The Gene Flow between Plasmids and Chromosomes: Insights from Bioinformatics Analyses

Isabel Maida, Marco Fondi, Maria Cristiana Papaleo, Elena Perrin and Renato Fani*

Laboratory of Microbial and Molecular Evolution, Department of Evolutionary Biology, Via Romana 17-19, University of Florence, I-50125 Firenze, Italy

Abstract: Despite the key-role of plasmids in the horizontal spreading of genetic information within the prokaryotic community, their evolutionary dynamics has been poorly explored. Recently, a computational pipeline (Blast2Network) based on similarity networks reconstruction has been developed. Adopting this strategy we analyzed the genome of *Synechococcus* sp. PCC 7002, a cyanobacterium that possesses six plasmids and one chromosome and that can be considered as a model microorganism to analyze the interrelationships existing between different plasmids and between plasmids and chromosome. The aims of our work were: i) the identification and the analysis of the possible evolutionary relationships existing within and among the six *Synechococcus* sp. PCC 7002 plasmids at both intra- and intermolecular level, ii) the analysis of the gene flow (if any) existing between chromosome and plasmids of different size inhabiting the same cytoplasm. A deep analysis of the obtained networks suggested that i) intra-molecular rearrangements did not occur very recently in evolution of plasmids; ii) evolutionary relationships exist between both plasmids and plasmids and chromosome; iii) these exchanges may involve a single gene, operons or gene clusters. Besides, the overall degree of sequence identity/similarity shared by interconnected proteins suggests that plasmids are more prone to recombine between them rather than with chromosome.

Keywords: Blast2Network, gene sharing, molecular remodelling, plasmids, gene duplication.

INTRODUCTION

In the last decade, the total number of completely sequenced prokaryotic genomes has grown exponentially and, at the time of this writing, there are 1083 publicly listed bacterial and archaeal genome projects (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) at different stages of progress [1].

The arisal of next-generation sequencing technologies (NGST) has immediately raised the question on how to store, update, and (probably more interestingly) interpret the whole amount of available data. Bioinformatics, that is the interdisciplinary field that blends computer science and biostatistics with biological and biomedical sciences, is expected to fill these gaps. Indeed, bioinformatics has affected a large number of biological fields and the *in silico* analysis of genomic data currently available has provided significant advances in our understanding of a number of important themes, including bacterial diversity and populations characteristics, also providing a number of challenges in understanding the ecology of, as yet, undiscovered bacterial worlds. These approaches can also help in gaining a deeper understanding of the evolutionary forces that have shaped genomes architecture, from the origins of new genes to their clustering into operonic structures [2]. These key advances are tightly coupled to both NGST and bioinformatics packages for sequence data analysis and led to the arisal of the very intriguing challenge of reconstructing the main evolutionary steps of (in principle) each gene sequenced by sequencing machinery. This task is usually achieved through phylogenomics, that is the integration of both phylogenetic methods and genomics. Interestingly, phylogenomics and comparative genomics approaches based on the availability of a large number of completely sequenced genomes have highlighted the importance of non-vertical transmission in shaping genomes, that is the possibility that (in some cases) genes may not follow classical vertical inheritance but, rather, may be directly transferred (horizontally) between different cells. This process is usually referred to as horizontal gene transfer (HGT) and, despite its extent is still under debate [3, 4], it is thought to have played a major role (at least) in the early stages of bacterial evolution [5-7].

From a molecular perspective, HGT is usually carried out by different mechanisms and is mediated by a mobile gene pool (the so called "mobilome") that comprises plasmids, transposons and bacteriophages (all of which usually referred to as mobile genetic elements, MGEs) [8, 9], the major players of this process. In fact, plasmids and other MGEs can be transferred between micro-organisms, representing natural vectors for the transfer of genes and functions they code for [8]. Usually MGEs do not accommodate any of the "core" genes required by the cell for basic growth and division, but rather they carry genes that may be useful periodically to enable the cell to exploit particular environmental conditions, such as the survival in the presence of a potentially lethal antibiotic [10]. This flexibility is mostly due to the abundance of transposable elements they harbour and that facilitate intra- and

*Address correspondence to this author at the Laboratory of Microbial and Molecular Evolution, Department of Evolutionary Biology, Via Romana 17-19, University of Florence, I-50125 Firenze, Italy; Tel: +390552288244; Fax: +390552288250; E-mail: renato.fani@unifi.it

1874-1363/11

intermolecular recombination by creating homology regions. In this way, a single DNA fragment (possibly embedding one or more coding genes) can be exchanged between the MGE harbouring it and other informational molecules (including chromosomes and/or other MGEs). In this context, it is particularly interesting to find that, in some cases, chromosomes and plasmids inhabiting the same cell can share sequences possessing a very high degree of similarity, probably as the result of recombination events [11]. In some cases, also chromosomes and plasmids belonging to different strains/species were found to share a number of homologous sequences, probably as the result of one (or more) HGT event(s) [11]. This latter event has important biological drawbacks since it may allow the transfer of previously plasmid-encoded functions to the chromosome(s) and, in turn, permit to the corresponding genes to be spread in the bacterial community through vertical inheritance.

Despite the key-role of plasmids in the prokaryotic world, their evolutionary dynamics has been poorly explored, mainly because of the lack of extensive similarities between them, except for genes involved in replication and transfer functions [12, 13], which hampers classical phylogenetic analyses based on gene genealogy and synteny [14]. However, a computational biology approach (Blast2Network) based on similarity networks reconstruction and phylogenetic profiling (Fig. **1**) has been recently proposed and applied in a study-case to depict the similarities among plasmids from *Enterobacteriaceae* [8] and further utilized to analyze the *Acinetobacter* pan-plasmidome [11].

Briefly, the main workflow of the program (Fig. **1**) starts from a file containing protein or nucleic acid sequences in standard NCBI fasta format. This is used as an input to gather information on source sequences from the NCBI website. Input sequences are then screened one against each

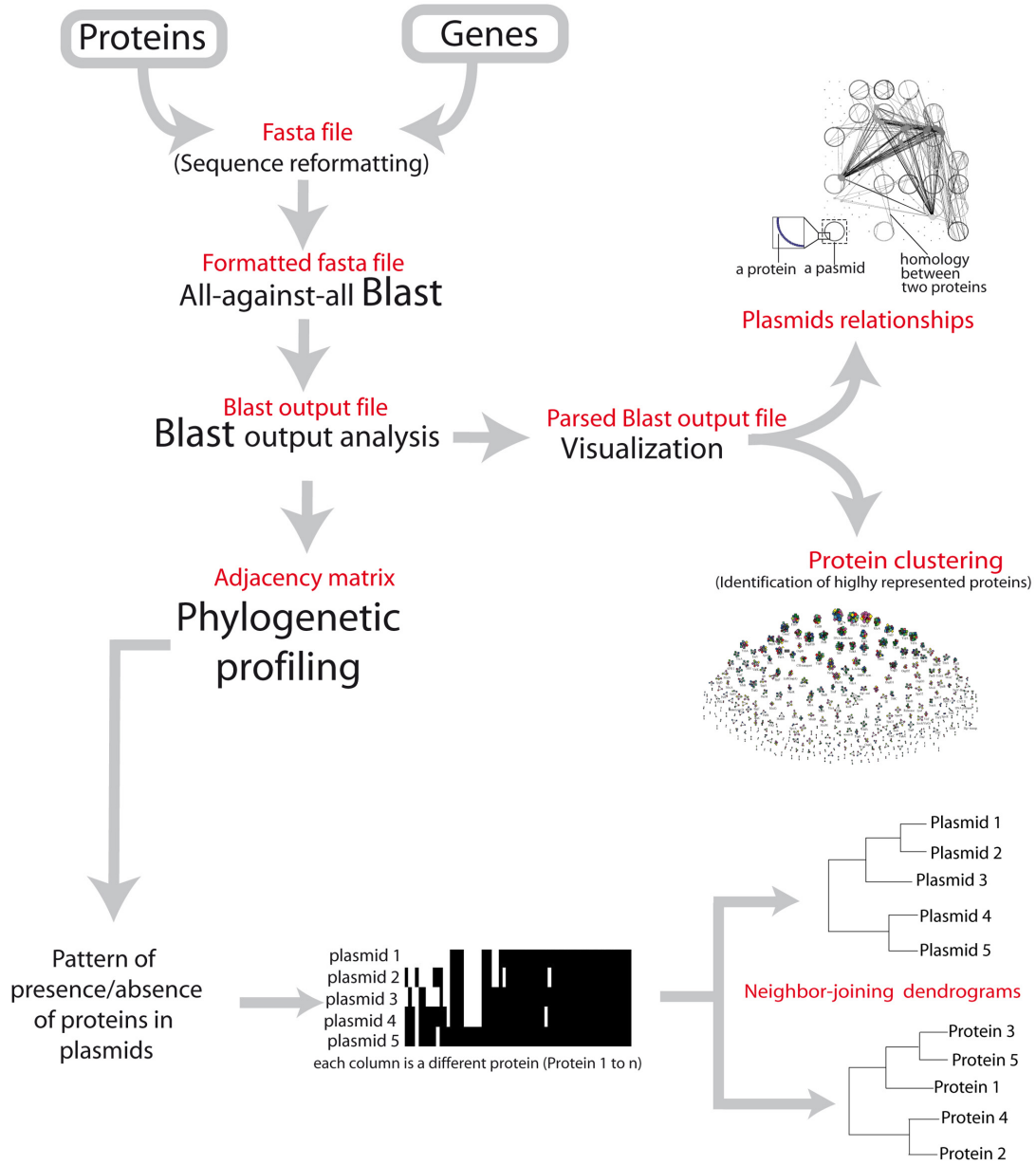

Fig. (1). Workflow of the program Blast2Network (redrawn from [8]).

