A system for automated genome annotation

Gomes, A. F. M.

October, 2010

Abstract

Eucalyptus globulus is a tree used to produce paper pulp. With the aim to study the genes responsible for wood formation in *E. globulus*, which are not yet fully understood although known to have an impact on the quality of paper pulp, GenEglobwq project was defined.

One of the project goals is the development of an annotation module, to automatically annotate the gene sequences obtained. Therefore, the aim of this thesis was to extend a web information system, GEDI, used to manage the data from GenEglobwq project, in order to provide a tool for automated genome annotation. In this sense the annotation module in the GEDI system was created, which makes use of an automatic genome annotation pipeline, named MAKER, which can be used to annotate any type of organism.

In order to validate the annotation module operation, the chloroplast genomes of *Eucalyptus globulus* and *Eucalyptus grandis* was annotated. However, like any automatic genome annotation pipeline, not all the genes were annotated, being necessary a manual curation of the results.

Key Words: Genome, Annotation, Biomarkers, Chloroplast, *Eucalyptus globulus*.

1 Introduction

Portugal is a major producer of eucalypt (*Eucalyptus globulus*), which provides most of the raw material used to produce paper pulp. The final pulp and paper quality is highly affected by the wood properties. It is accepted that variations of wood properties are regulated by a high number of genes and proteins during xylem differentiation. However the xylogenesis (wood formation) is not yet fully understood. Sequencing, mapping and annotating the *E. globulus* genome will help better understand this process. With that objective in mind, the GenEglobwq [1] project was created. The GenEglobwq project is a collaboration between IBET, RAIZ, IST and INESC-ID and aims to identify and characterize the genomic regions that underlies strong effect QTL (Quantitative Trait Loci) for pulp yield, in *E. globulus*, combining an array of genomic tools and transcriptomic approaches. The main role of KDBIO group in this project is to provide an easy way to collect, manage and process these huge amount of data. Therefore KDBIO group developed a web information system, GEDI - Genomic Data Information System, to fulfill this demand.

1.1 Genome Annotation

In the study of a particular organism, the complete genome sequence provides only partial and raw information. More importantly, scientists need to find out where the genes are, what they do, how they are related, etc. This is where the annotation process intervenes to attach this information to the genome sequence. Genome annotation is thus the process of attaching biological information to the genome sequences and starts by identifying the positions of structural genomic elements, like genes, exons, introns, repeated regions, promoters, etc. This process can be defined as structural annotation. After identifying these elements a secondary annotation to provide biochemical and biological function information to these elements is necessary and this process is called functional annotation [2].

The quality of the annotation is very important for future experiments. If an annotation is correct, then experiments, such as, RNAi, PCR, gene expression arrays, targeted gene knockout, or ChIP [2], that need information from these annotation, are greatly facilitated. The quality of the annotation, on its turn, depends, among other things, on the annotation pipeline that is employed, some more detailed and accurate than others, and also on the skills and experience of those operating the pipeline. It is also important for a correct annotation to have a curation and review process with an expert on the biology of each genome.

2 Annotation Module

In order to extend the GEDI system to allow its integration with a tool that provide for automated genome annotation, it was created the annotation module. For that it was necessary to decide which automatic annotation tool should be used in the GEDI system. The software should fulfill certain requirements, such as, to be open source, to be preferably written in Java, to be well documented and to present good results for any type of organism.

During the process of choosing the software, several tools were tested. Special attention was given to Blast2GO [3], DAS [4] and MAKER [5]. The first options were Blast2GO and DAS, mainly because they are written in Java, which would facilitate integration. DAS is not really a annotation tool but a client-server system to exchange biological annotation. In case we would like to annotate a new genomic sequence not in any DAS server, this system would be useless. Blast2GO is a good tool for functional annotation but we wanted a tool for structural annotation. Another option would be to use gene predictors, but they are usually specific to one type of prokaryotic or eukaryotic organism. Therefore, MAKER became a good choice because it is a structural annotation pipeline that can be used to annotate equally prokaryotic and eukaryotic organisms, and can also be used to re-annotate a genome sequence as well as to annotate a new one. The only drawback of this tool is that it is written in Perl. However, during the development of the module, this has proven to be a minor issue.

