Finding the Core-Genes of Chloroplasts

Bassam Alkindy*‡, Jean-François Couchot*‡, Christophe Guyeux*‡

Arnaud Mouly†‡, Michel Salomon*‡, Jacques M. Bahi*‡

*FEMTO-ST Institute, UMR 6174 CNRS, DISC Computer Science Department
†Chrono-Environnement Lab., UMR 6249 CNRS
‡University of Franche-Comté, France

***Abstract*—Due to the recent evolution of sequencing techniques, the number of available genomes is rising steadily, leading to the possibility to make large scale genomic comparison between sets of close species. An interesting question to answer is: what is the common functionality genes of a collection of species, or conversely, to determine what is specific to a given species when compared to other ones belonging in the same genus, family, etc. Investigating such problem means to find both core and pan genomes of a collection of species, *i.e.*, genes in common to all the species vs. the set of all genes in all species under consideration. However, obtaining trustworthy core and pan genomes is not an easy task, leading to a large amount of computation, and requiring a rigorous methodology. Surprisingly, as far as we know, this methodology in finding core and pan genomes has not really been deeply investigated. This research work tries to fill this gap by focusing only on chloroplastic genomes, whose reasonable sizes allow a deep study. To achieve this goal, a collection of 99 chloroplasts are considered in this article. Two methodologies have been investigated, respectively based on sequence similarities and genes names taken from annotation tools. The obtained results will finally be evaluated in terms of biological relevance.**

***Index Terms*—Chloroplasts, Coding sequences, Clustering, Genes prediction, Methodology, Pan genome, Core genome**

I. Introduction

Identifying core genes may be of importance either to understand the shared functionality and specificity of a given set of species, or to construct their phylogeny using curated sequences. Therefore, in this work we present methods to determine both core and pan genomes of a large set of DNA sequences. More precisely, we focus on the following questions by using a collection of 99 chloroplasts as an illustrative example: how can we identify the best core genome (that is, an artificially designed set of functional coding sequences as close as possible to the real biological one) and how to deduce scenarios regarding their genes loss. In other words, how to deduce scenarios regarding the gene increasing compared to the core genome?

Chloroplasts found in Eucaryotes have an endosymbiotic origin, which means that they come from the incorporation of a photosynthetic bacteria (Cyanobacteria) within an eucaryotic cell, which means they are the 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 allow 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 large scale, 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 [8]. 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. Such investigations have already been started in in [9], where core genome for photosynthetic productivity in *Cyanobacteria* (*Synechococcus* and *Prochlorococcus*) has been regarded. Authors identified core photosystem II genes in cyanophages, which may increase viral fitness by supplementing the host production of some specific types of proteins. The study also proposed evidences of the presence of photosystem I genes in the genomes of viruses that affect cyanobacteria. However, the circumscription of the core chloroplast genomes for a given set of photosynthetic organisms needs bioinformatics investigations using sequence annotation and comparison tools, for which choices are available.

Our intention in this first research work regarding the methodology in core and pan genomes determination is to investigate the impact of these choices. A general presentation of the approaches detailed in this document is provided in the next section. Then we will study in Section III-A the use of annotated genomes from NCBI website [1] with a coding sequences clustering method based on the Needleman-Wunsch similarity scores [2]. We will show that such an approach

based on sequences similarity cannot lead to satisfactory results, biologically speaking. We will thus investigate name-based approaches in Section III-B, by using successively the gene names provided by NCBI and DOGMA [4] annotations, where DOGMA is a recent annotation tool specific to chloroplasts. Finally, a discussion based on biological aspects regarding the evolutionary history of the considered genomes will finalize 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.

II. General presentation

Figure 1 presents a general overview of the entire proposed pipeline for core and pan genomes production and exploitation, which consists of three stages: *Genomes annotation*, *Core extraction*, and *Features Visualization*.

