What genomes have to say about the evolution of the Earth

Bastien Boussau, Manolo Gouy

What genomes have to say about the evolution of the Earth

Bastien Boussau[1,2] and Manolo Gouy[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, Université de Lyon, Université Lyon 1, CNRS, UMR5558, Villeurbanne, France
[2]Department of Integrative Biology, UC Berkeley

Abstract

The geological record provides an irreplaceable account of the joint history between the Earth and living organisms. Extant living organisms also contain in their phenotypes and most importantly in their genomes information about their history, and about the history of the Earth. In this review we explain how biologists attempt to extract this information and draw inferences about past history, using statistics, computer algorithms, and molecular biology. We show that inferred ancestral gene contents provide insights into ancient metabolisms, ancestral genome composition in bases or amino-acids provide information about ancient growth temperatures, and protein resurrection offers means to investigate the function of proteins long disappeared. All these inferences throw a new light on organism and Earth evolution. Their combination and the use of statistical models integrating both genomic and geologic histories hold great promises for unveiling more of the past 4 billion year history on Earth.

Introduction

For most of its history, Earth has been home to evolving organisms, which recorded in their genomes traces of its transformations, and contributed to these transformations. Since 1995 and the first genome sequence of a cellular organism (Fleischmann et al., 1995), this information has been made available for a large number of species. The comparative analysis of these genome sequences provides means to infer ancient events in the history of the Earth. Questions that can be addressed using genomes include: what were the ecological properties of the ancestors of extant organisms? When did a particular metabolic function arise? When did major radiations or extinctions occur? These questions have a direct bearing on geological history: the ecology of an organism informs about the nature of its environment, its metabolism informs about the chemicals available and produced, and the rock record may contain explanations as well as traces of massive radiations and extinctions.

Here, we review the historical information that can be gained from the study of genome sequences, and the methods that are used to extract this information. To set the stage, we rapidly sum up our knowledge of the early evolution of life on Earth, according to the geological record. This quick summary shows that there are considerable gaps in the geological record, especially in the Archaean, and highlights the need to use all the data at hand to learn more about the early co-evolution of life and the Earth. Following sections review different ways to analyze genetic data and draw inferences about ancestral organisms, and hence the ancient Earth. First, we review

how molecular phylogeny provides means to reconstruct and date the tree of life, which may hint at past events of radiations and extinctions, and importantly gives means to discover the succession of organisms in the past few billion years. Second, we describe methods to reconstruct ancestral characters of living organisms based on the knowledge of the characters of extant organisms. Such methods have been used on morphological characters, but can also be used on genomic data. This provides simple ways to infer lifestyles of organisms that are no longer available for observation. Third, we dive deeper into genomics, and show that it is possible to infer gene contents of ancient organisms, using methods described in the preceding section, or using more sophisticated methods specifically designed for modeling gene content evolution. Ancestral gene contents provide powerful means to assess ancient metabolic capacities, as we show in a few examples. Fourth, we introduce another type of genome analysis, where ancestral nucleotide or amino-acid contents are used to learn about ancient ecologies. We show how such a technique has been used to infer the history of growth temperature along the tree of life. Fifth, we end our review with a gene-specific type of analysis, gene resurrection. In this approach, molecular biologists use computational methods to estimate the ancestral sequence of a gene, then synthesize the protein, and eventually perform wet-lab experiments on this protein to establish its properties. Such an approach can give insight into the ecology, metabolism or inner functioning of an organism, depending on the gene under study. Although each of these five methods can independently bring crucial insight into the early co-evolution of life and the Earth, we plead for an integrative approach, where both biologists and geologists cooperate to improve our knowledge of our very distant past on this planet.

What rocks say about early evolution

Ancient rocks suggest that life has existed on Earth for more than 3.5 billion years (Schopf, 2006). Some authors argue that the Earth could have been habitable a mere 10-20 My after the moon-forming impact (Zahnle et al., 2007) about 4.5 bya, and that micro-organisms could have survived the late heavy bombardment about 3.8 bya (Abramov and Mojzsis, 2009). It is therefore possible that extant organisms are the descendants of a cell or group of cells (LUCA, the Last Universal Common Ancestor, the ancestor of al extant life forms on Earth) that lived about 4 bya. Unfortunately, so far there is no evidence to reject the hypothesis that LUCA lived at 3.8, 4, or 4.5 bya. In this first section, we briefly review some of the inferences drawn from the rock record that have a direct bearing on the co-evolution between Earth and living organisms in the last 4 by. Notably, we focus on the geological record of temperature and atmospheric oxygen content on Earth, as both factors have a major impact on living organisms. We then briefly list fossil evidence for some ancient groups of organisms.

Figure 1 shows a phylogenetic tree of life and a timeline of the evolution of the Earth, where some of the major early events have been highlighted.

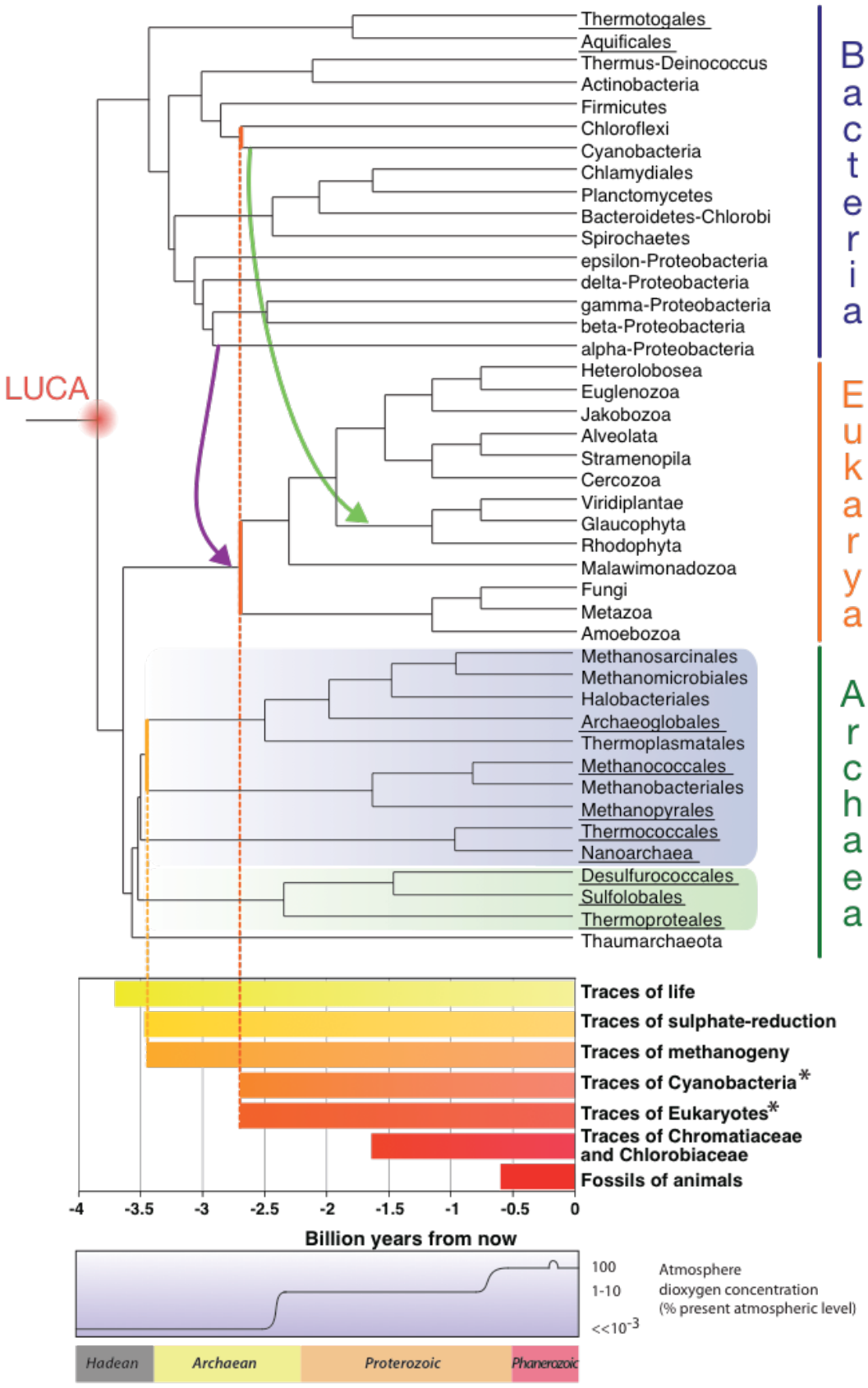

Thermotogales
Aquificales
Thermus-Deinococcus
Actinobacteria
Firmicutes
Chloroflexi
Cyanobacteria
Chlamydiales
Planctomycetes
Bacteroidetes-Chlorobi
Spirochaetes
epsilon-Proteobacteria
delta-Proteobacteria
gamma-Proteobacteria
beta-Proteobacteria
alpha-Proteobacteria
LUCA
Heterolobosea
Euglenozoa
Jakobozoa
Alveolata
Stramenopila
Cercozoa
Viridiplantae
Glaucophyta
Rhodophyta
Malawimonadozoa
Fungi
Metazoa
Amoebozoa
Methanosarcinales
Methanomicrobiales
Halobacteriales
Archaeoglobales
Thermoplasmatales
Methanococcales
Methanobacteriales
Methanopyrales
Thermococcales
Nanoarchaea
Desulfurococcales
Sulfolobales
Thermoproteales
Thaumarchaeota
Bacteria
Eukarya
Archaea
Traces of life
Traces of sulphate-reduction
Traces of methanogeny
Traces of Cyanobacteria*
Traces of Eukaryotes*
Traces of Chromatiaceae and Chlorobiaceae
Fossils of animals
-4
-3.5
-3
-2.5
-2
-1.5
-1
-0.5
0
Billion years from now
100
1-10
<<10-3
Atmosphere dioxygen concentration (% present atmospheric level)
Hadean
Archaean
Proterozoic
Phanerozoic

Figure 1: Some elements of the evolution of life and of the Earth. From top to bottom: phylogeny of the tree of life, timeline showing the earliest evidence for a few phyla, atmospheric oxygen content through time, and geological ages. The phylogeny of Bacteria is as in (Boussau, Guéguen, et al., 2008b), the phylogeny of Archaea has been compiled from (Gribaldo and Brochier-Armanet, 2006; Brochier-Armanet et al., 2008; Elkins et al., 2008), and the phylogeny of Eukarya from (Rodríguez-Ezpeleta et al., 2007). Some nodes of the tree have been constrained to agree with dates from the fossil record. Dates associated with non-constrained nodes should be ignored. Primary endosymbioses have also been represented with arrows indicating the direction of transfer: purple for the origin of mitochondria, and green for the origin of chloroplasts. Phyla harboring hyperthermophilic organisms have been underlined. Crenarchaeota are on a green background, and Euryarchaeota on a blue background. *: controversial evidence (see main text). The evolution of atmospheric oxygen has been simplified from (Lyons et al., 2009).

