The ancestor of eukaryotes, an already complex organism

6/1/12

The term "simple" is often associated with "ancient" in the field of evolution. Indeed, the more simple an organism, the more we tend to consider it ancient (and vice versa). This excessively limited vision of things has once again been put to the test by researchers in Liège. The latter have discovered that the machinery of the last common eukaryotic ancestor was already highly complex, especially in terms of the RNA maturation process.

A cell is comparable to a miniature factory. Just like a factory, the cell needs energy and raw materials to function and, thanks to a well-organised production line, it generates finished products as well as waste. The cells' machinery is essentially programmed to create proteins. While all cells are equipped with this sophisticated machinery, the type of proteins they produce depends on their function.

There are various stages along the protein production line, in particular, transcription, splicing and translation.
During transcription, the **DNA** double helix unwind to allow the **RNA** polymerase, an **enzymatic** complex, to come and read the sequence on the single strand of DNA corresponding to the **gene** that encodes the protein to be produced. As it moves along the DNA, the RNA polymerase synthesizes a precursor RNA called the pre-mRNA. As well as containing all the information required to generate the said protein, it also includes superfluous information that must be eliminated before the translation of the **mRNA** into the **amino acids** composing the protein. "**In eukaryotes, the genes are split up or arranged in a patchwork, i.e., they are constituted of coding and regulating sequences, the exons, and non-coding sequences, the introns**," explains Professor Patrick Motte, in charge of the **Laboratory of Functional Genomic and Molecular Plant Imaging** at ULg. "**The exons are the parts you find on the mature mRNA while the introns will be eliminated**," continues Patrick Motte. Thus, the pre-mRNA will undergo a sort of cleaning process during which it will be relieved of its introns and after which it will then be mature and ready to be translated. "**The process during which the introns are recognised then eliminated, and the exons are linked to one another, is called splicing**," Patrick Motte elaborates.

**Several proteins for a single gene**

Up until recently, it was commonly agreed that a gene encodes a specific protein through the intermediary of a mature mRNA that is generated after the constitutive splicing of a pre-mRNA. Constitutive splicing means that every intron is eliminated and every exon is retained in the final mRNA. This traditional set-up was completely turned upside down several years ago by the discovery of another type of splicing: alternative splicing. Contrary to constitutive splicing, this second form of splicing doesn't strictly result in the exclusion of all the introns and the inclusion of all the exons in the mature mRNA. Depending on various as yet unknown events, during alternative splicing, some exons may not be retained and some introns may be included in the mature mRNA. Hence, the expression of a gene may lead to the formation of several different mRNA and therefore the production of several different proteins, or protein isoforms.

"**The expression of a large number of genes is regulated by alternative splicing. In human, for instance, recent studies have shown that this concerns 95 % of genes**," Patrick Motte explains. This alternative splicing is at the origin of the great diversity of proteins produced from a limited number of genes. "**In the case of human, during the sequencing of the human genome, scientists were surprised to only find 30 000 genes. They found this to be very little considering the complexity of our organism. Indeed, the genome of the smallest flowering plant species also contains 30 000 genes... We now know that alternative splicing is the preponderant process leading to the great diversity of proteins**," continues the professor.
Retracing the evolutionary history of SR proteins

Whether it is constitutive or alternative, splicing requires a complex macromolecular edifice known as the spliceosome, which is itself composed of around a hundred different proteins. Among them are the SR proteins, whose name is derived from their domain rich in serine/arginine dipeptides. "The SR proteins participate in the assembly of the spliceosome, the selection of splicing sites and the inclusion or elimination of introns and exons in the mRNA", Patrick Motte points out.
It is within the framework of the new research theme concerning alternative splicing, developed by Patrick Motte, that he began to focus on SR proteins. While he was working on these proteins, the scientist aimed to determine their origin over the course of evolution and, in particular, he wanted to know if they were present in all plants. "At my humble level, I started to compare and align several sequences of the green lineage (green algae and land plants) with a view to studying their evolution, but this quickly became very complex. That's why I asked Professor Denis Baurain if he would be interested in working with me on this study", explains Patrick Motte.

Once contact was established, the researchers decided to do things on a large scale. Rather than limiting themselves to analysing the presence of SR proteins within the green lineage, they decided to research the presence of SR proteins on the scale of the tree of life! "We extended this study to numerous species whose proteome, i.e., the entire set of proteins, was already known at the time of the study", specifies Patrick Motte. The idea was, among other things, to be able to answer the following question: were SR proteins already present in the last common eukaryotic ancestor?

