Membrane bioenergetics at major transitions in evolution

Víctor SOJO MARTÍNEZ

Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX)

and

Department of Genetics, Evolution and Environment (GEE)

University College London

A Thesis submitted in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

London, January 2016

I, Víctor SOJO MARTÍNEZ confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

29 / Jan / 2016 Víctor SOJO MARTÍNEZ

date signed

ABSTRACT

This thesis presents theoretical and computational studies of four major evolutionary transitions in which cellular membranes and their embedded proteins played crucial roles:

1)The divergence of archaea and bacteria

Archaea and bacteria are the basal domains of life, so it is important to understand how they came to diverge. They share several core traits, such as transcription, translation, and the genetic code. They also share the chemiosmotic exploitation of ion gradients across membranes, yet they do not share the membranes themselves. Notably, the phospholipid backbone is *sn*-glycerol-1-phosphate in archaea but the enantiomer *sn*-glycerol-3-phosphate in bacteria. The synthesising enzymes are unrelated. I used mathematical modelling to propose an explanation for this divergence in the context of natural proton gradients in alkaline hydrothermal vents, plausible scenarios for an autotrophic origin of life. Results show that early membranes had to be leaky, so both pumping and glycerol-phosphate backbones (which drastically decrease permeability) evolved later, and independently, in archaea and bacteria.

2)The evolution of Homochirality

The “dual homochirality” of lipids suggests that the stereospecificity of bioorganic catalysis itself, not prebiotic physics or chemistry, is behind the origin of handedness in life’s molecules (e.g. L-amino acids and D-sugars).

3)The evolution of membrane proteins

I report that membrane proteins are less shared across the tree of life. Faster evolution of outside-facing regions and true gene losses point to a common cause: as cells adapt to new environments selective pressure is stronger on the outside, while the inside, subject to strong homeostasis, evolves more slowly.

4)The bacterial nature of eukaryotic membranes

Eukaryotes arose from a merger of a bacterium into an archaeon, so the first eukaryote must have had an archaeal plasma membrane and bacterial (proto)mitochondrial membranes; yet all modern eukaryotes have exclusively bacterial membranes. I suggest that archaeal membranes were lost and bacterial ones kept because of the bioenergetic adaptation of mitochondrial proteins to the bacterial membrane.

Every one knows how greedily a theorist pounces on a fact,
highly favourable to his views (...)

Charles Darwin (1845)

ACKNOWLEDGEMENTS

I am forever grateful to Abigail Espie and Guy Moss for their support in securing my continuation at UCL, along with Chris Langridge, Gemma Ludbrook, and Laura Heekin at the CoMPLEX office, and Fiona Williamson at GEE.

In alphabetical order, these people made London my home: Karina Álvarez, Kostas Angelis, Will Ashworth, Bobby Bentham, Sami Bouremoum, Anna Czarkwiani, Daniel Dalquen, Jon Emberey, Johannes Girstmair, Julia Hill, Tim Lucas, Steven Müller, Dan Manson, Claire Walsh, Alex Whicher, Tom Wyatt, Anne Zakrzewski, and Gabriel Zamora. Thanks specifically again to Anne and Julia for their help printing, binding and submitting this document.

I benefitted greatly from a collaboration with Jan Koch, and from constant discussions with Zena Hadjivasiliou, Bram Kuijper, Jez Owen, Eloi Camprubí, and Barry Herschy. I was fortunate to work with Dan Jeffares as my supervisor during my Master's year, and with Christophe Dessimoz during the latter part of my PhD.

I would not have made it nearly as far in chasing my own dream of becoming a scientist without the constant support of my mother and father, my brother, my aunts and uncles, my departed grandparents, and my cousins; I will not thank them here, their love is beyond thanks and I can only try to repay in the same currency.

Luisana has been a pillar throughout and I am lucky she chose to bless my life by sharing hers with me. I look forward to a life full of adventures, some academic, some not.

I cannot be more grateful to my supervisors, Andrew Pomiankowski and Nick Lane, for their generosity, academic and otherwise. They constantly encouraged and inspired me to think for myself, while providing active guidance throughout. This is invaluable for someone who wishes to follow in their footsteps and become a world-class scientist, and I will remember it in the hope that I can do the same for my own students some day, should I succeed.

Finally, I would like to, but unfortunately cannot, thank my cousin and brother Edgar, for inspiring me to think better and more, about anything and everything. Every day with him was a lesson and a joy, and I sorely miss him.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND ABBREVIATIONS

amphiphile: a molecule with two regions, one of which has an affinity to water (is **hydrophilic**), while the other is apolar and typically rich in C-C and C-H bonds, with low affinity for water (**hydrophobic**) but relatively high affinity to similar regions within the same or in other molecules. Many amphiphiles, including fatty acids and phospholipids, spontaneously self-assemble into micelles (spheres with apolar ends pointing inwards and polar ends facing the solvent) or vesicles (bilayers that wrap an aqueous core inside, such as in a cell).

archaea (singular: '**archaeon**'; adjective '**archaeal**'**):** one of the three domains of life, and members within, referred to either as individual organisms, species, or higher taxa. Also called 'archaebacteria' in the literature for historical reasons. See **bacteria**.

archaeal: pertaining to the **archaea**. Not to be confused with '**Archean**', a geological eon.

archaeon: singular of '**archaea**', i.e. one organism or species thereof. Not to be confused with '**Archean**', a geological eon.

Archean: the second of the four major geological eons of Earth's history, between 4 **Ga** and 2.5 **Ga**, following the **Hadean** and preceding the **Proterozoic**, together forming the **Pre-Cambrian** super-eon. Not to be confused with '**archaea**', '**archaeal**' or '**archaeon**'. Also spelled 'Archaean', but this spelling is avoided here as it can lead to confusion due to the use of the exact same term in some of the literature to refer to '**archaeal**' (pertaining to the **archaea**).

ATPase: adenosine triphosphate (ATP) synthase. Throughout this document the ATPase is assumed to have been operating as an ATP-synthesising enzyme unless otherwise noted.

bacteria (singular: '**bacterium**; adjective '**bacterial**)**:** one of the three domains of life, and one of the two prokaryotic domains, together with the **archaea**. In this document, 'bacteria' is never used to refer to the archaea, although such indiscriminate usage is unfortunately still common in the literature. For historical reasons bacteria are also called 'eubacteria' in the literature, to distinguish them from the 'archaebacteria' (i.e. archaea). In this document only 'archaea' and 'bacteria' are used.

BLAST: Basic Local Alignment Search Tool, a bioinformatics algorithm used to detect homologous sequences in a database (the 'targets'), by their similarity to a reference sequence (the 'query'). Variants used in this thesis include: `blastp`, for detecting homologous protein sequences using a protein sequence as query, and `tblastn`, for searching a protein sequence query against all 6 possible reading frames in a database of nucleotide sequences.

domain: used in this document to refer to any of the three main sub-types of life on Earth: the archaea, bacteria, and eukaryotes. The term "domain" is also used independently in structural biology to describe functional sub-portions of proteins, but this usage is avoided here in favour of the phylogenetic one, to avoid confusion.

Ech: the energy-converting hydrogenase, an enzyme involved in carbon and energy metabolism in several microoganisms. In this document, Ech is suggested to have been in **LUCA**, where it operated as a gradient-exploiting bioenergetic enzyme, reducing ferredoxin for carbon fixation in a natural ion gradient. Methanogens (considered here to be the ancestors of all **archaea**) would keep this ancestral function and evolved a separate ion pump. Conversely, acetogens (which would be the ancestors of all **bacteria**) reversed flux, converting Ech into a membrane pump, in turn being forced to evolve a new pathway for the reduction of ferredoxin and for carbon fixation.

endosymbiotic gene transfer (EGT): a transfer of genetic material from an "endosymbiont" genome to a "host" genome, such as from mitochondria and plastids to the nucleus in the evolution of eukaryotes.

eukaryotes: The group that contains all known complex cells and organisms. Chronologically the third domain of life, assumed here to have originated from an endosymbiotic association of a bacterium into an archaeon, the first of which would gradually evolve into the mitochondrion. See **FECA** and **LECA** for the first and last eukaryotic common ancestors, respectively.

FECA: The first eukaryotic common ancestor. Throughout this document generally considered to have arisen directly after the endosymbiotic association between an archaeal host and a bacterial endosymbiont that would become the mitochondrion. It therefore had an archaeal plasma membrane and bacterial proto-mitochondrial membranes. See **LECA**.

G1P: *sn*-glycerol-1-phosphate, the moiety at the backbone of most archaeal phospholipids, in contrast to the enantiomer **G3P**, prevalent in bacteria (and eukaryotes by inheritance).

G1PDH: *sn*-glycerol-1-phosphate dehydrogenase. The enzyme that in archaea catalyses the stereospecific synthesis of **G1P** from dihydroxyacetone phosphate and NAD(P)H. Unrelated to bacterial **G3PDH**, although both use the same pair of substrates.

G3P: *sn*-glycerol-3-phosphate, the moiety at the backbone of bacterial (and by inheritance eukaryotic) phospholipids, in contrast to the enantiomer **G1P**, present in archaea.

G3PDH: *sn*-glycerol-3-phosphate dehydrogenase. The enzyme that in bacteria catalyses the stereospecific synthesis of **G3P** from dihydroxyacetone phosphate and NAD(P)H. Unrelated to archaeal **G1PDH**, although both use the same pair of substrates.

Ga: Giga-annum; one thousand million (one billion, 10^9) years; in this document, always in reference to the past (i.e. numbers are implicitly negative). Also in the literature as Gy (Giga-years), or Gya (Giga-years ago).

Hadean: First of the geological eons of the Earth, leading from the formation of the planet approximately 4.5 **Ga** to the beginning of the **Archean**' eon ~4 **Ga**. Life is thought to have arisen at some point in or shortly following the late Hadean, but the eon is characterised by an equivocal or inexistent rock record.

Heterotypic: of different kinds. In this document used to refer to lipids of the archaea and bacteria, in particular with regards to the origin of eukaryotes: a membrane composed purely of archaeal lipids would be **homotypic**, while a hybrid membrane composed of archaeal and bacterial lipids would be heterotypic.

homologous: see **homologue**.

homologue: two genes (or proteins, structures, etc.) are homologues if they share common ancestry, i.e. if the two sequences are descended from one same ancient sequence. If this relation is by a gene-duplication event, the genes are termed **paralogues**, and they can either retain the same function or perform different functions within the same species. If, on the other hand, the relation between the two genes is by a speciation event, the genes are said to be **orthologues**, and often (but not necessarily) they perform the same function in two related species. The term **xenologue** is some times (but much less frequently) used to refer to genes that were acquired by a horizontal gene transfer, and thus have no vertical ancestry in the ancestors of the recipient species, but in a foreign species. This type of ancestry is remarkably common across the tree of life and often confounds phylogenetic analyses. There are other types of homologues, such as epaktologues, which share ancestry only of fragments of the full sequence, such that two or more different and independently derived portions are assembled into a gene product of complex ancestry. This type of homology can arise through a combination of acquisitions of any the other three types, followed by independent remodelling of the whole composite sequences, which share only partial ancestry.

homotypic: of the same kind. See **heterotypic**.

hydrophilic: a molecule or part of a molecule with high affinity to water, e.g. the polar headgroups at one end of a phospholipid.

hydrophobic: a molecule or part of a molecule with low affinity to water, e.g. the apolar tails at one end of a phospholipid or fatty acid.

LECA: The last eukaryotic common ancestor. This cellular form gave rise to all extant eukaryotes, and is thought to have been fully eukaryotic in possessing traits such as a nucleus, actin/tubulin cytoskeleton with associated kinesin and dynein motors, flagellum, mitosis and meiosis, sexual reproduction, mitochondria, Golgi apparatus, endoplasmic reticulum, endocytosis and phagocytosis, bacterial-type membrane lipids, (largely) bacterial-type metabolism, and (largely) archaeal-type information processing. See **FECA**.

LUCA: the last universal common ancestor of all life on Earth. The term has different meanings to different researchers. Some assume it to necessarily have been a cellular organism, and indeed some claim the 'C' in the acronym to mean 'cellular', whereas others see it as an intermixing population either without clear-cut boundaries (cellular membranes), or with individual boundaries but genetic and reproductive commonness and fluidity, such that modern cells do not descend from any particular one of them, but from the group as a whole. In this work, LUCA is deduced to have been cellular, but with membranes leaky to protons and other small ions, unlike the modern impermeable phospholipid membranes of extant archaea, bacteria and eukaryotes.

Ma: Mega-annum; one million (10^6) years; in this document, always in reference to the past (i.e. numbers are implicitly negative). Also in the literature as My (Mega-years), or Mya (Mega-years ago).

MBGD: the Microbial Genome Database. A database of homologous sequences with a very wide sampling of species of all three domains (mostly microbial), hosted at mbgd.genome.ad.jp.

mito-nuclear transfer: the transfer of genes from the mitochondrial to the nuclear genomes in the evolution of eukaryotes. Modern mitochondrial genomes have few genes (37 in the case of humans), but the nuclear genomes retain many of the ancestrally bacterial genes, which are deduced to have been imported from the proto-mitochondrion after the endosymbiotic origin of eukaryotes. More generally, mito-nuclear transfers are referred to as **endosymbiotic gene transfers**, as are transfers from the plastids in plants and algae.

monophyletic, paraphyletic, polyphyletic: members of a group are **monophyletic** if they all cluster together in a tree, i.e. if they all have a single common ancestor, and the tree (or tree section) includes only members of the group. If a group is defined independently (e.g. "reptiles") and there is at least one member in a tree that is not also a member of the group as defined (e.g. "birds"), the distribution of the containing group is said to be **paraphyletic** (the internal group may itself be monophyletic, such as in birds, which all cluster together within the reptiles, but are not considered to be reptiles themselves). If the members of the group have different ancestors in the tree, their distribution is **polyphyletic**. See section 1.3.2 for further details.

MP: Membrane protein, as opposed to **WS**, a water-soluble protein.

OMA: from ‘Orthologous MAtrix'. A database of **orthologous** genes used in this work (chiefly in Chapter 5) and created in the group of Christophe Dessimoz, collaborator in Chapters 4 and 5 of this work.

orthologue (U.S. spelling: ‘**ortholog**’)**:** see **homologue**.

paralogue (U.S. spelling: ‘**paralog**’)**:** see **homologue**.

paraphyletic: See **monophyletic**.

polyphyletic: See **monophyletic**.

Precambrian (alternative spelling: ‘**Pre-Cambrian**’)**:** the super-eon composing the first three of the four eons of Earth’s geological history, in successive order the **Hadean**, **Archean** and **Proterozoic**. At the end of the latter, approximately 542 **Ma**, complex life including animals and plants became abundant, a phenomenon termed the “Cambrian explosion” (giving the super-eon its name).

SPAP: sodium-proton antiporter. Also be abbreviated in the literature as SPA, NHE (for Na/H$^+$ exchanger), or NHA (for Na/H$^+$ antiporter).

TACK: a super-phylum of archaea that contains the Thaumarchaeota, Aigarchaeota, Crenarchaeota, and Korarchaeota, where the acronym is derived from, in addition to other recently discovered phyla, such as Lokiarchaeota.

TMH: trans-membrane helix, or a trans-membrane helical protein. Typically a protein with α-helices that span a biological membrane. In Chapters 4 and 5 of this document, TMH proteins are used as a proxy for membrane proteins.

TMHMM: trans-membrane helix Markov model. An algorithm that detects trans-membrane α-helical structures in proteins. It was used in this thesis (Chapters 4 and 5) as a proxy to infer membrane proteins. It is accessible online, although in this thesis a locally run version was used on a computation cluster due to the vast amounts of data involved.

TOPCONS: a website that includes results from multiple software packages for the detection of trans-membrane α-helical structures in proteins (not including TMHMM).

tree of life: an attempt to depict the relationships between all living species on Earth, in its simplest form intrinsically assuming one single common origin of life and binary speciation processes. The widespread horizontal and endosymbiotic gene transfers between all domains of life cause the relationships between the clades to not be tree-like. For simplicity, however, the phrase ‘tree of life’ is used generically in this document to refer to the relationships between all clades and chiefly between the three domains (**archaea**, **bacteria** and **eukaryotes**), regardless of whether or not they are strictly tree-like.

WS: water-soluble protein, as opposed to **MP**, a membrane protein.

xenologue (U.S. spelling: ‘**xenolog**’)**:** see **homologue**.

1 INTRODUCTION

This thesis discusses the role of cellular and organellar membranes in defining the shape of the tree of life. In this introduction, I first discuss the unsuccessful quest for a satisfying definition of life. Although lacking a definition as such, I observe that life is typically characterised by membranes and the disequilibria across them, which is the overarching topic of this thesis. I then go on to discuss the single origin of life on Earth and the three domains that arose from it, namely the archaea, bacteria, and eukaryotes (or eukaryota). I discuss some common and differing traits among the three domains, and how they relate to each other in the tree of life. Next I take a brief look at some phylogenetics definitions relevant to this work, and then discuss membranes and membrane bioenergetics, the main theme of this work. I close this introduction with a brief overview of each subsequent chapter. There are more specific introductions at the beginnings of each of the chapters of this thesis.

1.1 Life

The quest for a satisfying definition of life is perhaps one of the least successful endeavours in the history of theoretical biology. Although multiple attempts exist, none enjoys universal acceptance. In terms of everyday language, the Oxford English Dictionary defines life as:

> *The condition that distinguishes animals and plants from inorganic matter, including the capacity for growth, reproduction, functional activity, and continual change preceding death.*
>
> Oxford Dictionaries (2015a)

Although nothing in the definition is intrinsically wrong, it is more a description of what life *is not* than of what it actually *is*. Inorganic matter is of course not alive, while plants and animals almost tautologically are; similarly, only things that have been alive can die, so these elements of the definition take the reader no closer to enlightenment. The other elements are more useful: the capacity for growth, reproduction, functional activity, and continual change, are indeed good descriptors of life, although none is unequivocal. Not even the combination is: fire, for example, would fit the definition rather well in that it grows, reproduces, can be said to have many a function, and changes continually until it is put out (or "dies"). Yet, intuitively, fire is not alive.

Definitions from within the field are hardly any better. In the 1990s scientists at NASA famously defined life as "a self-sustained chemical system capable of Darwinian evolution" (Joyce 1994). Although useful for astrobiologists seeking life elsewhere in the Universe, this "working definition" eschews the central aspect of energy flows: life is not self-sustained (Schrödinger 1944; Lane 2010; Lane 2015); the thermodynamic working of the universe implies that it cannot be. Heterotrophs depend on autotrophs, and autotrophs depend on external sources of energy, be it the sun, simple inorganic chemicals, or rocks. Life thus depends on sources of energy that are always foreign to the living being itself (Skulachev 1988), and everything that happens in everything that lives involves the flow of energy down a gradient.

These energy fluxes come closer to, if not a definition, at least a significant descriptive aspect of life. Sustained disequilibrium across boundaries, or homeostasis, is a remarkable trait common to all life, be it parasitic, autotrophic, or heterotrophic, and it may indeed have arisen on the early Earth as part of the geochemical processes that led to the origin of life (Allen 2010).

1.1.1 A timeline for the geological evolution of Earth and life on it

The Earth has existed for approximately 4.54 Ga (thousand million, or U.S. billion, years) (Dalrymple 2001). The early past of the planet has been divided by geologists into three "eons": Hadean, Archean, and Proterozoic, which together span the first 4 billion years of the planet's history and form the Precambrian super-eon. This was followed by the current Phanerozoic eon, which started some 541 million years ago. Life is thought to have arisen early on, at some point between the Hadean and Archaean, approximately 4 Ga in the past. Figure 1 summarises a number of key events in the evolution of the Earth and life on it.

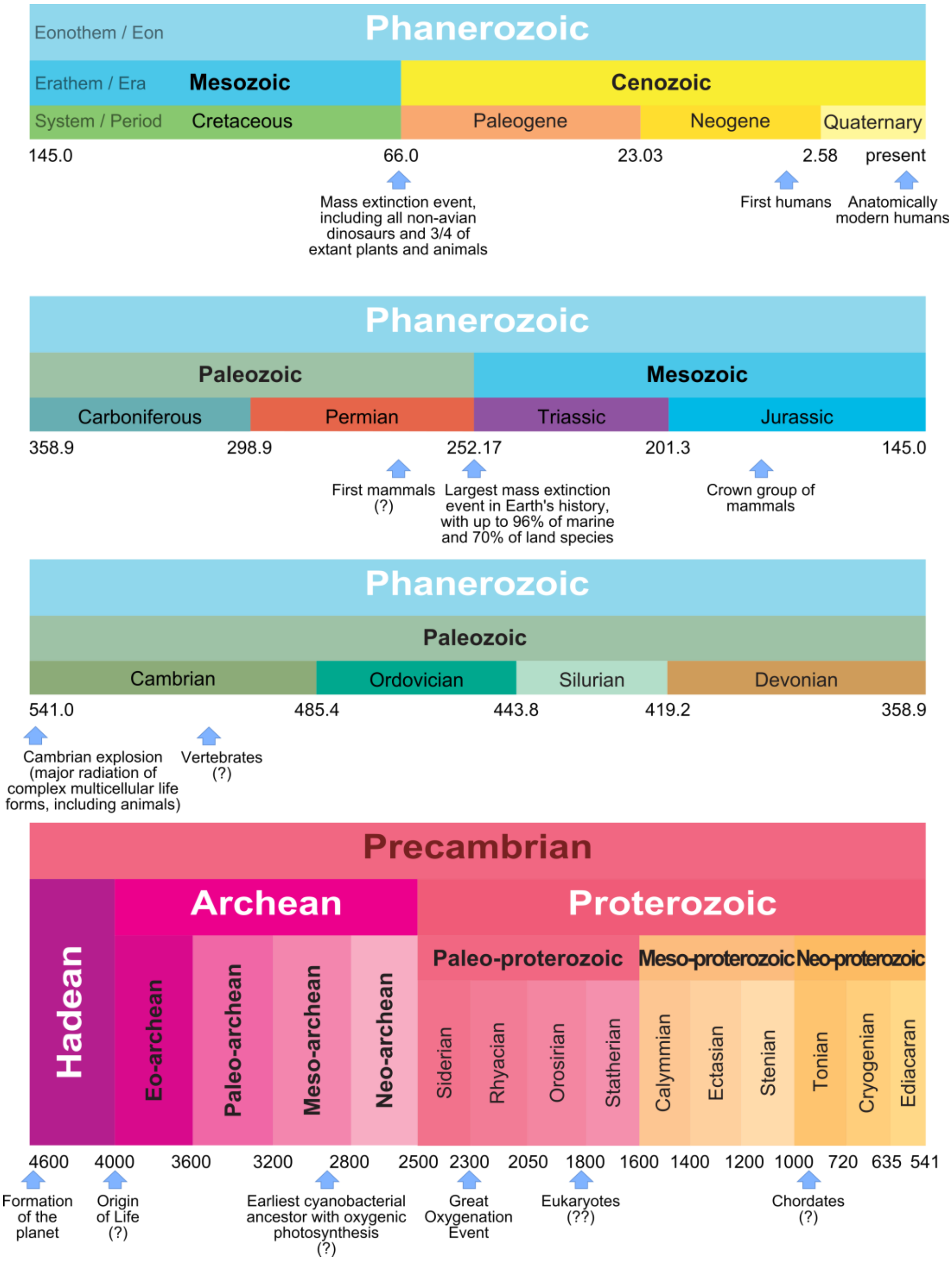


Figure 1. A brief chronology of earth's bio-geological history

The inference of geological dates depends on processes such as fossilisation (which involves more than one element of luck) and mineral deposition (which is often difficult to interpret). Many of the dates for events denoted with arrows are highly contentious. Times are in millions of years (Ma), with a non-linear scale. Adapted from the International Commission on Stratigraphy (Cohen et al. 2013).

Notably, the date of origin of eukaryotes is so contentious that it is challenging to assign a date in the chart above; recent estimates for the last eukaryotic common ancestor (LECA) range extremely widely between 943 and 2,094 Ma (Eme et al. 2014), while the first eukaryote (FECA) may have been as early as 2.2 Ga - 2.7 Ga (Brocks 1999). Another major unknown date is the origin of life, which occurred at some point between the late Hadean and early Archean, a period from which the rock record is either equivocal or inexistent.

1.1.2 The Origin of Life (OoL)

Historically, theories for the origin of life have focused mostly on the chemical problem of bio-organic synthesis from inorganic material, namely the origins of monomers of proteins and nucleotides from simple chemical precursors (Miller 1953). Although crucial, this is only one aspect of the problem, and one that can render misleading conclusions altogether by fostering the consideration of environments that could have provided the chemistry to synthesise the molecules, but not the energy to actually bring them to life.

Life is about disequilibrium (Harold 1986), and all life on Earth is constantly in chemiosmotic and redox disequilibria, with sustained electrochemical gradients across cellular membranes (Allen 2010; Lane and Martin 2012; Sousa et al. 2013). Many of the core bioenergetic traits are shared across all domains of life. In fact, the ATP synthase, the enzyme used by all three domains in the synthesis of life's chief "energy currency", ATP, is thought to be ancestral to all extant life (Gogarten et al. 1989; Hilario and Gogarten 1993). So, it is reasonable to extrapolate that the first cells were, like all modern organisms, chemiosmotic (Lane and Martin 2012; Sojo et al. 2014).

But there are crucial problems with this theory, main of which is a chicken-or-egg conundrum: did early life waste vast amounts of energy by generating a gradient before it had the enzymes to exploit it, or did it start exploiting a gradient before it had the ability to generate it? Such a wasteful scenario as the one presented by the first option would seem unlikely to provide an ecological advantage, but the alternative seems like a dead end. A possible explanation is that the ATP synthase is indeed ancestral, but its job was as a proton pump in adaptation to acidic environments (de Duve 1995); yet this leaves the question of the ubiquitous distribution of chemiosmosis unanswered.

One elegant solution to this puzzle has come from the prediction (Russell et al. 1989; Russell et al. 1993; Russell et al. 1994) and later finding (Kelley et al. 2001) of alkaline hydrothermal vents. In the context of an early Earth, these geological systems would have provided not only a natural proton gradient at the interface between their alkaline fluid and the relatively acidic ocean, but also the chemical reagents needed to start the abiotic production of biochemical material (Martin and Russell 2003; Martin and Russell 2007). Nevertheless, details of the specific processes, as well as experimental demonstration, are notably lacking (Sojo et al. 2016).

In this document in general, the Origin of Life as an on-going evolutionary process is loosely considered to span from the advent of the first replicating units to the Last Universal Common Ancestor (LUCA) and the origin of the first true modern cells. It is considered throughout this document that Russell and co-workers' alkaline-vent theory provides the most plausible suggested scenario for the origin of life on the early Earth.

Whatever the details of the origin, there is no question that all life on Earth is related. As far as can be traced, there is only one common type of life that survived to this day. Whether this says something about chemistry or about ecology is difficult to ascertain, due to the impossibility of drawing conclusions from negatives in science. It is possible that life was a "lucky" combination of chemical events, and that we are largely or even entirely alone in the universe, which would in turn explain why more than 60 years of prebiotic chemistry research have come nowhere close to producing a fully-fledged independently living cell in the lab. Alternatively, it is possible that the unity of life on Earth is a matter of ecology, and that extant life either wiped out any independently derived competitors, or has perpetually prevented them from ever arising. Darwin favoured this latter view, when considering the spontaneous abiotic formation of the first biological molecules:

> *(...) at the present-day such matter would be instantly devoured or absorbed, which would not have been the case before living creatures were formed (...)*
>
> Charles Darwin (1863)

If this were so, then the as-yet limited success of prebiotic chemists could be simply because we have been asking the wrong questions, or just haven't yet come up with the right answers.

Either way, and whether life is a "lucky accident" or essentially inevitable wherever the right conditions are met, it is evident that there is only one type of life on Earth. This single form of life is divided into three sub-types or "domains", and their similarities are as stunning as their differences.

1.2 The three domains of life: archaea, bacteria, and eukaryotes

While all life on Earth is related, there are three main sub-types, or "domains": the archaea, the bacteria, and the eukaryotes (also frequently called "eukaryota", and somewhat less frequently "eukarya" or "eucarya"). The first two domains together are called the "prokaryotes", a term that in itself has different meanings to different researchers within the field.

Literally, "prokaryote" is Greek for "before the nucleus" so, like in the case of "life" above, it could actually be considered an anti-definition that simply alludes to the fact that species of these two domains are not eukaryotes. But it is the "before" part that has historically led to bitter arguments, mainly because consensus on the branching pattern of the three main arms of the tree of life remains unattainable. In fact, this confusion chiefly stems from the shared and unshared properties between the three domains of life, and the further sub-groups (or "phyla") within them.

1.2.1 Some differences and similarities between the three domains

Inferring phylogenetic relationships between branches that not only diverged billions of years ago, but that have been exchanging genes on a frequent basis ever since, is not a trivial task. Table 1 below, although not fully inclusive, attempts to show a general picture of some seemingly clear-cut similarities and differences between the three domains, extracted from diverse sources in the literature.

Table 1. Some relevant differences and similarities between the three domains

Trait	Archaea (A)	Bacteria (B)	Eukaryota (E)	Comments
Lipid membrane	**A**	**B**	**E**	
Phospholipid backbone	*sn*-glycerol-1-phosphate in most, but 4-carbon and 5-carbon alternatives have been reported, potentially in crenarchaeota (Zhu et al. 2014)	*sn*-glycerol-3-phosphate	(Like **B**)	**ABE** all have glycerol-phosphate as lipid backbone, but **A** have mirror structure of that used by **BE.** Synthetic enzymes are unrelated (Koga et al. 1998).
Phospholipid tail (or chain) linkage	Typically ether	Typically ester, but ether has been observed (Lombard et al. 2012a)	(Like **B**)	
Phospholipid tail (or chain) composition	Typically polymers of isoprene	Fatty acids synthesised from acetyl-CoA (Nelson and Cox 2013)	(Like **B**)	**BE**: di-fatty acid esters. No cyclic hydrocarbons. **BE** do use isoprene units when synthesising steroids (e.g. cholesterol) (Madigan et al. 2011: 81)
Cycles in tails	Chains may have cyclic hydrocarbons	Not observed	(like **B**)	
Bilayer joining	Lipids from opposite sides of the bilayer can be covalently joined (e.g. crenarchaeol, biphytanyl). Membranes can be composed of these or mix of both. Monolayers provide stability at high temperatures.	Always bilayers (no monolayers observed)	(like **B**)	**A** can form monolayers by covalently joining the tails of two opposing lipids (Hanford and Peeples 2002). **BE** membranes are always bilayers; covalently joined monolayers have not been observed.
Polar heads & Glycolipids	Most are phospholipids, as in **B**, but in *Euryarchaeota*, particularly methanogens, the polar head is commonly a carbohydrate	Phospholipids: the polar head is typically composed of a phosphate bound to an organic molecule, e.g., choline	Often like **B**, but polar headgroup can be composed of sugar polymers (glycolipids as opposed to phospholipids), similar to some **A**. Cardiolipin has two lipids joined by an isopropanol moiety at the polar head	Although mostly bacterial, eukaryotic lipids have some elements in common with some archaea, such as sugar polar heads
Cholesterol	Absent, but cholesterol is related to isoprenoids, the typical archaeal phospholipid tail	Almost entirely absent, but e.g. Mycoplasma require cholesterol for growth (Razin and Tully 1970)	One of the three main types of **E** lipids, the other two being glycolipids and phospholipids	Chiefly eukaryotic, but built from largely archaeal components (isoprenoids). Also observed in a few bacteria

Beyond the plasma membrane	**A**	**B**	**E**	
Peptidoglycan Composition	Have **pseudomurein** (aka **pseudopeptidoglycan**) instead, which is similar, but has a β-1,3-glycosidic bond as opposed to β-1,4-glycosidic bonds of peptidoglycan, and N-acetyltalosaminuronic acid is used instead of N-acetylmuramic acid. Some have different wall (remarkably, *Methanosarcina*), and a few have no wall at all (*Thermoplasma*).	**Peptidoglycan** is made of N-acetylglucosamine in β-1,4-glycosidic bond to N-acetylmuramic acid. A tetrapeptide is attached, made of L-Ala, D-Ala, D-Glu, Lys/diaminopimelic acid (DAP) D- amino acids and DAP have never been observed in **AE**		Bacterial Peptidoglycan and archaeal pseudo-peptidoglycan seem convergent and entirely unrelated (Madigan et al. 2011)
S-layer	Found in representatives of most groups of **A**. Cell wall of methanogen *Methanococcus jannaschii* consists **only** of the S-layer (i.e., it is enough to withstand osmotic pressure)	Present in several species of bacteria, both Gram-positives and negatives		S-layer is always the outermost layer. Composed of interlocking proteins or glycoproteins, arranged in paracrystalline structures. Can withhold osmotic pressure and filter out viruses. Reportedly, S layers could hold or at least help hold a H$^+$ or Na$^+$ gradient (Arbing et al. 2012). Some of the proteins and glycoproteins of **A** and **B** may be related (Baranova et al. 2012; Rohlin et al. 2012), but mostly they are not, even within domains.
Flagellum	Archaellum. ATP-powered. Several types of proteins, independent from those of **B**, but related to Type-IV pilins of **B**. Flagellum of *Halobacterium* (**A**) seems to be H$^+$-powered, like that of **B** (Streif et al. 2008).	One type of proton-powered flagellin, unrelated to that of **A**.	The "9+2" structure is different from those in known **A** and **B**, and appears to have evolved independently, although several proteins have homologues in archaea (Spang et al. 2015)	Although several involved proteins have homologues across the tree of life, **ABE** flagella seem unrelated (Zillig 1991; Madigan et al. 2011). Haloarchaea might be an exception, but this is likely due to HGT from bacteria (Nelson-Sathi et al. 2012).

Outer membrane	Outer membranes are not observed in general, but *Ignicoccus* has an outer sheath resembling the **B** outer membrane (Rachel et al. 2002).	In gram-negative **B**.	Not observed in the plasma membranes, but both mitochondria and chloroplasts are evidently derived from outer-membrane-bearing bacteria.	

DNA	**A**	**B**	**E**	
DNA replication origin	Multiple (Kelman and Kelman 2004)	Multiple replisomes have been observed, but typically single origin (Mott and Berger 2007)	Multiple	
DNA repair	shared with **E**	independent	shared with **A**	**AE** share replication and repair mechanisms for the most part
Supercoiling	Most **A** do negative, but some hyperthermophilic **A** do positive supercoiling (Forterre et al. 1996)	Negative	Negative	Mostly negative in **ABE**, but some **A** do positive
Chromosomes	One circular (Allers and Mevarech 2005)	Typically one circular (although some with more than one and some linear are known)	Multiple linear, mitochondrial and plastid circular	**AB** share circular chromosomes **E** are linear
Plasmids	✓	✓	✓ (some species)	
Histones	✓	✖ have different packing proteins called “histone-like” proteins	✓ except dinoflagellates, the only eukaryotes known to lack histones (Rizzo 2003)	
Nucleus	✖	✖ but planctomycetes surround their chromosome with membranes; known to be completely independent from eukaryotic nucleus (McInerney et al. 2011)	✓	Despite membranes observed around the chromosomes of some prokaryotes (e.g. planctomycetes), the nucleus is a purely eukaryotic trait, unrelated to any known prokaryotic structure
RNA primers in DNA synthesis	✓	✓	✓	DNA synthesis never starts *de novo*; DNA polymerases require a 3’ end to attach to (Alberts et al. 2007). Conversely, RNA polymerases only need a template strand.

Introns & splicing	Intragenic non-translated sequences exist, but no formal spliceosomal machinery exists.	Intragenic non-translated sequences exist, but no formal spliceosomal machinery exists.	Have typically many more and much longer introns than **AB**. Spliceosome is exclusive to **E**.	Although **AB** have non-translated intragenic sequences akin to introns, the spliceosomal machinery is exclusive to **E**
Discontinuous lagging-strand synthesis (Okazaki fragments)	✓	✓	✓	Okazaki fragments are universal, but not necessarily homologous (Leipe et al. 1999).
DNA replication enzymes	**A**	**B**	**E**	
Primase	Homologous to **E**, not to **B**	dnaG Primes new strands of DNA	Homologous to **A**, not to **B**	**B** primase is related to topoisomerases type I, II, VI but not to primases of **AE**
DNA polymerase III [B] DNA polymerase B [AE]	Homologous to **E**, not to **B**	Non homologous	Homologous to **A**, not to **B**	
DNA polymerase I		polA Excises RNA primer and fills in gaps		Only the 3'→5' exonuclease domain is conserved. Everything else seems unrelated to **AE**
Information flow	**A**	**B**	**E**	
Translation origin	First Met is formylated, like in **B**	First Met is formylated (N-formylmethionine)	Met is first, but tRNA for first Met is different	**AB:** N-formylated Met is start of translation **E**: Translation initiated at normal Met with special tRNA. Significantly, Met is suggested to be a late addition to the code
Operons & Polycistronic mRNA	✓	✓	✖	
tRNA	There are post-modifications of the tRNAs. The 7-deazaguanosine derivative archaeosine (G+) is always at position 15 in **A** tRNAs (Phillips et al. 2010)	Largely homologous to those of **AE**, but less related	Cytosolic are related to those of **A**, whereas mitochondrial ones are **B**	
Ribosomes	Similar in size to **B** but in sequence to **E**	Similar in size to **A**	Similar in sequence to **A**	Shared between **AE**, but clearly related to **B** (Woese et al. 1990)

Horizontal Gene Transfer (HGT)	Widespread	Widespread	Rare in macroscopic eukaryotes, importance still controversial in microscopic ones, but present (Katz 2015)	**AB** have undergone large amounts of HGT (Doolittle et al. 2003; Dagan and Martin 2007)
Others	**A**	**B**	**E**	
Replication	Asexual by fission (or fragmentation /budding)	Like **A**	Mostly sexual by meiosis and cell fusion	No meiosis or true recombination-sex in **AB**
Spores	Not observed	Some	Some, but unrelated	**A** do not appear to form endospores, like **B** and some **E** do
General phylogenetics	Appear to be as ancient as bacteria. Some clades are closely related to eukaryotes	As ancient as archaea. Similarly, some clades are closely related to eukaryotes	More recent than archaea and bacteria. Chimeric genomes with archaeal, bacterial, and unique components	Although the data overwhelmingly shows that **A** and **B** are older and **E** arose from a merger of the other two (Williams et al. 2013; Ku et al. 2015; McInerney et al. 2015), the topic is still debated by some (e.g. Doolittle and Mariscal 2015)
General Metabolism	**A**	**B**	**E**	
Energy and chemical sources	Typically chemotrophic. Although some, like *Halobacterium*, can use light in the production of ATP. Methanogenesis present in some, but not acetogenesis	As well as chemotrophy, **B** developed photosynthesis. They are extremely versatile biochemically, but no **B** are known to be methanogens (Baymann et al. 2003). Some perform the similar process of acetogenesis	Most phyla are heterotrophs. Plants and algae inherited photosynthetic carbon-autotrophy from cyanobacteria by endosymbiosis; however, they cannot fix nitrogen	No **A** has been found that fixes carbon using energy from light (photosynthesis) (Madigan et al. 2011). Similarly, no methanogenic bacteria are known, but acetogens (**B**) perform a chemically similar pathway, which may be related (Sojo et al. 2016). The acetyl-CoA pathway is the only CO_2-fixation pathway common to **A** and **B** (Fuchs 2011), suggesting it may have been ancestral in an autotrophic origin of life
Photosynthesis	No photosynthesis, but *Halobacterium* have bacteriorhodopsins and halorhodopsins that pump ions	Green (non)sulphur, purple (non)sulphur do anoxygenic. Cyanobacteria do oxygenic	Oxygenic, inherited from cyanobacteria[**B**] in plants and algae, and by secondary endosymbiosis in	Although many **A** can fix carbon directly from simple sources such as CO_2 or CH_4, and others get energy

	across the membrane due to light-driven conformation changes (Lanyi 1998). However, this is not directly coupled to carbon fixation, so it is photo-phosphorylation but not photosynthesis. Also, it is clearly horizontally acquired from bacteria (Nelson-Sathi et al. 2012)		e.g. euglena (Henze et al. 1995)	from light, none is truly photosynthetic.
Calvin cycle	Not observed (Berg et al. 2010)	Common in cyanobacteria. *Chlorobium* do reverse Krebs cycle as opposed to Calvin cycle (Fuchs et al. 1980)	In plants and algae, inherited from cyanobacteria (Martin and Schnarrenberger 1997)	Some **E** inherited from **B**. No **A** are known that do it.
ATP synthases	V-type, as in organelles of **E** (except mitochondria or chloroplasts, which have **B**-like F type)	Type F, like in mitochondria and chloroplasts of **E**	**Archaeal** V type operates as a pump (breaking up ATP) in vacuoles, endosomes, lysosomes, and secretory vesicles. **Bacterial** F type is main ATP synthesiser in mitochondrial inner membranes and chloroplast thylakoid membranes	All ATPases are related (Mulkidjanian et al. 2007). **E** have bacterial-like F-ATPases in mitochondria and chloroplasts, but archaeal-like V-ATPases in their other organelles (Senior and Wise 1983). Only the **B**-derived ATPases produce ATP in **E**; the **A**-derived V-ATPases consume ATP as proton pumps, e.g. in acidifying the vacuole
Carbon metabolism	**A**	**B**	**E**	
Glycolysis and gluconeogenesis	Saccharolytic **A** do variations of the Embden-Meyerhof-Parnas (EMP) and Entner-Doudoroff (ED) pathways that seem to have evolved independently (Verhees et al. 2003). Gluconeogenesis appears to (largely) be shared with **BE**	EMP pathway, which yields 2 ATP and 2 NADH, is more common, but the ED pathway, which yields half the ATP, is also observed	Mostly EMP pathway, with enzymes homologous to those of **B**, except for enolase, which is **A** (Hannaert et al. 2000)	The observation that gluconeogenesis enzymes (with only one exception) are shared across the tree of life but glycolysis enzymes are not (Verhees et al. 2003; Siebers and Schönheit 2005; Bräsen et al. 2014) suggests that glycolysis is a later development than gluconeogenesis

Fatty acid synthesis	FAS (fatty-acid synthase) is not present, but there are homologues that suggest fatty-acid synthesis may be ancestral (Lombard et al. 2012a; Lombard et al. 2012b)	✓	✓	**E** inherited fatty acid synthesis from **B**. **A** don't normally synthesise fatty acids, but they may have done so ancestrally

The constant discovery and re-classification of clades in all branches of the tree of life attests that the diversity of all three domains remains massively undersampled. As a simple demonstration, at the beginning of the work described in this thesis the NCBI taxonomy browser (www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi) reported 5 distinct top-level archaeal groups (or "phyla"). Towards the second year, the number had increased to 10. At the time of writing it stood at 13, including the recent discovery of the Lokiarchaeota, considered to be the closest known living relative of the original eukaryotic host (Spang et al. 2015); and by the time this document was being proof-read prior to submission the number had been collapsed back to only three major branches (the classic Euryarchaeota and TACK groups, plus the notoriously challenging (Williams and Embley 2014) DPANN group[i].

One major endeavour of biology is thus to elucidate the relationships between the different organisms that live and have lived on the Earth, a discipline called "phylogenetics".

1.3 Phylogenetics

Phylogenetics seeks to infer the evolutionary history and relationships between organisms, both extant and extinct. It is as such one of the two parts of the discipline of systematics, the other part being taxonomy, i.e. the assignment of names, description of the properties, and classification of the different organisms, without

[i] The branches leading to members of the DPANN group are particularly long, potentially leading to classification artifacts. Depending on the method and model used, the DPANN branch as sisters of the euryarchaeota, within the euryarchaeota (together with all other archaea), or as sisters of all archaea (Williams and Embley 2014). In this latter scenario they would currently be the closest known group to the original archaeal root, the last archaeal common ancestor (LACA).

regards to their evolutionary relationships (Michener et al. 1970, p. 3). In spite of the name, phylogenetics is not limited to using nucleotide sequences; in fact the use of the term “gene” in the name is far older than the modern discovery of genetic sequences, and relates instead to the inference of genealogies (e.g. Haeckel 1874). Literally, the word stems from the greek *φῦλον* (phýlon) for *tribe*, or *clan*, and *γένεσις* (génesis) for *source* or *origin* (Oxford Dictionaries 2016). Because of the degeneracy of the genetic code and the fact that only four options exist in each position of a DNA sequence versus twenty in amino-acid polymers, protein sequences tend to accumulate changes more slowly than nucleotide sequences, and are therefore useful for inferring relationships between more diverged sequences and species.

