编者按: 合成生物学(synthetic biology) 近年来已成为生命科学、医学的一大热点, 其中微生物基因组研究为合成生物学的发展奠定了重要基础。2002 年 7 月 11 日 Wimmer 等在 *Science* 网上发表了第 1 篇关于人工化学合成脊髓灰质炎病毒的论文, 引起世人轰动。近来又有合成蝙蝠 SARS-CoV 的报道。为尽早介绍这一领域中的理论问题, 应本刊邀请, 对细菌基因组学深有造诣的 Antoine Danchin 教授(法国巴斯德研究所) 根据比较细菌基因组学的大量资料, 结合生物遗传学, 对比细胞与计算机的复制与繁殖问题, 为本刊撰写了英文综述, 阐述了他的观点, 前瞻性地提出了合成细胞生命体的物理理论基础, 强调了功能基因及其介导的信息累积过程, 即利用能量避免功能基因产物降解, 并使新的有功能个体替换无功能的旧个体。Danchin 称之为自然选择。为便于国内读者阅读与理解, 复旦大学上海医学院医学分子病毒学实验室博士后孟哲峰博士作了中文编译。

Towards synthetic cells: are cells computers making computers?

Antoine Danchin

Genetics of Bacterial Genomes / CNRS URA 2171

Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France

Abstract: Understanding life supposes that one will, one day, reconstruct cells. A deep analysis of what life is shows that a cell is similar to a computers making computer. This asks for several orginal levels of organisation. First, the cell needs to be seen as a machine separated from the genetic program, which it runs. Over generations the machine reproduces, while the program replicates. Reproduction is a process which is able to accumulate valuable information over generations. Extracting valuable information from an ocean of noise requires an energy-dependent process which uses energy to prevent degradation of functional entities. Analysis of bacterial genomes shows that the core set of genes which persist in most genomes code for the functions needed to perform this process of ratchet-like information accumulation. It also suggests that a mineral, polyphosphates, could be a ubiquitous (and stable) energy source essential for the process.

Key words: Synthetic cell; Reproduction and replication; Information

The goal of Synthetic Biology is to try and reconstruct life, following engineering principles[1]. The underlying assumption is that a cell is simply some kind of automaton, constructed along principles that are quite similar to those of electronic engineering. However, electronic devices do not produce copies of themselves. It is therefore essential to go deeper into what life is to see whether we can reasonably apply engineering principles to the building up of synthetic cells. As pointed out by Dyson in his book *Origins of Life*[2], one should not consider that perpetuation of life is the result of one single process that would duplicate individual organisms. When considering the process permitting a cell to have a progeny, we need to separate between the cell and the genetic program. This is because it is necessary to separate between *reproduction* of the cell s machinery and *replication* of the cell s program. Reproduction is a process that perpetuates something abstract, the overall relationships between the individual components, in space and in time, of the organism. By contrast, replication is an exact duplication of the hereditary material, usually preserved in the form of nucleic acids, best illustrated in the DNA double helix. In this context we need to carefully explore the meaning of words. It is essential to remark that the mixing up of several concepts that should be well identified and separated from each other culminates in the widely spread vocabulary of Systems Biology with the consequence that it is difficult to define the borders of the discipline. Indeed, the word “system” is remarkably vague, and covers a variety of sometimes contradictory concepts more or less

通信作者: Antoine Danchin, E-mail: antoine. danchin@ normalesup. org
Corresponding author: Antoine Danchin, E-mail: antoine. danchin @ normalesup. org

associated with the idea of integrating together a variety of processes. In contrast, "synthetic" emphasizes the role of artifice in the construction of cells[3]. In this context we need to stress the role of integration in the new trends of biology. We think that the word "symplectic" constructed from the Greek, πλεκτειν, to weave, and συν, together, would be more appropriate, while encompassing most of what is deemed systemic or synthetic[4]. Also, this word has no connotation associated with it, which would prevent intrusion of irrational discussions in a purely scientific context (as was the case of the unfortunate connotation associated with Genetically Modified Organisms, http://www.normalesup.org/～adanchin/causeries/GMOs.html). And this kind of societal consideration is important for the future of the discipline. In this fuzzy context, most investigators tend to use alternatively reproduction and replication as if interchageable, and this has enormous conceptual consequences when reflecting about the very nature of what life is. Indeed, while replication must be as accurate as possible to perpetuate the same molecule over generations, it tends to irremediably accumulate errors. And, in the absence of recombination or exchange with external sources permitting to trace back the situation before errors occurred, noise will invade the DNA sequence and result in the error catastrophe described by Orgel[5,6], or before him, by Muller (Muller's ratchet)[7]. In contrast, reproduction is not doomed to decay, and as shown by Dyson with a simple toy model, it can even improve over time[2]. The background idea here is that during the reproduction process alternative entities (molecules, complexes, dynamics) can play the same role, so that while they are replacing previous entities, they may uncover novel properties (novel interactions in particular). In a nutshell, reproduction permits accumulation of *useful information* while most of the information created during replication is noise. As I use this word here in a popular sense I will have to begin this article with a reflection about the concept, to see how this can be articulated with what we know about genomes, microbial genomes in particular. A second feature of this view is that we have to separate, in the cell (and in the organism in general, but I will concentrate here on prokaryotic organisms to make the reflection simpler) between two structures: the cell machinery, with metabolism and envelope, and the cell's genetic program, with its chromosome(s). Reproduction will be deeply associated with the machine, while replication, as usual, will be associated with the chromosome. Describing the machine in detail is difficult, and we lack much understanding about its organisation (despite progresses in recent times, which show that contrary to expectation, the bacterial cytoplasm is certainly not a tiny test tube, but is extremely well organised[8-10]). Fortunately, in contrast, we have in the recent years gained much knowledge about the chromosome, and about its organisation, thanks to the considerable number of genome programmes which have kept accumulating (4 213 at the time of writing). And because macromolecular components of the cell are coded in the genome, we begin to have a fairly complete knowledge of what they are, and hints about their organisation. The second part of this article will be devoted to analysis of this organisation. Finally, I will try to put together this reflection about the nature of information, and the analysis of bacterial genome content and organisation to propose a completely novel view of living organisms. I will try to convince the reader that information is an authentic category of reality, on the same level as matter, energy, space and time, and show that living organisms are constructed to behave as information traps. And to make this conjecture more realistic I will show how some experiments may help us to understand how information is articulated with the standard categories. In conclusion, I will suggest how to extend this reflection to other domains of biology, including information-processing activities such as learning and memory.

