Finding the Core-Genes of Chloroplasts based on Gene Similarity Approaches

Bassam Al Kindy[1,3], Christophe Guyeux[1,3], Jean-François Couchot[1,3], Michel Salomon[1,3] and Jacques M. Bahi[1,3]

[1]FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department, Université de Franche-Comté, France

February 20, 2014

Abstract

Investigating in the evolution of genomes become a hard task due to the amount of evolutionary techniques and the amount of genomes that raises every day. The important question to understand here is: how can we clusterize large amounts of chloroplast species?, and what are the common genes that play a role in the process of evolution among these species?. Clusterizing collection of species aims to find the common genes that share the same functionality properties. In other words, clustering helps us to find the core and pan genome among species that share a common properties, such us gene name, gene sequence, family, ..., etc. According to other studies, finding such core and/or pan genome is not an easy task due to a large amount of computation, and requiring a rigorous methodology. To achieve this goal, a collection of 99 chloroplasts are considered in this article. Two methodologies will be investigated, respectively based on sequence similarities and from annotation tools. The obtained results will finally be evaluated in terms of performances and biological relevance.

Keywords: Core genome, Methodology, Pan genome, Genes prediction, Coding sequences clustering, Chloroplasts, Gene quality test.

1 Introduction

The idea behind the importance of identifying core genes is to understand the shared functionality of agiven set of species. We introduced in previous work (see [6]) two methods for discovering core and pan genes based on sequence similarity method and alignment based approache method. However, to determine both core and pan genomes of a large set of DNA sequences, we consider in this work compare the same clustering algorithm of sequence similarity method

proposed in previous work with new method as an improvement of alignment based approach by considering sequence quality control test. More precisely, we focus on the following questions using a collection of 99 chloroplasts as illustrative example: how can we identify the best core genome (that is, an artificially designed set of coding sequences as close as possible to the real biological one) and how to deduce scenarios regarding their genes loss.

The existance of Chloroplasts is behind the fact that chloroplasts found in Eucaryotes have an endosymbiotic origin, meaning that they come from the incorporation of a photosynthetic bacteria (Cyanobacteria) within an eucaryotic cell. They are fundamental key elements in living organisms history, as they are organelles responsible for photosynthesis. This latter is the main way to produce organic matters from mineral ones using solar energy. Consequently photosynthetic organisms are at the basis of most ecosystem trophic chains. Indeed photosynthesis in eucaryotes has allowed a great speciation in the lineage, leading to a great biodiversity. From an ecological point of view, photosynthetic organisms are at the origin of the presence of dioxygen in the atmosphere (allowing extant life) and are the main source of mid to long term carbon storage, which is fundamental regarding current climate changes. However, the chloroplasts evolutionary history is not totally well understood, at least large scale speaking, and their phylogeny requires to be further investigated.

A key idea in phylogenetic classification is that a given DNA mutation shared by at least two taxa has a larger probability to be inherited from a common ancestor than to have occurred independently. Thus shared changes in genomes allow to build relationships between species. In the case of chloroplasts, an important category of genomes changes is the loss of functional genes, either because they become ineffective or due to a transfer to the nucleus. Thereby a small number of gene losses among species indicates that these species are close to each other and belong to a similar lineage, while a large loss means distant lineages. Phylogenies of photosynthetic plants are important to assess the origin of chloroplasts and the modes of gene loss among lineages. These phylogenies are usually done using a few chloroplastic genes, some of them being not conserved in all the taxa. This is why selecting core genes may be of interest for a new investigation of photosynthetic plants phylogeny. However, the circumscription of the core chloroplast genomes for a given set of photosynthetic organisms needs bioinformatics investigations using sequence annotation and comparison tools, and various choices are available.

Our intention in this research work regarding the methodology in core and pan genomes determination is to investigate the impact of these choices. on the results. A general presentation of the approaches detailed in this document is provided in the next section. Then we will study in Section 3.1 the use of annotated genomes from NCBI website [9] with a coding sequences clustering method based on the Needleman-Wunsch similarity scores [15]. While the second method will be proposed in Section 3.2.1, which intends to use gene name and sequence comparisons. Information regarding computation time and memory usage are provided in Section 4. Finally, a discussion based on biological aspects regarding the evolutionary history of the considered genomes will final-

ize our investigations, leading to our methodology proposal for core and pan genomes discovery of chloroplasts This research work ends by a conclusion section, in which our investigations will be summarized and intended future work will be planned.