Evolution of temperature at the surface of the Earth

Temperature is a major factor influencing all chemical reactions and consequently life forms: the timeline of temperature on the Earth should therefore be intimately linked to the evolution of life. The analysis of oxygen and silicon isotopes first suggested that the archean oceans around 3.5 bya were hot, with temperatures above 55°C (Knauth and Lowe, 2003; Robert and Chaussidon, 2006). This was viewed as supporting the idea that early life may have been thermophilic or hyperthermophilic, *i.e.* lived at high (>60°C) or very high temperature (>80°C). However, more recent studies support instead cooler oceans at 3.42 bya, with temperatures below 40°C (Hren et al., 2009; Blake et al., 2010). Such temperatures obviously do not preclude a mesophilic lifestyle for the earliest organisms, but can still be consistent with a thermophilic origin of life, earlier than 3.42 bya, or next to hydrothermal vents. Overall, throughout the Archaean, except for some glaciations (Kasting and Ono, 2006) and the occasional meteorite, temperature at the surface of the Earth may on average have been not very different from what it is today.

In the last billion year, the Earth may have gone through "Snowball Earth" stages, notably around 715–680 mya and around 680–635 mya (Maruyama and Santosh, 2008). This latter episode immediately predates the radiation of the Ediacaran fauna, and may thus have contributed to the emergence of this important multicellular fauna.

Contrary to earlier records, the fossil record for the last 550 my contains fossils of phosphatic and calcitic shells, from which isotopic measurements are used to estimate complete and more precise timelines for temperatures at the surface of the Earth (Veizer et al., 2000; Zachos, 2001; Royer et al., 2004)(Bijl et al., 2010). These data show that there is a strong relation between temperature and CO2 content in the atmosphere (e.g. (Bijl et al., 2010)), and suggest some climatic "aberrations", in which temperature reached an extreme value for a short amount of time (1000 to 100,000 years), and that seem to have had an impact on life, triggering radiations and extinctions for different groups of organisms (Zachos, 2001).

Evolution of the atmosphere

The composition of the early atmosphere, before life appeared, is a matter of active debate (e.g. (Tian, 2005; Kasting and Howard, 2006; Ueno et al., 2009; Wolf and Toon, 2010)). However, it may have contained a mixture of CO2 and CH4 (Kasting and Howard, 2006), creating a greenhouse effect. This greenhouse effect may be sufficient to explain the "faint young sun paradox" (Sagan and Mullen, 1972), which describes the contradiction between an early sun less active than it is today, and the presence of

liquid water at the surface of the earth, as attested by oxygen isotopes in ancient zircons. Living organisms later drastically changed the composition of the atmosphere, by producing more methane (see (Kasting and Howard, 2006) for a more thorough discussion), and of course oxygen.

Oxygen was injected into the atmosphere mainly through oxygenic photosynthesis, a metabolism invented by Cyanobacteria. Uncovering when oxygen was first introduced into the Earth atmosphere would thus help date the origin of this important group of Bacteria, but would also set an upper time constraint for the emergence of many groups of organisms that depend on oxygen, as Eukaryotes, which need oxygen notably to produce sterols present in their cell membranes (Summons et al., 2006). Analyses of the geological record generally support a Great Oxidation Event around 2.4 bya (Farquhar et al., n.d.), perhaps preceded by local oxygenations as early as 2.8 bya (although some researches suggest there may be traces of oxygenated oceans as early as 3.46 bya (Hoashi et al., 2009)).

A second large increase in the oxygen content of the atmosphere occurred around 0.85–0.54 Ga (Holland, 2006), and has therefore been linked to episodes of Snowball Earth and to the ediacaran radiation. Oxygen levels then reached a maximum at the end of the carboniferous period, at about 300 mya. Such oxygen levels could be linked to the emergence of vascular land plants, and explain the appearance in the fossil record of giant insects (Lane et al., 2010).

Appearance of major groups of organisms

The oldest pieces of evidence for particular groups of organisms come from (Ueno et al., 2006) and (Philippot et al., 2007), who used carbon and sulfur isotopes respectively to propose that methanogens (archaea) and sulfur-disproportionating bacteria existed as early as around 3.5 bya. Therefore, as early as 3.5 bya, Archaea and Bacteria had already diverged, and started to diversify.

(Rasmussen, 2000) proposes evidence that thermophilic organisms existed by 3.235 bya, in a submarine hydrothermal setting. 2.707 to 2.685 bya, (Ventura et al., 2007) provide evidence that bacteria and archaea co-existed in a subsurface hydrothermal environment. These works suggest that hyperthermophilic organisms are very ancient, possibly about 3.2 by old.

The Great Oxidation event at around 2.4 bya provides evidence that Cyanobacteria had already been pumping oxygen into the atmosphere in large quantities by that time. The earliest clues for their existence may come from molecular biomarkers (Summons et al., 2006; Eigenbrode et al., 2008; Waldbauer et al., 2009; Schopf, n.d.) dating from 2.67 to 2.46 bya, although these results have been disputed (Rasmussen et al., 2008). In addition, stromatolites have been found in rocks dating from 3.46 bya (Allwood et al., 2006). However, although stromatolites nowadays are dominated by Cyanobacteria, one cannot link these oldest fossil stromatolites to early Cyanobacteria, as they could have been produced by other types of bacteria instead (Schopf, n.d.).

Biomarkers were also used to propose that two groups of Bacteria, Chromatiaceae and Chlorobiaceae, were present in a 1.64 billion year old basin in Australia (Brocks et al., 2005). The presence of these bacteria indicates that by 1.64 bya, there were anoxic and sulphidic deep water basins on Earth.

Eukaryotes are thought to have entered the Earth stage later than Bacteria and Archaea, but it remains difficult to date their appearance. The same disputed molecular biomarker studies that were used to date Cyanobacteria also hint at an appearance for

Eukaryotes (Eigenbrode et al., 2008; Waldbauer et al., 2009) around 2.67 to 2.46 bya. However, (Knoll et al., 2006) place the earliest morphological evidence of Eukaryotes at about 1.8 bya. This state of affair is further complicated by the recent discovery of large complex fossils that may be remnants of multicellular Eukaryotes from 2.1 bya (Albani et al., n.d.), and by large unicellular fossils that might be ancient Eukaryotes from 3.2 bya (Javaux et al., 2010). If these fossils are correctly interpreted as Eukaryotes, then Eukaryotes may have appeared before the Great Oxidation Event, even if their membranes contain sterols, whose synthesis requires oxygen (Summons et al., 2006)(Summons et al., 2006). In this case, these early Eukaryotes may have not used sterols in their membranes, or may have lived in particular environments where local concentrations of oxygen were high enough for sterol synthesis.

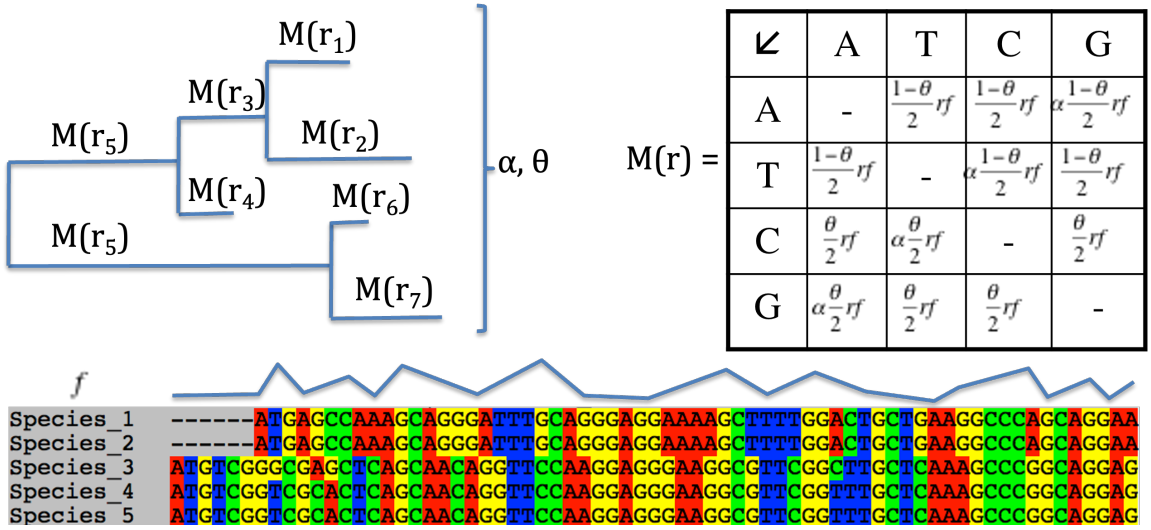
This rapid summary of current geological knowledge about the early evolution of the Earth shows that although data is accumulating quickly, much remains controversial and uncertain. It is notably very difficult to date the origin of particular groups of organisms. In such a context, insights that can be gained from the study of the genomes of extant organisms are very valuable. Genomes contain millions (for Bacteria and Archaea) to billions (for some Eukaryotes) of DNA bases whose evolution may have recorded footprints from ancient environments. In this manuscript we will review how the study of the evolution of organisms and of their genomes has informed and could inform about the evolution of lifestyles in the last 4 billion years, and notably about temperature and oxygen content on Earth.