1 Briefly revisiting information

As stated above, we keep using the word "information" as if we understood exactly what it means. This is an important difficulty in biology as, if we use a sloppy concept we can only derive sloppy conclusions from our experimental observations. It is therefore of the utmost importance to get more insight in the concept. To begin with, physicists already know that there is an intrinsic difficulty in the handling of the physical reality with the familiar categories of matter, energy, space and time. Indeed, as suggested by Einstein himself, to explain the quantum world would require the existence of "hidden variables" pertaining to the standard categories of reality, and he suggested experiments to identify them. Interestingly, in the past

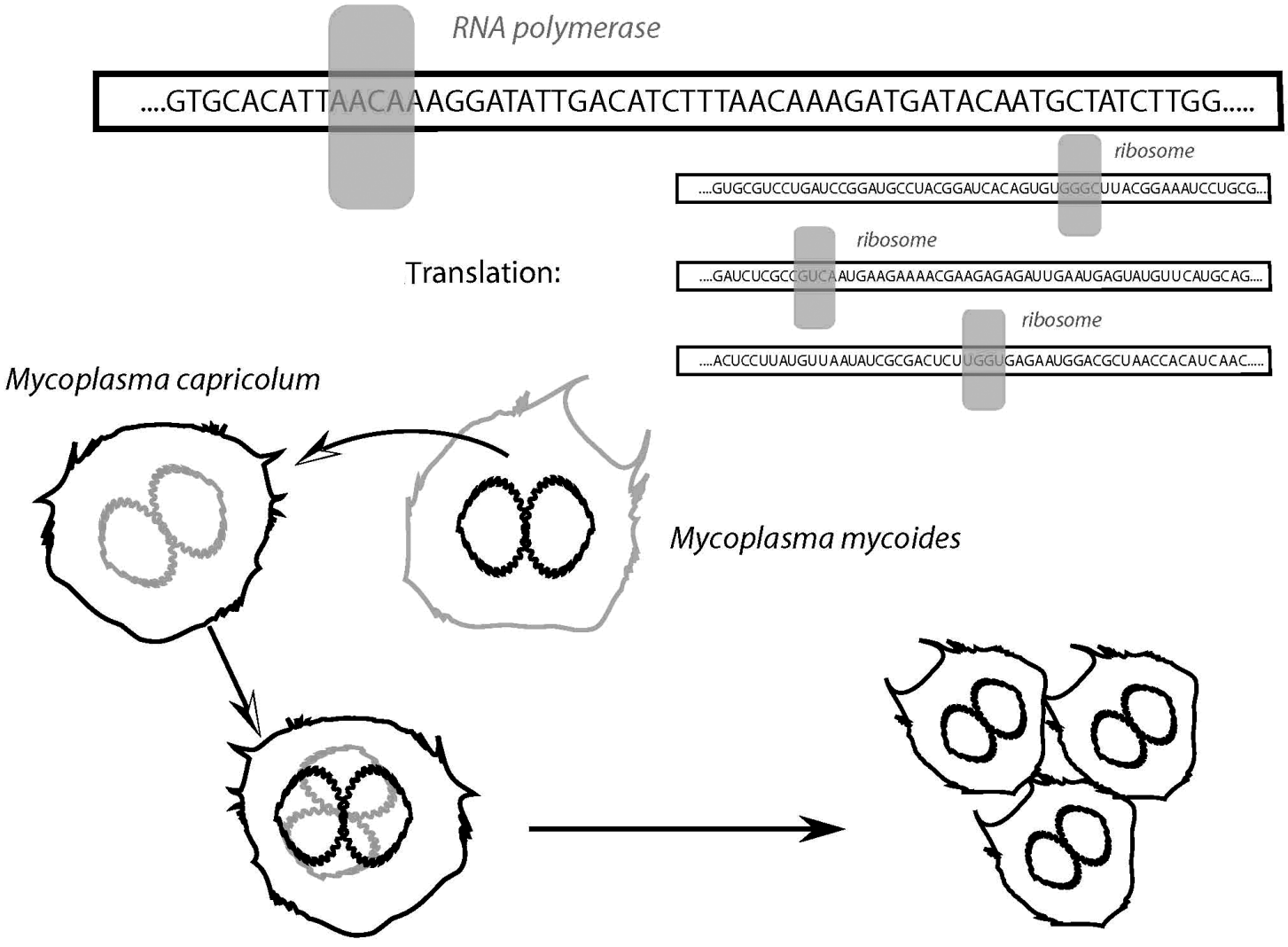
few decades all the predictions of Einstein have been refuted by carefully designed experiments. And this lead physicists such as Steane to contend that we need to include another feature in reality, information, whatever it is[11]. This is the path I will follow here, noting however that it cannot be question in such a short article to provide a novel mathematical description of an entirely new category of reality. I will therefore use "information" following the most recent formalisation of the concept, although I must stress that they are certainly inadequate, as they all have been created as if information was derived from the four standard categories of reality. This is not too much of a problem for us, fortunately, because, even if we cannot state explicitly what information is, we can nevertheless try to see how it is articulated with what we are familiar with, matter, energy, space and time. And I will concentrate on two features of information, its relationships with energy, with matter and with time.