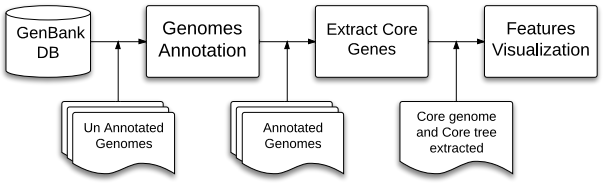

Fig. 1: A general overview of the annotation-based approach

As a starting point, the pipeline uses a DNA sequences database like NCBI's GenBank [1], the European *EMBL* database [5], or the Japanese *DDBJ* one [3]. 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. Various cost-effective annotation tools [15] that produce genomical annotations at many detailed levels have been designed recently, some reputed ones being: DOGMA [4], cpBase [6], CpGAVAS [9], and CEGMA [11]. Such tools usually use one out of the three following methods for finding gene locations in large DNA sequences: *alignment-based*, *composition based*, or a combination of both [11]. The alignment-based method is used when trying to predict a protein coding sequence by aligning a genomic DNA sequence with a cDNA sequence coding an already known homologous protein [11]. This approach is used for instance in GeneWise [13]. The alternative method, the composition-based one (also known as *ab initio*) is based on probabilistic models of genes structure [12].

Using such annotated genomes, we will detail two general approaches for extracting the core genome, which is the second 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. 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.

TABLE I: List of chloroplast genomes of photosynthetic Eucaryotes lineages from NCBI

F.	#	Acc. No	Scientific Name
Brown Algae	11	NC_001713.1	*Odontella sinensis*
		NC_008588.1	*Phaeodactylum tricornutum*
		NC_010772.1	*Heterosigma akashiwo*
		NC_011600.1	*Vaucheria litorea*
		NC_012903.1	*Aureoumbra lagunensis*
		NC_014808.1	*Thalassiosira oceanica*
		NC_015403.1	*Fistulifera sp*
		NC_016731.1	*Synedra acus*
		NC_016735.1	*Fucus vesiculosus*
		NC_018523.1	*Saccharina japonica*
		NC_020014.1	*Nannochloropsis gadtina*
F1	3	NC_000925.1	*Porphyra purpurea*
		NC_001840.1	*Cyanidium caldarium*
		NC_006137.1	*Gracilaria tenuistipitata*
Green Algae	17	NC_000927.1	*Nephroselmis olivacea*
		NC_002186.1	*Mesotigma viride*
		NC_005353.1	*Chlamydomonas reinhardtii*
		NC_008097.1	*Chara vulgaris*
		NC_008099.1	*Oltmannsiellopsis viridis*
		NC_008114.1	*Pseudoclonium akinetum*
		NC_008289.1	*Ostreococcus tauri*
		NC_008372.1	*Stigeoclonium helveticum*
		NC_008822.1	*Chlorokybus atmophyticus*
		NC_011031.1	*Oedogonium cardiacum*
		NC_012097.1	*Pycnococcus provaseolii*
		NC_012099.1	*Pyramimonas parkeae*
		NC_012568.1	*Micromonas pusilla*
		NC_014346.1	*Floydiella terrestris*
		NC_015645.1	*Schizomeris leibleinii*
		NC_016732.1	*Dunaliella salina*
		NC_016733.1	*Pedinomonas minor*
F2	3	NC_001319.1	*Marchantia polymorpha*
		NC_004543.1	*Anthoceros formosae*
		NC_005087.1	*Physcomitrella patens*
F3	2	NC_014267.1	*Kryptoperidinium foliaceum*
		NC_014287.1	*Durinskia baltica*
F4	2	NC_001603.2	*Euglena gracilis*
		NC_020018.1	*Monomorphina aenigmatica*
Ferns	5	NC_003386.1	*Psilotum nudum*
		NC_008829.1	*Angiopteris evecta*
		NC_014348.1	*Pteridium aquilinum*
		NC_014699.1	*Equisetum arvense*
		NC_017006.1	*Mankyua chejuensis*
F5	1	NC_007288.1	*Emiliana huxleyi*
F6	2	NC_014675.1	*Isoetes flaccida*
		NC_006861.1	*Huperzia lucidula*