MAKER is a structural annotation pipeline, being one of the Generic Model Organism Database (GMOD) [6] components. It was developed to allow researchers to easily annotate eukaryotic and prokaryotic genome sequences and to create genome databases. This system makes use of existing software tools that can produce *ab initio* gene predictions, align ESTs and proteins to a genome and identify and mask repeat elements, combining their output and produce what it believes to be the best possible annotation [2], and also produce the evidences alignments that support those annotations.

2.1 Use Case and Pipeline

A *use case* [7] is a means of specifying required usages of a system. It consists of a subject, upon which the action unfolds; an actor, which is a user or any other system that interacts with the subject; and a specification of the required behavior of the subject in a scenario of interaction with the actor. For the annotation module, two use cases were defined. The diagrams presenting in Figure 1 and Figure 2 indicate the use cases concerning the *annotation run manager* and the *curation manager*, respectively.

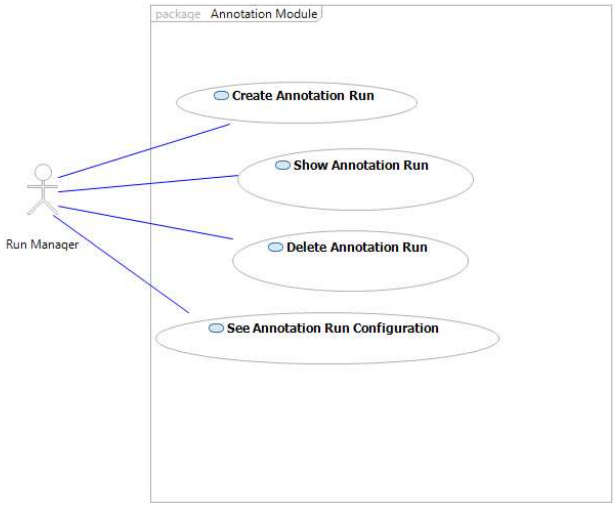


Figure 1: Run management use case diagram.

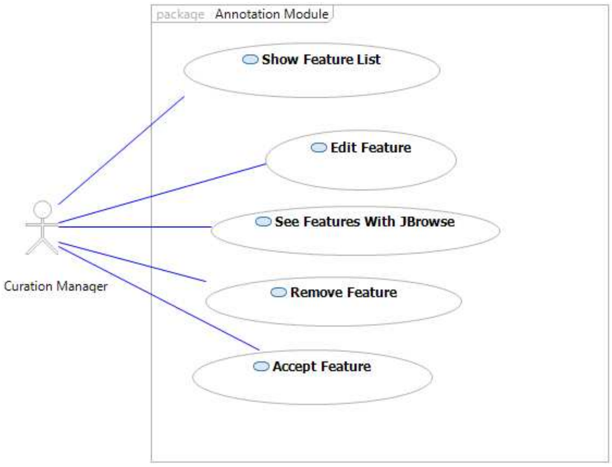


Figure 2: Curation management use case diagram.

Concerning the management of the annotation procedures (Figure 1), we need to distinguish the following interaction scenarios:

- Show Annotation Run: In this scenario the user wants to see all the annotation runs that have been executed, along with their status. The system displays all the runs, one per line.
- Create Annotation Run: In this scenario the user wants to create an annotation run. The system displays a form to be filled with basic information on the annotation run. After fulfilled the system can create an annotation run.
- See Annotation Run Configuration: In this scenario the user wants to see configuration of an existing annotation run. The system displays the annotation run configuration.
- Delete Annotation Run: In this scenario the user wants to remove an annotation run. The system removes the annotation run after confirmation.

As for the annotation curation, we need to distinguish the following interaction scenarios:

- Show Feature List: In this scenario the user wants to see all the features of an annotation run. The system displays the list of features.
- Edit Feature: In this scenario the user wants to edit a existing feature. The system displays a form with the feature data to be edited. After edited the system updates the feature.

- Remove Feature: In this scenario the user wants to remove a feature. The system removes the feature.
- Accept Feature: In this scenario the user wants to validate a feature. The system associates effectively the feature to the respective sequence.
- See Features with JBrowse: In this scenario the user wants to see the features with the visualization tool JBrowse. The system displays the features on JBrowse.