other using Blast [15]. The resulting output is parsed in the form of an adjacency matrix that describes the global sequence similarities in the dataset where each entry wij reflects the similarity existing between protein *i* and *j* (Fig. **1**). B2N also outputs the Blast post processing results as a network in Visone format (http://visone.info/), a freely available software for network visualization and analysis. By doing so, B2N transforms the Blast output into a similarity network in which the nodes represent proteins whereas the links indicate the existence of a given degree of sequence similarity between them (Fig. **1**). Moreover, in the resulting network, all the nodes belonging to the same plasmid source are circularly arranged and filled with the same color.

Finally, the adjacency matrix obtained by parsing the Blast output is the input for the phylogenetic profile method allowing the analysis of co-occurrence patterns, metabolic reconstruction and so on (see Materials and Methodology). The proposed methodology is general enough to be used for comparative genomic analyses of bacteria and was recently implemented in a more comprehensive computational pipeline in order to study the extent and the dynamics of HGT of antibiotic resistance determinants within the whole bacterial community (Fondi and Fani, *unpublished results*).

In this work, we have applied the above mentioned workflow to the analysis of the whole genome of *Synechococcus* sp. PCC 7002 a cyanobacterium that possesses six plasmids and one chromosome (whose sequences are publically available) and that, for this reason, can be considered as a model microorganism to analyze the interrelationships existing between different plasmids and between plasmids and chromosome. Remarkably, *Synechococcus* genus has recently gained a well-recognized importance from both physiology and molecular biology viewpoints since representatives of this genus have a ubiquitous distribution in oceanic waters [16], are estimated to be responsible for around a quarter of the primary production in some regions and possess a well-recognized importance within the global carbon cycling [17].

MATERIALS AND METHODOLOGY

Sequence Data Source

The dataset used in this work is composed of all the proteins encoded by the completely sequenced *Synechococcus* sp. PCC 7002 genome (a single chromosome and six plasmids) that were downloaded from the NCBI ftp websites ftp://ftp.ncbi.nih.gov/genomes/Bacteria/ and ftp://ftp.ncbi. nih.gov/refseq/release/plasmid (Table **1**).

Network Construction

Similarity, identity based, networks were constructed using the tools implemented in the software B2N [8]. Networks, whereby the nodes represent the proteins and the links connecting them represent the shared identity values, were visualized and analyzed using the software Visone (http://visone.info/).

Functional Assignment

The putative functional role of unassigned proteins was automatically retrieved according to the first best hit (FBH) in a similarity search (using Blast algorithm [15]) in the PFAM database (http://pfam.sanger.ac.uk/). The standalone version of the databases was used, and default parameters were used during Blast probing.

RESULTS AND DISCUSSION

Overall Strategy

The overall strategy used in this work is reported in Fig. (**2**) and starts with the retrieval of the aminoacid sequence of all the proteins encoded by the chromosome and the six plasmids of the cyanobacterium *Synechococcus* sp. PCC 7002. We have chosen this micro-organism as a model system since its genome is composed of one chromosome and six plasmids of different length (ranging from 4,809 to 186,459 nt and coding for a number of proteins comprised between 3 and 165) (Table **1**). In this way it should be possible to analyze both the (eventual) cross-talk between chromosomes and plasmid(s) and also between different plasmids inhabiting the same cytoplasm.

The second step of the strategy is the construction of networks showing the correlations existing at intra- or intermolecular level, i.e. within the same DNA molecule (plasmids) or between one or more DNA molecules [plasmid-plasmid(s) or plasmid(s)-chromosome] (Fig. **2**). The analysis of the networks will allow the identification of events of duplication and/or recombination involving one

Table 1. List of *Synechococcus* sp. PCC 7002 DNA Molecules Analyzed in this Work

Strain	Plasmid				
	Name	Accession Number	Length (nt)	N° of Genes	N° of Proteins
Synechococcus sp. PCC 7002	pAQ1	NC_010476	4,809	3	3
	pAQ3	NC_010477	16,103	17	17
	pAQ4	NC_010478	31,972	30	30
	pAQ5	NC_010479	38,515	39	39
	pAQ6	NC_010480	124,030	109	109
	pAQ7	NC_010474	186,459	165	165
	Chromosome				
	Chromosome 1	NC_010475	3,008,047	2,875	2,824

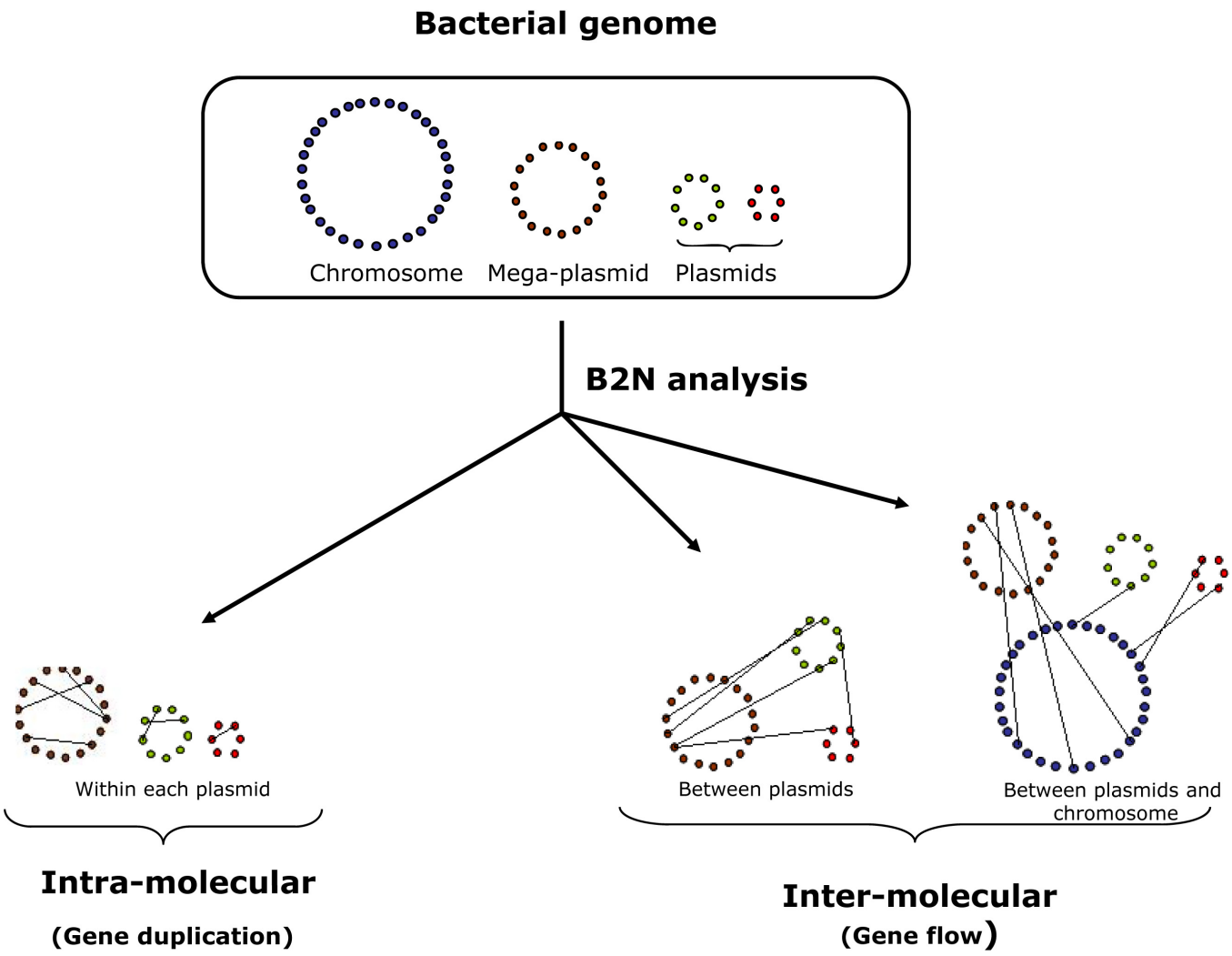

Fig. (2). Overall strategy used in this work.

gene, cluster of genes and/or entire operons as well as transposition events.