Morphological data is also often used in the analysis of evolutionary relations, and for a long time it was the only way of addressing the question of the relationships between species. Lamarck, Darwin and Wallace were of course unaware of the information in genetic sequences when they started making some of the first formal attempts at drawing out these relationships. Morphological data has however the caveats of convergence: multiple independent lineages can arrive at the same solution through independent means (such as wings in insects, pterosaurs, birds, and bats), a problem that extends into protein structure (Lupas et al. 2001). This problem is less common in biopolymer sequences, but it is by no means absent. For example, since only four bases exist in DNA sequences, a position can change from A to G, then to C, and then back to A (with any other succession in between), and therefore look to be ancestrally identical to a truly non-diverged sequence. Conversely, when two unrelated or distantly related biological sequences evolve rapidly, they can accumulate multiple identical changes due to convergence or sheer chance. A direct consequence of this phenomenon is *long-branch attraction* (LBA), a type of systematic error present in all types of phylogenetic tree-reconstruction methods (Philippe et al. 2005). In analyses affected by LBA, sequence positions that are identical due to chance or convergence can occlude the true phylogenetic relationship by being confused for common ancestry. This causes distantly related sequences to incorrectly appear to be close to each other in a tree. This type of error is of great importance in phylogenetic analyses, and it has confounded and may continue to confound the inference of ancient relationships (Lake 2015).

The comparative analysis of biological sequences leads to the identification of sequence homology, i.e. the detection of individual ancestral relationships between genes or proteins within and between species. There are multiple types of homology, the main of which are discussed in the next section.

1.3.1 Homologues, Orthologues, Paralogues, Xenologues

Homologue

The most generic of the terms discussed in this section, *homologue* simply defines two sequences that share an ancestry, whichever that may be. They can be sequences within the same genome, or in genomes from different species. At the simplest, homologues can be classified into three main sub-classes: *paralogues* arise by a gene-duplication event, *orthologues* arise by a speciation event, and *xenologues* by a horizontal transfer of genetic material. Importantly, two sequences (or portions of a sequence) that are similar (or even identical) by chance are *analogous*, not homologous, although technically they may be indistinguishable from such (Zhang and Kumar 1997).

Paralogue

Gene duplications are major forces of evolution (Ohno et al. 1968; Conant and Wolfe 2008). After a gene duplication event, the genome has two copies of the original sequence; these are said to be *paralogous*. The new copy may retain the original function, or it may acquire a new role. If the gene-duplication event occurred at the level of the whole genome, the term *ohnologue* (in honour of Susumu Ohno) is sometimes used; these large-scale gene duplications can have major effects in the evolution of complexity and novel traits, chiefly in eukaryotes (Ohno et al. 1968).

Orthologue

After two species diverge, many genes will be shared. These genes, derived by shared inheritance from a common ancestor following a speciation event are called orthologues. The identification of orthologues is confounded by gene duplications and gene losses, both before and after species divergence.

Xenologue

This term, less frequently used than those above, denotes genes that have been acquired by horizontal gene transfer. If undetected, these have a strong potential to confound the inference of phylogenetic relationships between species (i.e. they may appear to be orthologues, acquired by common inheritance).

There are other more complex types of homologous relationships not further discussed in this document. For example, *epaktologues* share an ancestry through the common independent acquisition of *fragments* of genetic material that are remodelled into a final product. This type of relationship is observed in the independent acquisition of protein domains followed by reshuffling into full proteins that therefore share an ancestry only partially (Haggerty et al. 2014).

1.3.2 Monophyly, Paraphyly, and Polyphyly

The members of a group or "clade" (including genes, species, families, or taxa in general) are *monophyletic* if they all cluster together in a section of a phylogenetic tree, i.e. if they all have a single common ancestor, and if the section of the tree that contains all of them includes only members of the clade. For example, in the animal tree all mammals cluster together under a common ancestor, and there are no members of this sub-tree that aren't mammals themselves.

If there is at least one taxon in the section of the tree that contains the clade that is not also a member of the clade as it was defined, then the tree is *paraphyletic*. For example, all the birds cluster together under a common ancestor that lies within the reptiles, but they are not themselves regarded as reptiles. Therefore, the tree is *paraphyletic* for the reptiles and *monophyletic* for the birds.

If two or more members of a defined group have different ancestors, the group is said to have a *polyphyletic* distribution. This happens frequently with artificially defined groups; for example, bats, insects, pterosaurs and birds constitute the "flying animals", yet naturally they do not cluster together in the tree of animals. Although obvious in this case, assigning members to groups is much less trivial in prokaryotes and the basal branches of the eukaryotes. Uncertain groupings are common in deep evolutionary analyses. Computational artefacts such as long-branch attraction, and biological phenomena such as horizontal and endosymbiotic gene transfers, can obscure the inference of relationships between species, the most fundamental of which is the search for a common ancestor of all extant species on Earth.

1.3.3 LUCA: the last universal common ancestor

Darwin was quick to realise that extrapolating backwards in time his theory of natural selection, by which common descent with modification underlies the origin of new species, inevitably led to few and possibly only one common ancestor of all life:

> *Analogy would lead me one step further, namely, to the belief that all animals and plants have descended from some one prototype. But analogy may be a deceitful guide. Nevertheless all living things have much in common, in their chemical composition, their germinal vesicles, their cellular structure, and their laws of growth and reproduction. We see this even in so trifling a circumstance as that the same poison often similarly affects plants and animals; or that the poison secreted by the gall-fly produces monstrous growths on the wild rose or oak-tree. Therefore I should infer from analogy that probably all the organic beings which have ever lived on this earth have descended from some one primordial form, into which life was first breathed.*
>
> Charles Darwin (1859 p. 484)

This "primordial form" has received several other names, including simply "universal ancestor" (e.g. Woese 1998), "cenancestor" (e.g. Edgell and Doolittle 1997), "progenote" (e.g. Woese 1998), "commonote" (e.g. Kagawa et al. 1995), "last universal ancestor" (LUA, e.g. Brooks and Fresco 2002), "last universal cellular ancestor" (LUCA, e.g. O'Donnell et al. 2013), "last universal cellular ancestral state" (LUCAS, e.g. Koonin 2009), and "last universal common ancestor" (e.g. Penny and Poole 1999). Although the different terms can have very different meanings, not discussed in this document, here the last of these will be used to refer simply to the last ancestor from which all extant life arose. In this thesis LUCA is assumed to have been cellular in nature, although its membranes and topology were unlike those of any living cell today. This is discussed in detail in Chapter 2.

1.3.4 The tree of life

The idea of the tree of life is familiar to biologists, dating back decades before the publication of the *Origin*, the only illustration of which is famously an evolutionary tree. Darwin himself, however, might have preferred a different analogy:

> *The tree of life should perhaps be called the coral of life, base of branches dead; so that passages cannot be seen.– this again offers contradiction to constant succession of germs in progress (...)*
>
> Charles Darwin (1838, pp. 25-26)

Darwin’s preoccupation follows from the potential confounding effect of a limited sample, taken from modern life, to infer the characteristics of the ancestors. Darwin, being a respected geologist long before achieving acclaim in biology (Herbert 2005), was well aware of multiple extinctions of entire clades of animals in the geological record, and the implication that unseen extinctions had on his theory. His concern translates directly into modern systematics: since so many branches of the ‘coral’ are dead, and the ‘bases’ and ‘passages’ leading to those that live today have been obliterated by evolution itself, drawing relationships between the deepest traceable branches is far from a trivial task. The three extant domains of life may have arisen in a number of ways, some of which are described in Figure 2.

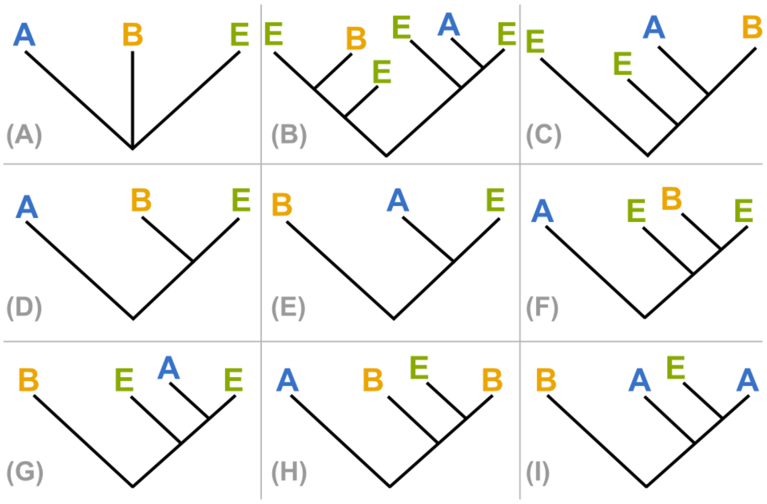


Figure 2. Possible branching patterns for the three domains of life in strictly vertical tree-like diagrams

The many shared and unique characteristics of the three domains (Table 1) confound the inference of their ancestral relationships. With all trees rooted at the lowest vertex, in the simplest possible scenario (**A**) the three domains arose directly from the same common ancestor (LUCA). (**B**) shows a "eukaryotes-first" scenario in which both the archaea and bacteria arose from within branches of the eukaryotes by reductive simplification and are thus convergently prokaryotic; this would explain why eukaryotes share some traits with archaea and others with bacteria (Table 1). (**C**) is also eukaryotes first, but a now extinct "last prokaryotic common ancestor" evolved initially, which underwent reduction and later gave rise to the archaea and bacteria; alternatively, the common prokaryotic ancestor could have arisen directly from LUCA (diagram not shown), i.e. the inner eukaryotic branch in (**C**) would be absent and eukaryotes are sister to the prokaryotic ancestor, as opposed to ancestral themselves. In (**D**) and (**E**) the eukaryotes are sister taxa to one of the two prokaryotic domains, and they share a now extinct common ancestor that was itself sister to the other prokaryotic domain; (**E**) is equivalent to Woese's ribosomal-based 3-domains tree, described below. Similar to (**D**) and (**E**), in (**F**) and (**G**) the eukaryotes are sister to one of the two prokaryotes, but the other prokaryotic domain is more recent, arising from within the eukaryotes. Finally, (**H**) and (**I**) have the eukaryotes being derived from *bona fide* members of the bacteria or the archaea, respectively. None of these branching patterns can readily explain all the similarities and differences in traits between the three domains. Several other scenarios are not shown (e.g. archaea arising from within bacteria or vice versa).

The three domains share multiple properties and their corresponding genes (Table 1). In general, eukaryotic genomes are more complex and seem to share several core traits with the archaea (mostly in information processing, i.e. DNA replication, transcription, and translation), and several equally important others with the bacteria (notably metabolism and membranes). All three domains have multiple unique traits, many of them fundamental. In the scenario of Figure 2A, the three domains arose essentially simultaneously from a common ancestor that would have had all the features that extant organisms share at present, and potentially others that are now lost. This ancestor was not necessarily a cellular independent entity in the modern sense, but may have instead been an unstructured population with vast horizontal transfer of genetic material, and in which genotype and phenotype were not clearly differentiated; that is, LUCA was a 'progenote' (Woese 1987; Woese 1998). The first archaeon picked some of the traits, the first bacterium some others (part of which were shared with the archaeon), and the first eukaryote picked more than its two sisters' share, ending up having more in common with both. This option has the problem of inevitably leading to a 'genome of Eden' (Doolittle et al. 2003): the LUCA that this scenario requires would have had copies of essentially every common gene in present-day organisms, which would have made it far more complex than any creature that followed. Although the idea of a progenote population with vast horizontal transfers makes this problem more tractable and is still prevalent in work by leading researchers in the field (e.g. Damer et al. 2016), a genetically undifferentiated progenote poses problems for a selective interpretation of life's origins. A population with vast and essentially uncontrolled horizontal transfers would be prone to invasion by 'cheats' (Maynard Smith and Szathmáry 1999), and therefore unviable. LUCA must have been both individual and cellular.

Alternatively, the eukaryotes may have been the earliest domain of life, such that essentially eukaryotes were LUCA. In this scenario, presented in Figure 2B, archaea and bacteria may have arisen by posterior simplification or 'streamlining', and they are as such "convergently prokaryotic" (i.e. the nucleus, mitochondria, Golgi, meiosis, and many other features were lost in parallel). Although far from being mainstream, this convoluted scenario is still considered plausible by some researchers (Doolittle and Mariscal 2015). A slight modification (Figure 2C) would consider the possibility that there was a common prokaryotic ancestor from which both the archaea and the bacteria arose, but the drastically different genetics of the two domains would rule this out immediately.

Coming into more mainstream views, another possibility is that the eukaryotes share a common ancestor with one of the two prokaryotic domains, and that this ancestor was itself sister to the other of the prokaryotic domains (Figure 2D-E). The familiar modern tree of life, as introduced by Woese and collaborators (1990), follows such a pattern, specifically the one in Figure 2E. The three distinct branches have the archaea and eukaryotes as sister clades, and their common ancestor stemming near the base, as a sister group to the bacteria (Figure 3).

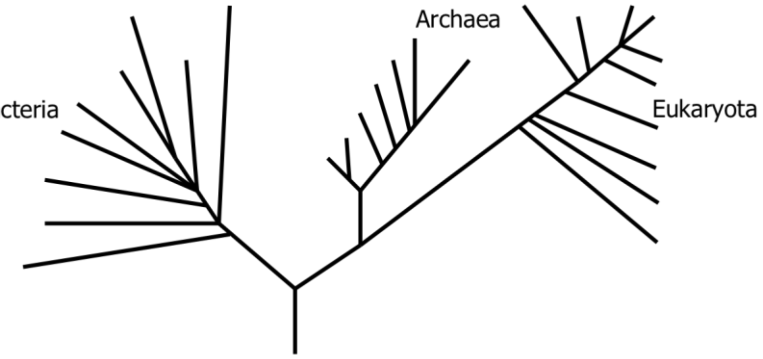


Figure 3. Woese's ribosomal RNA tree of life (the "3-domains" tree)

According to Woese's small-ribosomal-subunit tree (Woese et al. 1990), the archaea and eukaryotes are sister clades, and their common ancestor was in turn sister to the bacteria. If so, the eukaryotes are indeed more recent than the bacteria, but no earlier than the archaea, such that the "pro" in "prokaryotes" is unjustified and they should perhaps be called "akaryotes" instead. As discussed below, Woese's three distinct domains hold, but the branching pattern has been overthrown by recent findings: prokaryotes are indeed more ancient than eukaryotes. Adapted from the original by Woese and collaborators (1990).

This seminal work by Carl Woese and collaborators elaborated on Pauling and Zuckerkandl's (1965) ideas of molecular phylogeny, becoming the first formal attempt to classify all of life using genetic data in a systematic way. Woese used the small ribosomal subunit because of its distinctive domain-defining characteristics, and because of its high level of conservation when compared to most other genes (Woese et al. 1990). Similarly, the small ribosomal subunit, Woese assumed, is such an important gene that every living being will have a copy, and this copy will have been inherited ancestrally with little chance of horizontal gene transfer, so it should betray the evolutionary history of the species as a whole. With the data and methods available at the time, Woese's tree was a solid effort that painted the diversity of Earth's life into three well-defined colours that survive to this day. The strokes, however, have changed since.

Although still prevalent in modern reference textbooks (e.g. Lehninger's Principles of Biochemistry by Nelson and Cox 2013 p. 4), a reinterpretation of ribosomal sequences with modern available genomes and methods unequivocally places the eukaryotes within the archaea (Williams et al. 2012). Beyond this purely vertical view, it now seems clear that the relationships are more complex than Woese's original tree suggests, if however the clear-cut separations between the three domains that he and his colleagues introduced still remain.

In recent years, the evidence is becoming overwhelming that the basis of the eukaryotic cell was a fully-formed archaeon. The recent discovery of the Lokiarchaeota, an archaeal group that is closest in many traits to the eukaryotes than any other known prokaryotic lineage (Spang et al. 2015), seems to rule out the basal sisterhood of archaea and eukaryotes as domains, apparently making the scenario in Figure 2I the most likely. Lokiarchaea do not stem at the base of archaea, belonging instead to a branch well within the TACK superphylum.

However, the unequivocal assignment of eukaryotes to a branch within the archaea (Figure 2I) leaves out an equally crucial portion of the eukaryotic genome, namely and largely, phospholipid membranes and bioenergetic metabolism. In fact, archaeal-ancestry genes constitute the minority of the eukaryotic genome: most eukaryotic genes with a detectable prokaryotic ancestor are actually bacterial (Esser et al. 2004). And just as there is now very little doubt that the genome of modern eukaryotes, and therefore LECA, was archaeal in its information-processing machinery, it is equally clear that its membrane-forming and metabolic enzymes, including most of those for glycolysis, gluconeogenesis, fatty-acid synthesis and breakage, the citric acid cycle, and oxidative phosphorylation and respiration, are largely bacterial. The relation is again not one of sisters at the deepest level of the tree (Figure 2D), as it is clear that the eukaryotic bacterial ancestor was a proteobacterium, and likely an α-proteobacterium (Gray et al. 1999; Pisani et al. 2007), although potentially with a complex ancestry itself. This makes scenario H instead of I in Figure 2 more likely. So, do the eukaryotes branch within the archaea or within the bacteria? The answer is the simplest one, given the complexity of the data: both.

1.3.5 The ring of life

The dual archaeal and bacterial ancestry of eukaryotes suggests that the genomes of the latter evolved from a merger of genomes from the two prokaryotic domains. This resonated with the endosymbiotic theory of Lynn Margulis (Sagan 1967), elaborated on the work of Mereschkowski, who had introduced the ideas of symbiogenesis and endosymbiosis by considering that chloroplasts constituted a 'double plasma' in plant cells (Mereschkowsky 1905; Mereschkowsky 1910). Margulis successfully and correctly argued that both mitochondria and chloroplasts (amongst, incorrectly, other parts of the eukaryotic cell) had originated by endosymbiosis from bacteria[ii] (Sagan 1967).

The view gradually emerged that not only eukaryotic genomes but eukaryotes themselves as a domain are hybrid, and so there can be no tree of life, assuming traditional tree branches that can only depict vertical inheritance (Figure 2). Instead, the chimeric nature of eukaryotes turns the deepest relations between the three domains into a 'ring' as opposed to a tree (Rivera and Lake 2004) (Figure 4).

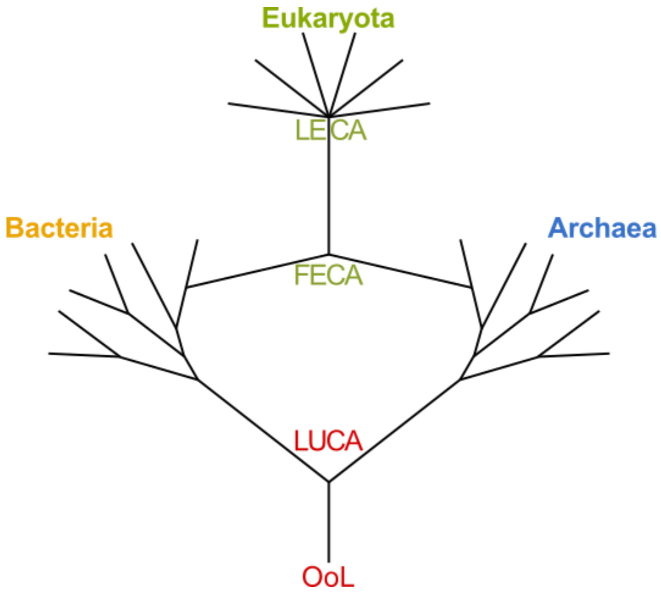


Figure 4. Lake and Rivera's "ring of life" (a "2-domains" tree)

This metaphor takes into account the contributions of both clades into forming the chimeric eukaryotic cell and genome. Due to their endosymbiotic origin, eukaryotes branch within both the archaea and the bacteria. The host of the endosymbiosis was an archaeon, specifically an "eocyte" (most simply in modern terms, a member of the archaeal TACK superphylum), while the endosymbiont was bacterial, specifically an alpha-proteobacterium. Adapted from the original by the authors.

Although some researchers still consider that the idea of eukaryotes arising first (earlier than the prokaryotes) may have some credence (Doolittle and Mariscal 2015),

[ii] The discussion of endosymbiosis is continued in more detail in the introduction to Chapter 4.

evidence in favour of the ring-of-life hypothesis is becoming too overwhelming to deny (Williams et al. 2013; McInerney et al. 2015).

Throughout this thesis, Rivera and Lake's ring (Figure 4) is assumed to be the one correct depiction of the broad relationships between the three domains of life. The main goal of this thesis is to shed light on some of the causes for this peculiar shape.

1.3.6 Horizontal (or Lateral) gene transfer (HGT/ LGT) and the "web" of life

Even if consensus about the general shape of the tree or ring of life were to be reached soon, an additional confounding factor in the inference of phylogenetic relationships is the widespread occurrence of horizontal gene transfer amongst prokaryotes (Doolittle et al. 2003; Martin and Roettger 2012; Koonin 2015). To a lesser extent, this phenomenon also has a role in the evolution of eukaryotic genomes, including multicellular (Katz 2015; Soucy et al. 2015).

Although known since before the structure of DNA (Tatum and Lederberg 1947), and indeed long before then (Griffith 1927), the massive phylogenetic significance of horizontal transfers became apparent towards the turn of the century (Aravind et al. 1998; W.F. Doolittle 1999; W. Doolittle 1999) with the advent of the genomics revolution, when closer analysis of individual gene trees failed to produce the expected pattern of canonical ribosomal species trees such as Woese's. Several of these conflicting trees had very strong support, so it was apparent that the discrepancies were not due to methodological errors (Koonin et al. 2001).

These transfers often have an ecological role; for example, the genomes of a number of hyperthermophilic bacteria show more imported archaeal genes than the average bacterium (Aravind et al. 1998), and the medical significance of horizontal mechanisms in the acquisition of bacterial resistance have been known for decades (Griffith 1927; Maynard Smith 1993). Both archaea and bacteria have evolved multiple dedicated mechanisms that mediate the acquisition of genes from the environment (Koonin 2015), which allows them to escape genome degeneration into "mutational meltdown" (or "Muller's ratchet"). Eukaryotes do not have these; the role of avoidance of Muller's ratchet is taken over by sex. However, massive horizontal transfers at the origin of the domain are evidenced by the chimeric nature of eukaryotic genomes discussed above. These large-scale influx of foreign genes from the symbiont into the host at the base of the eukaryotes is referred to as “endosymbiotic gene transfer” (Timmis et al. 2004), a sub-class of horizontal gene transfers (although the distinction is not always made in the literature). A further endosymbiotic transfer would occur later in the evolution of some eukaryotic branches, after the acquisition and endosymbiotic association of a cyanobacterium that would evolve into the modern-day plastids of plants, algae, and others.

Extensive horizontal gene transfers were recently suggested to have played a crucial ecological role in the evolution and remarkable adaptation to tolerance of environmental extremes in the multicellular eukaryote *Hypsibius dujardini* (Boothby et al. 2015), a tardigrade (a group of animals related to the arthropods). These animals tolerate extreme temperatures, radiation, and pressure, so it was suggested that HGTs from multiple organisms could allow them to survive in these environments, and that this should be a common mechanism in extremophilic animals (Boothby et al. 2015). However, these results were quickly and emphatically challenged by a separate group working on the genome of the same organism (Koutsovoulos et al. 2015). At the time of writing, the importance and extent, if any, of horizontal gene transfers in the evolution of modern eukaryotic multicellular lineages, and in particular animals, remains to be elucidated.

Conversely, the extent of horizontal transfers in archaea and bacteria is so large that early researchers wondered whether it was even possible to define the relationships between the clades (Stanier and van Niel 1962). In fact, the mechanisms of gene gain and loss play a greater role in prokaryotic evolution than point mutations (Koonin 2015). This massive and constant exchange of genetic material across unrelated branches means that the links between the varied species of life on Earth may resemble a “web” or “network” more than a tree or ring after all (W.F. Doolittle 1999; Doolittle 2000). This makes the prospect of phylogenetic analysis daunting and advances the question of whether phylogenetic analyses of microscopic organisms is a worthy use of researcher and computer time. It is, but with caveats.

In spite of widespread horizontal transfers, there is still an inherent verticality to the processes of inheritance and speciation in a large number of cases. Case-by-case gene transfers don’t exhibit the same systematic patterns as the simple bifurcations of vertical inheritance. Consequently, their effect on the shape of an overall species tree is not as strong. This translates into a “statistical” tree of life (Koonin 2015), i.e. an underlying vertical pattern that can generally be recovered for a full genome even if some or even all of its genes don’t fully exhibit it (Figure 5).

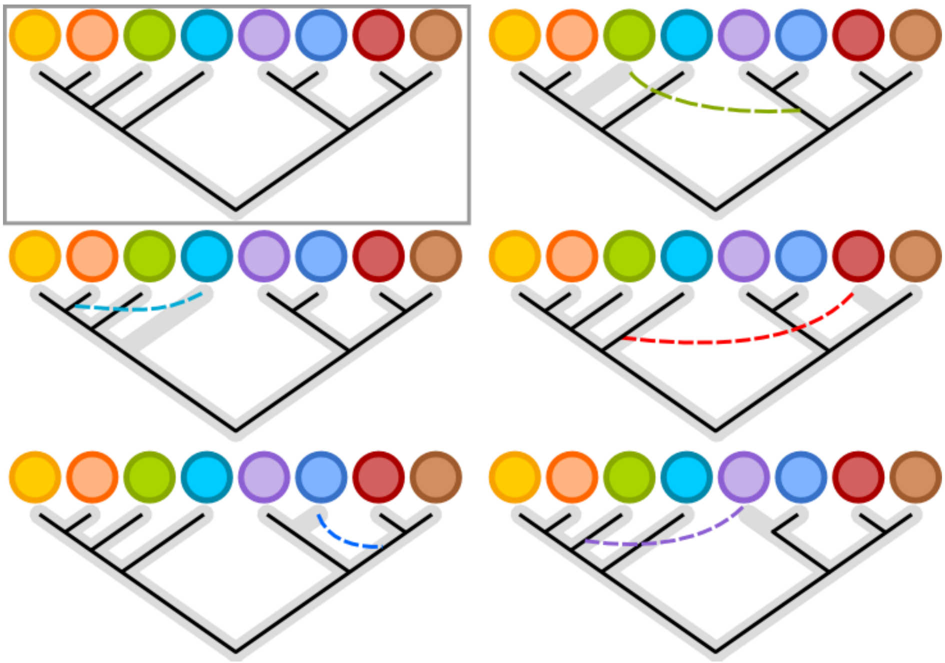


Figure 5. A “statistical” tree of life for full genomes, and trees for several genes with individual HGTs

The box at the top left represents a purely binary and vertical species-divergence tree. Even though all of the genes (subsequent diagrams) have an instance of horizontal transfer (coloured dashed branches) and therefore none reproduce the original tree, the overall tree (grey background pattern in all trees) can still be recovered from the ensemble. HGTs will weaken the support for the branches in the overall tree, represented in each individual gene tree by an empty grey branch in the position of the species branch in the overall tree. Since the transfers are not systematic across all genes, the tree at the top left still emerges. Sporadic gene losses have a similar effect to HGTs in that they weaken support overall.

Nevertheless, when transfers are systematic or massive, such as those in endosymbiotic events, or those suggested at the origin of most if not all modern archaeal clades (Nelson-Sathi et al. 2012; Nelson-Sathi et al. 2015), the inference of such a statistical tree becomes impossible, and models that allow for non-vertical inheritance become necessary (McInerney et al. 2015).

Energy underlies the advantages and consequent spread of many evolutionary innovations, both through vertical and horizontal mechanisms. The following section discusses bioenergetics and the roles that biological membranes play in it.

1.4 Membrane bioenergetics

Bioenergetics can be defined as the study of the transduction of energy from external sources into forms that can be utilised by the organism in performing various types of work (Skulachev 1988). This immediately makes bioenergetics a matter of membranes, since any source of energy, including light, will first encounter the exterior of the cell. It is then no surprise that membrane proteins typically play crucial roles directly in most and ultimately in all energy-transduction mechanisms across the tree of life.

The role of membrane bioenergetics in the major evolutionary transitions was pointed out over half a century ago (Mitchell 1957), but it has generally received limited attention until recently (Lane and Martin 2010; Lane and Martin 2012). Life is about disequilibrium; living beings are kept steadily at their far-from-equilibrium states by a continuous flow of both matter and energy (Harold 1986) that is operated across membranes, typically by membrane proteins. Energy flows in particular are mediated in the vast majority of cases through the generation, homeostatic maintenance, and exploitation of electrochemical and redox gradients across membranes (Harold 1986; Skulachev 1988; Allen 2010) or, in bioenergetic terminology, chemiosmotic coupling (Mitchell 1961; Nicholls and Ferguson 2013).

1.4.1 Chemiosmosis

The process of chemiosmosis was first discovered and advocated by Mitchell (1961; 1966), who broke away from what at the time was the consensus idea of substrate-level phosphorylation as the main (and only) source of ATP in the cell. Mitchell argued that the non-stoichiometric relation between oxygen and phosphorous in respiration, and the association of phosphorylation to membranes in mitochondria (amongst other observations), suggested a mechanism in which transfer of protons and electrons across the membrane was coupled to the production of ATP. In this process, protons are imported spontaneously into the mitochondrion (in the case of eukaryotes) down a pH and electrochemical gradient through the ATP synthase (ATPase). The influx of protons in turn decreases the strength of the gradient, which is recovered by pumping the protons back out (or by neutralising them) in a process powered by the spontaneous oxidation of a substrate. In this way, phosphorylation of ADP into ATP is coupled to the oxidation of a substrate, but the relation is not direct. The oxidation of the substrate can occur in multiple steps catalysed by independent proteins, each of which acts as a redox pump that uses the energy released by oxidation to extrude protons (Figure 6). This is termed an "electron transport chain".

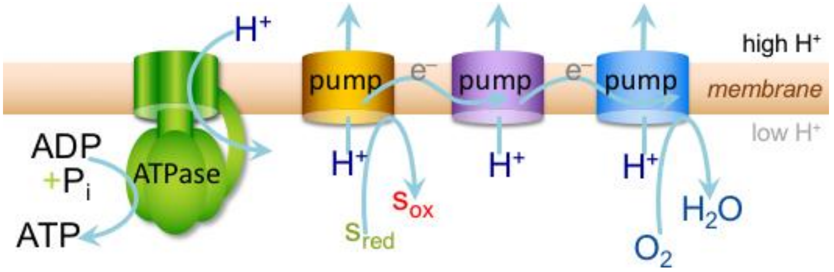


Figure 6. A simplified view of chemiosmotic coupling and the electron transport chain in oxygen-respiring mitochondria

Protons (H^+) travel through the ATP synthase (ATPase) from a region of high to a region of low H^+ concentration; the spontaneous flux is used to synthesise ATP. The two regions are separated by a H^+-impermeable membrane. Imported protons are then exported by redox pumps, which are powered by the multi-step oxidation of a reduced substrate, with oxygen as the final electron (e^-) acceptor. Adapted from the original by Mitchell (1961).

Mitchell's contributions, revolutionary in themselves, also served to highlight the importance of membranes to energy transduction and therefore life in general as more than simple containers of biological material.

1.4.2 Membranes

Membranes and their embedded proteins play a major role in most cellular processes (Singer and Nicolson 1972). All living organisms have at least one membrane composed of amphiphilic[iii] molecules, most generally phospholipids. A typical membrane phospholipid has four parts: a polar headgroup, a backbone that holds the different parts together, the links to the tails, and the tails themselves (Figure 7).

[iii] An *amphiphile* or *amphiphilic molecule* has two ends, one of which is *hydrophilic*, i.e. has an affinity to water, while the other is *hydrophobic*, i.e. it is generally apolar and has no affinity to water. Although related to solubility, the terms are not equivalent. Multiple salts (e.g. $BaSO_4$, AgCl, or FeS) are made of polar ions, yet have very low solubility in water; importantly, these salts are not *hydrophobic*, in spite of their poor affinity for water. The term *hydrophobic* is generally used to describe interactions such as van der Waals forces between apolar organic molecules or portions of these molecules.

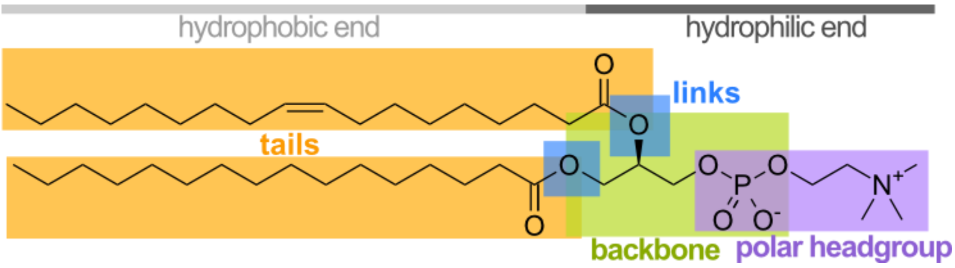


Figure 7. The four parts and two ends of a typical membrane phospholipid

The figure shows a typically eukaryotic phosphatidylcholine membrane phospholipid. The lipids of bacteria share a generally similar composition (although phosphatidylcholines are rare). Archaeal lipids are similar in overall structure in that two ends and four main components are present and have the same functions, but the biochemistry and genetics are entirely different (Chapter 2). More complex lipids exist, including dimeric structures joined at the heads (e.g. cardiolipin in mitochondria) or at the tails (e.g. crenarchaeol in crenarchaeota), multimeric structures with polysaccharic heads (in archaea and some bacteria), and others, but the general amphiphilic construction is retained.

The amphiphilic nature of phospholipids makes them assemble spontaneously into shapes that minimise energy. This can happen in a number of ways, chiefly micelles (simple spheres in which every polar headgroup faces outward towards the aqueous solvent, and every tail faces inward into each other) or, in the case of cellular membranes, by tightly packing lipids into parallel layers or sheaths, and two sheaths in an anti-parallel fashion. In this arrangement, the polar headgroups of one sheath face into the aqueous cytosol, while those in the opposing sheath face out toward the also aqueous environment; meanwhile, each hydrophobic tail is surrounded by the hydrophobic tails within its own sheath, and at the far end is met by the far end of the tails in the opposing sheath. This produces a “fluid mosaic” bilayer, into which membrane proteins are typically embedded (Figure 8).

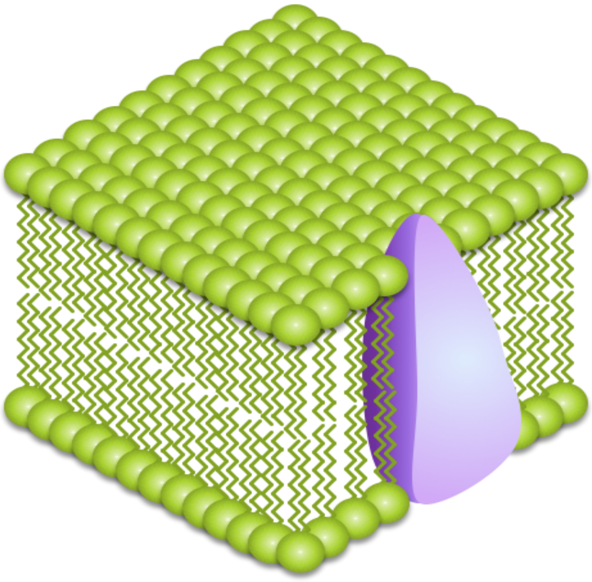


Figure 8. "Fluid mosaic" model of a membrane bilayer (section) with an embedded trans-membrane protein

Membranes work as "fluid mosaics" (Singer and Nicolson 1972), in which each phospholipid (green spheres with green tails) would be analogous to a tile. Membrane proteins (purple) can be embedded as active parts of this mosaic. The figure shows only one protein; the density of proteins in a real biological membrane can be expected to be higher than depicted.

In spite of vast differences in the lipids of archaea, bacteria, eukaryotes, and also within domains and even within different cells of the same individual in multicellular eukaryotes, the traditional model in Figure 8 is maintained throughout.

The role of embedded membrane proteins is central to life, be it in transport, cell-to-cell communication, cell division, and, crucially, bioenergetics. This thesis attempts to contribute to the emerging view of the central importance of membranes, their embedded proteins, and the disequilibria across these to life and living.

1.5 A brief outline of this thesis

This PhD thesis addresses the role of membranes and membrane proteins in shaping the evolution of life on Earth. Although connected by the common thread of membrane bioenergetics, the chapters are generally meant to stand on their own. A reader following the thesis from cover to cover will therefore encounter a number of repetitions throughout the document.

A number of questions remain unanswered in deep evolutionary biology; broadly, the details of the origin of life, the divergence of domains and species, and the origin of eukaryotes. The chapters in this thesis seek to contribute to elucidating these processes in the following ways:

Chapter 2 deals with the deepest split in the tree of life, namely the origin of archaea and bacteria from the last universal common ancestor (LUCA), and in particular the differences in their membranes. I explain the construction of and show results of a mathematical model that allows me to argue that the membranes of LUCA had to be leaky to H^+.

Chapter 3 is a theoretical work that follows the arguments in Chapter 2 to tackle the evolution of homochirality in general. After analysing relevant elements in the literature, I conclude that invocations to neither physical nor chemical prebiotic mechanisms are necessary to explain the origin of single-handedness in biochemistry. The *dual homochirality* of phospholipid backbones in archaea and bacteria suggests that homochirality is indeed the simplest evolutionary scenario, an intrinsic result of biochemical catalysis.

Chapter 4 returns to the differences between archaeal and bacterial membranes in relation to the origin of the eukaryotic cell. If eukaryotes arose from the endosymbiosis of a bacterium into an archaeon, then the first eukaryotic common ancestor (FECA) must have had an archaeal plasma membrane and bacterial proto-mitochondrial membranes. Yet, all modern eukaryotes have exclusively bacterial membranes. I argue that the reason for this convoluted evolutionary process was bioenergetic: as the mitochondrion became specialised as the powerhouse of the eukaryotic cell, energy production came to rely increasingly upon it, and the physiological adaptation of its bioenergetic proteins to the bacterial membrane became correspondingly crucial. I argue that replacing the original mitochondrial membranes with archaeal ones would have led to decreased fitness, so the bacterial membranes were retained, and the archaeal ones lost.

Chapter 5 closes this thesis by discussing adaptation in general in the context of membrane proteins and homeostasis. I report that membrane proteins have on average fewer detectable orthologues than water-soluble proteins, across the tree of life. I demonstrate that both divergence beyond sequence recognition and true gene losses have occurred. I conclude that, as emerging species specialise in new environments and functions, selective pressure is stronger on the outside. This leads to faster evolution outside the cell, and to the loss of a number of membrane proteins that are rendered useless in the new environment.

A general discussion of how each of these topics fit into and have contributed to the shape of the tree of life as we know it is presented in the final discussion in **Chapter 6**, along with a number of open questions that stem from the results of this thesis.

2 THE DIVERGENCE OF ARCHAEAL AND BACTERIAL MEMBRANES

Note: this chapter contains large sections adapted directly from the original research article "*A bioenergetic basis for membrane divergence in archaea and bacteria"*, published in *PLoS Biology* in August 2014 in collaboration with PhD supervisors Prof. Andrew Pomiankowski and Dr. Nick Lane (Sojo et al. 2014).

2.1 Summary

At its simplest, life is defined by its cellular nature, the separation of the living being's interior from the environment, as provided by biological membranes. It is therefore surprising that the two oldest domains of life, archaea and bacteria, which share multiple fundamental traits such as transcription and translation, do not share plasma membranes, a feature that must have been crucial for early life and predictably for the last universal common ancestor (LUCA). Archaeal phospholipids are typically composed of isoprenoid chains ether-linked to a glycerol-phosphate backbone, while bacterial phospholipids are typically composed of fatty-acid tails in ester linkage, also to a glycerol-phosphate backbone. However, the stereochemistry of the backbones is inverted: while archaea use *sn*-glycerol-1-phosphate, bacteria use the enantiomer *sn*-glycerol-3-phosphate. Both molecules are synthesised from dihydroxyacetone phosphate (DHAP) and NAD(P)H, but the synthesising enzymes are unrelated. Selective explanations have been put forward for the disparities in tails and links, but the subtler difference in backbones remains unexplained. A possible resolution of this paradox is advanced here by studying energy fluxes in early protocells that depended on a geochemically generated ion gradient, such as those in alkaline hydrothermal vents. Results show that geochemical ion gradients could power carbon and energy metabolism, but only if the membranes were leaky: a modern, more impermeable membrane, would insulate protocells from the geochemical gradient, preventing energy flow. The development of modern membranes with glycerol-phosphate backbones had to wait until after the origin of pumping. However, pumping protons across a leaky membrane offers no advantage: energy is wasted extruding protons that readily flow back in through the permeable membrane. Since sodium ions are considerably less permeable than protons, a sodium-proton antiporter (SPAP) could

have been a crucial adaptation preceding the evolution of both pumping and modern membranes. Still without active pumps, SPAP would have generated a sodium gradient in addition to, and powered by, the natural proton gradient. This would have enabled survival in smaller gradients, facilitating ecological spread and the incipient divergence between the proto-archaea and proto-bacteria. SPAP would also have provided a sustained advantage to decreasing membrane permeability after the origin of pumping, after which the evolution of modern proton-tight phospholipid membranes became advantageous. Crucially, this would have happened independently in the archaea and bacteria, explaining the vast differences in the phospholipids of the two domains, and in particular the subtlest of these: the backbones of the two domains have opposing stereochemistries because they were developed independently from a LUCA that had neither, and whose membranes were consequently much more permeable to protons than modern ones.

2.2 Introduction

There is now all but universal agreement that archaea and bacteria are the deepest branches of the tree of life, with eukaryotes arising considerably later from an endosymbiosis between *bona fide* members of the two older domains[iv] (Rivera and Lake 2004; Pisani et al. 2007; Dagan et al. 2010; Williams et al. 2012; Williams et al. 2013; McInerney et al. 2015) (although see Doolittle and Mariscal 2015 for an alternative view). Understanding this deepest split, the divergence of the two prokaryotic domains, should help shed light on the nature of the last universal common ancestor (LUCA), and potentially on the origin of life. Archaea and bacteria share important core components of their biochemistry, including transcription (Werner and Grohmann 2011), the genetic code (Koonin and Novozhilov 2009), and the ribosomal translation process and machinery (Londei 2006), yet they differ for unknown reasons in equally fundamental traits, including DNA replication (Leipe et al. 1999) and,

[iv] Note that the endosymbiotic origin of eukaryotes, as well as widespread horizontal and endosymbiotic gene transfers between all domains of life, mean that the relationships between the clades are not tree-like (i.e. simple vertical inheritance with divergence from a common ancestor does not suffice as an evolutionary model to explain the genomes and traits of extant species; gene flows are often not tree like). For simplicity, however, the phrase 'tree of life' is used generically in this thesis to refer to the relationships between clades, regardless of whether or not they are literally tree-like.

notably, their membrane lipid composition (Koga et al. 1998). Given that cells and life itself are defined by membranes that separate the biological material inside from the environment that surrounds it, this observation is surprising.

Phospholipid side chains are typically isoprenoids in archaea and fatty acids in bacteria, and the links to the glycerol-phosphate backbones are typically ethers in archaea and esters in bacteria. While these differences could reflect adaptive evolution (Valentine 2007), archaea and bacteria also differ in the stereochemistry of the glycerol-phosphate backbones (Koga et al. 1998). Archaeal lipids have an *sn*-glycerol-1-phosphate (G1P) backbone, while bacteria use the mirror structure (or enantiomer) *sn*-glycerol-3-phosphate (G3P) (Figures 9 and 10).