Taking into consideration the use cases above, an annotation pipeline was created for the GEDI system, which is represented in Figure 3. This pipeline represents a recipe of what a user has to do to annotate a sequence. First, the user needs to set up the annotation configuration, next it needs to choose the sequences that he wishes to annotate and then the MAKER software is executed. After the annotation run is finished the user proceeds to the curation process.

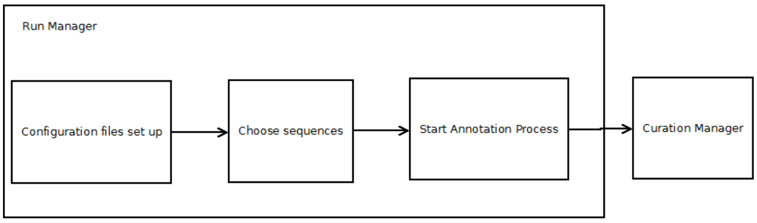


Figure 3: GEDI annotation pipeline.

The implementation of the use cases described above required the development of three web interfaces: an interface to visualize the list of annotation runs, an interface to create a annotation run, and another to visualize the annotation run results. These interfaces will be described in more detail in the next sections.

2.2 Annotation Runs List

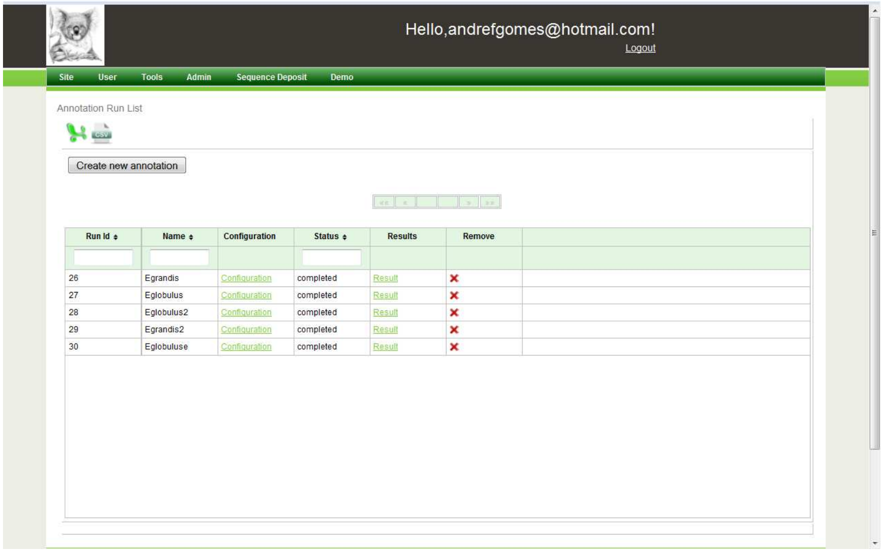
In this section, the “Annotation Runs List” web page is described (Figure 4). This page is used to visualize the list of annotation runs.

This page is composed by a table with a list of annotation runs and a button on top of the table to create a new annotation. The table is composed by the columns: “Run Id”, with the unique identifier of the run; “Run Name”, with a human-readable name for the run; “Configuration”, with a link to the configuration page of the annotation run; “Status”, with an indication of the progress status of the run, which can be *completed*, *in progress* or *failed*; “Results”, with a link to the results page, in the case of completed annotation runs; and finally, “Remove”, where the user can mark runs to be deleted.

When an annotation run is launched, its status is *in progress*, and so there is not a link for the results in the “Results” column. After being completed, the “Status” column needs to be changed. In order to do that, the system checks the progress and automatically refreshes the page every minute.

2.3 Annotation Run Configuration

In this section, the “Annotation Run Configuration” web page is described . This page serves to create an annotation run or see an annotation run configuration. In order to visualize this page, the user needs to