F.	#	Acc. No	Scientific Name
Angiosperms	45	NC_007898.3	*Solanum lyopersicum*
		NC_001568.1	*Epifagus virginiana*
		NC_001666.2	*Zea Mays*
		NC_005086.1	*Amborella trichopoda*
		NC_006050.1	*Nymphaea alba*
		NC_006290.1	*Panax ginseng*
		NC_007578.1	*Lactuca sativa*
		NC_007957.1	*Vitis vinifera*
		NC_007977.1	*Helianthus annuus*
		NC_008325.1	*Daucus carota*
		NC_008336.1	*Nandina domestica*
		NC_008359.1	*Morus indica*
		NC_008407.1	*Jasminum nudiflorum*
		NC_008456.1	*Drimys granadensis*
		NC_008457.1	*Piper cenocladum*
		NC_009601.1	*Dioscorea elephantipes*
		NC_009765.1	*Cuscuta gronovii*
		NC_009808.1	*Ipomea purpurea*
		NC_010361.1	*Oenothera biennis*
		NC_010433.1	*Manihot esculenta*
		NC_010442.1	*Trachelium caeruleum*
		NC_013707.2	*Olea europea*
		NC_013823.1	*Typha latifolia*
		NC_014570.1	*Eucalyptus*
		NC_014674.1	*Castanea mollissima*
		NC_014676.2	*Theobroma cacao*
		NC_015830.1	*Bambusa emeiensis*
		NC_015899.1	*Wolffia australiana*
		NC_016433.2	*Sesamum indicum*
		NC_016468.1	*Boea hygrometrica*
		NC_016670.1	*Gossypium darwinii*
		NC_016727.1	*Silene vulgaris*
		NC_016734.1	*Brassica napus*
		NC_016736.1	*Ricinus communis*
		NC_016753.1	*Colocasia esculenta*
		NC_017609.1	*Phalaenopsis equestris*
		NC_018357.1	*Magnolia denudata*
		NC_019601.1	*Fragaria chiloensis*
		NC_008796.1	*Ranunculus macranthus*
		NC_013991.2	*Phoenix dactylifera*
		NC_016068.1	*Nicotiana undulata*
Gymnosperms	7	NC_009618.1	*Cycas taitungensis*
		NC_011942.1	*Gnetum parvifolium*
		NC_016058.1	*Larix decidua*
		NC_016063.1	*Cephalotaxus wilsoniana*
		NC_016065.1	*Taiwania cryptomerioides*
		NC_016069.1	*Picea morrisonicola*
		NC_016986.1	*Gingko biloba*

where lineages F1, F2, F3, F4, F5, and F6 are *Red Algae*, *Bryophytes*, *Dinoflagellates*, *Euglena*, *Haptophytes*, and *Lycophytes* respectively.

The final stage of our pipeline, only envoked 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 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 [1] as shown in Table I. These genomes lie in the eleven type of chloroplast families. 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.

III. Core genomes extraction

A. Similarity-based approach

The first method, described below, considers a distance-based similarity measure on genes' coding sequences. Such an approach requires annotated genomes, like the ones provided by the NCBI website.

1) Theoretical presentation: We start with the following preliminary definition.

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. In other words, vertices are DNA sequences, and two sequences are connected with an edge is their similarity is larger than a predefined threshold. Remark that this graph is generally not connected for sufficiently large thresholds.

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. To say this another way, if there is a finite sequence $s_1, ..., s_k$ of vertices (DNA sequences) such that g_i is similar to s_1, which is similar to s_2, etc., and s_k is similar to g_j.

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 $\left\{g_1^G, ..., g_{m_G}^G\right\}$ 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 $\left\{\pi(g_1^G), ..., \pi(g_{m_G}^G)\right\}$. 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.

Remark finally that this first method requires the calculation of all similarities between all allele sequences in all species under consideration. So, even in the case of ≈ 100 organisms and with a focus on a specific family or function, this is a computationally heavy operation. In a future work, the authors' intention is to take benefits from the very large set of already computed similarities, to develop heuristic approaches using this basis of knowledge, specific to chloroplastic genes, making it possible to build rapidly this similarity graph.

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 [16]) 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 [2]. 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