Molecular Phylogeny

Phylogenetics aims at reconstructing the evolutionary relationships among organisms (Felsenstein 2004). The resulting relationships are most often in the form of a tree, that is, studied organisms are positioned within an acyclic graph whose nodes represent ancestral or extant organisms and whose edges represent evolutionary lineages from an ancestor to one of its descendants. Although phylogenetics has been initially applied to morphological data, it is now most often applied to molecular data, *i.e.,* genomic and protein sequences.

Molecular phylogenetics works by applying tree-building methods to molecular sequence data. The sequences under study must be homologous, that is, must all derive from a common ancestral sequence that has been transformed during evolutionary time along the lineages that connect this ancestral sequence to studied sequences. Sequences are next aligned, that is, homologous residues (nucleotides or amino acids) that share a common ancestral residue, are identified in all sequences. The final step of a phylogenetic analysis infers a phylogenetic tree from the data set of aligned sequences. This can be done through one of a variety of methods that all make some kind of regularity assumptions about the evolutionary process, without which no reconstruction of ancient evolutionary events from extant sequence data would be possible. These methods can be usefully distinguished according to the regularity assumptions they make. The parsimony method has been historically employed first. It assumes that evolution occurs slowly

enough that the possibility that the same site has been transformed several times in the same lineage can be neglected, and that all parts of the molecule and all lineages evolve at similar rates (Felsenstein, 1988). Under these assumptions, the tree that requires the smallest total number of changes to explain the observed sequences is the phylogenetic tree of the compared sequences. The assumptions made by parsimony are unrealistic when evolutionary relationships between deeply divergent organisms are sought. In this case, the most efficient methods presently known are those that are based on an explicit probabilistic model of the evolutionary process. The probabilistic modeling of phylogenetics (Felsenstein, 1981) has been developed in the maximum likelihood (Guindon et al., 2010; Jobb et al., 2004; Stamatakis et al., 2005; Swofford, 2003) and in the Bayesian frameworks (Ronquist & Huelsenbeck, 2003). Fig. 2 gives an example of a probabilistic model often used to describe the molecular evolutionary process at the DNA level.



↙	A	T	C	G
A	-	$\frac{1-\theta}{2}rf$	$\frac{1-\theta}{2}rf$	$\alpha\frac{1-\theta}{2}rf$
T	$\frac{1-\theta}{2}rf$	-	$\alpha\frac{1-\theta}{2}rf$	$\frac{1-\theta}{2}rf$
C	$\frac{\theta}{2}rf$	$\alpha\frac{\theta}{2}rf$	-	$\frac{\theta}{2}rf$
G	$\alpha\frac{\theta}{2}rf$	$\frac{\theta}{2}rf$	$\frac{\theta}{2}rf$	-

Figure 2: A data set of five homologous, aligned DNA sequences and the probabilistic model used to represent their evolutionary history. In this model, the evolutionary process is characterized at each site of the molecule and each branch of the tree by a rate matrix M whose terms are the probability per unit time of the replacement of a nucleotide by another one. M depends of four positive parameters: parameters r and f determine the overall rate of evolution, and parameters θ and α modulate the relative values of substitution rates between different nucleotides (θ varies between 0 and 1). The first parameter (r) is free to vary at each branch of the tree, thus allowing for changes of the evolutionary rate among lineages. Parameter f is supposed to vary across sites according to a gamma distribution of mean 1, thus allowing for a molecule that does not evolve with equal rate at all its sites. The last two parameters (θ and α) are shared by all branches and all sites, which is a typical regularity assumption made by phylogenetic methods.

In most phylogenetic trees, the length of a branch represents the quantity of molecular evolution that occurred between the two branch extremities. A branch

length is expressed in substitutions/site, the average number of times each site of the molecule has been transformed along the branch. Consider two branches that stem from an ancestral organism and end at two extant organisms. The evolutionary times spanned by both branches are equal. Therefore, length differences between these two branches express substitution rate differences between the two lineages. In some cases, branch lengths differences do not exceed the variation expected from a random process of molecular evolution ticking at the same rate on all tree branches. This is the evolutionary clock (Zuckerkandl and Pauling, 1965). If the clock assumption holds, and if the fossil record gives a date for an ancestral node of a molecular phylogenetic tree, it becomes possible to date all the other nodes of the tree using a simple proportionality rule between evolutionary time and molecular divergence. A phylogenetic tree with dates for all its nodes is called a chronogram.

The evolutionary clock assumption has been extensively applied because it allows one to estimate the date of origin of groups of organisms for which no fossil record is available, most remarkably microorganisms. It also enables extrapolating back in time using the geological ages of later fossils: for example, the date of the divergence between jawed and jawless vertebrates has been extrapolated from the paleontology-given age of the birds/mammals split (Kumar and Hedges, 1998).

The evolutionary clock assumption, however, is known not to be realistic in a number of cases (Graur and Martin, 2004). In general terms, this assumption rarely holds when the species under consideration are not all very closely related. The current method to combine molecular trees and dates given by the geological record is to apply a so-called relaxed clock where the evolutionary rates are allowed to vary between lineages according to a pre-defined model (Kishino et al., 2001; Drummond and Rambaut, 2007). The autocorrelated relaxed clock model is based on the observation that the molecular evolutionary rate is heritable. Therefore, two sister lineages will tend to have more similar molecular evolutionary rates than two evolutionarily distant lineages (Kishino et al., 2001). The rates of two sister branches are modeled by randomly sampling in a Gaussian distribution centered on the rate of the parental branch. The variance of these distributions is assumed to be the same for all branches, and its value is fitted to the sequence data. Alternatively, the uncorrelated relaxed clock assumes that each branch evolves with a rate drawn from a unique *a priori* probability distribution (Drummond et al., 2006). These clock models have been implemented in the Bayesian statistical framework. Fossil record dates, expressed as temporal intervals containing given internal nodes of the phylogenetic tree, are included in the model by altering the prior distribution of divergence dates. This approach has been used, for example, to date the molecular tree of the eukaryotic domain, using 129 protein coding genes sequenced from 36 eukaryotic species sampling the eukaryotic diversity, calibrated with six paleontological dates (Douzery et al., 2004).

A chronogram can furthermore be used to date the origin of ecological traits for which no fossil record evidence is available. This type of approach has been followed by Bytebier *et al.* (2011) to date the origin of the modern fire regime of the Cape flora in South Africa. They built a dated molecular phylogeny of orchids, and placed the origin of the fire regime at the age given by the phylogeny to the last common ancestor of two obligate post-fire flowering clades.

Overall, the knowledge of the species phylogenetic tree is a crucial foundation for the proper understanding of the evolution of all biological characters that can be linked to the varying environment of the ancient earth. It is particularly useful to have a dated phylogenetic tree, where nodes can be related to particular strata in the geologic record. Unfortunately, the geologic record offers few identifiable fossils for the first few billion years of evolution on Earth. As a consequence, new techniques need to be developed to date the tree of life when fossils are rare, and much research still has to be done.

Accurately dated phylogenetic trees can be used to study the dynamics of species diversification and extinction. If a species tree contains a good representative sample of extant species, then it may be possible to identify when events of diversification and extinction took place, by looking at the number of extant species, and at the distribution of speciation events through time. This type of phylogenetic tree shape analysis is currently the subject of much research, and it is still unclear how much information can be extracted. For instance, it has been recently shown that extinction rates could not be estimated from phylogenetic trees when speciation and extinction rates were not constant over a given phylogeny (Rabosky, 2010). However, such tree shape analyses have been used to study diversification in different groups of species. For instance, Alfaro et al. (2009) studied the pattern of diversification in vertebrates, and identified six clades where the diversification rate had been particularly high, and three clades where the diversification rate had been particularly low. These results could be linked to the geological record, to test whether global changes are correlated with changes in patterns of diversification.

Reconstruction of ancestral phenotypic characters

If the phylogenetic tree of a group of species is known, and if the state of a particular phenotypic character (*e.g.,* capacity to live in the presence of oxygen, or at high temperature) is known for each species, one can attempt to estimate the states of this character at all ancestral nodes of the tree. The parsimony method (Felsenstein, 2004) can be readily employed for discrete characters: the ancestral character states that minimize the total number of character changes along the tree are computed and used as estimation of ancestral character states (Fig. 3a). As previously mentioned, the parsimony model requires that the character changes slowly to be valid. If the phylogenetic tree is a chronogram (with dated nodes), a good method is to use a probabilistic model that quantifies the rates of changes between character states per unit time (Fig. 3b). The model allows then to estimate these rates, and the probabilities of each character state at each tree node.

These methods, though, have several shortcomings. The reconstruction of ancestral phenotypic characters by parsimony often results in uncertainties about the character state at some ancestral nodes because several states are equally parsimonious. More generally, a single character mapped at the tip of a phylogenetic

tree is only a small amount of information, too small to reconstruct ancestral states by parsimony or using a probabilistic method, especially if the rate of change of this character varies along the tree.

