HeliaGene, a bioinformatics portal for *Helianthus* sp. genomics

Sébastien Carrere, Jérôme Gouzy, Nicolas Langlade, Pascal Gamas, Patrick Vincourt
Laboratoire Interactions Plantes Micro-organismes UMR441/2594,
INRA/CNRS, F-31320 Castanet Tolosan, France, E-mail: Patrick.Vincourt@toulouse.inra.fr

ABSTRACT

A bioinformatics portal, called HeliaGene (http://www.heliagene.org) has been developed for in-depth analyses of *Helianthus sp.* EST data. This portal uses the same approach as that already developed for *Medicago truncatula* (MENS database, http://medicago.toulouse.inra.fr/MENS), and provides a variety of pre-computed analyses and tools for EST clusters and for exploring gene function and protein families in a user-friendly fashion. The prediction of EST cluster-encoded peptides is supported by FrameD, a program originally developed for prokaryotic gene prediction. Analyses at the protein level, such as signature and domain searches, can be helpful to make predictions about gene function and to annotate EST clusters. The HeliaGene portal provides interactive access to the annotation of tens of thousands of clusters and their corresponding peptides. Generic workflows for similarities searches versus plant databases or protein family phylogenies are provided as well as specific workflows, like the detection of potential SNPs, based on the between and within *Helianthus* species sequence polymorphisms. In the future, HeliaGene will integrate more tools and results, including genetic maps, characterization of genetic resources and core collections, and integration of sequence-based expression data with transcriptomics experiment results.

Key words: annotation – bioinformatics.

INTRODUCTION

Due to its global adaptation to a wide range of southern European water-scarce environments as well as to the introduction by breeding of a quality trait now being required for biofuel production (“high oleic” type), the sunflower crop *Helianthus annuus* is able to occupy an increasing place among the environmentally safe crops devoted to the production of raw material for the “first generation” biofuels. *Helianthus annuus* is not a model plant, and less genomic resources than for other agronomic crops like corn have been developed. But particularly thanks to an important effort by U.S. laboratories (Compositae Genome Project, http://compgenomics.ucdavis.edu/) but also by French teams (Genoplante program, https://gpi.versailles.inra.fr/), a relatively large number of *Helianthus* sp. ESTs are available in the public databases (284,251 at NCBI by January 18th, 2008). Besides *Helianthus annuus*, six other *Helianthus* species have been used to produce these ESTs, which are derived from a variety of cDNA libraries, providing information on gene transcription in a number of developmental and physiological contexts: various organs at different developmental stages (buds, roots, stems, leaves, seeds…), responses to abiotic stresses, and interaction with various pathogens, etc.

Using the EST-clusters consensus dataset generated by Mike Barker (http://msbarker.com/data.htm), which is the current reference set of sequences on which is based the design of the first generation of sunflower chips, we have developed a user-friendly portal, “HeliaGene”, which provides a variety of pre-existing or specifically developed tools and pre-computed searches to conduct in-depth analyses at different levels. The navigation system provided makes it possible (i) to rapidly visualize EST cluster characteristics, (ii) to explore gene function, (iii) to analyse gene and protein families, (iv) to detect potential SNPs, based on the between and within *Helianthus* species sequence polymorphisms.

Whilst graphical representations are provided for immediate access to analysis summaries, raw results, as well as a number of links, are also provided to conduct in-depth searches whenever necessary. This important feature enables the HeliaGene user to examine the validity of annotations that have been automatically entered for thousands of EST clusters, and to propose a different annotation wherever judged appropriate.

The scope of this paper is to present the HeliaGene navigation system.

MATERIALS AND METHODS

Implementation

The web server has been developed with PERL/CGI and is run on a linux cluster. Sequence data and corresponding annotation sheets are indexed using a lucene-based search engine to allow complex queries.

Similarity searches and automatic annotation

Sequence comparisons against the protein databases UniProt (Apweiler et al., 2004), ProDom (Bru et al., 2005) and HuSep2007 were performed using NCBI-BLASTX and NCBI-BLASTP release 2.2.13 (Altschul et al., 1997) with default parameters, except for the penalty values to create a gap (-G) (set to 9 instead of 11) or to extend a gap (-E) (set to 2 instead of 1), and the threshold for the expected value (-e) (set to 0.1 instead of 10). InterproScan (Quevillon et al., 2005) software has been executed with default parameters on the peptide database in order to identify InterPro (Mulder et al., 2007) domains and families. Then, raw results have been analysed to generate, whenever possible, a synthetic description of the peptide function based on InterPro domain content.

