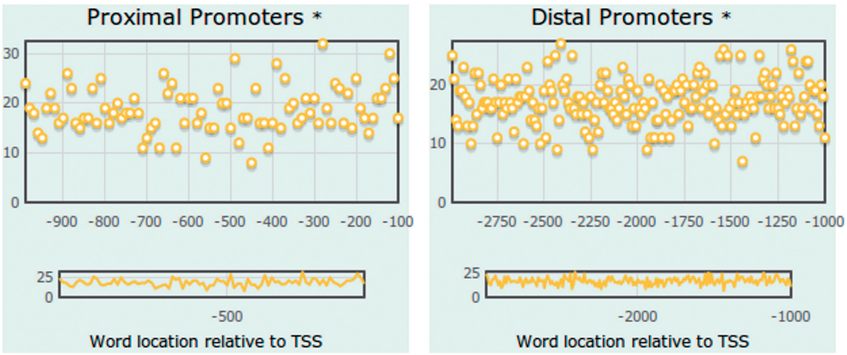


Figure 2. Results of querying word frequencies. Occurrence, scoring and positional distribution of a given word is calculated for various genomic segments. (**A**) A query is initiated by entering a word 5–15 letters (e.g. CTAACCA) in length into the search box. (**B**) The user is notified if the queried word matches a known CRE available in AtcisDB. (**C**) Total occurrences include possible multiple occurrences within a single sequence. For instance, CTAACCA is found a total of 138 times in 5′-UTR region sequences, and 135 genes contain this word in their 5′-UTR regions. Expected occurrences are calculated by a Hidden Markov Model, as described in ref. 14. The unique sequence occurrences column is hyperlinked to a list of genes that contain the word in the corresponding segments. (**D**) Flot (http://code.google.com/p/flot/), a Javascript plotting library, was used for visualization of positional distribution of a word within different sets. Only Proximal and Distal Promoters segments are shown here. Charts are interactive and are capable of zooming into a desired region.

to display gene regulatory information on genomic regions in previous releases of AGRIS, was replaced with the Generic Genome Browser (13) (http://gmod.org/wiki/GBrowse). The Generic Genome Browser framework allows the integration of AtcisDB with the visualization and integration of any genomic data, for instance word counts and CRE location information.

The latest AGRIS release makes genome-wide word counting an integral component of AtcisDB (14). Using the WordSeeker (http://word-seeker.org/) enumerative word discovery approach, putative CREs can be identified *de novo* (without prior knowledge) by investigating their over-representation and correlation with functional (e.g. gene expression) studies. Accessible through AtcisDB, the word landscape analysis of the *Arabidopsis* non-coding regions for word lengths 5–15 (http://arabidopsis.med.ohio-state.edu/words/) is available for complete download or for query with particular motifs (Figure 2), with over-representation scores calculated, as previously described (14).

In the context of WordSeeker approach, a 'word' is defined as a strings of letters [Adenine (A), Guanine (G), Cytosine (C) and Thymine (T)], each of which is found a specific number of times in a particular genome. Due to the integration of the WordSeeker tool, AGRIS is able to provide a complete word enumeration summary of user-queried words in non-coding domains of the *Arabidopsis* genome, including promoter regions, introns, and 3′ and 5′ untranslated regions (3′ UTR and 5′ UTR). Furthermore, the positional distribution of selected words in all genomic segments is visually displayed. This represents an important step toward fully cataloging the functional elements of the *Arabidopsis* genome.