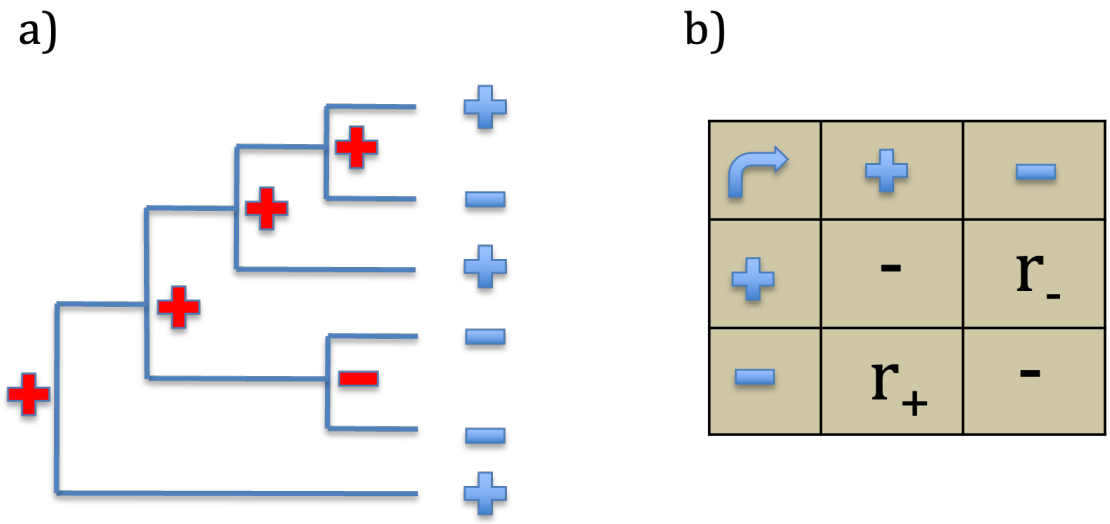


Figure 3: Phenotypic character evolution.
a) A two-state phenotypic character (blue + and – signs) mapped on the tips of a phylogenetic tree. The parsimony principle allows inferring ancestral states (red signs). b) The evolution along a chronogram of a two–state phenotypic character can be modeled using the matrix of the rates of state changes r_+ and r_- per unit time.

Reconstruction of ancestral Gene contents

Reconstructions based on numbers of genes in extant genomes

Instead of directly reconstructing ancestral phenotypic characters, and with the availability of an increasing number of genome sequences, biologists have also resorted to reconstructing ancestral genomic characteristics. In addition to gaining insight into genomic evolution, such reconstructions provide the opportunity to infer ancient phenotypic characters as well, because the phenotype of an organism is encoded in its genome. Given the genome of an organism, one can in principle infer its lifestyle and metabolism, for instance.

In practice, instead of reconstructing the entire genome sequence for ancestral organisms, researchers often only reconstruct ancestral gene contents, *i.e.* the number or the presence/absence of genes in ancestral genomes. In fact, reconstructing entire genome sequences is more difficult than reconstructing gene contents: on one hand, every single base of a genome must be reconstructed, which amounts to reconstructing the evolution of millions or billions or characters, whereas on the other hand, reconstructing ancient gene contents only requires analyzing a few thousand characters.

Genes can be grouped in gene families, *i.e.* groups of genes that descend from a common ancestral gene. A gene family can be present in all organisms, like the small subunit ribosomal RNA (rRNA) gene family, or can be present in just a few species, like genes coding for collagen, which are not found outside animals. Therefore, in a given species, a gene family can be present or absent. If present, it can appear in several copies: for instance there are several sorts of the collagen gene in humans, which differ by some bases, and have slight functional differences. Such facts show that a gene can be created, like collagen when animals appeared, lost, or duplicated (Figure 4a-c). In addition, a gene can be transferred from one organism to another, as is commonly observed when pathogenic bacteria acquire genes providing resistance against antibiotics from other bacteria (Figure 4d). Overall, gene family evolution is difficult to model accurately, and only recently have realistic probabilistic descriptions been proposed.

The first approaches to inferring ancestral gene contents considered genes as phenotypic characters (Figure 4e). Under this assumption, the same models that were used to infer ancient phenotypic characters could be used to infer ancient gene contents. Therefore, parsimony or model-based approaches have been used to infer the presence/absence or the number of genes in ancient organisms. For instance, (Snel et al., 2002) and (Mirkin et al., 2003) used a simple parsimony approach to infer gene contents along the tree of life. They tested different penalty scores associated to the events of gene duplication, gene loss, and gene gain, gain accounting for both gene creation and gene acquisition from another organism, and chose the "best" penalty scores as the ones that appeared consistent with *a priori* expectations. Although such approaches are subjective, they are probably able to extract the most robust signals contained in gene content data. For instance, (Mirkin et al., 2003) were able to reconstruct metabolic pathways that looked functional for LUCA. More precisely, using the set of penalty scores they trusted most, the genes predicted to have been present in LUCA enabled assembling fairly complete metabolic pathways, able to generate energy and essential amino-acids and DNA bases for a free-living organism. (Boussau, Karlberg, Frank, Legault, and Andersson, 2004a)(Boussau, Karlberg, Frank, Legault, and Andersson, 2004b) also used this type of approach to infer the ancestral gene content of the ancestor of alpha-Proteobacteria, and based on this gene content, concluded that this ancestor was probably an aerobic organism.

Recently, statistically more rigorous approaches have been developed. These approaches explicitly model the processes of gene family evolution in a statistical framework. For instance, (Hahn et al., 2005) have developed a model of gene duplication and loss that can be used to reconstruct ancestral gene numbers when gene transfer events can be neglected; (Cohen and Pupko, 2010) have developed a model of gene gain and loss, that can be used to infer the presence or absence of genes in ancestral genomes; (Csűrös and Miklós, 2006) have developed the most complete model so far, that accounts for gene transfer, gene duplication, and gene loss. These models have been used to characterize the dynamics of genome evolution, for instance investigating what characteristics of gene families correlate with their transferability. However, these models could also be used to investigate ecological evolution, inferring ancestral ecologies based on ancestral gene contents.

In turn, these ancestral ecologies could be contrasted with data from the geological record, and perhaps suggest that some biomarkers may be present in particular geological settings where they were thought unlikely.

Approaches using gene sequences

Approaches presented above and based on the presence or number of genes in extant genomes certainly can provide interesting insights into ancient ecological evolution, for gene families whose evolution contains very few events of gene transfers. Indeed, gene transfers are very difficult to infer correctly based on numbers of genes in extant genomes only (Figure 3d). Instead of using only gene family presence or absence, or gene number in a given gene family, one can use the information contained in the sequences of these genes. As explained in part 2, gene sequences can be used by molecular phylogenetic techniques to reconstruct the history of these sequences. The result of such an analysis applied to a gene family is a gene family phylogenetic tree. In this gene family tree, internal nodes correspond to events of speciation, duplication, creation, or transfer. The type of event associated to these internal nodes can be identified through a careful manual comparison of the gene family tree to the species tree, as has been done several times for a few gene families.

For instance, (Castresana et al., 1994) and (Brochier-Armanet et al., 2009) made phylogenetic analyses of the cytochrome oxidase families, genes involved in the reduction of dioxygen to water, and found that the ancestor of all life forms (LUCA, for Last Universal Common Ancestor) probably already contained a member of this gene family. If the ancestral gene had the same function it has today, this suggests that metabolisms able to tolerate at least trace amounts of oxygen may have existed before the great oxidation event (Gribaldo et al., 2009). Using similar manual approaches, (Becerra et al., 2007) analyzed the evolution of the Protein disulfide oxidoreductases gene family, genes important for protein stabilization at high temperatures, and found that this gene family was likely absent from LUCA. They concluded that their results did not support the hypothesis according to which LUCA was a thermophilic or hyperthermophilic organism.

These careful single gene family analyses require tedious manual analyses, and therefore cannot be applied at a genomic scale, to thousands of gene families. Fortunately, automated methods are now starting to be developed, that can reconstruct gene family trees, compare these trees to putative species trees, and therefore reconstruct the events of gene creation, duplication, loss and transfer that have affected gene families. Such an approach was recently applied to 100 genomes from Bacteria, Archaea and Eukaryota (David and Alm, 2010). It was found that a short period around 3 bya seems to have contributed about 27% of all extant gene families, and that these gene families seem to be preferentially involved in respiratory chains. Genes appearing after this short period are predicted to have a function consistent with an increasingly oxygenated atmosphere. Although these results need to be confirmed by further analysis, it is likely that other research groups will develop similar or more sophisticated approaches to study early genomic evolution. One limitation of the (David and Alm, 2010) work is that it is not fully statistical, but relies on the parsimony assumption, so that penalties have to be

subjectively associated to different types of events; future approaches will probably adopt a fully statistical approach, where the data itself dictates the gene creation, duplication, transfer and loss rates. Such approaches could be used to accurately reconstruct ancestral gene contents, at a genomic scale.

Hypothetically, these ancestral gene contents could then be analyzed through methods that use gene contents to reconstruct metabolic pathways and networks (Feist et al., 2008)(Feist et al., 2008). Such metabolic networks provide crucial insights into a species ecology. For instance (Borenstein et al., 2008) used network analyses in a set of present-day organisms whose genome was available to establish what metabolites they required for proper growth. Such an analysis done on ancestral gene contents could shed an interesting light upon what metabolites were available billions of years ago.