Plasmids Networks

The first aim of the work was the identification and the analysis of the possible evolutionary relationships existing among the six *Synechococcus* sp. PCC 7002 plasmids at both intra- and intermolecular level.

Intramolecular Analysis

During this stage of the analysis, we aimed at identifying the interrelationships existing between proteins coded for by genes harbored by the same plasmid molecule. In other words, we checked the existence of single/multiple internal duplications involving a stand-alone gene, one (or more) operon(s) or cluster(s) of genes. In this way we may shed some light also on the role that paralogous gene duplication [18, 19] might have had on the construction of plasmids. To this purpose, all the protein sequences encoded by each *Synechococcus* sp. PCC 7002 plasmid were used as input for the B2N software (see Materials and Methodology) and, as a result, a set of networks showing all the sequence identities existing among these proteins was obtained (Fig. **3**). In these networks nodes represent proteins, whereas links indicate the existence of sequence identity among them. The degree of sequence identity threshold is *a priori* selected. In principle, the higher the threshold used, the lower the number of links existing between proteins encoded by different plasmids. In addition, it can be assumed that the higher the degree of amino acid identity between two proteins, the more recent would be the event (recombination/transposition/duplication/ vertical transmission) responsible for the presence of the two orthologous/paralogous coding genes in different plasmids. We selected a minimum of 40% identity threshold since this degree of sequence identity is sufficiently high to guarantee that in most cases the interconnected proteins perform the same function (i.e., they are encoded by orthologous genes) [20, 21].

The networks obtained by reiterating the analysis using different identity thresholds are shown in Fig. (**3**), whose analysis revealed that no trace of paralogy was found in each of the four shortest *Synechococcus* sp. PCC 7002 plasmids (pAQ1, pAQ3, pAQ4, and pAQ5). Paralogous genes were found only in the larger plasmids (pAQ6 and pAQ7). The analysis of paralogous genes within the same molecule revealed that:

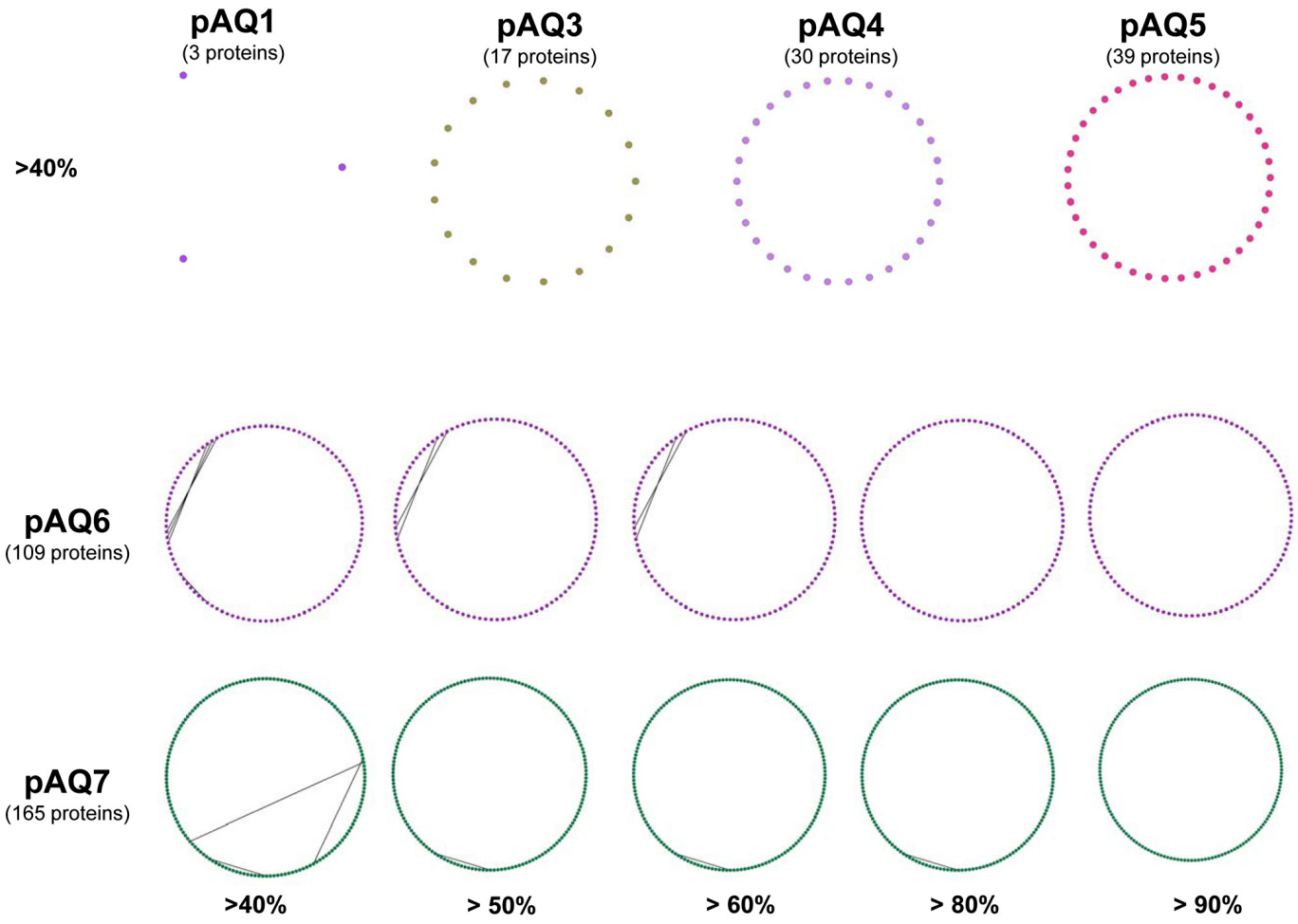

Fig. (3). Identity based networks showing the intra-molecular relationships existing within each the six *Synechococcus* sp. PCC 7002 plasmids.

1. At 40% identity threshold, two events of paralogous duplication were disclosed in plasmid pAQ6 (Fig. **4**). The first one involves two genes (GI code: 170079618 and 170079625) coding for TPR-repeat containing proteins. The two proteins share a degree of sequence identity/similarity of 33.0% and 48.8 %, respectively (Table **2**). This finding might suggest that the duplication event of the ancestral gene cannot be traced in recent times. However, the two proteins strongly differ in their length with protein coded for by gene with GI code 170079618 being 1,402 residues, whereas the other one (GI 170079625) 1,106 residues long. The ClustalW [22] alignment of the two amino acid sequences revealed that they mainly differ at the N-terminus (data not shown).

   The second duplication exhibited by plasmid pAQ6 did not involve a single gene, but a cluster of three genes (GI codes 170079610-611-612 and 170079590-591-592, respectively). A further analysis of these sequences revealed that they code for hypothetical proteins with unknown function (Table **2**). Since they are organized in the same relative order, we suggest that the two clusters are the outcome of a duplication event involving (at least) the three cistrons. We cannot *a priori* exclude the alternative possibility, i.e. that each of the three ancestral genes underwent single independent duplication events, placing the three copies in the same relative order. However, in our opinion, this possibility is much less probable than the previous one. Besides, even though no transcriptional data is available for these plasmids, it is quite possible that the two triads are the outcome of an operon duplication event. This relies on the finding that the three genes of each triad are very close to each other and separated by short intergenic regions. Moreover, two genes (GI 17007590-591) of one triad overlap at a considerable extent. This *per se* implies a translational coupling, strongly suggesting that they belong to the same transcriptional unit, i.e. an operon (Fig. **5a**).

2. In plasmid pAQ7, at 40% identity threshold, five events of paralogous duplication involving five pairs of genes were disclosed (Figs. **3**, **4**). Four of these gene pairs encode proteins involved in iron metabolism. The degree of sequence identity between the proteins of each pair is very close (Table **2**), a finding that might suggest that the four duplication events could have occurred in a short timescale. One

of this duplication event is an in tandem rearrangement.

The fifth paralogous gene pair codes for hypothetical proteins sharing a very high degree of sequence identity/similarity (84.6-91.2%, respectively) rising the possibility that the duplication event of the ancestor gene occurred in recent times (and that this event is more recent than the duplication involving the other four gene pairs).

3. No multiple duplications were found in the two plasmids.

The whole body of data reported in this section revealed that some plasmid-borne genes underwent paralogous duplications in different evolutionary stages. The finding that none of the genes involved in paralogous duplications is connected neither to chromosomal or other plasmids genes (see below) strongly suggests that the duplications are the outcome of a recombination event occurring within the same molecule rather than a recombination between plasmids and chromosome or between two different plasmids.