RESULTS

General organization of EST data mining system

The system is organized around two databases, corresponding to EST cluster DNA sequences and predicted protein sequences, respectively. The set of 87,237 EST-clusters was primarily generated by Mike Barker at http://msbarker.com/data.htm from a total set of 284,251 ESTs available on GenBank (9 Sep 2007), most of them having been produced in the frame of the Compositae Genome Project (http://compgenomics.ucdavis.edu/) and by Genoplante (https://gpi.versailles.inra.fr/), and additional sequences being provided by Steve Knapp lab. The results of various analyses, conducted both at the DNA and protein sequence levels, are provided and can be used to annotate EST clusters and corresponding gene products (see below).

The system can be entered in a variety of ways: via queries based upon annotations, keywords (using a lucene-based retrieval system) as well as similarity searches. Results are presented with links allowing for an easy navigation through different sources of information.

EST cluster analysis at the DNA level

An overview of the various types of information provided for EST cluster analysis is shown in Fig. 1. A general control panel gives access to a synthetic summary of similarity results, to the predicted peptide annotation sheet and to several workflows enabling the execution of more complex pipelines on the current sequence.

The cluster annotation is found below the control panel. Summaries of WU-blastn searches (Gish, W. (1996-2002) http://blast.wustl.edu) using HuSep2007 consensus sequences against other HuSep2007 clusters and of NCBI-blastx searches against the UniProt protein database are then shown, with links to complete raw results and database entries.

Prediction of EST coding regions

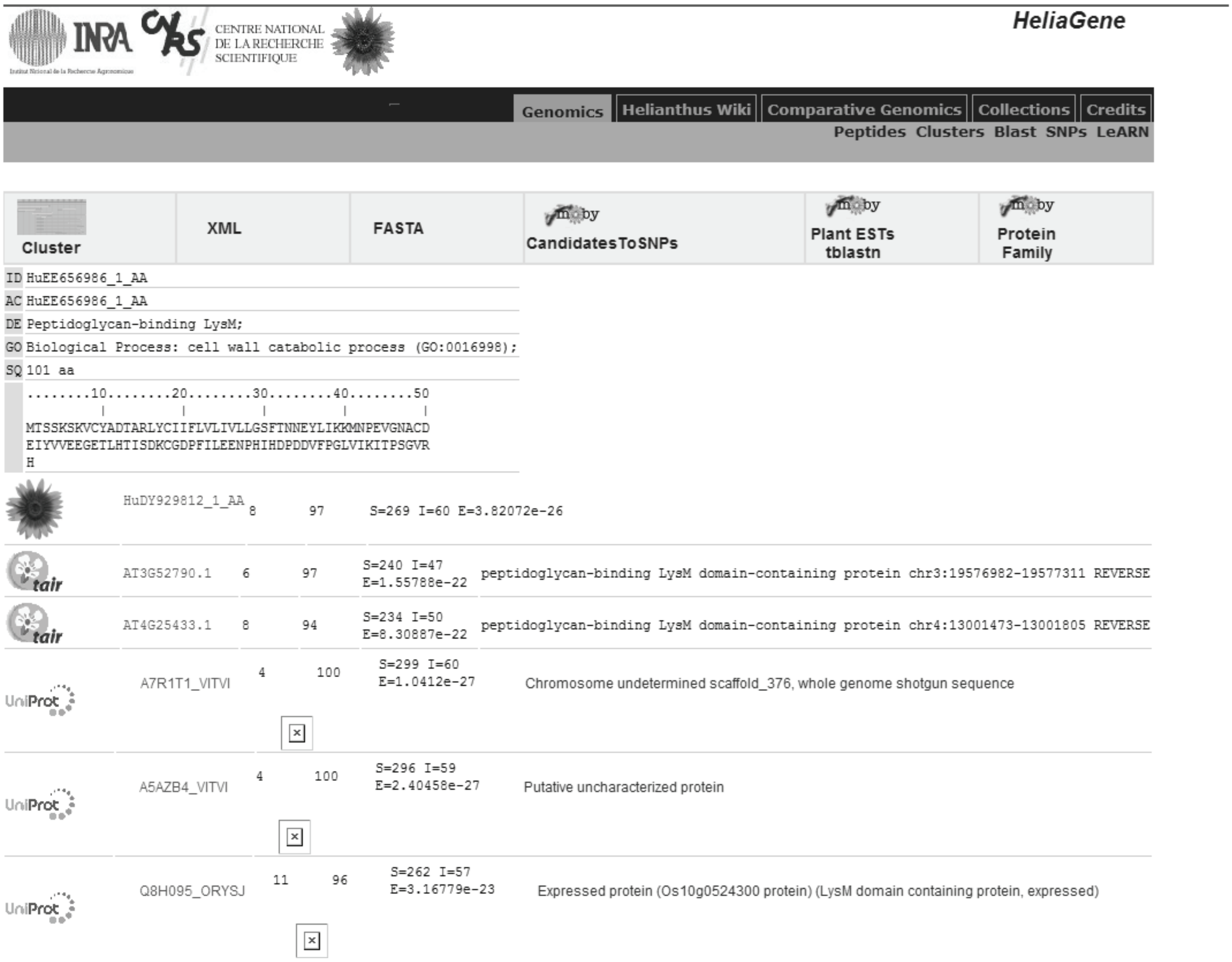
The starting point for the prediction of coding regions is the FrameD program originally designed for prokaryotic sequences (Schiex et al., 2003). Prediction of coding regions from eukaryotic transcripts is somewhat similar to prokaryotic gene prediction, but additional difficulties arise from the fact that (i) EST clusters are of different sizes, with depth from 1 to almost 100 for a given nucleotide, and, consequently, of a variable robustness in the consensus cDNA sequence prediction. To manage this heterogeneity in consensus quality, FrameD was repeatedly applied to each cluster using similarity information and several combinations of parameters aiming at handling different frameshift sensitivities. By this means it was possible to predict a protein sequence for 83% of the assembled clusters, which corresponds to 72,372 peptides from 87,237 EST-clusters (Table 2). Prediction failures were mainly due either to a too short coding fragment (threshold 29 aa) or to the absence of a parameter set fitting the sequence.

