REVIEW

The extant core bacterial proteome is an archive of the origin of life

Antoine Danchin, Gang Fang and Stanislas Noria

Génétique des Génomes Bactériens, Institut Pasteur, Paris, France

Received: June 16, 2006
Revised: December 8, 2006
Accepted: December 11, 2006

Genes consistently present in a clique of genomes, preferring the leading DNA strands are deemed persistent. The persistent bacterial proteome organises around intermediary and RNA metabolism, and RNA-related information transfer, with a significant contribution to compartmentalisation. Despite inevitable losses during evolution, the extant persistent proteome displays functions present early on. Proteins coded by genes staying clustered in a majority of genomes constitute a network of mutual attraction made up of three concentric circles. The outer one, mostly devoted to metabolism, breaks into small pieces and fades away. The second, more continuous, one organises around class I tRNA synthetases. The well-connected inner circle comprises the ribosome and information transfer. This reflects the progressive construction of cells, starting from the metabolism of coenzymes, nucleotides and fatty acids-related molecules. Subsequently, a core set of aminoacyl-tRNA synthetases scaffolded around RNA, connected to cell division machinery and organised metabolism around translation. This remarkable organisation reflects the evolution of life from small molecules metabolism to the RNA world, suggesting that extant microorganisms carry the marks of the ancient processes that created life. Further analysis suggests that RNA degradation, associated to the presence of iron, still plays a role in extant metabolism, including the evolution of genome structures.

Keywords:
Homeotopic transformation / Oxygen toxicity / Paleocells / Synthetic biology / YycF

1 Introduction

All major domains of science have evolved from a start point where data were collected in a given field of the physical world, then organised by some process of taxonomy or systematics. Mendeleieff has established the catalogue of atoms present in the Universe. Painstaking work has gathered an ever-growing catalogue of stars, a catalogue of plants and animals has been (and still is being) constructed, organising our knowledge of the living world. In the same way, before trying to understand the depths of what life is, it was necessary to get a complete chemical description of what makes a cell and this endeavour was at the core of genome projects [1]. Sequencing genomes was a major step towards identifying all of the macromolecules in the cell because, using the genetic code, one has access not only to the chromosome sequence, but also to its proteome. However, in the same way as the map of the sky is not cosmology, the genome text does not constitute genomics.

Genomes code for an unlimited set of functions, contributing to the fitness of the organism. In this work, we identify and then focus on the core functions that associate together to make every cell alive. We first discuss the concept of function (an action adapted to some 'goal' [2]), and we further try and identify the core proteome that has been recruited in the course of evolution to fulfil that particular task extending a previous approach to a large set of genomes [3]. This approach helps us to place life in context and explore

Correspondence: Professor Antoine Danchin, Génétique des Génomes Bactériens, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France
E-mail: adanchin@pasteur.fr
Fax: +33-1-4568-8948

Abbreviation: PMA, potential of mutual attraction

whether extant organisms are palimpsests [4], precluding any deep insight about the origin of life, or whether they comprise archaeological domains that would tell us much about our past.

The way goals are constructed through evolution – they are not decided *a priori* or from the outside but appear *a posteriori* – is not discussed here, but we retained the constraint that any function is mediated by interactions, and that, in fact, the core of living organisms rests on relationships between objects. In brief, a boat made of wooden planks differs from a simple heap of planks, and its function (floating on water while carrying a content) is directly associated with the way the planks are assembled [5]. We aim in this work at developing more robust concepts and the premises of a dedicated terminology to deal with and categorise biological functions, an essential prerequisite for the transition from symplectic biology (biology focusing on the relationships between objects: from συν, together, and πλεκτειν, to weave) to synthetic biology.

Since the onset of genome programmes, it has been common to try and identify minimal requirements, *i.e.* those genes and gene products that are needed to make a cell alive. This endeavour was at the very root of microbial genome programmes, and in the late 1980s, looking for functions required for life (the functional approach we are advocating in the present work) one could calculate that approximately 400 genes were necessary to allow the construction of a cell [6]. Assuming that an average protein comprises 300 residues, the average gene length would be 1 kb, making the smallest genome for an autonomous bacterium some 400–500 kb long. This short length (only ten times that of bacteriophage λ, sequenced in 1982 using the first shotgun procedure [7]) was used to justify the bacterial genome sequencing endeavour. Later, a structural approach based on the newly acquired full genome sequences was developed. Looking for genes conserved in all organisms, Mushegian and Koonin [8] using the recently deciphered *Mycoplasma genitalium* genome, evaluated to less than 300 the number of genes required to form a minimal genome, *i.e.* a genome that would allow a cell to live on a medium supplied with all possible metabolites under stable physicochemical conditions (pH, humidity, gasses and temperature). Finally, in an experimental approach, the *Bacillus subtilis* European–Japanese consortium disrupted the genes of the organism and predicted that 277 genes were essential for growth on plates supplemented with a rich medium in the laboratory [9]. Several other studies consistently ended with a more or less similar number of genes [10] either when comparing closely related genomes [11] or when taking into account the frequent cotranscription which makes most gene disruption techniques to have a polar effect on the distal part of operons. And this culminates in the endeavour, starting from the genome of *M. genitalium*, to construct one instance of a minimal free-living organism [12].

Growth under well-defined laboratory conditions as well as straightforward genome comparison, however, introduces highly restrictive constraints to functional requirements as this takes into account only very specific environments (in particular, for genomes which have witnessed reductive evolution, because they correspond to extremely narrow growth conditions [13, 14]). When investigating the functions needed for life to develop, we are in fact interested in all functions that are commonly contributing to fitness, not functions that allow growth in highly restricted niches. In summary, we favour a functional analysis over a structural analysis: rather than choosing the intersect of all that exists – and which may ultimately reduce to nothing – we would like to know what is commonly present, in short, what is persistent in a clique of organisms. With this view in mind, we identified in a study involving 28 genomes about twice as many persistent genes (persistent genes are consistently present in a clique of genomes, see Section 2 for a technical definition) as those strictly making the core essential genes [3]. In the present study, when we included many more genomes (228 genomes, see Section 2), this figure did not change significantly, showing that the concept of gene persistence is robust.

Before proceeding to identify core functions using genome comparisons, we first constructed a list of the functions expected to be required for sustaining life (Fig. 1). The organisation of the functions is split into three major processes indispensable to permit life: metabolism, compart

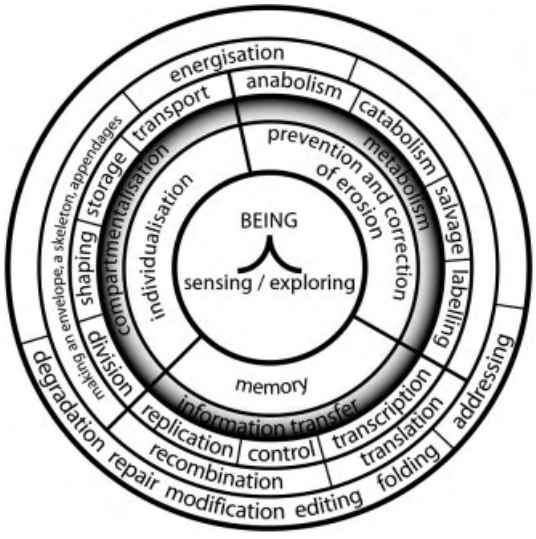

Figure 1. Basic functions for life. Duration in time (being) is performed by compartmentalisation of a metabolism that allows a cell to sustain the constraints of erosion. RNA metabolism, initially a part of intermediary metabolism, allowed cells to discover the complementarity law and then the genetic code, leading to information transfer processes. As an example of functional assignment, the function of the proteins of the lactose operon, *lacI*, *lacZ*, *lacY*, and *lacA* can be illustrated as follows: *lacI*: information transfer, control; *lacZ*: metabolism, degradation; *lacY*: compartmentalisation, transport; *lacA*: metabolism, labelling (lactose is labelled to be transported by a security valve in order to prevent building up of a lethal osmotic pressure, when too much lactose is transported by lactose permease).