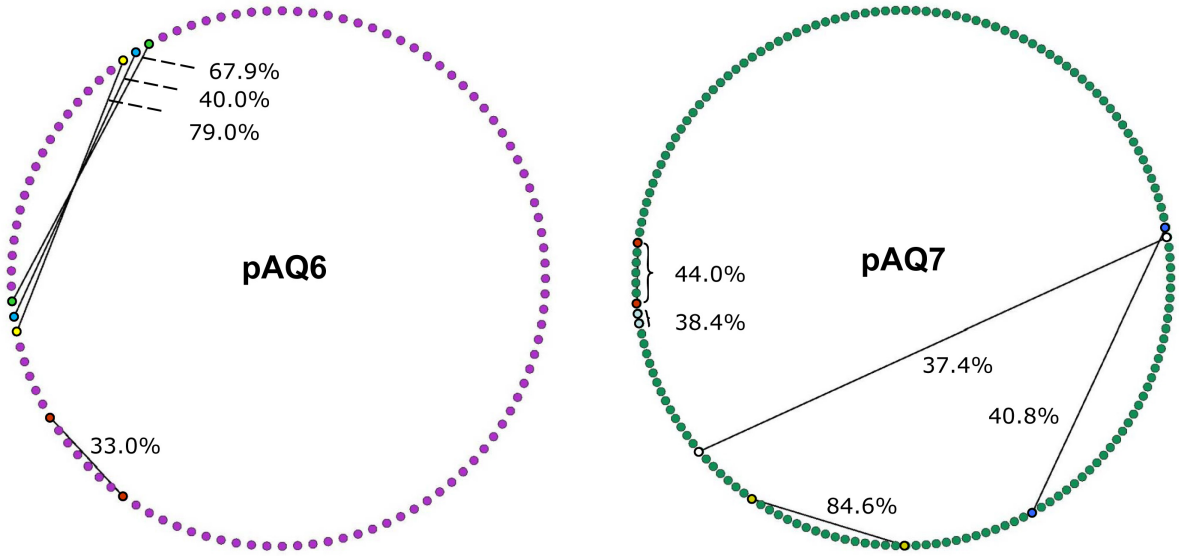

Fig. (4). Paralogous genes in *Synechococcus* sp. PCC 7002 pAQ6 and pAQ7 plasmids. Percentages represent the degree of sequence identity shared by each paralogous gene pair. Paralogous gene are in the same colour.

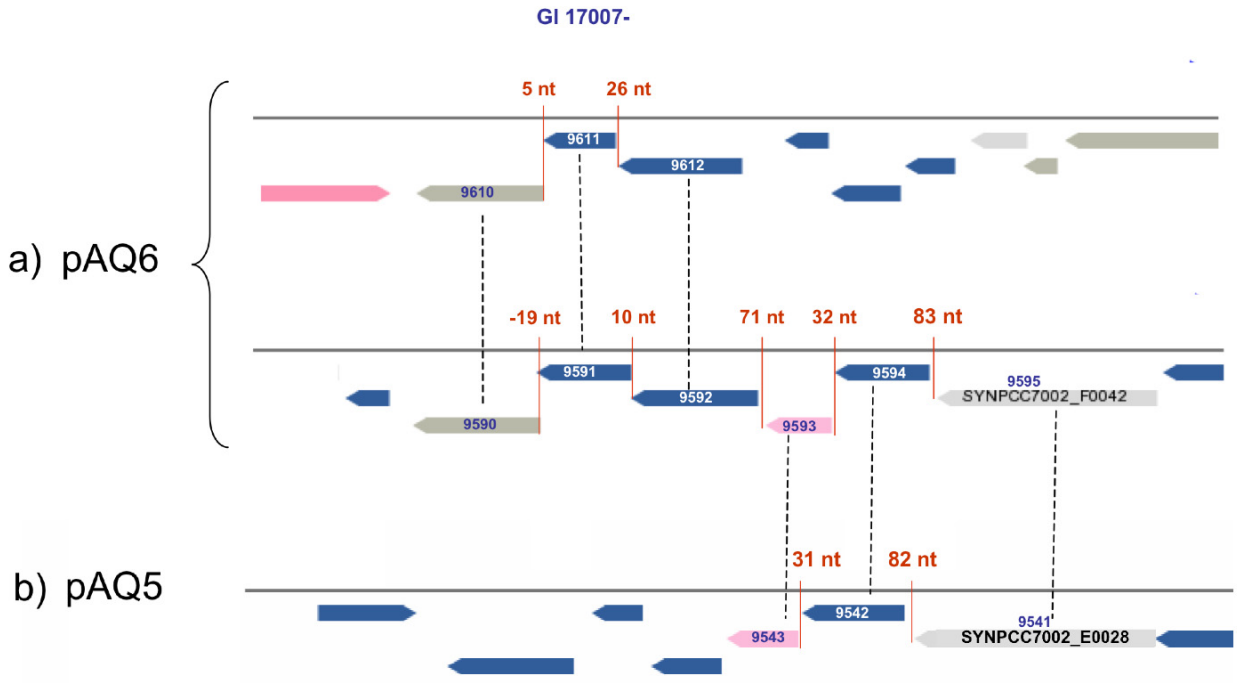

Fig. (5). Schematic representation of interrelationship existing between genes located on plasmids pAQ5 and pAQ6. **a**) paralogous operon in plasmid pAQ6, **b**) homologous operons in pAQ5 and pAQ6. Dotted lines connect homologous genes. Numbers in red represent the length of the intergenic region between two cistrons; negative numbers indicate the overlapping length between two cistrons.

Table 2. List of Paralogous Genes Encoded by *Synechoccoccus* sp. PCC 7002 Plasmids pAQ6 and pAQ7

Plasmid	Gene		Paralogous Genes		
	gi	Putative Function	Pair	% Identity	% Similarity
pAQ6	170079612	Hypothetical Protein	170079612-170079592	67.9	79.1
	170079592	Hypothetical Protein			
	170079611	Hypothetical Protein	170079611-170079591	40.4	53.8
	170079591	Hypothetical Protein			
	170079610	Pentapeptide repeat-containing Protein	170079610-170079590	79.1	89.1
	170079590	Hypothetical Protein			
	170079618	TPR repeat-containing Protein	170079618-170079625	33.0	48.8
	170079625	TPR repeat-containing Protein			
pAQ7	170076550	ATP-binding protein of ABC transporter for iron	170076550-170076556	44.0	61.1
	170076556	ATP-binding protein of ABC transporter for iron			
	170076557	Iron ABC transporter	170076557-170076558	38.4	57.9
	170076558	Iron ABC transporter			
	170076572	Periplasmic binding protein; iron siderophor	170076572-170076475	37.4	55.6
	170076475	Periplasmic binding protein			
	170076608	TonB dependent siderophore receptor	170076608-170076476	40.8	58.9
	170076476	TonB dependent siderophore receptor			
	170076579	Hypothetical Protein	170076579-170076595	84.6	91.2
	170076595	Hypothetical Protein			

Intermolecular Analysis

The intermolecular analysis, i.e. the analysis of the evolutionary relationships existing among the six *Synechococcus* sp. PCC 7002 plasmids, was carried out in order to identify those genes (if any) shared by two or more plasmids. To this purpose, all the 363 retrieved sequences of *Synechococcus* sp. PCC 7002 plasmid-encoded proteins were used as input for the B2N software (see Material and Methodology), generating a set of networks showing all the sequence identity relationships existing among these proteins. The networks obtained at >40% and 100% identity thresholds are shown in Fig. (**6**) (the entire set of networks constructed at >40, >50, >60, >70, >80, >90, 100% identity threshold is reported in Supplementary file 1).

Links analysis. The analysis of links revealed that:

a. Four of the six plasmids (the largest ones, i.e. pAQ4, pAQ5, pAQ6 and pAQ7) are interconnected (although at different extent), suggesting that they shared at least some common steps in their evolutionary pathway. The other (and shorter) two plasmids (pAQ1 and pAQ3) harbor genes encoding proteins that do not share any link neither between them nor with other proteins in the network, suggesting that they are not related to the other plasmids inhabiting the same cytoplasm.

b. As expected, the number of links and interconnected nodes decreased with the increasing of the identity threshold (Table **3**). At 40% of sequence identity, 32 (about 10%) out of the 363 plasmid-encoded proteins were linked together. The other 331 proteins remained isolated because each of them did not share any link with the others and they were excluded from further analyses. The number of linked proteins decreased to 8 at 100% sequence identity threshold. The finding that the number of links connecting the 32 proteins is 19 suggests that some proteins are connected to more than one other protein. Indeed, a deeper analysis of networks shown in Fig. (**6**) revealed that two proteins are encoded by genes shared by three plasmids (pAQ4, pAQ6 and pAQ7). The two genes are adjacent in the three plasmids and very likely belong to an operon (data not shown).

c. Quite interestingly, four gene pairs (Fig. **6**, lower part and Table **4**) code for proteins sharing the very same amino acid sequence (100% identity). This might indicate that the plasmids sharing at least one of these genes underwent recombination events very recently in time.

d. Plasmid pAQ5 is connected only to pAQ6 *via* three cistrons coding for proteins sharing a very high degree of sequence identity (see Table **4**), which in turn suggests that the molecular rearrangements involving (the ancestors of) plasmids pAQ5 and pAQ6 is recent. The three cistrons are contiguous and likely organized in an operon (Fig. **5b**). It is noteworthy that the three cistrons of plasmids pAQ6, homologous to the triads of pAQ5, are adjacent to one of the two triads of plasmid pAQ6 involved in the paralogous operon duplication described above.