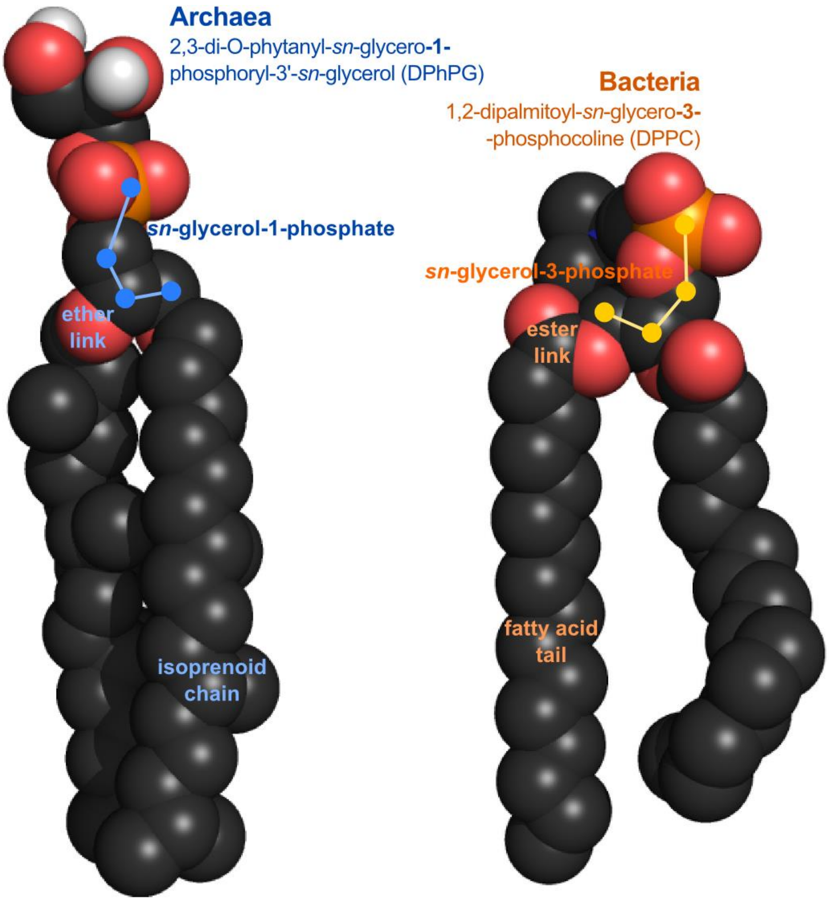

Archaea
2,3-di-O-phytanyl-*sn*-glycero-**1**-phosphoryl-3'-*sn*-glycerol (DPhPG)
sn-glycerol-1-phosphate
ether link
isoprenoid chain
Bacteria
1,2-dipalmitoyl-*sn*-glycero-**3**--phosphocoline (DPPC)
sn-glycerol-3-phosphate
ester link
fatty acid tail

Figure 9. Archaeal and bacterial membrane lipids (3-D structures)

Examples of extant phospholipid structures of archaea (left) and bacteria (right), with carbon-bound hydrogen atoms removed. Images rendered in PyMOL (Schrödinger LLC 2010) from PDB files downloaded from Lipidbook (Domański et al. 2010). See details in the skeletal chemical structures of Figure 10.

No selective reasons can readily explain these opposite stereochemistries (Koga et al. 1998; Sousa et al. 2013). The synthesising enzymes, *sn*-glycerol-1-phosphate-dehydrogenase (G1PDH) in archaea and *sn*-glycerol-3-phosphate-dehydrogenase (G3PDH) in bacteria, both reduce dihydroxyacetone phosphate (DHAP) (Figure 10), but they do not seem to have any phylogenetic relation. This has led to the surprising conclusion that they must have arisen independently in the two domains (Koga et al. 1998); this would imply that LUCA either had no membrane at all, or that its membranes were unlike modern ones. Yet life is defined by its cellular nature, that is, the separation of the inside of the living being from its environment as provided by amphiphiles assembled into a membrane. These large differences in the boundaries of the two basal domains of life therefore constitute a significant unresolved evolutionary problem (Koga et al. 1998; Martin and Russell 2003; Peretó et al. 2004; Lane and Martin 2012).

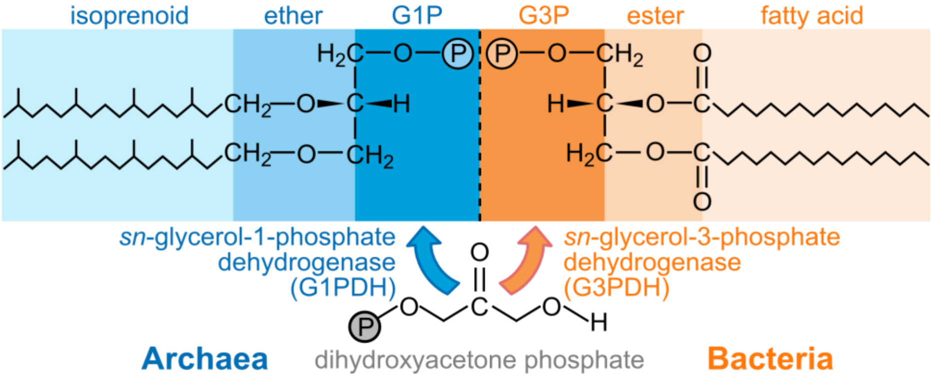


Figure 10. Archaeal and bacterial membrane lipids (skeletal structure)

Archaeal lipids (left) are typically composed of isoprenoid chains linked by ether bonds to an *sn*-glycerol-1-phosphate (G1P) backbone. The chirality of the two glycerol backbones is fully conserved within each clade, not only in structure but in their unrelated synthetic enzymes. Although ether linkages have been observed in bacterial membranes (Lombard et al. 2012a) and isoprenoids are common to all three domains, bacterial lipids (right) are typically composed of fatty acids in ester linkage to an *sn*-glycerol-3-phosphate (G3P) backbone. Despite widespread horizontal gene transfer, no bacterium has been observed with the archaeal enantiomer, or vice versa. See 3-Dimensional structures in Figure 9.

The significance of this puzzle is emphasised by the fact that archaea and bacteria (and indeed eukaryotes) share chemiosmotic bioenergetics (Lane and Martin 2012). That is, all known cells depend on the powering of ATP production in proteins such as the ATP synthase (ATPase) by ion gradients (i.e. differences in the concentrations of H^+ or Na^+) across membranes (Mitchell 1961). Surprisingly, the highly complex ATPase is as universally conserved as the ribosome (Stock et al. 1999), and the phylogenetic split between the two domains is equally deep for both structures, suggesting that both were already present in LUCA (Gogarten et al. 1989; Mulkidjanian et al. 2007; Lane et al. 2010).

However, the idea of a fully chemiosmotic LUCA is questionable, not least of which due to the vast differences in membranes described above. First, the proteins that pump ions across membranes are sophisticated, albeit no more than the ATPase and ribosomes themselves; but unlike these two, no ion pumps seem to be universally conserved. The synthetic pathways for both haem and quinones, the major respiratory protein cofactors, are in general different between the two domains (Sousa et al. 2013). Widespread horizontal gene transfers in prokaryotes obscure the reconstruction of the evolutionary history of pumps in the respiratory electron transport chain, but it seems likely that both active ion pumping and modern phospholipid membranes evolved independently in the two domains. These observations render the idea of a chemiosmotic LUCA with an ATP synthase powered by an ion gradient difficult to accept, in spite of the universality of this protein.

A possible resolution is that LUCA exploited geochemically sustained proton gradients such as those in alkaline hydrothermal vents (Martin and Russell 2003; Sousa et al. 2013). These vents are formed of microporous structures with an alkaline interior and a comparatively acidic exterior (Russell et al. 1994; Kelley et al. 2001; Martin et al. 2008). It has thus been suggested that the first cells might have arisen in this kind of hydrothermal environments (Martin and Russell 2003; Martin and Russell 2007; Lane et al. 2010; Ducluzeau et al. 2014; Lane 2014). However, the hypothesis that natural proton gradients could drive early carbon and energy metabolism faces a serious drawback, specifically that the influx of H^+ down a concentration gradient in the absence of active membrane pumps implies that electrical charges and concentration differences will swiftly counterbalance each other. Every positive charge (H^+) that comes into the cell increases the internal charge and makes the transfer of a subsequent charge more difficult, and eventually impossible in terms of net flux, i.e., the cell reaches an electrochemical “Donnan” equilibrium (Nicholls and Ferguson 2013). A possible resolution comes from the power of geochemically sustained proton gradients to offset the influx of positive charges from the acidic side (H^+) with an influx of negative charges from the alkaline side (OH^-) or, similarly, efflux of protons into the alkaline side. This power, which could have driven carbon and energy metabolism in LUCA, should depend on membrane permeability (Lane and Martin 2012), but the feasibility of this potential depends on dynamics of ion fluxes that are unknown. I have built a model to estimate qualitative differences in free energy (–ΔG) across organic membranes of different permeability exposed to geochemically generated proton gradients.

The system was designed as simple protocells lodged in vent pores and exposed simultaneously to flows of alkaline vent fluids on one side, and relatively acidic ocean waters on the other, giving a continuously replenished pH gradient sustained not by biology but by geology (Figure 11).

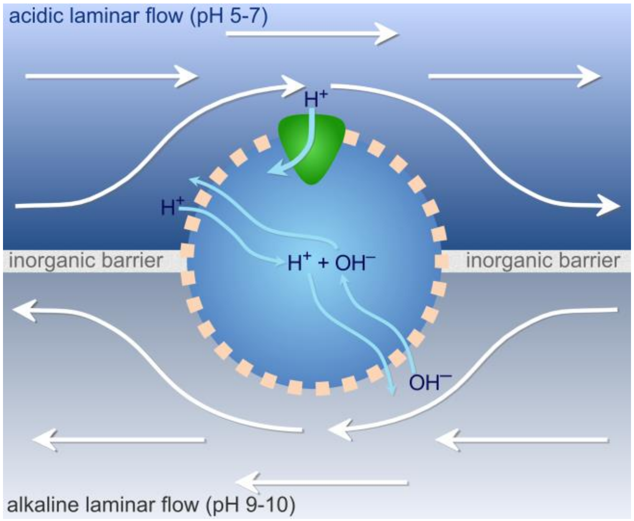


Figure 11. The model

A cell with a semi-permeable membrane sits at the interface between an alkaline and an acidic fluid. The fluids are continuously replenished and otherwise separated by an inorganic barrier. Protons (H^+) can flow into the cell from the acidic side (above) by simple diffusion across the membrane down their concentration gradient, with hydroxide ions (OH^-) entering in a similar manner from the alkaline side. Other ions (Na^+, K^+, Cl^-, not shown) diffuse similarly, as a function of their permeability, charge, and respective internal and external concentrations on each side. Inside the protocell, H^+ and OH^- can neutralise into water, or leave towards either side. Internal pH accordingly depends on the water equilibrium and relative influxes of each ion. A protein capable of exploiting the natural proton gradient (in green) sits on the acidic side, allowing energy assimilation via ATP production, or carbon assimilation via CO_2 fixation.

I here assume that these protocells already had genes and proteins, and that their membranes contained a primitive ATPase that relied on geochemically sustained H^+ gradients. In addition, I consider three other membrane proteins: the energy-converting hydrogenase (Ech), an enzyme central to energy metabolism in H2-dependent prokaryotes, and potential ancestor of respiratory complex I (Meuer et al. 2002); a simple H_2-powered proton pump; and a sodium-proton antiporter (SPAP) (Figure 12, Ech not shown).

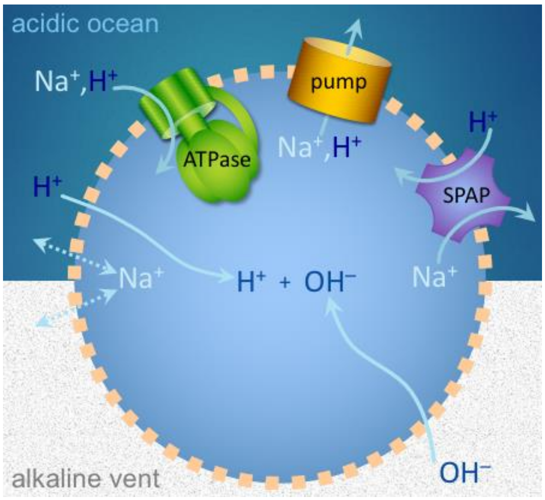


Figure 12. Three proteins considered in the model

The model starts with an ATP synthase (ATPase) promiscuous to Na^+ and H^+, as observed in methanogens (Schlegel et al. 2012). A generic pump (e.g. Ech operating in reverse direction) capable of extruding either H^+ or Na^+ is added next, followed by a 1:1 non-electrogenic sodium-proton antiporter (SPAP), whose effects are analysed with and without the presence of the pump. Ech, has a similar behaviour to ATPase when operating in the forward direction (not shown), in that it exploits the H^+ gradient, albeit for CO_2 fixation as opposed to ATP production.

In a number of archaea, Ech uses the proton-motive force to reduce ferredoxin, which in turn drives carbon fixation (Buckel and Thauer 2013); I therefore consider whether Ech could drive carbon reduction by H_2 in geochemically sustained proton gradients. In bacteria, Ech operates in reverse as an H_2-powered H^+ pump (Buckel and Thauer 2013), so I take it as a possible analogue of an active ion pump. Finally, I consider the effect of a non-electrogenic $1Na^+/1H^+$ antiporter on free-energy availability in natural proton gradients. Exchanging Na^+ for H^+ does not alter membrane charge directly, but the difference in permeability of the two ions alters ion flux, with significant effects on membrane potential $\Delta\psi$ and ultimately free energy

–ΔG. The findings allow the drawing of a bioenergetic route map from leaky protocells to the first archaea and bacteria, with divergent membranes and pumping mechanisms.

2.3 Results

2.3.1 Free-energy availability depends on membrane permeability

The early oceans may have been mildly acidic, as low as pH 5, while alkaline fluids would have been as high as pH 11 or 12 (Martin et al. 2008). I conservatively set a 3 pH-unit gradient, with the ocean at pH 7 and alkaline fluids at pH 10. The model shows that modern membranes were unviable: protocells with a membrane area covered by 1% ATPase embedded in proton-tight phospholipids with glycerol-phosphate backbones, giving a H^+ permeability $<10^{-5}$ cm/s, equivalent to extant archaea and bacteria (Deamer and Bramhall 1986), would collapse the proton gradients within seconds (Figures 13A and 13B). Collapse of the gradient was evident in proton-tight membranes across a range of gradients (Figure 13B). Following the dynamics of the system shows that the reason is as follows: protons enter through the ATPase faster than they can exit or be neutralised by OH^-, so H^+ influx rapidly reaches electrochemical equilibrium. In contrast, leaky protocells (equivalent to fatty-acid vesicles without glycerol-phosphate backbones) in a 7:10 pH gradient with 1% ATPase in the membrane retain nearly all the free energy available (Figure 13A), having a –ΔG only ~17% lower than an open system (i.e. a closed cellular system is energetically comparable to an open, non-encapsulated system). This is because proton flux through the ATPase is ~4 orders of magnitude faster than through the lipid phase, even with a high proton permeability of 10^{-2} cm/s (based on the kinetics of proton-flux through the ATPase, see Table 3 and Methods in section 2.5). Closed but leaky cell-like vesicles within vent pores incur only a small energetic cost (see comparison to the open system in Figure 13A), while providing the major advantage of retaining useful biomolecules, such as amino acids, nucleotides, or ATP.

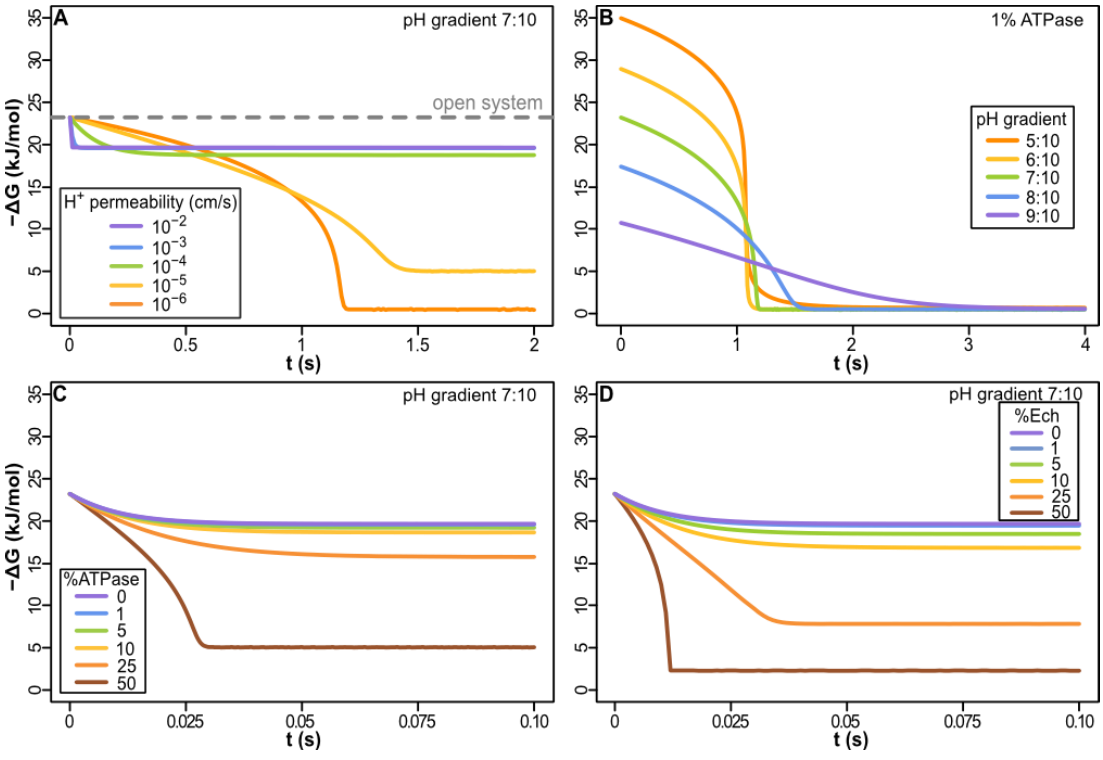


Figure 13. Dynamics of free-energy change (–ΔG) in cells powered by natural proton gradients

(**A**) Proton-permeable vesicles (H+ permeability $\geq 10^{-4}$ cm/s) have only a small loss of free energy compared with an open system (pH gradient 7:10, 1% ATPase). Reduced membrane H^+ permeability ($<10^{-4}$ cm/s), including permeabilities equivalent to modern membranes ($\leq 10^{-5}$ cm/s), collapse the gradient within seconds. (**B**) At low H^+ permeability (10^{-6} cm/s), –ΔG collapses regardless of gradient size. Within seconds, H^+ flux through ATPase equilibrates with the acidic ocean. (**C**) The collapse of –ΔG is more extensive the greater the amount of membrane-bound ATPase, even with a H^+–leaky membrane (10^{-3} cm/s). (**D**) With Ech, the collapse of the natural gradient is similar to that of the ATPase, showing that natural proton gradients can power energy (ATPase) and carbon (Ech) metabolism, given 1-5% coverage of enzyme in the membrane. Na^+ permeability was kept 6 orders of magnitude higher than that of H^+ throughout all simulations in this and all figures in this chapter. Except in (B), all results were calculated in a pH gradient 7:10.

Regardless of how leaky to H^+ the lipid phase is, the vesicles are sensitive to the amount of membrane protein, with higher proportions of ATPase collapsing the gradient (Figure 13C). In this case, the rate of H^+ entry through ATPase covering 10–50% of the membrane surface area is substantially faster than the rate of clearance of H^+ from inside the cell by neutralisation with OH^- or extrusion to the alkaline side, collapsing –ΔG. However, 1–5% ATPase in a leaky membrane (10^{-3} cm/s) retains a –ΔG of close to 20 kJ/mol (Figures 13A and 13C). With 3–4 protons translocated per

ATP synthesised (Table 3), this gives a –ΔG for ATP production of 60 to 80 kJ/mol, similar to modern cells and sufficient to drive intermediary biochemistry, including aminoacyl adenylation in protein synthesis (Pascal and Boiteau 2011).

Much the same applies to the Energy-converting hydrogenase (Ech), which in the ‘forward’ direction draws on the proton gradient to drive carbon metabolism via the reduction of ferredoxin (Buckel and Thauer 2013). As with the ATPase, protocells with 1–5% Ech in the membrane retain most of the free energy available from a 7:10 pH gradient (Figure 13D). Higher concentrations of Ech (10–50%) collapse –ΔG even more than the ATPase, as the rate of proton flux through Ech is double that of the ATPase, and its surface area is slightly smaller, so there are more proton pores per unit surface area (Table 3). Such high concentrations of Ech or ATPase are in any case improbable, but they demonstrate the range of conditions in which natural gradients can in principle drive carbon and energy metabolism.

Given a 7:10 pH gradient, it is therefore feasible to have 1–5% Ech and 1–5% ATPase in the membrane, driving both carbon and energy metabolism in cells with leaky membranes. But incorporation of either G1P or G3P backbones, or racemic mixtures of archaeal and bacterial lipids (which, surprisingly, can be as impermeable to protons as standard membranes (Shimada and Yamagishi 2011)), is not favoured because these backbones would significantly decrease permeability and therefore collapse the energetic driving force.

2.3.2 Pumping across leaky membranes gives no sustained increase in –ΔG

If leaky protocells with low amounts of ATPase and Ech (1–5%) are viable in natural proton gradients, but protocells with phospholipid membranes are not, then the evolution of active pumping becomes a paradox: pumping protons across a proton-permeable membrane does not significantly increase –ΔG, because the protons immediately return through the porous membrane. Modelling 1% Ech in the ‘reverse’ mode as a primitive H_2-powered proton-pump shows that in a 7:10 pH gradient –ΔG falls as membrane permeability decreases from 10^{-2} to 10^{-6} cm/s (Figure 14A). –ΔG here depends on two factors: active pumping and the natural pH gradient. As membrane permeability falls, the contribution of the natural pH gradient also falls, undermining –ΔG. In contrast, the benefit of pumping increases, as fewer protons return through the lipid phase. The balance between these two factors depends on the strength of pumping (which equates to the number of pumps, or % surface area). However, even when the pump occupies 5% of the membrane surface area, pumping H^+ gives no advantage until a modern permeability of $10^{–5}$ cm/s, there is no benefit to improving permeability across a 1000-fold decrease in permeability (Figure 14A). Therefore, there is no selective pressure to drive either the origin of pumping or the evolution of modern proton-tight membrane lipids in natural proton gradients.

Pumping Na^+ works better with leaky membranes (Figure 14B), as membranes are ~6 orders of magnitude less permeable to sodium than to protons (Deamer and Bramhall 1986). However, as with pumping H^+, $–\Delta G$ drops as the membrane becomes less permeable, because the contribution of the natural gradient falls, giving no continuous selective advantage to pumping Na^+. With a proton permeability $<10^{-5}$ cm/s, there is no advantage to pumping Na^+ at a pump density of 1–5% surface area compared with leaky protocells lacking a pump. Pumping Na^+ therefore offers an initial advantage, but there is no sustained selective pressure for tightening membrane permeability towards modern values.

Nor is there any advantage in the absence of a natural pH gradient, for example in the margins of the vent or when cut off from active flow. (This would also apply to the evolution of chemiosmotic coupling in the 'outside world' without natural gradients.) Under this condition, pumping either H^+ (Figure 14C) or Na^+ (Figure 14D) does offer a steadily amplifying advantage as membrane permeability falls. However, without an external pH gradient, $–\Delta G$ is well below the 20 kJ/mol required by modern cells to drive processes like aminoacyl adenylation for protein synthesis. Cells with permeable membranes ($10^{–2}$–$10^{–4}$ cm/s) are therefore unlikely to be viable unless powered by some other means (Mulkidjanian et al. 2007; Mulkidjanian et al. 2012). Hence, in either the presence or absence of pH gradients, there is no sustained selective pressure to drive the evolution of either active pumping or modern membranes.

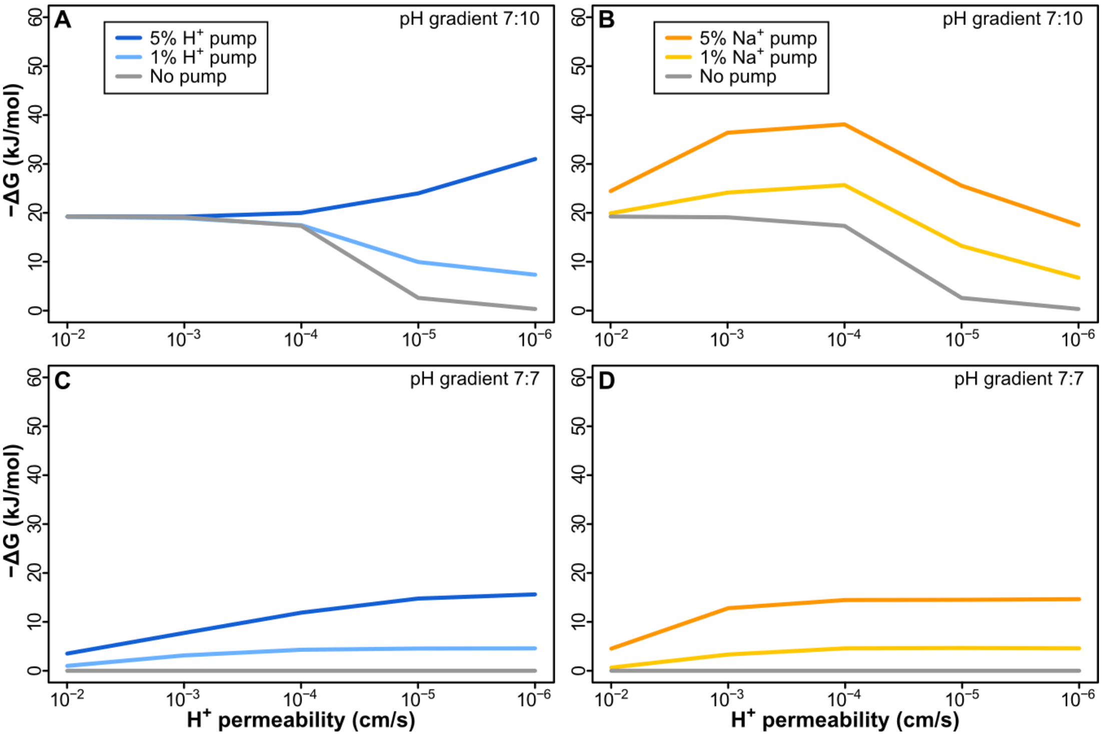


Figure 14. Pumping H^+ or Na^+ does not offer a sustained selective advantage

(**A**) Pumping H^+ in a membrane with 1% ATPase causes a sustained loss in –ΔG as membrane H^+ permeability decreases, with 1% pump. Even with 5% pump, –ΔG does not change over 3 orders of magnitude, and pumping only improves –ΔG near modern H^+ permeability ($\leq 10^{-5}$ cm/s). (**B**) Pumping the less-permeable Na^+ ion is initially better, adding to the natural gradient, but the early benefit is lost as membranes become tighter, due to the collapse of the natural H^+ gradient. In the absence of a gradient, pumping both H^+ (**C**) and Na^+ (**D**) offers a sustained advantage to tightening up membranes, but given a minimal requirement of around 20 kJ/mol for early life (Pascal and Boiteau 2011), the energy attained is not sufficient to power intermediary biochemistry.

2.3.3 Promiscuous H^+/Na^+ bioenergetics facilitates spread and is prerequisite for active pumping

The model presented here shows that leaky membranes were necessary to survive in natural proton gradients, but that pumping protons across such leaky membranes was fruitless. Yet free-living cells require ion-tight membranes and active pumping for bioenergetics. What drove this evolutionary change?

The hypothesis put forward here is that a necessary first step was adding Na^+ as an additional ‘promiscuous’ coupling ion. A non-electrogenic sodium-proton antiporter (SPAP), found widely in cells of all three domains of life, could in principle use a natural H^+ gradient to generate a biochemical Na^+ gradient. Because lipid membranes are ~6 orders of magnitude less permeable to Na^+ than to H^+, fewer Na^+ ions can pass through the lipid phase of the membrane (Deamer and Bramhall 1986), so the Na^+ gradient does not dissipate as quickly. As a result, Na^+ flux becomes more tightly funnelled through membrane proteins, improving the coupling of the membrane without changing its chemistry (Lane and Martin 2012). Because the H^+ gradient is sustained geochemically, SPAP simply adds a Na^+ gradient to the natural H^+ gradient. Taking advantage of mixed Na^+/H^+ gradients requires promiscuity of membrane proteins for both ions (Figure 12), which is indeed the case for several contemporary bioenergetic proteins in methanogens, including the ATPase (Schlegel et al. 2012) and Ech (Buckel and Thauer 2013).

SPAP increases proton influx, initially lowering $-\Delta G$ (Figure 15A). However, the coupled extrusion of the relatively impermeable Na^+ ions ultimately increases $-\Delta G$ by ~60% within minutes in a 7:10 gradient, saturating when SPAP covers 5% of the membrane surface area (Figure 15A). Importantly, the free energy available from pH gradients declines in more acidic conditions. $-\Delta G$ is greatest with a 7:10 gradient, lower at 6:9, and nearly zero with a 5:8 gradient, despite the three-order-of-magnitude correspondence (Figure 15B). This asymmetry arises because the pH scale is neither linear nor symmetrical (except around pH 7), and H^+ and OH^- flux through the membrane depend on concentrations as well as gradient sizes (Hodgkin and Katz 1949). At a pH 5-8 gradient, H^+ concentration is 10^{-5} in the acidic side, while OH^- concentration in the alkaline side is 10^{-6}; that is, in terms of concentrations the gradient is 10:1 in favour of H^+. Conversely, in an apparently equally sized 6-9 gradient, the situation is reversed, with the H^+ concentration being 10 times smaller than that of OH^-.

Comparatively high acidity and low alkalinity increases H^+ influx but hinders OH^- neutralisation, collapsing the H^+ gradient. Because Na^+ extrusion through SPAP depends on the natural H^+ gradient, SPAP increases $-\Delta G$ in relatively alkaline regions (pH 7–10 and 6–9) but has little effect on $-\Delta G$ in more acidic regions (pH 5–8), making acidic regions less favourable for colonisation, even with SPAP. When the rate of H^+ influx does not collapse the proton gradient, SPAP significantly increases $-\Delta G$, allowing survival in shallower pH gradients (Figure 15C). Taking $-\Delta G > 20$ kJ/mol as a marker for survival, 10% SPAP allows protocell viability to extend over 10-fold weaker gradients (i.e. 7:9 and 8:10; Figure 15C), a significant ecological advantage. In the context of hydrothermal vents, SPAP may have facilitated the colonisation of other parts of the same vent or possibly contiguous vents that had a weaker gradient.

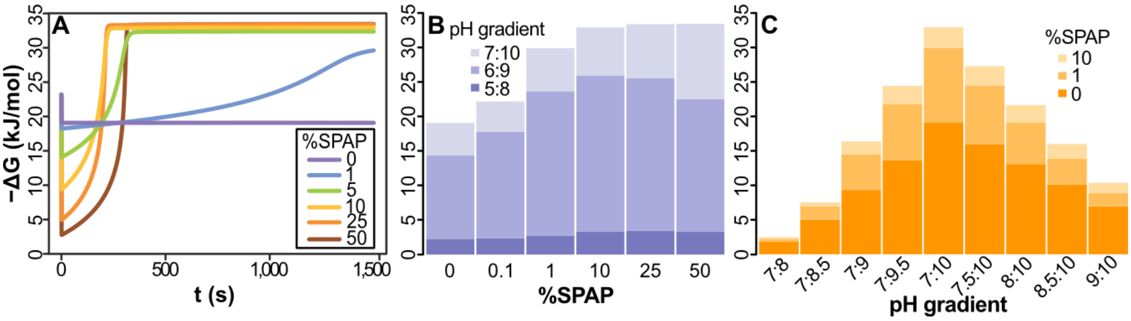


Figure 15. SPAP significantly increases free energy

(**A**) Because external Na^+ concentration (0.4 M) is higher than H^+ concentration (10^{-7} M), SPAP initially collapses –ΔG, and it takes minutes for the 1:1 H^+:Na^+ exchange to increase –ΔG significantly; eventually it renders an increase of ~60%. (**B**) The greatest increases are attained in relatively alkaline pH 7:10 environments, saturating as % surface area rises. Despite equivalent gradient sizes, the absolute difference in H^+ and OH^- concentrations means a 6:9 gradient gives a lower –ΔG, as the rate of H^+ influx is greater while neutralising OH^- influx is lower. A 5:8 gradient undermines –ΔG further, with or without SPAP. (**C**) SPAP facilitates colonisation of environments with weaker proton gradients. 1% SPAP pushes –ΔG above 20 kJ/mol in a 7.5:10 gradient, whereas 10% SPAP salvages an otherwise unviable 8:10 gradient. All simulations with 1% promiscuous ATPase, no pump, no Ech, and H^+ permeability 10^{-3} cm/s.

Crucially, SPAP is also a necessary preadaptation for the active pumping of protons, and for decreasing membrane permeability towards modern values. Whereas pumping H^+ in the absence of SPAP gives no sustained benefit to decreasing permeability in terms of –ΔG, the presence of SPAP in a leaky membrane allows pumping of H^+ to pay dividends. –ΔG now markedly increases with decreasing permeability (Figure 16A), for the first time giving a sustained selective advantage to tighter membranes. Again, –ΔG depends on the power of the pump, which varies with the proportion of surface area covered. As in the absence of SPAP, –ΔG depends on two factors: active pumping and the natural pH gradient. As membrane permeability falls, the contribution of the natural pH gradient also falls, undermining –ΔG. But in the presence of SPAP, 5% H^+ pump gives a steadily amplifying advantage to lowering membrane permeability, whereas 1% pump cannot sustain –ΔG when the contribution of the gradient is lost. Much the same applies to pumping Na^+, even with 5% pump (Figure 16B). As in the absence of SPAP, the lower permeability of Na^+ gives an initial benefit to pumping this ion, but this is lost as the membrane becomes tighter. The lower efficacy of pumping Na^+ relates to the much higher external concentration of Na^+: protons are being pumped against a 10^{-7} mol/L concentration, while the external Na^+ concentration is over six orders of magnitude higher at 0.4 mol/L.

With active pumping, tighter membranes, and SPAP, cells could colonise more acidic regions (Figure 16A) with weaker gradients (Figure 16C) and ultimately survive in the absence of a gradient altogether (Figure 16D). With no external pH gradient, SPAP interconverts efficiently between H^+ and Na^+, making it feasible to pump either ion (Figure 16D). These cells are now modern in that they have a fully functional chemiosmotic circuit and proton-tight membranes, and hence could evolve the traits required to leave the natural gradients provided by vents. The idea put forward here is that this process occurred independently in divergent populations that had spread widely using SPAP and colonised regions with weaker gradients (see Discussion in section 2.4). These independent populations subsequently evolved into the two main branches of early life, the archaea and the bacteria.

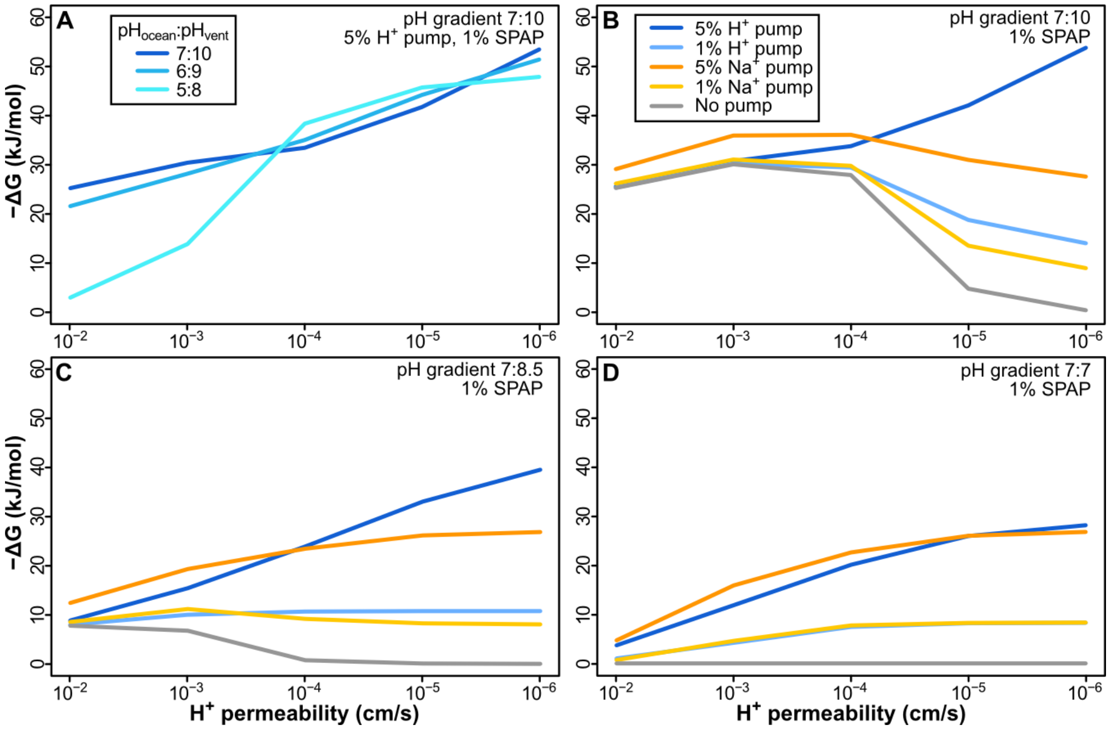


Figure 16. SPAP gives a sustained benefit to pumping, favouring tighter membranes and allowing free living

(**A**) Transition from highly H$^+$–permeable gradient-powered systems on the left to pump-powered with low H$^+$ permeability on the right. As in Figure 15B, relatively acidic environments (5:8) fail to replenish OH$^–$ from the natural gradient at high permeability (10$^{–2}$ cm/s). Tightening the membrane facilitates pumping but collapses the natural gradient, so the 5:8 system gains more at intermediate permeability (~10$^{–4}$ cm/s). With tighter membranes (10^{-6} cm/s) the cell is powered by its own pumping machinery. The opposing H$^+$ concentration is greater at 5:8, making it harder to pump against than in 6:9 or 7:10 gradients. (**B**) As seen in A, 5% H$^+$ pump provides sufficient power to make the sustained improvement of membranes advantageous. Conversely, 1% pump is insufficient either with H$^+$ or Na$^+$. 5% Na$^+$ pumping remains above the minimum 20 kJ/mol threshold, but the advantage to decreasing permeability is not sustained. Since SPAP interconverts between Na$^+$ and H$^+$ gradients, lowering the size of the gradient (**C**) reduces the difference between pumping H$^+$ and Na$^+$, ultimately making it equivalent to pump either ion in the absence of a gradient (**D**). Cells could not survive without a gradient unless relatively tight membranes are already in place, as –ΔG falls well below 20 kJ/mol. All simulations assume 1% ATPase and no Ech. Legend in B is common to C and D.

2.4 Discussion

The model presented in this chapter puts forward a resolution to the long-standing paradox of universal membrane bioenergetics but fundamentally different membranes (Lane and Martin 2012). In so doing, the model gives a striking insight into the deep evolutionary split between archaea and bacteria. It reveals that the late and divergent evolution of impermeable membranes arises as a simple outcome of the exploitation by a protocell[v] of natural proton gradients, such as those found in alkaline hydrothermal vents. These vents function as electrochemical flow reactors and, in the anoxic waters of the Hadean and Archean, provided suitable conditions for abiotic synthesis (Amend and McCollom 2009) and concentration (Baaske et al. 2007) of organics at the origin of life. Here I have assumed that protocells contained DNA, RNA, and proteins, but not modern membranes (yet they were still cellular in nature, having an organic boundary, albeit leaky to H^+). The model shows that, given the membrane proteins Ech and ATPase, natural proton gradients can in principle sustain both carbon and energy metabolism (Figures 13C and 13D). However, only leaky protocells with membrane H^+ permeability equivalent to fatty-acid vesicles can escape electrochemical equilibrium in natural proton gradients (Figure 13A). The results show that pumping either H^+ or Na^+ over leaky membranes gives no sustained advantage to decreasing permeability, even when this decrease is 1000-fold (Figure 14A). Early protocells, up to LUCA, could have been sophisticated in terms of genes and proteins, but the evolution of modern phospholipid membranes was a later development in evolutionary history.

The actual permeability of fatty-acid vesicles and modern phospholipid membranes is difficult to determine experimentally, as H^+ permeability depends in part on the permeability of counter-ions, and therefore varies with the composition of solutions used in measurements. Values of liposome permeability range from 10^{-4} cm/s (Deamer and Nichols 1983) to 10^{-10} cm/s (Nozaki and Tanford 1981; van de Vossenberg et al. 1995), with a consensus favouring a value of around 10^{-4}–10^{-6} cm/s (Deamer and Bramhall 1986). The proton permeability of fatty acids is higher, but again hard to constrain. The values used here (10^{-2}–10^{-3} cm/s for fatty-acid vesicles and $\leq 10^{-5}$ cm/s for modern phospholipid membranes) are necessarily approximate. But the argument relates to the principle of energy transduction in geochemical proton gradients and not to the specific values used for either membrane permeability or

[v] Here, a protocell is defined as a fully independent entity with a membrane and membrane proteins. However, the membranes are unlike those of modern cells, and the topology has the protocell embedded in a mixed microfluidic system with different pH levels at each of two sides, unlike any known modern cell.

enzyme kinetics (which also affect permeability). The key point is that leaky membranes were essential to transduce natural proton gradients in LUCA, and there was no advantage to be gained by the evolution of proton-tight membranes. Specifically, this means that glycerol-phosphate backbones, which would drastically decrease permeability, were a late addition to cell membranes.

Here I have suggested that the evolution of a sodium-proton antiporter (SPAP) was the key innovation that transformed the selective landscape before the evolution of pumping and modern membranes became consistently advantageous. SPAP adds a Na^+ gradient to the geochemically sustained H^+ gradient. As even proton-leaky lipid membranes are relatively impermeable to Na^+, these ions preferentially flow back through membrane proteins as opposed to the membrane itself, and thereby increase free-energy availability by up to 60% (Figure 15A). This enabled protocells to survive in 10-fold lower gradients (Figure 15C), facilitating the spread and divergence of protocell populations within vents. In addition, and for the first time, SPAP gave an advantage to actively pumping protons even across a leaky membrane (Figure 16A). This advantage amplified steadily as membrane permeability decreased, all the way towards values for largely impermeable modern membranes (Figure 16A).

This would imply that the SPAP is ancestral and must have been present in LUCA. Preliminary phylogenetic analysis is consistent with this prediction. BLAST (Altschul et al. 1990) results, presented in Table 2, show a significant match for archaeon *Methanococcus jannaschii*'s Mj1275 SPAP to an equivalent or very closely related protein in at least one member of 31 out of all 35 prokaryotic phyla known to the date of this analysis.

Table 2. BLAST search results for matches of the archaeal *M. jannaschii* Mj1275 SPAP to at least one member of each of the 35 prokaryotic phyla known at the time of this analysis

Results show matches of *M. jannaschii* SPAP gene Mj1275 to an equivalent or very closely related protein in at least one member of 31 out of all 35 known prokaryotic phyla. One archaeal (Nanoarchaeota) and three bacterial (Caldiserica, Dictyoglomi, Armatimonadetes) clades failed to give a match to a SPAP. These four are to date single-member phyla whose only known species may have either lost the gene over time, had it diverge beyond observable similarity to the *M. jannaschii* orthologue, or not have been fully annotated to date. Two further bacterial clades (Thermotogae and Tenericutes) do contain an Mj1275-matching SPAP gene, but the result has an E-value considerably larger than 10^{-10}, a conservative cut-off for deep phylogenetics (Sousa et al. 2013). Both phyla also had a single member species at the time of this analysis.