The first quantitative mathematical description of the concept came from the work of Claude Shannon, when he analysed the limits of communication of strings of symbols[12]. The theory of communication he established was interesting only in measuring whether the integrity of the string of symbols (a message) was preserved, without taking into account its meaning[13]. In the biological context, this is exactly what replication does, when a DNA molecule is duplicated into an (almost) identical molecule. This is however very far from what we would like to say about information, and this is certainly very far from the way information is exploited at the level of transcription or translation. In a short work published posthumously, interested in the coupling between medicine and physics, the physician (not physicist!) Henry Quastler further developed a theory of biological organisation starting with the enigma of the origin of life, which was further developed, as stated in the introduction, by Dyson. As Quastler did, I wish to emphasize here the problem of the creation of information in simple cells, a question of central importance[14]. In a nutshell, most investigators at the time (and still many today) thought intuitively that creation of information required energy (for a historical discussion, see[15]), in such large amounts that what we could observe in life appeared quite mysterious. This is not so however: as demonstrated by Landauer, working on the limits of computation in computers, creation of information is reversible and the process does not require energy. In contrast, accumulating useful information required energy[16]. This remarkable property of the physical world, which I will not discuss here in details (see[15,17] for a general presentation), has the consequence that accumulation of information by living organisms is not a paradox and can be associated with explicit molecular processes, as we shall see below. Finally, there is a wealth of other processes where biologists contemplate a direct involvement of information in their topic of interest or when one deals with properties of the central nervous system, thus substantiating information as an authentic category of reality. This is the case of the way molecular motors perform their task, and even molecules can act as ? molecular information ratchets?[18]. In the same way, analysis of the exploration of space by insect males looking for their female involves a process of infotaxis, which computes ways to access the source of a specific pheromone in a highly turbulent environment, using information as the driving element permitting identification of the target[19]. In what follows we try to relate information to the way bacterial genomes are organized. In summary, when considering information as an authentic category in the domain of life sciences, we need to see both how accumulation of useful information can be linked to the consumption of energy, and to the concrete material processes that unfold in space and time in the cell. As geneticists, this means that we need to identify genes involved in the process. To this aim it is an excellent approach to use comparative genomics, with the huge number of genome sequences we now possess, to see whether we can identify the relevant genes and functions.

2 Cells and computers

In our man-made artificial world we have constructed machines, the computers, which have been as their central goal manipulation of information. Do we find similarities between cells and computers, and if so, can we identify functions (and genes) that make the comparison significant? Let us first summarise what a computer is. In a famous paper (available on the Internet), Alan Turing proposed that all computations involving whole numbers, as well as all operations of logic, could be performed by a simple machine reading and modifying a tape carrying a linear sequence of symbols, the Universal Turing Machine[20]. This

concept is at the root of the way our modern computers are constructed. The core feature of Turing's model is the requirement for a physical separation between a linear string of symbols, the data/program, and a machine endowed with specific properties which enable it to manipulate (read and write on) the string of symbols. I wish to stress here that the organisation of the cell matches this description: the genetic program is carried out by the linear string of nucleotides that make up the DNA molecule. If we consider a cell in terms of Turing machines, this raises the straightforward question: can we consider the program to be a separate entity in the cell, and if so, to what extent? Many observations support this contention. The basis of genetic engineering is the manipulation of DNA molecules (real or artificially constructed ones) and expression in foreign cells: this is a first proof of concept. Pieces of a genetic program can be transformed from one organism into another: many bacteria now produce human proteins. Furthermore, cells that perform logical tasks have been experimentally constructed[21, 22]. Yet, all these experiment make use of only a small part of the genetic program: can the analogy be extended further? The unexpected identification of extensive horizontal gene transfer in the genomes of bacteria gave further substance to the separation between the program and the machine, as it was clear that a large number of genes coming from the outside can be expressed by any type of bacterium[23-25]. This widespread observation[26] was of considerable importance. Yet it did not provide final proof that the program defining an organism could be extracted as a whole and placed in another environment where it could act as a functional program. The ultimate proof was provided by the recent transplantation of an entire genome from a given species to a different one[27]. This conceptual advance perfected the analogy of the cell as a Turing machine by showing a complete separation between the cell machinery and the program that drives its functioning (Fig. 1).



