BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 12 2005, pages 2906–2908
doi:10.1093/bioinformatics/bti435

Genome analysis

Rapid and selective surveillance of *Arabidopsis thaliana* genome annotations with Centrifuge

Florence Armand[1], Philipp Bucher[2,3], C. Victor Jongeneel[2,4] and Edward E. Farmer[1,*]

[1]Gene Expression Laboratory, Plant Molecular Biology, University of Lausanne, Biology Building, 1015 Lausanne, Switzerland, [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [3]ISREC, 1066 Epalinges, Switzerland and [4]Office of Information Technology, Ludwig Institute for Cancer Research, 1015 Lausanne, Switzerland

Received on December 3, 2004; revised on March 22, 2005; accepted on April 5, 2005
Advance Access publication April 7, 2005

ABSTRACT

Summary: Centrifuge is a user-friendly system to simultaneously access Arabidopsis gene annotations and intra- and inter-organism sequence comparison data. The tool allows rapid retrieval of user-selected data for each annotated Arabidopsis gene providing, in any combination, data on the following features: predicted protein properties such as mass, pI, cellular location and transmembrane domains; SWISS-PROT annotations; Interpro domains; Gene Ontology records; verified transcription; BLAST matches to the proteomes of *A.thaliana*, *Oryza sativa* (rice), *Caenorhabditis elegans*, *Drosophila melanogaster* and *Homo sapiens*. The tool lends itself particularly well to the rapid analysis of contigs or of tens or hundreds of genes identified by high-throughput gene expression experiments. In these cases, a summary table of principal predicted protein features for all genes is given followed by more detailed reports for each individual gene. Centrifuge can also be used for single gene analysis or in a word search mode.

Availability: http://centrifuge.unil.ch/

Contact: edward.farmer@unil.ch

INTRODUCTION

Genomic sequencing of the key model plant *Arabidopsis thaliana* yielded results of particularly high quality (Arabidopsis Genome Initiative, 2000). Gene structure predictions as well as most functional annotations and many related datasets were provided by The Institute for Genomic Research (http://www.tigr.org) and can be accessed at The Arabidopsis Information Resource (TAIR, Rhee *et al*., 2003; http://www.arabidopsis.org/) and the Munich Information Center for Protein Sequences (MIPS) *Arabidopsis thaliana* database (MAtDB; http://mips.gsf.de/proj/thal/db/). In addition to providing genome-derived data and powerful tools for data analysis, TAIR, the largest community resource, also offers links to genetic resources such as ecotype collections and pools of knock-out mutants, etc. A current problem is that, despite the fact that annotations of the *A.thaliana* sequence are frequently updated, annotations for many genes, particularly those unique to plants, are still of poor quality (Gutiérrez *et al*., 2004) and will evolve greatly in the future. Centrifuge is complementary to current tools for the scrutiny of genome annotation at TAIR but has the advantage of having a single entry page that can accommodate tailored requests for data. This simple tool is highly flexible, allowing the analysis of individual genes, lists of genes from gene expression studies or genomic regions. The fast analysis of gene lists from microarray studies is greatly facilitated with Centrifuge.

SYSTEM OVERVIEW

Centrifuge is the web interface of a pre-assembled *Arabidopsis* database which is automatically compiled from the sources listed in Table 1. The database is compiled as follows: firstly, general annotations for each gene are retrieved from the Reference Sequence (RefSeq) collection from NCBI. TAIR, which employs algorithms to calculate isoelectric point (pI) and mass as well as TargetP, for the prediction of cellular location, and TMHMM (transmembrane hidden Markov model) to predict transmembrane domains was also mined for these features (Table 1). We also use subsets of the annotations in UNIPROT (protein functional characteristics and related literature) and InterPro (protein domains with annotations). Gene Ontology Consortium information was incorportated into the database. In addition, for every *Arabidopsis* gene, we performed and incorported BlastP sequence comparisons (Altschul *et al*., 1990) with and without a low-complexity region filter (SEG: segment sequence(s) by local complexity) against the following organisms: Arabidopsis itself, as well as *Oryza sativa*, *Homo sapiens*, *Caenorhabditis elegans* and *Drosophila melanogaster*. A maximum of 20 BlastP results per inter-organism comparison was retained with a default value of 5 to capture essential information. Lastly, information on whole genome-level transcription (Yamada *et al*., 2003) was included.