2 An Overview

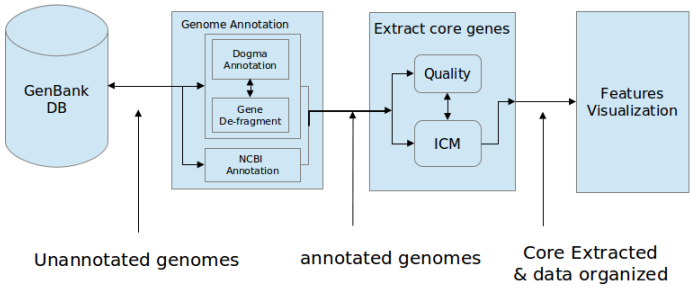


Figure 1: A general overview of the annotation-based approach

In previous work [6], we proposed a pipeline for the extraction of core genome. In this work, the pipline is considered with quality test method in extracting core genes, for more details (see figure 1). As a starting point, an annotation uses a DNA sequences database such as NCBI's GenBank [9], the European *EMBL* database [3], or the Japanese *DDBJ* one [19]. Further more, It is possible to obtain annotated genomes (DNA coding sequences with gene names and locations) by interacting with these databases, either by directly downloading annotated genomes delivered by these websites, or by launching an annotation tool on complete downloaded genomes. Obviously, this annotation stage must be of quality if we want to obtain acceptable core and pan genomes.

Using such annotated genomes, we will detail two general approaches for extracting the core genome, which is the third stage of the pipeline: the first one uses similarities computed on predicted coding sequences, while the second one uses all the information provided during the annotation stage.

instead of considering only gene sequences taken from NCBI or DOGMA, a quality test process is take place by working with gene names and sequences to produce quality genes. However, we will show that such a simple idea is not so easy to realize, and that it is not sufficient to only consider gene names provided by such tools while it gives good results in previous work [6]. Annotation, which is the first stage, is an important task for extracting gene features. Indeed, to extract good gene feature, a good annotation tool is obviously required. Indeed, such annotations can be used in various manners (based on gene names, gene sequences, protein sequences, etc.) to extract the core and pan genomes. We will subsequently propose methods that use gene names and sequences for extracting

core genes and producing chloroplast evolutionary tree.

The final stage of our pipeline, only invoked in this article, is to take advantage of the information produced during the core and pan genomes search. This features visualization stage encompasses phylogenetic tree construction (see [6] for more details) using core genes, genes content evolution illustrated by core trees, functionality investigations, and so on.

For illustration purposes, we have considered 99 genomes of chloroplasts downloaded from GenBank database [9]. These genomes lie in the eleven type of chloroplast families (see [6] for more details).Furthermore, two kinds of annotations will be considered in this document, namely the ones provided by NCBI on the one hand, and the ones by DOGMA on the other hand.

3 Core genes extraction

3.1 Similarity-based approach

We still need to propose good methodology to predict good core genes that reflect natural biological relationships among species. Proposing new methods and compare them with previous ones can give us an indicator of which method can produce good functional genes. In this section, we will recall the definition with a fast revision of similarity based method (see [6] for further details). This method, considers annotated genomes from NCBI and DOGMA and uses a distance-based similarity measure on genes' coding sequences. Such an approach requires annotated genomes, like the ones provided by the NCBI website.

3.1.1 Theoretical presentation

We start by fast revision of first method by the following preliminary definition [1, 6].

Definition 1 Let $A = \{A, T, C, G\}$ be the nucleotides alphabet, and A^* be the set of finite words on A (*i.e.*, of DNA sequences). Let $d : A^* \times A^* \to [0, 1]$ be a function called similarity measure on A^*. Consider a given value $T \in [0, 1]$ called a threshold. For all $x, y \in A^*$, we will say that $x \sim_{d,T} y$ if $d(x, y) \leqslant T$.

Let be given a *similarity* threshold T and a *similarity measure* d. The method begins by building an undirected graph between all the DNA sequences g of the set of genomes as follows: there is an edge between g_i and g_j if $g_i \sim_{d,T} g_j$ is established. This graph is further denoted as the "similarity" graph. We thus say that two coding sequences g_i, g_j are equivalent with respect to the relation $\mathcal{R}$ if both g_i and g_j belong in the same connected component (CC) of this similarity graph, *i.e.*, if there is a path between g_i and g_j in the graph.