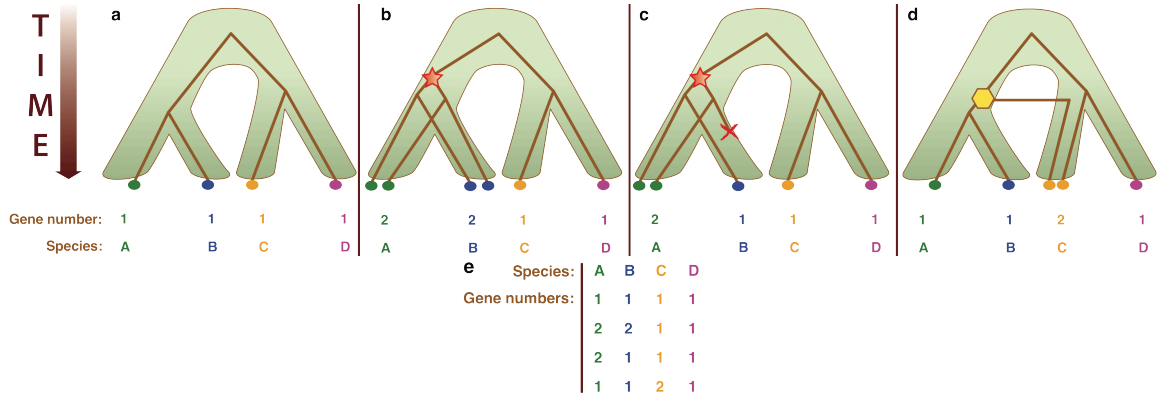


Figure 4: Genome evolution with 4 gene families and 4 species. Gene families (gene family phylogenetic trees in brown) evolve alongside a species tree (large structure in green). **a)** Gene family evolution without events of duplication, loss, or transfer. Each species A to D only has one gene. **b)** Gene family evolution with one gene duplication event ancestral to species A and B. Species A and B have 2 copies of this gene. **c)** Similar to b, with an extra gene loss event ancestral to species B. Species B therefore only has one copy of the gene. **d)** Gene family evolution with a transfer event from an ancestor of species A and B to an ancestor of species C. Species C therefore has two copies of the gene in its genome. Based on these numbers of genes only, it is impossible to establish whether a gene transfer or a gene duplication occurred in the example d. This shows that using gene tree topologies instead of numbers of genes in extant species provides better insights into genome and species evolution. **e)** Gene number matrix. This matrix contains numbers of genes for each gene family in the genomes of the four species in our example.

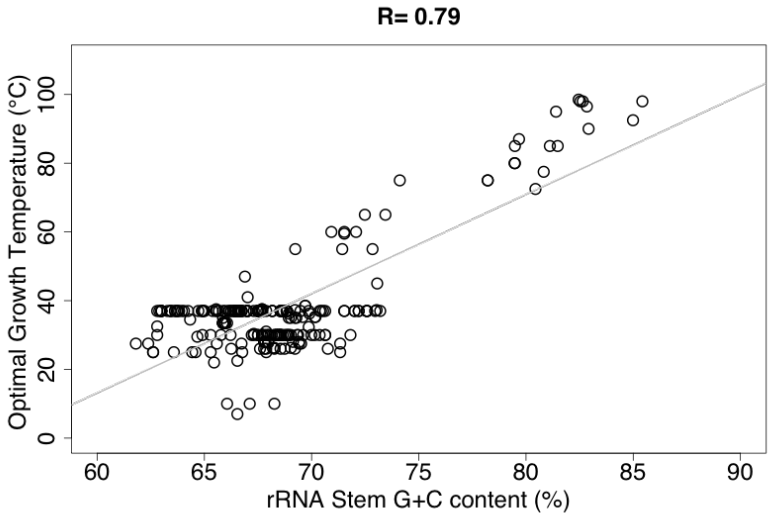
Reconstruction of ancestral nucleotide or amino-acid content

In addition to the presence/absence or number of genes in a genome, global characteristics like sequence composition in bases or in amino-acids can also point at the ecology of an organism. In fact, it was first discovered that the base composition of the ribosomal RNA (rRNA) genes in Bacteria and Archaea was correlated to the optimal growth temperature, *i.e.* the growth temperature at which the organism grows the fastest (Galtier and Lobry, 1997). More precisely, in Bacteria and Archaea, the proportion of bases Guanine (G) and Cytosine (C) in rRNA genes is higher as optimal growth temperature increases (Figure 4a). Using this correlation, it is possible to predict, based on the sequence of a rRNA gene, at what temperature

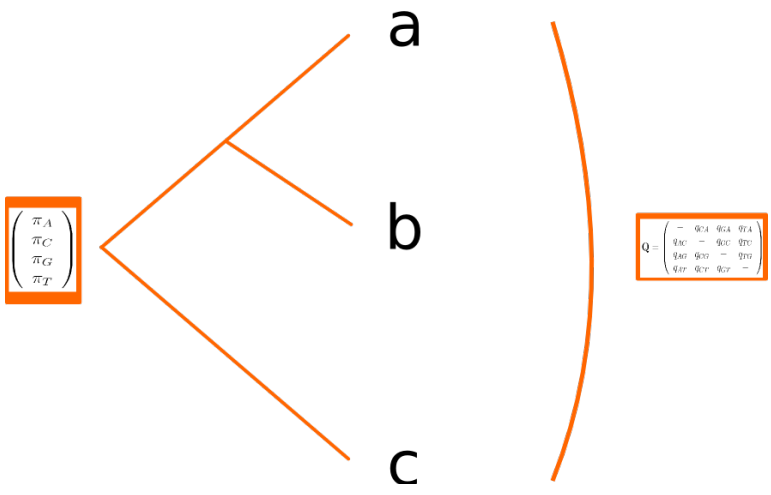
an organism grows best, and consequently approximately what is the average temperature of its environment. More recently, a correlation was also found between the amino-acid composition of all the proteins in a Bacteria or an Archaea and its optimal growth temperature (Zeldovich et al., 2007). Given the protein sequences of an organism, one can therefore infer the temperature prevailing in its environment.

This provides powerful methods to study the evolution of growth temperature along the tree of life: given ancestral sequences reconstructed using phylogenetic models (see part 2), one can infer ancestral growth temperatures. This approach was first used by (Galtier, 1999), using rRNA sequences alone, and then by (Boussau, Blanquart, et al., 2008a) using both rRNA and protein sequences. Importantly, both studies used sophisticated models of sequence evolution, which do not assume that the evolutionary process has been constant through time (Figure 4b), but instead can have changed from one branch of the tree of life to the next (Figure 4c). Both studies inferred that LUCA probably was not a hyperthermophilic organism. Further, (Boussau, Blanquart, et al., 2008a) found evidence for parallel increases in growth temperature from LUCA to the ancestor of Bacteria and the ancestor of Archaea (Figure 4d). Crucially, rRNA and protein data were consistent in their prediction: the fact that two independent markers of growth temperature provide the same inferences considerably reinforces these results. In addition, Fournier and Gogarten (Fournier and Gogarten, 2010), through careful manual reconstruction of sites in a few proteins, found that the composition for LUCA was consistent with a mesophilic lifestyle.

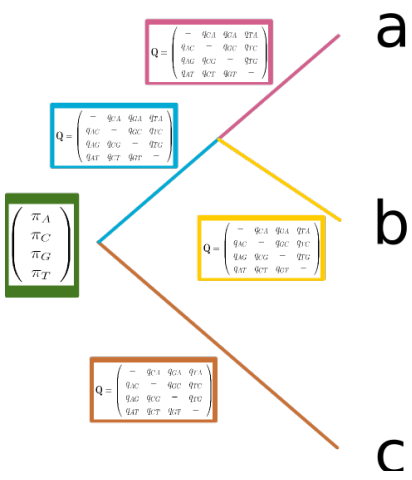
a



b



c



d

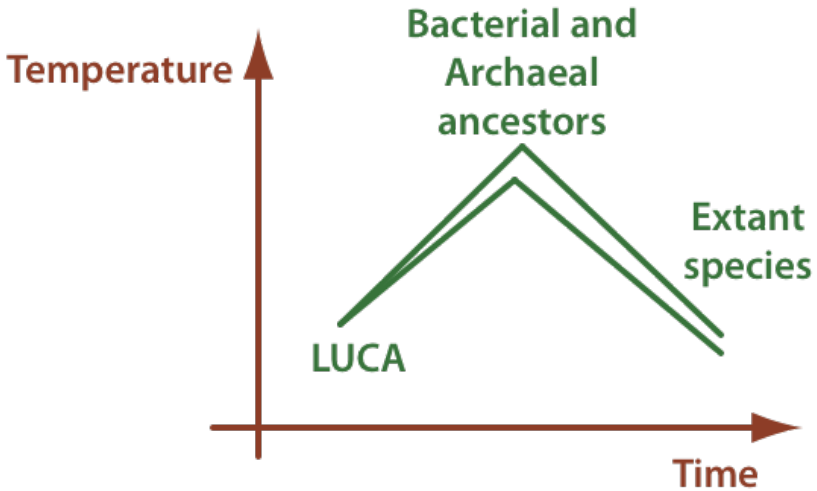


Figure 4: The compositional approach to reconstructing ancient growth temperatures. a) Correlation between the G+C content of the stem portion of ribosomal RNAs and optimal growth temperature in prokaryotes. **b)** Branch-homogeneous model of sequence evolution: all branches in the tree are associated to the same model of sequence evolution. Starting from particular base

compositions at the root of the tree, which can be estimated, such models assume that the evolutionary process has been the same on all branches of the tree. Such an assumption is especially problematic for very ancient phylogenies, as it seems unrealistic to assume that the evolutionary process has been constant over billions of years and very different types of organisms. **c)** Branch-heterogeneous models of evolution: each branch of the tree is associated to a particular substitution matrix. Such models can accommodate very different evolutionary processes through time and among species. However, they are difficult to implement and use. **d)** Branch-heterogeneous models find evidence for a two-step evolution in the history of growth temperatures along the tree of life. First, growth temperatures increased from the Last Universal Common Ancestor to the Bacterial and Archaeal ancestors. Then, growth temperatures seem to have independently decreased in the two prokaryotic kingdoms.

Other ecological parameters may also leave traces in genomes. One such parameter of interest is oxygen, as it exerts a strong pressure on organisms, some of them requiring oxygen in their environment to grow, when for others it is a poison. Recent studies suggest that there may be a very weak influence of oxygen on genome sequences (Naya et al., 2002; Vieira-Silva and Rocha, 2008). Perhaps sophisticated statistical analyses could extract the origin of this weak signal, and use it to infer whether oxygen was present in the environment of ancestral organisms whose genome sequence can be reconstructed using phylogenetic methods.