mentalisation and information transfer [15]. The latter is further split into the major functions of the 'central dogma' of molecular biology (replication, transcription and translation) [16]. Further functions (dispensable under highly restricted conditions) should be associated which certainly contribute to fitness: editing (proof-reading) [17, 18], control (in particular, transcriptional control) [19], scaffolding (shaping) [20], addressing (this is important for all cell's components, including RNAs [21] and proteins [22]) and maintenance [23]. Metabolism, associated to compartmentalisation [24], results in the management of energy, through osmotic pressure, electrochemical gradients and electron transfers. Anabolic processes allow construction of biomass. Degradative processes recover energy, and create basal level building blocks for anabolism. Salvage pathways may be dispensable, but they could have an obvious contribution to fitness. Furthermore, when considering the analogy with man-made machines, it is clear that there is a need for cleaning up the system, as no metabolic process can be error-free. Finally, compartmentalisation is required to control diffusion, besides allowing energisation. It is also required to make a 'casing' for the machine, with three major components, an envelope, a skeleton and a variety of appendages. Compartments separate the inside from the outside, and transport is a major function in this respect (which should be split into import and export). The shaping casing will also have a protective function, and it may be split into a variety of parts, some of which devoted to storage or to specific metabolic pathways (for anabolism or catabolism) as particular nanomachines (such as ATP synthase [25] or sulphur assimilation [26]). How is this reflected in the gene products that are present in all or most of the genomes? This is the question we try to answer in the present work.

After placing in perspective what life is, using this set of extant bacterial genome sequences, we identify the core persistent proteins of bacterial life. Finally, using the knowledge derived from interactions, we propose hypotheses to account for some of those core functions which remain unknown, despite in some cases, the fact that the 3-D structure of the corresponding proteins is known.

1.1 Conceptual background

1.1.1 What life is

Material systems submitted to the trio variation/selection/amplification evolve. While evolving, they recruit and select objects performing actions which result in their stabilisation, or rather we only witness the existence of those that have recruited the proper objects to become selectively stabilised into forms that survived. This somewhat haphazard situation arises because there is no intelligent design, and therefore, no prescribed goal to the evolution of natural systems. Evolving systems usually recruit pre-existing structures [27] rather than come up with *de novo* magic constructs that fulfil the needs of the action required by the stabilisation goal. *De novo* created orphan proteins, however, might go through progressively enhanced functional properties, starting from the stabilisation function of complexes as 'gluons', while they use the intrinsic gluing capacity of aromatic amino acids [28]. All of this accounts for the somewhat disconcerting 'tinkering' aspect of most biological constructs [29].

Among such continuously evolving systems, some are endowed with life, and they display highly original features which associate together. Briefly, three major processes are required to make a living entity: (i) metabolism (ongoing chemical processes that transform molecules into other molecules: life is expressed in a dynamic state), (ii) compartmentalisation (the cell with its inside and its outside is the atom of life) and (iii) information transfer.

In what follows, we try to uncover how these processes are implemented in bacterial proteomes. At this point, it is clear that many different objects could perform the same function, suggesting that any typology should first concentrate on categorising the concept of function, before identifying the objects that perform the functions. This also shows that, looking for objects and processes common to all cells, we miss many important ones, because acquisitive evolution relies on the selection of several from different origins, depending on the organism. For this reason, we considered proteins present in a clique of proteomes ('persistent proteins') rather than proteins conserved everywhere ('conserved proteins'), the number of which will steadily decrease as new compact genomes are discovered [14].

1.1.2 Core functions of life

We here look at living organisms in exactly the same way as we consider artefacts meant to perform an action. In short, what is the 'purpose' of a living organism? The simplest answer is Shakespearean. It is to be, to survive for the longest possible time. Note that the constraints imposed by sheer existence are enormous, and more often than not, underevaluated. Existence can be fulfiled by absolute stability, as in rocks, but dynamic systems also found a way out. They found that a state of flux of chemicals could be maintained stable in time [30]. This is indeed the basis of metabolism: there is a state in between life and death, dormancy, but the spore or the seed will be said to be alive only when its metabolic flux begins to show up. It must be recognised, however, that dynamic systems need objects to be implemented, asking again either for rock stability or for continuous maintenance. Metabolism combines both constraints as it can repair eroded objects but also be constructed in such a way that it has dynamic stability in time. Once compartmentalised and able to use electron transfer gradients to manage energy [24], metabolism was discovered as an untapped source of dynamic stability based on a limited number of building blocks, while it could compensate for inevitable weathering. This resulted in the selection of a fundamental function present in most living organisms, reproduction (which is a kind of rejuvenation), with the unwanted consequence, however, that the

progeny is not its parents, so that the goal of stability is transferred throughout generations. In parallel, exploration, used initially to forage to get food as source of building blocks and energy was probably rapidly created.

Stability in time requires coping with extremely variable conditions: living systems are those dynamic material systems that worked out that to survive means to prepare for the unexpected, the unforeseeable. This is an extremely difficult constraint. It assumes that the system, somehow, is able to build an image of the world, memorise it and then compare the real world to that image, while subsequently producing a relevant behaviour. This means that, in addition to dynamic properties of the processes involving metabolites, it possesses a kind of 'cognitive' representation of itself and of its environment, to sense and to memorise its environment (Fig. 1). Remarkably, the central discovery of living material systems is that this process is best – at least, until we discover some other better means, which we certainly cannot exclude – performed by an algorithmic representation, memorised in the genome text [31]. This generic property is most likely at the origin of genomes and the genetic code, with the associated 'alphabetic' management of string of symbols [32]. Indeed, replication, transcription and translation can be visualised as generic, highly parallelised, algorithmic processes. In short, the information transfer processes present in living organisms have evolved to fulfil these particular tasks: memorising, representing and predicting. The atom of life, the cell, can then be visualised as a Turing Machine, separating the analogue part (the machine itself) from the algorithmic alphabetic (symbolic) part (the genetic programme). Thus, well before the invention of the nervous system, cells were organised around a sensory–motor system (sensing and exploring), integrated in what was to become the repository of their memory, their genome, constructing what appears to be a computer-like system.

This is a first, superficial, view. A deeper insight will ask whether the computer metaphor is relevant. Can an organism be considered as a living computer? What constraints would be needed to permit computers to make computers? Do we find in living organsims the equivalent of a machine and of a programme? How fit is the genetic programme metaphor? This has been discussed elsewhere, and it was suggested that there are good reasons to take the metaphor of the turing machine as highly significant not only in its 'read' function, but also in its 'write' function, corresponding to programmed alteration of the programme itself (which we do not further discuss here) [31]. Since there are no external users, living organisms would be computers that use their computing ability (including programmed modification of the genetic programme) to try and persist under conditions where everything tends to waste. They have created a vast array of functions to do so, and we try, in what follows, to identify the core functions involved in the process, those which are present (almost) ubiquitously (Fig. 1). Further work will deal with functions specific to a particular niche and constituting a self-consistent ensemble, organised along Kielmayer–Cuvier's correlation of forms rules, the cenome (from κοινος, community).

2 Materials and methods

2.1 Genomes

We extracted bacterial genomes from the EBI entry point of the International Nucleotide Sequence Database Collaboration (INSDC) (http://www.ebi.ac.uk/genomes/) as on 1st May 2006. We examined the conservation of the genome context in terms of persistent genes (see below). Since a limited genome size would introduce a strong bias in clustering genes together, for sheer size constraints, we put aside the 57 genomes with less than 1500 coding DNA sequences (CDSs). Most of the latter genomes correspond to obligatory endosymbionts, having therefore suffered selective pressure for genome reduction [13, 14]. The genomes of a further 15 bacteria, lacking proper 16S rRNA annotation and poorly annotated, were also excluded from the study. As the final result, we included 228 bacterial genomes into our analysis (Supplementary Table 1). In this collection, 23 genomes are made of two or three chromosomes, which we concatenated for the study. However, this particular structure of the genome may alter the overall gene organisation as some of the chromosomes may originate from a plasmid [33]. Furthermore, this collection of genomes is uneven, as some distant species are represented by a single genome while some species are represented by many instances (*e.g. Staphylococcus aureus* in our sample). In any event, the genome sample we have is heavily biased by investigators' prejudices. Possessing both very distant sequences and closely linked sequences allowed us to test for the robustness of the concept of persistent gene [3] by comparing a set with all genomes and a reduced set with only phylogenetically divergent genomes (Fig. 2). Constructing a reduced set is not straightforward. In particular, in a collection of highly related genomes (for example, different *Escherichia coli* and *Shigella* sp. strains), there is no particular criterion to decide which one should be retained. For those species with more than one strains, we therefore randomly selected one as the representative, making a core sample of 144 genomes that we used as a control for the overall study (Fig. 2 and Supplementary Table 1, grey boxes).