Table 1. Summary of the peptide predictions

Total	72,372
predicted full-length CDS	24,799
N-term fragment (translation start only)	14,504
C-term fragment (translation stop only)	23,145
Fragment (start and stop are missing)	9,924
Number of frameshifts detected/corrected	24,053
Min-Max peptide length	29 aa-1,466 aa
Mean/Median peptide length	181 aa-155 aa

Protein sequence analyses

Protein prediction allows searches of structural or functional domains and motifs to be conducted (Fig. 1), which can be particularly informative when trying to decipher gene function. Queries of InterPro (Integrated Resource of Protein Domains and Functional Sites) were carried out to look for protein domains and amino acid signatures. Information about possible subcellular location and overall protein structure are provided with results from SignalP (Bendtsen et al., 2004) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/).



ID	HuEE656986_1_AA
AC	HuEE656986_1_AA
DE	Peptidoglycan-binding LysM;
GO	Biological Process: cell wall catabolic process (GO:0016998);
SQ	101 aa

```
........10........20........30........40........50
         |         |         |         |         |
MTSSKSKVCYADTARLYCIIFLVLIVLLGSFTNNEYLIKKMNPEVGNACD
EIYVVEEGETLHTISDKCGDPFILEENPHIHDPDDVFPGLVIKITPSGVR
H
```

	HuDY929812_1_AA	8	97	S=269 I=60 E=3.82072e-26	
tair	AT3G52790.1	6	97	S=240 I=47 E=1.55788e-22	peptidoglycan-binding LysM domain-containing protein chr3:19576982-19577311 REVERSE
tair	AT4G25433.1	8	94	S=234 I=50 E=8.30887e-22	peptidoglycan-binding LysM domain-containing protein chr4:13001473-13001805 REVERSE
UniProt	A7R1T1_VITVI	4	100	S=299 I=60 E=1.0412e-27	Chromosome undetermined scaffold_376, whole genome shotgun sequence
UniProt	A5AZB4_VITVI	4	100	S=296 I=59 E=2.40458e-27	Putative uncharacterized protein
UniProt	Q8H095_ORYSJ	11	96	S=262 I=57 E=3.16779e-23	Expressed protein (Os10g0524300 protein) (LysM domain containing protein, expressed)

Fig. 1. Typical annotation sheet providing a synthetic view of the functional annotation, and a summary of the similarities with access both to raw results (database icon) and to database entries (hypertext link).

Remora Workflows

The user interface provides access to several analysis pipelines based on Remora, which is a workflow manager (Carrere and Gouzy, 2006) able to create and run workflows based on BioMoby web-services. From the protein annotation sheet, the system provides the user with three Remora workflows:

- for searching SNPs "CandidatesToSNPs": as ESTs have been produced and made available on seven different species (*H. annuus, H. petiolaris H. argophyllus, H. paradoxus, H. exilis, H. tuberosus, H. ciliaris*), which is quite unusual in public databases, HeliaGene proposes a workflow starting from an amino-acid sequence, for example of candidate genes with a proven function in a model crop like *Arabidopsis*, to exploit the between and within species sequence polymorphism of ESTs to try to detect potential positions of SNPs.
- for identifying similar EST on other plants "Plant ESTs tblastn"
- and for aligning the current protein with other members of the same family "Protein Family".

From the Cluster annotation sheet an additional workflow is provided permitting the identification and the alignment of plant proteins belonging to the same family ("Family analysis").