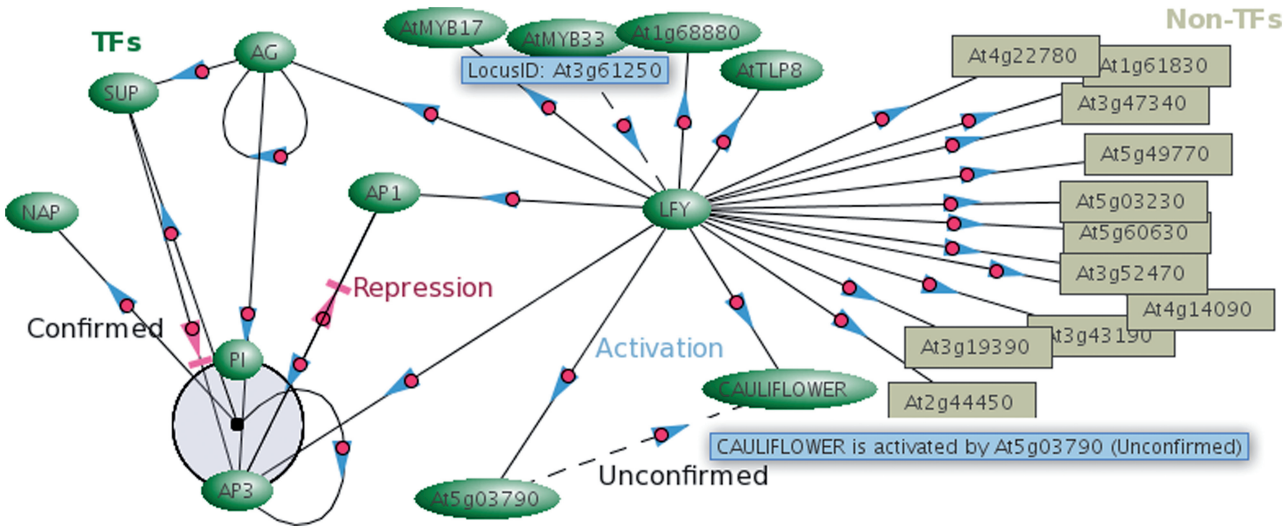


Figure 3. Screenshot from AGRIS depicting features of ReIN. For each Locus ID or gene name available in the AtRegNet database, ReIN displays all interactions involving the searched gene. Detailed descriptions of features are available in REIN tutorial page (http://arabidopsis.med.ohio-state.edu/REIN/ReINtutorial.html).

AtRegNet and ReIN

The *Arabidopsis thaliana* regulatory network database (AtRegNet) documents and visualizes networks formed by TFs and their direct target genes only (8). Currently, AtRegNet contains information on physical direct regulatory interactions between 8070 target genes, 64 TFs and three TF complexes. A complex in AtRegNet is defined as more than one transcriptional regulator recruited simultaneously, and often in a synergistic fashion, to DNA. The total gene regulatory network contains 8100 nodes, comprising 814 TF and 7286 non-TF encoding genes, connected by 11 123 edges. The information was parsed from 76 published studies, and derived from a combination of experimental approaches, including data generated from high-throughput *in vivo* DNA-binding techniques such as ChIP–Chip and ChIP-Seq.

Within these 11 123 edges, a set of 650 interactions was classified as 'confirmed', because they fulfilled the following criteria: A TF was shown to bind to the regulatory region of the target gene, AND *in vivo* evidence of regulation of the gene by the TF was available OR a TF directly regulates the target gene, AND *in vivo* evidence of regulation of the gene by the TF was available. For example, consider the direct targets of the basic helix–loop–helix TF AtbHLH15 which includes 750 direct targets identified by ChIP–Chip (15). Out of these 750 putative targets, gene-expression microarray experiments showed 11 as regulated by AtbHLH15, and were demonstrated to be bound by AtbHLH15 in ChIP–PCR experiments. Thus, these 11 genes are considered as 'confirmed' AtbHLH15 direct targets and the rest were classified as 'unconfirmed', awaiting further studies.

As a custom tool to view TF networks, we developed the Regulatory Networks Interaction Module (ReIN, http://arabidopsis.med.ohio-state.edu/REIN), an interactive tool capable of integrating AtRegNet data with the AtcisDB and AtTFDB databases and expanding networks on user's demand (Figure 3). ReIN complements the visualization of AtRegNet by more conventional network visualization applications, such as Cytoscape (16) (http://www.cytoscape.org).

In ReIN, nodes and edges are empowered with dynamic links that allow users to obtain additional information. For example, in the case of a node that corresponds to a TF target, the user can explore available information for this gene in AtcisDB, TAIR and AtTFDB, when appropriate. In addition, the user can expand the network by displaying, for example, targets of a target TF. Edges are linked to the abstract of the corresponding publication on PubMed that supports the relationship. ReIN also has the capability to add information provided by the user to any network displayed, and networks can be saved in a graphic format, or as TXT or XGMML files, the latter providing facile import into Cytoscape.

ReIN is accessible by clicking the 'Regulatory Networks' link in the title bar of all AGRIS pages, or through the 'Total Direct Interactions' column on the summary page obtained when displaying specific TFs or list of TF families from AtTFDB. This information is only shown for those TFs for which information is available in AtRegNet. To fully explore the capabilities of ReIN, a tutorial is available at http://arabidopsis.med.ohio-state.edu/REIN/ReINtutorial.html.

ACCESS TO THE DATABASE

Contents of AGRIS are freely accessible online. After a free registration process, users can download the database contents as plain text. Registration and downloads are possible through http://arabidopsis.med.ohio-state.edu/downloads.html.