The cell can be considered as a highly parallel Turing Machine, where transcription and translation operate as algorithms (begin — core action — check point — repeat — end). In a Turing machine it is essential that the program is physically independent from the machine itself. The experiment of the group of Craig Venter which demonstrated the feasibility of genome transplantation, starting from a *Mycoplasma capricolum* genome to transmute it into a *M. mycoides* genome is an experimental demonstration of that separation[27].

Fig 1. The cell as a Turing Machine and genome transplantation (modified from Figure 1 in reference[29])

3 Giving life to a Turing Machine

A Turing Machine is just an abstract entity. It forgets about matter and operates only within the realm of the category information and there is no question about where it comes from. When a Turing Machine is made into a concrete computer, it needs to be linked to matter and energy, and placed in the standard tridimensional world. Importantly, the "information" it manipulates is only that of the program, not that of the machine which permits the program to run. The machine is supposed to be pre-existing, with specific features for reading and writing on the program, and there is no investigation of the way it has been created: this is why the fact that it could reproduce has not been much explored by computer scientists (see however the reflection of von Neumann[28] and my comments about this situation[29]). This lack of adequacy between the abstract world of pure information and the material part of reality is not without consequences as can be seen in the following example[4]. The operating system is an abstract program making a computer run. To be usable it must be carried by concrete objects, such as a compact disk (CD). A CD left lying for some time in a car's rear window in the sun will be deformed, and despite the fact that the program it carries is unaltered, it will no longer be read by the computer's laser beam. As a consequence, the computer cannot use it to start up. In other words, although in the abstract world in which Turing Machines exist, the separation between hardware and software is rigorous, in practice there must be a physical support for each entity, so that we cannot completely separate the hardware from the software in any real implementation of the Turing Machine. This is an important constraint, which may create difficulties in transplantation experiments such as those where an artificial *Mycoplasma* genome has been synthesised, using *Saccharomyces cerevisiae* as an intermediary host[30]. The role of matter, space and time is even more complicated for the machine itself, and, following Dyson, I contend that this is where valuable information can be progressively increased over generations[29]. Indeed this is where the reproduction process is able to create novel information, as is demonstrated by the paradoxical observation, so obvious that nobody takes notice: babies are born very young! This means that old "machines" can produce new ones during the reproduction process, and this happens all the time, and everywhere. This implies that living organisms have ubiquitous functions that permit them to produce novel, useful information. In cells, functions are coded by genes, and postulating the existence of ubiquitous functions requires postulating the existence of corresponding genes. This is what we shall now see. We face a first difficulty: in living organisms, there is never a one to one correspondence between a function and a gene, as different genes can code for the same function. The consequence is that it is *not possible to infer gene ubiquity from function ubiquity*. Nevertheless living organisms, when they carry a gene for a useful function, tend to propagate them down generations. This implies some kind of "stickiness" in genes, which can be measured and used, and that we named *persistence*[31]. A thorough study of gene persistence in bacterial genomes led us to identify some 500 persistent genes. Unexpectedly, this is approximately twice the number of genes deemed essential for life[31]. What is the function of these extra persistent non-essential genes? Briefly, they can be grouped into three major classes: genes coding for metabolic "patches", which code functions taking care of chemical incompatibilities between standard metabolic intermediates (alpha-dicarbonyl molecules are extremely reactive, for example[32]); genes coding for maintenance and repair; and genes coding for degradative functions which work using energy.

4 Creating valuable information: Maxwell's demon's genes

This latter family provides us with a hint about the way information can be progressively accumulated over generations, permitting creation and reproduction of a concrete Turing Machine. Indeed, this family is the exact counterpart of the process described above as essential to accumulate valuable information in a computer[16]. This family of genes creates a process — this is exactly what we name "natural selection" — which measures what is functioning, using energy to avoid destroying it, while it makes room for new entities by destroying non-functional entities[17]. It is remarkable that these genes code exactly for what is expected from the counterpart of Maxwell's demon[16], using energy in a way somewhat similar to that described in the model of kinetic proofreading developed to account for the accuracy of the translation