2.2 Persistent genes

A persistent gene refers to a gene which exists in all or most (a clique) of the genomes retained in the present study and which displays a leading replication strand preference [3]. Briefly, for each gene, we first searched its orthologues from all the genomes (an orthologue is defined using the straightforward bidirectional best hit (BBH) strategy [34], with amino acid similarity $\geq$40% and protein length difference $\leq$20%). This provided an ordered list of genes for each

Figure 2. Unrooted 16S RNA phylogenetic tree of the species used in this study.

genome. Subsequently, from each genome, we extracted a dataset made of the 400 genes with the largest number of orthologues (the selection of 400 as a relevant threshold is presented below). In order to evaluate the contribution of individual strains within a given species, we also kept a series of strains of individual species, noting that the definition of the bacterial species is fuzzy (9 for *S. aureus* and 11 for the *E. coli*/*Shigella* complex).

The genes in these 228 datasets vary from bacteria to bacteria and the contribution from individual strains of a given species affects the last 10% of a dataset, justifying the inclusion of several strains of a given species in this study (see below and Fig. 3 and Supplementary Fig. 1). We therefore retained as common persistent genes only the 332 genes which were present in more than 160 of the datasets (Supplementary Table 2). As a control we considered a sample where only one strain of a given species was retained: we obtained a highly similar gene list, substantiating the robustness of the approach (Table 1 and Supplementary Table 2).

2.3 Mutual attraction

It has been reported that highly conserved genes tend to cluster together in bacterial genomes [35]. To see whether this observation extended to persistent genes, we applied a Kuiper's test [36] to examine the coupling between gene's persistence and clustering tendency. The genes with the most BBHs were the most persistent ones. For each bacterium, genes were sorted in descending order in terms of persistence and grouped into batches of 100 genes each. In parallel, the genes coordinates in the circular chromosome were computed as angular measurements in radians. Kuiper's test uses a discrepancy statistic which was calculated as the

Table 1. Borderline persistent genes

	Added		Deleted
allB, ybbX	Subunit of allantoinase	*aroK*	Shikimate kinase
argJ	Ornithine acetyltransferase/amino acid acetyltransferase	*atpF*	ATP synthase (subunit b)
aroL	Shikimate kinase II	*clpA*	ATP-dependent Clp protease-like (class III stress gene): protein degradation
atpE	ATP synthase (subunit c)	*dnaN*	DNA polymerase III (β subunit)
bcp	Thiol peroxidase	*folA*	Dihydrofolate reductase
bioA	Subunit of adenosylmethionine-8-amino-7-oxononanoate aminotransferase	*hemA*	Subunit of glutamyl-tRNA reductase
dtd, yihZ	Subunit of D-Tyr-tRNA(Tyr) deacylase	*hflX*	'Putative GTPase; possible regulator of HflKC'
fabI	Enoyl-ACP reductase (NAD[P]H)/enoyl-ACP reductase (NADPH)/enoyl-ACP reductase (NADH)enoyl-ACP reductase (NAD[P]H)/ enoyl-ACP reductase (NADPH)/enoyl-ACP reductase (NADH)	*hupB*	DNA-binding protein HU-, NS1 (HU-1), subunit of HU
ftsE	Cell division protein FtsE	*lpd*	Subunit of dihydrolipoate dehydrogenase/ dihydrolipoamide dehydrogenase, 2-oxoglutarate dehydrogenase complex, gcv system and pyruvate dehydrogenase multienzyme complex
hisB	Imidazoleglycerol-phosphate dehydratase	*metB*	Subunit of *O*-succinylhomoserine lyase/ *O*-succinylhomoserine(thiol)-lyase
hisC	Histidine biosynthesis	*prmA*	Methylation of 50S ribosomal subunit protein L11
hom	Homoserine dehydrogenase	*rpiA*	Subunit of ribose-5-phosphate isomerase A
hpt	Guanine phosphoribosyltransferase/hypoxanthine phosphoribosyltransferase	*rpsP*	Ribosomal protein S16
ilvC	Ketol-acid reductoisomerase	*sodA*	Subunit of superoxide dismutase (Mn)
iscS, yfhO	Cysteine desulphurase	*sucC*	Succinyl-CoA synthetase, subunit, subunit of succinyl-CoA synthetase
ispB	Octaprenyl diphosphate synthase	*tatC*	Subunit of TatABCE protein export complex
ispG, gcpE	1-Hydroxy-2-methyl-2-(E)-butenyl 4-diphosphate synthase	*ycfF*	Inhibitor of cell division: control of cell division
lysA	Lysine biosynthesis *via* diaminopimelate	*yeaZ*	Hypothetical peptidase
lysC	Homoserine biosynthesis and lysine biosynthesis *via* diaminopimelate	*yfdZ*	'Aspartate aminotransferase; PLP-dependent enzyme, of unknown function; similar to YkrV in *B. subtilis*, that could be involved in diaminopimelate aminotransferase and in methionine aminotransferase'
mreB	Subunit of longitudinal peptidoglycan synthesis/chromosome segregation-directing complex	*ygiH*	'Unknown; inner membrane protein'
mrp	Putative ATPase	*yhgF*	'Unknown; transcription accessory protein involved in RNA metabolism'
nuoK	Subunit of NADH dehydrogenase I	*ynhE, sufB*	Subunit of SufB-SufC-SufD cysteine desulphurase (SufS) activator complex
phoB	PhoB-Phosphorylated transcriptional dual regulator	*yqgF, ruvX*	'Resolvase homologue, function unknown; has an RNase H-like fold'
prfC	Peptide chain release factor RF3	*yycF*	Transcription two-component response regulator
pstA	Subunit of phosphate ABC transporter	*znuB*	Subunit of ZnuA/ZnuB/ZnuC ABC transporter

Table 1. Continued

	Added		Deleted
rpmE	Ribosomal protein L31	*zwf*	Glucose 6-phosphate-1-dehydrogenase
rpmH	Ribosomal protein L34		
smc	Chromosome condensation and segregation SMC protein		
smf	'Unknown; conserved protein'		
sodB	Subunit of superoxide dismutase (Fe)		
trpC	Indole-3-glycerol phosphate synthase		
upp	Subunit of uracil phosphoribosyltransferase		
xseB	Exonuclease VII, small subunit, subunit of Exonuclease VII		
yhdE	'Unknown; conserved protein, synteny close to mreB'		
yqhS	Similar to 3-dehydroquinate dehydratase		

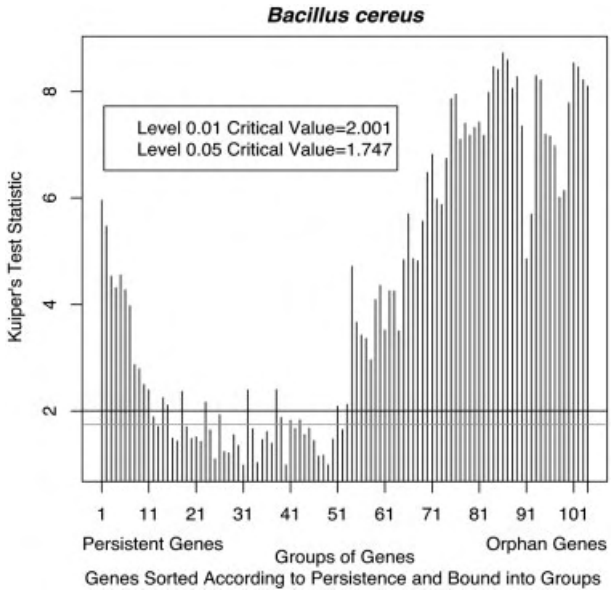

Figure 3. Association between gene persistence and clustering tendency. Kuiper's test was performed on the genes of *B. cereus* sorted according to persistence and clustered into groups of 100 genes each (see Supplementary Fig. 1 for the other genomes). Two adjacent groups were chosen so as to overlap by 50 genes. The *y*-axis is the output of the test statistic. The *x*-axis labels groups of genes according to the number of genomes in which they are present. Persistent genes [3] are on the left and genes present in a few genomes only on the right (orphan genes are those present in a single species). The red line indicates the critical statistical cut-off at the level of $\alpha = 0.01$ and the blue line at the level of $\alpha = 0.05$. Persistent genes and orphan genes are significantly clustered.

deviation of the cumulative distribution function under scrutiny with an expected cumulative uniform distribution to examine the significance of clusters. We performed Kuiper's test (CircStats R package, http://cran.r-project.org/src/contrib/Descriptions/CircStats.html) for all batches from each bacterium (see an example in Fig. 3, and for all bacteria see Supplementary Fig. 1). Interestingly, persistent genes on the one hand, and orphan or quasi-orphan genes on the other hand, were the genes which significantly clustered. Table 2 shows the summary of the significantly clustered persistent gene batches in the 144 control genomes. At the level of $\alpha = 0.05$ (Kuiper's test), in 73 (>50%) bacteria, the top 400 persistent genes are significantly clustered. We therefore extracted the top 400 persistent genes from each bacterium for further analyses.