It is not hard to see that this relation is an equivalence relation whereas $\sim$ is not. Any class for this relation is called a "gene" in this article, where its representatives (DNA sequences) are the "alleles" of this gene, such abuse of language being proposed to set our ideas down. Thus this first method produces

for each genome G, which is a set $\{g_1^G, ..., g_{m_G}^G\}$ of m_G DNA coding sequences, the projection of each sequence according to π, where π maps each sequence into its gene (class) according to $\mathcal{R}$. In other words, a genome G is mapped into $\{\pi(g_1^G), ..., \pi(g_{m_G}^G)\}$. Note that a projected genome has no duplicated gene since it is a set.

Consequently, the core genome (resp., the pan genome) of two genomes G_1 and G_2 is defined as the intersection (resp., as the union) of their projected genomes. We finally consider the intersection of all the projected genomes, which is the set of all the genes $\dot{x}$ such that each genome has at least one allele in $\dot{x}$. This set will constitute the core genome of the whole species under consideration. The pan genome is computed similarly as the union of all the projected genomes.

	Method 1				Method 2	
	NCBI		DOGMA		NCBI and DOGMA	
Threshold	core	pan	core	pan	core	pan
50	1	163	1	118	**5**	**245**
51	1	291	1	194	-	-
52	1	412	1	258	-	-
53	1	508	1	321	-	-
54	4	617	2	372	-	-
55	**5**	**692**	2	409	-	-
56	**5**	**761**	**3**	**445**	-	-
57	4	832	**3**	**459**	-	-
58	4	905	2	477	-	-
59	4	976	2	497	-	-
60	2	1032	2	519	4	242
61	2	1113	2	553	-	-
62	2	1186	2	580	-	-
63	2	1264	2	607	-	-
64	2	1352	2	644	-	-
65	1	1454	2	685	-	-
66	1	1544	1	756	-	-
67	0	1652	1	838	-	-
68	0	1775	1	912	-	-
69	0	1886	1	1007	-	-
70	0	2000	1	1116	3	242
80	0	3541	0	2730	1	242
90	0	5703	0	5181	0	241

Table 1: Size of core and pan genomes w.r.t. the similarity threshold, first and second approache.

3.1.2 Case study

Let us now consider the 99 chloroplastic genomes introduced earlier. We will use in this case study either the coding sequences downloaded from NCBI website or the sequences predicted by DOGMA. DOGMA, which stands for *Dual Organellar GenoMe Annotator*, has already been evoked in this article. This is a tool developed in 2004 at University of Texas for annotating plant chloroplast and animal mitochondrial genomes. This tool translates a genome in all six reading frames and then queries its own amino acid sequence database using Blast (blastx [2]) with various ad hoc parameters. The choice of DOGMA is natural, as this annotation tool is reputed and specific to chloroplasts.

Each genome is thus constituted by a list of coding sequences. In this illustration study, we have evaluated the similarity between two sequences by using a global alignment. More precisely, the measure d introduced above is the similarity score provided after a Needleman-Wunch global alignment, as obtained by running the *needle* command from the *emboss* package released by EMBL [15]. Parameters of the *needle* command are the default ones: 10.0 for gap open penalty and 0.5 for gap extension.

The number of genes in the core genome and in the pan genome, according to this first method using data and measure described above have been computed using the supercomputer facilities of the Mésocentre de calcul de Franche-Comté. Obtained results are represented in Table 1 with respect to various threshold values on Needleman-Wunsch similarity scores. Remark that when the threshold is large, we obtain more connected components, but with small sizes (a large number of genes, with a few numbers of alleles for each of them). In other words, when the threshold is large, the pan genome is large too. No matter the chosen annotation tool, this first approach suffers from producing too small core genomes, for any chosen similarity threshold, compared to what is usually expected by biologists. For NCBI, it is certainly due to a wrong determination of start and stop codons in some annotated genomes, due to a large variety of annotation tools used during genomes submission on the NCBI server, some of them being old or deficient: such truncated genes will not produce a large similarity score with their orthologous genes present in other genomes. The case of DOGMA is more difficult to explain as, according to our experiments and to the state of the art, this gene prediction tool produces normally good results in average. The best explanation of such an under-performance is that a few genomes are very specific and far from the remainder ones, in terms of gene contents, which leads to a small number of genes in the global core genome. However this first approach cannot help us to determine which genomes must be removed from our set of data. To do so, we need to introduce a second approach based on gene names: from the problematic gene names, we will be able to trace back to the problematic genomes.