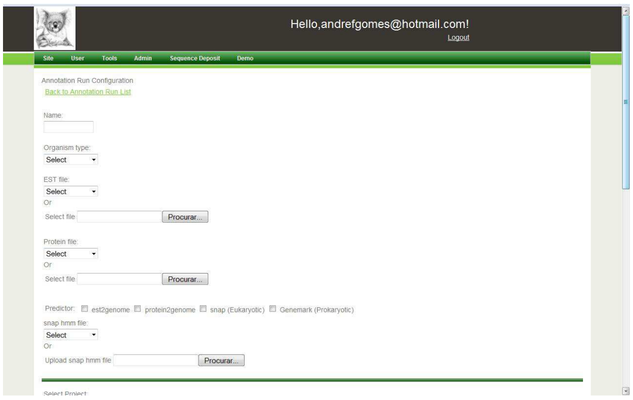
Annotation Run List

Create new annotation

Run Id	Name	Configuration	Status	Results	Remove
26	Egrandis	Configuration	completed	Result	✗
27	Eglobulus	Configuration	completed	Result	✗
28	Eglobulus2	Configuration	completed	Result	✗
29	Egrandis2	Configuration	completed	Result	✗
30	Eglobuluse	Configuration	completed	Result	✗

Figure 4: Annotation Run List Page.

press the “create new annotation” button in the “Annotation Run List” page or to press the “Configuration” link in the table from the “Annotation Run List” page.



Annotation Run Configuration
Back to Annotation Run List

Name:

Organism type:
Select

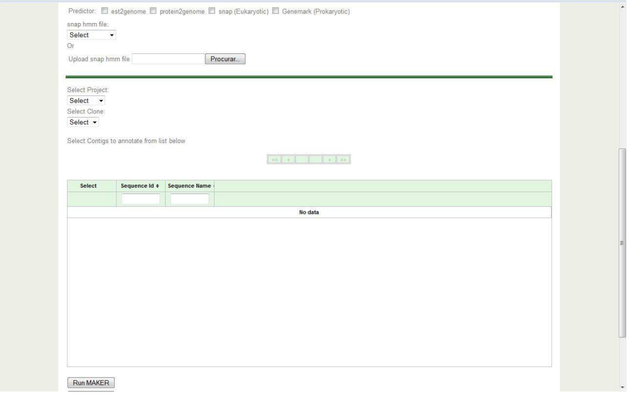
EST file:
Select
Or
Select file Procurar...

Protein file:
Select
Or
Select file Procurar...

Predictor: □ est2genome □ protein2genome □ snap (Eukaryotic) □ Genemark (Prokaryotic)
snap hmm file:
Select
Or
Upload snap hmm file Procurar...

Select Project:

Figure 5: Configuration Annotation Run Page.


Predictor: □ est2genome □ protein2genome □ snap (Eukaryotic) □ Genemark (Prokaryotic)
snap hmm file:
Select
Or
Upload snap hmm file Procurar...

Select Project:
Select
Select Clone:
Select

Select Contigs to annotate from list below

Select	Sequence Id	Sequence Name
No data		

Run MAKER

Figure 6: Configuration Annotation Run Page.

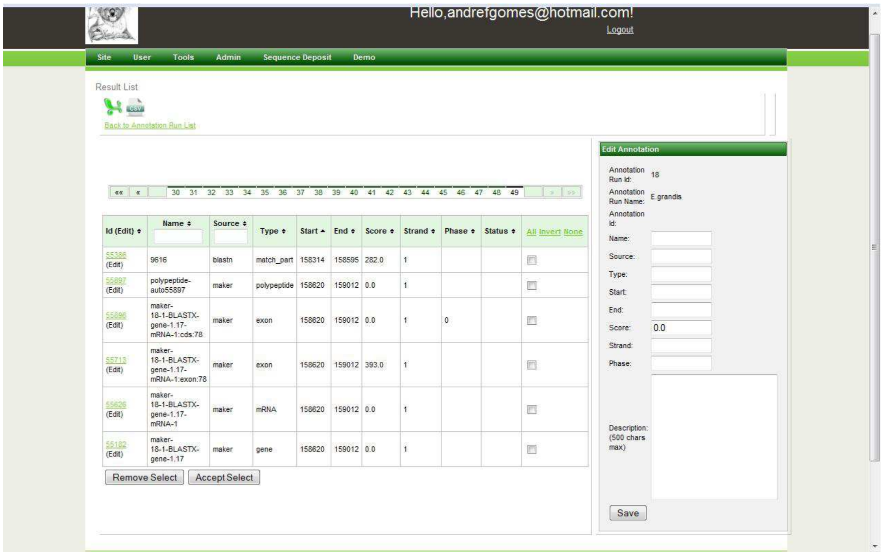
When creating a new annotation run, several fields need to be filled. These fields are: “name”, the name a user wants to give to an annotation run; “organism type”, the type of organism to be annotated, prokaryotic or eukaryotic; “EST file”, an existing or newly uploaded EST file in FASTA format, only necessary if the est2genome option is selected in predictor field; “protein file”, an existing or newly uploaded a protein file in FASTA format, only necessary if the protein2genome option is selected in predictor field; and “predictor”, the methods that can be used to generate the annotations. The annotation methods currently available are the *est2genome* (a MAKER program which allows to EST alignments to become gene annotations), *protein2genome* (a MAKER program which tries to construct gene models directly from protein alignments), *SNAP* [8] (a gene predictor for eukaryotic organisms), or *GeneMark* [9] (a gene predictor to prokaryotic organisms). If the user chooses the SNAP program, there is a field where he can upload a HMM file or select one already added to the database.