Subsequently, we examined the distance between persistent genes in all the genomes, in an effort to find out those genes conservatively coded together in the chromosomes. The distance between two genes in one chromosome was denoted by $d_{ij} = \frac{N_{ij}}{N/2-1}$, where N_{ij} is the number of intercalated genes between gene i and gene j, and N is the total number of genes of that chromosome. A pair of genes retaining low d_{ij} values in most bacteria signifies that they are conservatively located together, as if there were forces attracting each other. The bacterial genomes available are not equidistantly distributed in the phylogenetic tree (Fig. 2). To take this inevitable fact into account, we used a 20% trimmed mean of the d_{ij} (the smallest and largest 10% d_{ij} were ruled out) acquired from all chromosomes, to measure the strength of such attractions. We named this measurement potential of mutual attraction between gene i and j (PMA_{ij}). Figure 4a shows the distribution of PMAs of all combinations of persistent genes grouped pairwise. The small peak close to 0 hinted that some pairs of genes were indeed conservatively located together. To substantiate this observation we further applied an expectation maximisation method (Mclust R package, http://www.stat.washington.edu/mclust) on the supposed mixture models. This allowed us to identify

Table 2. Summary of the results of Kuiper's test performed on 144 genomes

Group ID	Persistent genes range[a)]	No. of bacteria in which clustering in this group is significant	
		$\alpha = 0.01$	$\alpha = 0.05$
1	1–100	136	140
2	51–150	103	124
3	101–200	85	104
4	151–250	72	98
5	201–300	71	89
6	251–350	61	84
7	301–400	49	73
8	351–450	35	57
9	401–500	20	46
10	451–550	19	41
11	501–600	19	34
12	551–650	16	33
13	601–700	9	27
14	651–750	13	31
15	701–800	11	32
16	751–850	8	22
17	801–900	6	17

a) Genes were sorted in descending order in terms of persistence. The first group included the top 100 persistent genes from each bacterium, and the second group included the 51st–150th persistent genes, and so on.

unambiguously two classes of PMAs. Figure 4b shows the distribution of genes in Class A in which PMAs were all very small. Figure 4c shows the distribution of genes in Class B. This latter class fits the normal distribution quite well, which is confirmed by the normal quantile plot for Class B shown in Fig. 4d. As a consequence, we retained genes located three times the SD away of the mean of PMAs in Class B as our threshold (PMA$<$0.25) to define mutually attracted gene pairs (MAGP). In this way, we defined 1140 pairs of MAGP (Supplementary Table 3).

A deeper statistically validated approach to mutual attraction will be presented elsewhere (Fang, Rocha, Vergassola and Danchin, unpublished). The network of mutually attracted genes has been constructed using the Cytoscape algorithm [37].

3 Results

3.1 Persistent processes in extant bacteria

Analysis of persistent processes results in a list of functions that are highly connected to the list of functions we retained as necessary for life (Fig. 5 and Supplementary Table 2). Because in our sample some species are over-represented and some genomes are split into two or three chromosomes, we replicated this analysis with a set of 144 non-redundant spe-

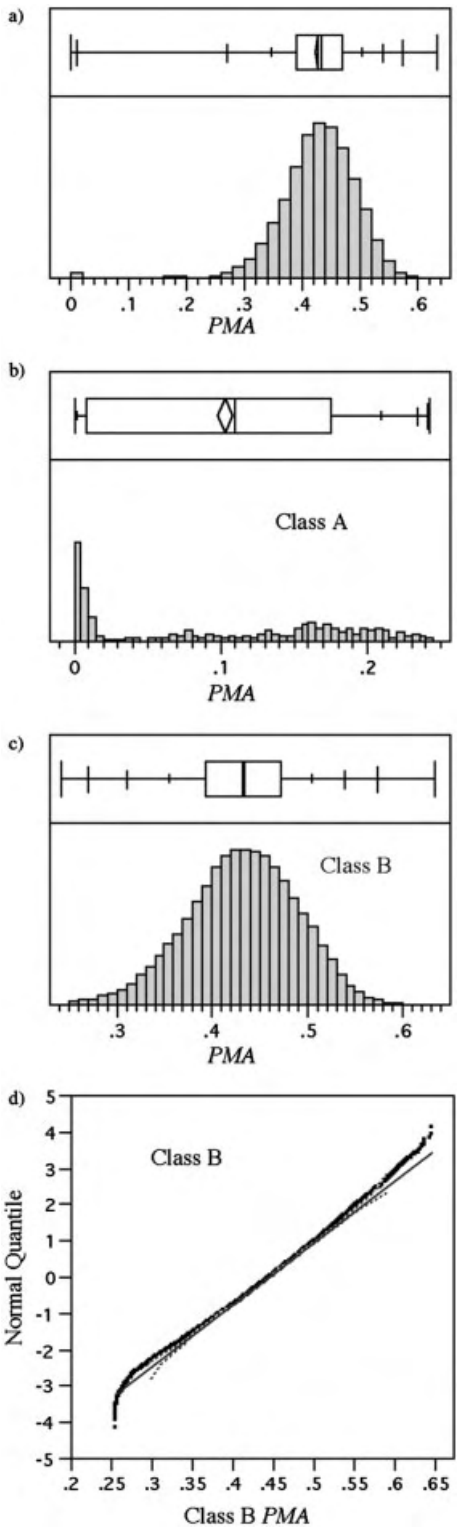

Figure 4. Distribution of potential for mutual attraction. (a) The distribution of PMAs is computed from the total of 53 956 pairs of persistent genes. (b, c) The whole set of PMAs was evaluated and separated into two classes. (d) Displays the normal quantile plot showing that Class B follows a normal distribution. The mean value of PMA in Class B is 0.43 and the SD is 0.06.