Phylum	G.I.	Description	%id	S	E
Archaea					
Euryarchaeota*	294496655	sodium/proton-potassium antiporter	30.57	157	3.00e-42
Thaumarchaeota	563488844	putative Na(+)/H(+) antiporter	28.72	146	2.00e-39
Korarchaeota	170290145	sodium/hydrogen exchanger	25.44	91.7	1.00e-21
Chrenarchaeota	352683176	Na(+)/H(+) antiporter	27.85	95.9	2.00e-21
Nanoarchaeota[†]	490715315	Type II restriction enzyme, methylase subunit	25.37	29.3	0.15
Bacteria					
Cyanobacteria	515885330	hypothetical protein. Sodium/hydrogen exchanger family	30.59	168	7.00e-45
Firmicutes	15893735	Na/H antiporter NapA	33.51	160	2.00e-42
Bacteroidetes-Chlorobi	548235349	putative uncharacterised protein. Sodium/hydrogen exchanger family	30.34	149	1.00e-37
β-Proteobacteria	490375968	Na+/H+ antiporter	28.12	140	2.00e-37
δ-Proteobacteria	493978264	Kef-type K+ transport system, membrane component. Sodium/hydrogen exchanger family	27.89	135	3.00e-34
Deinococcus-Thermus	297624885	sodium/hydrogen exchanger	28.00	130	2.00e-32
Chloroflexi	156742237	sodium/hydrogen exchanger	26.67	122	9.00e-31
Spirochaetes	517350815	hypothetical protein. Sodium/hydrogen exchanger family	27.13	125	1.00e-29
ε-Proteobacteria	390940331	Kef-type K+ transport system membrane protein. Sodium/hydrogen exchanger family	28.75	122	1.00e-29
Aquificae	225849059	Na+:H+ antiporter, NhaA family	27.66	115	6.00e-29
Elusimicrobia	189485528	NapA type Na+/H+ antiporter	27.55	111	1.00e-28
Fibrobacteres-Acidobacteria	522212591	hypothetical protein. Sodium/hydrogen exchanger family	25.68	114	5.00e-28
γ-Proteobacteria	495083969	sodium/hydrogen exchanger	27.79	120	6.00e-28
Fusobacteria	310779803	sodium/hydrogen exchanger	27.35	116	2.00e-27
Nitrospirae	206891081	Na/H+ antiporter	28.23	110	3.00e-27
Thermodesulfobacteria	551229848	sodium:proton antiporter	26.72	106	3.00e-26
Chlamydiae-Verrucomicrobia	494656847	sodium/hydrogen exchanger	26.08	109	4.00e-26
Synergistetes	357419296	transporter, CPA2 family. Sodium/hydrogen exchanger family	27.34	108	2.00e-25

Actinobacteria	501185687	putative Na+/H+ antiporter CPA2 family	26.67	112	5.00e-25
Planctomycetes	283779895	sodium/hydrogen exchanger	25.18	102	2.00e-23
Chrysiogenetes	317050984	sodium/hydrogen exchanger	26.08	94.7	1.00e-22
α-Proteobacteria	334343506	sodium/hydrogen exchanger	27.88	100	8.00e-22
Deferribacteres	555548272	hypothetical protein. Sodium/hydrogen exchanger family	28.20	90.5	2.00e-20
Gemmatimonadetes	226226447	putative sodium/hydrogen transporter	25.92	68.9	2.00e-13
Nitrospinae	491148743	Kef-type potassium transporter	25.07	60.5	2.00e-10
Thermotogae‡	389842970	NhaP-type Na+(K+)/H+ antiporter	24.80	35.0	0.061
Tenericutes‡	493942442	na(+)/h(+) antiporter	24.81	36.2	0.069
Caldiserica†	383788853	putative formate dehydrogenase subunit alpha	22.38	30.8	0.073
Dictyoglomi†	206901223	glycyl-tRNA synthetase, beta subunit	37.21	30.0	0.29
Armatimonadetes†	512551780	NADH dehydrogenase subunit M	32.35	28.5	0.93

Only the highest-scoring sequence for each clade is shown.
Phylum: each of the 35 known prokaryotic phyla, considering each of the proteobacteria separately.
G.I.: unique NCBI/GenBank sequence identification number.
Description: a brief summary of the annotation for the highest-matching protein found in the search.
%id: the percentage identity of the sequence to Mj1275.
E: a measure of the likelihood of finding such a match with score S by chance, in the given database.
S: the “bit score”. A measure of the quality of the alignment and match between the sequences.
* Euryarchaeota excluding the *Methanococcus* genus.
† Grayed-out phyla produced unsuccessful results.
‡ These phyla produced a match to a SPAP, but the result is below the threshold of significance ($E\leq10^{-10}$).

These results provide support to the suggestion of the universality of SPAP in spite of the stark dissimilarity in membranes, and pave the way for closer phylogenetic analysis of these antiporters as well as other related proteins.

The early operation of SPAP should have had the effect of lowering the intracellular Na^+ content substantially below the environmental concentration. The operation of Na^+ and K^+ antiporters, driven by natural proton gradients, could in principle have modulated intracellular ionic composition to the low-Na^+–high-K^+ characteristic of most modern cells, leading to selective optimisation of protein function without the need for a specific terrestrial environment with a particular ionic balance (Mulkidjanian et al. 2012).

In conclusion, these findings suggest that the membranes of LUCA were necessarily leaky, composed of simple amphiphiles, possibly fatty acids, which readily dissipate H^+ gradients by flip-flop (spinning of a protonated acid from the more acidic side into the opposite, more alkaline side, where the proton is released). Importantly, these ancestral membranes lacked glycerol-phosphate backbones (Figure 17).

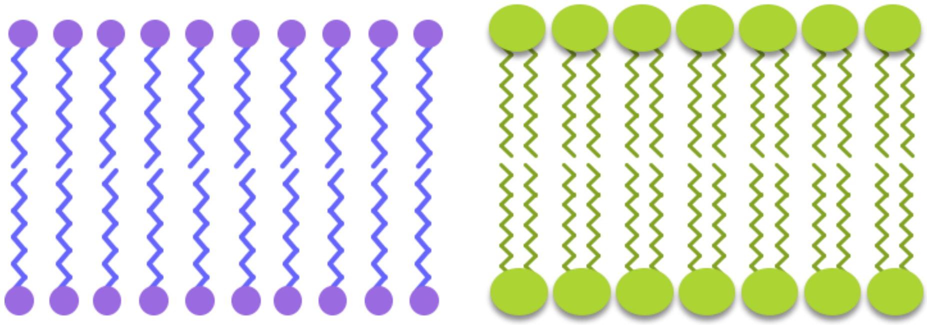


Figure 17. A proposed permeable ancestral fatty acid bilayer (left) and a modern impermeable phospholipid bilayer (right)

Fatty acid vesicles, left, are much more permeable than modern phospholipid bilayers, right, due to the presence of the glycerol-phosphate backbones in the latter (green spheres on the right).

Fatty-acid vesicles have long been considered plausible protocells because of their simplicity, stability and dynamic ability to grow (Hanczyc et al. 2003; Mansy et al. 2008; Budin et al. 2009), but are generally thought unsuitable for chemiosmotic coupling due to their high proton permeability (Deamer and Weber 2010; Mulkidjanian et al. 2012). Leaky membranes have therefore generally been interpreted in terms of heterotrophic origins of life (Deamer 2008). In contrast, the results presented here show that high proton permeability was in fact indispensable to drive both carbon and energy metabolism in natural proton gradients, consistent with an autotrophic origin of life. This requirement for leaky membranes in turn delayed the early evolution of glycerol-phosphate backbones and modern phospholipid membranes (Figure 18). The results of the model offer a selective basis for the universality of membrane bioenergetics and the ATPase, while helping to elucidate the paradoxical differences in membranes and active ion pumps. The deep disparity between archaea and bacteria in carbon and energy metabolism, and in membrane lipid stereochemistry, reflects two independent origins of active pumping in divergent populations (Figure 18). The core proteins involved – Ech and SPAP – are predicted to be central to membrane bioenergetics in archaea and bacteria, and indeed both are integral to respiratory complex I (Hedderich 2004; Sazanov and Hinchliffe 2006; Marreiros et al. 2013). Since the bacterial replicon is attached to the membrane during cell division (Jacob et al. 1966), the deep split between archaeal and bacterial DNA replication (Leipe et al. 1999) may also be linked to the late origin of phospholipid membranes, for these bioenergetic reasons.

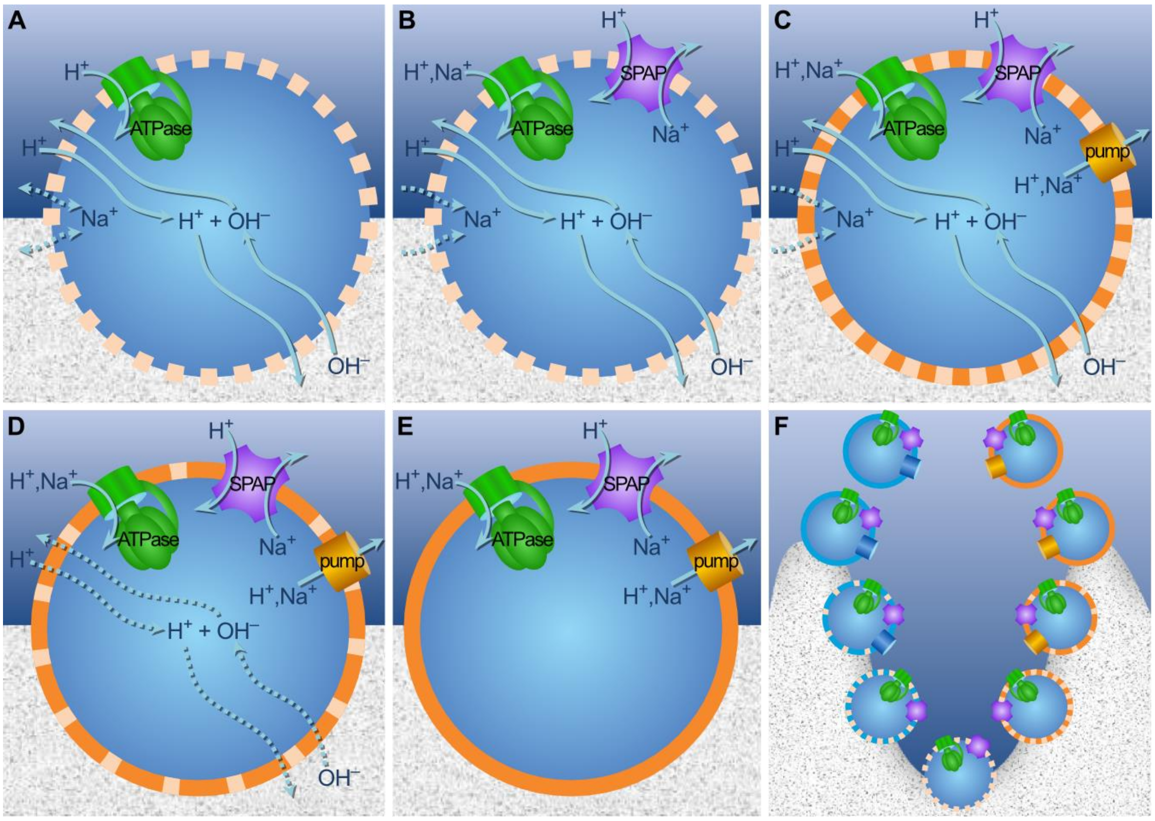


Figure 18. Divergence of archaea and bacteria

(**A**) Ions cross the membrane in response to concentration gradients and electrical potential. OH^- neutralises incoming protons. The H^+ gradient drives energy metabolism via ATPase, and carbon metabolism via Ech (not shown). (**B**) SPAP generates a Na^+ gradient from the H^+ gradient. As Na^+ is less permeable than H^+, SPAP improves coupling, given promiscuity of membrane proteins for H^+ and Na^+. (**C**) Membrane pumps generate gradients by extruding H^+ or Na^+ ions. (**D**) Exploiting natural gradients demands high membrane permeability, but pumping with SPAP drives the evolution of tighter membranes, facilitating colonisation of less alkaline environments. (**E**) Impermeable membranes funnel ion flow through bioenergetic proteins, independent of natural gradients. (**F**) From bottom up, SPAP favours divergence, selection for active pumping and tighter membranes. Pumping and phospholipid membranes arose independently in archaea and bacteria.

2.5 Methods

2.5.1 General description of the model

Protocells were modelled as half-embedded in the alkaline fluid, with the other half exposed to the comparatively acidic ocean. This produced an inward proton gradient from the acidic side, sustained by the constant replenishment of alkaline fluids and ocean water, which could be exploited by membrane proteins for carbon and energy metabolism. Equation [1] describes the various ways in which protons could enter or leave the protocell at every time step: by simple diffusion across the membrane on either side, and through any of the membrane proteins, namely the ATPase, SPAP, pump, or Ech.

$$\mathrm{N_H} = \mathrm{N_{H(ocean)}} + \mathrm{N_{H(vent)}} + \mathrm{N_{H(ATPase)}} + \mathrm{N_{H(SPAP)}} + \mathrm{N_{H(pump)}} + \mathrm{N_{H(Ech)}} \qquad [1]$$

where N_H represents the total number of protons that enter or leave the protocell in a given time step, and each of the $N_{H(i)}$ represent the number of protons that enter through a given surface or protein in that time step. Positive N values imply an influx of protons, whereas a negative N implies protons are leaving the protocell. Time was modelled as a discrete succession of such time steps.

Total concentrations were calculated at every time step by neutralisation and equilibration to the dissociation constant of water. External concentrations were assumed to be constant, as a result of continuous hydrothermal flow and convection in the ocean. Analogous equations were used for all other ions ($OH^–$, K^+, Na^+, $Cl^–$). Table 3 describes the parameters chosen for the results presented in the text, unless otherwise stated above.

Table 3. Parameters in the model and references

Parameter	**Value**	**Comment**	**Reference**
Concentrations*	**[M]**		
H^+_{ocean}	10^{-7}	pH 7. May have been as low as pH 5	(Arndt and Nisbet 2012)
H^+_{vent}	10^{-10}	pH 10. Can be as high as pH 11 at present	(Arndt and Nisbet 2012)
Na^+	0.4	Could have been as high as 0.8 M	(Pinti 2005)
K^+	0.01	Could have been as high as 0.02 M	(Pinti 2005)
$Cl^–$	0.41	Chosen to balance out the concentrations of Na^+ and K^+	(Pinti 2005)
H_2	0.015	Can be as high as 0.02 M	(Proskurowski et al. 2006)
Permeabilities†	**[cm/s]**		
H^+	$10^{–3}$	Default value unless otherwise noted	(Deamer and Bramhall 1986)
$OH^–$	$10^{–3}$	Assumed equal to H^+	(Deamer and Bramhall 1986)
Na^+	$10^{–9}$	In general kept six orders of magnitude less permeable than H^+ throughout	(Deamer and Bramhall 1986)
K^+	$10^{–9}$	Assumed equally permeable to Na^+	(Deamer and Bramhall 1986; Deamer and Dworkin 2005)
$Cl^–$	$10^{–7}$	In general, two orders of magnitude more permeable than Na^+	(Nichols and Deamer 1980)

Turnover rates	**[s^{-1}]**		
ATPase	270	Parameterised from mitochondria	(Etzold et al. 1997; Yoshida et al. 2001)
Ech	700	Parameterised from a soluble NiFe hydrogenase	(Liebgott et al. 2010)
SPAP	1500	Parameterised from *E. coli*'s NhaA SPAP	(Hunte et al. 2005)
Pump	200	Parameterised from mitochondrial Complex I	(Vinogradov 1998)

Surface areas	**[m^2]**		
ATPase F_o subunit	$4{\cdot}10^{-17}$	Estimated from PDB entry: 1C17	(Stock et al. 1999; Yoshida et al. 2001)
Ech	$3{\cdot}10^{-17}$	Relevant subunits of Complex I, estimated from PDB:4HEA	(Baradaran et al. 2013; Marreiros et al. 2013)
SPAP	$1.5{\cdot}10^{-17}$	Estimated from PDB:1ZCD	(Hunte et al. 2005)
Pump	$3{\cdot}10^{-17}$	Assumed to be similar to Ech	
Others			
H^+ per ATP	3.33	This many H^+ enter the ATPase in the synthesis of 1 ATP	(Ferguson 2010; Nicholls and Ferguson 2013)
Protocell diameter	1 μm	Small diameter of *E. coli*	(Moran et al. 2010)
Temperature	298.15 K	Standard temperature	
Embedment	50%	Protocell is exactly half-embedded in the alkaline side	

[*] Excluding H^+ and OH^-, all concentrations were assumed equal in the alkaline and acidic sides.

[†] In all simulations Na^+ and Cl^- permeabilities were kept respectively six and four orders of magnitude lower than the permeability of H^+.

It is reasonable to suppose that enzymes would not have reached their current reaction-rate values at the early stages of evolution considered here, so for the results presented above I consistently used 10% of the current turnover rates referenced in Table 3. A series of results using modern (100%) turnover rates are presented in Figure 19 for comparison.

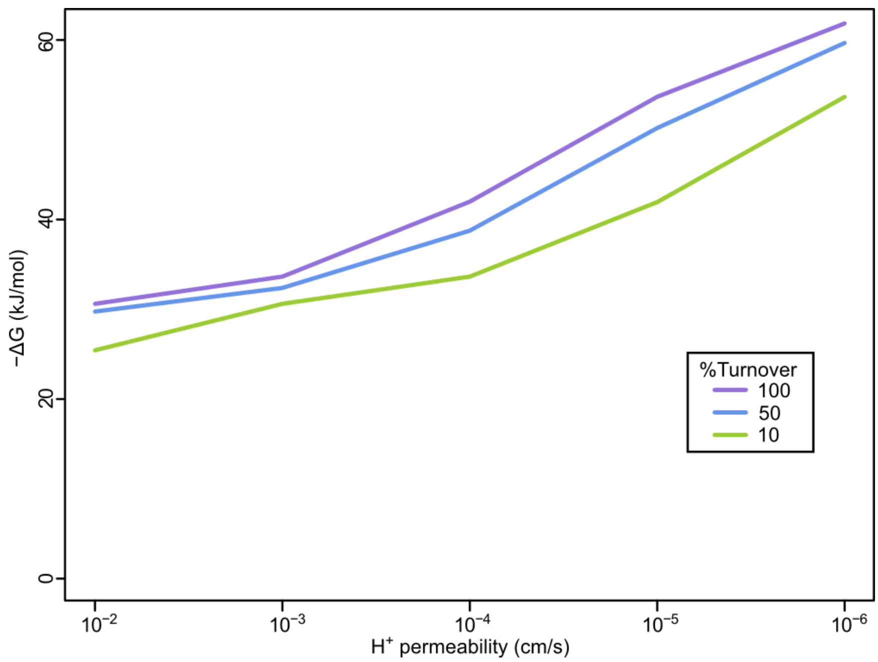


Figure 19. Modern (100%) turnover rates are comparable to the 10% rates used elsewhere

Membrane proteins were assumed to have lower turnover rates than in current cells, so their parameterised turnover rates were kept at 10% of modern values (see Table 3). The figure shows that with ATPase, SPAP and pump operating at modern speeds, the behaviour is similar to that operating at 10%. Rates of 50% give intermediate results. Parameters: 5% pump, 1% ATPase, 1% SPAP, pH gradient 7:10.

Flux through the membrane

Membrane flux J_S of a neutral substance S was modelled using a traditional passive diffusion equation (Lodish et al. 2000)

$$J_S = P_S A([S]_{ext} - [S]_{int}) \qquad [2]$$

where P_S is the permeability of the substance, A is the area of the membrane and $[S]_{ext/int}$ are the external/internal concentrations. To account for the effect of membrane potential Δψ on the behaviour of charged particles, ion diffusion was modelled using the Goldman-Hodgkin-Katz flux equation (Goldman 1943; Hodgkin and Katz 1949)

$$J_S = P_S z_S^2 \frac{\Delta\psi\, F}{RT} \frac{[S]_{int} - [S]_{ext}\, e^{-\frac{z_S\, \Delta\psi\, F}{RT}}}{1 - e^{-\frac{z_S\, \Delta\psi\, F}{RT}}} \qquad [3]$$

where z_s is the charge of the substance, F and R are the Faraday and gas constants, respectively, and T is the temperature. Electrical membrane potential $\Delta\psi$ was in turn modelled using the Goldman-Hodgkin-Katz voltage equation (Goldman 1943; Hodgkin and Katz 1949)

$$\Delta\psi = \frac{RT}{F} \ln\left(\frac{\sum P_{cation}[cation]_{ext} + \sum P_{anion}[anion]_{int}}{\sum P_{cation}[cation]_{int} + \sum P_{anion}[anion]_{ext}}\right) \qquad [4]$$

for the permeability and concentration of each ion present.

Internal protons and hydroxide were equilibrated using the dissociation constant of water ($K_w = 10^{-14}$).

2.5.2 Free energy (–ΔG) calculations

The available free energy –ΔG from the H^+ gradient was modelled with the equations used by Mitchell (1961)

$$\Delta G_{H^+} = -F\,\Delta\psi + RT \ln\left(\frac{[H^+]_{int}}{[H^+]_{ext}}\right) \qquad [5]$$

An analogous equation was used for the Na^+ gradient. In this way, the natural H^+ gradient serves the role that redox potential serves in modern chemiosmotic cells, e.g. in oxygen-respiring mitochondria. Namely, in mitochondria the electrochemical gradient is maintained by the oxidation of reduced substrates (Figure 6). Conversely, the geochemical disequilibrium between the volcanic acidic ocean and the serpentinising Earth crust sustained the imbalance in the alkaline hydrothermal vent.

The power of ATP to catalyse biochemical reactions in the cell comes not specifically from hydrolysis of the molecule itself, but from the degree to which the ATP/ADP ratio is shifted from thermodynamic equilibrium (Nicholls and Ferguson 2013); that is, the energy available from ATP hydrolysis varies with the ATP/ADP ratio. The equilibrium constant and consequently the energy required for ATP synthesis depends on the concentrations of ADP, phosphate, and magnesium ion, as well as pH (Mitchell 1961; Nicholls and Ferguson 2013), but with the exception of pH these values are unknown for the systems modelled, as are the rates of ATP hydrolysis. I have therefore used equation [5] to calculate the size of the electrochemical gradient (–ΔG) as a function of the H^+ and Na^+ gradients and the electrical membrane potential

($\Delta\psi$). The steady-state $-\Delta G$ in turn gives an indication of how far from equilibrium the ATP/ADP ratio could be pushed. With 3-4 protons translocated per ATP, a steady-state $-\Delta G$ of 20 kJ/mol is large enough to drive the ATP/ADP ratio to a disequilibrium of 10 orders of magnitude, equivalent to that found in modern cells (Nicholls and Ferguson 2013).

Steady-state $-\Delta G$ was calculated as a function of the size of the H^+ and Na^+ gradients and the electrical membrane potential ($\Delta\psi$) between the ocean and the inside of the cell. These factors in turn depend on steady-state rates of proton flux into and out of the cell via the lipid phase of the membrane (specified by its H^+ and Na^+ permeability and surface area) and through the ATPase. I calculated the maximum flux of H^+ or Na^+ through the ATPase based on the maximum possible number of ions translocated per second. Maximum ion flux is based on the reported maximum turnover rate of ATPase (Table 3), i.e. the maximum number of ATP molecules that each ATPase unit can synthesise in one second when operating at top speed, multiplied by 3.3, the number of H^+ or Na^+ required to synthesise 1 ATP (Table 3). This number was then multiplied by the number of ATPase units in the system, estimated from the membrane surface area assigned to this protein in each simulation (e.g. 1%, 5%, etc.) and the reported surface area of the membrane-integral F_O subunit (Table 3).

I further assumed that the actual flux rate of H^+ and Na^+ through the ATPase would also depend on the driving force itself, $-\Delta G$, i.e. the size of the H^+/Na^+ gradient and the electrical membrane potential ($\Delta\psi$). The ATPase was assumed to obey hyperbolic Michaelis-Menten dynamics, commonly the case in enzyme kinetics (Alberts et al. 2007) and reported for the ATPase (Hammes and Hilborn 1971), such that H^+/Na^+ flux asymptotically approaches the maximum turnover rate when the driving force is large, again assuming that flux rate is unconstrained by ADP availability. Increasing $-\Delta G$ beyond a threshold cannot increase H^+/Na^+ flux beyond the maximum turnover rate, so flux rate must saturate. The hyperbolic curve was modelled to reach saturation slightly beyond 20 kJ/mol, a gradient large enough to drive the ATP/ADP ratio to 10 orders of magnitude disequilibrium in modern cells (Nicholls and Ferguson 2013) and equivalent to a membrane potential of around 200 mV, close to a maximum for modern lipid membranes, given the low capacitance of thin lipid membranes. This number, between zero and one, was finally multiplied by the maximum flux of H^+ or Na^+, described above, to determine the influx of each of the two ions through the ATPase. When added to H^+/Na^+ flux rates across the lipid phase, the steady-state H^+/Na^+ flux through the ATPase gave a steady-state $-\Delta G$ available to drive ATP synthesis.

Full promiscuity of the ATPase to Na^+ and H^+ was assumed, with preference of one ion over the other depending solely on their respective gradient sizes. The energy-converting hydrogenase (Ech) was modelled similarly.

2.5.3 Modelling the sodium-proton antiporter (SPAP) and pump

SPAP was modelled to respond to the H^+ and Na^+ gradients, exchanging ions in the direction determined by the larger of the two gradients. $\Delta\psi$ was assumed to affect SPAP speed but not direction (Bassilana et al. 1984). Since the H^+ gradient is reversed on the alkaline side, I assumed that gene expression controls allowed the ATPase, SPAP, and Ech to operate only on the acidic side.

The pump was modelled as a generic system able to extrude either H^+ or Na^+, dependent on the concentration of hydrogen gas (H_2), and responding to the difference in concentrations of the respective ion, thereby making it easier to pump protons against a comparatively alkaline fluid, and more difficult against a comparatively acidic fluid.

2.5.4 Source Code

A running example of the code can be downloaded and ran locally from

http://github.com/UCL/membranedivergence

This code can be run directly from any typical computer with a regular web browser (e.g. Chrome, Firefox, Safari, or Internet Explorer).

2.5.5 BLAST searches

The primary amino acid sequence of the *M. jannaschii* Mj1275 Na^+/H^+ antiporter (SPAP) was obtained from the NCBI protein sequence database. Mj1275 is one of three known SPAP genes in archaeon *M. jannaschii* (Hellmer et al. 2002), the other two being Mj0057 and Mj1521. The first belongs to the NapA family, while the latter two are in the NhaP family. Phylogenetic analysis was performed on these three genes as well as the two common *E. coli* SPAP genes, NhaA and NhaB (Taglicht et al. 1991; Taglicht et al. 1993), using the NCBI-BLASTp server (Altschul et al. 1990) with standard parameters, filtering for each prokaryotic phylum (considering each of the proteobacteria as a separate clade). Results for Mj1275 showed the highest hit rate, possibly hinting that it is closest to the ancestral form of the SPAP. However, this is only a preliminary result that ignores the possibility of horizontal gene transfers, and it will require more detailed analysis in the future.

2.6 Appendix: The evolution of haem synthesis

Haem (or heme) is a crucial prosthetic group of many respiratory and photosynthetic proteins across all three domains, so as an initial step in the study of the early evolution of pumping as part of this work I analysed the distribution of the enzymes involved in the synthesis of this porphyrin group. I used the Microbial Genome Database for Comparative Analysis (MBGD) (Uchiyama 2003; Uchiyama 2007; Uchiyama et al. 2010) which allows the simultaneous comparison of multiple species for the presence or absence of genes.

Storbeck et al. (2010) suggested that the pathway is shared by archaea and bacteria up to the synthesis of uroporphyrinogen (urogen) III, a molecule that is itself involved in the synthesis of chlorophylls, cobalamin (vitamin B_{12}), sirohaem, haem, haem d_1, and coenzyme F430. The synthesis of urogen-III is catalysed by the enzyme uroporphyrinogen III synthase (UROS), a product of the gene *hemD*. From there, many archaea, as well as sulphate-reducing bacteria such as *Desulfovibrio vulgaris*, have a separate synthetic pathway. I analysed Storbeck et al.'s results and expanded them to include a large number of archaea. Table 4 presents the results of this analysis.

Table 4. Presence or absence of traditional (bacterial and eukaryotic) haem-synthesis genes in archaea

The first three columns correspond to bacterial species, used for comparison. *E. coli* and *P. aeruginosa* represent the "classical" pathway in most bacteria (and eukaryotes), while *D. vulgaris* exemplifies the sulphate-reducing bacteria, which lack the post-urogen genes. The remaining 97 columns, separated by a leading empty column, are archaea. Please note that the table has been divided into three consecutive blocks for reasons of space. Green: gene is present, Red: absent.

gene	*Escherichia coli*	*Pseudomonas aeruginosa*	*Desulfovibrio vulgaris*		*Acidilobus saccharovorans*	*Aeropyrum pernix*	*Desulfurococcus kamchatkensis*	*Desulfurococcus mucosus*	*Ignicoccus hospitalis*	*Ignisphaera aggregans*	*Staphylothermus hellenicus*	*Staphylothermus marinus*	*Thermosphaera aggregans*	*Hyperthermus butylicus*	*Pyrolobus fumarii*	*Acidianus hospitalis*	*Metallosphaera cuprina*	*Metallosphaera sedula*	*Sulfolobus acidocaldarius*	*Sulfolobus islandicus*	*Sulfolobus solfataricus*	*Sulfolobus tokodaii*	*Thermofilum pendens*	*Caldivirga maquilingensis*	*Pyrobaculum aerophilum*	*Pyrobaculum arsenaticum*	*Pyrobaculum calidifontis*	*Pyrobaculum islandicum*	*Thermoproteus neutrophilus*	*Thermoproteus tenax*	*Thermoproteus uzoniensis*	*Vulcanisaeta distributa*	*Vulcanisaeta moutnovskia*	*Archaeoglobus fulgidus*	*Archaeoglobus profundus*	*Archaeoglobus veneficus*
Pre-Urogen																																				
hemA	🟩	🟩	🟩		🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemL	🟩	🟩	🟩		🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemB	🟩	🟩	🟩		🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemC	🟩	🟩	🟩		🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemD	🟩	🟩	🟩		🟩	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟩	🟩	🟩
Post-Urogen																																				
hemE	🟩	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemF	🟩	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemN	🟥	🟥	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemFNX*	🟥	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥
hemGX	🟥	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemY	🟩	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemH	🟩	🟩	🟥		🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

gene	*Ferroglobus placidus*	*Halalkalicoccus jeotgali*	*Haloarcula hispanica*	*Haloarcula marismortui*	*Halobacterium sp.*	*Haloferax volcanii*	*Halogeometricum borinquense*	*Halomicrobium mukohataei*	*Halopiger xanaduensis*	*Haloquadratum walsbyi*	*Halorhabdus utahensis*	*Halorubrum lacusprofundi*	*Haloterrigena turkmenica*	*Natrialba magadii*	*Natronomonas pharaonis*	*Methanobacterium sp.*	*Methanobrevibacter ruminantium*	*Methanobrevibacter smithii*	*Methanosphaera stadtmanae*	*Methanothermobacter marburgensis*	*Methanothermobacter thermautotrophicus*	*Methanothermus fervidus*	*Methanocaldococcus fervens*	*Methanocaldococcus infernus*	*Methanococcus jannaschii*	*Methanocaldococcus vulcanius*	*Methanotorris igneus*	*Methanococcus aeolicus*	*Methanococcus maripaludis*	*Methanococcus vannielii*	*Methanococcus voltae*	*Methanothermococcus okinawensis*	*Methanocella paludicola*	*Methanocorpusculum labreanum*	*Methanoculleus marisnigri*	*Methanoplanus petrolearius*
Pre-Urogen																																				
hemA	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemL	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemB	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemC	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
hemD	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
Post-Urogen																																				
hemE	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥
hemF	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemN	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemFNX*	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemGX	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemY	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
hemH	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥

gene	*Methanospirillum hungatei*	*Methanoregula boonei*	*Methanosphaerula palustris*	*Methanosaeta concilii*	*Methanosaeta thermophila*	*Methanococcoides burtonii*	*Methanohalobium evestigatum*	*Methanohalophilus mahii*	*Methanosalsum zhilinae*	*Methanosarcina acetivorans*	*Methanosarcina barkeri*	*Methanosarcina mazei*	*Methanopyrus kandleri*	*Pyrococcus abyssi*	*Pyrococcus furiosus*	*Pyrococcus horikoshii*	*Pyrococcus yayanosii*	*Thermococcus barophilus*	*Thermococcus gammatolerans*	*Thermococcus onnurineus*	*Thermococcus sibiricus*	*Picrophilus torridus*	*Thermoplasma acidophilum*	*Thermoplasma volcanium*	*Korarchaeum cryptofilum*	*Nanoarchaeum equitans*	*Cenarchaeum symbiosum*	*Nitrosopumilus maritimus*	**% Archaeal species with gene**
Pre-Urogen																													
hemA	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	82
hemL	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	82
hemB	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	82
hemC	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	82
hemD	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	80
Post-Urogen																													
hemE	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	16
hemF	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	0
hemN	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	0
hemFNX*	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟩	🟩	7
hemGX	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	4
hemY	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	0
hemH	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	7

The table shows a clear divide in the pre-urogen and post-urogen enzymes between archaea and traditional bacteria, confirming the results of Storbeck et al. (2010). A few cases, most trivially attributable to horizontal gene transfers, show the potential synthesis of haem in some archaea via the “classical” bacterial pathway, but the post-urogen genes are largely missing otherwise. Predictably, the pre-urogen genes were present in LUCA, and had a role there unrelated to haem synthesis (e.g. cobalamin synthesis); consistent with the prediction of the model presented in this chapter, archaea and bacteria developed haem-associated pumps independently.

The ultimate goal of this project was to research the predicted independent origin of pumping in archaea and bacteria, specifically the origin of quinones and cytochromes; however, a project similar to this had been started earlier and in parallel in another lab, and was published in the meantime (Sousa et al. 2013). Indeed, Sousa et al. (2013) report that quinones show a similar pattern (in spite of a number of horizontal gene transfers), and conclude that pumping via electron transport chains with quinones and cytochromes, although admittedly ancient, evolved after the divergence of methanogens and acetogens, which they see as the earliest ancestors of all modern day archaea and bacteria, respectively.

3 THE EVOLUTION OF HOMOCHIRALITY

Note: large fragments of this chapter are adapted directly from the original research article "*On the biogenic origins of homochirality"*, published in the journal *Origins of Life and Evolution of Biospheres (OLEB)* in March 2015 (Sojo 2015).

3.1 Summary

Homochirality, the single-handedness of optically asymmetric chemical structures, is present in all major biological macromolecules. Terrestrial life's preference for one isomer over its mirror image in D-sugars and L-amino acids has both fascinated and puzzled biochemists for well over a century. But the contrasting case of the equally fundamental phospholipids has received less attention. Although the phospholipid glycerol backbones of archaea and bacteria are both exclusively homochiral, the stereochemistries between the two domains are opposite. Here I argue that the reason for this "dual homochirality" was a simple evolutionary choice at the independent origin of the two synthesising enzymes. More broadly, this points to a trivial biogenic cause for the evolution of homochirality: the enzymatic processes that produce chiral biomolecules are stereospecific in nature. Once an orientation has been favoured, shifting to the opposite is both difficult and unnecessary. Homochirality is the simplest and most parsimonious evolutionary case.

3.2 Introduction

3.2.1 Chirality and homochirality

The existence of polarity in molecules has puzzled biochemists from the very origins of the discipline. The word "chirality" itself comes from the Greek *kheir* for "hand" (Oxford Dictionaries 2015b), and thus it describes the "handedness" of structures, not only in chemistry but in the universe in general, all the way from the electro-magnetic spin of sub-atomic particles, to the helices of DNA, through the shells of snails, to the spirals of galaxies. Of interest here is chirality in organic molecules, first observed by Pasteur almost 170 years ago in his studies of tartrate (1848).

In the vast majority of cases, each carbon atom can bond in one of three types of patterns to other atoms in a molecule: *sp*^{3}, *sp*^{2}, or *sp*, summarised visually in Figure 20.

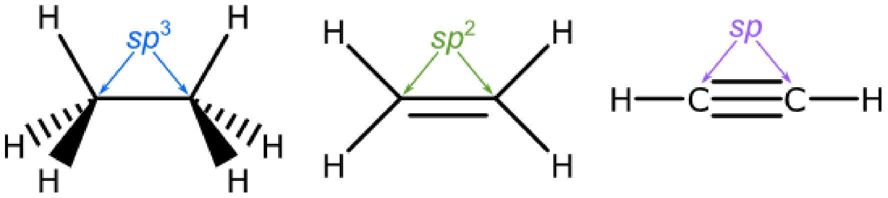


Figure 20. *sp*^{3}, *sp*^{2}, and *sp* symmetries in carbon atoms

In molecular orbital theory, carbon atoms usually occur in one of three bonding patterns in most of their molecules. *sp*^{3} carbons are bound to four separate atoms, each time with a single bond; *sp*^{2} carbons also have four bonds, but two of these are to the same atom, another carbon in the case of this figure; finally, *sp* carbons have a single bond to one atom, and a triple bond to another, again giving a total of four bonds.

All three bonding patterns are relevant in biology, and many biochemical reactions involve the conversion of one type into another by oxidations and reductions, additions and eliminations. The structures of both *sp*^{3} and *sp*^{2} carbon atoms imply that there is more than one way to organise the substituents if they are different. In particular, a molecule with a single *sp*^{3} carbon atom that has four different substituents can be arranged in two different manners, as shown in Figure 21.

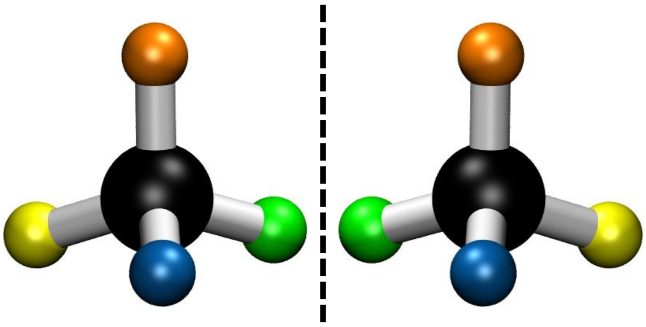


Figure 21. Chirality in carbon atoms with sp^3 (tetrahedral) symmetry

Although the four substituents are the same in both molecules, the distribution is different. The two molecules are mirror images of each other, and cannot be super-imposed.

Carbons with *sp*^{2} symmetry are not themselves chiral (although they can be asymmetric), but an addition reaction that produces an *sp*^{3} symmetry can give rise to a chiral molecule if the four final substituents are different. That is, molecules with an asymmetric *sp*^{2} carbon are pro-chiral, as shown in Figure 22.

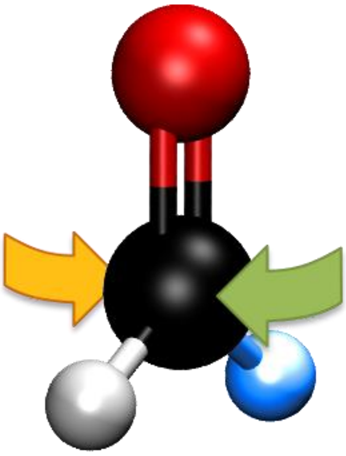


Figure 22. A prochiral molecule is not itself chiral, but can form chiral products upon reaction

The mechanisms of certain organic reactions mean that pro-chiral molecules can only produce a chiral product. A nucleophilic attack from the right (green arrow) would push the three substituents to the back (i.e. into the page, as the molecule is drawn), whereas attacking from behind (yellow arrow) would push the three substituents to the front (or out of the page).

3.2.2 Homochirality in sugars and amino acids

Chirality is ubiquitous in biology; macroscopically, it is most obviously present in structures such as snail shells (which either spiral one way or the other, normally in species-characteristic fashion), while at the macro-molecular level the most noticeable example are perhaps DNA double-helices and protein alpha-helices (both of which are typically right-handed). At the atomic level, the monomers are themselves chiral, with DNA composed exclusively of right-handed (D) sugars, while a similar situation is observed in proteins, composed of left-handed (L) amino acids (Figure 23). The mirrored structures, or *enantiomers*, do play roles in some organisms, but their biochemical relevance is minor (Krebs 1935; Corrigan 1969).

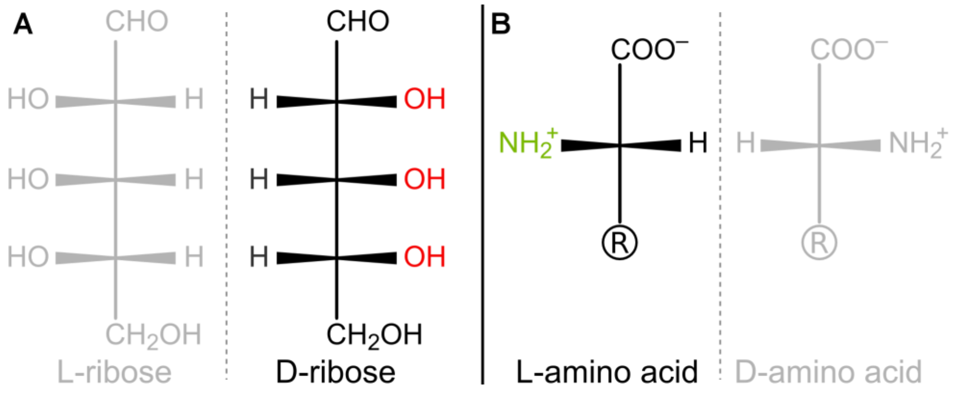


Figure 23. Homochirality in sugars and amino acids

(A) The backbone of DNA and RNA is formed exclusively from D-sugars. **(B)** Although D-amino acids occur sparsely in certain organisms, proteins of all domains of life are formed almost entirely of L-amino acids. ‘R’: amino acid side chain, e.g. ‘H’ for glycine or ‘-CH_2-SH’ for cysteine.

The observation of this absolute bias in life’s most fundamental molecules has led to several quests to detect a possible non-biological cause for these preferences in either prebiotic chemistry or physics. The assumption is that, since biology is already homochiral, then the orientation biases in modern life must have begun before biology itself, by asymmetric physical forces or chemical interactions operating on the monomers.

3.2.3 Homochirality in meteorites: physical and chemical causes for the evolution of homochirality

A most remarkable slight bias towards biological-type enantiomers reported in the Murchison and Murray meteorites (Engel and Macko 1997; Pizzarello and Cronin 2000) has fuelled the search for intrinsic physical causes behind the origin of terrestrial homochirality. Parity violations in radioactive β-decay from electroweak nuclear interactions (Mason 1984; Kondepudi and Nelson 1985), spontaneous autocatalytic symmetry breaking (Blackmond 2004; Kawasaki et al. 2006), adsorption onto chiral surfaces (Karagounis and Coumoulos 1938; Bonner et al. 1975), and asymmetric photochemical reactions caused by polarised light from supernovae in the interstellar medium (Jorissen and Cerf 2002) have all been put forward as plausible physical forces behind a potential pre-biotic origin of homochirality in D-sugars and L-amino acids. Much research in prebiotic chemistry seeks to explain this bias in chemical orientation, the notion being that the starting material for life must have been biased already, for one of the several reasons suggested above, or another that remains to be elucidated.

However, a case of *dual homochirality* is known in an equally fundamental group of biomolecules. The backbone of membrane phospholipids has opposite handedness in archaea and bacteria, the two basal domains of life.

3.2.4 The dual evolution of homochirality in lipids

As discussed in Chapter 2, the glycerol-phosphate backbone of phospholipids can come in two orientations and, intriguingly, *sn*-glycerol-1-phosphate (G1P) is exclusive to the archaea, while the enantiomer *sn*-glycerol-3-phosphate (G3P) is unique to bacteria (and eukaryotes by inheritance) (Figure 24). The synthesising enzymes, *sn*-glycerol-1-phosphate dehydrogenase (G1PDH) and *sn*-glycerol-3-phosphate dehydrogenase (G3PDH), are unrelated (Koga et al. 1998). I suggest that an explanation for this dichotomy may help elucidate some of the fundamental principles behind the origin and maintenance of homochirality.

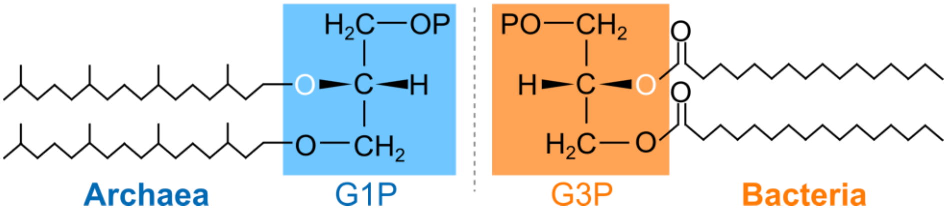


Figure 24. Dual homochirality in membrane phospholipids

The backbone moiety *sn*-glycerol-1-phosphate (G1P) is exclusive to the archaea, while its enantiomer *sn*-glycerol-3-phosphate (G3P) is exclusive to the bacteria (and eukaryotes by inheritance). No archaea have been observed with G3P, or vice versa. That is, both domains are exclusively homochiral, but the stereochemistries are inverted.

Note that, while possessing chimeric genomes of both archaeal and bacterial ancestry, no archaeal membranes have been observed in eukaryotes. However, it can be predicted that the plasma membrane of the first eukaryotic common ancestor must have been archaeal, an aspect I will return to in detail in Chapter 4.