At the bottom of this page, Figure 6, the user needs to choose the sequence to be annotated. For that, the user needs to select a project and a clone, to retrieve its sequence.

After completing all the required fields, the user simply needs to press the “run MAKER” button to create an annotation run. If all the fields were correctly filled, then the user is redirected to the “Annotation Run List” page, otherwise an error message will appear.

2.4 Result List

In this section, the “Results List” web page is described. This page serves to visualize the annotation run results. For that, the run needs to be successfully finished and the user needs to press the “Result” link in the annotation runs table from the “Annotation Runs List” page.



Id (Edit)	Name	Source	Type	Start	End	Score	Strand	Phase	Status	All Invert None
55386 (Edit)	9616	blastn	match_part	158314	158595	282.0	1			□
55897 (Edit)	polypeptide-auto55897	maker	polypeptide	158620	159012	0.0	1			□
55896 (Edit)	maker-18-1-BLASTX-gene-1.17-mRNA-1:cds:78	maker	exon	158620	159012	0.0	1	0		□
55713 (Edit)	maker-18-1-BLASTX-gene-1.17-mRNA-1:exon:78	maker	exon	158620	159012	393.0	1			□
55626 (Edit)	maker-18-1-BLASTX-gene-1.17-mRNA-1	maker	mRNA	158620	159012	0.0	1			□
55182 (Edit)	maker-18-1-BLASTX-gene-1.17	maker	gene	158620	159012	0.0	1			□

Edit Annotation

Annotation Run Id:	18
Annotation Run Name:	E.grandis
Annotation Id:	
Name:	
Source:	
Type:	
Start:	
End:	
Score:	0.0
Strand:	
Phase:	
Description: (500 chars max)	

Figure 7: Annotation Run Results Page.

This page presents the results of an annotation run in the form of a table and it is where here the curation process begins. In the bottom of the page, there are two buttons, “Accept Selected” button and “Remove Selected”, here a user can save or remove a selected feature, respectively. Saving a feature will effectively associate it to the respective sequence in the database. Still on this page, the user can edit a feature by pressing the “Edit” button in the table and then editing the feature information on the displayed form. In order to save the edited feature, the user needs to press the “save” button on the form.

3 Results

3.1 *Eucalyptus globulus*

In this chapter, we report on the results of our case study, which consisted in using the implemented annotation pipeline on the chloroplast genome of the *Eucalyptus globulus*. To assess the quality of our findings, we compare them to those obtained with the GeneMark annotation suite [9].

The chloroplast genome of *E. globulus* was released in 2005. It is a genome with 160 286 bp with a GC-content of 36.9%. It has 135 genes documented, coding for 45 structural RNAs, with 8 of them containing one intron, what give a total of 53 exons; and 90 genes coding for proteins, with 11 of them containing one intron and 4 containing two introns, what give a total of 108 exons (one exon is shared by two genes). The genome is available on GenBank with accession number AY780259.

3.1.1 Pipeline Configuration

The input provided to the annotation pipeline consisted of the chloroplast genome sequence, retrieved from GenBank, as well as an EST and a protein file. The EST file consisted of 90 CDS, 37 tRNA and 8 rRNA of the *E. globulus*. It was obtained from the Chloroplast Genome Database [10]. The protein file consisted of 90 proteins from the *E. globulus*, also obtained from the Chloroplast Genome Database.