process[33, 34]. The idea is that particular proteins interact with relevant entities and measure whether they are functioning (i. e. are actively engaged in catalysis, properly folded, etc.). If this is the case, then energy is used to dislodge the proteins before they display their degradative power. If not they degrade their substrate, making room for synthesis of further, newly synthesised (and presumably functional) entities (RNA and proteins in general, but this could also be membranes or more complex structures). The outcome of this process is that functional entities will progressively accumulate, replacing the aged or non-functional ones. An interesting feature of this degradative process is that it does not compare its substrate to any template, but, rather, measure whether it is working or not working. The consequence is that any event that gave rise to a functional entity, from whatever source, will lead to retaining that entity in the system, at the expense of non-functional entities. Functional replacement, rather than structural replacement will be the rule. Overall the cell of interest will progressively build up a progeny which will be richer in valuable information that its ancestors. And remarkably, this positive trend will develop independently of any goal. This explains the pervasive observation that living organisms evolved constantly towards higher complexity. The remaining question, here, is therefore the nature of the energy source that will be used to make these energy-dependent processes work properly. ATP is an obvious candidate, and under standard metabolic conditions, when energy availability is not limiting, it is probably used in the process (as well as other energy-rich nucleotides). However, the most important moment when a cell has to be able to generate a progeny is after some kind of hard times, typical of the long stationary phases of growth witnessed by most bacteria in the environment. Under such conditions it is likely that a stable energy source is required. Our analysis of gene persistence has suggested that polyphosphates might play this essential role[17]. Polyphosphates is a mineral, and therefore resistant to all kinds of extremely harsh environments, in particular desication, reactive oxygen species or irradiation. It will be of major interest to explore whether the conjecture of their involvement in the process of accumulation of information over generation.

5 Provisional conclusion

To sum up, bacterial genomes comprise a set of genes developing life along two major processes, sustaining life, thanks to genes which form a core set of approximately 250 genes, and a second set of persistent genes which permit perpetuation of life, as well as maintenance and coping with the inevitable idiosyncrasies of the molecules that make cells. Life, as defined by the core essential genes is probably limited to a few generations, as it rests on the maintenance of the cell machinery and the replication of the genetic program (see references in[17]). However, when genes aiming at collecting and accumulating novel valuable information are included, the cell will make the most of what it gets, and will progressively build up an information-richer progeny. This process is remarkable as it is a concrete implementation of the Maxell's demon. Finally, because this line of analysis of bacterial genomes is very general in its abstract form, it can be pursued in other domains of biology. In particular, on the sad side, it appears to provide a straightforward explanation for the induction of cancer, when differentiated cells, doomed to multiply, age and die, discover that they have the opportunity to accumulate information. In contrast, on the happy side, the same process could well be at the root of learning and memory in the brain[35], where, rather than proteins and RNA molecules, specific synapses are taken into consideration: an energy-driven process, meant to inactivate those synapses which are not working will use energy to prevent degradation of the functional ones. The underlying genetic processes will probably been uncovered in the next few years.

Acknowledgements

This work results of three decades of discussion with many people, often members of the the Stanislas Noria network (http://www.normalesup.org/~adanchin/causeries/causeries_en.html). *In silico* and *in vivo* experiments have been supported by the PROBACTYS programme, grant CT-2006-029104 and the TARPOL programme, grant KBBE-2007-212894 in an effort to define genes essential for the construction of a synthetic cell.

REFERENCES

[1] Endy D. Foundations for engineering biology [J]. Nature, 2005, 438(7067): 449-453

[2] Dyson FJ. Origins of Life [M]. Cambridge, UK: Cambridge University Press, 1985

[3] Barrett CL, Kim TY, Kim HU, *et al.* Systems biology as a foundation for genome-scale synthetic biology[J]. Curr Opin Biotechnol, 2006, 17(5): 488-492

[4] de Lorenzo V, Danchin A. The discovery of new worlds and new words[J]. EMBO Rep, 2008, 9(9): 822-827

[5] Orgel LE. The maintenance of the accuracy of protein synthesis and its relevance to aging[J]. Proc Natl Acad Sci USA, 1963, 49(4): 517-521

[6] Orgel LE. The maintenance of the accuracy of protein synthesis and its relevance to ageing: a correction[J]. Proc Natl Acad Sci USA, 1970, 67(3): 1476

[7] Muller H. Some genetic aspects of sex. Am Nat, 1932, 66(703): 118-138

[8] Danchin A, Guerdoux-Jamet P, Moszer I, *et al.* Mapping the bacterial cell architecture into the chromosome[J]. Philos Trans R Soc Lond B Biol Sci, 2000, 355(1394): 179-190

[9] Errington J. Dynamic proteins and a cytoskeleton in bacteria[J]. Nat Cell Biol, 2003, 5(3): 175-178

[10] Jun S, Mulder B. Entropy-driven spatial organization of highly confined polymers: lessons for the bacterial chromosome[J]. Proc Natl Acad Sci USA, 2006, 103(33): 12388-12393

[11] Steane A. Quantum computing[J]. Rep Prog Phys, 1998, 61(2): 117-173

[12] Shannon C, Weaver W. The Mathematical Theory of Communication[M]. Urbana: University of Illinois, 1949

[13] Cover T, Thomas J. Elements of Information Theory[M]. New York: Wiley, 1991

[14] Quastler H. The Emergence of Biological Organization[M]. New York: Yale University Press, 1964