In proteins, the advantage of a more stable secondary structure in the combination of twenty different amino acids may in itself account for the prevalence of one orientation over the other (Brack et al. 1979). Similarly, a higher stability of homochiral RNA has also been demonstrated (Urata et al. 2005). It is therefore possible that at the origin of the first catalytic and informational biopolymers, competition between homochiral and heterochiral molecules ensued, with the pure ones out-performing the hybrids. Still this does not explain why one enantiomer prevailed over the other, but the dual homochirality of phospholipid backbones suggests that the prebiotic bias hypothesis may be unnecessary.

A number of plausible scenarios for the *lipid divide* have been suggested. Most simply, it is possible that the last universal common ancestor (LUCA) was not cellular in the modern sense and had no genes for specifying either type of lipid; not only the glycerol-phosphate backbones but the specific enzymes required to synthesise all parts of the lipids evolved later, and independently, in archaea and bacteria (Martin and Russell 2003). However, the broad conservation of a number of membrane proteins, including the signal-recognition particle and the ATP synthase, would make a lipid-free scenario unlikely (Koonin and Martin 2005; Mulkidjanian et al. 2009). Early lipids may have been produced by abiotic means (Deamer et al. 2002; Martin and Russell 2003), and certain parts of the lipid-synthesising machinery may have already been present in the common ancestor (Peretó et al. 2004), but the absence of the full machinery for lipid synthesis in LUCA would account for the fundamental differences between archaeal and bacterial membranes, chiefly the opposing stereochemistries of G1P and G3P.

Another possibility is that both types of glycerol-phosphate backbones were present in LUCA, G1P being later favoured at the origin of archaea, and G3P at the origin of bacteria (Wächtershäuser 2003). The LUCA of this scenario had a heterochiral membrane, either racemic or not (Peretó et al. 2004). Heterochiral membranes were predicted to be unworkable (Wächtershäuser 2003), but experiments have shown that they are in fact viable (Shimada and Yamagishi 2011). In this hybrid scenario, it is likely that the ancestor of one of the two enzymes (G1PDH or G3PDH) evolved first. Since both G1P and G3P are well known to be viable and effective in a plethora of environments, it is difficult to see why a second enzyme would arise once the first one was in place, only to completely eradicate the other after the archaea-bacteria split.

Alternatively, a non-stereospecific ancestral enzyme existed first that produced a heterochiral mixture of the two glycerol-phosphate backbones. However, all known $NAD(P)^+$-dependent CH-OH dehydrogenases (E.C. number 1.1.1), are exclusively stereospecific in their hydrogen transfers (You et al. 1978; Benner 1982). Within this large supergroup to which both G1PDH and G3PDH belong, two classes exist. Class 1 dehydrogenases exclusively transfer the pro-*R* hydrogen of NAD(P)H, whereas Class 2 are stereospecific for the pro-*S* hydrogen (Figure 25). These redox reactions are intrinsically stereospecific both in their coenzymes and substrates (Fisher et al. 1953; Arnold et al. 1976).

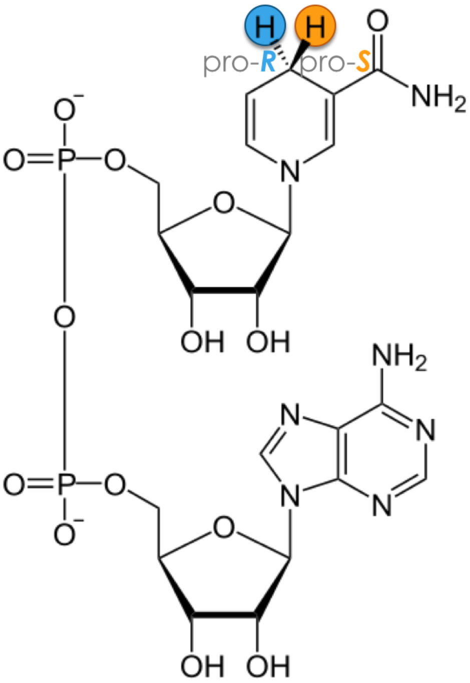


Figure 25. The pro-*R* and pro-*S* hydrogens of NADH

Although written NADH (with a single H), in fact the molecule has two available hydrogens, but only one of them is used in each reaction. All of the OH-dehydrogenase enzymes that use this ubiquitous biochemical reductant as substrate are stereospecific in their choices of only one of the two available hydrogens. Aptly, these hydrogens are termed pro-*R* and pro-*S*, with regards to the stereochemistry of the product they generate when hydrogenating a pro-chiral sp^2 carbon to sp^3.

The carbonyl centre of dihydroxyacetone phosphate (DHAP), from which both G1P and G3P are formed, is prochiral: hydrogenation from one side of the double bond produces G1P, while reacting from the opposite side gives G3P. At the atomic level, the amino acids of the active site of G3PDH face the pro-*S* hydrogen of NADH, whereas the G1PDH active site has recently been reported to exhibit a pro-*R* geometry (Koga et al. 2014). The idea of a non-stereospecific glycerol-phosphate synthase is difficult to reconcile with biochemical knowledge of the enzymes that catalyse these reactions.

A simpler explanation is that LUCA, although cellular in nature, had neither of the two enzymes, and so no glycerol-phosphate backbone (Sojo et al. 2014; Chapter 2). Early membranes were a mixture of more rudimentary amphiphiles, most simply fatty acids. This would have made such membranes leaky to ions and other small molecules, and indeed may have been a requirement for the early evolution of membrane bioenergetics and free-living cells, as discussed by Lane and Martin (2012), and in Chapter 2. In this scenario, G1PDH and G3PDH had independent origins after the divergence of archaea and bacteria.

In the evolution of the two novel enzymes, the stereochemistry of the respective ancestral proteins would be maintained (Hanson and Rose 1975). G1PDH was recruited from an ancestor of the alcohol-dehydrogenase/dehydroquinate-synthase/glycerol-dehydrogenase superfamily (Peretó et al. 2004). Like all extant members of this superfamily (You et al. 1978), this must have been a pro-*R* enzyme. Independently, G3PDH was derived from an ancestor of the UDP-glucose-6-dehydrogenase/3-hydroxyacyl-CoA-dehydrogenase superfamily (Peretó et al. 2004); analogously, like all members of this family (You et al. 1978), this would have been a pro-*S* enzyme (Figure 26).

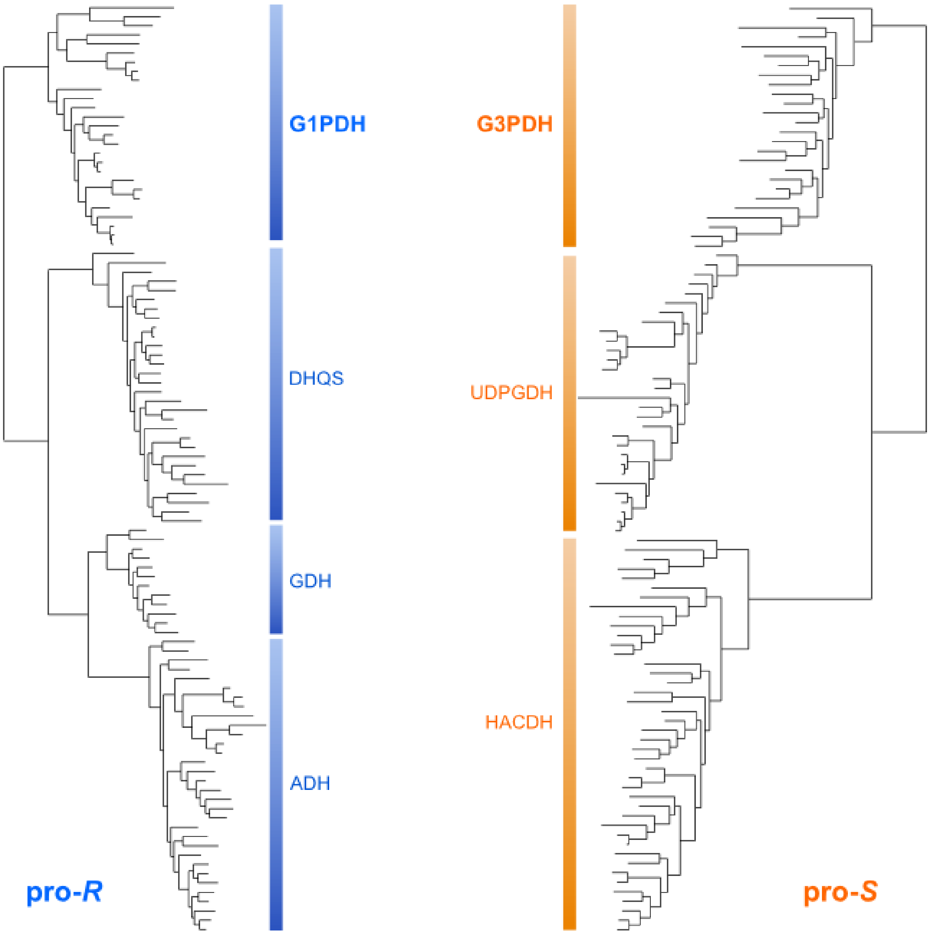


Figure 26. All proteins within the respective phylogenetic families of both G1PDH and G3PDH share the same stereospecificity

The two trees show seven families, all of which contain both archaeal and bacterial sequences. Sequences in one tree are completely unrelated to those in the other, and are shown together here only because they catalyse similar reactions (oxidation/reductions using NAD(P)$^+$/NAD(P)H as a substrate). All sequences in the G1PDH tree use the pro-*R* hydrogen of NAD(P)H, while all sequences in the G3PDH tree use the pro-*S* hydrogen. Abbreviations as follows. DHQS: dehydroquinate synthase. GDH: glycerol dehydrogenase (not to be confused with glycerol-phosphate dehydrogenase). ADH: alcohol dehydrogenase. UDPGDH: UDP-glucose 6-dehydrogenase. HACDH: 3-hydroxyacyl-CoA dehydrogenase. Trees adapted from Peretó et al. (2004; see original article for details of the sequences and species), and stereochemical classifications by You et al. (1978).

These two independent origins of DHAP reduction gave rise to the two opposing configurations of the glycerol-phosphate products, G1P and G3P. A non-stereospecific glycerol-phosphate synthase was unlikely. In fact, its postulation is unnecessary.

3.3 Homochirality as the simplest evolutionary scenario

This dual origin of single-handedness in fundamental biological molecules provides a crucial insight into the evolution of homochirality in general. Whether or not pre-biotic molecules were enriched in one enantiomer, life itself would naturally choose one catalytic orientation over the other (Martin and Russell 2003). This simply reflects the orientations that the ancestral enzymes had and their evolutionary availability for duplication, divergence, and neo-functionalisation. The question, if any, lies in why terrestrial life went in one specific direction towards L-amino acids and D-sugars, rather than the opposite. Subtly, homochirality of a given molecule is in itself biologically trivial, while the specific orientation may or may not be. Life would have chosen only one orientation either way.

The catalytic success of enzymes depends on their specific binding to substrates and cofactors, and these highly selective orientations largely account for the evolution of stereospecificity in enzymes (Hanson 1972). Certain structures, such as cyclic molecules, are intrinsically obliged to react stereospecifically (Hanson 1972), such that chiral exclusivity is in fact a general principle of biochemical catalysis. The independent evolution of DHAP reduction by G1PDH and G3PDH sheds light on the prevalence of one orientation over the other. If the ancestral enzyme had an R-favouring orientation, the duplicated enzyme would have inherited this preference (Hanson and Rose 1975), and once an orientation had been favoured, there would be no selective pressure to develop the opposite one. Such a process would not only be evolutionarily challenging but often impossible, and ecologically superfluous.

3.4 Implications for the origin of life

The choice of an orientation early in the evolution of a biochemical pathway imposes this preference on any subsequent reactions. Enantiospecificity thus becomes a “frozen accident” at the key steps in which chirality is introduced (or removed). Any arising enzymes that use the same product or its chiral derivatives would have to adapt to the chosen orientation.

It is tempting to draw analogies between early biochemistry and classic non-biological synthetic chemistry. However, free-solution chemistry is not directly comparable to enzymatic catalysis, largely because of the highly specific binding of substrates, water, and cofactors to enzymes (Hanson and Rose 1975).

At the atomic level, enantiomers look essentially identical when considering simple inorganic catalysts such as platinum or iron, of common use in synthetic chemistry. But from the point of view of biochemical catalysis, the lock-and-key fashion in which enzymes facilitate reactions (Fischer 1894) makes it apparent that enantiomers are two entirely different molecules when it comes to folding around them in the way that enzymes typically do (Figure 27).

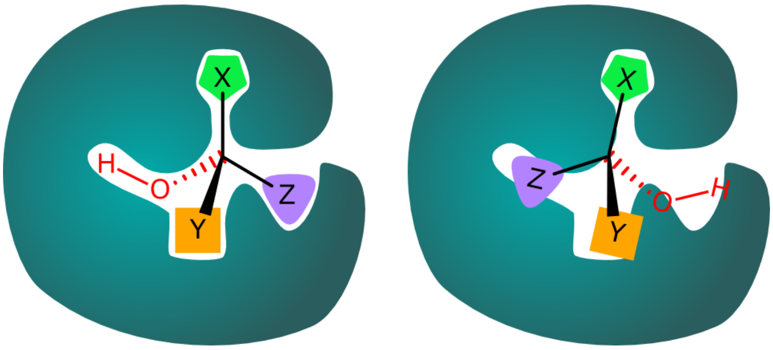


Figure 27. The enantiomer on the right is not a match to the active site of the hypothetical enzyme that catalyses the reaction of the enantiomer on the left

In terms of enzymatic catalysis, enantiomers are two completely different molecules at the local atomic level. Despite the apparent chemical similarity, the three-dimensional distribution of atoms in space means that enzymatic catalysis must be intrinsically stereospecific.

The overwhelming homochirality of terrestrial biochemistry seems to suggest that life could not have started in a racemic mixture (Cline 2005; Breslow 2011). However, recent findings of cross-chiral RNA-polymerising ribozymes (Sczepanski and Joyce 2014) allow for both D- and L- enantiomers potentially playing a key role in the origin of life. In a plausible early system in which simple amino acids or short polypeptides were chelated to metals or larger mineral structures and started catalysing reactions (Russell and Martin 2004), it is not challenging to envision a parsimonious explanation for a single orientation being favoured: the chirality of the amino acids themselves, or the particular arrangement of the primary sequence, would eventually but inevitably lead to stereospecific synthesis. Many reactions, indeed, might only occur stereospecifically or not at all (Hanson 1972), and it is far easier structurally for an efficient enzyme to catalyse a reaction stereospecifically than not, again a simple implication of the traditional lock-and-key principle of biochemistry (Fischer 1894).

An early D-ribozyme that naturally started selecting for L-amino acids would suffice as an explanation for L-homochirality in proteins (Martin and Russell 2003; Martin and Russell 2007). An independent origin of life that synthesised D-proteins may have occurred later, but it could not succeed since it would have had to compete against the earlier and by then more efficient mirror form. If an enantiomeric ribozyme arose within the same organism instead, it would have been useless, not only because the function was already covered by the original one, but crucially because any related enzymes would have already adapted to the chiral preference of the earlier one. That is, an emerging fully functional proto-ribosome would have inexorably imposed its chiral preference across life and outcompeted less efficient alternatives. On Earth, this successful proto-ribosome was itself D-homochiral and preferred L-amino acids. Any competing L-ribozymes and D-polypeptides were swiftly outcompeted as soon as the D-sugar/L-amino-acid association was established.

3.5 Discussion

In the face of widespread horizontal gene transfer between the two domains, it is remarkable that not a single case of the archaeal G1P in bacterial phospholipids, or vice versa, has ever been observed. This is in spite of both G1PDH and G3PDH having been observed across domains (Peretó et al. 2004), their role being catabolic rather than anabolic (Rawls et al. 2011). Both enantiomers are self-evidently viable, so no deeper explanation of this dual homochirality is necessary: independently derived enzymes catalyse stereochemically opposite reactions in archaea and bacteria, giving two molecules that perform the same job equally well. These enzymes, and all members of their evolutionary superfamilies, are stereospecific, and this is an intrinsic characteristic of many enzymatic reactions. More broadly, the very functional onset of a catalytic enzyme may require the selection of an orientation when it comes to chiral, pro-chiral, and asymmetric molecules in general.

Slight spontaneous enantioselections may have occurred prebiotically, potentially clearing a path for one orientation to prevail over the other. However, even in the presence of such a small advantage for one of the two mirror images, the development of a successful bio-catalyst that preferred the disfavoured one would still have sufficed to make it more prevalent, even if the other enantiomer was slightly more stable chemically. Homochirality would have arisen either way, as it is indeed the simplest solution in terms of both biochemistry and evolution. Having a heterochiral product is not only structurally disadvantageous, it is biochemically cumbersome, ecologically superfluous, and evolutionarily challenging. In many cases, it is actually impossible. A chiral choice early in the development of a pathway would inevitably impose the selected orientation on all subsequent reactions.

The explanation for the origin and maintenance of homochirality may lie simply in the structural and spatial differences between inorganic and biochemical catalysis.

4 THE EVOLUTION OF EUKARYOTIC MEMBRANES

Note: some of the contents of this chapter will be part of the original research article "*Why do eukaryotes have bacterial membranes?"*, created as first author, in collaboration with PhD supervisors Prof. Andrew Pomiankowski and Dr. Nick Lane, as well as with Prof. Christophe Dessimoz and his Master's student Jan Koch. Manuscript in preparation.

4.1 Summary

Recent phylogenetic evidence supports the modern endosymbiotic hypothesis for the origin of eukaryotes, by which a bacterium embedded within an archaeal host gave rise to the ancestral eukaryotic cell. If so, this original eukaryote must have had an archaeal plasma membrane and bacterial (proto)mitochondrial membranes; yet all known modern eukaryotes have exclusively bacterial membranes, both in their mitochondria and other organelles as in their boundary to the exterior. For some yet unknown reason the archaeal phospholipid synthesis machinery was lost and the bacterial one retained. The membranes of archaea and bacteria are significantly different, but, given that members of the two domains coexist in many environments, it is not clear why one type of phospholipid would have been favoured over the other in eukaryotes. In fact, genes for bacterial lipid biosynthesis had to be transferred from the (proto)mitochondrion into the archaeal genome that formed the basis of what now is the eukaryotic nucleus. Why wasn't the archaeal lipid machinery kept instead, given that it was already in the (proto)nucleus? I hypothesise that this was due to bioenergetic constraints: while mitochondria became specialised as the powerhouses of the eukaryotic cell, energy production came to rely increasingly on them. The physiological adaptation of bioenergetic mitochondrial proteins to their membrane meant that the bacterial phospholipids had to be kept; replacing them with archaeal analogues would have led to a loss of efficiency in energy conversion, potentially even deleterious leakage of reactive oxygen species and, more generally, decreased fitness. There should be evidence of similar effects in extant archaea and bacteria: membrane proteins should be less likely to be transferred horizontally across the prokaryotic domains, since they would have to sit on a foreign membrane, while water-soluble proteins are transferred between aqueous media and should have an easier adoption. In this chapter I provide evidence that supports this prediction.

4.2 Introduction

Eukaryotes constitute all complex life on Earth, from animals to plants to fungi, as well as multiple unicellular lineages. Several details of their origin are amongst the main unresolved questions in evolutionary biology. Eukaryotes are not only complex macroscopically. At the cellular level, unicellular eukaryotes are as complex as multicellular ones, and all types are structurally more complex than any members of either of the two prokaryotic domains.

Eukaryotic cells contain a number of domain-specific traits, including a nucleus, actin/tubulin cytoskeleton with associated kinesin and dynein motors, eukaryote-specific flagella, mitosis and meiosis, sexual reproduction, mitochondria, Golgi apparatus, endoplasmic reticulum, endocytosis and phagocytosis, bacterial-type membrane lipids, (largely) bacterial-type metabolism, and (largely) archaeal-type information processing (Hartman and Fedorov 2002; Koumandou et al. 2013; Doolittle 2014; McInerney et al. 2015). Since these traits are common to the whole diversity of lineages in the domain, all of these are thought to have been present by the advent of the last eukaryotic common ancestor (LECA).

4.2.1 Hypotheses for the evolution of eukaryotes

There are two main types of hypotheses for the evolution of eukaryotes: gradual or endosymbiotic. In the gradual hypotheses, a classical Darwinist process of accumulation of favourable mutations over time led to the ancestor of eukaryotes, including the slow but steady development of a nucleus, cytoskeleton and other organelles, and the invention of phagocytosis, in turn leading to the acquisition of mitochondria (and later plastids). In the endosymbiotic scenarios, a “big-bang” style process was put in motion by the association between a host archaeon and the bacterium that would become the mitochondrion (as well as other endosymbionts in Margulis’ serial endosymbiosis theory (2004)). This association of ancestors, which were not much different from any other two prokaryotes, led to a unique type of consortium that allowed, for the first and so far only time in Earth’s history, the high complexity of the eukaryotic cell, ultimately leading to complexity at the multicellular level as well.

Simple calculations show that eukaryote-type complexity is impossible for prokaryotes, regardless of size or membrane organisation, due to bioenergetic constraints (Lane and Martin 2010). Specifically, the limited energy available for gene expression does not allow archaea or bacteria to accumulate and particularly express genomes of eukaryotic complexity. In eukaryotes, the compartmentalisation and specialisation of bioenergetic conversion into mitochondria and their membranes provides hundreds of thousands of times more available energy per gene, which translates into the potential for greater complexity that eukaryotes have capitalised so notably. Endosymbiotic theories therefore provide a more credible scenario, yet there are many flavours of endosymbiosis.

The origins of the theory can be traced back to a Russian botanist by the name of Constantin Mereschkowsky (1905; translated by Martin and Kowallik 1999), whose main focus was actually on chloroplasts, which he suggested had been derived from an ancient symbiotic association of cyanobacteria with a host that had no plastids.

This theory was elaborated in the 1960s by Lynn Margulis[vi], who suggested that mitochondria had been derived by a similar endosymbiotic association (Sagan 1967). Since mitochondria are the hub of oxidative phosphorylation in the eukaryotic cell, Margulis reasonably assumed that the driver of the initial association had been the proto-mitochondrion's ability to use oxygen as a final electron acceptor. However, the existence of anaerobic mitochondria and hydrogenosomes makes this assumption unlikely.

There are many alternatives to the classical endosymbiotic theory, not all endosymbiotic themselves, most of which have been thoroughly and critically reviewed in several recent articles (Doolittle and Mariscal 2015; López-García and Moreira 2015; Martin et al. 2015; McInerney et al. 2015), but in general the best supported ones match the Lake and Rivera model of the ring of life (2004). Bill Martin and Miklós Müller proposed the "hydrogen hypothesis" (1998) which, in brief, states that the eukaryotic cell arose from a symbiotic association of a H_2-dependent archaeon with a H_2-producing bacterium that would go on to become the mitochondrion. Importantly, the host was not a eukaryote, it was just as prokaryotic as the endosymbiont, and eukaryotes only existed after and directly because of the association between the two prokaryotes. This hypothesis neatly accounts for

[vi] Known at the time as Lynn Sagan.

hydrogenosomes and anaerobic mitochondria, and it has stood the test of time well. But, while the hydrogen hypothesis remains solid after almost two decades (a very long time in this volatile field), the exact natures of both the host and the symbiont remain uncertain. Whatever the case, three facts seem clear, if still not universally accepted in spite of overwhelming data (McInerney et al. 2015): (1)the host was an archaeon, (2)the endosymbiont was a bacterium, and (3)the eukaryotic cell arose only *after* and *because of* the association between the host and the endosymbiont (i.e. neither the host nor the symbiont had eukaryotic complexity, or many of the uniquely eukaryotic traits).

4.2.2 Why keep a mitochondrial genome?

Although the prokaryotic ancestors of eukaryotes must both have had a full-sized prokaryotic genome at the early stages of their association, modern eukaryotes have only a very reduced mitochondrial genome, with hundreds or thousands of genes of bacterial ancestry now in the nucleus following extensive mito-nuclear transfers over the millions of years since FECA. Humans, for example, have only 37 genes in their mitochondrial genome, 22 of which encode transfer RNAs (tRNAs), 2 ribosomal RNAs (rRNAs), and 13 proteins, all bioenergetic (Iborra et al. 2004); other eukaryotes typically have similarly small genomes. But, given that eukaryotic ancestors transferred such vast amounts of genetic material to the nucleus, why not transfer all of it?

There are several explanations for the permanence of a genome in mitochondria and chloroplasts, carefully reviewed by Allen (2003). Most simply, it is possible that either there is no particular reason for the present arrangement (this is simply how it is and the distribution is due to chance), or that the organelle-nucleus transfer is an ongoing process, that the handovers are not yet complete and that the current mitochondrial (and chloroplast) genome is merely a vestige that will continue to be reduced and repositioned in a relatively short evolutionary time (since only very little is left). However, given the long divergence times between the several eukaryotic lineages, these non-selective explanations would predict the distribution of genes to be random, yet the same bioenergetic proteins (a “conserved core”) appear over and over again in the organellar genomes of different lineages.

Moving on to selective explanations, it is possible that the evolution of an unrelated cellular mechanism may have "frozen" the transfer state of some genes. In one example of this scenario, the origin of exocytosis and protein secretion made it impossible for certain genes to be transferred to the nucleus, because of N-terminal targeting sequences that would direct them to the export pathway instead of the mitochondria (von Heijne 1986a). However, the high specificity of protein targeting mechanisms from the nucleus to mitochondria (Rehling et al. 2004) makes this explanation unlikely. A somewhat similar hypothesis suggests that certain proteins need to be assembled directly in the sub-cellular location where they exert their function, because otherwise they may end up performing it in the wrong place, a particularly important problem for membrane proteins; yet this ultimately suffers from the same drawbacks, namely that the mitochondrial importing machinery is now known to be so remarkably sophisticated that it can be expected to cope with such limitations with ease.

Another factor that could hinder some specific transfers is the slight difference between the mitochondrial and nuclear genetic codes. A number of codons have different meanings for the two translation machineries, some times drastically so; for example, the triplet UGA works as a stop codon for cytosolic ribosomes, but it indicates tryptophan within Opisthokont mitochondria (Alberts et al. 2007). Therefore, certain genes could not be transferred to the nucleus since their translation with cytosolic ribosomes would be ineffective and potentially deleterious. However, if this were due to an ancestral alternative code in the original bacterium before endosymbiosis, then only few of the genes should have been transferred, yet most were.

Another selective explanation for why this particular set are remarkably difficult to transfer, is that the physical properties of some proteins mean that they have to be expressed in the compartments where they are active. Most notably, the solubility of all the mitochondrially-encoded proteins is very low (they are all bioenergetic membrane proteins), such that importing them into the mitochondrion would be too difficult (von Heijne 1986a). However, there are several counter-examples to this prediction (Allen 2003), including the large subunit of RUBISCO in chloroplasts, which is in general organelle-encoded but water-soluble, and conversely the nuclearly-encoded but hydrophobic subunits of the light-harvesting complexes (LHC) I and II.

There is an alternative hypothesis that provides a strong selective explanation for organelle genomes. The "**C**o-location **o**f **R**edox **R**egulation for gene expression", or "CoRR" hypothesis (Allen 1993; Allen 2003; Allen 2015) suggests that the reason mitochondria and plastids kept the specific set of proteins in their reduced genomes is that the bioenergetic state of the organelle needs to be controlled locally and swiftly; waiting for a cytosolic response would be inefficient and potentially ineffective, since the nucleus has no direct information about the energetic state of each particular organelle. Instead, the expression of a number of key bioenergetic genes can be controlled by the redox state of the corresponding gene products within the organelle itself. This hypothesis implies that the preservation of bioenergetic function requires redox control from within the same compartment. CoRR accounts for the predictions of other hypotheses, and it is also consistent with multiple experimental evidence (Allen 2015).

4.2.3 From FECA to LECA

At the early stages of the endosymbiotic association, the first eukaryotic common ancestor (FECA) would have had two sets of genes for performing most fundamental cellular tasks, including DNA replication, transcription, translation, cellular membranes, and metabolism. Several of these duplicated tasks still remain in modern eukaryotes: DNA replication, transcription, the proton-powered ATPase, tRNAs, and ribosomes within mitochondria are different from their cytosolic equivalents, even though most genes whose products are active in mitochondria are actually nuclearly encoded, and their ancestry is unequivocally bacterial. Yet other redundant functions were streamlined. Eukaryotic metabolism is almost entirely bacterial (with some uniquely eukaryotic traits), and notably, so are membranes.

It seems certain that by the time the major extant eukaryotic lineages diverged, their common ancestor (LECA) had already developed most of the typically eukaryotic features, including the nucleus (with its envelope, nucleoli, and pore complexes), an actin/tubulin cytoskeleton with associated kinesin and dynein motors, flagellum, mitosis and meiosis, sexual reproduction, mitochondria, Golgi apparatus, endoplasmic reticulum, endocytosis and phagocytosis, (largely) bacterial-type metabolism, (largely) archaeal-type information processing (except in the mitochondrion), and bacterial-type membrane lipids (Hartman and Fedorov 2002; Koumandou et al. 2013; Doolittle 2014; McInerney et al. 2015).

No archaea have been observed with bacterial membranes, and similarly no bacteria with fully archaeal membranes are known (Chapter 2). So, if the eukaryotic cell arose from the endosymbiosis of a *bona fide* bacterium into a *bona fide* archaeon, the first eukaryotic common ancestor (FECA) must have had an archaeal plasma membrane and bacterial (proto)mitochondrial membranes. Yet all extant eukaryotes have exclusively bacterial phospholipids in all of their encapsulated cellular structures, including the nuclear envelope, mitochondria, vesicles, peroxisomes, lysosomes, and the plasma membrane. The membranes of the last eukaryotic common ancestor (LECA) therefore must have been exclusively bacterial. Somehow along the evolutionary path from FECA to LECA the archaeal membranes were lost and replaced with bacterial ones.

Genes for bacterial lipid biosynthesis now sit in the nucleus and have no remainder in the mitochondrion, while the functions of the archaeal counterparts are largely missing from the nucleus (although isoprene synthesis is kept). It is not yet known why this would be the case. In fact, tracing the genomic path shows that this was the long and convoluted way round: archaeal genes for membrane phospholipid biosynthesis were already in the (proto)nucleus, such that the most parsimonious evolutionary process would have simply involved the gradual loss of the bacterial genes from the mitochondrion, with corresponding replacement of the mitochondrial phospholipids by archaeal ones. Instead, the bacterial lipid genes got transferred from the mitochondrial chromosome into the nuclear genome (along with many other genes), and were eventually lost from the mitochondrion itself. The archaeal analogous functions were correspondingly lost from the nucleus, and the plasma membrane became bacterial, as did all of the organellar, vesicular, and nuclear membranes (Martin et al. 2015) (Figure 28).

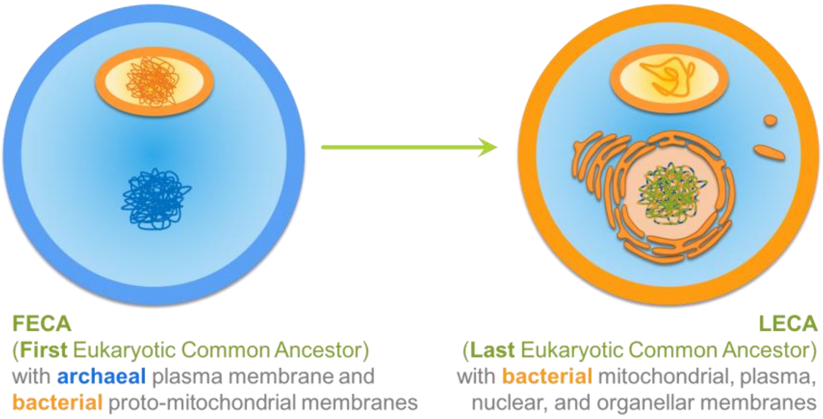

FECA
(First Eukaryotic Common Ancestor)
with archaeal plasma membrane and
bacterial proto-mitochondrial membranes
LECA
(Last Eukaryotic Common Ancestor)
with bacterial mitochondrial, plasma,
nuclear, and organellar membranes

Figure 28. Membranes and genomes from FECA to LECA

The first eukaryotic common ancestor (FECA, left) had archaeal plasma membranes and bacterial proto-mitochondrial membranes, predictably with complete or almost complete archaeal (blue knot) and bacterial (yellow knot) genomes in the (proto)nucleus and (proto)mitochondrion, respectively. Modern eukaryotes, and by inference the last eukaryotic common ancestor (LECA, right) have bacterial plasma, nuclear, mitochondrial, and organellar membranes. The mitochondrial genome is massively reduced but still unequivocally bacterial, whereas the nuclear genome is chimeric, with components from both the archaea (blue) and bacteria (yellow), plus many that cannot be traced to either domain and are therefore deemed eukaryote-specific (green). This view is minimalistic, amongst others, in that the effect of horizontal gene transfers prior to the endosymbiotic event are ignored.

4.3 Why not keep both lipid sets?

One of the two sets of lipids was lost between FECA and LECA. Why didn't eukaryotes keep both? The simplest explanation would be a physiological or ecological tendency for reduced genomes, a major evolutionary force in prokaryotes (Mira et al. 2001), and particularly in symbiotic or parasitic ones (Moran 2002). Although this could suffice as an explanation, redundant sets of genes were in fact kept for many other mitochondrial functions, including DNA replication, transcription, and translation (ribosomal RNAs, ribosomal proteins, and tRNAs). Most of these genes now sit in the nuclear genome and are targeted to the mitochondrion after translation. In fact, the genomes of eukaryotes are less subject to energetic restrictions of size than prokaryotic genomes are (Lane and Martin 2010), so it does not seem obvious that either a pressure or a spontaneous tendency for genome-reduction would cause the loss of one of the two phospholipid synthesis sets.

An alternative explanation lies in physiology. As the eukaryotic cell originated, the two types of lipids would have been expressed simultaneously, so it is reasonable to assume that they would have mixed spontaneously. In fact, in modern eukaryotes a number of phospholipids synthesised in the cytosol spontaneously make their way into the mitochondria, including phosphatidylcholine and phosphatidylinositol (van Meer et al. 2008). Similarly, there is constant exchange of phospholipids between the inner and outer mitochondrial membranes (Tatsuta et al. 2014), such that both the plasma and mitochondrial membranes may have been hybrid as consequence. A reasonable prediction is that such heterotypic membranes would have been unstable (Wächtershäuser 2003), however the demonstration that systems with hybrid G1P and G3P backbones could actually be viable (Shimada and Yamagishi 2011) came as a surprise. Yet the same authors also reported that other properties of the phospholipids, chiefly the varying lengths of the tails, have a large adverse effect on the behaviour of the resulting membrane. So, there may yet be credence to the prediction that heterotypic (hybrid) membranes are, at the very least, less effective than homotypic (purely archaeal or purely bacterial) ones, and therefore deleterious in an ecological sense. The possible competing organisms are described in Figure 29.

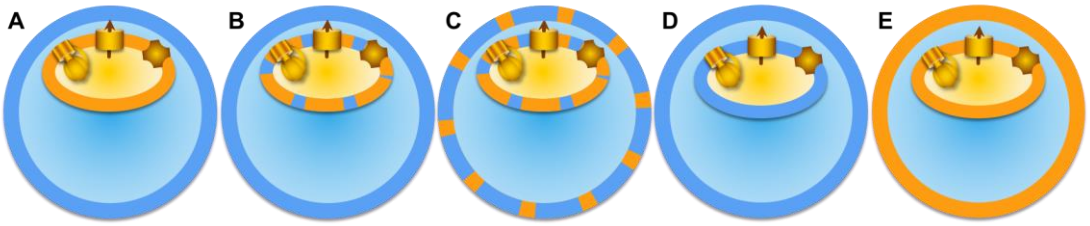


Figure 29. Competition between homo- and heterotypic intermediates

Heterotypic interactions between archaeal and bacterial phospholipids may have constrained the fitness of FECA. Following expression in a common space, it is reasonable to predict that the archaeal (blue) and bacterial (orange) lipids would have spontaneously mixed, such that (**A**) was not possible and cytosolic archaeal lipids leaked into the mitochondrion (**B**). Similarly, both sets of membranes (plasma and mitochondrial, ignoring whether the nucleus, peroxisomes, or other organelles were present at this stage) may have been hybrid if bacterial lipids also leaked out of the mitochondrion (**C**). Either way, a pressure for the more stable homotypic membranes immediately ensued. Although the archaeal option (**D**) was genomically simpler, the bacterial one (**E**) was selected instead. Yellow/golden structures on the mitochondrial membrane represent (chiefly bioenergetic) bacterial membrane proteins.

Even if maintaining a hybrid membrane were viable (Figure 29B-C), removing one of the two phospholipids and keeping a homotypic membrane would have been advantageous (Shimada and Yamagishi 2011). As described above, there were two ways of doing this, of which keeping the archaeal lipids and replacing them with bacterial ones was the easier of the two since, before the mito-nuclear transfer of the bacterial genes, archaeal lipid-synthesis genes were already in the (proto)nucleus. In addition, mitochondria potentially had higher mutation rates than the nucleus, just like at present (Berg and Kurland 2000), such that losing their genes would have been easier. If instead the loss happened after the mito-nuclear transfer of the bacterial genes, then both genes were being expressed and the membranes were hybrid. As discussed above, this would have caused a pressure to form homotypic membranes and therefore silence and potentially lose one of the two sets.

4.4 Why do eukaryotes have bacterial membranes?

Whatever the reason for the unity of eukaryotic lipids, between FECA and LECA bacterial genes for membrane phospholipid biosynthesis were transferred to the nucleus, their archaeal counterparts curtailed, and the original bacterial copies in the mitochondrion lost. Why?

4.4.1 Hypothesis: bacterial lipids were crucial for mitochondrial bioenergetic proteins, so they had to be kept

I suggest here that the reason for this evolutionary process was bioenergetic: as the early eukaryotic cell became increasingly reliant on mitochondrial energy conversion, the physiological adaptation of the mitochondrial bioenergetic proteins to their bacterial membranes became correspondingly indispensable. Replacing the bacterial phospholipids with archaeal ones would have meant maladaptation and decreased fitness; bacterial lipids had to be kept (Figure 30).

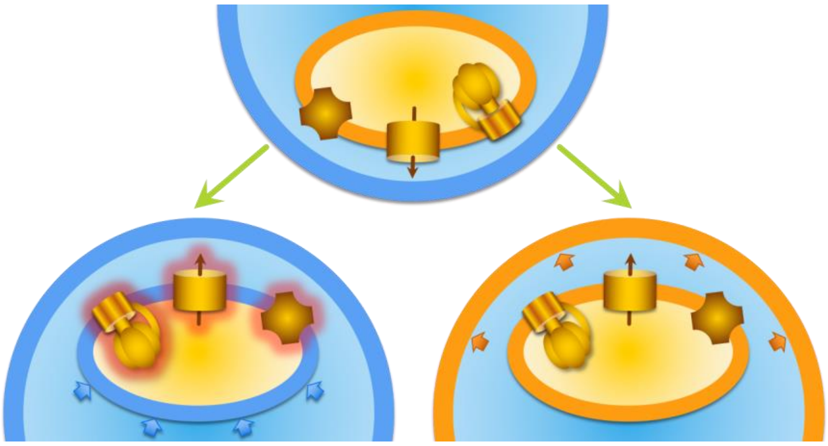


Figure 30. Adaptation of bacterial bioenergetic proteins to an archaeal or a bacterial membrane in two alternative versions of LECA

From a FECA with heterotypic membranes (top centre), two homotypic LECAs could evolve (bottom left and right). Mitochondrial bioenergetic proteins (yellow/golden structures on the internal membrane) would have been less efficient on an archaeal membrane (bottom left; blue arrows represent the imposition of archaeal lipids on the mitochondrial membrane). Instead, the plasma membrane was changed to a bacterial one (bottom right; orange arrows represent imposition of bacterial lipids on the plasma membrane).

In modern eukaryotes ATP yield is higher (up to 13 times) with respiring mitochondria than without them (Rich 2003). All known eukaryotes either have and heavily depend on mitochondrial ATP production, or had and lost mitochondria and now parasitise organisms that do have them (Tovar et al. 2003). As the early eukaryotic cell evolved, it came to rely increasingly upon mitochondrial energy production (Lane and Martin 2010). This is a complex process that depends on the tight coupling between the bioenergetic proteins and the membranes on which they sit (Mitchell 1961).

At an ecological level, a mitochondrial replacement of its original bacterial membranes (even if hybrid) with purely archaeal ones would have meant decreased fitness in local competitions with individuals that either had kept the original membrane or that moreover were instead replacing the archaeal phospholipids in the plasma membrane to the bacterial analogues.

4.4.2 Testing the hypothesis

The endosymbiotic event meant a massive horizontal transfer of genes from the mitochondrion to the nucleus, often referred to as mito-nuclear transfers, and the event as a whole more generally as an endosymbiotic gene transfer (EGT) (Timmis et al. 2004). Horizontal gene transfers (HGTs) should thus provide a useful tool for testing the hypothesis: if correct, and with all other things being equal, then HGTs across the prokaryotic domains should be easier for cytosolic proteins, which have to adapt from an aqueous to another aqueous medium, than for membrane proteins, which would have to sit in a foreign membrane. That is, if a gene is being imported by archaea from bacteria, or vice versa, it should be less likely to be kept if it is a membrane protein (Figure 31).

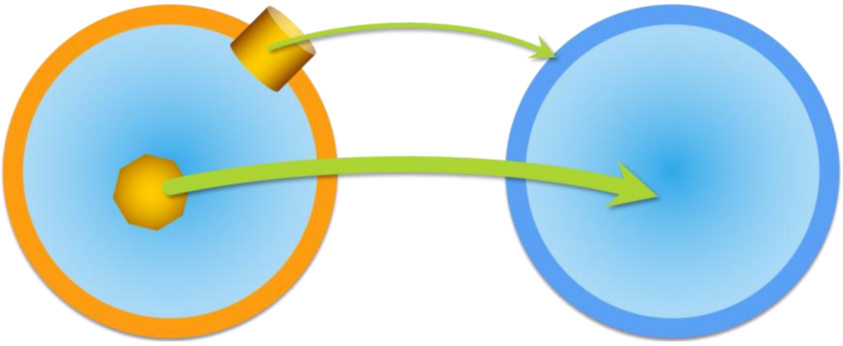


Figure 31. Prediction 1: horizontal gene transfers across the prokaryotic domains should be less common for membrane proteins

Picking up a foreign gene should be more difficult for membrane-bound proteins (cylinder), which would have to sit in a heterologous environment, than for water-soluble proteins (octagon).

This difficulty in importing foreign membrane proteins is frequently observed in the challenging biosynthetic and structural studies involving heterologous expression of membrane proteins: unsurprisingly, it is generally far easier to express foreign water-soluble proteins; and similarly, it is more straightforward to clone and express a foreign membrane protein in a host that is as close as possible to the original species (Schlegel et al. 2012). This naturally extrapolates into the hypothesis put forward here: in HGTs across domains it should be generally more difficult to pick up membrane proteins than water-soluble proteins. Results for this test are presented in section 4.5.1.

Separately, and from a structural perspective, the hypothesis also predicts that bacterial membrane proteins should be energetically and structurally less stable in archaeal membranes, and vice versa (Figure 32).

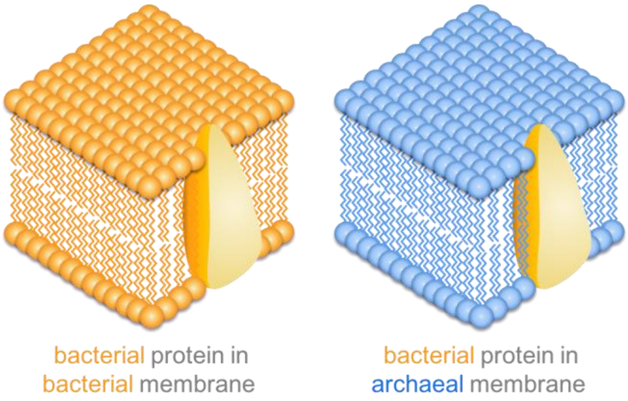


Figure 32. Prediction 2: bacterial membrane proteins should be more stable in bacterial lipids than in archaeal lipids (and vice versa)

A bacterial membrane protein (yellow/orange globule) should have lower energy and be more stable in a bacterial membrane (yellow/orange sphere heads and tails, left) than in an archaeal one (blue sphere heads and tails, right).

I have used molecular modelling techniques, including classical molecular dynamics (reviewed as an appendix in section 4.8), to test this prediction. Results are presented in section 4.5.2.