The parameters for the algorithm, mainly for MAKER, were set to their default values, except for those modifiable through the "Annotation Run Configuration" web page, namely the predictor and organism type parameters. For the predictor parameter, the "est2genome" and "protein2genome" options were used. For the organism type, we used the "prokaryotic" option.

The "prokaryotic" option was chosen as organism type because plastids, in particular chloroplasts, are commonly thought to have prokaryotic origin. Indeed, according to the endosymbiotic theory, these organelles evolved from a prokaryotic cell that was originally ingested as food by the eukaryotic cell, and somehow these prokaryotic cells were not digested and became part of their hosts. Since they have prokaryotic origin, their structure remains similar to the prokaryotic cells; they have their own DNA, which is circular and with a single strain, they don't have a nucleus and they have their own population of ribosomes similar to the prokaryotic ribosomes [11]. However, with evolution, the chloroplast gene expression machinery changed, and it became distinct from prokaryotic, eukaryotic and phage [12]. Nevertheless, it still more similar to the prokaryotic gene machinery, hence choice of prokaryotic organism type.

The "est2genome" and "protein2genome" option were chosen because, since we supply a ESTs and proteins file from *E. globulus* and we wanted these programs generate gene annotations from the ESTs and proteins alignments.

The GeneMark suite used in our evaluation was run with a heuristic model retrieved from the software web page corresponding to the GC-content of the chloroplast genome.

3.1.2 Measures of performance

The metrics used to analyze the annotation pipeline performance were the sensitivity and the specificity [13]. The sensitivity is given by

$$SN = \frac{TE}{AE}, \tag{1}$$

and specificity is defines as

$$SP = \frac{TE}{PE}, \tag{2}$$

where AE is the number of annotated exons that are documented, PE is the number of predicted exons by the gene predictor, and TE is the number of true exons, i.e. the number of predicted exons that are correctly annotated, meaning that both boundaries of the predicted exon match with an annotated exon.

3.1.3 Results

The results were divided into two main categories, “RNA” prediction and “Protein” prediction. This was done because heuristic models, for the GeneMark, do not include parameters for structural RNAs binding sites.

The results from the performance of the annotation pipeline, as well as GeneMark are represented in Table 1.

Table 1:
Performance of the annotation pipeline (AP) and GeneMark on the chloroplast genome of *E. globulus*.

	AE	PE	TE	PCE	OE	NDE	FE	SN	SP
AP RNAs	53	37	35	2	0	16	0	0.66	0.95
AP Protein	108	89	59	30	0	19	0	0.55	0.66
GeneMark Protein	108	115	36	33	10	26	36	0.33	0.31
AP	161	89	59	30	0	72	0	0.37	0.66
AP with evidence	161	136	103	33	0	25	0	0.64	0.76

AE stands for annotated exons, PE stands for predicted exons and TE stands for true exons. Partially corrected exons (PCE) is the number of predicted exons with only one boundary matching that of an annotated exon. Overlap exon (OE) is the number of predicted exons with no boundary exactly matching that of an annotated exon, but still having some overlap. Non-detected exons (NDE) is the number of annotated exons that were not annotated by any predicted exon. False exons (FE) is the number of predicted exons that don’t overlap any annotated exon. SN stands for sensitivity and SP stands for specificity.

The first line, corresponds to the genes coding for structural RNAs. Our annotation pipeline did not identify any gene coding for structural RNAs, the results obtained were from the evidence alignments that MAKER produces. So for the RNAs results we considered a evidence alignment, as a PE, predicted exon.

By analyzing the first line, we notice that 16 exons were not detected, which made the SN value low. These exons were all from the 8 genes containing one intron and they were not detected because it was considered the input sequence from a prokaryotic organism, therefore it didn’t take splice sites into account.

The second line, corresponds to the genes coding for proteins. Here, the exons predicted by the annotation pipeline are considered as PE. We notice that it didn’t identify 19 exons, with 9 of them being from genes with one intron or more. Once again, the main reason for the annotation pipeline being unable to identify these exons was because prokaryotic organisms don’t have introns. Regarding the other 10 exons, 9 of them had some information from evidence alignments. The reason why it didn’t annotate these exons is possibly because it gives a low score to these evidences. If some filtering parameters were changed, it could have detected these exons. The low values of SN and SP are mainly because of the high number of partially correct exons that weren’t correctly annotated.