Besides oxygen concentration and temperature, the rarity of particular elements in the environment of an organism has been reported to affect the composition of the genome and of the proteins of an organism (see (Elser et al., 2011) for a review). For instance, plants have a genome and protein content in Nitrogen smaller than animals, presumably because Nitrogen is much harder to acquire, and is therefore more limiting, for primary producers than for animals (Elser et al., 2006; Acquisti et al., 2009). Consistent with this idea, domesticated plants, for which soil Nitrogen is complemented by the use of fertilizers and is therefore less limiting, use more Nitrogen in their genomes and proteins than non-domesticated plants. Unicellular organisms may also show in their genomes depletions in the elements (Nitrogen, Carbon, Oxygen, Sulfur) that are most limiting to their growth. In this case, it would thus be possible, based on reconstructed ancient sequences, to infer relative concentrations of these elements in past environments. Based on this idea, (Acquisti et al., 2006) studied the oxygen content of extra-cellular proteins in 19 prokaryotes and eukaryotes, and drew conclusions regarding the early evolution of these organisms. However (Acquisti et al., 2006) did not reconstruct ancestral sequences to draw inferences about early evolution but considered instead that sequences of extant organisms were representative of conditions that prevailed on Earth billions of years ago. This approach could certainly be misleading, as the sequences of extant organisms have had time to adapt to modern conditions, and cannot be considered as an unaltered testimony of conditions past. A better approach would have used ancestral sequence reconstruction to study ancient history.

Such compositional approaches to inferring ancient ecological variables provide an opportunity for integrating geological measurements with genomic evolution, by incorporating the correlations that are found between genomic characteristics and ecological parameters into the model of evolution. Such an

approach could use methods akin to the one used by (Lartillot and Poujol, 2010)(Lartillot and Poujol, 2010), in a study of correlations between life history traits and molecular evolution in mammals. These authors analyzed together several life history traits such as mass or longevity, and inferred potential correlations between these traits and different characteristics of the process of molecular evolution. Importantly, their model does not assume that there are correlations in their data: correlations between two variables can be 0. However, in cases where correlations are found in the data, these correlations are then used to reconstruct ancestral states. For instance, because they found a correlation between mass and the rate of sequence evolution, (Lartillot and Poujol, 2010) were able to use sequence data to reconstruct masses for ancestral mammals, and found that the sequence data changed some of the ancestral mass estimates. A similar approach could be used to reconstruct ancestral temperatures for instance, using correlations that exist between sequence composition and this variable. Importantly, information coming from geological estimates of ancient temperatures could also be included in the model. This way, sequence data would help interpolate ancient temperatures between dated geological estimates. In addition, the same geological estimates would help date the tree of life. Such models remain to be developed, but constitute a natural framework for interactions between earth scientists and biologists, where geological estimates and genomic sequences are analyzed together in a unified statistical model.

Ancestral gene resurrection

So far, we have considered genome-wide traits such as overall base or amino-acid composition, or the presence/absence of particular genes. However, the sequence of particular genes itself can also provide crucial information into the functioning of an organism.

Studies performed on populations of extant organisms often show how one single base mutation affects the capacity of an organism to thrive in a particular environment. A striking example has been recently found in humans. Tibetans live at very high altitude, and for this reason have acquired the capacity to live in an environment with much less oxygen than the rest of the species. (Yi et al., 2010)(Yi et al., 2010) found that a single mutation close to a gene involved in the response to hypoxia was much more frequent in Tibetans than in other groups of humans. This was taken as evidence that this single mutation provides an important selective advantage to individuals leaving where oxygen is scarce. Although this result was found in an extant species, it shows that single mutations can have important effects on the fitness of an individual in a particular environment. The knowledge of such mutations provides the opportunity to predict, just based on the sequence of a gene, whether an organism could thrive in an environment or not.

The characterization of such mutations usually requires wet-lab experiments comparing two versions of a gene, one with, and one without a particular mutation. In some cases, wet-lab experiments are performed jointly with ancestral sequence reconstruction. When the natural two-step approach would be as follows: first, use molecular biology experiments to identify in a gene mutations diagnostic for a

particular lifestyle, and second, reconstruct ancestral sequences and use the formerly defined mutations to predict ancestral lifestyles, the joint approach avoids pinpointing precisely the nature and position of the mutations involved. Instead, ancestral sequences are inferred, and then "resurrected" in the lab using techniques from molecular biology. These "resurrected" sequences are then characterized alongside sequences from extant organisms: the properties of ancestral sequences are therefore determined and the lifestyles of the extinct organisms that contained them can be inferred.

(Dean and Golding, 1997; Zhu et al., 2005) used this type of approach to study the evolutionary history of the isocitrate dehydrogenase gene family, a central metabolic enzyme. They found that a similar mutation occurred independently three times, putatively around the time Eukaryotes first appeared. This series of mutations appears to be diagnostic of an ecological change: the ancestral enzyme was tuned to work well where energy-rich nutrients like glucose were available, and the mutations changed it into an enzyme working well in energy-poor conditions, for instance when acetate is the main energy source. More precise dating of these shifts would allow for querying the geological record and look for global events that may explain these shifts.

(Gaucher et al., 2003) used a similar approach to study the temperature at which the bacterial ancestor lived. They resurrected the ancestor of the Elongation factor Tu gene family, a protein important for the proper functioning of the cell, and assessed its function at various temperatures. Because in extant species the temperature at which the protein functions best is very close to the temperature at which the host organism lives, one can use the optimal temperature for the ancestral protein to estimate ancestral growth temperature. Gaucher et al. found that the bacterial ancestor was probably thermophilic. In a later study (Gaucher et al., 2008), they pursued their analyses and established growth temperatures for several ancestral organisms sampled from the tree of Bacteria. Based on these results, they concluded that growth temperatures had decreased from a thermophilic bacterial ancestor to extant species, which are mostly mesophilic. They related this decreasing tendency to the precambrian temperature trend of (Robert and Chaussidon, 2006), a result now contradicted by more recent analyses (Hren et al., 2009; Blake et al., 2010). Interestingly, Gaucher et al.'s results appear to be consistent with the study by (Boussau, Blanquart, et al., 2008a), even though both the methods and the genes used in these studies are different. Overall, genomic data point towards similar inferences for the evolution of growth temperature in the bacterial kingdom.

These resurrection studies have a lot of appeal, as they give the impression that actual fossil sequences are being re-generated. Unfortunately, several hurdles have to be overcome. These ancestral sequences are now usually inferred in a probabilistic framework using computer programs. A first difficulty with reconstructing precise ancestral sequence comes from the fact that it has been shown that the most probable sequence can be a biased estimate of the ancestral sequence (Williams et al., 2006), biased towards more stable sequences. The most probable ancestral sequence may therefore not be the best estimate of the ancestral sequence. The second difficulty is linked to the very large space of probable

ancestral sequences. In practice, several positions in the sequence are often recovered with low probabilities, especially when the sequences that are reconstructed are ancient, meaning that several bases or amino-acids are about equally probable to fit in this position. In these cases, a very large number of different sequences could be nearly equally likely. To be able to infer ancestral characteristics of an ancestor, several if not all of these putative ancestral sequences should therefore be tested, which represents a prohibitive amount of work. To deal with this issue, (Gaucher et al., 2008) first computationally generated 10000 probable sequences for the bacterial ancestor EF-Tu. From these 10000 sequences, they randomly picked five different sequences for the bacterial ancestor, and found that the optimal temperatures these sequences showed were all consistent with a thermophilic ancestor. However, most of the probable sequences have not been tested, and it remains possible that some probable ancestral sequences give a very different optimal temperature. The third difficulty has to do with the models of sequence evolution used to infer ancient sequences. The models of sequence evolution that have been used to reconstruct ancestral sequences are usually simplistic, in that they notably assume that the process of sequence evolution has been the same through time and lineages. More sophisticated models are now available (Blanquart, 2006; Dutheil and Boussau, 2008), which might improve upon former studies. The fourth difficulty is a common problem faced by studies of single, isolated proteins outside of the host organism. In fact, a gene usually functions in the context of a cell, interacting with lots of different genes and metabolites. Because ancestral genes are characterized in isolation and not in the context of the ancestral cell, it is unclear whether test-tube experiments always give a faithful representation of the behavior of the ancestral protein in its ancestral cellular context.

A very recent study by Perez-Jimenez et al. (Perez-Jimenez et al., 2011) may suffer from several of these biases: the authors resurrected ancestral thioredoxin enzyme sequences at seven different nodes along the tree of life over the last 4 billion years, and characterized their activity with respect to temperature and acidity. They found that the resurrected proteins for the oldest nodes were the most stable at high temperature, and were able to perform their enzymatic function at low PH. They interpret this result as showing that the ancient oceans may have been hot and acid. However, several tests should be done to confirm these results: only the most likely sequence was used for resurrected proteins, which might explain their great stability (Williams et al., 2006), so other less likely ancestral sequences should also be tested. In addition, a model that assumes that the process of sequence evolution has been constant through time was used, even though the results obtained by the authors suggest that the environment may have changed through time. More sophisticated models of sequence evolution that can deal with differences in the process of sequence evolution through time should therefore be used (Blanquart, 2006; Dutheil and Boussau, 2008). Overall, this study by Perez-Jimenez et al. (Perez-Jimenez et al., 2011) illustrates how difficult resurrection studies are: although the authors synthesized several ancestral sequences in vitro and performed sophisticated molecular experiments, it is unclear whether those results are polluted by reconstruction artifacts or not.

To sum up, these considerations suggest that resurrection experiments can be performed only for well-conserved proteins whose function can be easily tested *in vitro*. Moreover, several technical difficulties make such approaches extremely time-consuming. However, resurrection studies also provide the opportunity to test the activity of ancestral proteins much more precisely and exhaustively than what can be achieved with current computational methods, and therefore remain extremely precious.