4.5 Results

4.5.1 Membrane proteins are less likely to be horizontally transferred across domains than water-soluble proteins

To determine whether there is a difference in the likelihood of horizontal gene transfers across domains for water soluble proteins versus membrane proteins, I used the “default” dataset of the Microbial Genome Database (MBGD, mbgd.genome.ad.jp) (Uchiyama 2003; Uchiyama 2007; Uchiyama et al. 2010), which at the moment of this analysis included protein orthologue groups for a total of 787 species, 83 of which are Archaea, 659 Bacteria, and 45 Eukaryotes. These species form a total of 375,229 orthologue groups with any 2 or more members. 6,648 of these groups include at least one archaeon and one bacterium, and 10 or more prokaryotic species in total. To allow detection of horizontal gene transfers only between the prokaryotic domains, eukaryotic sequences were ignored. The MBGD “orthologue” groups often contain multiple sequences for the same species; therefore, where more than one sequence was present within the same group for the same species, the one with the shortest average pairwise distance to the sequences of all other species was kept. The sequences chosen in this way were aligned, phylogenetic trees built, and each protein classified as either membrane-bound or water-soluble (see details in the Methods, section 4.7).

Detecting HGTs using the proportions of archaea vs. bacteria

The prediction of the hypothesis is that detectable trans-domain horizontal gene transfers (tdHGTs) across the prokaryotic domains should be less frequent for membrane proteins than for water-soluble proteins. However, detecting ancient horizontal gene transfer events is not a trivial task. A simple decision rule could be based on a proportion *threshold*: any orthologue clusters in which only *few* sequences belong to one of the two domains is likely to be a horizontal gene transfer from the other domain. Setting such a threshold to define tdHGTs at 90% (i.e. if 9 or more of every 10 members of an OG belong to the same domain), and using a binary classification for each orthologue group as either membrane protein or water-soluble protein, produces the results in Table 5.

Table 5. Contingency table for water-soluble (WS) versus membrane-bound proteins (MP) grouped by trans-domain horizontal gene transfers (tdHGTs) vs. non-tdHGTs, identified by simple proportional composition of the orthologue group with a 90% threshold

Setting the threshold at 90%, if an orthologue group has 10 sequences and a composition of 9 bacteria and 1 archaeon (or vice versa), it is classified binarily as containing a tdHGT. Results allow testing whether membrane proteins are less likely to be transferred across domains, as the hypothesis predicts.

Proportion 90%	non-tdHGT	tdHGT	
Water-soluble	*a* 2138	*b* 3001	*c* 5139
Membrane proteins	*d* 528	*e* 676	*f* 1204
	g 2666	*h* 3677	*i* 6343

Approximately 58.0% of the orthologue groups are classified in this way as containing at least one tdHGT. The hypothesis predicts that the proportion of genes that have been transferred across domains should be higher for water-soluble proteins than for membrane proteins; that is, *b/c* in Table 5 should be greater than *e/f*. Whilst this is indeed the case ($3001/5139 > 676/1204 \Rightarrow 0.584 > 0.561$), the difference is slight, with a high *p* value under Fisher's exact test of 0.146. Relaxing the proportion cut-off to 80% produces the results in Table 6, analogous to those in Table 5.

Table 6. Contingency results as in Table 5, classifying tdHGTs as groups containing 80% or more sequences of the same domain, and correspondingly 20% or fewer from the other domain

Results were produced by lowering the cut-off to 80% proportion bias (i.e. orthologue groups composed by at least 80% bacteria and under 20% archaea, or vice versa, were classified as trans-domain HGTs).

Proportion 80%	non-tdHGT	tdHGT	
Water-soluble	*a* 884	*b* 4255	*c* 5139
Membrane proteins	*d* 254	*e* 950	*f* 1204
	g 1138	*h* 5205	*i* 6343

Here, 82.1% of the orthologue clusters in the mixed group are classified as containing tdHGTs, as opposed to 58.0% above. As above, the prediction that tdHGTs should be lower for membrane proteins is met (4255/5139 > 950/1204 ⇒ 0.828 > 0.789); this time the *p* value drops to 0.00174. However, the size of the effect remains small (~5% difference in the proportion of tdHGTs for the two types of protein, versus ~4% above).

This approach has the caveat that choosing different thresholds produces different results with varying p-values and effect sizes, and it is difficult to interpret these differences. An alternative method of detecting tdHGTs is described next.

Detecting HGTs by monophyly of archaea and bacteria

A potentially more satisfactory approach for detecting tdHGTs is to assess whether the archaea and bacteria each form a monophyletic group, i.e. whether it is possible to root a phylogenetic tree of the aligned sequences in a manner the perfectly splits the two domains. If all archaea are contained within the same group with no bacteria within it, and vice versa, the tree can be assumed to not contain a trans-domain HGT event (which would immediately suggest it is ancestral and perhaps present in LUCA). Results for this approach are presented in Table 7.

Table 7. Contingency table, with tdHGTs determined by monophyly of archaea or bacteria, for orthologue groups shared by at least 1 archaeon and 1 bacterium

Monophyly-1	non-tdHGT	tdHGT	
Water-soluble	*a* 894	*b* 4245	*c* 5139
Membrane proteins	*d* 250	*e* 954	*f* 1204
	g 1144	*h* 5199	*i* 6343

Using this approach, the proportion of genes that have been transferred across domains is higher for water-soluble than for membrane proteins (b/c > e/f ⇒ 4245/5139 > 954/1204 ⇒ 0.826 > 0.792). Similarly, membrane proteins are more prevalent in the non-tdHGT group (d/g > e/h ⇒ 250/1144 > 954/5199 ⇒ 0.219 > 0.183), with a p-value of 0.00680 under a Fisher's exact test.

Trivially, however, genes shared by only one archaeon or only one bacterium would be monophyletic by definition, since in these analyses the root of the tree is always placed between the archaea and the bacteria. Filtering the Mixed group to include only genes shared by at least two species of each domain gives a total of 4,927 orthologues, and the results in Table 8.

Table 8. Contingency table, with tdHGTs determined by monophyly of archaea or bacteria, with orthologue groups shared by at least 2 archaea and 2 bacteria

Monophyly-2	non-tdHGT	tdHGT	
Water-soluble	*a* 869	*b* 3118	*c* 3987
Membrane proteins	*d* 245	*e* 695	*f* 940
	g 1144	*h* 3813	*i* 4927

This retains the relationship, with $p = 0.00552$.

Although these methods for tdHGT detection can be useful, they have the caveat of ignoring ecology (see Discussion in section 4.6). An alternative, purely biophysical and energetic approach is discussed next.

4.5.2 Molecular modelling: archaeal and bacterial membrane proteins in homotypic and heterotypic lipids

The hypothesis suggested in this chapter for the bacterial nature of eukaryotic membranes leads to a straightforward structural prediction: in extant prokaryotes, bacterial membrane proteins should be less stable in archaeal membranes than in bacterial ones, and vice versa. I set out to test this hypothesis computationally by performing molecular dynamics simulations of relevant systems in GROMACS (Berendsen et al. 1995; Lindahl et al. 2001; van Der Spoel et al. 2005; Hess et al. 2008), a classical molecular mechanics package of common use in computational chemistry. However, preliminary tests proved difficult to obtain reliable forcefield parameters (section 4.8.3, p. 134) for archaeal phospholipids. Any comparison between different phospholipid models depends strongly on the parameter values reported for each in the literature and molecular packages. So, to guarantee an accurate comparison of results, the approach used here focused on the subtlest yet the most constant difference between archaeal and bacterial phospholipids: the opposite stereochemistries of the glycerol-phosphate backbones (Chapter 2). Dipalmitoyl-phosphatidylcholine (DPPC) is a model bacterial phospholipid often used in computational simulations of bacterial membrane proteins. This phospholipid has two simple non-ramified and completely saturated tails, with only single bonds between carbons (C–C), and carbons and hydrogens (C–H), such that its only chiral centre is that of the glycerol-phosphate backbone, shown by the yellow circle in the upper part of Figure 33.

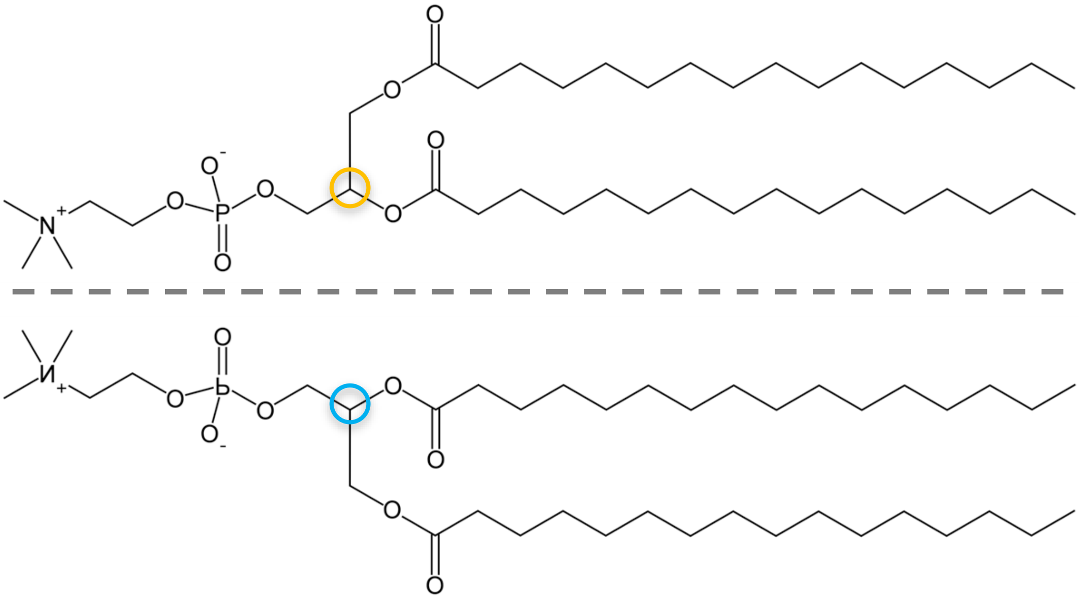


Figure 33. Dipalmitoyl phosphatidylcholine (DPPC), a model bacterial phospholipid (top), and its mirror image

DPPC (above the dashed line) has only one chiral centre: that of the glycerol-phosphate backbone (yellow circle). A mirror image of this structure (below the dashed line) would be identical in all aspects, except the stereochemistry of the backbone, which would be archaeal type (blue circle). Forcefield parameters for DPPC should behave similarly with the inverted structure, since all atoms and their connectivity are the same (this was tested, see Figure 35).

If DPPC were to be projected through a mirror plane, the resulting molecule (lower part of Figure 33) would be identical in all aspects except the stereochemistry of the mirrored glycerol-phosphate backbone. This opposite stereochemistry makes the phospholipid more archaeal-like. That is, DPPC has a bacterial-type *sn*-glycerol-3-phosphate backbone, whereas the inverted molecule has an archaeal *sn*-glycerol-1-phosphate backbone, but with bacterial ester linkage and non-ramified tails. It should be possible to use the parameters of the original molecule for the mirrored one, since all the atoms are identical. Ultimately, this should provide for a straightforward comparison between the two arrangements, to evaluate whether a membrane protein sitting in each of the two systems (Figure 34) has different energetic levels.

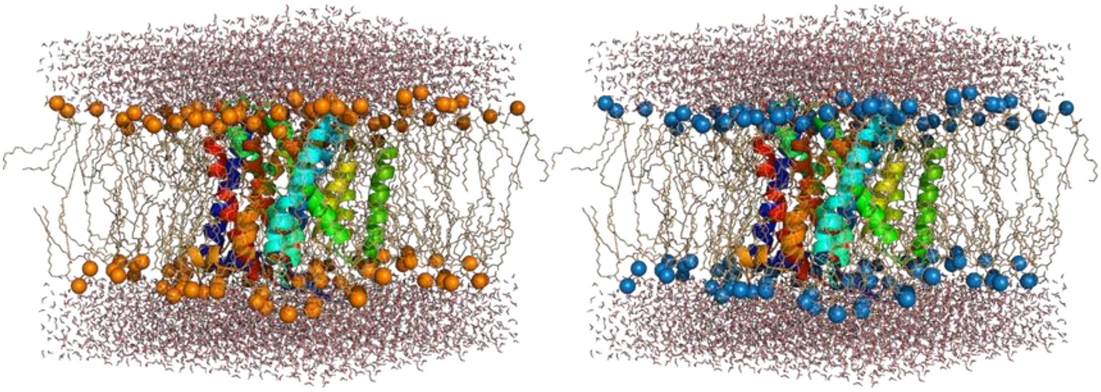


Figure 34. Membrane protein embedded in a bacterial (left) and pseudo-archaeal (right) lipid bilayer systems

A membrane protein (rainbow-coloured helices in the centre of the two figures) sits in the *forward* (orange sphere heads, left) and *reverse* (blue sphere heads, right) DPPC bilayer. The test being performed in this section is whether a membrane protein has a different behaviour and energetic level in the bacterial (left) versus the pseudo-archaeal (right) systems.

Such a "reverse" membrane was produced by inverting the Z coordinate of the bilayer description file (see Methods). An initial necessary test was to determine whether the two systems behave similarly before the inclusion of the membrane protein, i.e. whether the inverted phospholipids indeed respond correctly to the parameters of the original ones. Results in Figure 35 show that the pressure, density, kinetic energy, and potential energy of the two systems are indistinguishable.

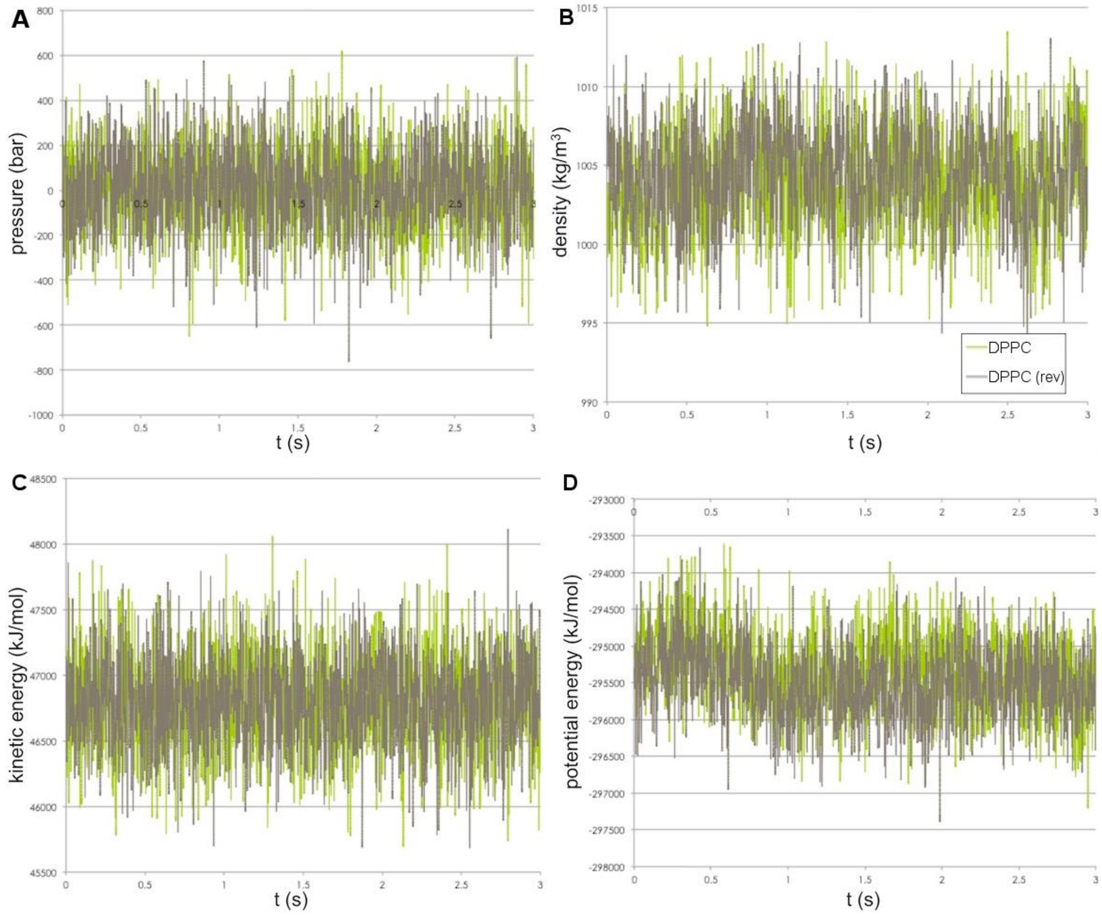


Figure 35. The physical and energetic properties of the *forward* (bacterial) and *reverse* (pseudo-archaeal) DPPC systems are equivalent

The pressure (**A**), density (**B**), kinetic energy (**C**), and potential energy (**D**) of the original bacterial DPPC bilayer system (green) and the mirrored pseudo-archaeal system (brown) are indistinguishable from each other.

Since the inverted DPPC bilayer behaves reliably similarly to the original one, the next step was to model proteins sitting in the two systems. To evaluate the viability of the two types of bilayer with embedded proteins, I first embedded an artificial peptide known as KALP-15. This is a simple 15-amino-acid membrane peptide of no biological relevance (or indeed affiliation, as it is an entirely artificial construction[vii]). Results, in Figure 36, show that, while the pressure and density of the two systems are comparable, the kinetic energy is slightly higher in the forward system, while the potential energy is noticeably higher.

[vii] KALP-15 and other similar model peptides of varying length are used regularly in computational chemistry to test molecular dynamics simulations of membrane systems (Kandasamy and Larson 2006).

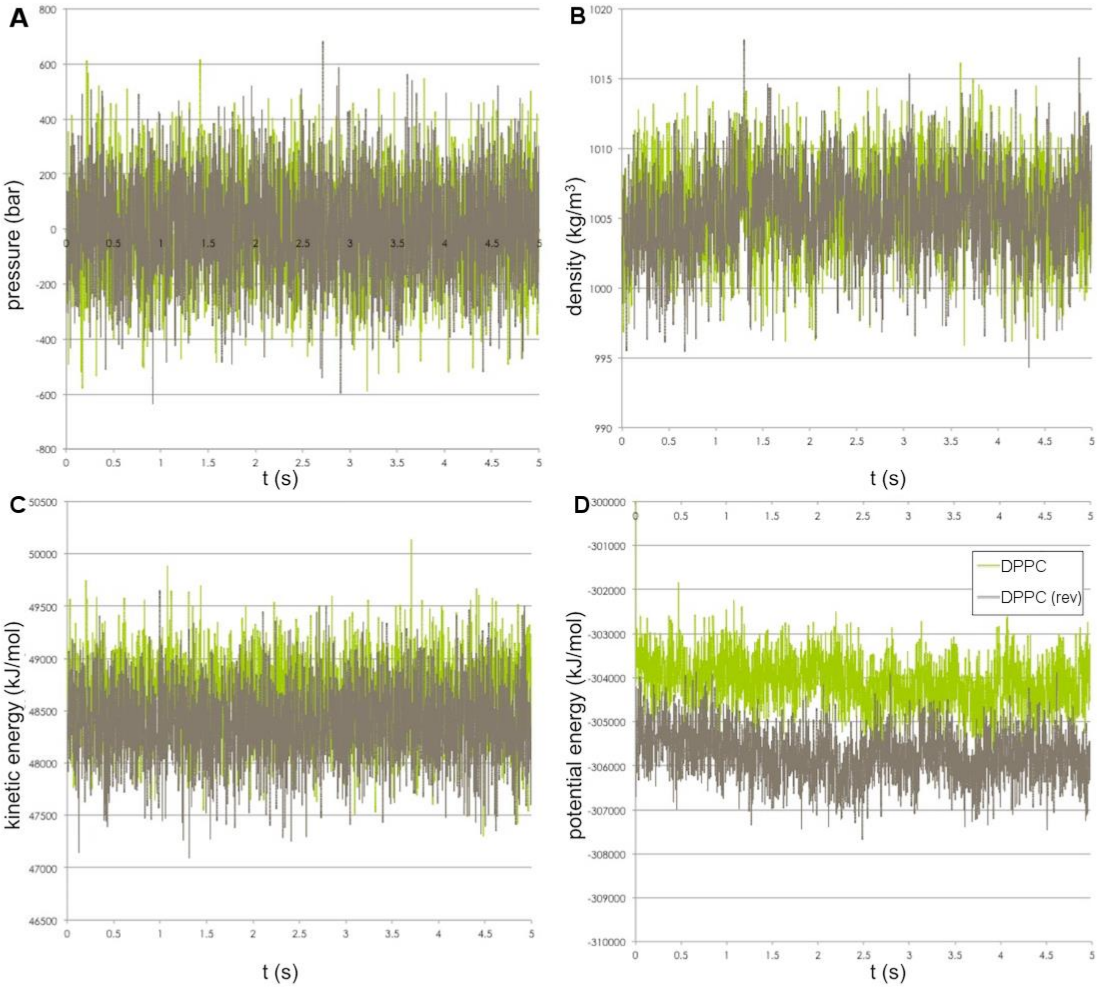


Figure 36. Pressure, density, kinetic energy and potential energy of artificial peptide KALP-15 in the *forward* and *reverse* membranes

While the spatial properties (**A**, **B**) are comparable in the bacterial (green) and pseudo archaeal (brown) systems, the energetic values are noticeably different for kinetic (**C**) and in particular for potential (**D**) energy.

Following this observation of a different behaviour with a pseudo-peptide in the bacterial (forward DPPC) and pseudo-archaeal (reverse DPPC) systems, the next step was to model a biologically relevant protein. I used the bacterial sodium-proton antiporter NhaA from *E. coli*. Figure 37 shows that, as for the artificial peptide KALP-15, the energy of the system is different in the bacterial and pseudo-archaeal membranes.

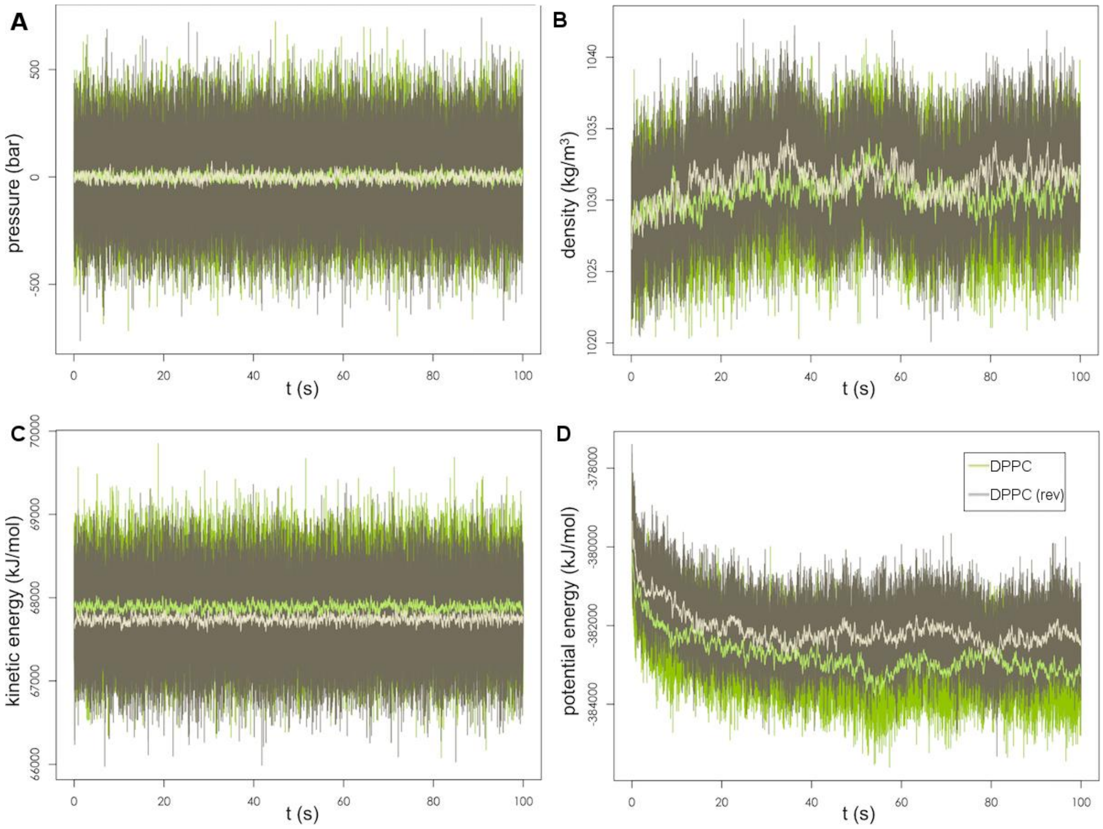


Figure 37. Bacterial Na^+/H^+ antiporter NhaA from *E. coli* produces different results in the two types of membranes

While the pressure (**A**) and density (**B**) appear comparable, the kinetic (**C**) and potential energies (**D**) are different, with the potential energy noticeably lower in the bacterial-type membrane. The total energy (not shown) is also lower in the bacterial membrane, since the contribution of the kinetic energy is small.

4.6 Discussion

The results provide support for the prediction that membrane proteins are less likely to be transferred across the prokaryotic domains, potentially due to energetic constraints.

As mentioned above, there are weaknesses associated with the approaches used here to determine trans-domain HGTs. The approach using the proportion of archaea vs. bacteria produces different results for different thresholds, with correspondingly varying *p* values and effect sizes. Similarly, one weakness of the detection approach by monophyly is that genes that are not classified as tdHGTs are intrinsically assumed to have been in LUCA. This means that genes shared by only one or a few members of one domain and many members of the other will be classed as LUCA if they branch monophyletically, while it is possible that they were instead acquired horizontally by the ancestor of the members of the smaller group (Figure 38).

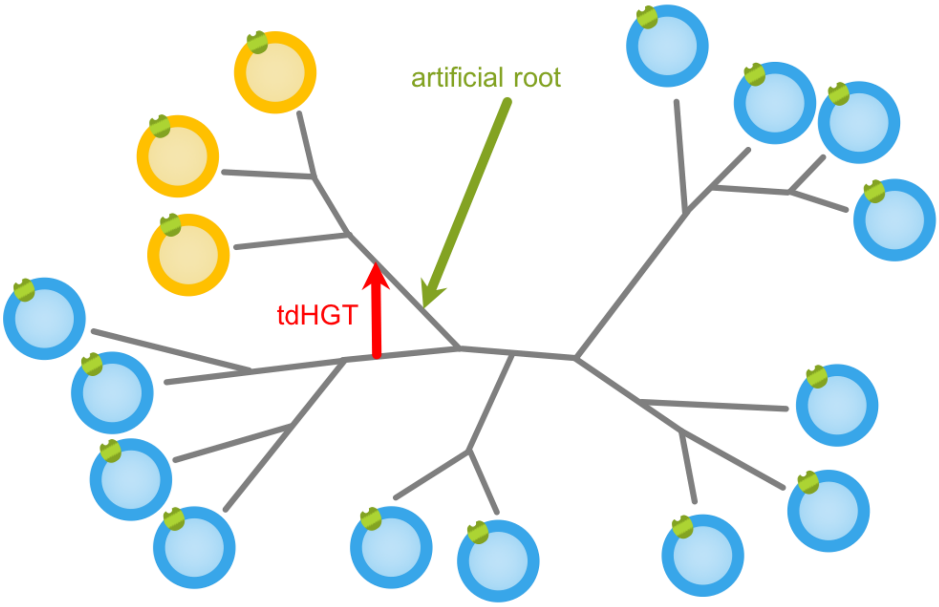


Figure 38. Monophyletic distributions of small groups can lead to an incorrect identification of LUCA genes

With yellow circles representing bacterial species and blue circles archaea, the green globules represent a protein for which a phylogenetic tree is being constructed. In the tdHGT detection analysis by monophyly described above, a horizontal transfer at the base of the bacterial clade (red arrow) may lead to a monophyletic distribution of all descendants of the species that acquired the gene. If a tree root is assigned at the split between the two domains (green arrow), the incorrect conclusion that the gene is ancestral (i.e., that it was present in LUCA) would be reached.

A possible solution would be to force both domains to be monophyletic for a non-tdHGT classification, but this would conversely over-represent tdHGTs. Although simple, the method of selecting only one domain to be monophyletic was preferred.

In general, both approaches exhibited a pattern that seems to support the hypothesis put forward here, but the effects are small and the *p* values, although below a traditional cut-off of 0.05 in most cases, are still questionable. In fact, it has been suggested that using a traditional 0.05 cut-off for *p* can lead to no less than 30% incorrect reports of positive discovery (Colquhoun 2014). Similarly, the American Statistical Association (ASA) has recently issued a “warning” stating that *p* values can determine neither whether a result is important nor whether it confirms a hypothesis

(Wasserstein and Lazar 2016). But conversely, the ASA also warns that *p* values below or above a given threshold do not binarily make a hypothesis true or false (Wasserstein and Lazar 2016). In that sense, the results presented above do not and cannot confirm the hypothesis, but they do encourage the development of further analyses to establish whether tdHGTs are indeed less likely for membrane proteins than for water-soluble proteins.

The molecular modelling computations also produce encouraging results, but a number of subsequent computational tests are required to provide final support for the hypothesis. First, it is important to establish whether the pattern reported here for the bacterial NhaA antiporter remains for other bacterial proteins and lipid systems. Similarly, it is possible that the pattern is cause simply by a difference in the numbers of molecules in the two systems, specifically water molecules, a possibility that needs to be investigated further.

Next, archaeal proteins will need to be tested to determine whether the bacterial lipids with artificially generated archaeal *sn*-glycerol-1-phosphate headgroups produce a lower energy than the regular bacterial lipids. I attempted to perform such calculations (data not shown), but the limitation was that all the suitable archaeal membrane proteins I procured from the Protein Data Bank have been crystallised after expression in bacterial membranes (e.g. PDBs, 4XXJ and 4PXK), a process that could alter their initial structure in precisely the way I wished to evaluate. I observed no relevant differences between the two systems for several systems constructed in this manner.

A further relevant test would be to use a fully archaeal lipid system to determine whether membrane proteins are more stable in an archaeal membrane than in a bacterial one. I also made multiple starts along this line during several months, using four different forcefields (GROMOS 53a6, GROMOS 43a1p, CHARMM27, and CHARMM36) but ultimately failed, possibly due to poor parameterisations of the archaeal lipids. Each of these calculations took several weeks, and in each one the systems failed to converge (i.e. the mean square distances between atoms in the structures and energy values continued to increase over time). There are extremely limited archaeal lipid systems and parameters available in the internet or literature (only one source I am aware of at the time of writing, Lipidbook (Domański et al.

2010) at lipidbook.bioch.ox.ac.uk). In fact, the parameterisation of archaeal phospholipids has such notorious limitations that molecular modelling of archaeal proteins is regularly done in bacterial lipid systems instead (e.g. Araya-Secchi et al. 2011). It is therefore likely that this kind of work will have to wait until more and better forcefield parameters for archaeal phospholipids become available.

A more formal method of determining energies in embedded membrane-protein systems may be of use. A method called *umbrella sampling* is available that determines the binding energy of a peptide to a reference group in the system; in the case of an embedded membrane protein, the method allows the computation of the binding energy of the peptide to the surrounding phospholipids (Lemkul and Bevan 2011). The method works by pulling the membrane protein perpendicularly away from the lipid membrane into the solvent, performing molecular dynamics calculations at a number of positions along the trajectory, and integrating the differences to obtain the binding energy, i.e. the difference in energy between the systems with the embedded and free proteins. Since each of the steps along the trajectory is itself a molecular dynamics calculation, this method is considerably more time-consuming than single MD calculations. I performed a respective pair of simulations using this method for a bacterial protein embedded in the forward and reverse bacterial DPPC systems, but could detect no differences between the two (data not shown). Once more, a conclusive analysis of this type will require a fully archaeal membrane system, for which parameters and forcefields do not exist at present.

These computations are exceedingly time consuming, each taking several weeks even after the successful acquisition of a suitable protein structure and lipid bilayer from the literature and internet databases, followed by assemblage of the system. The most time-consuming steps following construction of the system include the preliminary stabilisation, molecular dynamics, analysis of output, and, most significantly, testing for selection of a suitable forcefield (which requires repeating all of the previous steps). This is added to the waiting times on the calculation clusters, which at several points during the work described in this thesis were in the weeks due to down times of the server.

In spite of the negative results, the positive ones described above are encouraging, and allow the suggestion of a number of potential laboratory experiments. Specifically, it should be relatively straightforward, at least within the typical limitations of expressing membrane proteins heterologously (Schlegel et al. 2010), to clone archaeal membrane proteins into bacterial models in order to evaluate a converse scenario in which archaeal membrane proteins are forced to interact with bacterial phospholipids. This has been done multiple times for other purposes, mainly biotechnology and crystallography. The prediction is that archaeal proteins should have a reduced function in bacterial membranes, and the effect should be more significant than for similar heterologously expressed bacterial proteins. Less straightforward, but still viable, it could be possible to express bacterial membrane proteins in an archaeal model such as *Methanosarcina* or *Halobacterium* (Allers and Mevarech 2005), to evaluate a system closer to the origin of eukaryotes. Incidentally, *Halobacterium* already contains a large number of bacterial membrane proteins embedded in an archaeal membrane (Nelson-Sathi et al. 2012), as discussed below.

Bioinformatics results can also be expanded upon, for example by analysing well-supported HGT events between archaea and bacteria and determining whether there is a significant bias in favour of cytosolic over membrane proteins. Such an approach, as well as the one used here, has the caveat of ignoring ecology. That is, it is possible that the prediction is correct (that membrane proteins are indeed less stable in a foreign membrane), yet ecological constraints mean that they will still be picked up because it is more advantageous to have a sub-optimal membrane protein than to not have it at all. A well-known example of this would be halorhodopsin and bacteriorhodopsin, a pair of light-powered proton pumps crucial to the survival of species in the Halobacteria, the ancestor of which was unequivocally acquired from bacteria (Nelson-Sathi et al. 2012). Still the question is worth pursuing.

In all, this chapter provides a testable bioenergetic and selective explanation for why eukaryotes, though evolving from an archaeal host, have bacterial membranes. This is one of many long-standing puzzles in the endosymbiotic theory for the origin of the eukaryotic cell.

4.7 Methods

4.7.1 Obtaining orthologues from the MBGD dataset

The MBGD default dataset contains 375,229 proteins with orthologues detected in any 2 or more of the 787 species (83 archaea, 659 bacteria, and 45 eukaryotes). Ignoring eukaryotes, and filtering for proteins shared by 10 or more prokaryotic species gave 19,741 orthologue groups, 6,648 of which are shared by at least 1 archaeon and 1 bacterium. Table 9 presents a summary of these results.

Table 9. Analysis of 742 species (705 prokaryotes) on the MBGD database

Numbers of orthologue groups (OGs) per subset	
375,229	orthologue groups shared by any 2 or more species out of the 787 (including eukaryotes)
19,741	shared by any 10 or more prokaryotes
6,648	shared by at least 1 bacterium and 1 archaeon and 10 or more total prokaryotes
1,204	are membrane proteins (by TMHMM predictions. This is 19.0% of the 6,648)
4,964	are tdHGTs (by monophyly of archaea and bacteria. This is 82.0% of 6,648)

4.7.2 Selection of one sequence per species

The MBGD database often contains multiple sequences for the same species within the same “orthologue” cluster. To select a single sequence per species, the full set of sequences was aligned using Clustal-Ω (`clustalo`, or Clustal Omega) (Sievers et al. 2011). Where a species had multiple sequences, the pairwise distances of each of its sequence to all sequences from other species was determined. The sequence with the shortest distance (fewest differences) was kept and all others for the same species ignored.

4.7.3 Construction of multiple-sequence alignments and trees

The selected sequences were unaligned and re-aligned using Clustal-Ω. These alignments were used as a source for building phylogenetic trees using FastTree (Price et al. 2010).

4.7.4 Determination of trans-domain Horizontal Gene Transfers (tdHGTs)

Trees were analysed using BioPython module `ete2` (Huerta-Cepas et al. 2010), to determine whether the archaea and bacteria formed respective monophyletic groups (Figure 39).

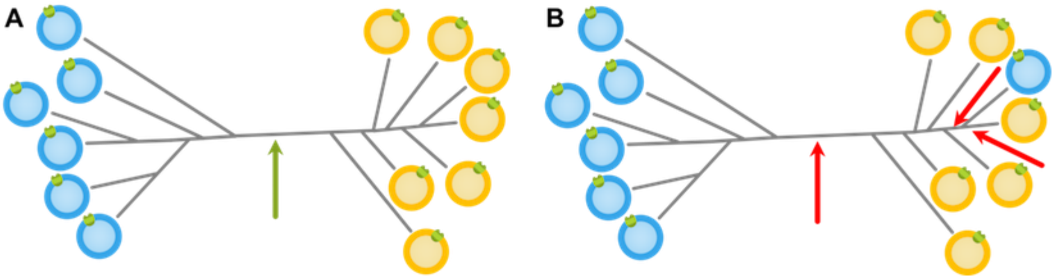


Figure 39. Monophyletic trees were inferred as non-tdHGTs, paraphyletic/polyphyletic ones as tdHGTs

(A) If it was possible to artificially place a root (green arrow) in the tree such that all of the archaeal orthologues (in blue, left of the green arrow) clustered together, and correspondingly so did the bacterial ones (in yellow, right of the green arrow), the encoding genes were assumed not to have been transferred horizontally across domains. (B) If, instead, there was no way of placing the root (red arrows) such that all the archaea clustered together, the tree was inferred to contain at least one trans-domain Horizontal Gene Transfer. Here, the archaea are polyphyletic (they branch from more than one ancestor), and the bacteria paraphyletic (they all branch from the same common ancestor, but the descendants include members that are not bacteria). See definitions of monophyly, paraphyly and polyphyly in section 1.3.2.

If no way could be found of re-rooting the tree such that the archaea and bacteria each formed a self-contained group with a single ancestor, the orthologue group was assumed to contain a trans-domain Horizontal Gene Transfer. In this way, orthologue groups were classified binarily as either tdHGTs or non-tdHGTs.

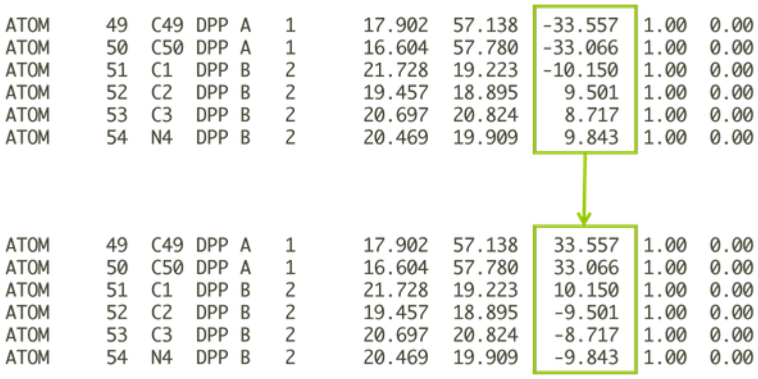
Importantly, this method ignores the effect of any ancient horizontal gene transfers (i.e. it has a potentially high rate of false negatives), which may have played massive and function-defining roles in the evolution of several prokaryotic clades (Nelson-Sathi et al. 2012; Ku et al. 2015), but it is conservative in that the number of false positives should be low.

4.7.5 Classification of membrane proteins

Membrane proteins were annotated using the predictions of the TMHMM algorithm (Krogh et al. 2001), which exclusively identifies trans-membrane helices. Gene Ontology (GO) annotations could be used in addition to the predictions of the TMHMM algorithm; however, the GO annotations in MBGD (and in similar databases in general) are incomplete: several recognisable membrane-bound proteins (e.g. a number of transporters, ion channels, and transposases), although correctly identified by TMHMM, do not have the corresponding GO terms for membrane proteins and thus fail to be identified as such. Additionally, multiple proteins do not have any GO annotations whatsoever, both in the OMA and MBGD databases. Since including the GO annotations produced only a small effect (data not shown), it was concluded that they were not a useful predictor The effect of including GO information could be unpredictable, since these are not systematically annotated. TMHMM, although biased towards helices, is systematically so, such that results are reliably comparable. This in turn depends on the proportions of membrane proteins identified by TMHMM being comparable between archaea and bacteria, which is indeed the case (19.53% in archaea versus 22.95% in bacteria).

4.7.6 Mirroring of lipid bilayer

To produce a geometry file with DPPC molecules with an archaeal *sn*-glycerol-1-phosphate backbone instead of the original bacterial *sn*-glycerol-3-phosphate, I created a simple python script that multiplied the Z-coordinate of every atom in a PDB file by –1, as described in Figure 40.



```
ATOM     49  C49 DPP A   1       17.902  57.138 -33.557  1.00  0.00
ATOM     50  C50 DPP A   1       16.604  57.780 -33.066  1.00  0.00
ATOM     51  C1  DPP B   2       21.728  19.223 -10.150  1.00  0.00
ATOM     52  C2  DPP B   2       19.457  18.895   9.501  1.00  0.00
ATOM     53  C3  DPP B   2       20.697  20.824   8.717  1.00  0.00
ATOM     54  N4  DPP B   2       20.469  19.909   9.843  1.00  0.00

ATOM     49  C49 DPP A   1       17.902  57.138  33.557  1.00  0.00
ATOM     50  C50 DPP A   1       16.604  57.780  33.066  1.00  0.00
ATOM     51  C1  DPP B   2       21.728  19.223  10.150  1.00  0.00
ATOM     52  C2  DPP B   2       19.457  18.895  -9.501  1.00  0.00
ATOM     53  C3  DPP B   2       20.697  20.824  -8.717  1.00  0.00
ATOM     54  N4  DPP B   2       20.469  19.909  -9.843  1.00  0.00
```

Figure 40. *Black-box* description of a simple script to mirror a PDB file by multiplying every Z coordinate by –1

The script (not shown) simply parses a PDB file searching for every “ATOM” entry, and inverts the sign of every Z coordinate, leaving the X and Y coordinates, and everything else, unaltered. Effectively, this produces a mirrored image in which every chiral centre will be turned to its enantiomer and every other molecule will be ultimately unaltered (see Figure 33).

The script was used on a regular 128-lipid DPPC bilayer description file obtained from Peter Tieleman’s website at wcm.ucalgary.ca/tieleman/downloads, and all calculations performed in GROMACS as described by Lemkul and Bevan (2011) and in the website www.bevanlab.biochem.vt.edu/Pages/Personal/justin/gmx-tutorials/membrane_protein.

Briefly, the lipid coordinates were expanded horizontally using inflateGRO (Schmidt and Kandt 2012), a script that makes the lipids separate artificially from each other on the membrane plane to provide space for the protein. The protein was then inserted into the middle of the system, with any overlapping lipids removed to avoid the possibility of two atoms on the same space, which would lead to infinite repulsive forces and therefore the collapse of the simulations. The lipids were then shrunk back gradually using inflateGRO, to a total area per lipid within the reported 62.9–64.0 Å^2 (Nagle et al. 1996), and applying strong force constraints on the protein to keep it stable in the vacuum while the lipids wrapped around it using successive energy minimisations in GROMACS. That is, the expanded lipids were allowed to gradually re-form a normal bilayer, this time around the protein, which was located at the centre. Periodic boundaries were applied to make the lipids at the borders interact with the ones on the opposite side, such that the system as a whole behaved as an infinite membrane. Up to this point the system had no water molecules. The system was then hydrated, and any water molecules that ended within the membrane were removed via Tcl/Tk scripting in VMD (Humphrey 1996). Where charges were present, the total charge was neutralised by replacing water molecules with the smallest possible number of either sodium (Na^+) or chloride (Cl^-) ions. The energy of the system was minimised with force constraints on the protein, to allow the relaxation of lipids, ions, and water molecules around the protein. Two equilibration processes were performed next, first an NVT (for Number of particles, Volume, and Temperature), and then an NPT (where P: pressure). This allowed the correct distribution of water molecules, lipids, and ions, and prepared the system for the molecular dynamics simulation. The simulation was then started and run for the times shown in the relevant figures above.

4.8 Appendix: Molecular modelling

Note: this section has been adapted from MRes project "Conformation-Activity Relationship of G-Protein-Coupled Receptors: Computational Modelling of the human P2Y1 receptor", submitted to UCL 02 May 2012, and references therein (Young 2001; Cramer 2004; Hinchliffe 2008; Jensen 2010).