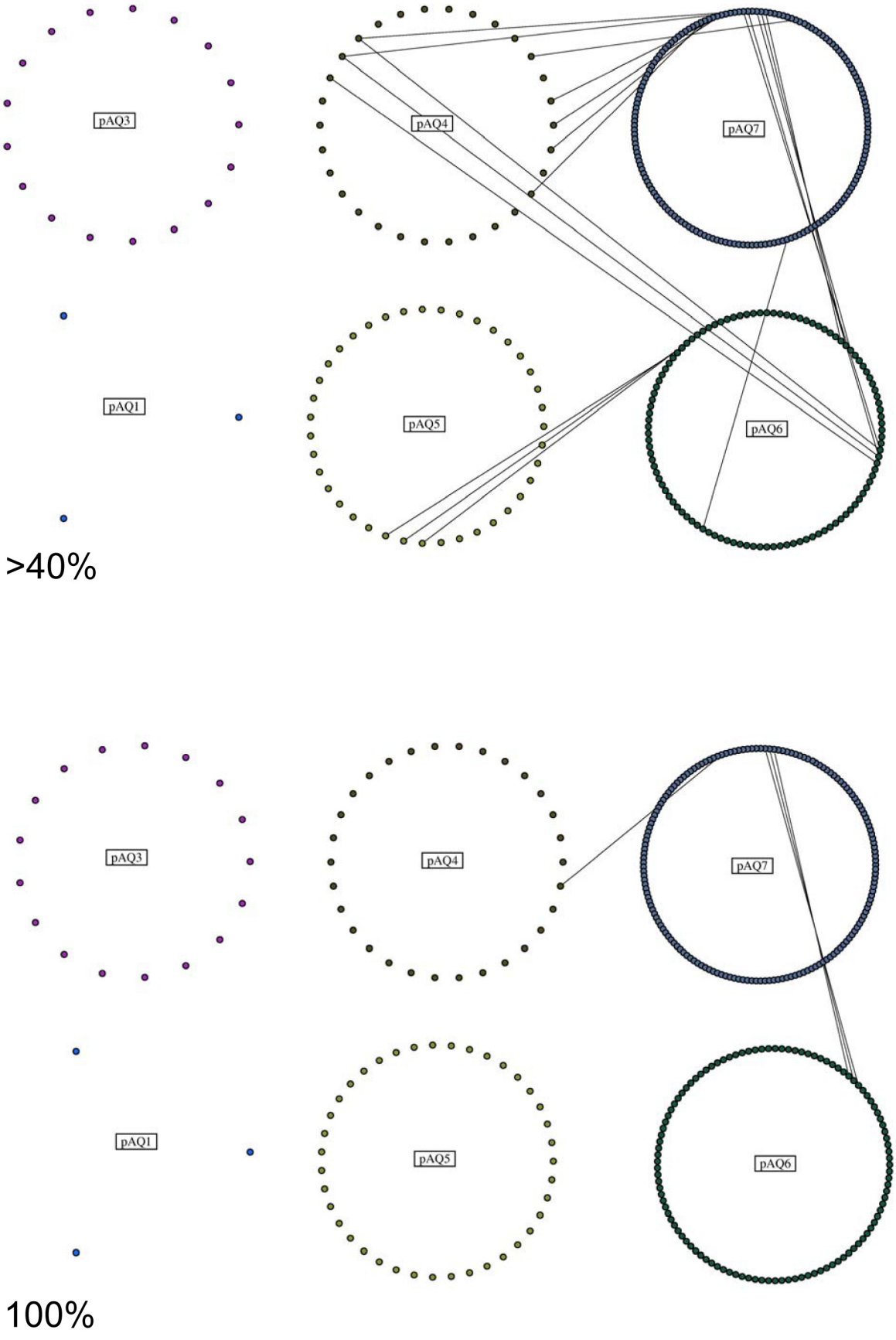

Fig. (6). Identity based networks of the 363 Synchecococcus sp. PCC 7002 plasmid encoded proteins. All the proteins belonging to the same plasmid (nodes) are circularly arranged and are linked to the others according to their identity value. The resulting pictures for two different identity thresholds (>40% and 100%) are shown.

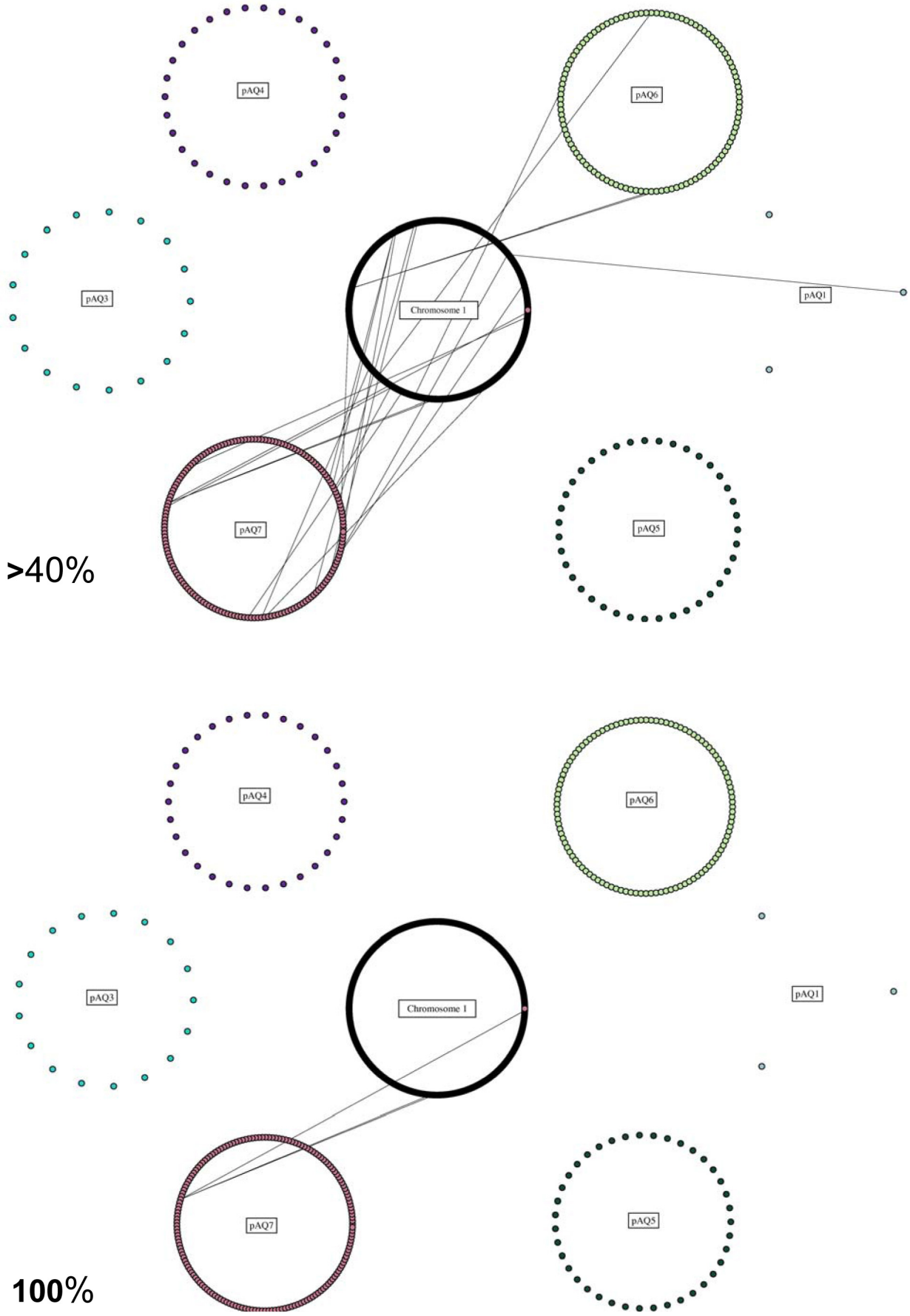

Fig. (7). Identity relationships between the proteins of the *Synechococcus* sp. PCC 7002 plasmids and chromosome proteins.

