Beilstein-Institut

Systems Chemistry, May 26th – 30th, 2008, Bozen, Italy

Frustration: Physico-chemical Prerequisites for the Construction of a Synthetic Cell

Antoine Danchin* **and Agnieszka Sekowska**

Genetics of Bacterial Genomes – CNRS URA2171,
Institut Pasteur, 8 rue du Docteur Roux, 75724 Paris, France

E-Mail: *antoine.danchin@normalesup.org

Received: 7th August 2008 / Published: 16th March 2009

Abstract

To construct a synthetic cell we need to understand the rules that permit life. A central idea in modern biology is that in addition to the four entities making reality, matter, energy, space and time, a fifth one, information, plays a central role. As a consequence of this central importance of the management of information, the bacterial cell is organised as a Turing machine, where the machine, with its compartments defining an inside and an outside and its metabolism, reads and expresses the genetic program carried by the genome. This highly abstract organisation is implemented using concrete objects and dynamics, and this is at the cost of repeated incompatibilities (frustration), which need to be sorted out by appropriate «patches». After describing the organisation of the genome into the paleome (sustaining and propagating life) and the cenome (permitting life in context), we describe some chemical hurdles that the cell as to cope with, ending with the specific case of the methionine salvage pathway.

An Introduction: Three Revolutions and the Birth of Synthetic (Symplectic) Biology

The past century witnessed a remarkable development of biological sciences, split between three overlapping revolutions. The 1944 – 1985 period saw the formation and development of molecular biology, with the creation of all the central concepts of modern biology, combining biochemistry and genetics, and decyphering the rules of replication and of gene expression with its regulation. The advent of gene sequencing technologies in 1975 permitted access to the exact text of genes. Subsequently, the years 1985 – 2005 saw a fascinating development of genomics with the central discovery (in 1991 at a EU meeting in Elounda, in Crete, but published later [1, 2]) that, contrary to expectation, a huge number of genes was of unknown function. At the time of writing (July 22nd, 2008), there is 3,887 ongoing genome sequencing projects, 833 already completed, (mostly from microbes, among which 686 of Bacteria, more or less correctly annotated) and 209,035,780,490 nucleotides are registered at the International Nucleotide Sequence Database Collaboration (INSDC). In this collection, 10% correspond to the core genome (“persistent” genes), while 40 – 50% coding DNA sequences (CDSs) do not correspond to known functions, showing that we lack understanding of a considerable fraction of what makes a living system.

As time elapsed the importance of the relationships between the objects of life – not necessarily the objects themselves – was recognized as absolutely central to any attempt to understand biological processes [3]. Nevertheless, the most recent avatar of molecular biology, Synthetic Biology, was launched initially as an engineering (and teaching) attempt meant to explore the following question:

> *“Can simple biological systems be built from standard, interchangeable parts and operated in living cells? Or is biology simply too complicated to be engineered in this way?”*
> (iGEM home page: http://parts2.mit.edu/wiki/index.php/About_iGEM).

This technological aim tries to class and normalise “biobricks”, basic components of living organisms, while the most important conceptual aim of Synthetic Biology is to reconstruct life, in an endeavour to explore whether we understand what life is and learn missing entities (physical objects and dynamic processes) from our failures. Keeping the abstract laws defining life, a second aim (sometimes named “orthogonal” synthetic biology) tries to reconstruct artificial living systems made from objects of a physico-chemical nature differing from that of the building blocks of extant life [4].

In all these aspects of Synthetic Biology, dynamic processes and rules of interactions appear to be essential, so that we should rather think of this new area of science as *Symplectic* Biology – from the Greek equivalent of the latin *complexus*, meaning ‘to weave together’ – which would combine the efforts of systems biology with engineering biology [4]. A further (sociological) argument to prefer the latter name is to avoid the fuzzy (and self-contra-

dictory) connotations associated to the word ''complexity''. Another reason to prefer a term that does not have strong connotations outside biology is that we need to avoid the confusion that plagues the understanding by the general public of the construction of genetically modified organisms. A connotation in geometry (''symplectic geometry'' is a lively domain of mathematics) will not interfere. Constructing a synthetic cell requires to understand what life is. In what follows I try to point out features that will be essential to take into account to make a cell *de novo*.

What Life is

Since the time when Schrödinger proposed his famous metaphor of the ''aperiodic crystal'' major discoveries accumulated that result today in a way to consider life as the association of a machine and of a program. The machine, which expresses the program, is made of a casing that defines an inside and an outside and drives exchanges within and without. It is also a chassis constraining the form of the living organism, with the cell as its ''atom''.

Compartmentalisation is essential to life and there are two major scenarios associated to this process. Either the cell is made of one single entity, encased in a more or less complex envelope (this corresponds to the domain prokaryotes), or the organism multiplies membranes and skins, even at the cell level, which comprises a nucleus and a variety of organelles (this corresponds to the domain eukaryotes).

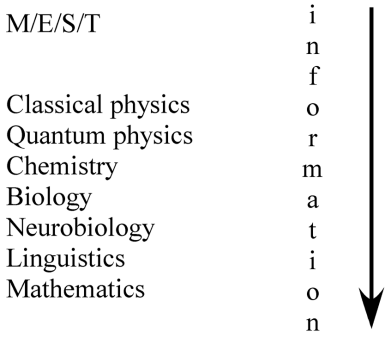
The machine also organizes chemical processes – *metabolism* – that build up, salvage and turn over all the required elements making the cell as well as the energy needed to make it work. Metabolic activities are at the root of the reproduction process, which preserves the relationships between pathways, in time and space but not necessarily in their ultimate details. At least 800 small molecules, assimilating C, H, N, O, S, P in the presence of specific ions (note that the role of iron is probably underestimated, as ferrous iron oxidizes extremely rapidly in the presence of dioxygen, and then precipitates in neutral or alkaline water [5]) are involved in the building up of the biomass. Energy is managed via the turnover of ATP and electron transfers. While the number of basic building blocks is small, many investigators tend to forget the importance of co-factors (co-enzymes and prosthetic groups), that are present generally at quite low concentrations but are essential for life. In this respect it is amusing to remark that most studies claiming to work on the origin of life forget about cofactors.

