

Pruning the Tree of Life

As classification goes, life forms can be put into three categories or domains; either you're a bacterium, an archaeon or an eukaryote. Or is there a mysterious fourth lineage?

Life, as we all know, is seldom simple. Classifying life, it seems, is even trickier and has been a matter of debate for the last 300 years or so. Back in 1735, the phylogenetic tree of life had only two branches: 'vegetabilia' aka plants and 'animalia'. It was none other than world-famous botanist Carl Linnaeus, who crowned these first two kingdoms. The two endured for more than 100 years until Ernst Haeckel recognised the 'royal claims' of a new life form, the protists (tiny eukaryotic microorganisms).

Fluctuating numbers

The tree of life Haeckel planted in 1866 now featured three kingdoms. Sixty years later, Édouard Chatton wasn't content with kingdoms anymore and established two empires, prokaryota and eukaryota. Then everything went very quickly, Herbert Faulkner Copeland proposed four kingdoms in 1938, including monera (prokaryotic organisms), protists, plants and animals. In 1969, Robert Whittaker gave fungi their own realm, the fifth kingdom. And a sixth kingdom was annexed by Woese *et al.* in 1977, the archaeobacteria. However, those six kingdoms only reigned for 13 years. In 1990, Carl Woese, Otto Kandler and Mark Wheelis revised the six taxonomic groupings and proposed a three-domain system based on ribosomal RNA (rRNA) sequencing analysis (*PNAS*, 87(12):4576-9). These three domains are nowadays the most widely accepted classification system. It differentiates at its highest level between bacteria, archaea and eukarya. Lately, though, new discoveries and advancements in analytical methods caused the leaves of the currently thriving tree of life to rustle distinctly and the question has arisen, how long will it continue to withstand the winds of change?

"There is still an ongoing debate as to whether there really are three domains – it is unclear, whether we should treat archaea and eukaryotes each as a separate domain, or whether eukaryotes are, in phylogenetic terms, just an odd branch of the archaea," says Anthony Poole from the University of Canterbury, New Zealand, who together with his colleagues from the universities of Paris, Lyon and Marseille, France has studied the origin of eukaryotes and their evolutionary relationship with the Archaea (*Nature Rev Microbiol*, 8:743-52).

But there are also efforts to expand the three-domain system to include a fourth. Recently, not one but two groups claimed to have found members of that mysterious fourth domain, different from all thus-far discovered life forms. One of the groups is headed by Jonathan Eisen from the University of California Davis (fittingly, Eisen also hosts the popular science blog *The Tree of Life*), the other group, led by Didier Raoult, is located at the

Université de la Méditerranée in Marseille, France. Interestingly, results from both groups were published in the open access journal *PloS ONE* – the Eisen paper: *PloS ONE*, 6(3):e18011 and the two Raoult papers: *PloS ONE*, 5(12):e15530, *PloS ONE*, 6(4):e18935.

Let's have a look first at the findings of the guys from California, which could well have been under the motto "Anyone who goes traveling has stories to tell"...

And it came to pass in the years 2004-2006 AD that a certain Mr. Venter embarked, with his 29m private yacht 'Sorcerer II', on a "voyage of microbial discovery". The scientific sail or, officially, the Global Ocean Sampling (GOS) expedition took him around the world and Venter and his team ended up with huge amounts of metagenomic data. This data, of course, needed to be analysed, which is where Jonathan Eisen and his team came in. Using phylogenetic marker genes (*ss-rRNA*, *recA*, *rpoB*), he went on a hunt for evolutionarily novel sequences. Even though rRNA didn't turn out to be very suitable for building "robust phylogenetic trees", the other two genes, *recA*, also dubbed the 'evolution gene' – coding for a protein involved in DNA recombination and repair – and *rpoB* – coding for a subunit of the RNA polymerase – however, appeared to be good for some surprises.

A microbial voyage

Eisen remembers in his blog, "I built a better RecA tree by first pulling out all possible homologs of RecA and RecA-like proteins from the GOS data and then building an alignment and a tree. And there they were! Some very fascinating novel RecAs – distinct from any previously known RecA-like proteins, as far as I could tell." He continues, "We then looked at RpoBs and found novel ones, too."