ACKNOWLEDGEMENTS

We would like to thank to Arabidopsis community for their support and their help in the curation of the data available through AGRIS.

FUNDING

National Science Foundation (MCB-0418891 to E.G.); and National Institutes of Health (5 T32 CA106196-05 to A.Y.). Funding for open access charges: Ohio State University discretionary funds (to E.G.).

Conflict of interest statement. None declared.

REFERENCES

1. Schlitt,T. and Brazma,A. (2007) Current approaches to gene regulatory network modelling. *BMC Bioinformatics*, **8(Suppl. 6)**, S9.
2. Koornneef,M. and Meinke,D. (2010) The development of *Arabidopsis* as a model plant. *Plant J.*, **61**, 909–921.
3. Moreno-Risueno,M.A., Busch,W. and Benfey,P.N. (2010) Omics meet networks - using systems approaches to infer regulatory networks in plants. *Curr. Opin. Plant Biol.*, **13**, 126–131.
4. Obayashi,T., Hayashi,S., Saeki,M., Ohta,H. and Kinoshita,K. (2009) ATTED-II provides coexpressed gene networks for *Arabidopsis*. *Nucleic Acids Res.*, **37**, D987–D991.
5. O'Connor,T.R., Dyreson,C. and Wyrick,J.J. (2005) Athena: a resource for rapid visualization and systematic analysis of *Arabidopsis* promoter sequences. *Bioinformatics*, **21**, 4411–4413.
6. Perez-Rodriguez,P., Riano-Pachon,D.M., Correa,L.G., Rensing,S.A., Kersten,B. and Mueller-Roeber,B. (2010) PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.*, **38**, D822–D827.
7. Davuluri,R., Sun,H., Palaniswamy,S., Matthews,N., Molina,C., Kurtz,M. and Grotewold,E. (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of *Arabidopsis cis*-regulatory elements and transcription factors. *BMC Bioinformatics*, **4**, 25.
8. Palaniswamy,S.K., James,S., Sun,H., Lamb,R.S., Davuluri,R.V. and Grotewold,E. (2006) AGRIS and AtRegNet. A platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol.*, **140**, 818–829.
9. Spannagl,M., Noubibou,O., Haase,D., Yang,L., Gundlach,H., Hindemitt,T., Klee,K., Haberer,G., Schoof,H. and Mayer,K.F. (2007) MIPSPlantsDB–plant database resource for integrative and comparative plant genome research. *Nucleic Acids Res.*, **35**, D834–D840.
10. Alonso,J.M., Stepanova,A.N., Leisse,T.J., Kim,C.J., Chen,H., Shinn,P., Stevenson,D.K., Zimmerman,J., Barajas,P., Cheuk,R. *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science*, **301**, 653–657.
11. Swarbreck,D., Wilks,C., Lamesch,P., Berardini,T.Z., Garcia-Hernandez,M., Foerster,H., Li,D., Meyer,T., Muller,R., Ploetz,L. *et al.* (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nucleic Acids Res.*, **36**, D1009–D1014.
12. Sun,H. and Davuluri,R.V. (2004) Java-based application framework for visualization of gene regulatory region annotations. *Bioinformatics*, **20**, 727–734.
13. Stein,L.D., Mungall,C., Shu,S., Caudy,M., Mangone,M., Day,A., Nickerson,E., Stajich,J.E., Harris,T.W., Arva,A. *et al.* (2002) The generic genome browser: a building block for a model organism system database. *Genome Res.*, **12**, 1599–1610.
14. Lichtenberg,J., Yilmaz,A., Welch,J.D., Kurz,K., Liang,X., Drews,F., Ecker,K., Lee,S.S., Geisler,M., Grotewold,E. *et al.* (2009) The word landscape of the non-coding segments of the Arabidopsis thaliana genome. *BMC Genomics*, **10**, 463.
15. Oh,E., Kang,H., Yamaguchi,S., Park,J., Lee,D., Kamiya,Y. and Choi,G. (2009) Genome-wide analysis of genes targeted by PHYTOCHROME INTERACTING FACTOR 3-LIKE5 during seed germination in Arabidopsis. *Plant Cell*, **21**, 403–419.
16. Cline,M.S., Smoot,M., Cerami,E., Kuchinsky,A., Landys,N., Workman,C., Christmas,R., Avila-Campilo,I., Creech,M., Gross,B. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nat. Protoc.*, **2**, 2366–2382.