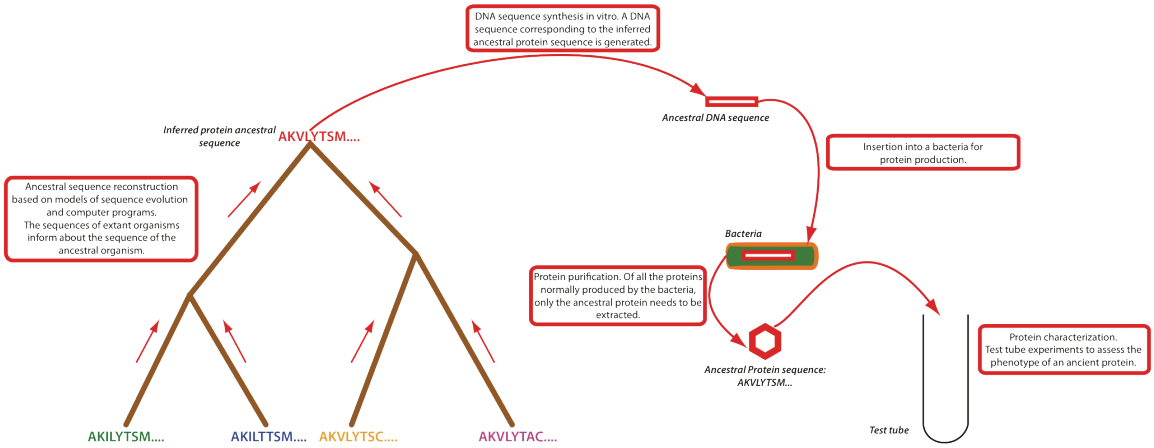


Figure 5: Protein resurrection. This figure describes the major steps involved in ancestral protein resurrection. Using computer programs and models of sequence evolution, an ancestral protein sequence is inferred. A DNA sequence corresponding to this protein sequence is synthesized in vitro. This DNA sequence is then inserted into a plasmid (not shown), and the plasmid is inserted into a bacteria for protein production. The bacteria produces the protein in large quantities. The protein is purified from the bacteria, and tests are performed to study the characteristics of the resurrected protein.

Conclusion

We have reviewed several approaches that use extant genomes to address questions related to the co-evolution between life and the Earth. Molecular phylogeny is used to build the tree of life and to date it, character reconstruction is used to infer characteristics of ancient organisms based on the characteristics of extant organisms, gene content and metabolic pathway reconstruction provide insight into ancient metabolisms and ecologies, nucleotide and amino acid content reconstruction can indicate at what temperature ancient organisms lived, and protein resurrection is a mean to assess whether an ancestral organism was able to perform a particular feat. Taken separately, each approach possesses particular weaknesses, and can only answer particular questions (Table 1). Together, they can contribute a crucial amount of information to complement geological data. One exciting avenue involves developing statistical models of genome evolution that include geological measurements. Such models will take advantage of both types of data to draw inferences about the history of life and of the Earth. They will require, perhaps more than ever, a strong collaboration between biologists and geologists.

Table 1: Merits and limits of various methods to infer the characters of ancient organisms based on extant species and their genomes.

Method of analysis	Advantages	Drawbacks	Questions
Phylogenetic reconstruction.	Solid statistical framework. Biases and best practices known.	Sequences can be saturated, which leads to biases in phylogenetic reconstruction. Dating is very difficult when few fossils are available for calibration.	What is the topology of the tree of life? What is the age of particular groups of organisms? When did radiations/extinctions occur?
Ancestral character reconstruction by parsimony.	Principle easy to understand, akin to Occam's razor.	Known to perform poorly when characters evolve quickly. Need to specify costs *a priori*.	When did this particular character appear in the tree of life? Did this ancestor have this character?
Ancestral gene content reconstruction.	Can provide deep insights into ancient organism characteristics.	No fully satisfying model that would use gene sequences in a statistical framework has been published so far.	What kind of metabolism had this ancestor? What kind of molecular machinery was present in this ancestor?
Ancestral base or amino-acid composition reconstruction.	Can provide genome-wide and thus robust estimates.	Requires sophisticated models of sequence evolution. Only a limited number of questions can be asked.	At what temperature lived this ancestral organism? What element was limiting in the environment of this ancestral organism?
Ancestral gene resurrection	Can provide very detailed analysis of a particular ancestral gene. Lots of genes in principle could be analyzed with	Requires sophisticated models of sequence evolution. Analysis has to be done one gene at a	Any question that can be asked on extant proteins can be asked on resurrected proteins.

	this approach, and lots of questions answered.	time. Very time consuming. Very large space of ancestral sequences needs to be sampled correctly.	

References

Abramov, O., Mojzsis, S.J., 2009. Microbial habitability of the Hadean Earth during the late heavy bombardment. Nature 459, 419.

Acquisti, C., Kleffe, J.U.R., Collins, S.E.A., 2006. Oxygen content of transmembrane proteins over macroevolutionary time scales. Nature 445, 47.

Acquisti, C., Elser, J., Kumar, S., 2009. Ecological Nitrogen Limitation Shapes the DNA Composition of Plant Genomes. Molecular Biology and Evolution 26, 953.

Albani, A.E., Bengtson, S., Canfield, D.E., Bekker, A., Macchiarelli, R., Mazurier, A., Hammarlund, E.U., Boulvais, P., Dupuy, J., Fontaine, C., Fürsich, F.T., Gauthier-Lafaye, F., Janvier, P., Javaux, E., Ossa, F.O., Pierson-Wickmann, A., Riboulleau, A., Sardini, P., Vachard, D., Whitehouse, M., Meunier, A., n.d. Large colonial organisms with coordinated growth in oxygenated environments 2.1 Gyr ago. Nature 466, 100.

Allwood, A., Walter, M., Kamber, B., Marshall, C., Burch, I., 2006. Stromatolite reef from the Early Archaean era of Australia. Nature 441, 714-718.

Becerra, A., Delaye, L., Lazcano, A., Orgel, L., 2007. Protein disulfide oxidoreductases and the evolution of thermophily: was the last common ancestor a heat-loving microbe? J Mol Evol 65, 296-303.

Bijl, P.K., Houben, A.J.P., Schouten, S., Bohaty, S.M., Sluijs, A., Reichart, G., Sinninghe Damste, J.S., Brinkhuis, H., 2010. Transient Middle Eocene Atmospheric CO2 and Temperature Variations. Science 330, 819-821.

Blake, R.E., Chang, S.J., Lepland, A., 2010. Phosphate oxygen isotopic evidence for a temperate and biologically active Archaean ocean. Nature 464, 1029.

Blanquart, S., 2006. A Bayesian Compound Stochastic Process for Modeling Nonstationary and Nonhomogeneous Sequence Evolution. Molecular Biology and Evolution 23, 2058-2071.

Borenstein, E., Kupiec, M., Feldman, M., Ruppin, E., 2008. Large-scale reconstruction and phylogenetic analysis of metabolic environments. Proceedings of the National Academy of Sciences 105, 14482.

Boussau, B., Karlberg, E., Frank, A., Legault, B., Andersson, S., 2004a. Computational inference of scenarios for α-proteobacterial genome evolution. Proc Natl Acad Sci USA 101, 9722.

Boussau, B., Karlberg, E.O., Frank, A.C., Legault, B., Andersson, S.G.E., 2004b. Computational inference of scenarios for alpha-proteobacterial genome evolution. Proc Natl Acad Sci USA 101, 9722-9727.
Boussau, B., Guéguen, L., Gouy, M., 2008a. Accounting for horizontal gene transfers explains conflicting hypotheses regarding the position of aquificales in the phylogeny of Bacteria. BMC Evol Biol 8, 272.
Boussau, B., Blanquart, S., Necsulea, A., Lartillot, N., Gouy, M., 2008b. Parallel adaptations to high temperatures in the Archaean eon. Nature 456, 942-945.
Brochier-Armanet, C., Boussau, B., Gribaldo, S., Forterre, P., 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Micro 6, 245-252.
Brochier-Armanet, C., Talla, E., Gribaldo, S., 2009. The multiple evolutionary histories of dioxygen reductases: implications for the origin and evolution of aerobic respiration. Molecular Biology and Evolution 26, 285.
Brocks, J.J., Love, G.D., Summons, R.E., Knoll, A.H., Logan, G.A., Bowden, S.A., 2005. Biomarker evidence for green and purple sulphur bacteria in a stratified Palaeoproterozoic sea. Nature 437, 866.
Castresana, J., Lübben, M., Saraste, M., Higgins, D.G., 1994. Evolution of cytochrome oxidase, an enzyme older than atmospheric oxygen. The EMBO Journal 13, 2516.
Cohen, O., Pupko, T., 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. Molecular Biology and Evolution 27, 703.
Csűrös, M., Miklós, I., 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. Research in Computational Molecular Biology 206-220.
David, L.A., Alm, E.J., 2010. Rapid evolutionary innovation during an Archaean genetic expansion. Nature 469, 93-96.
Dean, A., Golding, G., 1997. Protein engineering reveals ancient adaptive replacements in isocitrate dehydrogenase. Proc Natl Acad Sci USA 94, 3104.
Dutheil, J., Boussau, B., 2008. Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs. BMC Evol Biol 8, 255.
Eigenbrode, J., Freeman, K., Summons, R., 2008. Methylhopane biomarker hydrocarbons in Hamersley Province sediments provide evidence for Neoarchean aerobiosis. Earth and Planetary Science Letters 273, 323-331.
Elkins, J.G., Podar, M., Graham, D.E., Makarova, K.S., Wolf, Y., Randau, L., Hedlund, B.P., Brochier-Armanet, C., Kunin, V., Anderson, I., Lapidus, A., Goltsman, E., Barry, K., Koonin, E.V., Hugenholtz, P., Kyrpides, N., Wanner, G., Richardson, P., Keller, M., Stetter, K.O., 2008. A korarchaeal genome reveals insights into the evolution of the Archaea. Proc Natl Acad Sci USA 105, 8102-8107.
Elser, J.J., Fagan, W.F., Subramanian, S., Kumar, S., 2006. Signatures of Ecological Resource Availability in the Animal and Plant Proteomes. Molecular Biology and Evolution 23, 1946-1951.
Elser, J.J., Acquisti, C., Kumar, S., 2011. Stoichiogenomics: the evolutionary ecology of macromolecular elemental composition. Trends in Ecology & Evolution 26, 38-44.
Farquhar, J., Zerkle, A., Bekker, A., n.d. Geological constraints on the origin of