The third line is relative to the results obtained with GeneMark. If we compare these results to those of the second line, we can easily see that MAKER outperforms GeneMark, with a sensitivity of 0.55 against 0.33 and a specificity of 0.66 against 0.31, suggesting that our annotation pipeline, using MAKER, is indeed a better tool to annotate the chloroplast genome of *E. globulus*.

The fourth line considers the RNAs as undetected exons, which decreases, as expected, the sensitivity from 0.55 to 0.37.

The fifth line considers the evidence alignments as PE. This was done to demonstrate the results after a fast curation of the results, which raise the sensitivity value of 0.37 to 0.64 and the specificity value of 0.66 to 0.76.

3.2 *Eucalyptus grandis*

A further test of the annotation pipeline was done with the chloroplast genome of the *Eucalyptus grandis*. In particular, we were interested in assessing the impact of annotating a genome with ESTs and proteins from a different species, albeit from the same genus.

The chloroplast genome of *E. grandis* was sequenced and annotated by Paiva and colleagues [14] and available from the GenBank database under the accession id HM347959. It is a genome with 160 142 bp with a GC-content of 36.9%. This genome has 138 genes coding for 50 structural RNAs and 88 for proteins, with 8 of them containing one intron and 4 containing two introns, what gives a total of 103 exons (an exon is shared by two genes).

For this test it was used the same parameter values from the test before.

3.2.1 Results

The obtained results are presented in Table 2. If we compare them to the results from Table 1, we verify that there is not much difference between them, except for the RNAs results where the sensitivity and specificity is better for the *E. globulus*. The main reason why the results are similar, even using ESTs and proteins from other species, is because the species belong to the same *genus*, what makes their genomes very similar.

Table 2: Performance of the annotation pipeline on the chloroplast genome of *E. Grandis*.

	AE	PE	TE	PCE	OE	NDE	FE	SN	SP
AP RNAs	50	37	31	6	0	13	0	0.62	0.84
AP Protein	103	84	57	25	2	19	4	0.55	0.68
AP	153	84	57	25	2	69	4	0.37	0.68
AP with evidence	153	133	99	31	3	32	4	0.65	0.74

4 Conclusions and Future Work

This thesis consisted of extending a web information system, GEDI, so that an automatic annotation tool could be integrated in the system. The automatic annotation tool chosen was MAKER.

According to the literature, the performance of MAKER on eukaryotic organisms was proven to be almost at the same level as that of other gene finders, being Augustus [15] the one that presented better results, which is good because Augustus is one of the optional components of MAKER [16].

It was also our intention to test even further the performance of MAKER through the annotation pipeline, and to prove its versatility by testing it on the chloroplast genome of *E. globulus* and *E. grandis*. The results showed that out annotation pipeline obtained better results compared to GeneMark, a gene finder for all prokaryotic organisms, suggesting it is a better annotation tool. Regarding the performance on the chloroplast genome of *E. grandis*, the results were similar to those obtained with the chloroplast of *E. globulus*, showing that ESTs and protein sequences from sufficiently close species can be used without reducing significantly the performance.

According to the literature and the results obtained in this thesis, it can be concluded that MAKER is indeed a good option to be integrated into the annotation module, since it can yield good annotation results for prokaryotic organisms and it can be used as well to annotate any other type of organism. In addition, the ease to run and configure the tool, reinforces its choice. One setback is that it requires EST and protein database from the organism being annotated or from a related organism for more complete and sensible results. Moreover it is unable to identify genes coding for structural RNAs as genes.

With the choice of MAKER as an automatic annotation tool, it was possible to create a simple but effective annotation pipeline for the annotation module, making it a easy way to annotate the stored sequences in the GEDI system.

Since an automatic genome annotation is not 100% successful a manual curation of the results is required. The annotation module provides a initial process of curation by allowing the user to remove or accept the annotations results.

As future work, Augustus can be integrated into the annotation module in order to have better results for eukaryotic organisms. Other aspect that can possibly be extended, is the list of parameters made available for modification from the GEDI user interface. Since there were some missed exons in our case studies which could have been annotated if some parameters such as the filtering parameters were properly set. Nevertheless, the parameters that are currently available are sufficient to annotate a sequence.

An extra feature that can be implemented is to send a notification email to the user whenever an annotation run is finished. For a visualization tool of annotations it was installed JBrowse which still in underdevelopment. An alternative visualization tool, that can be installed, for the administration interface, could be Apollo, which also allows the edition of annotation.