[15] Danchin A. The Delphic Boat. What Genomes Tell Us[M]. Cambridge (Mass, USA): Harvard University Press, 2003

[16] Bennett C. Notes on the history of reversible computation[J]. IBM J Res Dev, 1988, 32(1): 16-23

[17] Danchin A. Natural selection and immortality[J]. Biogerontology, 2008, [Epub ahead of print]

[18] Serreli V, Lee CF, Kay ER, *et al.* A molecular information ratchet[J]. Nature, 2007, 445(7127): 523-527

[19] Vergassola M, Villermaux E, Shraiman BI. 'Infotaxis' as a strategy for searching without gradients[J]. Nature, 2007, 445(7126): 406-409

[20] Turing AM. On computable numbers, with an application to the Entscheidungsproblem[J]. Proc London Math Soc, 1936, 42: 230-265

[21] Basu S, Gerchman Y, Collins CH, *et al.* A synthetic multicellular system for programmed pattern formation[J]. Nature, 2005, 434(7037): 1130-1134

[22] Elowitz MB, Leibler S. A synthetic oscillatory network of transcriptional regulators[J]. Nature, 2000, 403(6767): 335-338

[23] Baumler AJ. The record of horizontal gene transfer in Salmonella[J]. Trends Microbiol, 1997, 5(8): 318-322

[24] Hilario E, Gogarten JP. Horizontal transfer of ATPase genes — the tree of life becomes a net of life[J]. Biosystems, 1993, 31(2-3): 111-119

[25] Médigue C, Rouxel T, Vigier P, *et al.* Evidence for horizontal gene transfer in Escherichia coli speciation[J]. J Mol Biol, 1991, 222(4): 851-856

[26] Moszer I, Rocha EP, Danchin A. Codon usage and lateral gene transfer in Bacillus subtilis[J]. Curr Opin Microbiol, 1999, 2(5): 524-528

[27] Lartigue C, Glass JI, Alperovich N, *et al.* Genome transplantation in bacteria: changing one species to another[J]. Science, 2007, 317(5838): 632-638

[28] von Neumann J. The Computer and the Brain[M]. New Haven: Yale University Press, 1958

[29] Danchin A. Bacteria as computers making computers[J]. FEMS Microbiol Rev, 2009, 33(1): 3-26

[30] Gibson DG, Benders GA, Andrews-Pfannkoch C, *et al.* Complete chemical synthesis, assembly, and cloning of a Mycoplasma genitalium genome[J]. Science, 2008, 319(5867): 1215-1220

[31] Fang G, Rocha E, Danchin A. How essential are nonessential genes[J]? Mol Biol Evol, 2005, 22(11): 2147-2156

[32] Munanairi A, O'Banion SK, Gamble R, *et al.* The multiple Maillard reactions of ribose and deoxyribose sugars and sugar phosphates[J]. Carbohydr Res, 2007, 342(17): 2575-2592

[33] Hopfield JJ. Kinetic proofreading: a new mechanism for reducing errors in biosynthetic processes requiring high specificity[J]. Proc Natl Acad Sci USA, 1974, 71(10): 4135-4139

[34] Ninio J. Kinetic amplification of enzyme discrimination. Biochimie, 1975, 57(5): 587-595

[35] Changeux JP, Courrège P, Danchin A. A theory of the epigenesis of neuronal networks by selective stabilization of synapses[J]. Proc Natl Acad Sci USA, 1973, 70(10): 2974-2978

向合成细胞迈进:细胞与计算机复制过程中的遗传问题

唐善 · 安东
法国巴斯德研究所

摘　要: 对生命机制的理解预示着人类总有一天能重建细胞。对于生命表征的深入分析显示,细胞的分裂、繁殖与计算机间的复制非常相似。这要求我们在亚细胞水平上进行分析比较。首先,细胞必须被看作是一个独立于其遗传物质的机器。在此意义上,遗传物质在数代间复制,而机器也可以再生。繁殖是一个可以在子代间累积有用信息的过程,从大量无用的信息海洋中提取有用信息是一个依赖能量的过程,并借助能量阻止功能性实体的降解。细菌基因组的分析显示,核心基因(即维持必需功能的基因)负责执行这种类似于棘齿运行的信息累积过程。本文提出,磷酸化矿物盐类可能是广泛的能量来源。

关键词: 合成细胞; 繁殖与复制; 信息

合成生物学的目的是遵循遗传学法则,尝试重建细胞或生命。这一过程根本的理论基础是细胞被看作某种简单的机器,它们被建造的法则与那些工程学的法则非常相似。但是机器之间是不能相互复制的,这就需要我们从更深层次去理解生命过程,并且为工程学法则在细胞建造上的运用寻找合理的依据。很明显,生命体的复制不是一个简单的拷贝过程,我们可以也必须把这一过程分成两个独立的事件:即细胞实体的繁殖与遗传信息的复制。繁殖体现的是一个抽象的概念,包括细胞成分和生命个体在时间和空间上的关系;复制则是遗传信息的准确拷贝。