Functional annotation overview

Table 2 lists the 25 top domains according to their representation in the predicted *Helianthus* protein sequences. When compared to the representation of the same domains in *Arabidopsis thaliana* (Uniprot database) some discrepancies appear: IPR009072 (Histone-fold, dominant role in regulating transcription), IPR000425 (involved in plant tonoplast intrinsic proteins), IPR001344 (involved in light-harvesting complex which delivers excitation energy to photosystems I and II) are over-represented. On the contrary, IPR001611 (Leucine-rich repeat:signal transduction, cell adhesion, DNA repair, recombination, transcription, RNA processing, disease resistance, and apoptosis) is clearly under-represented. These discrepancies are probably due to the differences in complexity between *Helianthus* and *Arabidopsis* genomes and/or to the specificities of cDNA libraries. Concerning the lack of representation of the LRR associated ESTs, it has to be mentioned that besides these ESTs, there is about 820 NBS-LLR resistance like fragments in the CoreNucleotide section of the NCBI database.

Table 2. InterPro top 25 entries. In the last column, the ratio between the number of predicted peptides in *Helianthus* having each domain and its occurrence in *Arabidopsis thaliana* proteins is calculated.

InterPro Accession	Number of occurrence	InterPro description	% of Arabidopsis in UniProt
IPR011009	1783	Protein kinase-like	97%
IPR001128	570	Cytochrome P450	130%
IPR001810	553	Cyclin-like F-box	74%
IPR013083	524	Zinc finger, RING/FYVE/PHD-type	62%
IPR000719	497	Protein kinase, core	27%
IPR012336	452	Thioredoxin-like fold	114%
IPR009057	425	Homeodomain-like	52%
IPR012677	339	Nucleotide-binding, alpha-beta plait	64%
IPR011046	334	WD40 repeat-like	72%
IPR009072	327	Histone-fold	234%
IPR002885	275	Pentatricopeptide repeat	42%
IPR001611	266	Leucine-rich repeat	20%
IPR000425	257	Major intrinsic protein	367%
IPR001344	246	Chlorophyll A-B binding protein	390%
IPR002213	244	UDP-glucuronosyl/UDP-glucosyltransferase	118%
IPR015609	243	Molecular chaperone, heat shock protein, Hsp40	116%
IPR002198	236	Short-chain dehydrogenase/reductase SDR	120%
IPR011992	235	EF-Hand type	96%
IPR000608	224	Ubiquitin-conjugating enzyme, E2	202%
IPR001087	223	Lipolytic enzyme, G-D-S-L	114%
IPR001806	221	Ras GTPase	140%
IPR001471	214	Pathogenesis-related transcriptional factor	140%
IPR008972	210	Cupredoxin	154%
IPR011050	207	Pectin lyase fold/virulence factor	77%
IPR003593	192	AAA+ ATPase, core	39%

DISCUSSION

The HeliaGene navigation system should prove useful for carrying out high throughput characterization of a large number of cDNA clones, and for collecting information complementary to expression profiling data. Indeed a major challenge in the coming years will be to determine the function of a vast number of genes by integrating as many sources of information as possible. As a first step towards this objective, we have developed a navigation system which links information derived from raw data, sequence analysis and database searches. Sequence annotation is facilitated by rapid access to various sources of documentation, and thus does not merely rely upon sequence homology detection and automated transfer of pre-existing annotation.

In future, we plan to integrate both expression profiling data from microarray experiments and genetic maps. Finally, tools for comparative genomics will be improved, especially to be able to transfer more information from other well studied model plants (*Arabidopsis thaliana*, *Medicago truncutula*, etc.).

REFERENCES

Altschul, S.F., T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Apweiler, R., A. Bairoch, and C.H. Wu. 2004. Protein sequence databases. Curr. Opin. Chem. Biol. 8:76-80.

Bendtsen, J.D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: signalp 3.0. J. Mol. Biol. 340:783-795.

Bru, C., E. Courcelle, S. Carrère, Y. Beausse, S. Dalmar, and D. Kahn. 2005 The prodom database of protein domain families: more emphasis on 3d. Nucleic Acids Res. 33:D212-D215.

Carrere, S., and J. Gouzy. 2006. Remora: a pilot in the ocean of biomoby web-services. Bioinformatics 22:900-901.

Mulder, N.J., R. Apweiler, T.K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P.S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A.N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J.D. Selengut, C.J. Sigrist, P.D. Thomas, F. Valentin, D. Wilson, C.H. Wu, and C. Yeats. 2007. New developments in the interpro database. Nucleic Acids Res. 35:D224-D228.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler, and R. Lopez. 2005. Interproscan: protein domains identifier. Nucleic Acids Res. 33:W116-W120.

Schiex, T., J. Gouzy, A. Moisan, and Y. de Oliveira. 2003. Framed: a flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. Nucleic Acids Res. 31:3738-3741.