Table 3. Number of Links Interconnecting Plasmid-Encoded Proteins at Different Threshold of Sequence Identity. The Number of Proteins Interconnected is Also Shown. Data are Retrieved from Fig. (6)

Plasmid		% of Sequence Identity						
		40	50	60	70	80	90	100
pAQ4	Number of proteins interconnected	8	8	7	6	5	4	1
	Number of links interconnecting proteins	10	8	7	6	5	4	1
pAQ5	Number of proteins interconnected	3	3	3	3	2	2	0
	Number of links interconnecting proteins	3	3	3	3	2	2	0
pAQ6	Number of proteins interconnected	10	8	8	7	6	6	3
	Number of links interconnecting proteins	12	8	8	7	6	6	3
pAQ7	Number of proteins interconnected	11	11	10	10	9	8	4
	Number of links interconnecting proteins	13	11	10	10	9	8	4
	Total number of proteins interconnected	32	30	28	26	22	20	8
	Total number of links interconnecting proteins	19	15	14	13	11	10	4

e. The most interconnected plasmid is pAQ6, which shares genes with all the other three connected plasmids.

Nodes analysis. The analysis of data shown in Table **4** revealed that the 32 connected proteins can be subdivided into fifteen groups, most of which (13) constituted by two members; the remaining two comprises three proteins (groups 14, 15). Concerning the function of the interconnected proteins, half of the groups (2, 6, 8, 10, 12, 14, and 15) include hypothetical proteins with unknown function. Three groups (1, 3, 11) are composed by proteins involved in DNA transposition, that in turn, exhibit a very high degree of sequence identity (99.4-100%) suggesting also a recent interaction between plasmids pAQ4, pAQ6 and pAQ7.

Chromosome–Plasmids Networks

In order to check for the existence of genes shared between plasmids and the chromosome of *Synechococcus* sp. PCC 7002 and to look for possible indications of past and/or recent rearrangements between them, we compared the 363 plasmid proteins at different identity thresholds with all the 2,824 proteins of the chromosome. The identity networks obtained at a sequence identity threshold >40% and 100% are shown in Fig. (**7**) (whereas those obtained at >50, >60, >70, >80, >90, and 100%) are reported in supplementary file 2.

Links analysis. As expected, the number of links decreased with the increase of identity threshold (ranging from 22 links at 40% to 3 links at 100% of threshold identity). Three of the plasmids (pAQ3, pAQ4 and pAQ5) did not share any link with the chromosome. Plasmids pAQ1 and pAQ4 share one and four genes with the chromosome, respectively. The highest number of connections (29) with the chromosome was exhibited by the largest plasmid (pAQ7).

The analysis of Fig. (**7**) also revealed that each connected chromosomal gene was linked to a single plasmid. Concerning the organization of the genes coding for connected proteins, only two genes are adjacent in the chromosome and in a plasmid (pAQ6) (Table **5**), that is GI 17007937 and -7938 and 170079636 and 9637, respectively. Since the two genes share the same orientation, the relative order (coding for proteins with unknown function) and are separated by a 3 nt long intergenic region, it can be surmised that they might form an operon (data not shown). Since the two pairs exhibit very similar degree of sequence identity, it is also quite possible that they are the result of an operon duplication, rather than two independent duplication events.

Nodes analysis. Data reported in Table **5** revealed that most of the groups of proteins are composed by two nodes, one of which belonging to a plasmid molecule and the other one to the chromosome, the only exception being represented by groups 16 and 17. In these two latter cases two proteins of the plasmid pAQ7 (with GI code 170076547 and 170076628, respectively) are connected to three and four chromosomal proteins, respectively. Concerning the function of the protein shared between plasmids and chromosome most of them are involved in different metabolic functions: three groups comprise proteins involved in membrane transport (group 6, 9 and 14) whereas groups 1 and 5 consist of proteins with unknown function.

The degree of sequences identity of the proteins embedded in this groups is not very high, with the exception of group 16 that includes proteins involved in DNA transposition that are 100% identical.

CONCLUSIONS

The aim of this work was to analyze the possible gene flow existing between different DNA molecules (one chromosome and six plasmids of different size) inhabiting the same cytoplasm using the cyanobacterium *Synechococcus* sp. PCC 7002 as a model system. The whole body of data reported in this work revealed that different molecular mechanisms have played an important role in shaping the plasmid(s) structure and that there is a "cross-talk" between genes belonging to: i) the same plasmid, ii) different plasmids, and iii) plasmid(s) and chromosome. In particular:

Table 4. Percentage Identity/Similarity and Function of Proteins Interconnected Between Different *Synechococcus* sp. PCC 7002 Plasmids

gi		Plasmid	% Identity	% Similarity	Group	Function
17007	6502	pAQ7	99.4	100.0	1	Is4 family transposase Tn10 like transposase
	9486	pAQ4				
	6518	pAQ7	53.0	60.5	2	Hypothetical Protein
	9484	pAQ4				
	6521	pAQ7	100.0	100.0	3	Integrase/recombinase
	9512	pAQ4				
	6519	pAQ7	97.9	100.0	4	ParA partitioning protein
	9483	pAQ4				
	6523	pAQ7	99.2	99.2	5	PilT domain containing protein PilT protein. N-terminal
	9510	pAQ4				
	9496	pAQ4	60.3	76.5	6	Hypothetical Protein
	9658	pAQ6				
	9595	pAQ6	75.3	86.7	7	Chromatin associated protein KTI12
	9541	pAQ5				
	9594	pAQ6	92.4	95.2	8	Hypothetical Protein
	9542	pAQ5				
	9593	pAQ6	92.1	97.5	9	Fe II oxygenase family oxidoreductase
	9543	pAQ5				
	6603	pAQ7	90.1	93.2	10	Hypothetical Protein
	9626	pAQ6				
	6509	pAQ7	100.0	100.0	11	Resolvase
	9569	pAQ6				
	6511	pAQ7	100.0	100.0	12	Hypothetical Protein
	9567	pAQ6				
	6510	pAQ7	100.0	100.0	13	Plasmid pRiA4b ORF-3-like
	9568	pAQ6				
	9494	pAQ4	9494-9660= 47.0 9494-6514= 74.0 9660-6514= 43.1	9494-9660=70.5 9494-6514=83.0 9660-6514=65.8	14	Hypothetical Protein
	9660	pAQ6				
	6514	pAQ7				
	9495	pAQ4	9495-9659= 41.6 9495-6513= 81.0 9659-6513= 45.1	9495-9659=53.0 9495-6513=86.3 9659-6513=57.2	15	Hypothetical Protein
	9659	pAQ6				
	6513	pAQ7				

1. Some genes of the largest plasmids (pAQ6 and pAQ7) underwent ***intra-molecular rearrangements*** (*i.e.* paralogous duplication events) of different size and complexity. The duplication events involved either single genes or (in one case) an entire operon. Some of these duplicatios are likely (very) old even though it is not possible to give an exact timing to these events. Other ones occurred later on. However, these internal rearrangements did not occur very recently in evolution as shown by the degree of sequence identity between each pair (Table **2**).

2. The analysis of **plasmids** ***inter-molecular rearrangements*** revealed that four plasmids (pAQ4, pAQ5, pAQ6 and pAQ7) exchanged, at variable extent, some of their genes. Again, the largest ones (pAQ6 and pAQ7) apperead to be more prone to interconnection than the others. In one case the same DNA region containing two gene has been exchanged between (the ancestors of) three plasmids (pAQ4, pAQ6 and pAQ7) (not shown). The exchange may involve a single gene, operons or gene clusters. It is

Table 5. Percentage Identity/Similarity and Function of Proteins Interconnected between *Synechococcus* sp. PCC 7002 Chromosome and Plasmids

gi		DNA Molecule	% Id	% Sim	Group	Function
17007	9461	pAQ1	37.3	53.9	1	Hypothetical Protein
	6845	Chromosome				
	9636	pAQ6	45.3	61.7	2	HEPN domain
	7938	Chromosome				
	9637	pAQ6	41.1	63.5	3	Nucleotydiltransferase
	7937	Chromosome				
	9581	pAQ6	40.1	57.0	4	pilT-domain containing protein
	8423	Chromosome				
	9606	pAQ6	59.0	67.4	5	Hypothetical Protein
	8633	Chromosome				
	6534	pAQ7	36.8	53.6	6	ATP binding protein of ABC transporter
	9425	Chromosome				
	6621	pAQ7	46.6	64.7	7	F0F1 ATP synthase subunit α
	7357	Chromosome				
	6614	pAQ7	49.4	65.2	8	F0F1 ATP synthase subunit β
	7372	Chromosome				
	6486	pAQ7	43.1	62.1	9	Sulfonate/nitrate transport system permease
	7483	Chromosome				
	6484	pAQ7	43.7	61.6	10	Sulfonate/nitrate transport system ATP binding
	7484	Chromosome				
	6597	pAQ7	52.8	66.0	11	Δ-aminolevulinic acid dehydratase
	8360	Chromosome				
	6483	pAQ7	39.3	57.3	12	Exopolysaccaride synthesis protein
	8506	Chromosome				
	6546	pAQ7	41.6	52.7	13	Cytochrome b6-f complex iron-sulfur subunit
	8514	Chromosome				
	6593	pAQ7	60.2	74.0	14	Bumetanide-sensitive Na-K-Cl cotransporter
	8568	Chromosome				Amino acid permease superfamily protein
	6598	pAQ7	55.6	69.0	15	GTP cyclohydrolase I
	9039	Chromosome				
	6547	pAQ7	100.0	100.0	16	Transposase
	9449	Chromosome				
	8739					
	8721					
	6628	pAQ7	6628-6771=42.6 6628-6849=41.5 6628-7473=36.6 6628-8061=42.1	6628-6771=59.5 6628-6849=55.5 6628-7473=53.3 6628-8061=61.8	17	Two component response regulator
	6771	Chromosome				
	6849					
	7473					
	8061					