3.2 Annotation based approach

3.2.1 Quality-test approach

Genome: NC_008114.1.fasta Threshold= 40
Genes in NCBI: 103 Genes in Dogma: 81 Common Genes: 52 NCBI: 50.49% Dogma: 64.20%

No	Gene	Len_NC	Len_Do	N.Start	N.Stop	D.Start	D.Stop	Score	Comments
1.	RPS11	393	389	ATG	TAA	TGG	GTT	98.982188	
2.	RPS14	303	299	ATG	TAA	ATG	TTG	98.679868	
3.	ACCD	939	908	ATG	TAA	ATG	TCC	96.698616	
4.	RPS19	279	275	ATG	TAA	TGT	CGT	98.566308	
5.	RPL5	543	536	ATG	TAA	TGA	AAA	98.710866	
6.	RPL2	837	824	ATG	TAG	TGG	AAA	98.446834	
7.	RPL20	345	338	ATG	TAA	ATG	GGT	97.971014	
8.	RPS7	471	467	ATG	TAA	ATG	TTT	99.150743	
9.	RBCL	1428	1426	ATG	TAA	TGG	TTA	99.510490	
10.	PETD	483	479	ATG	TAA	ATG	ATT	99.171843	
11.	PETG	114	107	ATG	TAA	TGG	AAT	93.859649	
12.	PETA	990	1001	ATG	TAA	GTT	TTT	98.306773	
13.	PETB	648	644	ATG	TAA	ATG	TCT	99.382716	
14.	RPOC1	4737	923	ATG	TAA	GTC	ATC	19.484906	Ignored, < Threshhold
15.	YCF4	549	545	ATG	TAA	AAT	TTT	96.057348	
16.	YCF3	504	485	ATG	TAA	TGC	AAG	96.230159	
17.	RPOB	6537	794	ATG	TAA	TAA	CTT	12.146244	Ignored, < Threshhold
18.	ATPI	744	710	ATG	TAA	AAT	TCA	95.430108	
19.	ATPH	249	239	ATG	TAA	ATG	ATT	95.983936	
20.	CLPP	597	578	ATG	TAA	ATG	AGC	96.817420	
21.	ATPB	1446	1436	ATG	TAA	TAT	TTA	99.308437	
22.	ATPA	1509	1471	ATG	TAA	AGA	TCA	55.349099	
23.	ATPE	393	401	ATG	TAA	TGA	ATA	97.512438	
24.	PSBE	252	245	ATG	TAA	TGG	ATT	97.222222	
25.	PSBD	1059	1036	ATG	TAA	TGA	GTA	97.828140	
26.	PSBF	129	119	ATG	TAA	CAA	CGT	92.248062	
27.	PSBA	1062	1045	ATG	TAA	TGA	GGC	87.533393	
28.	PSAJ	126	122	ATG	TAA	ATG	TTT	96.825397	
29.	PSBC	1386	1381	GTG	TAA	ACG	GAT	99.351585	
30.	PSBM	105	101	ATG	TAA	ATG	AGA	96.190476	

Figure 2: Part of the implementation of the second method, compaire the common genes from NCBI and DOGMA.

The second approach in this paper is an enhasement of the ICM*Intersection Core Matrix* proposed in [6] by considering gene names to find core genome. Based on gene names spelling, When they realizing simple homogenization of names provided by NCBI, they miss core genes which have slightly different name formats. To enlarge the size of the core genome, to be as close as possible to the true natural one, we propose to integrate a similarity distance on gene names. Each similarity will be computed between a name from DOGMA, which operates as a reference here, and a name from NCBI as shown in figure 2.

The proposed distance is the Levenshtein one, which is close to the Needleman-Wunsch, except that gap opening and extension penalties are equal. The same name is then set to sequences whose NCBI names are close according to this edit distance.