The datasets referenced in Table 1 are retrieved by FTP (File Transfer Protocol) on a server of the SIB (Swiss Institute of Bioinformatics). We wrote customized Perl scripts to automatically extract and assemble information relevant to *Arabidopsis* from this source into five files corresponding to the five *Arabidopsis* chromosomes. These files are indexed for rapid data retrieval. The resulting database that underlies Centrifuge is formatted in a SWISS-PROT-like format where each line is flagged with a four or five-letter code and each record begins with the AGI code and ends with a '----' terminator. Our *Arabidopsis* database itself (compiled as shown in Table 1) is updated a minimum of three times per year to keep pace with improved annotation at each of the different component resources (UNIPROT, InterPro, etc.). A second series of customized Perl scripts were used to allow a rapid search of gene annotations and predicted protein characteristics from the database. The user

[*]To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oupjournals.org

Table 1. Web resources used to compile the Arabidopsis database underlying Centrifuge

Type	Tool		Web addresses
Identifier	RefSeq (NCBI)		http://www.ncbi.nlm.nih.gov/RefSeq/
Contig			
Title			
Notes			
Gene Ontology data	Gene Ontology		http://www.geneontology.com/
Isoelectric point	TAIR		http://www.arabidopsis.org/
Molecular weight			
Cellular localization		TargetP	http://www.cbs.dtu.dk/services/TargetP/
Cellular localization scores			
Transmembrane domains		TMHMM	http://www.cbs.dtu.dk/services/TMHMM/
UniProt ID	UNIPROT		http://expasy.uniprot.org/
UniProt keywords			
UniProt title			
Medline and Pubmed ID			
Authors			
Abstract			
Journal references			
InterPro domains ID from UniProt features	InterPro		http://www.ebi.ac.uk/interpro/
InterPro domains title			
InterPro graphic html code			
InterPro ID			
Domain names			
Abstract			
Record terminator			

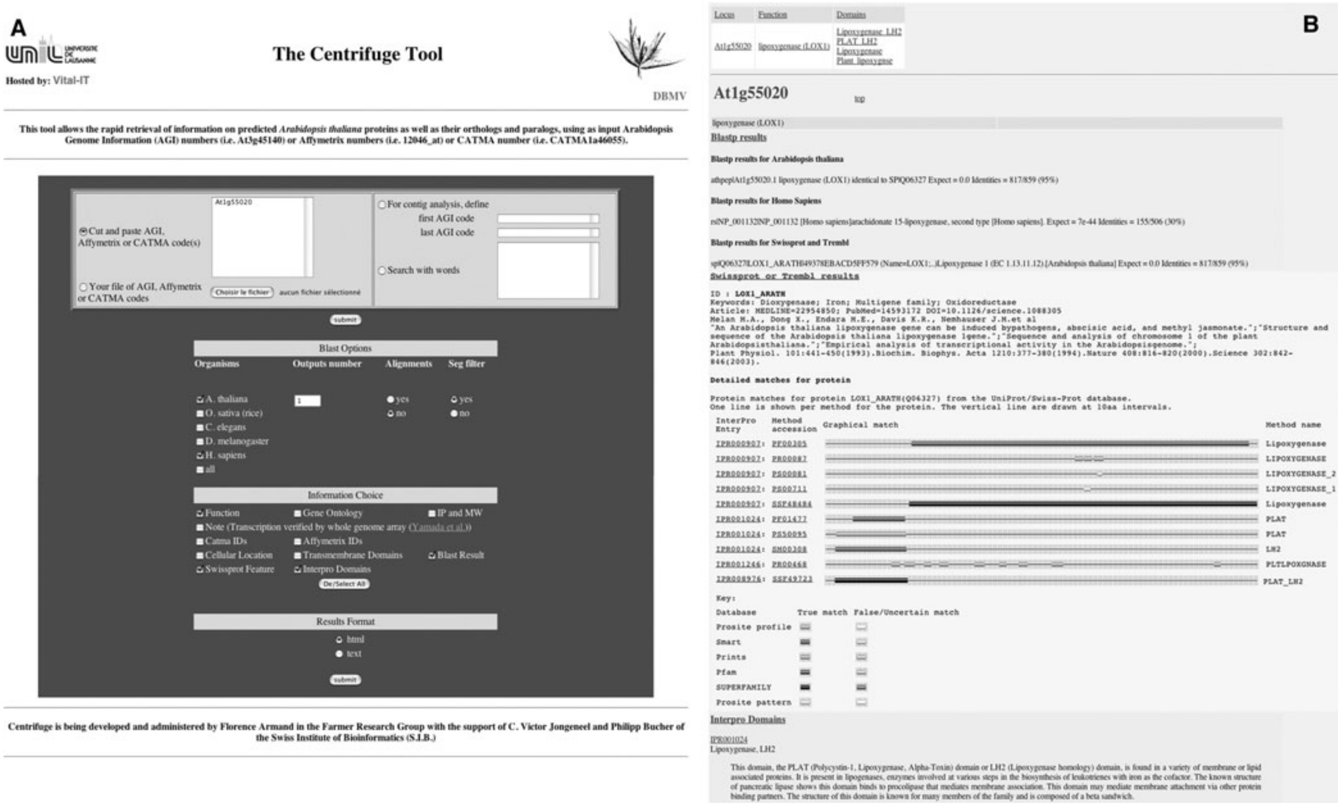
can retrieve all or part of this information for single genes, list of genes or contigs. These Perl scripts take into account the user's choices to parse the database via the index and display the output information on one page, either in an html format or in a tab-delimited text file. Simple word searches are an additional feature of the tool. Again, Perl was chosen as a query language for its text-processing capabilities. This allows the retrieval of selected words from all fields of individual gene records in Centrifuge. This feature has a low selectivity that we consider advantageous. For example if the word 'lipoxygenase' is employed, annotations containing 'lipoxygenase-like' or 'lipoxygenase-binding' will be recovered. On the output page, and for each individual gene, direct links to Massively Parallel Signature Sequencing (MPSS, http://mpss.udel.edu/at/) and *Arabidopsis thaliana* Genome Expression (AtGE, http://blast1.salk.edu/cgi-bin/AtGE) and the Nottingham Arabidopsis Stock Centre Arrays (NASCArrays) resource (http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl) web pages can be found. This allows easy access to several useful and growing current resources describing gene expression in *Arabidopsis*.