Molecular modelling is the use of computers to study the structure and function of chemical substances (Hinchliffe 2008). There are many tools within this discipline, the most detailed of which use quantum mechanics calculations to determine the properties of molecules with high accuracy. The problem is that these lead to equations that become intractable when more than a few atoms are involved (Dirac 1929). This makes pure quantum mechanics methods of little use for modelling whole systems of proteins, let alone membrane proteins embedded in a lipid bilayer, with thousands of solvent (water) molecules and ions at either side. Instead, approximations are needed. The simplest of these was first used successfully approximately 40 years ago (McCammon et al. 1977), and involves treating the protein and its surroundings as a classical Newtonian system.

4.8.1 Molecular Dynamics (MD)

Molecular Dynamics, in its classical-physics incarnation, is the application of Newton's equations of motion in computational chemistry to describe the movement of atoms and molecules over time. The main use of MD simulations in biochemistry is to study the behaviour of biological macromolecules (proteins and polynucleotides) either in solution or in interaction with the phospholipids in a bilayer membrane system.

Newton's second law describes the behaviour of a time(t)-depending force (F) as a function of the mass (m) of a particle and its acceleration (a)

$$F(t) = m \cdot a(t) \qquad [6]$$

Acceleration can be substituted as the second derivative of position (x) with respect to time, presented in one dimension in equation [7] for simplicity

$$F(t) = m\frac{\partial^2 x}{\partial t^2} \qquad [7]$$

It is of interest to know how the positions of the particles change in time. Thus, solving for the second-order differential term on the right-hand side of equation [7], and approximating to a discrete scenario gives

$$\frac{\partial^2 x}{\partial t^2} \approx \frac{x(t+\Delta t) + x(t-\Delta t) - 2x(t)}{\Delta t^2} = \frac{F(t)}{m} \qquad [8]$$

Solving for the position at each further time step, *x(t+Δt)*

$$x(t+\Delta t) = 2x(t) - x(t-\Delta t) + \frac{\Delta t^2 F(t)}{m} \qquad [9]$$

The process begins with an initial structure, normally minimised. Since this is defined as time zero, (i.e. there's no *x(t – Δt)*) and since there is no force on a minimised structure, the directions of the particles in the first step are normally chosen at random.

The duration of each step is determined from the temperature that has been chosen for the simulation, according to

$$k_B T = m\langle v_x^2 \rangle \qquad [10]$$

where k_B is Boltzmann's constant, $\langle \dots \rangle$ represents an average over all particles in the simulation, and v_x is the velocity, calculated traditionally as

$$v_x = \frac{x(t+\Delta t) - x(t)}{\Delta t} \qquad [11]$$

The size of the time step is chosen small enough so that temperature and total energy remain constant throughout the simulation, and typical values range between 1 and 10 fs.

Since gradient-based optimisation algorithms can easily get trapped in local minima, and the sizes and complexity of biological macromolecules make it all but impossible for researchers to determine whether a minimum is global or local, MD can be a remarkably useful tool in detecting the true minimum-energy structures. However, MD can also be used to estimate thermodynamic properties, as well as the biochemical behaviour of the system. Due to the computational costs involved, most simulations are allowed to run only within the nanosecond scale, although extending simulations into microsecond and even millisecond durations has proven worthwhile for properly exploring both potential energy surfaces and behaviour of biomolecules.

4.8.2 Molecular Mechanics (MM)

Although it is possible and valuable to run MD simulations using quantum-chemistry algorithms, the costs of doing so are prohibitively large for the vast majority of systems of interest with the computational resources typically available. It is therefore imperative to use approximations, and the classical one called Molecular Mechanics (MM) is chief of these in biological macromolecular modelling.

MM treats atoms as spheres and bonds as springs. This is clearly a very coarse approximation, as it implicitly allows unrestrictedly small variations in energy (as opposed to quantised values), it defines bonds in a fixed manner (as opposed to allowing electronic probability densities to arise from the calculations), and it ignores electronic transitions (therefore making it all but impossible to model chemical reactions). However, a large number of computational chemistry calculations are concerned with the determination of bond lengths and angles, i.e. chemical structure geometries; MM calculations can be remarkably accurate at this, and are thus highly regarded by the scientific community as a viable alternative to the intractable calculations that would be required in quantum descriptions of computational structural biology.

4.8.3 Forcefields

A typical MD simulation performs its calculations using what is called a Molecular Mechanics “forcefield”. In the balls-and-springs analogy, this is essentially a description of the force constants, equilibrium angles and lengths, and interactions between near neighbours, including Coulombic and other “nonbonded” interactions. Forcefields also involve a definition of “atom types” since, for example, a nitrogen in an amine group can be expected to behave differently from one in a nitro group.

This latter point draws attention to the validity of a forcefield when tackling a given problem. The accuracy that a researcher can expect will depend on the molecules that were used to parameterise the forcefield. Therefore, no forcefield is universal, some being more appropriate for nucleic acids, others for globular proteins, others for lipids, and so on.

A typical forcefield equation can be summarised by six main contributions to the total potential energy (U) of the system, namely bond stretching (bs), bond bending (bb), dihedral torsions (dh), out-of-plane torsions (op), electrostatic/Coulombic interactions (es), and other nonbonded interactions (nb), as follows

$$U = \sum_{AB} U_{bs} + \sum_{ABC} U_{bb} + \sum_{ABCD} U_{dh} + \sum_{AB(D)C} U_{op} + \sum_{AC} U_{es} + \sum_{AC} U_{nb} \qquad [12]$$

Consistent with MM's balls-and-springs approximation, the potential energy involved in **bond stretching** between atoms A and B is typically modelled using some adaptation of Hooke's law by defining a bond force constant k_{AB} and measuring the difference between the bond length r_{AB} and a pre-established equilibrium length $r_{eq,AB}$

$$U_{bs,AB} = \frac{1}{2} k_{AB} (r_{AB} - r_{eq,AB})^2 \qquad [13]$$

Analogously, **bond bending** is typically evaluated in terms of deviations of the angle θ_{ABC} between consecutively bonded atoms ABC from the equilibrium angle $\theta_{eq,ABC}$, weighed by harmonic force constant k_{ABC}

$$U_{bb,ABC} = \frac{1}{2} k_{ABC} (\theta_{ABC} - \theta_{eq,ABC})^2 \qquad [14]$$

As three consecutively bonded atoms define a plane, a subsequent fourth atom will define a **dihedral angle**. A common way of calculating the effect of alterations in the equilibrium torsional angle χ_{eq} is given by

$$U_{dh,ABCD} = \frac{U_0}{2} \left(1 - \cos\left(n(\chi - \chi_{eq})\right)\right) \qquad [15]$$

where n is a "periodicity parameter" based on symmetry of D about the B-C bond (e.g., n =3 if D is one of the three hydrogens of a methyl group).

Out-of-plane torsions also occur in groups of four atoms, but in this case three of the atoms (A,C,D) are bonded to a common central one (B). Considering that A-B-C lie on a plane, D defines an angle that can be distorted and inverted. This contribution arises from that inversion potential

$$U_{op,AB(D)C} = \frac{k}{2sin^2\psi_{eq}}(cos\psi - cos\psi_{eq})^2 \qquad [16]$$

where ψ_{eq} is the equilibrium angle.

Electrostatic interactions can be calculated in a traditional Coulombic fashion

$$U_{es,AC} = \frac{1}{4\pi\varepsilon_0}\frac{q_A q_B}{r_{AB}} \qquad [17]$$

Finally, other **nonbonded** (van der Waals) interactions are frequently modelled as Lennard-Jones 12-6 potentials(Jones 1924) and calculated for all pairs of atoms within certain cut-off distance

$$U_{nb,AC} = \frac{a_{AB}}{r_{AB}^{12}} - \frac{b_{AB}}{r_{AB}^{6}} \qquad [18]$$

where a and b are constants specific to atom types A and B.

Therefore, equation [12] can be re-written explicitly as

$$\begin{aligned} U \quad &= \sum_{bs,AB}\frac{1}{2}k_{AB}\left(r_{AB}-r_{eq,AB}\right)^2 \\ &+ \sum_{bb,ABC}\frac{1}{2}k_{ABC}\left(\theta_{ABC}-\theta_{eq,ABC}\right)^2 \\ &+ \sum_{dh,ABCD}\frac{U_0}{2}\left(1-\cos\left(n(\chi-\chi_{eq})\right)\right) \\ &+ \sum_{op,AB(D)C}\frac{k}{2sin^2\psi_{eq}}(cos\psi - cos\psi_{eq})^2 \\ &+ \frac{1}{4\pi\varepsilon_0}\sum_{es,AC}\frac{q_A q_B}{r_{AB}} \quad + \quad \sum_{nb,AC}\left(\frac{a_{AB}}{r_{AB}^{12}} - \frac{b_{AB}}{r_{AB}^{6}}\right) \end{aligned} \qquad [19]$$

These descriptions reinforce the importance of atom types. As an example, in order to determine the appropriate constant k_{ABC} and equilibrium angle $\theta_{eq,ABC}$ for a bond-bending calculation, it is imperative that all three atom types A, B, and C be specified in the forcefield. It is therefore necessary, as mentioned above, to choose a forcefield that appropriately describes the types of atoms being modelled, taking into consideration factors like hybridisation, formal charge, nearby atoms, and the solvent in which the system is being simulated.

5 THE LOW CONSERVATION OF MEMBRANE PROTEINS ACROSS THE TREE OF LIFE

Note: this chapter is adapted directly from the original research article "*Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life*", written as first author with PhD supervisors Prof. Andrew Pomiankowski and Dr. Nick Lane, in collaboration with Prof. Christophe Dessimoz (Sojo, Dessimoz, Pomiankowski, and Lane, *submitted*).

5.1 Summary

Membrane proteins are crucial in transport, signalling, bioenergetics, catalysis, and as drug targets. Here I show that membrane proteins have dramatically fewer detectable orthologues across the tree of life than water-soluble proteins, less than half in most species analysed, with the largest reductions in prokaryotes. This sparse distribution of membrane proteins could reflect rapid divergence, gene losses, or both. First, I show that membrane proteins evolve faster than water-soluble proteins, particularly in their exterior-facing portions. Second, I demonstrate the preferential loss of membrane proteins by comparing the presence/absence of predicted ancestral proteins within closely related species in both archaea and bacteria. The faster evolution of external portions and preferential loss of membrane proteins reflect increased adaptive evolution to varied environments, while stronger purifying selection operates in the homeostatic interior of the cell. These striking differences in conservation of membrane proteins versus water-soluble proteins have important implications for evolution and medicine.

5.2 Introduction

Biological membranes form the boundary between the cell and its surroundings, and their embedded proteins constitute an active link to the environment, with crucial roles in bioenergetics, transport, signalling, and catalysis (Mitchell 1957; Mitchell 1961; Hedin et al. 2011). Over half of all known drug targets are membrane proteins (Overington et al. 2006). Their study is therefore central to our understanding of the origins and evolution of life, as well as physiology and medicine.

Previous studies have shown that the subcellular localisation of a protein predicts its evolutionary rate. Extracellular proteins secreted from the cell evolve faster than intracellular proteins in both mammals and yeast, as do the external parts of membrane proteins, but the reasons are unclear (Tourasse and Li 2000; Julenius and Pedersen 2006). The pattern does not seem to depend on the essentiality of the gene product, suggesting that mechanisms other than purifying selection on crucial proteins are at play (Liao et al. 2010). Structural and packing constraints undoubtedly play a role, with the exposure of amino acid residues to the solvent (Oberai et al. 2009; Franzosa et al. 2013) and the sub-cellular localisation of the proteins and their fragments (Julenius and Pedersen 2006) being the strongest predictors of evolutionary rate. Membrane proteins also diverge faster than intracellular water-soluble proteins in parasites, where surface interactions evolve under pressure to avoid detection by the host (Volkman et al. 2002; Plotkin et al. 2004). This pattern may be specific to the 'red-queen' dynamics of parasitic interactions, i.e. the need for constant adaptation merely to maintain fitness. Taken together, however, these disparate findings suggest that evolution occurs faster outside the cell, and hint at the operation of a wider evolutionary mechanism.

Here I evaluate the simple hypothesis that protein evolution is faster outside the cell as a result of adaptation to changing environments (Figure 41). Over evolutionary time, the interior of the cell remains stable compared with the exterior, which is forced to change in response to shifting biogeochemical processes, migration and colonisation of new niches, and parasitic interactions. This leads to the faster evolution of secreted water-soluble proteins and outside-facing sections of membrane proteins. The utility of a protein will also depend on the specific environment, potentially leading to preferential losses of membrane-bound gene products over time as environments change (Figure 41). I have analysed large datasets of orthologues to evaluate the conservation of membrane proteins relative to water-soluble proteins across the entire tree of life, to test whether faster evolution outside the cell is driven by adaptation to new environments and functions.

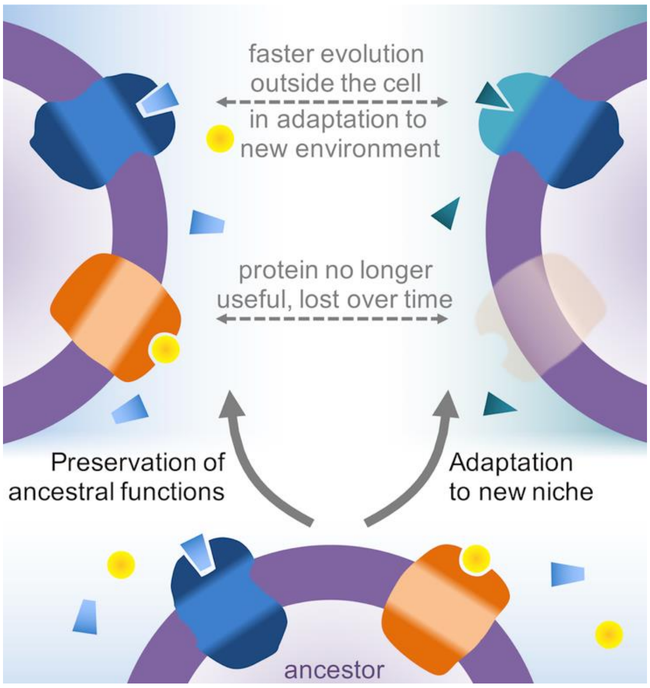


Figure 41. Two-fold effect of adaptation causes faster evolution of external sections and loss of homology in membrane proteins

Adaptation to new functions and niches causes faster evolution for outside-facing sections (top), potentially contributing to divergence beyond recognition. Other proteins may provide no advantage in the new environment, and would be lost entirely over time (centre). For simplicity, the species on the left is assumed to remain functionally identical to the common ancestor (bottom).

5.2.1 Identification of membrane proteins and the topology of their segments

The classification of proteins as either membrane-bound or water-soluble from their primary sequence is an important problem in structural biology and bioinformatics. Most simply, a succession of hydrophobic amino acids may betray the presence of a membrane-embedded segment, but such a gathering can also occur within buried portions of a water-soluble protein which are not exposed to the solvent.

In most proteins of the plasma membranes of the three domains, the stereotypical fundamental structure of membrane-embedded proteins is that of a trans-membrane α-helix (TMH) (Oberai et al. 2006). These helices often alternate with water-exposed loops at either side of the membrane in a sewing fashion, notably in 7-trans-membrane-helical structures, the best known subgroup of which are the G-protein-coupled receptors (GPCRs) in eukaryotes. A number of algorithms exist that predict both the presence of these helices and their topology (Krogh et al. 2001; Käll et al. 2005; Hessa et al. 2007; Bernsel et al. 2008; Reynolds et al. 2008; Viklund et al. 2008; Viklund and

Elofsson 2008; Tsirigos et al. 2015), most of them based on the properties of previously observed proteins and subsequent statistical methods that estimate the probability of each amino acid in a query sequence being part of a TMH, or exposed to the external or internal aqueous phases. These algorithms work well for predicting TMHs, but there are no equivalent methods for the β-sheet barrels in the outer membrane of gram-negative bacteria and mitochondria (e.g. porins). However, since, the majority of trans-membrane proteins are TMHs (Oberai et al. 2006), this may be an acceptable limitation.

The algorithms successfully detect transmembrane helices versus water-exposed loops. However, predicting the specific topology (i.e. whether aqueous loops are inside or outside) is a more difficult matter. One simple solution is the "positive inside rule" (von Heijne 1986b; von Heijne 1992), which stemmed from the observation that the number of (positively charged) lysine and arginine residues is four times higher in cytosol-facing loops than in their periplasm- or exterior-facing counterparts. Therefore, the rule posits that loops with more lys/arg can be expected to be facing towards the relatively negative inside of a bacterial cell membrane (the N-side), while the alternating low-lys/arg loops should face the exterior (the P-side)[viii].

This and other sources of information are incorporated into the packages mentioned above, and in particular in TMHMM (Krogh et al. 2001), the algorithm used in this and the previous chapter. My own results (below) show that, while the identification of TMHs and loops seems highly reliable, there is no consistency in the prediction of inside versus outside loops. Tests with TOPCONS (Bernsel et al. 2009), a consensus predictor that incorporates the results of several algorithms, were no better.

[viii] Significantly, however, no equivalent effect is observed for the negatively charged aspartate and glutamate (von Heijne 1986b).

5.3 Results

5.3.1 Membrane proteins are shared by fewer species in all domains of life

To study the evolution of membrane proteins across the tree of life, I downloaded the 883,176 pre-computed orthologue groups (OGs) for all species from the three domains of life present in the OMA database (Altenhoff et al. 2015). I then obtained the full list of 66 species in the EMBL-EBI list of reference proteomes (www.ebi.ac.uk/reference_proteomes), and extracted the OMA OGs for each protein of each species, where present. I classified each protein sequence as either a membrane protein (MP) or a water-soluble protein (WS) using the predictions of the TMHMM algorithm (Krogh et al. 2001). The number of orthologues found for each protein was determined (i.e. the size of the orthologue cluster, or OG, for each protein). I find that, in all cases, the mean number of orthologues is substantially smaller for MPs than for WSs (Figure 42); that is, membrane proteins are generally shared by fewer species (paired t-test: t=8.05; df=63; p=$2.88 \cdot 10^{-11}$; r=0.712).

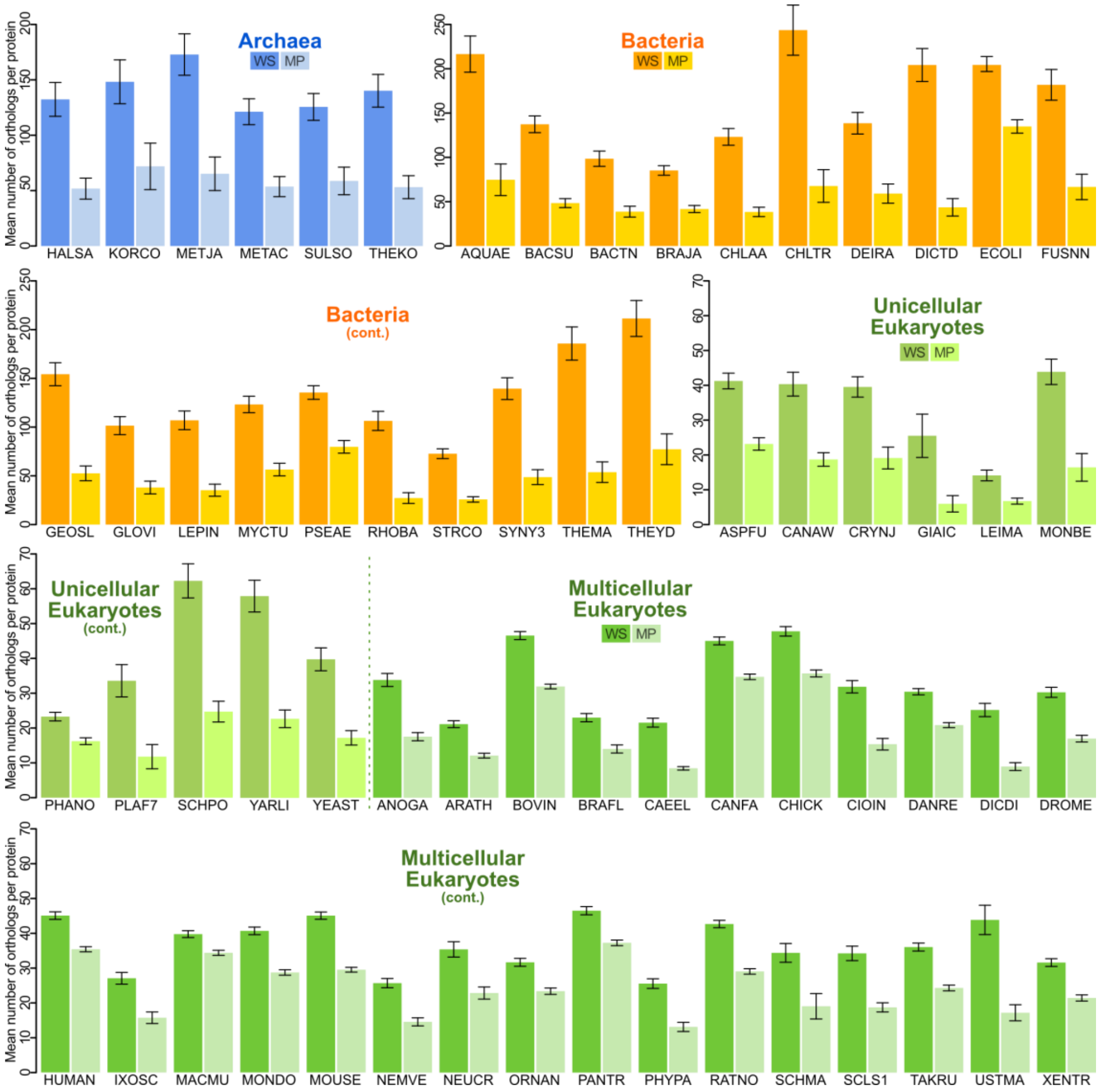


Figure 42. Membrane proteins have fewer detectable orthologues in all three domains of life

The average size of OMA Orthologue Groups (OGs) is substantially smaller for membrane proteins in all 64 species in the EMBL-EBI's list of reference proteomes studied (2 of the 66 species were not found in OMA at the time of this analysis). Five-letter codes are OMA species identifiers; details in Table 10. Dark shade: water-soluble (WS); light shade: membrane proteins (MP). Data represented as means of the number of orthologues that WS and MP of each genome have in OMA ± 2xSEM (standard error of the mean).

Table 10. Orthologue counts of water-soluble and membrane proteins in 66 reference species

In all cases, the mean size of an OMA orthologue group (i.e. the OG cluster size) is smaller for membrane proteins. Species correspond to the 66 in the EMBL-EBI's list reference proteomes, and Figure 42. Two of the 66 species were not found on the OMA database, and two others were replaced with related strains or species (details in the footnotes).

Species[a]	OMA code	Num. proteins in OMA	Num. MPs	Proportion of MPs	Mean WS OG size	Mean MP OG size
Archaea						
Halobacterium salinarum	HALSA	2241	472	0.211	66.2	25.9
Korarchaeum cryptofilum	KORCO	1192	213	0.179	74.1	35.9
Methanosarcina acetivorans	METAC	1639	285	0.174	86.4	32.6
Methanocaldococcus jannaschii	METJA	3514	878	0.250	60.6	26.8
Sulfolobus solfataricus	SULSO	2668	578	0.217	62.8	29.4
Thermococcus kodakaraensis	THEKO	2039	469	0.230	70.1	26.5
Bacteria						
Aquifex aeolicus	AQUAE	1393	271	0.195	216.5	74.7
Bacillus subtilis	BACSU	3984	1122	0.282	137.3	48.4
Bacteroides thetaiotaomicron	BACTN	3931	974	0.248	98.5	38.7
Bradyrhizobium japonicum[b]	BRAJA	6937	1723	0.248	85.2	41.7
Chloroflexus aurantiacus	CHLAA	3802	1104	0.290	123.1	38.4
Chlamydia trachomatis	CHLTR	889	219	0.246	243.6	67.7
Deinococcus radiodurans	DEIRA	2519	478	0.190	138.5	59.1
Dictyoglomus turgidum	DICTD	1673	454	0.271	204.2	43.6
Escherichia coli	ECOLI	4264	1045	0.245	205.4	134.9
Fusobacterium nucleatum	FUSNN	1661	352	0.212	181.9	66.7
Geobacter sulfurreducens	GEOSL	3066	823	0.268	154.2	52.6
Gloeobacter violaceus	GLOVI	3380	696	0.206	101.5	38.0
Leptospira interrogans	LEPIN	3645	1031	0.283	106.9	35.3
Mycobacterium tuberculosis	MYCTU	3933	797	0.203	123.2	56.5
Pseudomonas aeruginosa	PSEAE	5500	1327	0.241	135.5	79.7
Rhodopirellula baltica	RHOBA	3218	745	0.232	106.4	27.2
Streptomyces coelicolor	STRCO	7104	1719	0.242	72.7	25.8
Synechocystis sp.	SYNY3	3058	739	0.242	139.4	48.7
Thermotoga maritima	THEMA	1779	431	0.242	185.7	53.8
Thermodesulfovibrio yellowstonii	THEYD	1716	393	0.229	211.4	77.3
Eukaryota (unicellular)						
Aspergillus fumigatus[c]	ASPFU	8801	1826	0.207	41.2	23.1
Candida albicans[d]	CANAW	4932	949	0.192	40.3	18.7
Cryptococcus neoformans	CRYNJ	5679	1094	0.193	39.5	19.1
Giardia intestinalis	GIAIC	1181	211	0.179	25.5	6.0
Leishmania major	LEIMA	7858	1423	0.181	14.1	6.7
Monosiga brevicollis	MONBE	4184	775	0.185	43.9	16.4
Phaeosphaeria nodorum	PHANO	15023	2601	0.173	23.3	16.2
Plasmodium falciparum	PLAF7	1853	375	0.202	33.6	11.8
Schizosaccharomyces pombe	SCHPO	3541	602	0.170	62.3	24.7
Yarrowia lipolytica	YARLI	4222	862	0.204	57.9	22.7
Saccharomyces cerevisiae	YEAST	4811	926	0.192	39.7	17.2
Eukaryota (multicellular)						
Anopheles gambiae	ANOGA	9889	2390	0.242	33.8	17.5
Arabidopsis thaliana	ARATH	23989	6102	0.254	21.1	12.1
Bos taurus	BOVIN	19336	5432	0.281	46.5	31.9

Branchiostoma floridae	BRAFL	15318	3050	0.199	23.0	14.0
Caenorhabditis elegans	CAEEL	15719	5280	0.336	21.5	8.4
Canis familiaris	CANFA	18574	4805	0.259	45.0	34.7
Gallus gallus	CHICK	14226	3535	0.248	47.8	35.7
Ciona intestinalis	CIOIN	9346	1932	0.207	31.9	15.3
Danio rerio	DANRE	22138	5805	0.262	30.4	20.8
Dictyostelium discoideum	DICDI	8316	1793	0.216	25.2	8.9
Drosophila melanogaster	DROME	13582	3454	0.254	30.3	16.9
Homo sapiens	HUMAN	20221	5242	0.259	45.0	35.4
Ixodes scapularis	IXOSC	8470	1722	0.203	27.0	15.7
Macaca mulatta	MACMU	19771	4671	0.236	39.7	34.3
Monodelphis domestica	MONDO	18904	5168	0.273	40.6	28.7
Mus musculus	MOUSE	20457	6047	0.296	45.0	29.5
Nematostella vectensis	NEMVE	14935	2705	0.181	25.7	14.6
Neurospora crassa	NEUCR	6817	1241	0.182	35.4	22.8
Ornithorhynchus anatinus	ORNAN	14308	3247	0.227	31.7	23.4
Pan troglodytes	PANTR	18241	4551	0.249	46.5	37.3
Physcomitrella patens	PHYPA	13463	3126	0.232	25.6	13.1
Rattus norvegicus	RATNO	20502	5872	0.286	42.7	29.0
Schistosoma mansoni	SCHMA	4023	745	0.185	34.4	19.0
Sclerotinia sclerotiorum	SCLS1	8377	1576	0.188	34.2	18.7
Takifugu rubripes	TAKRU	17576	4480	0.255	36.0	24.3
Ustilago maydis	USTMA	3505	690	0.197	43.8	17.2
Xenopus tropicalis	XENTR	16021	4184	0.261	31.6	21.4

[a] Two of the 66 species, namely *Thalassiosira pseudonana* (THAPS) and *Trichomonas vaginalis* (TRIVA), were not found in OMA at the time of this analysis and were thus ignored.
[b] *Bradyrhizobium diazoefficiens* (BRADU) is in the EBI Reference Proteomes list, but was not found in OMA at the time of this analysis; *B. japonicum* (BRAJA) was used instead.
[c] Present in OMA as "*Neosartorya fumigata*" at the time of writing.
[d] The OMA code for the *Candida albicans* strain used was (CANAW) instead of the one in the EBI list (CANAL), not presently found in OMA.

Water-soluble OGs are on average 2.7 times larger than membrane-protein OGs in prokaryotes (Figure 43). Amongst the eukaryotes, the effect is also substantial but smaller in multicellular versus unicellular organisms; water-soluble OGs are on average 2.4 times larger than membrane-protein OGs in unicellular eukaryotes, whereas the factor decreases to 1.7 for multicellular eukaryotes (Figure 43; one-way analysis of variance: $F(2,61)= 21.07$; $p=1.1\cdot10^{-7}$; $\omega^2=0.149$).

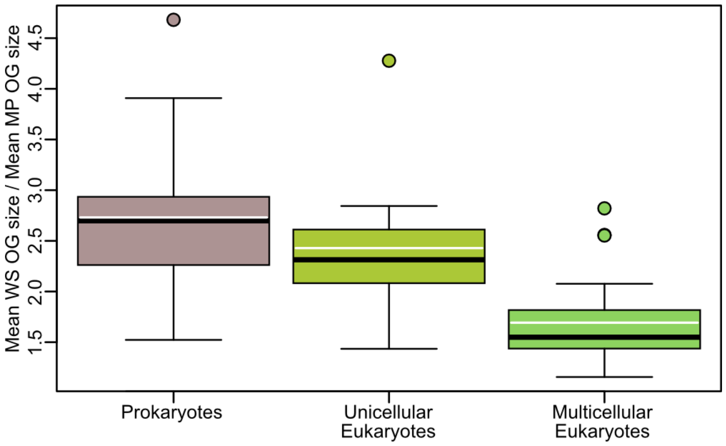


Figure 43. Water-soluble orthologue groups are substantially larger than membrane-protein groups, but the effect decreases as organismal complexity increases

The ratio of the mean orthologue group sizes of water-soluble over membrane proteins is always greater than 1 (i.e. each WS bar is always larger than its corresponding MP bar in Figure 42), but the effect decreases as cellular and organismal complexity increase, from prokaryotes to unicellular eukaryotes, to multicellular eukaryotes. Bold black lines represent the median, white lines the mean, boxes the inter-quartile range (IQR), and whiskers are R-package standard at a ±1.5*IQR threshold.

The findings in Figure 42 were confirmed with a protein-protein BLAST (blastp) search (Altschul et al. 1990) against the full non-redundant (nr) NCBI protein database, for each protein in the proteome of six species chosen from Figure 42. I picked well-annotated representatives of two distant clades from each domain, namely a euryarchaeon and a TACK-archaeon, a Gram-positive and a Gram-negative bacteria, and a unicellular and a multicellular eukaryotes. In all cases the mean number of blastp hits is lower for MPs than for WSs (Figure 44). These results consistently show that membrane proteins have fewer orthologues than water-soluble proteins across the tree of life. The fact that this is the case in all species studied suggests that an important evolutionary force is at play.

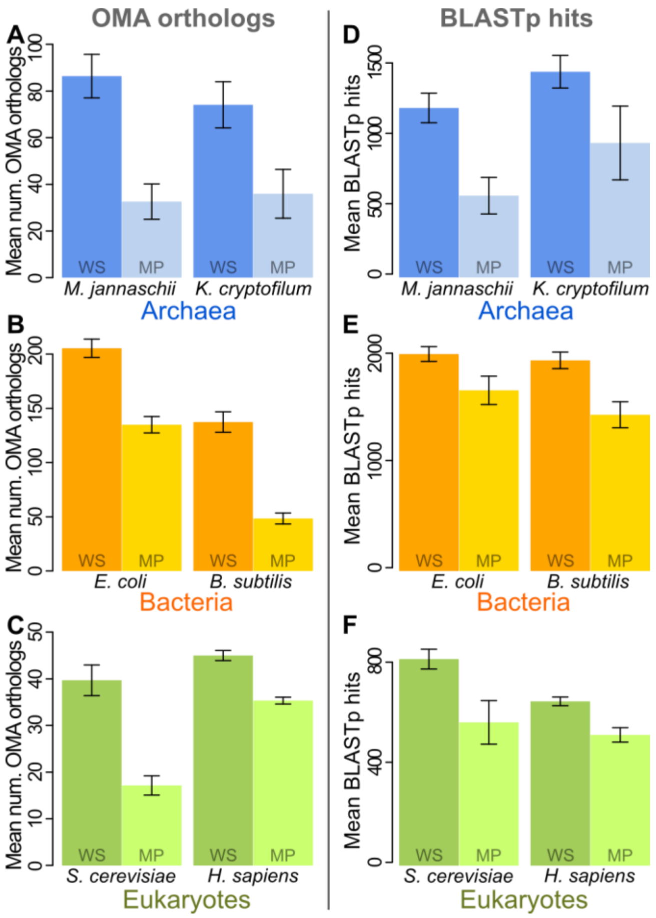


Figure 44. blastp search confirms lower homology of membrane proteins

A blastp search on the non-redundant (nr) NCBI protein database confirms the results in Figure 42 that membrane-bound proteins recover fewer homologs. Two well-annotated representative species from distant clades of each domain were chosen, namely a euryarchaeon and a TACK-archaeon, a Gram-positive and a Gram-negative bacteria, and a unicellular and a multicellular eukaryotes. Results from OMA in Figure 42 are repeated in (**A**, **B** and **C**), for comparison with blastp results in (**D**, **E** and **F**). Data presented as means ± 2SEM.

I performed a logistic regression on the OMA orthologue dataset to estimate the probability that a protein is membrane-bound as the number of clades sharing it increases. The number of clades was determined by choosing one orthologue from each clade at the sixth level of taxonomic differentiation according to the NCBI lineages (e.g. “Escherichia” for *E. coli*, and “Deuterostomia” for humans, see Methods in section 5.5). Of the 883,176 pre-computed OMA OGs, I picked the 228,148 that had representatives from at least three separate clades (to allow for later multiple sequence alignments). The results show that the probability of a protein being membrane-bound falls as the number of clades sharing it increases, both when considering the whole dataset, and for proteins shared exclusively within the archaea, bacteria, and eukaryotes (Figure 45). The more universal the protein, the less likely it is to be membrane-bound.

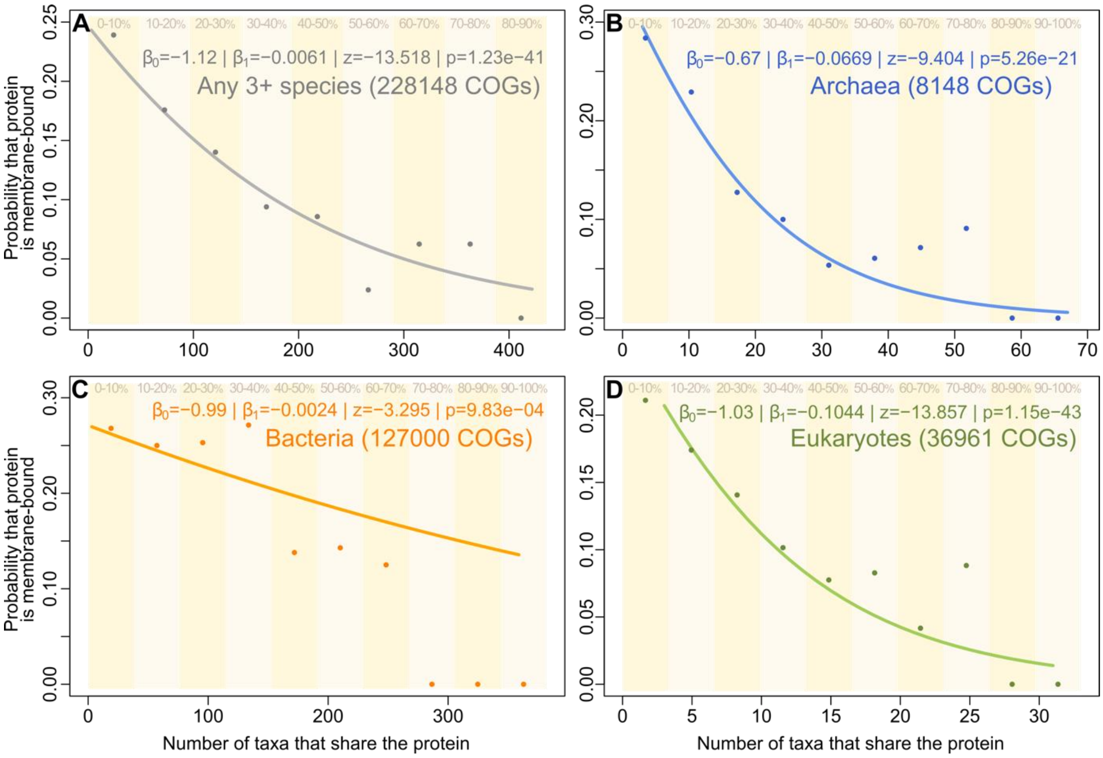


Figure 45. The probability of a protein being membrane-bound falls with wider distribution

(**A**) A logistic regression shows that the probability that a gene is a membrane protein falls significantly with increasing number of clades sharing it. The pattern remains when considering each of the three domains separately (**B**, **C** and **D**). The points and vertical stripes correspond to the proportions of MPs amongst genes shared by increasingly large numbers of clades, divided in 10% bins. No proteins retrieved were shared by over 90% of the 489 taxa in (A). In all cases the final bins have proportion zero, i.e. no highly shared proteins are membrane-bound. Note that the points and bins are provided for reference only: logistic regressions were performed on the individual orthologue clusters (i.e. the probability curves were derived independently. See Methods in section 5.5).

Since orthologue discovery depends on the successful detection of homologs using tools such as BLAST, the lower homology of membrane proteins I report could have two main causes (Figure 41). First, it is possible that membrane proteins evolve faster and hence their more divergent sequences are picked up less frequently by homology identification algorithms. Second, some of the absences may be true gene losses, such that the orthologues are not found because they are genuinely no longer there.

5.3.2 Faster evolution of membrane proteins and their outside-facing sections

To investigate whether the patterns above are due to membrane proteins having a higher divergence rate overall, I aligned the protein sequences of the 41,970 OGs shared by ten or more clades using MAFFT (Katoh and Standley 2013), computed the corresponding codon alignments using PAL2NAL (Suyama et al. 2006) and calculated the non-synonymous (dN) to synonymous (dS) rate ratio (ω=dN/dS) using the codeml program from the PAML suite (Yang 2007). The effect, although small, confirms previous reports on data sets with more limited phylogenetic ranges (Volkman et al. 2002; Julenius and Pedersen 2006; Sanders and Mittendorf 2011) that membrane proteins diverge more quickly than water-soluble proteins (Welch's t-test: t=14.08, df=14261.09; p=2.59·10^{-45}; r=0.12, Figure 46A) and this result was consistent across the three domains of life (archaea, bacteria and eukaryotes, Figure 46B-D).

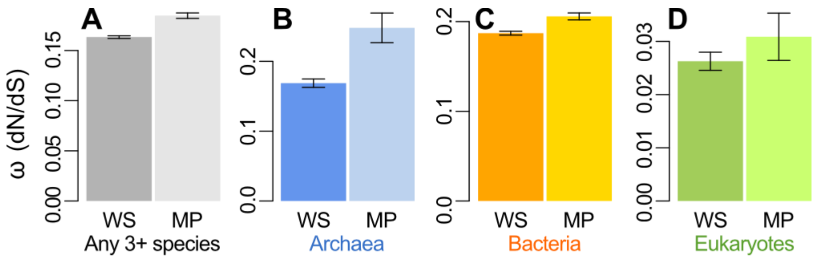


Figure 46. Purifying selection is weaker in membrane proteins

The ratio of non-synonymous to synonymous substitution rates (ω=dN/dS) is higher for membrane proteins in the full set of OMA OGs (**A**) as well as for the archaea (**B**), bacteria (**C**) and eukaryotes (**D**) separately, indicating that purifying selection is weaker on membrane proteins. Following the recommendations in the codeml manual (Yang 2007), only genes shared by 10 or more species were analysed.

While the TMHMM algorithm has been shown to infer trans-membrane helical regions with very high accuracy (Krogh et al. 2001), discerning the inside- versus outside-facing aqueous regions of TMH proteins is substantially more challenging. I therefore downloaded the full non-redundant set of sequences and annotations from the trans-membrane protein data bank (PDBTM, pdbtm.enzim.hu) (Tusnády et al. 2004), to assess the evolution of the three main regions of trans-membrane proteins: inside-facing aqueous, membrane-spanning, and outside-facing aqueous. Briefly, this database has annotations, where available, for the sub-cellular localisation of each amino acid in all membrane-protein structures deposited in the Protein Data Bank (PDB, www.rcsb.org) (Berman et al. 2000; Rose et al. 2015). I obtained homologs for each protein using blastp on the NCBI nr database, aligned the sequences with MAFFT, and sliced the alignments vertically to obtain the inside, membrane-spanning, and outside sections, plus an “aqueous” assemblage constructed by concatenating the inside and outside portions. I then built trees for each of these using FastTree (Price et al. 2010). See Methods for details. The mean of the branch lengths in each tree, which correspond to the number of substitutions per site, was used as an estimate of evolutionary rate. The results confirm that aqueous sections evolve faster than membrane-spanning ones (Figure 47; paired t-test: t=10.2109; df=371; p=1.40·10^{-16}; r=0.411). Amongst the aqueous sections, both of which evolve faster than the membrane spanning ones overall, the environment-facing sections evolve faster than the inside-facing ones (paired t-test: t=4.63; df=359; p=5.22·10^{-6}; r=0.237).

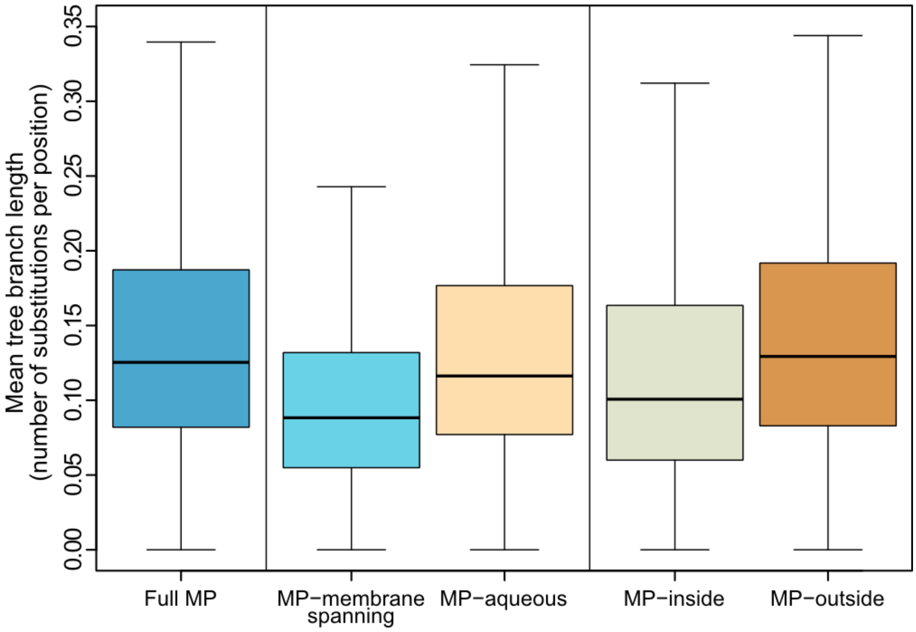


Figure 47. Evolutionary rates for sections of trans-membrane proteins annotated from PDB structures

For the 378 proteins studied from the PDBTM database, aqueous sections evolve faster overall than membrane-spanning sections. Splitting the aqueous sequences into outside- and inside-facing sections confirms that environment-exposed regions evolve faster than cytosol-facing ones. Values ranges as in Figure 44; outliers not shown.