许多学者选择性地使用繁殖和复制这两个名词,似乎它们之间可以随意替换,事实上不然。复制必须尽可能在各代间产生完全相同的分子,这带来不可逆转的累积错误;如果没有与外来 DNA 的重组与交换,错误信息或噪音最终会侵蚀 DNA 并带来毁灭性后果。繁殖不仅不会导致削减或毁灭,相反会使个体随时间进化。简单地说,繁殖能累积有用的信息,然而复制过程产生的多数信息都是无用的,甚至是噪音。

对于生命体所发生的繁殖和复制过程,繁殖与细胞器活动密切相关,而复制通常指的是染色体。详细描述细胞机器非常困难,因为我们缺乏对细胞组成形式的了解(即使是原核细胞,其细胞器组成也不仅仅是一个微小试管里的反应体系)。幸运的是,近年在原核细胞染色体基因组成上的大量进展为我们探寻这种细胞机器组成提供了线索,本文第 2 部分将会详细描述。

我们也需要理解信息的本质,比较基因组学和神经生物学的研究成果能很好地诠释信息是一种实体,它可以在概念上等同物质、能量、时间和空间。生命体被建造,也正是作为信息的载体和体现。

1　信息概念的重新理解

从爱因斯坦需要借助存在的“隐藏的变量”阐述量子世界起,物理学家们知道,以人们所熟悉的物质、能量、时间、空间等名词去诠释物质的现实世界,有着天然的障碍。同样,当我们把信息放在生物学中去解读时,以我们现在对信息这个名词的理解是远远不够的。当然,限于本文的篇幅,不可能重新发展一个全新的数学模式去描述它,但应该意识到,信息确实是与物质、能量、时间和空间等并列的物质实体。本文将集中描述信息的两个重要特征以及它与物质、能量和时间的关系。

第 1 个对信息的概念进行数学式定量描述的是 Claude Shannon,他所建立的通信理论集中在测试字符串的完整性是否保留,而不是去解释这些字符串的含义。在生物学的范畴里,这与 DNA 的复制过程相似,但与我们想讨论的信息相差甚远,特别是与信息在翻译和转录水平上的应用方式相差甚巨。

内科医生 Henry Quastler 尝试进行医学和物理学结合的研究,他由探讨生命起源之谜开始,建立了生命体组成的理论,并且该理论得到 Dyson 的进一步发展。正如他们所提到的,本人也将重点强调,信息创造是单细胞的核心难题。简而言之,多数研究

者认为,创造信息需要能量,在生命活动中观察到的大量事实使得这些过程好像很神秘。其实不然,计算机进行信息创造的过程可逆,且不需消耗能量;而累积有用的信息需要消耗能量。这一物理世界的显著特点与生命体的信息累积过程并不矛盾,而且可能与简单的分子过程相关。

最后,生物学家在他们的各自领域,对包含信息的生命过程所做的卓有成效的研究,为我们提供了很好的证据:即信息毫无疑问是一个现实世界的名词。比如神经生物学的研究证明了信息的存在及其引发的行为;对昆虫的生物学行为研究发现,雌、雄个体间产生的信息,使得昆虫在求偶路线上能作出最合理和有效的选择(即使他们处于非常 杂和无序的外部环境中)。

总而言之,当考虑在生物学领域把信息作为一个真实的范畴,我们必须看到信息的两个特征:信息的累积与能量消耗过程紧密联系,同时也与真实存在的生命过程紧密联系(已揭示的细胞生命周期过程)。作为遗传学家,这意味着需要去揭示参与这些过程的基因,对于这一目的,细菌基因组的研究提供了极大的便利。

2 细胞和计算机

在人造的物质世界中,人类发明了机器和计算机,后者在信息操作中发挥着中心的作用。人们是否能发现细胞与计算机的相似性?如果能,那么又能否验证那些实现了有意义比较的功能或基因?回答此问题,我们首先必须看一看什么是计算机。正如 Alan Turing 在一篇著名的论文中所提出的,所有的计算机,包括全数字的和逻辑的,能够通过一台简单的机器进行线性符号的读写和修正来实现,即通用图灵机。这也是现代计算机的理论根源。图灵机的核心特点是线性符号(数据和程序)与可操控的机器在物理上的分离。在这里要强调的是,细胞的组成形式符合这一特点:遗传信号由线性符号(ATCG 碱基)组成的 DNA 携带。如果考虑细胞是图灵机,马上要面对一个问题:细胞的遗传信息是否能以独立的实体存在?如果是这样,遗传信息的独立程度如何?大量基因工程操作,尤其是近来成功进行的全基因组移植和转导操作,后者使一个物种个体变为另一个不同的个体,可以证明遗传信息能独立存在,并能在其他个体操作和体现(类似于相同的程序在不同图灵机的使用)。这些进展使得细胞与图灵机的类比更加完美:即机器与线性符号的完全分离。