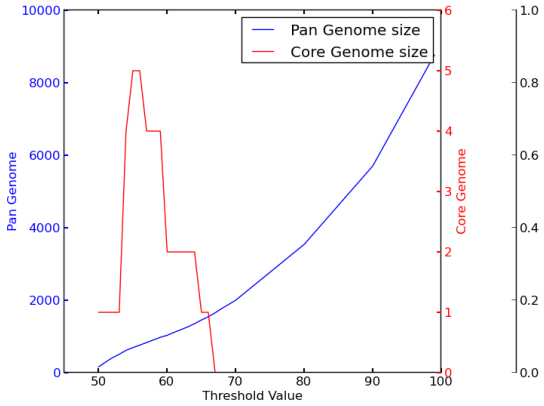

Fig. 2: Results based on NCBI annotation

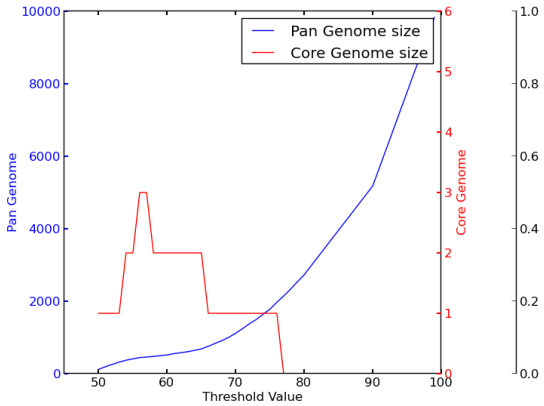

Fig. 3: Results based on DOGMA annotation

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 Figures 2 and 3 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 underperformance 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 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.

B. Annotation-based approach

1) Using genes names provided by annotation tools: Instead of using the sequences predicted by annotation tools, we can try to use the names associated to these sequences, when available. The basic idea is thus to annotate all the sequences using a given software, and to consider as a core gene each sequence whose name can be found in all the genomes. Two annotation techniques will be used in the remainder of this article, namely DOGMA and NCBI.

It is true that the NCBI annotations are of varying qualities, and sometimes such annotations are totally erroneous. As stated before, it is due to the large variety of annotation tools that can been used during each sequence submission process. However, we also considered it in this article, as this database contains human-curated annotations. To say this another way, DOGMA automatic annotations are good in average, while NCBI contains very good human-based annotations together with possibly bad annotated genomes. Let us finally remark that DOGMA also predicts the locations of *ribosomal RNA (rRNA)*, while they are not provided in gene features from NCBI. Thus core genomes constructed on NCBI data will not contain rRNA.

We now investigate core and pan genomes design using each of the two tools separately, which will constitute the second approach detailed in this article. From now on we will consider annotated genomes: either "genes features" downloaded from the NCBI, or the result of DOGMA.

2) Names processing: As DOGMA is a deterministic annotation tool, when a given gene is detected twice in two genomes, the same name will be attached to the two coding sequences: DOGMA spells exactly in the same manner the two gene names. So each genome is replaced by a list of gene names, and finding the common core genes between two genomes simply consists in intersecting the two lists of genes. The sole problem we have detected using DOGMA on our chloroplasts is the case of the RPS12 gene: some genomes contain RPS12_3end or RPS12_5end in DOGMA result. We have manually considered that all these representatives belong to the same gene, namely to RPS12.

Dealing with NCBI names is more complicated, as various annotation tools have been used together with human annotations, and because there is no spelling rule for gene names. For instance, NAD6 mitochondrial gene is sometimes written as ND6, while we can find RPOC1, RPOC1A, and RPOC1B in our chloroplasts. So if we simply consider NCBI data without treatment, intersecting two genomes provided as list of gene names often leads to duplication of misspelled genes. Automatic names homogenization is thus required on NCBI annotations, the question being where to draw the line on correcting errors in the spelling of genes? In this second approach, we propose to automate only obvious modifications like putting all names in capital letters and removing useless symbols as "_", "(", and ")". Remark that such simple renaming process cannot tackle with the situations of NAD6 or RPOC1 evoked above. To go further in automatic corrections requires to use edit distances like Levenshtein, however such use will raise false positives (different genes with close names will be homogenized). The use of edit distances on gene names, together with a DNA sequence validation stage, will be investigated in a second methodology article.

At this stage, we consider now that each genome is mapped to a list of gene names, where names have been homogenized in the NCBI case.

3) Core genes extraction: To extract core genes, we iteratively collect the maximum number of common genes among genomes, therefore during this stage an *Intersection Core Matrix* (ICM) is built. ICM is a two dimensional symmetric matrix where each row and each column corresponds to one genome. Hence, an element of the matrix stores the *Intersection Score* (IS): the cardinality of the core genes obtained by intersecting the two genomes. Mathematically speaking, if we have n genomes in local database, the ICM is an $n \times n$ matrix whose elements satisfy:

$$score_{ij} = |g_i \cap g_j| \tag{1}$$

where $1 \leq i \leq n$, $1 \leq j \leq n$, and g_i, g_j are genomes. The generation of a new core genome depends obviously on the value of the intersection scores $score_{ij}$. More precisely, the idea is to consider a pair of genomes such that their score is the largest element in the ICM. These two genomes are then removed from the matrix and the resulting new core genome is added for the next iteration. The ICM is then updated to take into account the new core genome: new IS values are computed for it. This process is repeated until no new core genome can be obtained.