Associated to the machine is a program, involved in processes which may be collectively summarized as ''*information transfers*''. It acts as a book of recipes, or, following the common metaphor, as a blueprint. A noteworthy feature of these information transfers is that they are *recursive*, using a code, a cypher, that permits one level of information to be translated into another level, the latter permitting synthesis of objects that can manipulate the program which encoded them. Life is therefore witnessing one of those exceptionally rich

"strange loops" (as they were recognised and named by Douglas Hofstadter [6]), which were used by Kurt Gödel, coding axioms and definitions of arithmetic as integers, to demonstrate the incompleteness of arithmetic.

A remarkable feature of this separation between the machine and the program is that it leads one to distinguish between *reproduction* and *replication*. While the latter inevitably accumulates errors [7, 8], the former can improve over time [9]. This is witnessed by the remarkable, but unobtrusive paradox that it is always an aged organism which give birth to young ones [10]. Hence, living organisms have an in-built capacity to generate information.

While the word "information" is currently used in biology, its meaning is not accurately defined [11]. This widespread use nevertheless emphasises the need to add a fifth entity to the four entities considered in classical physics to account for Reality, matter, energy, space and time, which are associated together in the remarkably concise equation proposed by Einstein, $\mathbf{E = mc^2}$. While not compatible with classical physics, Heisenberg's indeterminacy principle, $\Delta \mathbf{x}\ \Delta \mathbf{p} \geq \mathbf{h}/4\pi$, introduces information via "lack of information". In a nutshell, I contend that we are at the dawn of a new era in natural sciences, where *information* will play an ever increasing role as we will better understand and model the concept. The core of our future exploration will be to try and understand how information is articulated with matter, energy, space and time. This view implies a considerable change in the placing of biology in the Auguste Comte's hierarchy of sciences, according to increase in information, and progressively less influence of matter, energy, space and time:



With this view, biology is strongly linked to mathematics, and it needs to be perceived essentially as an information-related science. This also indicates that we are in considerable need, at present, to develop further views of what information is. Claude Shannon has investigated the constraints operating on communication of information, not on information itself [3, 11], and many further views have been developed, along a path which is certainly very preliminary but already quite rich conceptually [12, 13].

Computing

In the cell, information transfer is organized by the genetic program. If we take seriously the view just outlined, this process is much more than a metaphor: do we have the conceptual tools to push it to its ultimate consequences? Let us consider what computing is. As demonstrated by Alan Turing and many others [14 – 16], two entities are required to permit computing organised as a *machine* able to read and write a *program* on a physical support. The program is split by the human mind (not conceptually!) into two entities, the program itself (providing the "goal", in our anthropocentric view) and the data (providing the context). An essential point in this description is that the machine is physically distinct from the data/program and can be separated from it. Another point, which is not discussed here is that what we name "program" is *declarative* ("I am here", is enough to start running the program) not *prescriptive*.

Can we see cells as computers, or, asked otherwise, is the genetic program *separated* from the cell's machinery? At least four lines of evidence argue in favour of this view:

- Horizontal gene transfer is extremely widespread. In bacteria, it corresponds often to at least one fifth of the genome setup [17]. This indicates that the cell machinery can "understand" (i. e. read and express) a huge number of genes present in the environment. As a matter of fact, for a given bacterial species (with the caveat that "species" is difficult to define in the case of bacteria), the number of genes that can be horizontally transferred greatly outnumbers the average number of genes present in a given strain. For *Escherichia coli*, for example, taking into account the sequences of published strains, the number of genes that differ from strain to strain is already larger than 20,000, and this number keeps increasing as new strains' genomes are sequenced, while the average number of genes in any strain of this organism is slightly higher than 4,000.
- Viruses behave as pieces of program with a casing allowing them to recognize the machine they will parasite, and a process for coding for their own replication. In this case the metaphor went the other way around: computer scientists will speak about computer viruses, and this is a correct way to describe these invading, often noxious, pieces of programs. As in the case of biological viruses, computer viruses can not only replicate, but they can also carry information loads they extracted from previous infectious cycles. One notes that with this definition a virus is not living (it lacks the machine, and in particular the whole recursive translation machinery, even when it carries genes extracted from a variety of cells and coding for some functions involved in information transfers).
- A further way toward the "computing automaton" view of the cell is the process of genetic engineering. Here, not only do we have cases where genes are artificially associated together, but it is current practice to get DNA sequence pieces

that are entirely synthesised from scratch, after purposeful design (this is the only instance of real intelligent design...).

- Finally, the most interesting experiment demonstrating that the program is separated from the machine is the direct transplantation of a naked genome into a recipient cell with subsequent change of the recipient machine into a new one corresponding to the transplanted DNA [18] (Figure 1).