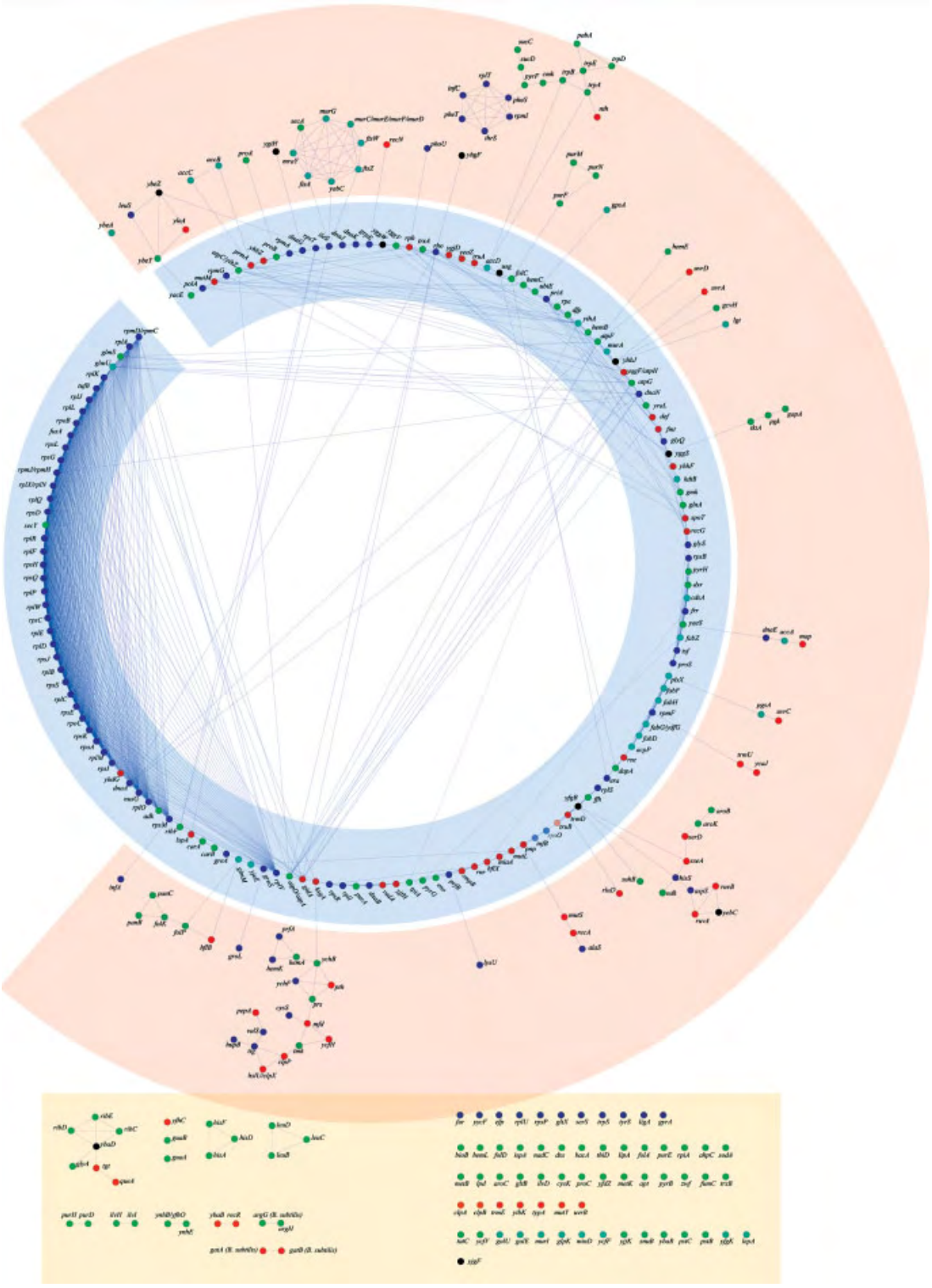

Figure 5. The network of mutual attraction in bacterial proteomes. Persistent proteins have been characterised from 228 complete genome sequences, and the conserved proximity of their genes in the genome has been measured, allowing creation of a mutual attraction index. Proteins related by mutual attraction are linked in a graph organised using the Cytoscape software. Functions are colourised as indicated. Note that the colour clustering supports functional clustering. Consistent groups of connectivity are shown as three circles: the inner, blue one is mostly organised around information transfer and RNA; the pink, middle one is organised around tRNA synthetases; and the discontinuous, beige rectangle, is organised mostly around biosynthesis of the major metabolic building blocks.

cies: only minor changes were observed (Table 2, and see Section 2). In the figure, the functions have been labelled using a colour code directly related to the function classification presented in Fig. 1, and they are connected using the Cytoscape software [37] to construct a network of 'mutually attracted' proteins in the proteome. The network can be analysed first according to the function types of Fig. 1, and then according to the way they are interconnected.

Remarkably, we find the following. There is a noteworthy absence of sensing and exploration (Fig. 1), except in the case of one unique signal transduction component, the regulator YycF [38], specific to the large class of monoderm organisms [39]. Consistent with this observation, YycF does not display strong mutual attraction with any other protein. Moreover, it is no longer present in the reduced 144 genomes set, showing that its persistence is borderline (Table 1). Unexpectedly, this protein has been shown to be essential for *B. subtilis* growth in rich media [9], and the present work hints that it should be considered in priority in experimental approaches, looking for a deep functional role. There is evidence that it plays a role in the control of membrane construction, which would couple its involvement in the sensing process (it is a member of a two-components system) and compartmentalisation [40]. In general, there is a significant lack of membrane components in the conserved proteome, except for the very basic components of the inner membrane and of those involved in cell division. This can be ascribed to two causes. On the one hand, bacteria split into several families that differ considerably in terms of their membrane structures: diderms have two membranes while monoderms have only one [39]. While we had taken this feature into consideration in our first analysis of persistent genes [3], we restricted the present study to functions present in all the bacteria. On the other hand, the amino acid 'vocabulary' for membrane proteins is more restricted in its usage than the one associated to cytoplasmic proteins, resulting often in difficulties to resolve the question of orthology for that class of proteins.

As expected, information transfer is represented by core proteins of the translation machinery, the transcription machinery and the replication machinery. These processes constitute the highly connected inner circular network of mutually attracted proteins. This core network is organised around RNA metabolism, as we recognise the ribosome, a couple of tRNA synthetases, translation factors, tmRNA-mediated proof-reading, the RNA-associated secretion machinery (Ffh), transcription factors and nucleotide-mediated translation-transcription control (SpoT) (note that this is one of the very few elements belonging to the 'control' category, suggesting that it might have another important metabolic function). Some components are borderline and either appear, or disappear when one compares the list of persistent genes from the 228 and 144 sample (Table 1 and Supplementary Table 2). We further find elements of the replication machinery, in particular, in DNA replication priming and Okasaki fragments-dependent replication (Fig. 5). tRNA synthetases also make the first outer circle (see below) and four of them (GluX, SerS, TrpS and TyrS) belong to the singleton class (Supplementary Table 2 and Fig. 5). Overall 16 tRNA synthetases are persistent, the missing ones being ArgS, AsnS, GlnS and MetS (but see Section 4 for the latter).

Interestingly, this makes the transition with intermediary metabolism, while most essential components of the synthesis of nucleotides belong to the core network of persistent proteins, most of the proteins needed to construct the building blocks of the memory of the organism, its chromosome, are absent from the picture. Indeed, the enzymes constructing deoxyribonucleotides, ribonucleotide reductases and thymidylate synthases, have been recruited several times in the course of evolution [41, 42]. However, thymidylate kinase is present. Another interesting connection between intermediary metabolism and information transfer is provided by homeotopic transformation, *i.e.* modification of residues carried by a transfer RNA molecule (formylation of initiator methionine, amidation of glutamate into glutamine, reduction of glutamyl-tRNA as a first step of heme biosynthesis), which may represent a very ancient trace of the coupling between metabolism and the RNA world [15, 43]. To this latter category, we might add the category of RNA and protein modifications. It is remarkable that RNA modifications (in particular, tRNA modifications [44]) are considerably conserved, while most of the protein modifications are not (Fig. 5 and Supplementary Table 2).

Intermediary metabolism is centred around the construction of major building blocks, as well as that of coenzymes and prosthetic groups. It is mostly arranged in an outer connecting circle and in singletons. Individual groups are connected in a metabolically significant way (see for example, connection between folate and purine biosynthesis, or paraminobenzoate and tryptophane synthesis). Not all steps and not all essential building blocks are represented, which may correspond either to the recruitment of a variety of different enzymes for the same catalytic activity, to fast evolution (as we can see in genes present in the 144 or 228 genome samples, Table 1 and Supplementary Table 2) or even to the involvement of RNA in the form of ribozymes, in some of the activities. It is therefore revealing that the recently uncovered non-mevalonate isoprene biosynthetic pathway, starting from deoxyxylulose-5-phosphate, is present in many of its steps (Dxr, Dxs, IspA, YaeS(IspU), YchB(IspE)) [45]. This pathway is essential to construct the quinones required for electron transfers, and it may also be involved in the synthesis of iron chelators [46].