The risk, by doing so, is to merge genes that are different but whose names are similar (for instance, ND4 and ND4L are two different mitochondrial genes but with similar names). The solution is thus to compare, in a second stage, the similarity of DNA sequences too (with a Needleman-Wunsch global alignment), and to simply ignore the gene if this similarity is below a given threshold.

By doing so, the second approach is designed, which takes the fundamental idea contained in the annotation-based approaches in the previous work. Remark that this approach is simply a deeper processing of the naming stage in the second approach in [6], the other stages being identical.

The DNA similarity computation raises another problem in the case of DOGMA: contrary to what happens with gene features in NCBI, genes predicted by DOGMA may be fragmented in several parts. Such genes are signaled in the GeneVision file produced by DOGMA, as each fragment is in this file and with the same gene name. A gene whose name is present at least twice in the file is thus either a duplicated gene or a fragmented one. Obviously, fragmented genes must be defragmented before the DNA similarity computation stage (remark that such a defragmentation has already been realized on NCBI website). As the orientation of each fragment is given in the GeneVision output, this defragmentation consists in concatenating all the possible permutations, and only keeping the permutation with the best similarity score to other sequences having the same gene name, if this score is larger than the given threshold.

To put it in a nutshell, the genomes list of gene names are firstly updated in this third approach, following the process detailed in Algorithm 2, while Algorithm 1 outlines the *geneChk* subroutine. These updated genomes are secondly sent to Algorithm [6], which will produce the desired core genomes, see Figure **??** for an updated pipeline.

Algorithm 1 Maximum similarity score between two sequences(geneChk)

Require: $g1, g2 \leftarrow$ NCBI gene sequence, DOGMA gene sequence
Ensure: Maximum similarity score
$score1 \leftarrow needle(g1, g2)$
$score2 \leftarrow needle(g1, Reverse(Complement(g2)))$
return $max(score1, score2)$

Algorithm 2 Extract new genome based on gene quality test

```
Require: Gname ← Genome Name, Threshold ← 60, RNGenes ← [ ], RDGenes ← [ ], PNGenes ← [ ], PDGenes ← [ ]
Ensure: geneList ← Quality genes
  for gene in NCBI genes of Gname do
    if gene in RNGenes then
      dir(NCBI_Genes) ← savePermutation(gene)
      PNGenes ← gene
    else
      RNGenes ← gene
    end if
  end for
  for gene in Dogma genes of Gname do
    if gene in RDGenes then
      dir(Dogma_Genes) ← savePermutation(gene)
      PDGenes ← gene
    else
      RDGenes ← gene
    end if
  end for
  geneList = empty list
  common = set(dir(NCBI_Genes)) ∩ set(dir(Dogma_Genes))
  for gene in common do
    scores ← [ ]
    if gene NOT in PNGenes AND gene NOT in PDGenes then
      ...
      scores ← geneChk(g1, g2)
    else if gene in PNGenes AND NOT gene in PDGenes then
      PGene ← loadPermutations('N', gene) ...
      for X in PGene do
        ...
        scores ← geneChk(g1, g2)
      end for
    else if gene in PDGenesANDgene NOT in PNGenes then
      PGene ← loadPermutations('D', gene)
      ...
      for X in PGene do
        ...
        scores ← geneChk(g1, g2)
      end for
    else if gene in PDGenesANDgene in PNGenes then
      for X in loadPermutations('N',gene) do
        for Y in loadPermutations('D',gene) do
          ...
          scores ← geneChk(g1, g2)
        end for
      end for
      score ← max(scores)
      if score > Threshold then
        geneList ← gene
      end if
    end if
  end for
  return geneList
```

4 Implementation

All different algorithms have been implemented using Python on a personal computer running Ubuntu 12.04 with 6 GiB memory and a quad-core Intel core i5 processor with an operating frequency of 2.5 GHz.

Table 2: Type of annotation, execution time, and core genes.