But does this really mean that the current taxonomic classification of life has to be revised on the spot? Not yet because the actual organismal source of the unknown sequences is still hidden in the dark depths of the oceans. But one can speculate. And so Eisen *et al.* came up with four possible explanations. The first one being that they are just artefacts (which can, of course, never be fully excluded) but assuming the sequences are real, three more alternatives exist. They could be derived from either novel viruses, be ancient paralogs of the marker genes or, as already suggested, be from a totally new domain of life. "We just don't know," writes Eisen and adds, "Obviously, if there is a novel lineage of cellular organism out there, well, that would be cool."

Didier Raoult, meanwhile, has already found what he was looking for. His contenders for being the first representatives of a

fourth domain of life are viruses – large DNA viruses to be exact or to be even more precise Nucleocytoplasmic Large DNA viruses (NCLDV).

F*%&ing awesome discoveries

But are viruses even “life forms”? Or are they the biological equivalent of ‘Schrödinger’s cat’ – dead and alive? Already in 2009, Jean-Michel Claverie and Hiroyukio Ogata, like Raoult from the Université de la Méditerranée in Marseille, France (and others) engaged in a hefty battle of words with David Moreira and Purificación López-García from the Université Paris-Sud, France. Both sides devised “ten (good) reasons” to either include or exclude viruses/giruses from the tree of life, respectively (*Nat Rev Microbiol.* 7(8):615; *Nat Rev Microbiol.* 7(4):306-11). There was no clear winner. Even a survey on a virology blog (www.virology.ws) didn’t create clarification: Out of 2,272 responses 34% of the survey participants said that viruses are not alive, while 29% said they are and 32% said they are “something in between”. Eckard Wimmer, German born virologist from the State University of New York, USA cuts right to the chase of the matter. When he, at a seminar, was asked whether viruses are dead or alive, he simply said “yes”.

Fact is, viruses are not able to self-replicate or self-maintain – two life-defining abilities. But, as one theory (Regressive Hypothesis) goes, maybe they once, a long, long time ago, were able to do so and only secondarily lost all their zest for life, evolving into a mere “biological entity”. Anyway, giant DNA viruses are different from your typical, garden-variety virus, in which they are, first and foremost, very large; so large that you can even see, at least one of them, by light microscopy. This giant answers to the name of Mimivirus, short for microbe-mimicking virus. It was first discovered, and back then misidentified as a bacterium, in a cooling tower of a power plant in Bradford, UK in 1992. Here, it spent most of its time infecting *Acanthamoeba polyphaga* amoebas. Only in 2003, did Raoult, Claverie *et al.* realise that the alleged bacterium was actually “A Giant Virus in Amoebae” (*Science*, 299:2033). Still to this day, it is the largest virus ever found, impressing with up to a record-breaking 750 nm in diameter and not less than 1.18 kilobases of genetic material coding for more than 1,000 genes. Backed by those unique characteristics, large DNA viruses obviously deserved their own name – and it was quickly found in “giruses”.

The viral giant

But where do giruses stand in the tree of life (or do they belong there, at all)? Curiously, a few of Mimivirus’ genes are shared with genes found in bacteria, archaea and eukaryota. But where do these genes come from? Was there a common ancestor (meaning the mimivirus family is monophyletic) or have

they been drawn from the giruses’ host by horizontal gene transfer, HGT? To find out, Raoult, who was once dubbed ‘Indiana Jones des microbes’ by the French news magazine *Le Point*, and his group in Marseille employed phylogenetic and phyletic studies of (shared) informational genes, genes that are involved with replication, transcription or translation. In contrast to house-keeping genes or operational genes, informational genes are said to be well-suited for those kinds of studies as they, presumably, are rarely subject to HGT. So the ‘Marseillais’ built “phylogenetic trees of eight proteins involved in different steps of DNA processing” including ribonucleotide reductase (RNR), topoisomerase II a (TopoIIA) and Proliferating Cell Nuclear Antigen (PCNA) as well as transcription factor II B (TFIIB). The result? Trees built from four of the analysed sequences supported their hypothesis that “the core genome of NCLDV is as ancient as the three currently accepted domains of life” and that it confirms the existence

of a NCLDV clade, which is “emerging from the rhizome of life with roots arising at the very beginning of life”.

Only a few months later, in April, the French boldly assigned a new member to their self-proclaimed fourth domain, the *Cafeteria roenbergensis* virus (CroV). This newly discovered family member of Mimivirus infects a marine flagellate. “The gene content of CroV unambiguously allows its classification as a new member of the fourth domain of life, along with other NCLDVs,” write the authors, pointing once again to the monophyletic origin of their favourite study objects and the ‘core genome’ they share with cellular life.