A considerable part of metabolism, significantly expressed in the second circle, is devoted to compartmentalisation, in particular, envelope construction, *via* synthesis of phospholipids, and *via* synthesis of the murein sacculus. A self-contained compartmentalisation process, associated to replication, is devoted to cell division. Surprisingly, it is connected to the inner network *via* genes assumed to code for isoleucine tRNA synthetase (but see Section 4), in particular, *via* the AdoMet-dependent methyltransferase YabC(MraW) [47], suggesting the involvement of RNA in compartmentalisa-

tion and division [15]. The metabolic pathways expressed at that level involve biosynthesis of aromatic amino acids, synthesis of coenzymes, synthesis of lipids and the core of carbon metabolism. It is worth noting that many of these steps appear to be attracted to tRNA synthetases, as we find nine of those enzymes at this level: Leu, Phe, Thr, His, Asp, Ala, Lys, Cys and Val. We also find several systems meant to clean up the cell or participate in DNA recombination and repair. Interestingly, AlaS, RecA and MutS connect to polynucleotide phosphorylase (Pnp) a core member of the degradosome involved in the synthesis of deoxyribonucleotides [48]. Two major processes associated to compartmentalisation, protein export and transport of small compounds are represented. However, the latter is extremely limited in number: this is the result of our procedure for identifying persistent proteins, which puts aside those which make very large families of paralogues, with a strong tendency to diverge during evolution. This methodological constraint makes particularly interesting those transporters that have been conserved, as this shows that conservation is very strong: the representatives of this class transport phosphate and divalent metals.

In addition to anabolic processes resulting in macromolecular syntheses, transport and construction of the cell's envelope, one observes a significant number of proteins involved in scaffolding (molecular chaperones) and maintenance (RNA degradation and proteolysis). The category of small molecules degradation and cleaning is also represented, in particular, in the degradation of modified nucleotides. However, whereas in whole genomes the category of catabolism is a major one, only a very limited set is persistent (with some elements involved in adaptation to ROS).

Finally, we notice that there are a few proteins of unknown function, for which, in many cases, a 3-D structure is known but for which we do not have a fair idea of the function. This observation makes those proteins of particular interest as their presence suggests that some very important functions have until now been overlooked.

3.2 From RNA metabolism to paleometabolism

As shown in Fig. 5, another way to see the network is to summarise how it is connected: one may see the network of mutual attraction of core persistent protein genes as organised as three layers, with the external network highly fragmented (and poorly connected), while the internal network is highly connected and arranged around an RNA-centred metabolism, with the ribosome as its major component. This latter circle comprises the core processes of the cell's life, translation, transcription and replication. RNA modification and degradation (the degradosome) also belong to that network.

A second, fragmented, layer is connected to the inner one *via* individual proteins. Quite interestingly, this layer is also functionally associated to RNA, especially tRNA, as nine tRNA synthetases play a major role in the organisation of the fragments present at that level, as well as peptidyl-hydrolase and several tRNA modification enzymes. Remarkably, there is also more emphasis on the role of the cell envelope and of proteins at that level, as many of the core components of cell envelope construction and division belong there, as well as the ATP-dependent proteolysis system (a primitive proteasome). DNA recombination and repair processes are also in majority found at that level.

Finally, we observe a series of much less connected mutually attracted clusters that could be considered as forming an outer circle which have lost or failed to create its connection in the course of evolution. This outer layer is made of small clusters and of singletons. The majority of the proteins making this outer circle is made of metabolic enzymes, involved in purine/histidine biosynthesis, riboflavin coenzymes synthesis and branched-chain amino acids biosynthesis and a cluster of proteins of recently characterised function YnhB/YfhO, YnhE (Suf, components of cysteine desulphurases [49, 50]). Four tRNA synthetases are singletons, including GluX, which can be related to the interesting presence of the major homeotopic modification found in most firmicutes (amidation of the glutamyl group of glutamyl-tRNA^{Gln} [51]), synthesis of the heme precursor aminolevulinate [52, 53] and tRNA modification (*e.g.* QueA, Tgt). This external network may also be considered as on the same level as GidA, probably involved in the biosynthesis of the hypermodified nucleotide 5-methylaminomethyl-2-thiouridine, also acting on tRNA holding molecules [15]. In the same way, two subunits of ATP synthase (AtpA and AtpD) may be considered as connecting the outer circle with the inner network. The other subunits are in the twilight zone that fluctuates between the study involving 144 pr 228 genomes (Table 1). Finally, there is a cluster of DNA repair proteins (RecR and YbaB, which may be involved in uracil repair [54]), as well as singletons in the same functional class (MutY, UvrB). This fragmented layer is also the one where most of the variations from the set of proteins identified using 144 or 228 genomes appear (Supplementary Table 2), consistent with the idea that this represents a set of proteins that have enjoyed a considerable time for evolution.

4 Discussion

The availability of many bacterial genome sequences allows investigation of the core proteome that may be essential to sustain life. However, it is unlikely that analysis of the simple combination of orthologues known as 'conserved' proteins will result in any constructive interpretation as we expect, because of the old age of life, genetic takeover has taken place [55], leading to gene recruitment and function replacement in many individual genomes. Hence, we might thus be faced, considering extant life, with the situation of a palimpsest where a variety of novel recruits have replaced the older actors [4], with the number of orthologues belonging to all species getting lower and lower as we get to know more and more genome sequences. We therefore considered another

approach, retaining a clique of orthologous genes, to construct the core class of persistent genes in bacterial genomes [3] and to study the way in which they functionally interact, using conservation of their proximity in the chromosome (mutual attraction) as a marker of their possible functional interactions. It is important to note that while direct protein–protein interactions [56, 57], or coregulation within transcriptional units [58], may be the cause of mutual attraction (this would require, however, a rational scenario to explain how they could uncover that they would need to interact or to be cotranscribed), we did not assume that those would be the only causes. Hence, the mutual attraction we uncovered can have a variety of causes, including persistence of a contingent association that might have existed since the origin of the paleocells, randomly fusing and splitting, and have resulted in the cells we know in extant life [59].

In the present work, following Cuvier, we have sought to organise persistent genes into functional classes, with the idea that we should consider function first, before analysing the structure of the objects that have been recruited to fulfil the function. This view implies that there is a certain 'correlation of forms' between the objects that have been recruited in the course of evolution, allowing us to understand the underlying phylogenetic raison d'être of the objects of interest. Having characterised persistent proteins, and their overall functions (Fig. 1 and Supplementary Table 2), we created a measure of their historical and functional relationships as their propensity to cluster together in a majority of genomes. Using the Cytoscape software to organise the corresponding network of interactions, we observed that the overall network could be seen as comprised of three concentric interconnected circles, the central one being almost continuous (Fig. 5). The three constitutive processes making life, metabolism, compartmentalisation and information transfer, are present in the network. Remarkably, the very fact that all three processes are present in most of their components indicates that it has generally not been possible, during the 3.8 billion years of life, to recruit objects other than those which are still extant to perform these essential functions.

The most fascinating feature displayed in Fig. 5 is that the overall network appears to be organised mostly around RNA, including RNA synthesis and degradation. DNA metabolism, and, in particular, error-prone DNA polymerases are, for example, missing from the picture. We find that the inner circle is organised around translation, with the ribosome, many of its ribosomal proteins and three tRNA synthetases: Ile, Gly and Pro, while a second circle is mainly organised around a further nine tRNA synthetases. Two of the four missing tRNA synthetases (AsnS and GlnS) can be accounted for by the persistence of homeotopic transformation, followed by recruitment of new proteins to fulfil the function of specific tRNA synthetases [15] and indeed proteins fulfiling this function (GatA and GatB) are present in the outer, more primitive, layer. The unexpected absence of MetS can be ascribed to the hypothesis that methionine either has been involved in translation recently or has been displaced by isoleucine when oxygen invaded the Earth's atmosphere: AUG is in the AUN isoleucine box, and IleS and MetS are highly related to such a point that misassignment is extremely likely is some genomes [60]. The consequence is that the category IleS/MetS should probably be considered as a unique one. In this respect, it may be revealing that the product labelled as IleS is closely connected to an enzyme using AdoMet as a substrate. The missing ArgS is of great interest: arginine is particularly rich in nitrogen, an often neglected element essential for the origin of life, and it could be that its involvement has followed a pathway that differs from the other amino acids. The behaviour of the *argS* gene does not follow the trend of the other tRNA synthetases genes in firmicutes [57]. Finally, bits and pieces of metabolism, directly connected to nucleotides metabolism (and hence to RNA) belong to that circle, which also comprises repair, degradation and maintenance. The prevalence of RNA is such that we could even propose that some of the metabolic activities that are not present in the picture could have been performed (and perhaps are still performed) by ribozymes [61].