quite interesting the finding that, in four cases (see groups 3, 11, 12, and 13 in Table **4**) the intermolecular rearrangements are very likely extremely recent, since the encoded proteins exhibit the same amino acid sequence (100% identity). The gene flow between these plasmids might be facilitated by the presence of genes coding for proteins involved in DNA transposition such as resolvase and integrase (see groups 1, 3, and 11 in Table **4**). Concerning the evolutionary pathway leading to the extant patterns of sequence similarity relationships, different and more or less complex scenarios can be depicted. On the basis of the available data it is not possible to trace the exact evolutionary steps for each plasmid. However, at least in the case of plasmids pAQ4, pAQ6 and pAQ7 we suggest that they might have met each others different times in the past, since the degree of sequence identity between proteins coded for by shared genes are strongly variable (in the range of 43.1 and 100.0 %). Indeed, it can be assumed that if two DNA undergo a single (unique) recombination event, all the proteins coded for by the exchanged genes should exhibit the same (or very similar) degree of sequence identity.

Table 6. Number of Connections/Plasmid Encoded Genes Shared by Plasmids or Plasmids and Chromosome at 40% Identity Threshold

Plasmid	Genes	Number of Genes Shared with						
		pAQ1	pAQ3	pAQ4	pAQ5	pAQ6	pAQ7	Chromosome
pAQ1	3	-	0	0	0	0	0	1
pAQ3	17	0	-	0	0	0	0	0
pAQ4	30	0	0	-	0	3	7	0
pAQ5	39	0	0	0	-	3	0	0
pAQ6	109	0	0	3	3	-	6	4
pAQ7	165	0	0	7	0	6	-	17/29*

*The number of genes connected is lower than 34 since two plasmid encoded genes (gi170076547 and gi170076628) are linked to three or four chromosomal genes, respectively.

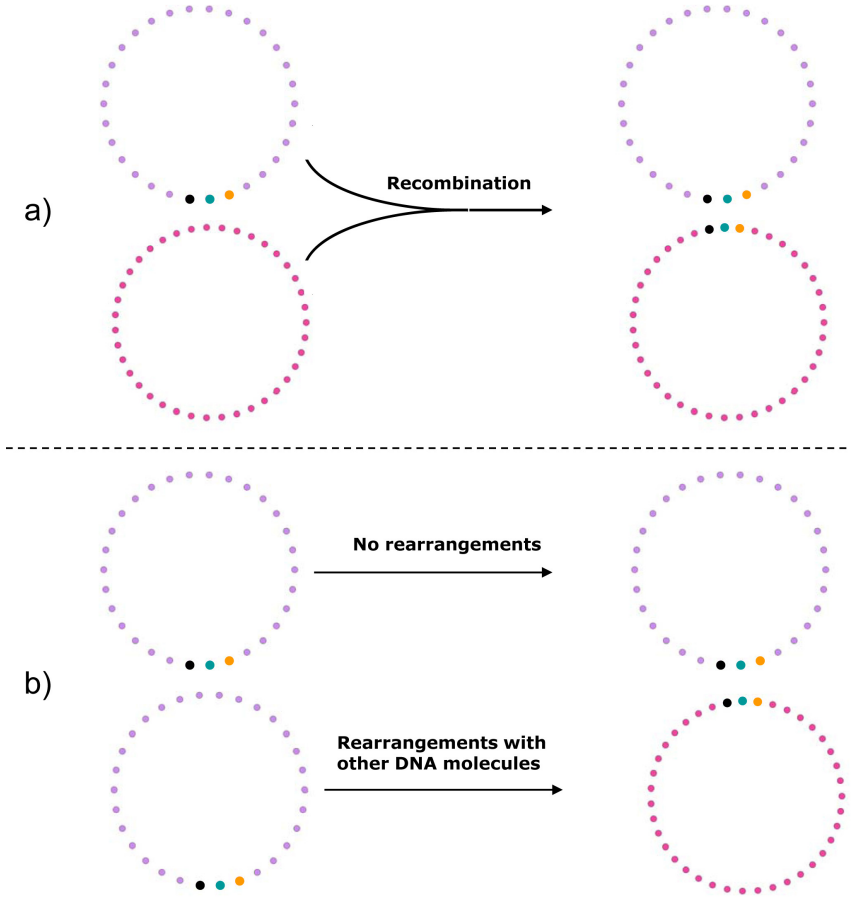

Fig. (8). Two hypothetical evolutionary pathways for the origin of genes shared by two (or more) DNA molecules.

3. A cross-talk between the cyanobacterial plasmids and the chromosome was also disclosed. The gene flow mainly occurred with the largest plasmid (pAQ7) and involved 29 genes coding for a plethora of different functions. Indeed, apparently, there was not a preferential flow of a given functional group of genes.

4. It is quite interesting that one of the six plasmids (pAQ3) did not exhibit the traces (either old or recent) of rearrangement events, at neither intra- nor inter-molecular level (Table **6**). This might suggest that this plasmid introgressed *Synechococcus* sp. PCC 7002 cells very recently.

5. Concerning the origin of genes shared by two (or more) DNA molecules, two different pathways can be proposed: i) the first one predicts (Fig. **8a**) that two different plasmids might have (transiently) inhabited the same cell; if a recombination occurred between them, this might have led to the gene sharing; ii) according to the alternative idea (Fig. **8b**), it is plausible that in the same cells multiple copies of the same plasmids might have inhabited; if one of them underwent a rearrangement with other DNA molecules (another plasmid or the chromosome), it might have acquired new genes in such a way that the structure of the originally identical copies (strongly) diverged. On the basis of the available data, it is not possible in the study case of this work to discern between the two possibilities.

6. It is also worth of noticing that most of the plasmid-borne genes (284, almost 80%) are not involved in any rearrangement and are not linked to any other chromosomal or plasmid gene (Fig. **9**). This opens the question of the origin of these genes. It is quite possible that they might have been recruited through recombination with other plasmids or chromosomes hosted by cells belonging to a different bacterial strain of the same or different species. This, in turn, highlights how the horizontal gene transfer may permit the flow and the exchange of genetic information between prokaryotic cells.

7. Concerning the mechanisms responsible for shaping the plasmid architecture, in addition to transposition and recombination events, it appears that gene duplication has played an important role. The importance of gene duplication for the development of metabolic innovations and shaping the genomes was firstly discussed by Lewis [23] and later by Ohno [24]. Indeed, duplication and divergence of DNA sequences of different size represents one of the most important forces driving the evolution of genes and genomes during the early evolution of life, since this process may allow the formation of new genes from pre-existing ones [19]. The idea that gene duplication may have played a major role in the shaping and in the assembly of genomes has been recently confirmed by the comparative analysis of complete sequences of archaeal, bacterial and eukaryal ones, showing that all of extant organisms harbor a remarkable proportion of paralogous genes and that many of them group into numerous families of different sizes [25-27]. DNA duplications may also concern entire operons or part