Cafeteria at the beginning of life

But “the presence of genes shared between viruses and cellular lineages suggests this holds for those shared genes but it does not readily follow that this explains the origin of those viruses, by vertical descent with modification from a cellular lineage, especially given they have picked up genes from a wide variety of sources. Establish-

ing the evolutionary origin of NCLDVs is a fascinating subject and the answer will undoubtedly be far more interesting than that suggested by the fourth domain hypothesis,” comments Anthony Poole. Further objection to the four domains hypothesis proposed by Raoult *et al.*, has recently been raised by a group from the University of Newcastle, UK (*PLoS ONE*, 6(6):e21080). Tom Williams, Martin Embley and Eva Heinz make this clear already in their paper title “Informational Gene Phylogenies Do Not Support a Fourth Domain of Life for Nucleocytoplasmic Large DNA Viruses”.

The trio re-analysed the data set of the French and found that the evolutionary models Raoult and co. used were “not fit” for a proper tree reconstruction. Analysing ancient evolutionary events is “notoriously difficult”, Williams *et al.* write. This, according to the authors, can be traced back to “variation in evolutionary rate and base composition among the lineages” being compared. One of the parameters not adequately modelled is homoplasy – meaning that two lineages “accidentally” evolved a similar gene sequence. As there are only four basic bases (that we know of), this



occurs way more often when the evolutionary rate is high. Thus, these similarities can lead one or the other scientist to believe that two branches of the phylogenetic tree are closely related, which, in fact, they are not.

Vertically or laterally?

Traditional evolutionary models (the ones Raoult and co. employed) do not properly take this into account but there are “more realistic” models that are able to do that and those were exactly the ones used by Williams *et al.* “Heterogeneous models, such as used in our study, have been developed only recently and are not widely used yet, but analyses like ours emphasise the importance of model-fitting in phylogenetics,” writes Eva Heinz in an email to *Lab Times*. Hence, “When features were accommodated by better models, the support for an NCLDV fourth domain effectively disappeared.” It turned out that the NCLDVs either emerged from within the eukaryotes or it was impossible to resolve their evolutionary relationships. Heinz concludes, “Our results are consistent with the null-hypothesis that these genes were acquired by lateral gene transfer from eukaryotes, and so there is no need to invoke a fourth domain. The concept of the tree of life depends on vertical descent as opposed to lateral gene transfer. To place the NCLDVs on the tree would require the identification of a set of genes vertically inherited from their ancestor with other forms of life. Our analyses suggest that the informational genes analysed by Boyer *et al.* may not represent such a set.” Unsurprisingly, neither Williams *et al.* nor *Lab Times* have so far heard anything back from the Raoult group.

So the Newcastle team clearly put the NCLDVs out of contention as the first members of the fourth domain. However, the organism that harbours Jonathan Eisen’s unknown sequence still has to be snatched from the moist clutches of the sea. In a Google knol, he recently revealed that the enigmatic sequence “is distinct” from those in large DNA viruses. But “as more and more unusual viruses are continuing to be discovered, the novel metagenomic sequences we found could be from novel uncharacterised viruses. However, we still are entertaining the possibility that some of these novel sequences are from unknown novel branches of cellular organisms”, the knol entry reads.

Timber!!

Eisen *et al.* therefore want to continue “exploring their metagenomic data” with more phylogenetic markers to come a little closer to the riddle’s solution. Anthony Poole, though, adds for consideration, “Establishing between a deep-diverging group and a different domain is not possible on phylogeny alone. Regarding the discovery of the archaea, their phylogenetic status as a domain was given weight because known archaea had biological characteristics that separated them from bacteria (they were all methanogens with a distinct cell wall and membrane lipids). Without this type of biological information, we cannot make the distinction.”

Thus, it’s unquestionable that many more compelling pieces of evidence first have to be brought to the table before any more conclusions can be drawn on the origin of life. Carl Woese, the man who came up with the three domains 21 years ago, however, thinks that time has already come to prune the tree of life again. “Serious consideration of the issue is long overdue,” he wrote in an email to *Lab Times*. Whether there will be two, three, four or even more domains in the future, only a resourceful taxonomist might one day be able to tell.