Considering this picture, one cannot but see a set of concentric circles, with the circles progressively fading out as one goes from the centre to the outside. This would be consistent with an inverse order in evolution, where the outer broken circle, being the most ancient one, is progressively taken over by new functions [55], or, perhaps progressively associated to the discovery of the role of template by RNAs leading to what became genes in the most recent developments of the evolution of cells. And one cannot escape visualising this organisation as a trace of the evolution predating the apparition of life as we know it, with metabolism first (including synthesis of coenzymes) [15, 62, 63], followed by a prototype of the RNA world, organised around tRNA-like molecules that would be carriers of the metabolic intermediates (ancestors of tRNA synthetases would be of major importance there) [15], to end up with the translation machinery, and the invention of replication, solving the chicken and egg paradox of the relationship between proteins and nucleic acids [59]. An intriguing observation, which relates to the structure of the network, is that class I and class II tRNA synthetases [64] do not distribute evenly in the circles, with the outer circle comprising a majority of class I enzymes, while the middle circle has mostly class II synthetases. This would suggest that class I tRNA synthetases predated the second class.

The functions present in the evanescent outer circle are quite interesting, as they are highly connected to metabolism of amino acids, nucleotides, coenzymes, sulphur metabolism and iron-related metabolism [62, 63]. Several are associated to the construction of an envelope. Heme biosynthesis connects the circles together, and comprises one step of homeotopic transformation, involving tRNA (HemA), which is at the limit of persistence as the synthesis of aminolevulinate recruited a non-tRNA-dependent pathway in many organisms [65]. The persistent functions that

are still unknown appear to be often related to electron transfer (Usg, YfgB, YggW), they also appear to involve iron and/or sulphur (YfgB, YggW, YjgF), and some may be related to RNA metabolism, degradation in particular (YbeZ, YhgF, YjgF).

The conserved proteome is a summary of the proteins in common to free-living bacteria, which have not been replaced with novel proteins during the course of evolution. It is expected to shrink as new genomes are sequenced as the process of gene recruitment is ubiquitous. To circumvent this effect, we have taken a functional approach and identified the core proteome as that made of persistent proteins, that are present in a clique of proteomes, while displaying particular features such as preferred location of their corresponding genes in the leading strand of the genome [3]. By contrast, because every organism is specific for a particular niche, the proteins that would be a landmark of the occupation of a niche (the cenome) will obviously be absent from the persistent set. Interestingly, as shown in Fig. 3 and Supplementary Fig. 1, the genes making the cenome, and therefore family-, genus- or species-specific are also clustered together, as are persistent genes. However, it does not follow that any particular essential function (displayed in Fig. 1) should be missing from the persistent list. Remarkably, while we observed several proteins involved in DNA maintenance and repair (*i.e.* slackening the pace of evolution) we did not observe proteins that would enhance the speed of evolution, such as DNA polymerases IV (DinB) and V (UmuCD). These enzymes would correspond to the 'write' function of a Turing machine and are expected to be important. We therefore released our criteria for the definition of persistent genes, retaining the first 500 persistent genes instead of 400: both proteins now appear to be present (data not shown). This stresses a limitation of our approach: because of the stringency of our criteria to define persistent genes we are probably missing a significant number of genes making the core genome because they evolved rapidly. These genes are among those present in the twilight zone in the middle of the graphs shown in Fig. 3 and Supplementary Fig. 1. The region is probably too noisy at this point to permit us to draw further firm conclusions.

The lack of persistent proteins involved either in the sensory–motor behaviour of cells or in specific transport strongly suggests that a variety of different objects have been systematically recruited in the course of evolution to fulfil that particularly important class of membrane-associated functions, despite the fact that all need to be connected to the three processes of life, represented as conserved persistent proteins. Furthermore, the split of Bacteria into monoderms (with one membrane) and diderms (with two membranes) [39] has certainly considerably blurred the picture of membrane proteins evolution: further work should explore separately those categories to identify relevant features of membrane proteins. This observation underscores that the couple organism/environment is the system submitted to selection pressure. It also suggests that exploration of new niches, with concomitant recruitment of novel proteins, is a major driving force in genome evolution. This will be the subject of a further study.

5 Conclusions: Ockham's razor, RNA turnover and iron metabolism

The most prominent feature of the creation of a list of persistent proteins using the genomes presently available is that they are organised in a consistent network coded by mutually attracted genes. Furthermore, this network is organised in a way that is consistent with the idea that extant living organisms are not as distant from the origin of life as one might have suspected, with an evolution starting from carbon, nitrogen and sulphur assimilation, going to compartmentalisation organised around tRNA-like molecules, and ending with the creation of the three essential processes of information transfer, translation, transcription and replication. Functions involved in proof-reading, maintenance, and cleaning and repair are also core functions, with some indication that molecular oxygen has played a significant role in the destructuring/structuring of the network, in particular, its outer borders, involved in small molecules metabolism. This underscores that the general functions that permitted the transition from the chemical world to the full development of life (sensing, exploration, replication/memory) have been later discoveries.

While one could always postulate the existence of important unknown functions that were not recorded in the persistent proteins list, it is good practice to follow Ockham's motto: *numquam ponenda est pluralitas sine necessitate* ('Plurality ought never be posed without necessity') and to try to make sense of most of the data we collected with a minimum set of hypotheses. Because persistent proteins belong to the vast majority of proteomes, it would have been expected that most would be known. Interestingly, many proteins of unknown or very recently characterised function belong to the list, and it is remarkable that they could often be proposed to be associated to RNA metabolism (recombination, modification, folding and degradation in particular) or iron chelation and electron transfers. A very simple way to put this together would be first to assume that DNA replication and recombination was preceded by a step of RNA replication and recombination, and that some relic of that step is still extant. This is not that far-fetched if one notes the extant weird replication initiation of DNA-lagging strand replication, which still involves an RNA primer. And second, noting that dioxygen was the first major pollutant of the Earth's atmosphere, we could assume that proteins involved in RNA metabolism used ferrous iron as a major cofactor. This divalent ion has a coordination sphere similar to that of magnesium, but has a different preference for ligands, with a significant preference for nitrogen, promoting easy interaction between proteins and RNA. The formation of cells, with a considerable control over their inside oxido-reduction poten-

tial allowing to maintain it significantly reducing, could have preserved some of that role of ferrous iron *in vivo*. By contrast, this physicochemical constraint should have precluded the construction of many biochemical experiments as we perform them today, by forbidding the opening of cells in the presence of atmospheric oxygen (ferrous iron becomes ferric in an extremely short time), if we were to keep the machinery intact. The present work suggests that some biochemical experiments in the absence of oxygen should be performed to assess the extent of that constraint in extant organisms.

A final observation comes from this work. The very fact that the core proteome of bacteria is organised in a highly non-random way must impact on the future of what is known as synthetic biology, where there is a need to develop robust concepts and a dedicated language to deal with and categorise biological parts, in order to construct novel and controllable cell types. We hope that the functional approach presented here will help to work in this direction.

This work is the result of three decades of discussion with many persons, often associated in the Stanislas Noria network (http://www.pasteur.fr/recherche/unites/REG/causeries.html). In silico experiments have been supported by the BioSapiens European Network of Excellence, grant LSHG-CT-2003-503265, the ACI IMPBIO Blastsets programme and the EuroPathoGenomics European Network of Excellence, grant LSHB-CT-2005-512061.

6 References

[1] Danchin, A., *Mol. Microbiol.* 1995, *18*, 371–376.

[2] Chaigneau, S. E., Barsalou, L. W., Sloman, S. A., *J. Exp. Psychol. Gen.* 2004, *133*, 601–625.

[3] Fang, G., Rocha, E., Danchin, A., *Mol. Biol. Evol.* 2005, *22*, 2147–2156.

[4] Benner, S., Allemann, R., Ellington, A., Ge, L. *et al.*, *Cold Spring Harb. Symp. Quant. Biol.* 1987, *52*, 53–63.

[5] Danchin, A., *The Delphic Boat. What Genomes Tell Us*, Harvard University Press, Cambridge, MA, USA 2003.