FUNCTIONALITY AND EXAMPLES OF USE

AGI (Arabidopsis Genome Initiative, 2000) codes (e.g. At3g45140), as well as Affymetrix (http://www.affymetrix.com/) and CATMA (Crowe *et al*., 2003, http://www.catma.org/) identifiers are accepted for input. Single genes or lists of genes from tab-delimited text files can be entered. Figure 1A shows the homepage into which we have added the single gene At1g55020 that encodes a lipoxygenase. On the input (home) page we selected several display options. The upper part of the output page for this example is shown in Figure 1B. In this example orthologs for the *Arabidopsis* gene chosen are found in humans. The protein's predicted domain structure is given, together with annotations for each individual domain. Being able to simultaneously display Uniprot features, Interpro domain descriptions and BlastP results against different organisms is unique and is a useful feature of the tool.

Using the word search capacity of Centrifuge with, for example, the terms 'endosperm', 'anther' or 'xylem' yields lists of gene products amongst which useful tissue-specific markers could be identified. 'Anther' currently gives 54 hits. The preponderance of genes predicted to encode lipase-like proteins, lipases and lipid transfer proteins indicates an active lipid metabolism and suggests that the tissue might be an attractive hunting ground for new lipid signals.

Centrifuge is most useful for the analysis of genomic regions (contigs) as well as gene lists from expression studies, e.g. microarrays. These types of experiments typically identify tens or hundreds of genes which meet the statistical criteria allowing them to be labeled as induced or repressed. Retrieved data are displayed in a summary table indicating the gene identification number, the predicted or known function, the predicted subcellular location, number of predicted transmembrane helices, and the predicted protein domains. Clicking on the column headings sorts gene products sharing the feature in question, e.g. chloroplast location. Clicking on any feature in the table itself brings up the full report for the gene product in question. We analyzed data from Table S2 in Reymond *et al*. (2004) to reveal that a large proportion (nearly 30%) of insect-inducible genes are



The Centrifuge Tool

This tool allows the rapid retrieval of information on predicted *Arabidopsis thaliana* proteins as well as their orthologs and paralogs, using as input Arabidopsis Genome Information (AGI) numbers (i.e. At3g45140) or Affymetrix numbers (i.e. 12046_at) or CATMA number (i.e. CATMA1a46055).

◉ Cut and paste AGI, Affymetrix or CATMA code(s): At1g55020

○ Your file of AGI, Affymetrix or CATMA codes

○ For contig analysis, define
first AGI code
last AGI code

○ Search with words

Blast Options

Organisms	Outputs number	Alignments	Seg filter
☑ A. thaliana	1	● yes	◉ yes
□ O. sativa (rice)		◉ no	● no
□ C. elegans			
□ D. melanogaster			
☑ H. sapiens			
□ all			

Information Choice

☑ Function □ Gene Ontology □ IP and MW
□ Note (Transcription verified by whole genome array (Yamada et al.))
□ Catma IDs □ Affymetrix IDs
□ Cellular Location □ Transmembrane Domains ☑ Blast Result
☑ Swissprot Feature ☑ Interpro Domains

Results Format

◉ html
● text

Centrifuge is being developed and administered by Florence Armand in the Farmer Research Group with the support of C. Victor Jongeneel and Philipp Bucher of the Swiss Institute of Bioinformatics (S.I.B.)