Method	Annotation		Features		Exec. time (min.)		Core genes		Bad genomes	
	N	D	Name	Seq	N	D	N	D	N	D
Gene prediction	√	√	-	√	1.7	-	?	-	0	-
Gene Quality	√	√	√	√	≃3 days + 1.29		4		1	

Table 2 presents for each method the annotation type, execution time, and the number of core genes. We use the following notations: **N** denotes NCBI, while **D** means DOGMA, and **Seq** is for sequence. The first two *Annotation*

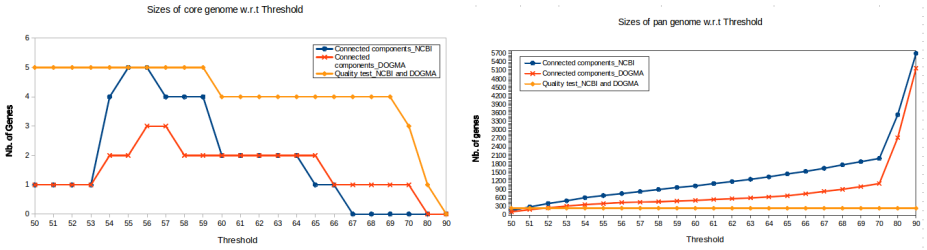
columns represent the algorithm used to annotate chloroplast genomes. The next two ones *Features* columns mean the kind of gene feature used to extract core genes: gene name, gene sequence, or both of them. It can be seen that almost all methods need low *Execution time* expended in minutes to extract core genes from the large set of chloroplast genomes. Only the gene quality method requires several days of computation (about 3-4 days) for sequence comparisons. However, once the quality genomes are well constructed, it only takes 1.29 minutes to extract core gene. Thanks to this low execution times that gave us a privilege to use these methods to extract core genes on a personal computer rather than main frames or parallel computers. The lowest execution time: 1.52 minutes, is obtained with the second method using Dogma annotations. The number of *Core genes* represents the amount of genes in the last core genome. The main goal is to find the maximum core genes that simulate biological background of chloroplasts. With NCBI we have 28 genes for 96 genomes, instead of 10 genes for 97 genomes with Dogma. Unfortunately, the biological distribution of genomes with NCBI in core tree do not reflect good biological perspective, whereas with DOGMA the distribution of genomes is biologically relevant. Some a few genomes maybe destroying core genes due to low number of gene intersection. More precisely, *NC_ 012568.1 Micromonas pusilla* is the only genome who destroyes the core genome with NCBI annotations for both gene features and gene quality methods.

The second important factor is the amount of memory nessecary in each methodology. Table 3 shows the memory usage of each method. In this table, the values are presented in megabyte unit and gV means genevision file format. We can notice that the level of memory which is used is relatively low for all methods and is available on any personal computer. The different values also show that the gene features method based on Dogma annotations has the more reasonable memory usage, except when extracting core sequences. The third method gives the lowest values if we already have the quality genomes, otherwise it will consume far more memory. Moreover, the amount of memory, which is used by the third method also depends on the size of each genome.

Table 3: Memory usages in (MB) for each methodology

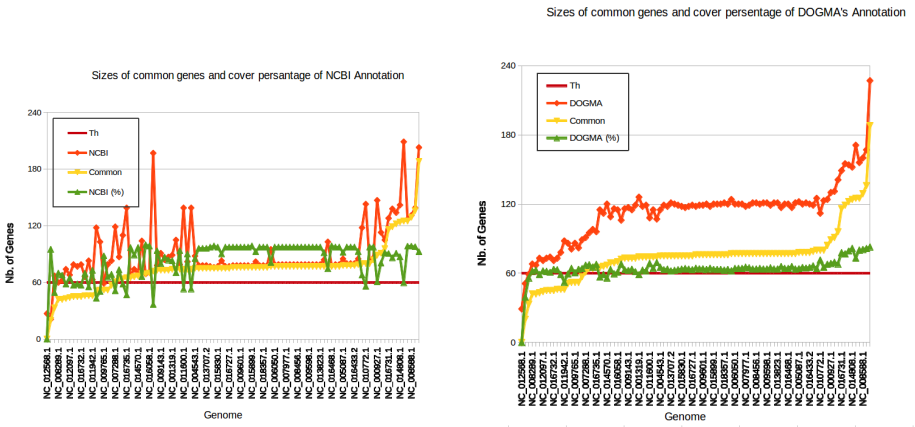
Method		Load Gen.	Conv. gV	Read gV	ICM	Core tree	Core Seq.
Gene prediction	NCBI	108	-	-	-	-	-
Gene Quality		15.3	$\leq$3G	16.1	17	17.1	24.4

Figure 4 represent the sizes of core and pan genomes produced from the two methods. In figure 3a core genes are predicted, note that max core genes do not mean good genes. We are looking for genes that meet it's biological principles. The core genes produced from the first method specially from DOGMA can reflect its biological meaning, we will explain later in the section of disscusion the reason why. In figure 3b, we can see that the values of pan genome from second method is still steady with different thresholds the second method, while in the first method pan genes increases when the threshold increased.