Finally, other complementary annotation tools can be integrated into the GEDI system, with the aim of providing more accurate and more complete information. For example, Blast2GO, can be integrated to provide functional annotation to the sequences, and DAS, to share and collate genomic annotation information.

References

[1] “Geneglobwq - scanning for candidate genes underlying a pulp yield qtl in eucalyptus globulus.” [Online]. Available: http://geneglob.inesc-id.pt/

[2] GMOD, “Maker tutorial,” Online Website, GMOD, August 2010. [Online]. Available: http://gmod.org/wiki/MAKER_Tutorial

[3] Blast2GO, “Blast2go,” Online Website, Bioinformatics and Genomics Department, August 2010. [Online]. Available: http://blast2go.bioinfo.cipf.es/

[4] R. Dowell, R. Jokerst, A. Day, S. Eddy, and L. Stein, “The distributed annotation system,” *BMC Bioinformatics*, vol. 2, no. 1, pp. 7+, 2001. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-2-7

[5] B. L. Cantarel, I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado, and M. Yandell, “Maker: An easy-to-use annotation pipeline designed for emerging model organism genomes,” *Genome Research*, vol. 18, no. 1, pp. 188–196, January 2008. [Online]. Available: http://dx.doi.org/10.1101/gr.6743907

[6] “Generic model organism database (gmod).” [Online]. Available: www.gmod.org

[7] O. M. Group, “Omg unified modeling language (omg uml), superstructure, v2.1.2,” Tech. Rep., November 2007. [Online]. Available: http://www.omg.org/spec/UML/2.1.2/Superstructure/PDF

[8] I. Korf, “Gene finding in novel genomes,” *BMC Bioinformatics*, vol. 5, no. 1, pp. 59+, May 2004. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-5-59

[9] A. V. Lukashin and M. Borodovsky, “Genemark.hmm: new solutions for gene finding,” *Nucl. Acids Res.*, vol. 26, no. 4, pp. 1107–1115, February 1998. [Online]. Available: http://nar.oxfordjournals.org/cgi/content/abstract/26/4/1107

[10] L. Cui, N. Veeraraghavan, A. Richter, K. Wall, R. K. Jansen, J. Leebens-Mack, I. Makalowska, and C. W. dePamphilis, “Chloroplastdb: the chloroplast genome database.” *Nucleic Acids Res*, vol. 34, no. Database issue, January 2006. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/16381961

[11] U. Vothknecht and J. Soll, “Chloroplast membrane transport: interplay of prokaryotic and eukaryotic traits.”

[12] M. G. Bausher, N. D. Singh, S.-B. Lee, R. K. Jansen, and H. Daniell, “The complete chloroplast genome sequence of citrus sinensis (l.) osbeck var ’ridge pineapple’: organization and phylogenetic relationships to other angiosperms,” *BMC Plant Biology*, vol. 6, pp. 21+, September 2006. [Online]. Available: http://dx.doi.org/10.1186/1471-2229-6-21

[13] K. Knapp and Y.-P. P. P. Chen, “An evaluation of contemporary hidden markov model genefinders with a predicted exon taxonomy.” *Nucleic acids research*, vol. 35, no. 1, pp. 317–324, January 2007. [Online]. Available: http://dx.doi.org/10.1093/nar/gkl1026

[14] P. E. V. S. S. D. S.-C. H. B. S. F. P. G. D. S. X. A. J. K. D. W. R. F. A. B. H. Paiva, J.A.P. and J. Grima-Pettenati, “Advancing eucalyptus genomics: identification and sequencing of lignin biosynthesis genes from deep-coverage bac libraries,” *Unpublished*, 2010.

[15] M. Stanke and S. Waack, “Gene prediction with a hidden markov model and a new intron submodel.” *Bioinformatics (Oxford, England)*, vol. 19 Suppl 2, October 2003. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/14534192

[16] A. Coghlan, T. Fiedler, S. McKay, P. Flicek, T. Harris, D. Blasiar, the nGASP Consortium, and L. Stein, “ngasp - the nematode genome annotation assessment project,” *BMC Bioinformatics*, vol. 9, no. 1, pp. 549+, December 2008. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-549