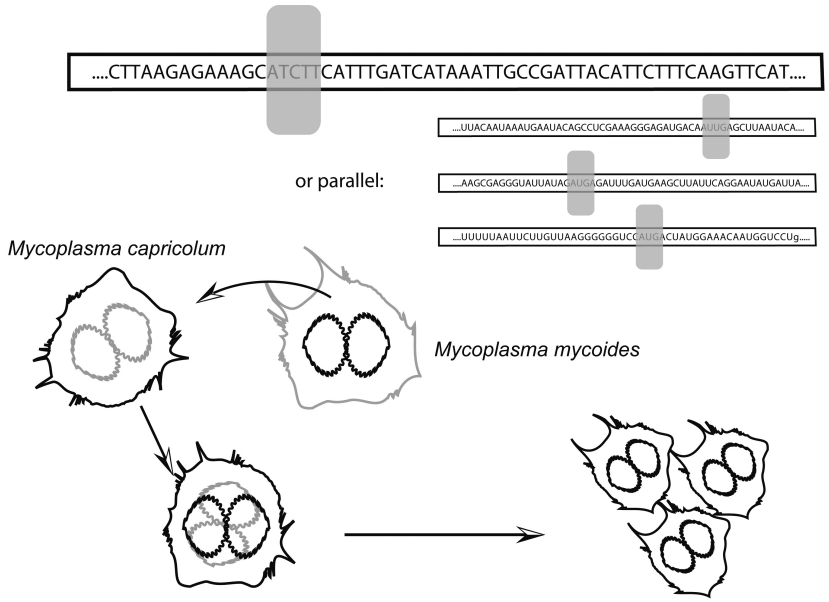


Figure 1. The Turing Machine and an experiment of chromosome transplantation. The Universal Turing Machine head reads and write on a linear string of symbols. Specific Turing Machines can work with parallel pieces of program (here illustrated in the case of protein translation starts). In a transplantation experiment DNA from *Mycoplasma mycoides* is transplanted in *M. capricolum* under selective conditions. The resulting colony is typical of *M. mycoides* [18].

All these observations point to an obvious separation between a ''machine'' (the cell factory) and ''data/program'' (the genome). This provides a convincing background to analyse the way information is transferred in living processes.

The Universal Turing Machine works on a program made of one linear string of symbols. Turing has further shown that this is equivalent to a machine with a parallel setup, where several pieces of program could run in parallel. The organisation of information transfer in the cell is more of the latter type, when many pieces of DNA are translated into proteins, for example. Parallel information processing requires coordination, or a clock. In general, biological information transfers are algorithmic in nature. Replication, transcription as well

as translation display a high parallelism, always expressed along the same pattern: ''Begin, control check-points, repeat, end''. The information transfer action is oriented, with a beginning and an end. Curiously, the processes of time dependent control (check-points, or clocks) are rarely taken into account (except for the replication/division processes [19]), but their role is essential to allow the coordination of multiple actions in parallel. This is a first prediction of the model of the cell-as-a-computer: it should prompt investigators to construct experiments to identify check-points in the processes of transcription and translation. Some experiments suggest that they do exist [20, 21].

A Map of the Cell in the Chromosome?

John von Neumann, trying to understand the functioning of the brain, suggested that, were a computer both to behave as a computer and to construct the machine itself, it should keep somewhere an image of the machine [16]. The metaphor does not appear to apply to the brain, does it apply to the cell? Linking a geometric program to the information of the genetic program may seem farfetched.

However we have one – unexplained – example of such a link. The homeogenes found in insects follow an order that exactly matches that of the segments of the insect. The comparison between insect and crustacean substantiates this observation: Geoffroy Saint-Hilaire in the middle of the 19th century showed that the body plan of crustacea was reversed under the thorax (the abdomen becomes the back and vice versa) as compared to that of insects, and this triggered a bitter controversy. This has now been proven and backed by the observation that modification of homeogenes between insects and crustacean affects their body plan [22, 23]. The same is true for vertebrates, where four sets of corresponding homeogenes also match the organisation of the adult organism.

We thus have the equivalent of the *homunculus* of preformists, but not as a full tiny organism, but, rather, as the algorithm for the construction of the organism. Can we think of a ''celluloculus''? Stated otherwise, is there an image of the cell in the genome? An analysis of the *mur-fts* clusters by Tamames and co-workers suggests that this may be as follows: a tree built up following the way the corresponding genes distribute in different bacterial genomes parallels the bacterial shape variations, not the 16S phylogenetic tree [24]. All this points to the need to explore the points of contact between the information setup and the material setup of living organisms.

Frustration of DNA Structures

The organisation just described is conceptual, it deals with immaterial information, while it needs to be implemented concretely, within the matter/energy/space/time dimensions. However, concrete objects have often properties that are not compatible with those of other

objects. This implies "frustration" of possible mutually exclusive entities (because of constraints in space or energy states) [25]. The cell factory will therefore require construction of appropriate "patches" to cope with these incompatibilities.

As a first example, the cell-as-a-computer model requires check-points for parallel gene expression, and this introduces a need for regulation (which may be seen as an important constraint at the origin of the creation of the various regulation systems that are pervasive in biology). This results in a large number of mutually exclusive constraints which are typical of a ubiquitous type of frustration, and explains why it is often so difficult to sort out transcriptional controls mediated by different factors interacting with the same promoter region.

A second physico-chemical constraint derives from the fact that the program needs to be physically separated from the machine. Interestingly, this particular feature matches a common objection raised against the model of the cell-as-acomputer: in living cells, it is not possible to completely separate between the hardware and the software. However, the objection cannot be retained as a strong one, as the same holds true for real computers. Indeed, these machines cannot be purely abstract entities either, but are very concrete entities. They run programs, but any program needs a physical support. For example it can be stored on a CD, and a CD is deformable, by heat for example. When deformed, and despite the fact that the program it carries is unaltered, the laser beam that is used to read it will not be able to do so, and the program will no longer be usable by the computer (Figure 2). This does not alter the very existence of either the computer or the abstract laws establishing what a computer is (a Turing Machine) but this tells us that in any concrete implementation of the Turing Machine, one cannot completely separate between the hardware and the software. This observation points to an important constraint that may explain the somewhat surprising lack of a transplantation experiment in the recent synthesis of an artificial *Mycoplasma* genome [26].

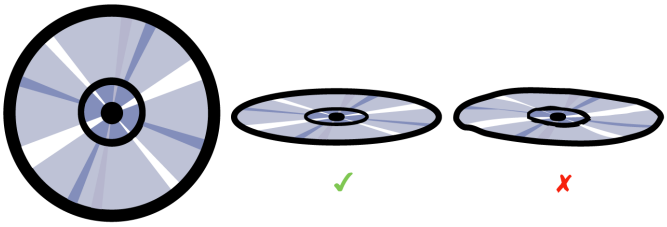


Figure 2. A computer's program must be carried by a physical support. Here, a deformed CD can no longer start a computer.