We can observe that the ICM is relatively large due to the amount of species. As a consequence, the computation of the intersection scores is both time and memory consuming. However, since ICM is obviously a symmetric matrix we can reduce the computation overhead by considering only its upper triangular part. The time complexity for this process is: $O(\frac{n.(n-1)}{2})$. Algorithm 1 illustrates the construction of the ICM matrix and the extraction of the core genomes, where

GenomeList represents the database storing all genomes data. At each iteration, this algorithm computes the maximum core genome with its two parents (genomes).

Algorithm 1 Extract Maximum Intersection Score

Require: $L \leftarrow$ genomes sets
Ensure: $B1 \leftarrow$ Max Core set
  for $i \leftarrow 1 : len(L) - 1$ **do**
    $score \leftarrow 0$
    $core1 \leftarrow set(GenomeList[L[i]])$
    $g1 \leftarrow L[i]$
    for $j \leftarrow i + 1 : len(L)$ **do**
      $core2 \leftarrow set(GenomeList[L[j]])$
      $core \leftarrow core1 \cap core2$
      if $len(core) > score$ **then**
        $score \leftarrow len(core)$
        $g2 \leftarrow L[j]$
      end if
    end for
    $B1[score] \leftarrow (g1, g2)$
  end for
  return $max(B1)$

4) Features visualization: The last stage of the proposed pipeline is naturally to take advantage of the produced core and pan genomes for biological studies. As this key stage is not directly related to the methodology for core and pan genomes discovery, we will only outline a few tasks that can be operated on the produced data.

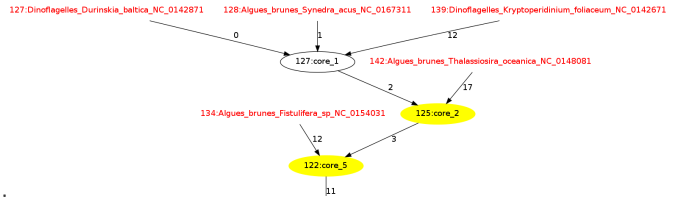

Fig. 4: Part of a core genomes evolutionary tree (NCBI gene names)

Obtained results may be visualized by building a core genomes evolutionary tree, simply called *core tree*. Each node in this tree represents a chloroplast genome or a predicted core, as depicted in Figure 4. In this figure, nodes labels are of the form *(Genes number:Family name_Scientific name_Accession number)*, while an edge is labeled with the number of gene loss when compared to its parents (a leaf genome or an intermediate core genome). Such numbers can answer questions like: how many genes are different between two species? Which functionality has been lost between an ancestor and its children? For complete core trees based either on NCBI names or on DOGMA ones, see http://members.femto-st.fr/christophe-guyeux/en/chloroplasts.

A second application of such data is obviously to build accurate phylogenetic trees, using tools like PHYML[7] or RAxML[14]. Consider a set of species, the least common core genome in a core tree contains all shared common genes among these species. To constitute a phylogenetic tree, core genes will be multi-aligned to serve as an input to the phylogenetic tools mentioned above. An example of such a phylogenetic tree for core 58 is provided in Figure 5. Remark that, in order to constitute the phylogenetic tree, a relevant outgroup is needed from *Cyanobacteria*. The process simply starts by blasting each gene in the core with outgroup genes, and then selects the relevant one.

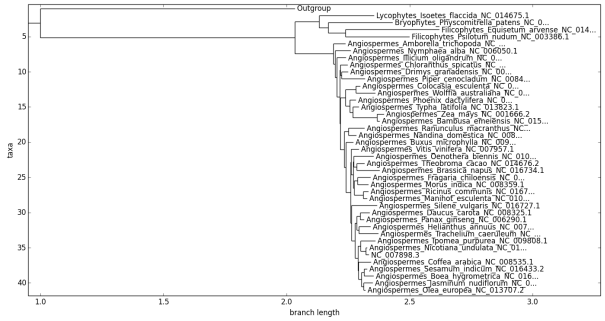

Fig. 5: Part of a phylogenetic tree for core 58 (NCBI gene names)

IV. Discussion

A. 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* [17].

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 [17].