[6] Danchin, A., in: Goffeau, A. (Ed.), *Complete Genome Sequencing: Future and Prospects*, pp. 1–24, *Sequencing the Yeast Genome. A detailed assessment. BAP* 1988–1989. Commission of the European Communities, Brussels 1989.

[7] Sanger, F., Coulson, A. R., Hong, G. F., Hill, D. F. *et al.*, *J. Mol. Biol.* 1982, *162*, 729–773.

[8] Mushegian, A. R., Koonin, E. V., *Proc. Natl. Acad. Sci. USA* 1996, *93*, 10268–10273.

[9] Kobayashi, K., Ehrlich, S. D., Albertini, A., Amati, G. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 4678–4683.

[10] Baba, T., Ara, T., Hasegawa, M., Takai, Y. *et al.*, *Mol. Syst. Biol.* 2006, *2*, doi: 10.1038/msb4100050.

[11] Gil, R., Silva, F. J., Pereto, J., Moya, A., *Microbiol. Mol. Biol. Rev.* 2004, *68*, 518–537.

[12] Glass, J. I., Assad-Garcia, N., Alperovich, N., Yooseph, S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2006, *103*, 425–430.

[13] Klasson, L., Andersson, S. G., *Trends Microbiol.* 2004, *12*, 37–43.

[14] Perez-Brocal, V., Gil, R., Ramos, S., Lamelas, A. *et al.*, *Science* 2006, *314*, 312–313.

[15] Danchin, A., *Prog. Biophys. Mol. Biol.* 1989, *54*, 81–86.

[16] Crick, F., *Nature* 1970, *227*, 561–563.

[17] Hopfield, J. J., *Proc. Natl. Acad. Sci. USA* 1974, *71*, 4135–4139.

[18] Ninio, J., *Biochimie* 1975, *57*, 587–595.

[19] de Lorenzo, V., Perez-Martin, J., *Mol. Microbiol.* 1996, *19*, 1177–1184.

[20] Ellis, R. J., *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 1993, *339*, 257–261.

[21] Condeelis, J., Singer, R. H., *Biol. Cell* 2005, *97*, 97–110.

[22] Emanuelsson, O., Nielsen, H., Brunak, S., von Heijne, G., *J. Mol. Biol.* 2000, *300*, 1005–1016.

[23] Nystrom, T., *Mol. Microbiol.* 2004, *54*, 855–862.

[24] Mitchell, P., *Eur. J. Biochem.* 1979, *95*, 1–20.

[25] Capaldi, R. A., Aggeler, R., *Trends Biochem. Sci.* 2002, *27*, 154–160.

[26] Wei, J., Tang, Q. X., Varlamova, O., Roche, C. *et al.*, *Biochemistry* 2002, *41*, 8493–8498.

[27] Thompson, L. W., Krawiec, S., *J. Bacteriol.* 1983, *154*, 1027–1031.

[28] Pascal, G., Médigue, C., Danchin, A., *Proteins* 2005, *60*, 27–35.

[29] Jacob, F., *The Logic of Life (Translation of La Logique du Vivant)*, Princeton University Press, Princeton, New Jersey, USA 1993 (1970).

[30] Delbrück, M., in: Lwoff, A. (Ed.), *Discussion of "Influence des Gènes, des Plasmagènes et du Milieu dans le Déterminisme des Caractères Antigéniques chez Paramecium Aurelia (Variété 4) by TM Sonneborn and GH Beale"*, pp. 33–35. Unités Biologiques Douées de Continuité Génétique. Centre National de la Recherche Scientifique, Paris 1949.

[31] Danchin, A., *ComPlexUs* 2004/2005, 61–70.

[32] Rocha, L. M., Hordijk, W., *Artif. Life* 2005, *11*, 189–214.

[33] Medigue, C., Krin, E., Pascal, G., Barbe, V. *et al.*, *Genome Res.* 2005, *15*, 1325–1335.

[34] Tatusov, R. L., Koonin, E. V., Lipman, D. J., *Science* 1997, *278*, 631–637.

[35] Martin, M. J., Herrero, J., Mateos, A., Dopazo, J., *Genome Res.* 2003, *13*, 991–998.

[36] Jammalamadaka, S., SenGupta, A., *Topics in Circular Statistics*, World Scientific, River Edge, USA; Singapore 2001.

[37] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S. *et al.*, *Genome Res.* 2003, *13*, 2498–2504.

[38] Ng, W. L., Tsui, H. C., Winkler, M. E., *J. Bacteriol.* 2005, *187*, 7444–7459.

[39] Gupta, R. S., *Crit. Rev. Microbiol.* 2000, *26*, 111–131.

[40] Mohedano, M. L., Overweg, K., de la Fuente, A., Reuter, M. *et al.*, *J. Bacteriol.* 2005, *187*, 2357–2367.

[41] Myllykallio, H., Lipowski, G., Leduc, D., Filee, J. *et al.*, *Science* 2002, *297*, 105–107.

[42] Jordan, A., Reichard, P., *Annu. Rev. Biochem.* 1998, *67*, 71–98.

[43] Wong, J. T., *Proc. Natl. Acad. Sci. USA* 1975, *72*, 1909–1912.

[44] Dunin-Horkawicz, S., Czerwoniec, A., Gajda, M. J., Feder, M. *et al., Nucleic Acids Res.* 2006, *34*, D145–D149.

[45] Eisenreich, W., Bacher, A., Arigoni, D., Rohdich, F., *Cell. Mol. Life Sci.* 2004, *61*, 1401–1426.

[46] Buss, K., Muller, R., Dahm, C., Gaitatzis, N. *et al., Biochim. Biophys. Acta* 2001, *1522*, 151–157.

[47] Carrion, M., Gomez, M. J., Merchante-Schubert, R., Dongarra, S. *et al., Biochimie* 1999, *81*, 879–888.

[48] Carpousis, A. J., *Biochem. Soc. Trans.* 2002, *30*, 150–155.

[49] Mueller, E. G., *Nat. Chem. Biol.* 2006, *2*, 185–194.

[50] Mihara, H., Esaki, N., *Appl. Microbiol. Biotechnol.* 2002, *60*, 12–23.

[51] Tumbula, D. L., Becker, H. D., Chang, W. Z., Soll, D., *Nature* 2000, *407*, 106–110.

[52] Schulze, J. O., Schubert, W. D., Moser, J., Jahn, D. *et al., J. Mol. Biol.* 2006, *358*, 1212–1220.

[53] Levican, G., Katz, A., Valenzuela, P., Soll, D. *et al., FEBS Lett.* 2005, *579*, 6383–6387.

[54] Lim, K., Tempczyk, A., Parsons, J. F., Bonander, N. *et al., Proteins* 2003, *50*, 375–379.

[55] Cairns-Smith, A., *Genetic Takeover and the Mineral Origin of Life*, Cambridge University Press, Cambridge, UK 1982.

[56] Dandekar, T., Snel, B., Huynen, M., Bork, P., *Trends Biochem. Sci.* 1998, *23*, 324–328.

[57] Nitschke, P., Guerdoux-Jamet, P., Chiapello, H., Faroux, G. *et al., FEMS Microbiol. Rev.* 1998, *22*, 207–227.

[58] de Daruvar, A., Collado-Vides, J., Valencia, A., *J. Mol. Evol.* 2002, *55*, 211–221.

[59] Danchin, A., *L'Oeuf et la Poule. Histoires du Code Génétique*, Fayard (translated into Japanese, and Portuguese Relogio d'Agua, Lisbon), Paris 1983.

[60] Sekowska, A., Denervaud, V., Ashida, H., Michoud, K. *et al., BMC Microbiol.* 2004, *4*, 9.

[61] Brackett, D. M., Dieckmann, T., *Chembiochem* 2006, *7*, 839–843.

[62] Granick, S., *Ann. N. Y. Acad. Sci.* 1957, *69*, 292–308.

[63] Wachtershauser, G., *Microbiol. Rev.* 1988, *52*, 452–484.

[64] Eriani, G., Delarue, M., Poch, O., Gangloff, J. *et al., Nature* 1990, *347*, 203–206.

[65] Atteia, A., van Lis, R., Beale, S. I., *Eukaryot. Cell* 2005, *4*, 2087–2097.