Still another constraint results from the dissymmetry of replication, conservative on the leading strand and semi-conservative on the lagging strand. This lack of symmetry implies that mutational errors and efficiency of repair will differentially affect the nucleotide and

gene composition of both strands of the double helix [27]. As a matter of fact genes that are found to be essential in the laboratory are systematically located in the leading replication strand [28]. This has considerable consequences in the amino acid composition of proteins [29]. In short, material implementation of a Turing Machine requires a variety of specific adjustments to manage material and temporal incompatibilities.

Frustration in Proteins Build-up

The major effectors of cell metabolism are proteins. Their activity usually requires functional interactions, and it is expected (and observed) that many proteins form complexes. Furthermore, translation appears to organise the chromosome structure, with specific islands corresponding to particular codon usage biases [21]. The consequence is that the amino acid composition of proteins cannot be random, and indeed there is a large bias in amino acid distribution among the different proteins making a proteome. A multivariate analysis of the proteome (correspondence analysis) of a large number of prokaryotes showed that proteins are grouped into clusters comprising a similar distribution in particular amino acids. A strong bias opposes charged residues to hydrophobic residues and permits one to identify with remarkable precision the protein located in the inner membrane of the cell (IIMPs) [30]. Two further biases, apparently universal, characterise the bacterial proteomes. There is a bias, perhaps not unexpected, created by the G+C composition of the genome, and another one, driven by the aromatic composition of the proteins. Interestingly, aromatic-rich proteins are most often without recognised function. This group is also highly enriched in ‘‘orphan’’ proteins [31]. An explanation to this observation is that proteins created *de novo* might indeed go through progressively enhanced functional properties, starting from the general function of stabilising complexes by acting as ‘‘gluons’’, where they use the intrinsic stickiness of aromatic amino acids [30].

A further bias appeared in proteins coded by psychrophilic organisms. Indeed their proteome is systematically enriched in asparagine, while the dioxygensensitive amino acids, cysteine, histidine and methionine are counter-selected [32, 33]. This bias corresponds to intrinsic properties of asparagine, which isomerises easily, leading to perhaps the major post-translational modification in all proteomes. Asparagine spontaneously isomerises in particular contexts into isoaspartate, with concomitant deamidation in a reaction which is still poorly understood. This reaction affects protein structures (and may affect their function). It may also have a role in regulating protein folding and it is a signal for degradation of intracellular proteins [34]. Aspartate and asparagine isomerisation is therefore another physico-chemical constraint that needs to be dealt with using appropriate metabolic patches. In many organisms (including *Escherichia coli* and *Homo sapiens*) there exists a process that can restore aspartate from isoaspartate after methylation and demethylation, an extremely costly repair system [35].

This observation leads us to revisit the inevitability of ageing. Indeed, be it only because of asparagine/aspartate isomerisation, proteins age, sometimes very fast (e. g. ribosomal protein S 11 from *E. coli*, within minutes at 37 °C [36]). As a consequence, it is always an aged cell (or multicellular organism) that gives birth to a young one. This implies that in the process of forming a progeny, there is creation of information. We need therefore to identify the genes acting in the process of accumulating information [10].

Revisiting Information

A natural way to consider information is to appreciate its ''value''. This implies intuitively that one will need energy to create information. This was indeed the common view until Rolf Landauer showed in 1961 that creation of information is reversible and therefore does not require any energy [37, 38]. This remarkable work, curiously widely ignored, showed however that reversibility was at a cost: an enormous amount of time or space was required to permit reversible creation of information. Hence, creation of information could only be tolerated if a process existed that permitted to ''make room'' for novel information to be further created. By contrast with reversibility of the creation of information, this process required consumption of energy.

In this context of physics, improvement of metabolism over time is therefore not an impossibility. It can be at least conceptually tolerated, as creation of information is reversible. However, in order to proceed efficiently, the corresponding process will require a specific process to ''make room'': how is this obtained? Can we identify in genomes the genes coding for the functions required to put this process into action?

In order to proceed with this investigation, which assumes the existence of a fairly ubiquitous process, we need to look for ubiquitous functions. However, with genome studies we have only direct access to sequences (and sometimes structures), while ''acquisitive evolution'' systematically masks functional persistence. Briefly, any system submitted to the trio *variation*/*selection*/*amplification* will evolve, as it will open windows for novel functions (note that this is creation of information). Functions however can only exist when a concrete object is recruited, so that many objects will fulfil a given function [39, 40]. The consequence is that it is not possible to identify the presumably ubiquitous genes which would correspond to ubiquitous functions, simply because they will not be ubiquitous.

From Functional Ubiquity to Gene Persistence

To sum up, functional ubiquity does not imply structural ubiquity. Fortunately, however, living organisms evolve by descent, and efficient objects tend to persist through time because their genes will tend to be conserved over generations. Briefly, there is some kind of stickiness in the adaptation of an object to a particular function. Hence, rather than look for ubiquity, we should look for ''persistence'', i. e. for the tendency of a gene to be present

in a given number of genomes. And looking for persistence will permit us to identify ubiquitous functions. We need to note here that any approach to this quest will heavily depend on the genome sample we possess, making it a fairly difficult enterprise. As in the quest for consensuses in sequences, we expect that the sampling bias will go through a maximum when the number of genomes increases, and then slowly decrease as more genomes are available (the exception makes the rule) [41]. Appropriate computing techniques can be set up to deal with this problem and it is possible to find out persistent genes from the present collection of genome sequences.