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 [18]. These species are *Cuscuta gronovii*, an Angiosperm (flowering plant) at the base of the DOGMA Angiosperm-Conifers branch, and *Epifagus virginiana*, also an Angiosperm, at the complete basis of this tree.

Another interesting result is that *Land Plants* that represent a single sub-lineage 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 larger 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.

V. 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)" were 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 successively constructed by selecting the maximum intersection score (IS) in this matrix, removing each time 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] Eric W. Sayers, Tanya Barrett, Dennis A. Benson, Evan Bolton, Stephen H. Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M. Church, Michael DiCuccio, Scott Federhen, Michael Feolo, Ian M. Fingerman, Lewis Y. Geer, Wolfgang Helmberg, Yuri Kapustin, David Landsman, David J. Lipman, Zhiyong Lu, Thomas L. Madden, Tom Madej, Donna R. Maglott, Aron Marchler-Bauer, Vadim Miller, Ilene Mizrachi, James Ostell, Anna Panchenko, Lon Phan, Kim D. Pruitt, Gregory D. Schuler, Edwin Sequeira, Stephen T. Sherry, Martin Shumway, Karl Sirotkin, Douglas Slotta, Alexandre Souvorov, Grigory Starchenko, Tatiana A. Tatusova, Lukas Wagner, Yanli Wang, W. John Wilbur, Eugene Yaschenko, and Jian Ye., "Database resources of the national center for biotechnology information". Nucleic Acids Res. 40(D1):D13-D25, January 2012.

[2] P. Rice, I. Longden, and A. Bleasby. "EMBOSS: the European Molecular Biology Open Software Suite", Journal Trends in Genetics, 16(6):276-277, 2000.

[3] 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.

[4] Robert K. Jansen, Stacia K. Wyman, and Jeffrey L. Boore, "Automatic annotation of organellar genomes with DOGMA", Bioinformatics, 20(17):3252-3255, 2004.

[5] A. Bairoch, R. Apweiler. "The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998". Nucleic Acids Research, 26(1):38-42, 1998.

[6] Javier De Las Rivas, Juan Jose Lozano, Angel R. Ortiz. "Comparative analysis of chloroplast genomes: functional annotation, genome-based phylogeny, and deduced evolutionary patterns", Journal Genome research, 12(4):567–583, 2002.

[7] Stephane Guindon, Franck Lethiec, Patrice Duroux, Olivier Gascuel. "PHYML Online—a web server for fast maximum likelihood-based phylogenetic inference", Journal of Nucleic acids research, 33(2):W557–W559, 2005.

[8] Jacques M. Bahi, Christophe Guyeux, Antoine Perasso. "Predicting the Evolution of two Genes in the Yeast Saccharomyces Cerevisiae", Journal of Procedia Computer Science, 11:4-16, 2012.

[9] Chang Liu, Linchun Shi, Yingjie Zhu, Haimei Chen, Jianhui Zhang, Xiaohan Lin, and Xiaojun Guan. "CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences", Journal BMC genomics, 13(suppl 1):715, 2012.

[10] Itai Sharon, Ariella Alperovitch, Forest Rohwer, Matthew Haynes, Fabian Glaser, Nof Atamna-Ismaeel, Ron Y Pinter, Frédéric Partensky, Eugene V Koonin, Yuri I Wolf, et al.. (2009). Photosystem I gene cassettes are present in marine virus genomes. Nature, 461(7261), 258-262.

[11] Genís Parra, Keith Bradnam, and Ian Korf. "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes", Journal of Bioinformatics, 23(9):1061–1067, 2007.

[12] Genís Parra, Enrique Blanco, Roderic Guigó. "Geneid in drosophila", Journal of Genome research, 10(4):511–515, 2000.

[13] Birney, Ewan, Michele Clamp, and Richard Durbin. "GeneWise and genomewise.", Genome research 14(5):988-995, 2004.

[14] Alexandros Stamatakis."The RAxML 7.0. 4 Manual", Department of Computer Science. Ludwig-Maximilians-Universität München 2008.