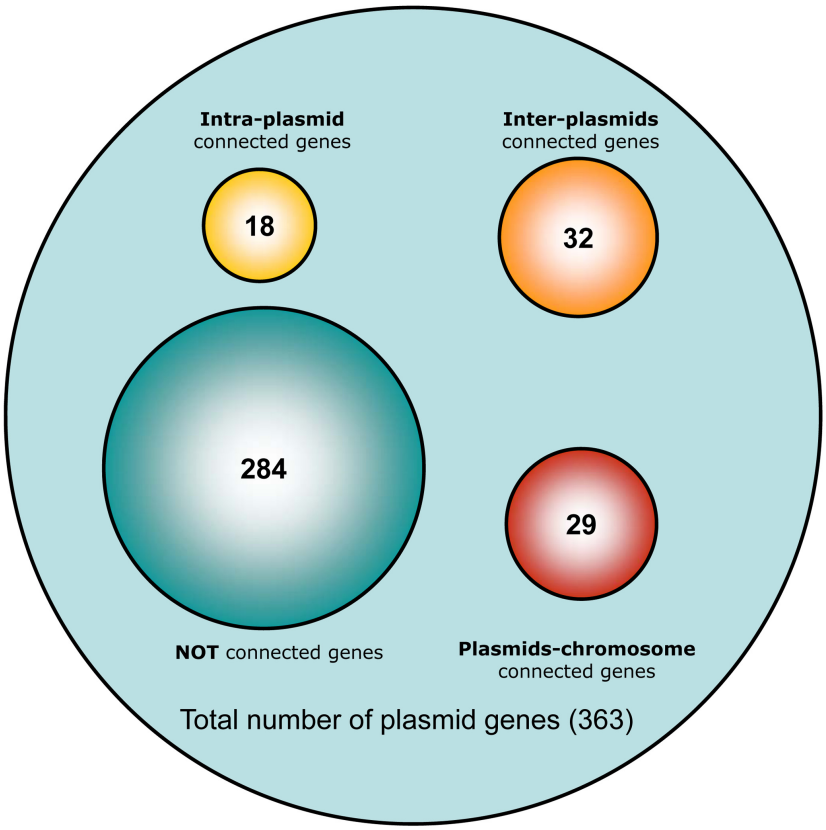

Fig. (9). Schematic representation of *Synechococcus* sp. PCC 7002 gene pool, showing the absence of shared genes by each sub-pool.

thereof, and several examples of single or multiple operon duplications have been recently reported [28]. This, in turn, can lead to the appearance of paralogous operon families, increasing the overall number of operons within a genome. The relative high number of paraologus genes (that is genes that are the outcome of an event of gene duplication) found during the analysis of *Synechococcus* sp. PCC 7002 plasmids networks (Fig. **6**) may suggest that the molecular mechanisms involved in the shaping of chromosome molecules across time might also act in plasmids evolution.

8. It is also worth of noticing that the overall degree of sequence identity/similarity is much higher between proteins shared by different plasmids than the other two possibilities (Table **7**). This strongly suggests that plasmid molecules are more prone to recombine between them rather than with chromosome or within them. The biological significance of this finding is still obscure.

Table 7. Mean Degree of Sequence Identity/Similarity between Nodes

Protein Shared	% Identity	% Similarity
Intra-Plasmids	51.7	66.2
Inter-Plasmids	78.5	85.5
Plasmid-Crhomosome	47.7	62.9

ACKNOWLEDGEMENTS

Marco Fondi is supported by a postdoctoral fellowship from “Fondazione Adriano Buzzati-Traverso”.

CONFLICT OF INTEREST

None declared.

SUPPLEMENTARY MATERIAL

This article is also accompanied with supplementary material and it can be viewed at publisher’s website.

REFERENCES

[1] T. T. Binnewies, Y. Motro, P. F. Hallin, O. Lund, D. Dunn, T. La, D. J. Hampson, M. Bellgard, T. M. Wassenaar, and D. W. Ussery, “Ten years of bacterial genome sequencing: comparative-genomics-based discoveries,” *Funct. Integr. Genome.,* vol. 6, pp. 165-185, Jul 2006.

[2] M. Lynch, “The frailty of adaptive hypotheses for the origins of organismal complexity,” *Proc. Natl. Acad. Sci. USA,* vol. 104, Suppl 1, pp. 8597-8604, May 2007.

[3] T. Dagan, and W. Martin, “Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution,” *Proc. Natl. Acad. Sci. U S A,* vol. 104, pp. 870-875, Jan 2007.

[4] T. Dagan, and W. Martin, “The tree of one percent,” *Genome. Biol.,* vol. 7, p. 118, 2006.

[5] C. Woese, “The universal ancestor,” *Proc. Natl. Acad. Sci. U S A,* vol. 95, pp. 6854-6859, Jun 1998.

[6] C. R. Woese, “Interpreting the universal phylogenetic tree,” *Proc. Natl. Acad. Sci. U S A,* vol. 97, pp. 8392-8396, Jul 2000.

[7] C. R. Woese, “On the evolution of cells,” *Proc. Natl. Acad. Sci. U S A,* vol. 99, pp. 8742-8747, Jun 2002.

[8] M. Brilli, A. Mengoni, M. Fondi, M. Bazzicalupo, P. Lio, and R. Fani, “Analysis of plasmid genes by phylogenetic profiling and visualization of homology relationships using Blast2Network,” *BMC Bioinformatics,* vol. 9, p. 551, 2008.

[9] L. S. Frost, R. Leplae, A. O. Summers, and A. Toussaint, “Mobile genetic elements: the agents of open source evolution,” *Nat. Rev. Microbiol.,* vol. 3, pp. 722-732, Sep 2005.

[10] P. M. Bennett, “Plasmid encoded antibiotic resistance: acquisition and transfer of antibiotic resistance genes in bacteria,” *Br. J. Pharmacol.,* vol. 153, Suppl 1, pp. S347-S357, Mar 2008.

[11] M. Fondi, G. Bacci, M. Brilli, C. M. Papaleo, A. Mengoni, M. Vaneechoutte, L. Dijkshoorn, and R. Fani, “Exploring the evolutionary dynamics of plasmids: the Acinetobacter pan-plasmidome,” *BMC Evol. Biol.,* vol. 10, p. 59, Feb 2010.

[12] R. Fernandez-Lopez, M. P. Garcillan-Barcia, C. Revilla, M. Lazaro, L. Vielva, and F. de la Cruz, “Dynamics of the IncW genetic backbone imply general trends in conjugative plasmid evolution,” *FEMS Microbiol. Rev.,* vol. 30, pp. 942-966, Nov 2006.

[13] M. A. Cevallos, R. Cervantes-Rivera, and R. M. Gutierrez-Rios, “The repABC plasmid family,” *Plasmid.,* vol. 60, pp. 19-37, Jul 2008.

[14] S. D. Bentley, and J. Parkhill, “Comparative genomic structure of prokaryotes,” *Annu. Rev. Genet.,* vol. 38, pp. 771-792, 2004.

[15] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, “Gapped Blast and PSI-Blast: a new generation of protein database search programs,” *Nucleic Acids Res.,* vol. 25, pp. 3389-3402, Sep 1997.

[16] F. Partensky, W. R. Hess, and D. Vaulot, “Prochlorococcus, a marine photosynthetic prokaryote of global significance,” *Microbiol. Mol. Biol. Rev.,* vol. 63, pp. 106-127, Mar 1999.

[17] D. J. Scanlan, “Physiological diversity and niche adaptation in marine Synechococcus,” *Adv. Microb. Physiol.,* vol. 47, pp. 1-64, 2003.

[18] M. Fondi and R. Fani, “The origin and evolution of metabolic pathways,” *Phys. Life Rev.,* vol. 6, pp. 23-52, 2009.

[19] M. Fondi, G. Emiliani, and R. Fani, “Origin and evolution of operons and metabolic pathways,” *Res. Microbiol.,* vol. 160, pp. 502-512, Sep 2009.

[20] W. Tian and J. Skolnick, “How well is enzyme function conserved as a function of pairwise sequence identity?,” *J. Mol. Biol.,* vol. 333, pp. 863-882, Oct 2003.

[21] F. A. Gonzalez, E. Bonapace, I. Belzer, I. Friedberg, and L. A. Heppel, “Two distinct receptors for ATP can be distinguished in Swiss 3T6 mouse fibroblasts by their desensitization,” *Biochem. Biophys. Res. Commun.,* vol. 164, pp. 706-713, Oct 1989.

[22] J. D. Thompson, D. G. Higgins, and T. J. Gebson, “CLUSTAL W: improving the sensitivity of progressive multiple sequence allignement through sequence weighting, position-secific gap penlties and weight matrix choice.,” *Nucleic Acid Res.,* vol. 22, no.22, pp. 4673-4680, 1994.

[23] Lewis, “Pseudoallelism and gene evolution.,” *Spring Harb. Symp. Quant. Biol.,* vol. 16, p. 15, 1951.

[24] S. Ohno, “Simplicity of mammalian regulatory systems,” *Dev. Biol.,* vol. 27, pp. 131-136, Jan 1972.

[25] R. de Rosa, and B. Labedan, “The evolutionary relationships between the two bacteria Escherichia coli and Haemophilus influenzae and their putative last common ancestor,” *Mol. Biol. Evol.,* vol. 15, pp.17-27, Jan 1998.

[26] R. Pushker, A. Mira, and F. Rodriguez-Valera, “Comparative genomics of gene-family size in closely related bacteria,” *Genome. Biol.,* vol. 5, p. R27, 2004.

[27] B. Labedan, and M. Riley, “Widespread protein sequence similarities: origins of Escherichia coli genes,” *J. Bacteriol.,* vol. 177, pp. 1585-1588, Mar 1995.

[28] R. P. Anderson, and J. R. Roth, “Tandem genetic duplications in phage and bacteria,” *Annu. Rev. Microbiol.,* vol. 31, pp. 473-505, 1977.

Received: January 7, 2010 Revised: September 14, 2010 Accepted: March 27, 2011