With this view, a set of 400 – 500 genes has been identified, that persist in bacterial genomes and, as expected, the vast majority of the genes labelled as ''essential'' (because they cannot be inactivated without complete loss of viability) belonged to this set [42]. Is, then, ''persistent'' a synonym of ''essential''? A remarkable feature of both categories of genes is that, as in the case of genes identified as essential in the laboratory, most persistent genes are located in the leading replication strand, suggesting that they respond to common selection pressures. In terms of functions the ~250 essential genes code for the bulk of the functions involved in information transfers. The functions in this list are not unexpected, as the list could be established very early on, and was indeed at the root of the interest of the European Commission for sequencing genomes, for example [43]. A list established using the most degenerate autonomous organisms also resulted later in a similar number [44].

The category of non essential persistent genes is interesting, both because it was not predicted by the latter studies, and because it is very much biased in particular functions. It codes for functions involved in stress, maintenance and repair, on the one hand, and for a few metabolic patches, in particular for serine degradation into pyruvate [42], on the other hand. An important feature of this particular set is that it codes for functions that may have a consequence in the long term, and have not, therefore, been studied properly under laboratory conditions. Indeed, studies investigating essentiality have just tested the capacity of mutants to grow and to generate a colony after individual gene inactivation.

Clustering of Persistent Genes

The location of persistent genes in the leading DNA strand (as does the subcategory of essential genes) combines with another specific feature in their organisation in genomes: they tend to cluster together. Using 228 genomes comprising more than 1,500 genes (to avoid sampling biases) and accurate annotations, we identified genes that tend to remain close to one another. This ''mutual attraction'' constructed a remarkable network made of three layers, building networks with differing connectivities. These layers can be grouped into a consistent picture in relation to the functional properties of the genes they are made of. A first network, made of genes coding for the construction of the building blocks of intermediary metabolism (nucleotides and coenzymes, lipids), is highly fragmented.

A second network is built around class I tRNA synthetases, and a third network, almost continuous, is organised around genes coding for the ribosome, for transcription and replication and for other functions managing information transfers [5].

How can we account for this clustering? Based on the observation that the gene flow mediated by horizontal gene transfer (HGT) must be high, we proposed that a purely passive clustering process is at work in genomes, noting that what has been interpreted as causes (co-transcription and formation of operons, and protein-protein interactions) are, instead, consequences of clustering. If genes are deleted as local bundles (to compensate for bundles of genes introduced by HGT), this results in a purely passive gene clustering, as the progeny of cells with clustering of the most important genes for sustaining life (persistent genes) are more likely to survive than that of cells with genes uniformly spread in the genome (uniform distribution is the largest deviation from clustering) [45].

A noteworthy feature of this organisation is that it emphasises the separation between metabolism and replication and is consistent with the scenario:

$$\text{Building blocks} \Rightarrow \text{nucleotides} \Rightarrow \text{tRNA} \Rightarrow \text{ribosome} \Rightarrow \text{DNA}$$

which is highly reminiscent of what could have happened at the origin of life. To better understand its meaning, let us briefly explore a mineral scenario for the origin of life [46, 47].

The surface of charged solids (e.g. pyrite (Fe-S) [48]) selects and compartimentalises charged molecules; this first step forms some amino acids, the main coenzymes, fatty acids and ribonucleotides; polymerisation with elimination of water molecules increases entropy and is therefore favoured on surfaces. Subsequently (once the nucleotides have been created) compartmentalised metabolism creates surface substitutes via polymerisation of ribonucleotides in the presence of peptides, with ancestors of tRNA (the RNA world), via a shift of role, from that of a *substrate* to that of a *template*. Then, RNAs develops its template role in translation with the invention of the genetic code, placing the ribosome as the core structure of nascent life. Finally, it further shifts away from its role as a substrate for metabolic reactions, to template for self-replication, discovering the complementarity law. Nucleic acids are further stabilised by the invention of deoxyribonucleotides, at the time when the rules controlling information transfer are discovered, first within the RNA world where vesicles carrying the ancestors of genes split and fuse randomly, before formation of the first genomes.

The Paleome and the Cenome

Coming back to the organisation of bacterial genomes, we now can see them as composed of two major parts. Persistent genes, with this scenario reminiscent of the origin of life, form the *paleome* (from παλαφοζ ancient) with genes coding for the basic functions permitting cells to survive and to perpetuate life. Bacteria need also to occupy a particular environment.

In 1877, Karl Möbius referred to the common pool of living species in a particular environment as a *biocenose* (see e. g. [49]). While the concept of gene did not exist at the time, we need now to relate this idea to that of the genes permitting the cell to occupy an ecological niche. These genes are acquired by HGT from a large unknown pool of genes and, as a consequence of the corresponding gene transfer processes (transformation, transduction, integration of prophages and conjugation), they are generally coming in genomes as gene clusters [50]. This very large class, the *cenome* (after κοφνοζ, common, as in biocenose [5]) tends to comprise novel members in different strains of the same species (Figure 3). Taking into account the concept of *pan-genome*, which puts together all the genes of a given species [51], the cenome of a given species is a subset of the pan-genome, comprising all the genes permitting any strain of that species to live in its favoured niche. As stated above, in a species such as *E. coli* the pan-genome is mostly made of genes forming the cenome of each individual strain, and is already larger than 20,000 genes with no sign of levelling off as new strain genomes are sequenced.

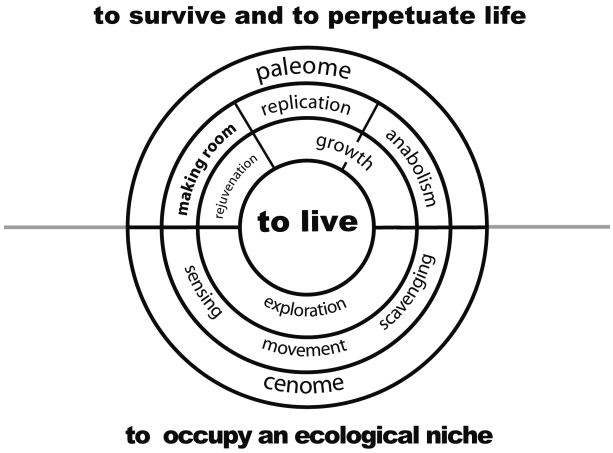


Figure 3. The paleome and the cenome. The paleome codes for functions necessary for survival and for propagation of life. Among its non-essential part one finds genes coding for metabolic patches meant to remedy conditions of metabolic incompatibilities (frustration). Such patches can also be found in the cenome as they are necessary when a pathway produces highly reactive molecules such as alpha-diketones.