[15] Peter Bakke, Nick Carney, Will DeLoache, Mary Gearing, Kjeld Ingvorsen, Matt Lotz, Jay McNair, Pallavi Penumetcha, Samantha Simpson, Laura Voss, et al., "Evaluation of three automated genome annotations for Halorhabdus utahensis". PLoS One, 4(7):e6291, 2009.

[16] 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.

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

[18] Xi Li, Ti-Cao Zhang, Qin Qiao, Zhumei Ren, Jiayuan Zhao, Takahiro Yonezawa, Masami Hasegawa, M James C Crabbe, Jianqiang Li, and Yang Zhong., “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.

Bassam ALKINDY was born in january 1978 in Baghdad, Iraq. In 2005, he defended his M.Sc. thesis in Robotics, from the University of Yarmouk, faculty of information technology and computer science, department of Computer Science, Jordan.

Since 2006, he became a full time lecturer in the University of Mustansiriyah, Baghdad-Iraq. Since 2012, he became a Ph.D. student in the University of Franche-Comté, Besançon-France, in Femto-ST/ DISC - Department of computer science. He is interesting in the domains of Artificial Intelligent, bioinformatics, and Machine learning.

Jean-François Couchot is an Assistant Professor in the Department of Computer Science (DISC) of the FEMTO-ST Institute (UMR 6174 CNRS) at the University of Franche-Comté. He received a Ph.D. in Computer Science in 2006 in the FEMTO-ST Institute and applied for a postdoctoral position at INRIA Saclay Île de France in 2006. His research focuses on discrete dynamic systems (with applications in data hiding, pseudorandom number generators, hash function) and on bioinformatics, especially in genes evolution prediction. He has written more than 20 scientific articles in these areas.

Christophe Guyeux has taught mathematics and computer science in the Belfort-Montbéliard university Institute of Technologies (IUT-BM) this last decade.

He has defended a computer science thesis dealing with security, chaos, and dynamical systems in 2010 under Jacques Bahi’s leadership, and is now an associated professor in the computer science department of complex system (DISC), FEMTO-ST Institute, University of Franche-Comté. Since 2010, he has published 2 books, 17 articles in international journals, and 27 articles dealing with security, chaos, or bioinformatics.

Arnaud Mouly was born in France, in October 1978. He completed his Ph.D studies at the National Herbarium of the Nation Museum of Natural History. After the Ph.D degree he occupied a postdoc position in systematic botany at the Bergius Foundation of Royal Swedish Academy of Sciences. He is currently assistant professor in plant systematics and ecology at the University of Franche-Comté, in Besançon, France. His research interest includes botany, diversity dynamics in insular systems, and usages of phylogenetic trees to answer biological/ecological questions. He is also the director of the Botanical garden of Besançon, France.

Michel Salomon is an Assistant Professor in the Department of Computer Science (DISC) of the FEMTO-ST Institute (UMR 6174 CNRS-UFC-ENSMM-UTBM) at the University of Franche-Comté (UFC). He received a Ph.D. in Computer Science in December of 2001 from the University of Strasbourg. His research focuses on dynamic systems, in particular machine learning approaches, with applications in various areas: wireless networks, radiotherapy (prediction of respiratory motion and the dose deposited), bioinformatics (protein folding and so on.), or active airflow control.

Jacques M. Bahi was born in July, 25th 1961. He received the M.Sc. and Ph.D. degrees in applied mathematics from the University of Franche-Comté, France, in 1991.

From 1992 to 1999, he was an Associate Professor of applied mathematics at the Mathematical Laboratory of Besançon. His research interests were focused on parallel synchronous and asynchronous algorithms for differential algebraic equations and singular systems. Since September 1999, he has been a full professor of computer science at the University of Franche-Comté. He published about 150 articles in peer reviewed journal and international conferences and 2 scientific books.

Dr. Bahi is the head of the Distributed Numerical Algorithms team of the Computer Science Laboratory of Besançon, he supervised 21 Ph.D. students. He is a member of the editorial board of 2 international journals and is regularly a member of the scientific committees of many international conferences. Currently, he is interested in: 1) high performance computing, 2) distributed numerical algorithms for ad-hoc and sensor networks and 3) dynamical systems with application to data hiding and privacy.