**Extensive analyses**

"In phylogeny, it is important to compare conserved sequences", says Professor Motte. And yet this is not the case of the domain rich in serine/arginine dipeptides, known as SR domain. But besides this domain, SR proteins have an RRM domain (which stands for RNA recognition motif), which is a RNA binding domain. Contrary to the SR domain, the former has the necessary "phylogenetic qualities". It is therefore the RRM domain that the researchers chose in order to study the evolutionary history of the SR proteins. "It is a small domain of 80-90 amino acids, but it is the only really useful one from a phylogenetic point of view", continues Patrick Motte.

Consequently, the researchers' (Denis Baurain, Sophie Califice, Marc Hanikenne and Patrick Motte) mission consisted, no less, of identifying and studying all the RRM proteins within the proteomes of more than 200 organisms! Both in prokaryotes (bacteria and archaea) and eukaryotes (plants, animals, fungi and protists)...

"We therefore found ourselves with more than 12 000 different sequences", Patrick Motte specifies. "This was very difficult to manage but essential to retrace the evolution of these proteins. The majority of the other studies on the subject only related to several hundred sequences at most", continues the scientist.

In order to be able to analyse this vast amount of data, the researchers had to develop complicated techniques. Among other things, Professor Denis Baurain elaborated new algorithms in order to complement the computer programs already available for such analyses. "The analyses required months of calculations on a super computer and a vast amount of interpretation work", explain the two scientists. "We had two approaches: automatic on the one hand, with the use of computers for data preprocessing, and manual on the other hand, with a meticulous examination of the results obtained", continues Denis Baurain. Hence, the gigantic phylogenetic trees created by the super computer were then scrutinised by the researchers looking for SR proteins. Besides this double approach, the researchers also used various phylogenetic models to support their results. It was indeed a challenge to work on such a short sequence as the RRM domain. "As a result, we increased the approaches because they each came up with rather uncertain statistical results. But since all of them converged towards the same tendency, this reinforced the validity of our conclusions", Denis Baurain explains.
SR proteins, an old innovation!

The results obtained measured up to the time and work the researchers put into this study, which began in 2007. Indeed, it revealed several important conclusions that were published in the *Plant Physiology* (1) journal. The analyses revealed that SR proteins are present in a great many eukaryotic organisms. Better still, the subfamilies of these proteins are found within different groups of eukaryotes. "This means that these subfamilies were already present in an ancestral eukaryotic form. Because when we find the same family or subfamily of proteins in organisms as distant as plants and animals, either this is a case of convergence or they have a common origin, as confirmed by our work on SR proteins", Patrick Motte emphasises. Thus, an ancestral sequence has evolved and led to different families and subfamilies of SR proteins that were already present in the last eukaryotic common ancestor (LECA). Besides the interest of this discovery in helping to understand the history of SR proteins, this also means that LECA already possessed highly complex RNA maturation machinery and was far from the simple organism that we might have imagined. "In general, we often think that if an organism is simple, it must be ancient (and vice versa) but this isn't always the case", Patrick Motte tells us. "Evolution doesn't always progress towards complexity", Denis Baurain continues. "Take baker's yeast for instance. This organism has no SR proteins. Yeast has evolved by simplifying its RNA maturation machinery compared with that of LECA and that of its current cousins", Denis Baurain specifies.

**Was the RRM domain already present in the common universal ancestor?**

Besides the light it sheds on the evolution of SR proteins, this gargantuan phylogenetic study also relates to the superstructure of the tree of life. By showing that a large number of bacteria present proteins containing a very similar RRM domain to that of eukaryotic SR proteins, it does indeed suggest that this domain potentially existed in the last universal common ancestor (LUCA). "This domain is extremely well conserved and presents identical motifs in eukaryotes and prokaryotes. Considering their complexity and phylogenetic proximity, it's rather unlikely that these RRM domains appeared (or were transferred) independently in these two groups", Denis Baurain explains.
While many scientists think that eukaryotes (far more complex, especially their machinery) result from the fusion between an archeon and a bacterium (hence, two prokaryotic organisms), others support the more traditional vision that archaea, bacteria and eukaryotes have evolved separately from LUCA. Even if the study
of RRM doesn't help us make a clear-cut decision between these two hypotheses, it does however invalidate the idea that LUCA was an organism with very simple machinery, which opens the way for the rehabilitation of the traditional hypothesis. "The RRM domain can be found in eukaryotes and some types of bacteria. We can therefore reasonably imagine that this domain was present in LUCA and that it then disappeared, during the evolution of archaea and some types of bacteria", concludes Patrick Motte.

The importance of the phylogenetic study conducted by the two teams of the professors in Liège has drawn the attention of the scientific community. Hence, it not only featured on the cover of the *Plant Physiology* journal last February, but it also attracted the attention of Faculty of 1000, an international scientific organisation with more than 10 000 experts in biology and medicine who assess the quality of publications in these domains. It was very well received ("must read") and recommended as an example for other phylogenetic studies to follow.