Provisional Conclusion: a Split Paleome

To live associates three major processes: to survive ageing processes, to perpetuate life while already aged, and to live in a particular context. The first two processes require presumably ubiquitous functions, which have been grouped in the set we named the paleome, because of the way it organises relative to a large number of bacterial genomes. Most of the functions in the paleome have been identified, at least at the biochemical level. They are often (as in the case of transcription and translation processes) the target of drugs that prevent propagation

of the relevant organisms. While these functions are conserved, they are often not resulting from similar structures, so that they can only be identified via analysis of gene persistence. The functions of the paleome may be split following a variety of specific characters. For example, half of its genes code for functions that are essential to permit formation of a colony on plates supplemented by rich medium: they are the functions of essential genes [52, 53]; the other half, while ubiquitous, does not have this property [42]. This latter half comprises mostly genes that are essential to perpetuate life, but are not essential in the short term [10]. Another split identifies functions which solve some of the metabolic incompatibilities in the cell, resulting from chemical constraints such as spontaneous isomerisation of aspartate and asparagine. This phenomenon of frustration is necessarily quite widespread, as a large number of chemical intermediates, such as alpha-diketones, are extremely reactive towards amino groups. This explains why the downstream section of the methionine salvage pathway, which recycles the methylthioadenosine formed from a variety of reactions derived from S-adenosyl-methionine, is highly variable [54]. The functions in this pathway, typical of what is found in the cenome of an organism, has only been solved in the case of reactions involving dioxygen, while it is certainly present in anaerobic organisms (Figure 4). Its very existence examplifies the type of unknown functions we should look for when exploring the cenome of organisms, in particular in metagenomic studies. This will require considerable imagination for the prediction of novel chemical reactions.

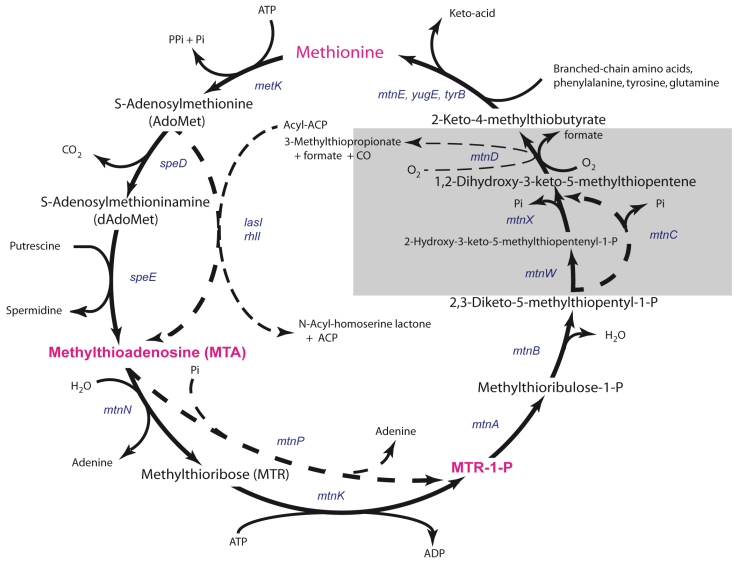


Figure 4. Present knowledge of the methionine salvage pathway. The pathway has been decyphered in [54]. The shaded region corresponds to the reactions downstream from the highly reactive 2,3-Diketo-5- methylthiopentyl-1-phosphate. They are chemically completely different in different organisms (MtnW is a RuBisCO-like enzyme with no relation with the MtnC enolase-phosphatase).

Acknowledgements

This work benefited from many years of continuous discussions with the Stanislas Noria group. Support for *in silico* analyses and experiments came from the PROBACTYS programme, grant CT-2006 – 029104 in an effort to define genes essential for the construction of a synthetic cell and the BioSapiens Network of Excellence, grant LSHG CT-2003 – 503265.

References

[1] Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., Aigle, M., Alberghina, L., Alexandraki, D., Antoine, G., Anwar, R., Ballesta, J.P., Benit, P., *et al.* (1992) The complete DNA sequence of yeast chromosome III. *Nature* **357**:38 – 46.

[2] Glaser, P., Kunst, F., Arnaud, M., Coudart, M.P., Gonzales, W., Hullo, M.F., Ionescu, M., Lubochinsky, B., Marcelino, L., Moszer, I., Presecan, E., Santana, M., Schneider, E., Schweizer, J., Vertes, A., Rapoport, G., Danchin, A. (1993) *Bacillus subtilis* genome project: cloning and sequencing of the 97 kb region from 325 degrees to 333 degrees. *Mol. Microbiol.* **10**:371 – 384.

[3] Danchin, A. (2003). *The Delphic boat. What genomes tell us*. Trans. A. Quayle. Harvard University Press, Cambridge (Mass, USA).

[4] de Lorenzo, V., Danchin, A. (2008) Synthetic biology: discovering new worlds and new words. The new and not so new aspects of this emerging research fields. *EMBO Reports* **9**:822 – 827.

[5] Danchin, A., Fang, G., Noria, S. (2007) The extant core bacterial proteome is an archive of the origin of life. *Proteomics* **7**:875 – 889.

[6] Hofstadter, D. (1979). "Gödel, Escher, Bach: an Eternal Golden Braid". Basic Books, New York.

[7] Muller, H. (1932) Some genetic aspects of sex. *The American Naturalist* **66**:118 – 128.