3 生命体的图灵机

当一个图灵机被建造为一个真实的计算机,它必须被赋予物质和能量,并且放置在一个标准的三维空间。重要的是,它操作的“信息”仅仅是程序,而不是允许程序运行的机器。这个机器被假定为预先存在,并具有读写功能,但没有关于它如何产生的调查:这也是为什么关于图灵机再生的事实很少被计算机科学家所探究的原因。程序是一个抽象的实体,它能让图灵机运行,但它必须要有另一个真实的物质支持(如 CD),而且当 CD 损坏时,虽然程序完好,图灵机也不可能运行和启动。换句话说,图灵机所存在的抽象世界中,软件和硬件的分离是严格的,而且实际操作中必须有物理的实体支持他们的存在,所以我们不可能在真实的实践中完全分离图灵机的软件和硬件。这也是一个很重要的限制条件,对细菌全基因组的移植操作设置了很大的障碍,这些操作必须要使用中间宿主(类似于用于装载程序的 CD)。

物质、时间和空间的机制比机器本身更复杂,本人坚持认为这些机制使得有价值信息能在各代个体间逐渐增加。真实的情况是,繁殖的过程能带来新的信息:即产生全新的个体。很少有人注意到这一事实:婴儿出生都是全新的生命。这意味着旧的机器能产生新的个体,并且每时每刻、每处都在发生。生命有一个普遍的机制:产生并保留有价值的信息。对细胞而言,功能由编码基因实现,因此推测存在的功能需要先去推测存在的基因。然而,这在细胞功能研究上有很大的障碍:细胞内同一功能可能需有多个基因联合作用得以实现,即功能与基因间没有一一对应的关系。但无论如何,对于生命体,当携带有功能的基因时,它总是倾向于在子代延续和保留该基因。这暗示着基因有某种我们所称之的坚持性(能够被检测到和使用)。对基因组的广泛比较和研究提示,细菌基因组中大约 500 个基因具备这种持续性,它们可以被分为 3 类:编码辅助调节代谢蛋白的基因、编码代谢和修复蛋白的基因以及编码行使降解功能蛋白的基因。

4 能量是创造有价值信息的麦克斯韦妖

麦克斯韦妖:麦克斯韦是分子运动论的奠基人,他设想有两箱温度相同的气体,两箱之间有个小孔,小孔上坐着一个精灵(某种装置),精灵可以控制小

孔上的一扇门打开或者关闭。对从左边箱子来的分子,如果运动速度很快,精灵就打开门让其运动到右边;如果速度较慢,精灵就让门处于关闭状态。反之从右边箱子来的分子,如果速度较慢,精灵便打开门,让它们运动到左边;如果速度较快,精灵就让门关闭。这样左边箱子里的气体将变冷,右边箱子里的气体将不断变热。但问题是我们是否能找到实现麦克斯韦妖并破坏热力学第二定律。

细菌基因组研究为我们探寻有价值信息在各代个体间累积的方式提供了线索:即类似执行计算机中有价值信息累积的过程,从而达到图灵机的创造和再生。这个过程是:检测有功能基因,利用能量避免其产物降解,并使新的有功能个体替换无功能的旧个体,我们称之为自然选择。值得关注的是,正是这些基因编码我们所寻找的麦克斯韦妖式的功能。这一思路应该是通过对特定蛋白与相关实体相互作用来检测该蛋白是否有功能来实现。如果是这样,能量将用于在功能性蛋白降解前使之转移。这一过程的结果是功能性实体累积,并替换无功能或年长的实体。该降解过程的一个显著特点是不对降解底物作比较,而是检测降解产物是否有用。因此生物细胞遵循的规则是功能性替换,而不是结构性替换。这一规则对细胞的好处是,在其子代中将出现有价值信息的逐渐累积,这也能解释生物的进化总是朝向更高等和更复杂的方向进行。

剩下的问题是驱动这一过程的能量来源,ATP 是一个明显的候选者。然而,细胞繁殖子代最重要的时期总是在一些困难时相之后,典型的例子是细菌增殖中有一个较长的稳定期(stationary phase)。在此种情况下,稳定的能量来源是必需的。对细菌基因持续性分析的结果提示,多磷酸盐可能发挥重要的作用。作为矿物质,在极端环境中,多磷酸盐依然很稳定。因此对于多磷酸盐参与信息累积过程的设想也值得好好研究。

5 结语

综上所述,细菌基因组用以发展生命的两个大基因家族分别施行特定的过程:一类大基因家族是组核心基因(大约 250 个),用来维持生命,另一大基因家族允许生命复制并编码涉及细胞再生过程的关键蛋白产物。前者定义为生命基因组(维持生命的核心基因组),仅仅能延续数代,因为它停留在维持对基因组和细胞器的复制;然而,当后者参与到细胞生命活动中时,通过对有价值信息累积的倾向,使得细胞能不断超越,并最终构建信息富集的子代。这个过程的引人注目之处还在于,它似乎是麦克斯韦妖作用的真实例证。最后,因为对细菌基因组分析的模式具有普遍的适用性,这一分析模式也将广泛地被运用于其他生物学领域。例如从坏的方面讲,可能引发肿瘤细胞的产生,是正常细胞一些注定应该消亡的信息累积的后果;从好的方面来说,是正常细胞获得累积信息的后果,可能对于脑神经的认知和记忆有着积极的意义,即消耗能量的过程可以清除无功能触突而保留有功能的触突。

(收稿日期:2009-01-13)