(a) Sizes of core genome (b) Sizes of pan genome

Figure 3: Sizes of Core and Pan genomes for first and second method.



(a) genes coverage of NCBI genomes (b) genes coverage of DOGMA genomes

Figure 4: Gene comparisons cover from NCBI and DOGMA, second method

Furthermore, we calculate the correlation coeffecient formula for the second method and the results shows that the correlation for the annotation from DOGMA was 0.97 while with NCBI was 0.69.

5 Discussion

5.1 Biological evaluation

It is well known that the first plants' endosymbiosis ended in a great diversification of lineages comprising *Red Algae*, *Green Algae*, and *Land Plants* (terrestrial). Several second endosymbioses occurred then: two involving a *Red Algae* and other heterotrophic eucaryotes and giving birth to both *Brown Algae* and *Dinoflagellates* lineages; another involving a *Green Algae* and a heterotrophic eucaryote and giving birth to *Euglens* [12].

The interesting point with the produced core trees (especially the one obtained with DOGMA, see `http://members.femto-st.fr/christophe-guyeux/en/chloroplasts`) is that organisms resulting from the first endosymbiosis are distributed in each of the lineages found in the chloroplast genome structure evolution. More precisely, all *Red Algae* chloroplasts are grouped together in one lineage, while *Green Algae* and *Land Plants* chloroplasts are all in a second lineage. Furthermore organisms resulting from the secondary endosymbioses are well localized in the tree: both the chloroplasts of *Brown Algae* and *Dinoflagellates* representatives are found exclusively in the lineage also comprising the *Red Algae* chloroplasts from which they evolved, while the *Euglens* chloroplasts are related to the *Green Algae* chloroplasts from which they evolved. This makes sense in terms of biology, history of lineages, and theories of chloroplasts origins (and so photosynthetic ability) in different Eucaryotic lineages [12].

Interestingly, the sole organisms under consideration that possess a chloroplast (and so a chloroplastic genome) but that have lost the photosynthetic ability (being parasitic plants) are found at the basis of the tree, and not together with their phylogenetically related species. This means that functional chloroplast genes are evolutionary constrained when used in photosynthetic process, but loose rapidly their efficiency when not used, as recently observed for a species of Angiosperms [8]. These species are *Cuscuta-grovonii*, an Angiosperm (flowering plant) at the base of the DOGMA Angiosperm-Conifers branch, and *Epipactis-virginiana*, also an Angiosperm, at the complete basis of this tree.

Another interesting result is that *Land Plants* that represent a single sublineage originating from the large and diverse lineage of *Green Algae* in Eucaryotes history are present in two different branches of the DOGMA tree, both associated with *Green Algae*: one branch comprising the basal grade of *Land Plants* (mosses and ferns) and the second one containing the most internal lineages of *Land Plants* (Conifers and flowering plants). But independently of their split in two distinct branches of the DOGMA tree, the *Land Plants* always show a higher number of functional genes in their chloroplasts than the *Green Algae* from which they emerged, probably meaning that the terrestrial way of life necessitates more functional genes for an optimal photosynthesis than the marine one. However, a more detailed analysis of selected genes is necessary to better understand the reasons why such a distribution has been obtained. Remark finally that all these biologically interesting results are apparent only in the core tree based on DOGMA, while they are not so obvious in the NCBI one.

6 Conclusion

In this research work, we studied two methodologies for extracting core genes from a large set of chloroplasts genomes, and we developed Python programs to evaluate them in practice.

We firstly considered to extract core genomes by the way of comparisons (global alignment) of DNA sequences downloaded from NCBI database. However this method failed to produce biologically relevant core genomes, no matter

the chosen similarity threshold, probably due to annotation errors. We then considered to use the DOGMA annotation tool to enhance the genes prediction process. The second method consisted in extracting gene names either from NCBI gene features or from DOGMA results. A first "intersection core matrix (ICM)" where built, in which each coefficient stored the intersection cardinality of the two genomes placed at the extremities of its row and column. New ICMs are then constructed by selecting the maximum intersection score (IS) in this matrix, removing the two genomes having this score, and adding the corresponding core genome in a new ICM construction.