[8] Orgel, L. (1963) The maintenance of the accuracy of protein synthesis and its relevance to aging. *Proc. Natl. Acad. Sci. U.S.A.* **49**:517 – 521.

[9] Dyson, F. J. (1985). *Origins of life*. Cambridge University Press, Cambridge, UK.

[10] Danchin, A. (2008) Natural Selection and Immortality. *Biogerontology (submitted)*.

[11] Cover, T., Thomas, J. (1991). *Elements of information theory.* Wiley, New York.

[12] Bennett, C. (1988) Logical Depth and Physical Complexity. In *The Universal Turing Machine: a Half-Century Survey* (R. Herken, ed.), pp. 227 – 257. Oxford University Press, Oxford.

[13] Danchin, A. (1996) On genomes and cosmologies. In *Integrative Approaches to Molecular Biology* (J. Collado-Vides, B. Magasanik, T. Smith, eds.), pp. 91 – 111. The MIT Press, Cambridge (USA).

[14] Turing, A. (1936 – 1937) On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* **42**:230 – 265.

[15] Turing, A. (1946 (1986)) A. M. Turing's ACE Report of 1946 and Other Papers. *In* Charles Babbage Institute reprint series for the History of Computing (B. Carpenter, R. Doran, eds.), Vol. 10. MIT Press, Cambridge (Mass).

[16] von Neumann, J. (1958 (reprinted 1979)). *The Computer and the Brain*. Yale University Press, New Haven.

[17] Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., Danchin, A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J. Mol. Biol.* **222**:851 – 856.

[18] Lartigue, C., Glass, J.I., Alperovich, N., Pieper, R., Parmar, P.P., Hutchison, C.A., 3rd, Smith, H.O., Venter, J.C. (2007) Genome transplantation in bacteria: changing one species to another. *Science* **317**:632 – 638.

[19] Bussiere, D.E., Bastia, D. (1999) Termination of DNA replication of bacterial and plasmid chromosomes. *Mol. Microbiol.* **31**:1611 – 1618.

[20] Thanaraj, T.A., Argos, P. (1996) Ribosome-mediated translational pause and protein domain organization. *Protein Sci.* **5**:1594 – 1612.

[21] Bailly-Bechet, M., Danchin, A., Iqbal, M., Marsili, M., Vergassola, M. (2006) Codon usage domains over bacterial chromosomes. *PLoS Comput Biol* **2**:e37.

[22] Averof, M., Akam, M. (1995) Hox genes and the diversification of insect and crustacean body plans. *Nature* **376**:420 – 423.

[23] Averof, M. (1997) Arthropod evolution: same Hox genes, different body plans. *Curr. Biol.* **7**:R634 – 636.

[24] Tamames, J., Gonzalez-Moreno, M., Mingorance, J., Valencia, A., Vicente, M. (2001) Bringing gene order into bacterial shape. *Trends Genet.* **17**:124 – 126.

[25] Kitao, A., Yonekura, K., Maki-Yonekura, S., Samatey, F.A., Imada, K., Namba, K., Go, N. (2006) Switch interactions control energy frustration and multiple flagellar filament structures. *Proc. Natl. Acad. Sci. U.S.A.* **103**:4894 – 4899.

[26] Gibson, D.G., Benders, G.A., Andrews-Pfannkoch, C., Denisova, E.A., Baden-Tillson, H., Zaveri, J., Stockwell, T.B., Brownley, A., Thomas, D.W., Algire, M. A., Merryman, C., Young, L., Noskov, V.N., Glass, J.I., Venter, J.C., Hutchison, C.A., 3 rd, Smith, H.O. (2008) Complete chemical synthesis, assembly, and cloning of a *Mycoplasma genitalium* genome. *Science* **319**:1215 – 1220.

[27] Rocha, E., Danchin, A., Viari, A. (1999) Universal replication biases in bacteria. *Mol. Microbiol.* **32**:11 – 16.

[28] Rocha, E., Danchin, A. (2003) Gene essentiality determines chromosome organisation in bacteria. *Nucleic Acids Res.* **31**:6570 – 6577.

[29] Rocha, E.P., Danchin, A. (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol. Biol. Evol.* **21**:108 – 116.

[30] Pascal, G., Médigue, C., Danchin, A. (2005) Universal biases in protein composition of model prokaryotes. *Proteins* **60**:27 – 35.

[31] Pascal, G., Médigue, C., Danchin, A. (2006) Persistent biases in the amino acid composition of prokaryotic proteins. *Bioessays* **28**:726 – 738.

[32] Riley, M., Staley, J.T., Danchin, A., Wang, T.Z., Brettin, T.S., Hauser, L.J., Land, M.L., Thompson, L.S. (2008) Genomics of an extreme psychrophile, *Psychromonas ingrahamii. BMC Genomics* **9**:210.

[33] Médigue, C., Krin, E., Pascal, G., Barbe, V., Bernsel, A., Bertin, P.N., Cheung, F., Cruveiller, S., D'Amico, S., Duilio, A., Fang, G., Feller, G., Ho, C., Mangenot, S., Marino, G., Nilsson, J., Parrilli, E., Rocha, E.P., Rouy, Z., Sekowska, A., Tutino, M.L., Vallenet, D., von Heijne, G., Danchin, A. (2005) Coping with cold: the genome of the versatile marine Antarctica bacterium *Pseudoalteromonas haloplanktis* TAC 125. *Genome Res.* **15**: 1325 – 1335.

[34] Shimizu, T., Matsuoka, Y., Shirasawa, T. (2005) Biological significance of isoaspartate and its repair system. *Biol. Pharm. Bull.* **28**:1590 – 1596.

[35] Clarke, S. (2003) Aging as war between chemical and biochemical processes: protein methylation and the recognition of age-damaged proteins for repair. *Ageing Res. Rev.* **2**:263 – 285.