oxygenic photosynthesis. Photosynthesis Research 1-26.
Feist, A.M., Herrg Aring Rd, M.J., Thiele, I., Reed, J.L., Palsson, B.O., 2008. Reconstruction of biochemical networks in microorganisms. Nat Rev Micro 7, 129.
Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B., Merrick, J., 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496.
Fournier, G.P., Gogarten, J.P., 2010. Rooting the Ribosomal Tree of Life. Molecular Biology and Evolution 27, 1792-1801.
Galtier, N., 1999. A Nonhyperthermophilic Common Ancestor to Extant Life Forms. Science 283, 220-221.
Galtier, N., Lobry, J.R., 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol 44, 632-636.
Gaucher, E.A., Thomson, J.M., Burgan, M.F., Benner, S.A., 2003. Inferring the palaeoenvironment of ancient bacteria on the basis of resurrected proteins. Nature 425, 285-288.
Gaucher, E.A., Govindarajan, S., Ganesh, O.K., 2008. Palaeotemperature trend for Precambrian life inferred from resurrected proteins. Nature 451, 704-707.
Gribaldo, S., Talla, E., Brochier-Armanet, C., 2009. Evolution of the haem copper oxidases superfamily: a rooting tale. Trends in Biochemical Sciences.
Gribaldo, S., Brochier-Armanet, C., 2006. The origin and evolution of Archaea: a state of the art. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 1007.
Hahn, M.W., de Bie, T., Stajich, J.E., Nguyen, C., Cristianini, N., 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res 15, 1153-1160.
Hoashi, M., Bevacqua, D.C., Otake, T., Watanabe, Y., Hickman, A.H., Utsunomiya, S., Ohmoto, H., 2009. Primary haematite formation in an oxygenated sea 3.46|[thinsp]|billion years ago. Nature Geoscience 2, 301.
Holland, H., 2006. The oxygenation of the atmosphere and oceans. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 903.
Hren, M.T., Tice, M.M., Chamberlain, C.P., 2009. Oxygen and hydrogen isotope evidence for a temperate climate 3.42 billion years ago. Nature 462, 205.
Javaux, E.J., Marshall, C.P., Bekker, A., 2010. Organic-walled microfossils in 3.2-billion-year-old shallow-marine siliciclastic deposits. Nature 463, 934.
Kasting, J., Howard, M., 2006. Atmospheric composition and climate on the early Earth. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 1733.
Kasting, J., Ono, S., 2006. Palaeoclimates: the first two billion years. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 917.
Knauth, L., Lowe, D., 2003. High Archean climatic temperature inferred from oxygen isotope geochemistry of cherts in the 3.5 Ga Swaziland Supergroup, South Africa. Geological Society of America Bulletin 115, 566.
Knoll, A.H., Javaux, E.J., Hewitt, D., Cohen, P., 2006. Eukaryotic organisms in Proterozoic oceans. Philosophical Transactions of the Royal Society B: Biological

Sciences 361, 1023.
Lane, N., Allen, J.F., Martin, W., 2010. How did LUCA make a living? Chemiosmosis in the origin of life. Bioessays 32, 271-280.
Lartillot, N., Poujol, R., 2010. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Molecular Biology and Evolution.
Lyons, T.W., Lyons, T.W., Reinhard, C.T., Reinhard, C.T., 2009. Early Earth: Oxygen for heavy-metal fans. Nature 461, 179.
Maruyama, S., Santosh, M., 2008. Models on Snowball Earth and Cambrian explosion: A synopsis. Gondwana research.
Mirkin, B., Fenner, T., Galperin, M., Koonin, E., 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. BMC Evol Biol 3, 2.
Naya, H., Romero, H., Zavala, A., Alvarez, B., Musto, H., 2002. Aerobiosis increases the genomic guanine plus cytosine content (GC%) in prokaryotes. J Mol Evol 55, 260-264.
Perez-Jimenez, R., Inglés-Prieto, A., Zhao, Z., Sanchez-Romero, I., Alegre-Cebollada, J., Kosuri, P., Garcia-Manyes, S., Kappock, T.J., Tanokura, M., Holmgren, A., Sanchez-Ruiz, J.M., Gaucher, E.A., Fernandez, J.M., 2011. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nature Structural & Molecular Biology .
Philippot, P., van Zuilen, M., Lepot, K., Thomazo, C., Farquhar, J., van Kranendonk, M.J., 2007. Early Archaean Microorganisms Preferred Elemental Sulfur, Not Sulfate. Science 317, 1534-1537.
Rasmussen, B., 2000. Filamentous microfossils in a 3,235-million-year-old volcanogenic massive
sulphide deposit. Nature 405, 676.
Rasmussen, B., Fletcher, I.R., Brocks, J.J., Kilburn, M.R., 2008. Reassessing the first appearance of eukaryotes and cyanobacteria. Nature 455, 1101-1104.
Robert, F., Chaussidon, M., 2006. A palaeotemperature curve for the Precambrian oceans based on silicon isotopes in cherts. Nature 443, 969-972.
Rodríguez-Ezpeleta, N., Brinkmann, H., Burger, G., Roger, A.J., Gray, M.W., Philippe, H., Lang, B.F., 2007. Toward resolving the eukaryotic tree: the phylogenetic positions of jakobids and cercozoans. Current Biology 17, 1420-1425.
Royer, D., Berner, R., Montañez, I., Tabor, N., Beerling, D., 2004. CO 2 as a primary driver of Phanerozoic climate. GSA Today 14, 4-10.
Sagan, C., Mullen, G., 1972. Earth and Mars: Evolution of Atmospheres and Surface Temperatures. Science 177, 52-56.
Schopf, J.W., 2006. Fossil evidence of Archaean life. Philos Trans R Soc Lond, B, Biol Sci 361, 869-885.
Schopf, J.W., n.d. The paleobiological record of photosynthesis. Photosynthesis Research 1-15.
Snel, B., Bork, P., Huynen, M., 2002. Genomes in flux: the evolution of archaeal and proteobacterial gene content. Genome Res 12, 17.
Summons, R.E., Bradley, A.S., Jahnke, L.L., Waldbauer, J.R., 2006. Steroids,

triterpenoids and molecular oxygen. Philosophical Transactions of the Royal Society B: Biological Sciences 361, 951-968.
Tian, F., 2005. A Hydrogen-Rich Early Earth Atmosphere. Science 308, 1014-1017.
Ueno, Y., Yamada, K., Yoshida, N., Maruyama, S., Isozaki, Y., 2006. Evidence from fluid inclusions for microbial methanogenesis in the early Archaean era. Nature 440, 516-519.
Ueno, Y., Johnson, M.S., Danielache, S.O., Eskebjerg, C., Pandey, A., Yoshida, N., 2009. Geological sulfur isotopes indicate elevated OCS in the Archean atmosphere, solving faint young sun paradox. Proceedings of the National Academy of Sciences 106, 14784-14789.
Veizer, J.A.N., Godderis, Y., Fran Ccedil Ois, L.M., 2000. Evidence for decoupling of atmospheric CO2 and global climate during the Phanerozoic eon. Nature 408, 698.
Ventura, G., Kenig, F., Reddy, C., Schieber, J., Frysinger, G., Nelson, R., Dinel, E., Gaines, R., Schaeffer, P., 2007. Molecular evidence of Late Archean archaea and the presence of a subsurface hydrothermal biosphere. Proceedings of the National Academy of Sciences 104, 14260.
Vieira-Silva, S., Rocha, E., 2008. An assessment of the impacts of molecular oxygen on the evolution of proteomes. Molecular Biology and Evolution 25, 1931.
Waldbauer, J., Sherman, L., Sumner, D., Summons, R., 2009. Late Archean molecular fossils from the Transvaal Supergroup record the antiquity of microbial diversity and aerobiosis. Precambrian Research 169, 28-47.
Williams, P., Pollock, D., Blackburne, B., Goldstein, R., 2006. Assessing the accuracy of ancestral protein reconstruction methods. PLoS Comp Biol 2, e69.
Wolf, E.T., Toon, O.B., 2010. Fractal Organic Hazes Provided an Ultraviolet Shield for Early Earth. Science 328, 1266-1268.
Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., Zheng, H., Liu, T., He, W., Li, K., Luo, R., Nie, X., Wu, H., Zhao, M., Cao, H., Zou, J., Shan, Y., Li, S., Yang, Q., Asan, Ni, P., Tian, G., Xu, J., Liu, X., Jiang, T., Wu, R., Zhou, G., Tang, M., Qin, J., Wang, T., Feng, S., Li, G., Huasang, Luosang, J., Wang, W., Chen, F., Wang, Y., Zheng, X., Li, Z., Bianba, Z., Yang, G., Wang, X., Tang, S., Gao, G., Chen, Y., Luo, Z., Gusang, L., Cao, Z., Zhang, Q., Ouyang, W., Ren, X., Liang, H., Zheng, H., Huang, Y., Li, J., Bolund, L., Kristiansen, K., Li, Y., Zhang, Y., Zhang, X., Li, R., Li, S., Yang, H., Nielsen, R., Wang, J., Wang, J., 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. Science 329, 75-78.
Zachos, J., 2001. Trends, Rhythms, and Aberrations in Global Climate 65 Ma to Present. Science 292, 686-693.
Zahnle, K., Arndt, N., Cockell, C., Halliday, A., Nisbet, E., Selsis, F., Sleep, N.H., 2007. Emergence of a Habitable Planet. Space Sci Rev 129, 35-78.
Zeldovich, K., Berezovsky, I., Shakhnovich, E., 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comp Biol 3, e5.
Zhu, G., Golding, G., Dean, A., 2005. The selective cause of an ancient adaptation. Science 307, 1279.