Locus	Function	Domains
At1g55020	lipoxygenase (LOX1)	Lipoxygenase LH2 PLAT LH2 Lipoxygenase Plant lipoxygnse

At1g55020

lipoxygenase (LOX1)

Blastp results

Blastp results for Arabidopsis thaliana

athpep|At1g55020.1 lipoxygenase (LOX1) identical to SP|Q06327 Expect = 0.0 Identities = 817/859 (95%)

Blastp results for Homo Sapiens

rs|NP_001132|NP_001132 [Homo sapiens]arachidonate 15-lipoxygenase, second type [Homo sapiens]. Expect = 7e-44 Identities = 155/506 (30%)

Blastp results for Swissprot and Trembl

sp|Q06327|LOX1_ARATH|49378EBACD5FF579 (Name=LOX1;..)Lipoxygenase 1 (EC 1.13.11.12).[Arabidopsis thaliana] Expect = 0.0 Identities = 817/859 (95%)

```
Swissprot or Trembl results

ID : LOX1_ARATH
Keywords: Dioxygenase; Iron; Multigene family; Oxidoreductase
Article: MEDLINE=22954850; PubMed=14593172 DOI=10.1126/science.1088305
Melan M.A., Dong X., Endara M.E., Davis K.R., Nemhauser J.M.et al
"An Arabidopsis thaliana lipoxygenase gene can be induced bypathogens, abscisic acid, and methyl jasmonate.";"Structure and
sequence of the Arabidopsis thaliana lipoxygenase 1gene.";"Sequence and analysis of chromosome 1 of the plant
Arabidopsisthaliana.";"Empirical analysis of transcriptional activity in the Arabidopsisgenome.";
Plant Physiol. 101:441-450(1993).Biochim. Biophys. Acta 1210:377-380(1994).Nature 408:816-820(2000).Science 302:842-
846(2003).

Detailed matches for protein

Protein matches for protein LOX1_ARATH(Q06327) from the UniProt/Swiss-Prot database.
One line is shown per method for the protein. The vertical line are drawn at 10aa intervals.
 InterPro    Method
 Entry       accession  Graphical match                                   Method name
 IPR000907:  PF00305                                                      Lipoxygenase
 IPR000907:  PR00087                                                      LIPOXYGENASE
 IPR000907:  PS00081                                                      LIPOXYGENASE_2
 IPR000907:  PS00711                                                      LIPOXYGENASE_1
 IPR000907:  SSF48484                                                     Lipoxygenase
 IPR001024:  PF01477                                                      PLAT
 IPR001024:  PS50095                                                      PLAT
 IPR001024:  SM00308                                                      LH2
 IPR001246:  PR00468                                                      PLTLPOXGNASE
 IPR008976:  SSF49723                                                    PLAT_LH2

 Key:
 Database          True match  False/Uncertain match
 Prosite profile
 Smart
 Prints
 Pfam
 SUPERFAMILY
 Prosite pattern
```

Interpro Domains

IPR001024
Lipoxygenase, LH2

This domain, the PLAT (Polycystin-1, Lipoxygenase, Alpha-Toxin) domain or LH2 (Lipoxygenase homology) domain, is found in a variety of membrane or lipid associated proteins. It is present in lipogenases, enzymes involved at various steps in the biosynthesis of leukotrienes with iron as the cofactor. The known structure of pancreatic lipase shows this domain binds to procolipase that mediates membrane association. This domain may mediate membrane attachment via other protein binding partners. The structure of this domain is known for many members of the family and is composed of a beta sandwich.

Fig. 1. (**A**) Homepage for Centrifuge. (**B**) Output data requested for the *LIPOXYGENASE 1* gene.

predicted to encode proteins that enter either the secretory pathway (20%) or the chloroplast compartment (15%). Furthermore, 44% of insect-inducible genes in the same dataset have related sequences in humans.

In summary, the main advantages of Centrifuge are its one page entry format, its rapidity and its flexibility, all of which greatly facilitate the simultaneous analysis of multiple genes. The scrutiny of long gene lists is accelerated with Centrifuge, allowing a higher throughput exploration of the datastream. With rapidly improving annotations, the tool will become even more powerful in the near future.

ACKNOWLEDGEMENTS

We thank Robin Liechti and Philippe Reymond for helpful discussion. This work was supported by the Swiss National Science Foundation and the University of Lausanne.

REFERENCES

Altschul,S.F. *et al*. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

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

Crowe,M.L. *et al*. (2003) CATMA—A complete *Arabidopsis* GST database. *Nucleic Acids Res.*, **31**, 156–158.

Gutiérrez,R.A. *et al*. (2004) Phylogenetic profiling of the *Arabidopsis thaliana* proteome: what proteins distinguish plants from other organisms? *Genome Biol.*, **5**, R53.

Reymond,P. *et al*. (2004) A conserved transcript pattern in response to a specialist and a generalist herbivore. *Plant Cell*, **16**, 3132–3147.

Rhee,S.Y. *et al*. (2003) The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community. *Nucleic Acids Res.*, **31**, 224–228.

Yamada,K. *et al*. (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. *Science*, **302**, 842–846.