[36] David, C.L., Keener, J., Aswad, D.W. (1999) Isoaspartate in ribosomal protein S 11 of *Escherichia coli*. *J. Bacteriol.* **181**:2872 – 2877.

[37] Landauer, R. (1961) Irreversibility and heat generation in the computing process. *IBM Journal of research and development* **3**:184 – 191.

[38] Bennett, C. (1988) Notes on the history of reversible computation. *IBM Journal of research and development* **44**:270 – 277.

[39] Thompson, L.W., Krawiec, S. (1983) Acquisitive evolution of ribitol dehydrogenase in *Klebsiella pneumoniae*. *J. Bacteriol.* **154**:1027 – 1031.

[40] Ashida, H., Danchin, A., Yokota, A. (2005) Was photosynthetic RuBisCO recruited by acquisitive evolution from RuBisCO-like proteins involved in sulphur metabolism? *Res. Microbiol.* **156**:611 – 618.

[41] Hénaut, A., Danchin, A. (1996) Analysis and predictions from *Escherichia coli* sequences or *E. coli in silico*. In *Escherichia coli and Salmonella, Cellular and Molecular Biology* (F. Neidhardt, ed.), Vol. 1, pp. 2047 – 2065. ASM Press, Washington.

[42] Fang, G., Rocha, E., Danchin, A. (2005) How essential are nonessential genes? *Mol. Biol. Evol.* **22**:2147 – 2156.

[43] Danchin, A. (1988) Complete genome sequencing: future and prospects. In *BAP 1988 – 1989* (A. Goffeau, ed.), pp. 1 – 24. Commission of the European Communities, Brussels.

[44] Mushegian, A.R., Koonin, E.V. (1996) A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. U.S.A.* **93**:10268 – 10273.

[45] Fang, G., Rocha, E.P., Danchin, A. (2008) Persistence drives gene clustering in bacterial genomes. *BMC Genomics* **9**:4.

[46] Granick, S. (1957) Speculations on the origins and evolution of photosynthesis. *Ann. N. Y. Acad. Sci.* **69**:292 – 308.

[47] Danchin, A. (1989) Homeotopic transformation and the origin of translation. *Prog. Biophys. Mol. Biol.* **54**:81 – 86.

[48] Wächtershäuser, G. (1988) Before enzymes and templates: theory of surface metabolism. *Microbiol. Rev.* **52**:452 – 484.

[49] Movila, A., Uspenskaia, I., Toderas, I., Melnic, V., Conovalov, J. (2006) Prevalence of *Borrelia burgdorferi* sensu lato and *Coxiella burnetti* in ticks collected in different biocenoses in the Republic of Moldova. *International Journal of Medical Microbiology* **296**:172 – 176.

[50] Lawrence, J.G., Roth, J.R. (1996) Selfish operons: horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**:1843 – 1860.

[51] Tettelin, H., Masignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Angiuoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., Deboy, R.T., Davidsen, T.M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J.D., Hauser, C.R., Sundaram, J.P., Nelson, W.C., Madupu, R., Brinkac, L.M., Dodson, R.J., Rosovitz, M.J., Sulli-

van, S.A., Daugherty, S.C., Haft, D.H., Selengut, J., Gwinn, M.L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K.J., Smith, S., Utterback, T.R., White, O., Rubens, C.E., Grandi, G., Madoff, L.C., Kasper, D.L., Telford, J.L., Wessels, M.R., Rappuoli, R., Fraser, C.M. (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pangenome". *Proc. Natl. Acad. Sci. U.S.A.* **102**:13950 – 13955.

[52] Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., Boland, F., Brignell, S.C., Bron, S., Bunai, K., Chapuis, J., Christiansen, L.C., Danchin, A., Debarbouille, M., Dervyn, E., Deuerling, E., Devine, K., Devine, S.K., Dreesen, O., Errington, J., Fillinger, S., Foster, S.J., Fujita, Y., Galizzi, A., Gardan, R., Eschevins, C., Fukushima, T., Haga, K., Harwood, C.R., Hecker, M., Hosoya, D., Hullo, M.F., Kakeshita, H., Karamata, D., Kasahara, Y., Kawamura, F., Koga, K., Koski, P., Kuwana, R., Imamura, D., Ishimaru, M., Ishikawa, S., Ishio, I., Le Coq, D., Masson, A., Mauel, C., Meima, R., Mellado, R.P., Moir, A., Moriya, S., Nagakawa, E., Nanamiya, H., Nakai, S., Nygaard, P., Ogura, M., Ohanan, T., O'Reilly, M., O'Rourke, M., Pragai, Z., Pooley, H.M., Rapoport, G., Rawlins, J.P., Rivas, L.A., Rivolta, C., Sadaie, A., Sadaie, Y., Sarvas, M., Sato, T., Saxild, H.H., Scanlan, E., Schumann, W., Seegers, J.F., Sekiguchi, J., Sekowska, A., Seror, S.J., Simon, M., Stragier, P., Studer, R., Takamatsu, H., Tanaka, T., Takeuchi, M., Thomaides, H.B., Vagner, V., van Dijl, J.M., Watabe, K., Wipat, A., Yamamoto, H., Yamamoto, M., Yamamoto, Y., Yamane, K., Yata, K., Yoshida, K., Yoshikawa, H., Zuber, U., Ogasawara, N. (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U.S.A.* **100**:4678 – 4683.

[53] Joyce, A.R., Reed, J.L., White, A., Edwards, R., Osterman, A., Baba, T., Mori, H., Lesely, S.A., Palsson, B.O., Agarwalla, S. (2006) Experimental and computational assessment of conditionally essential genes in *Escherichia coli*. *J. Bacteriol.* **188**:8259 – 8271.

[54] Sekowska, A., Denervaud, V., Ashida, H., Michoud, K., Haas, D., Yokota, A., Danchin, A. (2004) Bacterial variations on the methionine salvage pathway. *BMC Microbiol* **4**:9.