Core trees have finally been generated for each method, to investigate the distribution of chloroplasts and core genomes. The tree from second method based on DOGMA has revealed the best distribution of chloroplasts regarding their evolutionary history. In particular, it appears to us that each endosymbiosis event is well branched in the DOGMA core tree.

In future work, we intend to deepen the methodology evaluation by considering new gene prediction tools and various similarity measures on both gene names and sequences. Additionally, we will investigate new clustering methods on the first approach, to improve the results quality in this promising way to obtain core genes. Finally, the results produced with DOGMA will be further investigated, biologically speaking: the genes content of each core will be studied while phylogenetic relations between all these species will be questioned.

Computations have been performed on the supercomputer facilities of the Mésocentre de calcul de Franche-Comté.

References

[1] Bassam Alkindy, Jean-François Couchot, Christophe Guyeux, and Michel Salomon. Finding the core-genes of chloroplast species. Journées SeqBio 2013, Montpellier, November 2013.

[2] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[3] Rolf Apweiler, Claire O'Donovan, Maria Jesus Martin, Wolfgang Fleischmann, Henning Hermjakob, Steffen Moeller, Sergio Contrino, and Vivien Junker. Swiss-prot and its computer-annotated supplement trembl: How to produce high quality automatic annotation. *EUR. J. BIOCHEM*, 147:9–15, 1985.

[4] Ewan Birney, Michele Clamp, and Richard Durbin. Genewise and genomewise. *Genome research*, 14(5):988–995, 2004.

[5] Javier De Las Rivas, Juan Jose Lozano, and Angel R Ortiz. Comparative analysis of chloroplast genomes: functional annotation, genome-based phy-

logeny, and deduced evolutionary patterns. *Genome research*, 12(4):567–583, 2002.

[6] Alkindy B. *et al.* Find core-genes for chloroplasts. 2014.

[7] Bakke *et al.* Evaluation of three automated genome annotations for *Halorhabdus utahensis*. *PLoS ONE*, 4(7):e6291, 07 2009.

[8] Li *et al.* Complete chloroplast genome sequence of holoparasite cistanche deserticola (orobanchaceae) reveals gene loss and horizontal gene transfer from its host haloxylon ammodendron (chenopodiaceae). *PloS one*, 8(3):e58747, 2013.

[9] Sayers *et al.* Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 39(suppl 1):D38–D51, 2011.

[10] Zhang *et al.* Cpgavas, an integrated web server for the annotation, visualization, analysis, and genbank submission of completely sequenced chloroplast genome sequences.

[11] Stephane Guindon, Franck Lethiec, Patrice Duroux, and Olivier Gascuel. Phyml online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids research*, 33(suppl 2):W557–W559, 2005.

[12] Geoffrey Ian McFadden. Primary and secondary endosymbiosis and the origin of plastids. *Journal of Phycology*, 37(6):951–959, 2001.

[13] Genís Parra, Enrique Blanco, and Roderic Guigó. Geneid in drosophila. *Genome research*, 10(4):511–515, 2000.

[14] Genis Parra, Keith Bradnam, and Ian Korf. Cegma: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23(9):1061–1067, 2007.

[15] P. Rice, I. Longden, and A. Bleasby. Emboss: the european molecular biology open software suite. *Trends Genet*, 16(6):276–7, 2000.

[16] Robert K. Jansen Stacia K. Wyman and Jeffrey L. Boore. Automatic annotation of organellar genomes with dogma. *BIOINFORMATICS, oxford Press*, 20(172004):3252–3255, 2004.

[17] Alexandros Stamatakis. The raxml 7.0. 4 manual. *Department of Computer Science. Ludwig-Maximilians-Universität München*, 2008.

[18] Alexandros Stamatakis, Thomas Ludwig, and Harald Meier. Raxml-iii: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, 21(4):456–463, 2005.

[19] Hideaki Sugawara, Osamu Ogasawara, Kousaku Okubo, Takashi Gojobori, and Yoshio Tateno. Ddbj with new system and face. *Nucleic acids research*, 36(suppl 1):D22–D24, 2008.