To control for potential errors in the automatic annotations of PDBTM, I repeated the analysis by manually annotating the three main regions (inside, outside, and membrane-spanning) of twelve membrane proteins that are highly shared in OMA, including one outer-membrane beta-barrel porin and eleven trans-membrane helical proteins. The closest matching structural file was found by blastp search against the PDB subset on NCBI. The subcellular location of each amino acid residue was then assigned by inspecting the PDB structures against the information in the corresponding primary literature (Table 11). The sequences in the OMA OGs were aligned to the PDB sequence, alignments sliced and evolutionary rates estimated as described above. In all twelve proteins hand-annotated in this way, evolution occurs faster for outside-facing than for inside-facing aqueous regions (Figure 48).

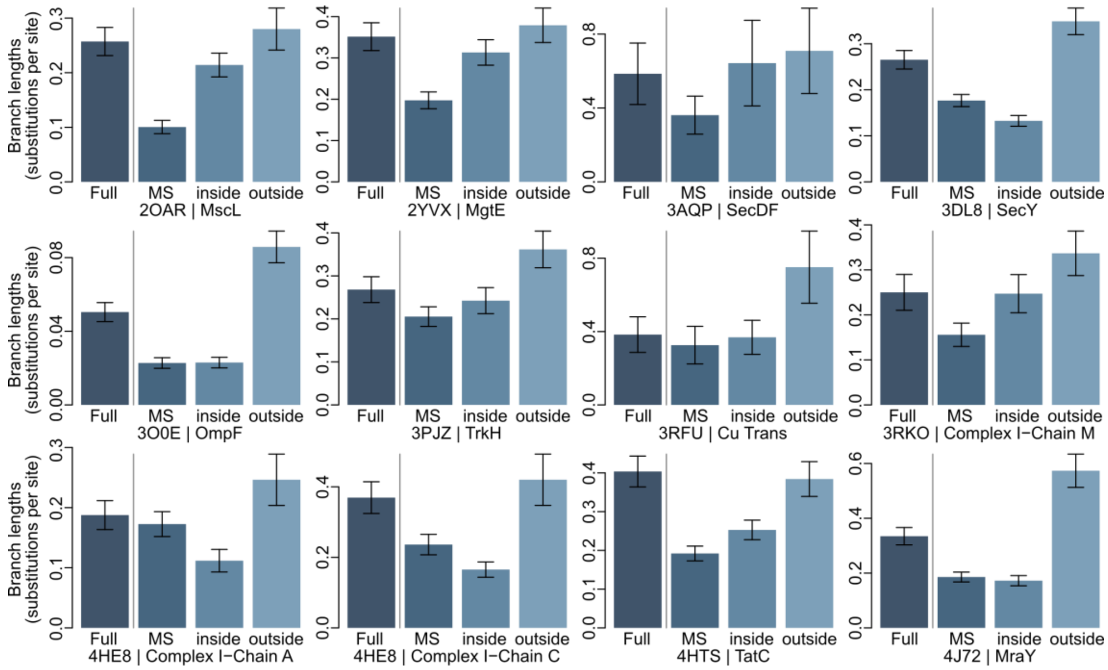


Figure 48. Outside-facing sections of membrane proteins evolve faster than inside-facing sections

In all proteins annotated by visually inspecting the PDB structure in relation to the original literature, evolution occurs faster for outside-facing than for inside-facing aqueous sections. This occurs both in outer-membrane and inner-membrane proteins. Four-character codes (e.g. “4HE8”) represent the PDB entry of the protein, followed by a short description of the protein name or function, as per the original literature. Full: the whole multiple-sequence alignment, without slicing. MS: membrane-spanning section (i.e. the lipid-exposed or “middle” section of the trans-membrane protein). Details of proteins and primary references in Table 11.

Table 11. Descriptions and primary references for PDB structures in Figure 48

PDB entry codes (3-D structures deposited on rcsb.org), descriptions, and references in the primary literature for the proteins in Figure 48.

PDB entry	Short description	Long description	Reference
2OAR	MscL	Mechanosensitive channel of large conductance	(Chang et al., 1998)
2YVX	MgtE	Magnesium transporter	(Hattori et al., 2007)
3AQP	SecDF	Translocon-associated membrane protein	(Tsukazaki et al., 2011)
3DL8	SecY	Bacterial protein translocation channel	(Zimmer et al., 2008)
3O0E	OmpF	Bacterial outer-membrane porin	(Housden et al., 2010)
3PJZ	TrkH	Potassium transporter	(Cao et al., 2011)
3RFU	Cu Trans	Cu-transporting PIB-type ATPase	(Gourdon et al., 2011)
3RKO	Complex I-M	Chain M of respiratory complex I	(Efremov and Sazanov, 2011)
4HE8	Complex I-A	Chain A of respiratory complex I	(Baradaran et al., 2013)
4HE8	Complex I-C	Chain C of respiratory complex I	(Baradaran et al., 2013)
4HTS	TatC	Twin arginine translocase receptor	(Ramasamy et al., 2013)
4J72	MraY	Polyprenyl-phosphate N-acetyl hexosamine 1-phosphate transferase	(Chung et al., 2013)

These findings are again widespread across the tree of life, and apply to multiple types of proteins. I note that these patterns hold true despite the fact that some aqueous proteins are exported from the cell and predictably evolve faster (Julenius and Pedersen 2006), whereas some membrane proteins, especially in eukaryotes, sit on organellar membranes (hence presumably evolve slower).

5.3.3 Membrane proteins have been lost more often within groups of closely related species

The results in Figure 46 suggest that the higher evolutionary rates of membrane proteins could, through divergence beyond recognition in tools such as BLAST, lead to the loss of homology reported above (Figures 42 and 44). But it is also possible that true gene losses have occurred in addition. I repeated the presence-absence analysis (Figure 42) on sets of predictably ancestral proteins within groups of closely related species and strains. If one or more species do not share a protein that is ancestral to the clade, I conclude that the encoding gene has been truly lost, under the assumption that, between closely related species, orthologues are unlikely to have diverged beyond recognition. I selected all prokaryotic clades with 10 or more closely related species in OMA (at the fifth taxonomic level of differentiation or higher according to the NCBI lineages). The number of orthologues in each cluster (OG) was then determined and filtered for OGs with at least half of the species within the group. I assumed that proteins shared by greater than half of the members of the clade were ancestral. Membrane proteins were then annotated using the consensus from the TMHMM predictions within each OG.

The results show that the mean numbers of species sharing each of these ancestral OGs are lower for membrane-proteins than for water-soluble ones across 31 of the 35 clades studied (Figure 49) (paired t-test: t=7.31; df=34; p=$1.81{\cdot}10^{-8}$; r=0.782).

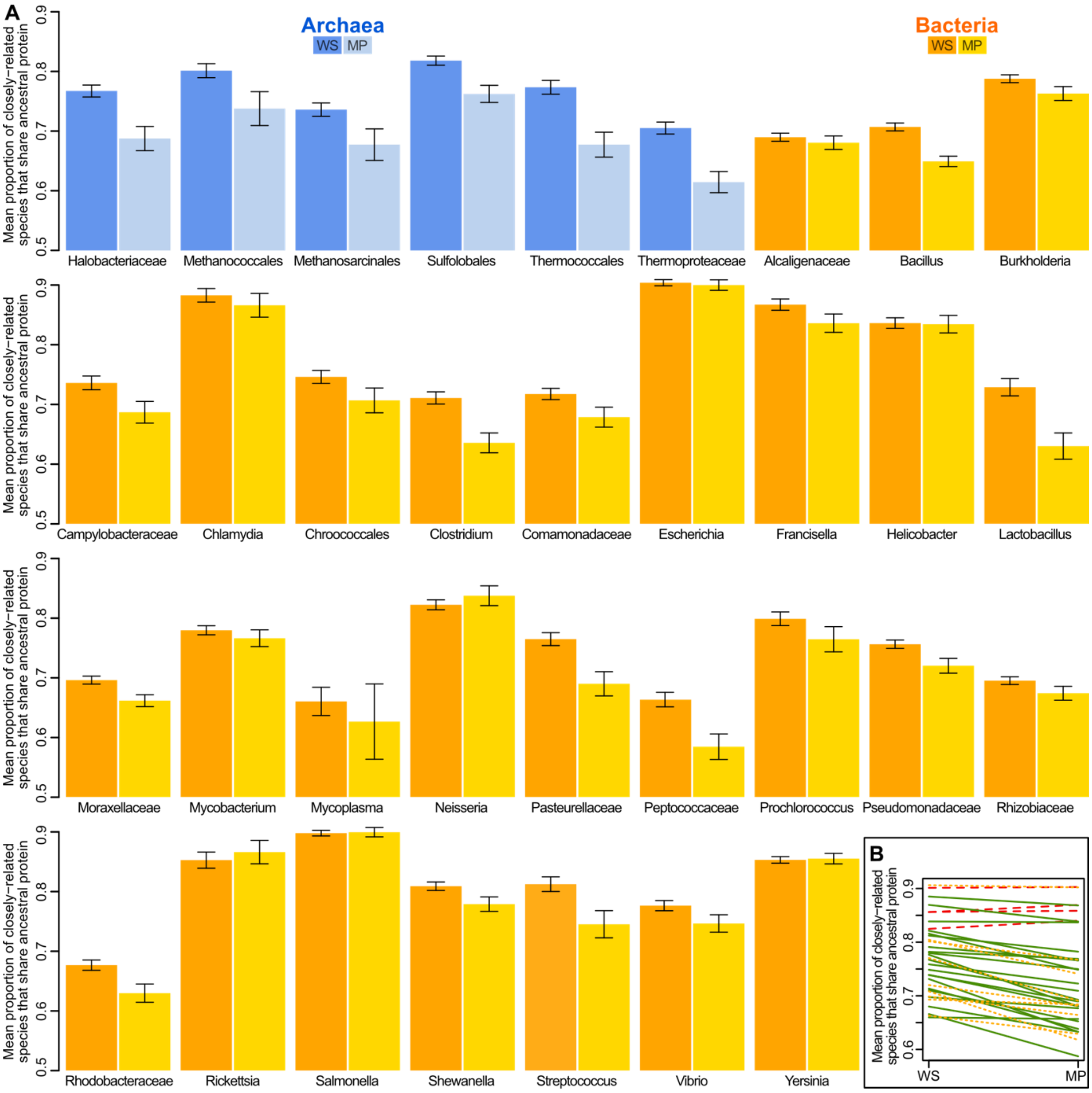


Figure 49. Ancestral membrane proteins have been lost more frequently

(**A**) Predicted ancestral proteins (defined as shared by at least half of the members of a clade), are shared by a smaller proportion of members in the clade if they are membrane proteins, for 31 of the 35 groups studied (exceptions are *Neisseria*, *Rickettsia*, *Salmonella*, and *Yersinia*). Dark shade: water-soluble; light shade: membrane proteins. First six pairs (blue) are archaeal clades, the remainder (yellow) are bacterial. Error bars: 2xSEM. (**B**) Values as in (A), paired and without error bars. Red-dashed: results with higher mean proportion of sharing species for MP than WS. Yellow-dotted: results with MP<WS but p-value over significance cut-off of 0.05 under a two-sample Welch t-test. Green-solid: statistically significant results.

That is, membrane proteins have been lost more often than water-soluble proteins within closely related clades, confirming that true gene losses can also account in part for the lower homology of membrane proteins reported above.

5.4 Discussion

I report that membrane proteins have fewer orthologues than water-soluble proteins across the entire tree of life (Figures 42-45). In principle this finding could be due to a higher evolutionary rate, which prevents sequence-search algorithms such as BLAST from detecting homologs beyond a given threshold. That is, since orthology detection ultimately relies on the successful identification of suitable homologues, and since membrane proteins have been reported to evolve faster than water-soluble proteins, they can be expected to cross a given detection threshold faster. This will lead to spurious loss of homology that would be confused for gene loss in databases such as OMA. Conversely, it is possible that some of the missing orthologues correspond to true gene loss, i.e. that homology detection algorithms fail to detect some of the genes because they are genuinely no longer present in the genomes. My results suggest that both mechanisms are at play.

First, I confirm that the evolutionary rates of membrane proteins are faster than for aqueous proteins, and extend these findings across the whole tree of life, and in each of the three domains of bacteria, archaea and eukaryotes independently (Figure 46). The evolutionary rates of membrane proteins are fastest in the outside-facing aqueous regions, comparatively slower in the inside-facing counterparts, and slower still in the transmembrane portions (Figures 47 and 48).

Second, the analysis of closely related species shows that predicted ancestral proteins are lost more frequently if they are membrane bound (Figure 49). This indicates that the lower homology of membrane proteins is not only due to divergence beyond sequence recognition, but also that true gene losses may have occurred. However, it is possible that closely related species may experience divergence beyond recognition (Haggerty et al. 2014), most significantly in short and rapidly evolving sequences, a possibility that must be evaluated further. One possible solution would be to consider only sequences above a certain length for the results in Figure 49. In addition, an analysis of the chromosomal vicinity of each gene could provide a clearer picture of orthology and help identify any true gene losses, but given the fluidity of prokaryotic genomes and the vast amount of data used here, this approach would be prohibitive at present.

The findings point to a general evolutionary principle: membrane proteins may evolve faster because they face stronger adaptive selection in changing environments, whereas cytosolic proteins are under more stringent purifying selection in the homeostatic interior of the cell (Figure 41). The outside-facing sections of membrane-spanning proteins are closely involved in adaptation to new environments and functions, and so are more likely to diverge over time than the cytosolic portions. As emerging species colonise novel environments or specialise in new tasks, the outside-facing sections are subject to stronger positive selection, while rate-limiting purifying selection prevails in the membrane-spanning and inside-facing portions (Figures 47 and 48). Novel or changing environments also render ancient membrane proteins useless, leading to loss over time, and accounting for the absences that observed even between closely related species (Figure 49). The hypothesis put forward here immediately suggests that this effect should be strongest in prokaryotes, weaker in unicellular eukaryotes (where intracellular membranes mean that membrane proteins can also face an internal homeostatic environment), and weakest in multicellular eukaryotes (where even extracellular proteins face a homeostatic environment provided by tissues and organs). This is indeed the case (Figure 43). Nonetheless, the difference in size of orthologue groups between membrane proteins and water-soluble proteins is substantial even in multicellular eukaryotes.

This broad evolutionary perspective provides a framework for interpreting a number of earlier findings that have proved difficult to generalise. Previous results show that water-soluble proteins secreted from the cell evolve faster than cytosolic proteins in mammals and yeast, and that the external portions of membrane proteins evolve faster than the internal domains (Julenius and Pedersen 2006). However, given the complexity of mammalian species, a focus on this taxonomic class does not lend itself to generalisations in terms of purifying selection or adaptation to changing extracellular environments. Similarly, the G-protein-coupled receptor superfamily is known to evolve faster in its extracellular portions than in the transmembrane and cytosolic regions, but this has again been interpreted in terms of particular functional and structural constraints (Tourasse and Li 2000; Lee et al. 2003). In Gram-negative bacteria, degradation of xenobiotic toxic substances occurs in the periplasmic space (Kawai 1999; Nagata et al. 1999), making evolutionary pressure stronger on the external regions than in the homeostatic interior. Signal peptides have been shown to evolve rapidly in both prokaryotes and eukaryotes, pointing to positive selection on these secretory membrane-targeting fragments (Li et al. 2009). Finally, parasitic interactions can promote the rapid evolution of membrane proteins, especially the external loops involved directly in antigen interactions (Volkman et al. 2002; Plotkin et al. 2004). Parasite membrane proteins face positive selective pressure from recognition by the host, but these red-queen dynamics have not been extended beyond parasite-host interactions. I argue that each of these specific instances can be generalised for membrane proteins as a class across the tree of life. When interpreted into a comprehensive context, all these observations point to faster evolution outside the cell in response to changing environments or functions.

I have not considered the effects of horizontal gene transfer (HGT), a major force in microbial evolution, as the unequivocal detection and ecological significance of ancient HGT events is still a hotly debated topic (Philippe and Douady 2003; Dagan and Martin 2007; Puigbò et al. 2014; Akanni et al. 2015; Katz 2015; Koonin 2015; Ravenhall et al. 2015; Soucy et al. 2015). Horizontally transferred genes tend to be integrated at the periphery of metabolic networks, while genes at the core tend to be more evolutionarily conserved (Pál et al. 2005). At the level of cellular gene networks, extracellular proteins could be considered peripheral, while intracellular proteins are more central, and so more conserved (Julenius and Pedersen 2006). However, the fact that membrane proteins evolve faster in their outside portions and more slowly on the inside is not consistent with the idea that membrane proteins evolve faster simply because they are peripheral to gene networks, but rather because selection operates differently outside the cell. I have also ignored the effect of exported water-soluble proteins, which as noted evolve faster than cytosolic proteins and even than the external sections of membrane proteins in mammals and yeast (Julenius and Pedersen 2006). Annotation deficiencies across the rest of the tree of life forced me to neglect these differences between water-soluble proteins inside and outside the cell. Conservatively, however, removing the relatively fast-evolving secreted proteins should magnify the difference in evolutionary rates between water-soluble proteins and membrane proteins, reinforcing the findings. In fact, the observation of true gene losses should be echoed by secreted proteins.

I conclude that adaptation to novel environments and functions underlies the lower homology of membrane proteins across the tree of life. Life is defined by its cellular nature: the inside of a living cell separated from its environment by an organic membrane. Cells must constantly interact with varying environments, while maintaining tight internal homeostasis. The interactions between the inside and outside of the cell are largely mediated by membrane proteins, so elucidating their evolution is central to understanding the origins and evolution of life. For the same reasons, membrane proteins have great medical importance. Over half of all known drug targets are membrane proteins, so these findings may help to explain why the progression of new drugs from animal models into human trials is so often unsuccessful (Holmes et al. 2011; Denayer et al. 2014). The results are also of practical importance in phylogenetics: if membrane proteins are less than half as likely to be conserved widely across the tree of life, then homology searches will often be confounded, as could molecular clocks. Faster evolution outside the cell makes simple intuitive sense, but the strength of this signal across the whole tree of life elevates what has been seen as an interesting sporadic pattern into a general principle of evolution.

5.5 Methods

The full set of orthologue groups (OGs) from the OMA database was downloaded from the OMA server at www.omabrowser.org/export, September 2014 release.

This set of 883,176 OGs includes multiple orthologues shared by repeated species (e.g. multiple strains of *Escherichia coli*), so, as a strategy to avoid oversampling in the phylogenetic distribution analysis of Figure 45, one gene was chosen per clade at the sixth level of taxonomic differentiation according to the NCBI taxonomy browser. Only OGs with 3 or more different such clades were kept. This left a total of 228,148 OGs. As an example, the full NCBI taxonomic lineage for *E. coli* is

1.Bacteria > 2.Proteobacteria > 3.Gammaproteobacteria > 4.Enterobacteriales > 5.Enterobacteriaceae > 6.Escherichia > 7.*Escherichia coli*, from where the sixth taxonomic level is "Escherichia"; similarly, the equivalent for humans is "Deuterostomia" from: 1.Eukaryota > 2.Opisthokonta > 3.Metazoa > 4.Eumetazoa > 5.Bilateria > 6.Deuterostomia. When multiple genes were found for the same clade, as defined above, the representative was chosen from well annotated species (e.g. *Escherichia coli*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Methanosarcina acetivorans*), where available, or at random.

Membrane proteins were annotated using the predictions of the TMHMM 2.0c algorithm (Krogh et al. 2001). This algorithm predicts only trans-membrane alpha helical proteins, so for Figure 45 undetected genes marked as membrane porins or integral membrane proteins in their descriptions in OMA were further annotated as MPs. Finally, Gene Ontology annotations, where available, were also used to identify MPs. All other proteins were assumed to be WS. This additional classification of MPs produced only minor changes (data not shown), so to ensure reproducibility and avoid unpredictable effects of the sparse annotations, in all other figures MPs were annotated using only the predictions of the TMHMM algorithm.

The species in the list of reference proteomes were obtained from EMBL/EBI at www.ebi.ac.uk/reference_proteomes, and the full proteomes of six representative species were procured from the same website.

Since I classify each protein in a binary fashion as either WS or MP, the logistic regressions of Figure 45 were produced by fitting a quasi-binomial model to the type of protein (0 for WS and 1 for MP), as predicted by the number of orthologues in the cluster (i.e. the size of the OMA OG, or more simplistically the number of clades that have an identifiable orthologue of the protein in OMA). The points were produced entirely independently by binning the data in 10% increments in terms of how many clades share each protein, or size of the OG. That is, for Figure 45A, the total number of clades is 489, so proteins in the first bin are shared by anything between 3 and 49 clades. The point represents the proportion of those proteins that are MPs.

The complete non-redundant (nr) protein database was downloaded from NCBI on 18 June 2015. The blastp algorithm was run locally for each protein in each of the six selected proteomes in Figure 44. Significant blastp matches were defined as having an e-value lower than 10^{-10} and a query coverage of at least 70%; when multiple hits were found for the same species, only the highest scoring hit was kept to avoid oversampling. blastp was also used to detect orthologues in Figures 47 and 48.

The entire non-redundant set of PDB structure sequences and annotations used in Figure 47 was downloaded from pdbtm.enzim.hu (Tusnády et al. 2004). This dataset is constantly updated to include all PDB structures for membrane proteins in the PDB database, and parse the files into annotations for the subcellular localisation of each amino acid in each of these structures, where the information is available (often the crystal structures have unresolved portions, notably loops, and in other cases the researchers do not report whether an aqueous section is inside- or outside-facing, in which case I ignored the protein altogether). At the time of this analysis there were 576 non-redundant integral membrane proteins in PDBTM (496 annotated as alpha helices and 80 as beta barrels), 378 of which unambiguously specified inside- versus outside-facing aqueous regions. To slice (or split vertically) the multiple-sequence alignments (MSAs) used in Figures 47 and 48 into the membrane-spanning, inside, outside and aqueous (which includes both inside- and outside-facing) sections, the PDBTM annotations (Figure 47) or the hand-annotated positions (Figure 48) for the reference PDB protein sequence were used to establish the sub-cellular location of each amino acid. Each position was then sliced as described in the example below:

```
                                    0        10        20
                                     12345678901234567890
i/m/o annotations of PDB structure   iimmmoomm--mmmiiiii-
Amino acid sequence of PDB structure ABCDEFGHI--JKLMNOPQ-
Amino acid sequence of orthologue 1  --CDEFWHIWWJXLMNOPQW
Amino acid sequence of orthologue 2  -BCDEFGHIXXJXLMXOPQX
Amino acid sequence of orthologue 3  ABCDEFXHI--JXLMNOPQX
Amino acid sequence of orthologue 4  ABC-ZFXHIZ-JXLMNZPQZ
```

where *i*, *m*, and *o* represent that the amino acid is annotated as inside, membrane-spanning, or outside (respectively), either in the PDBTM database for Figure 47 or the hand-annotated positions, or in the twelve annotations done by directly inspecting the PDB structure against the primary literature for Figure 48. In the example above, positions 1-2, 15-19 are inside; 6-7 are outside; 3-5, 8-9, 12-14 are membrane-spanning; and 10-11, 20 are ignored. The aqueous portions were constructed by concatenating the inside and outside alignments.

Python, BioPython (Cock et al. 2009) and R (R Core Team 2014) were used widely in the calculations and analyses in this chapter.

6 DISCUSSION, OPEN QUESTIONS, AND CONCLUSIONS

Throughout this thesis I have discussed the role of membranes and membrane bioenergetics in some of the major transitions in evolution, and in shaping the relationships between the three domains of life.

The very antiquity of the events discussed here hinders the achievement of absolute certainty about most of the proposed evolutionary scenarios. It is nevertheless possible and fruitful to gather and analyse data and develop models in order to determine which predictions hold up better to scrutiny. I have attempted to do so here.

As with any scientific endeavour, I have had to take certain views for granted. Some of the views I have assumed as true regarding a number of evolutionary events are still controversial, often bitterly so. For example, in Chapter 2 I assumed an autotrophic origin of life in alkaline hydrothermal vents, while in Chapter 4 the endosymbiotic origin of eukaryotes from a bacterium into an archaeon was considered to be true. I use these assumptions as starting points to test the respective models.

The study of evolution is the quest for understanding our origins (Maynard Smith and Szathmáry 1997, p. xiii), and as such it is a worthy goal on its own merits. But the study of evolution can transcend this purely philosophical role, and it often does. General and wide-ranging findings such as those in Chapters 3 (which discusses a general biochemical explanation for the origin of homochirality) and 5 (which highlights the different evolution of membrane-bound versus water-soluble proteins) can serve as an example of the potential of early-evolution studies to transcend their intrinsic blue-skies nature and provide useful knowledge to biology and the life sciences as a whole. Most notably, it is a well-known fact that the rates of success when progressing drugs from animal models into humans is excruciatingly low. Understanding why and how protein evolution occurs faster in the outside-facing sections of membrane proteins, which constitute over half of drug targets, could help tackle this problem in the future by highlighting the differences in membrane proteins between humans and animal models. It is my modest intention that some of the work in this thesis can serve as yet another example that studying early evolution can produce dividends beyond the acquisition of knowledge. This on its own, however, would more than suffice as a justification to pursue it (and to fund it).

6.1 The chapters of this thesis and the tree of life

Throughout this document I have focused chiefly on the relationships between the three domains of life, and how the tree (or ring) that links them has been shaped by some of the major transitions in evolution, all the while considering the roles that membranes, their embedded proteins, and the disequilibria across these have played throughout. The intended contribution of each of the four research chapters in this thesis to the understanding of the shape of the tree of life is summarised in Figure 50.

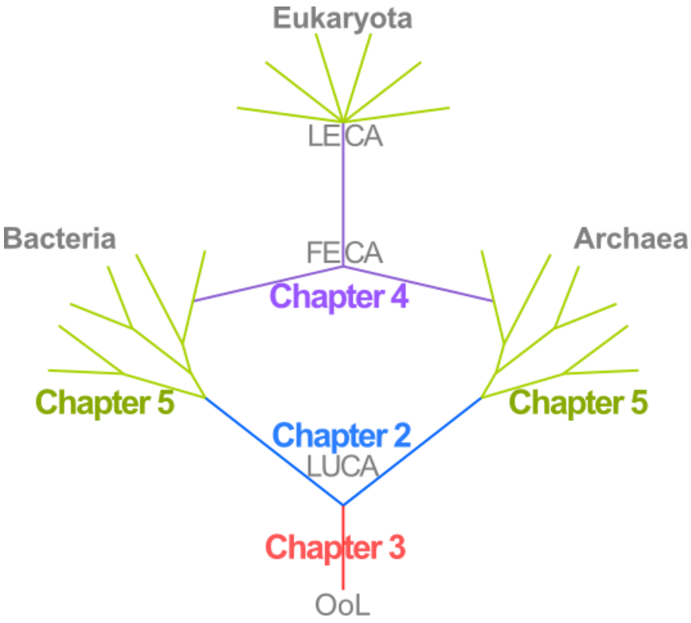


Figure 50. How the chapters in this thesis fit into the tree (or ring) of life

Chapter 2 dealt with the deepest branching in the tree: the divergence of archaea and bacteria. Chapter 3 followed this into a general discussion of homochirality and why only one enantiomer was favoured in the very early steps of evolution leading from the origin of life (OoL) to a last universal common ancestor (LUCA) that was completely homochiral in its D-sugars and L-amino acids. Chapter 4 returned to the differences between archaeal and bacterial membranes in the symbiotic origin of eukaryotes, and discussed why bacterial membranes were favoured in the transition from the first (FECA) to the last (LECA) eukaryotic common ancestors. Finally, Chapter 5 dealt with adaptation to new niches and speciation in general.

Again, and unless otherwise noted, my use of the expression “tree of life” throughout this thesis simply alludes to the relationships between species and domains, regardless of whether or not they are literally tree-like, i.e. pure vertical inheritance with modification from a common ancestor, as per Darwin’s original postulation in the *Origin* (Darwin 1859).

All the work presented here is in the general spirit set out by Mitchell (1961) in his chemiosmotic theory, which proposes an understanding of life as characterised by disequilibrium across membranes.

6.2 A recount of the findings and their implications

In **Chapter 2** I analysed the deepest split in the tree of life, namely the origin of archaea and bacteria from the last universal common ancestor (LUCA), and in particular the differences in their membranes and what these differences can indicate about the membranes of LUCA itself. I argue that the most crucial of the differences between the membranes of archaea and bacteria is also the subtlest: the dichotomy between the archaeal *sn*-glycerol-1-phosphate backbone versus the bacterial enantiomer *sn*-glycerol-3-phosphate. Assuming that life had an autotrophic start in an alkaline vent, and that LUCA depended on the spontaneous geologically generated proton gradients that these provide, I performed mathematical modelling that shows that early membranes had to be leaky. If attempting the evolution of modern membranes, the (proto)cell would have insulated itself into equilibrium. In terms of phospholipids this means that neither of the two enantiomers of glycerol-phosphate could be present. So, the lipids of LUCA would have been simpler amphiphiles such as single fatty acids (as opposed to the two fatty acids bound to a glycerol-phosphate backbone in modern bacterial and eukaryotic membranes; see Figure 17 in page 77).

However, the theory for the origin of life discussed in Chapter 2 is not universally accepted (Mulkidjanian et al. 2012; Le Page 2014; Deamer and Georgiou 2015). Importantly, the autotrophic alkaline hydrothermal vent theory depends on achieving high concentrations of synthesised products in what ultimately is an open system. The spontaneous self-assembly of amphiphiles into vesicles and the thermal gradients that occur in the pores of the vents provide a resolution through cycling and concentrating organics by several orders of magnitude (Baaske et al. 2007; Herschy et al. 2014). However, it has been argued that, while these processes works well in dilute aqueous solutions, results should be considerably less favourable in oceanic waters with their highly saline composition (up to 0.5 M NaCl in modern oceans), which adds to the considerable concentrations of divalent cations (10 mM Ca^{2+} and 54 mM Mg^{2+}), and predictably higher in alkaline vents, which are rich in both Mg^{2+} and Ca^{2+} (Deamer and Georgiou 2015). The problem is that NaCl tends to inhibit the spontaneous assemblage of membranes from amphiphiles such as simple fatty acids (Monnard and Deamer 2002); and similarly, even millimolar concentrations of divalent cations cause fatty acid vesicles to precipitate (Szostak et al. 2001), directly contradicting the model proposed here. These negative predictions have recently led some leading researchers in the field to dismiss the alkaline vent scenario in general (Damer and Deamer 2015; Deamer and Georgiou 2015; Damer et al. 2016) and the work discussed in Chapter 2 specifically (see comments by Prof. Jack Szostak in Le Page 2014), However, none of the relevant negative experiments mentioned have been done under pressure, or under the full chemical conditions of alkaline vents; and, significantly, recent preliminary results from two independent labs seem to suggest that forming organic membranes by spontaneous self-assembly of amphiphiles is indeed possible under alkaline vent conditions (Prof. Dieter Braun, personal communication). The evidence shows that these amphiphiles not only self-assemble, but that they can hold simulated pH gradients.

A potential reason for confusion in the work of Chapter 2, as published in the literature (Sojo et al. 2014) is my use of the term “leaky”. This is specifically used with relation to protons, which can cross the membrane by mechanisms different from those of other ions and molecules. Namely, protons can permeate by fatty-acid flip-flop which, as described previously, implies a fatty acid being protonated on the acidic side and then flipping to the alkaline side, where the proton is released. This mechanism is not available to Na^+ or other ions, let alone nucleotides, amino acids, or other molecules. The membranes I propose here would therefore serve as biological boundaries, much in the same way that modern phospholipid membranes do now, but with the advantage of allowing exploitation of the natural proton gradient in alkaline hydrothermal vents.

The predictions and conclusions of Chapter 2 are summarised in Figure 51.

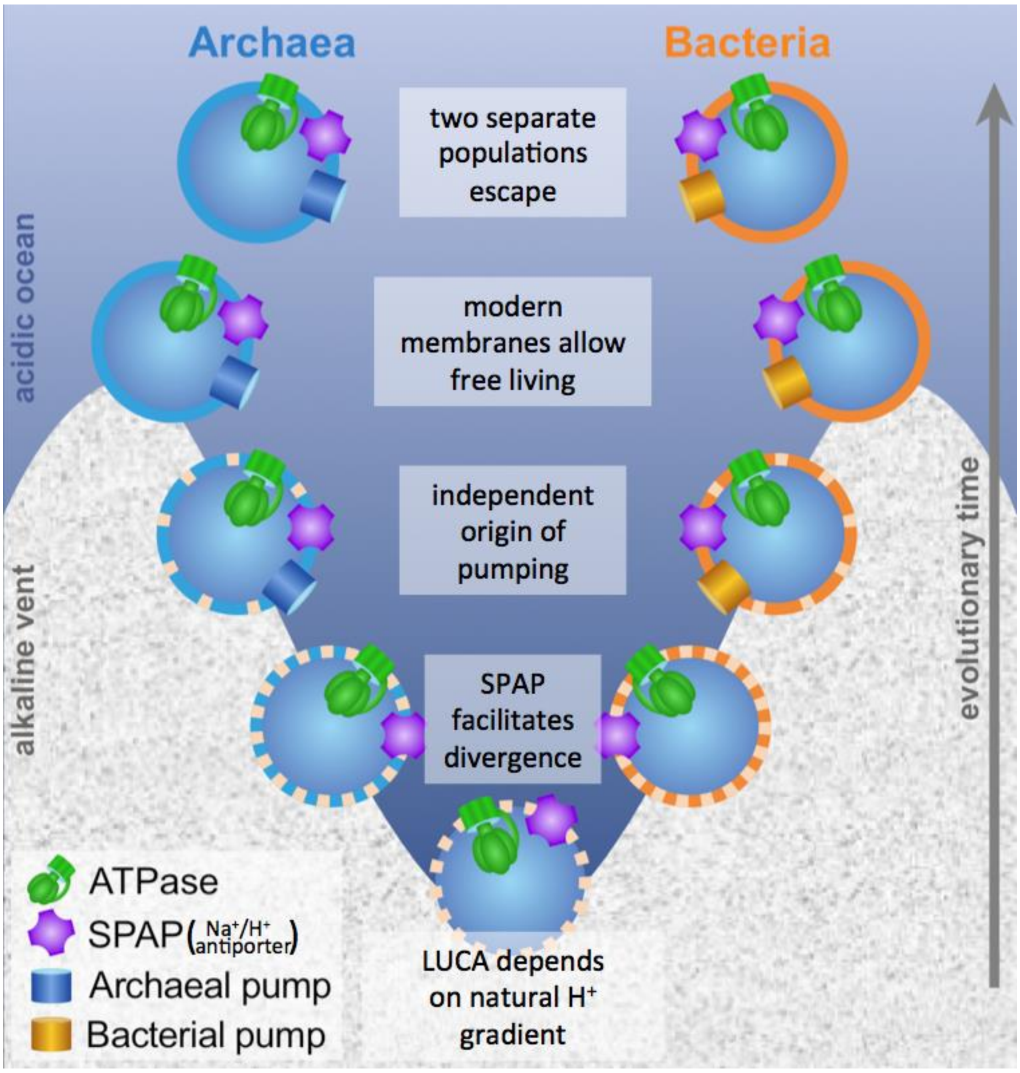

Archaea
Bacteria
two separate populations escape
modern membranes allow free living
independent origin of pumping
SPAP facilitates divergence
LUCA depends on natural H+ gradient
acidic ocean
alkaline vent
evolutionary time
ATPase
SPAP (Na+/H+ antiporter)
Archaeal pump
Bacterial pump

Figure 51. Chapter 2: The independent origin of modern membranes and pumping in archaea and bacteria

From bottom to top: starting with a LUCA that had a fully functional ATPase, the model shows that a sodium/proton antiporter (SPAP) would facilitate spread into areas of the vent that had weaker gradients. This may have also involved the differentiation of certain aspects of membranes, but the calculations show that membranes had to remain leaky to H^+. After a suggested independent origin of pumping in archaea and bacteria, the development of glycerol-phosphate backbones followed. Archaea and bacteria would do this independently as well, in turn explaining the opposite stereochemistries. With fully functioning pumping and H^+–impermeable modern membranes, the two types of cells were ready to escape the vent. Everything that lives on Earth today is a descendent of the two cells (or populations of cells) that escaped in this way.

Another challenge that can be brought against the idea of the cellular ancestor in an alkaline vent suggested here lies in electrochemical equilibration due to the influx of charges that are not neutralised, i.e. a Donnan equilibrium (Nicholls and Ferguson 2013): as a proton flows in, a positive charge is transferred into the cell, therefore opposing further transfers of positively charged ions. This would quickly render the system unworkable. However, the model I suggest escapes this scenario by providing electrical (as well as chemical) neutralisation towards the alkaline side, by either importing negatively charged OH^- ions into the cell, or spontaneously extruding the incoming positive H^+ ions into the alkaline fluid on the opposite side.

I conclude that the evolution of membrane-insulating glycerol-phosphate backbones had to wait until after the evolution of active ion pumping, to which the Na^+/H^+ antiporter (SPAP) was a necessary pre-adaptation. Archaea and bacteria would develop their glycerol-phosphate backbones independently (after the also independent origin of pumping), which explains the opposite stereochemistries: archaea developed the catalysis of dihydroxyacetone phosphate from one side of the planar molecule, bacteria from the other. Chemically speaking this was a 50:50 chance. Biologically speaking, however, the probability of developing either the *S* or *R* products was not simply 50:50, it depended on the catalysts that were available, as I discuss in the follow-up chapter, summarised below.

In **Chapter 3** I identified the *dual homochirality* of glycerol-phosphate backbones between archaea and bacteria as a fortunate event from the perspective of an evolutionary biologist, in that it sheds light on the evolution of homochirality in general. Hypotheses for the origin of this trait abound in physics and chemistry, from polarised light in interstellar radiation to stereo-selective interactions on catalytic clays. The case of the membrane-lipid glycerol-phosphate backbones described in Chapter 2, however, has received less attention, in spite of being arguably more interesting. Although both enantiomers are prevalent in extant life, the archaea and the bacteria are both exclusively homochiral, each domain having singularly picked one of the two enantiomers and never having been observed to use the other. I argue that this case of dual homochirality clarifies the evolution of single-handedness in life's molecules as a whole. Although both enantiomers are formed from the same substrate, dihydroxyacetone phosphate (DHAP), the respective synthesising enzymes, *sn*-glycerol-1-phosphate dehydrogenase (G1PDH) in archaea and *sn*-glycerol-3-

phosphate dehydrogenase in bacteria (G3PDH), are entirely unrelated. In terms of the reactions they catalyse, however, both belong to the E.C. 1.1.1 supergroup of NAD(P)H-dependent OH-dehydrogenases. Interestingly, all of these OH-dehydrogenases are known to be stereospecific, belonging to either the pro-*R*, or the pro-*S* kinds, aptly named with regards to the exclusively stereospecific nature of the reactions they catalyse. Figure 52 (a repetition of Figure 26) shows that all the members of the phylogenetic family of G1PDH are pro-*R* enzymes, while all in the G3PDH family are pro-*S* enzymes.

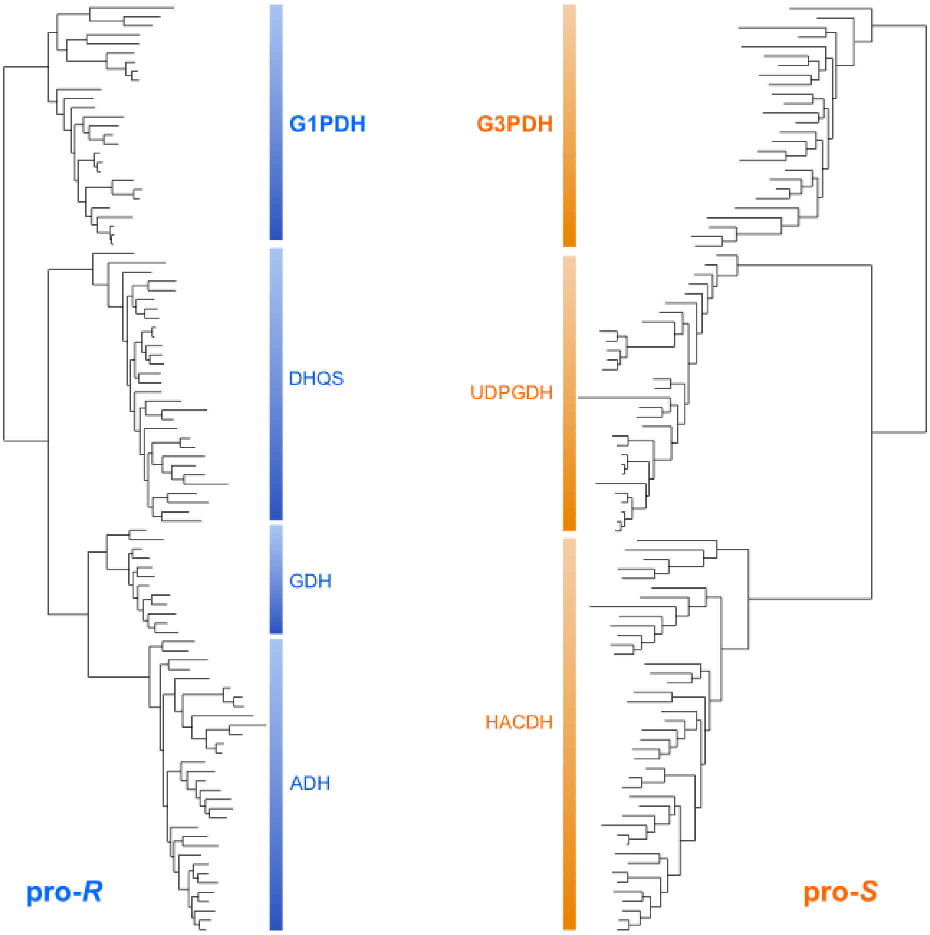


Figure 52. Chapter 3: all proteins in a phylogenetic family share the same chirality

If an ancestral enzyme or proto-enzyme was left-favouring, so will be all of its descendants, and vice versa. Homochirality is no great mystery, it is intrinsic to biochemical catalysis. Original tree by Peretó et al. (2004), and stereochemical classifications by You et al. (1978).

This brings the discussion to one of ancestry: self-evidently, the ancestor of G1PDH must have been a pro-*R* enzyme, and the G3PDH ancestor must have been pro-*S*. It seems clear that the reaction catalysed by these enzymes is stereospecific by nature; it can only proceed either exclusively one way or exclusively the other, never both. I suggest that this is general to enzymatic catalysis. This dual case of homochirality suggests that single-handedness is indeed the simplest evolutionary scenario and no prebiotic physical or chemical mechanisms need be invoked to explain it; the answer lies within biology itself.

In **Chapter 4** I returned to the differences between archaeal and bacterial membranes, this time with regards to the origin of the eukaryotic cell. Modern phylogenetic evidence strongly supports the theory that eukaryotes arose from the endosymbiosis of a bacterium into an archaeon. If this was the case, then the first eukaryotic common ancestor (FECA) must have had an archaeal plasma membrane and proto-mitochondrial bacterial membranes. Yet all known modern eukaryotes have exclusively bacterial membranes in their boundary to the exterior as well as in their internal vesicles, nucleus, and organelles, including the Golgi apparatus, peroxisomes, the smooth and rough endoplasmic reticula, and the mitochondrion. It is therefore reasonable to infer that the last eukaryotic common ancestor (LECA) had lost the archaeal membranes of its forebear.

Eukaryotes kept redundant sets of genes for many tasks, including building the ribosomal small and large subunits, ribosomal proteins, tRNAs, DNA replication, transcription, and ATP synthases, all of which are dually present in extant eukaryotes and have both an archaeal and a bacterial ancestry depending on whether it is the cytosolic or the mitochondrial version that is being described, respectively. Regardless of the origin, most of the encoding genes are now kept in the nucleus, which predictably was built upon the original archaeal chromosomal material. In spite of these many redundancies, archaeal phospholipids were entirely lost from the membranes. It is unlikely that this was due to pressures for genome reduction[ix]. Instead, a structural explanation seems more likely: heterotypic (or “hybrid”) membranes, although viable, were less stable than homotypic (or “pure”) ones, such that there was a pressure in FECA to get rid of one of the two copies, and by the time LECA arose only one of the two types of membrane remained. Why were the bacterial

[ix] In fact, isoprene synthesis and ether linkages are observed in modern eukaryotes.

phospholipids selected over their archaeal equivalents? I argue that the reason was bioenergetic: as the mitochondrion became specialised as the powerhouse of the eukaryotic cell and energy production came to rely increasingly upon it, the physiological adaptation of its bioenergetic proteins to the bacterial membrane would have become correspondingly indispensable. Replacing these membranes with archaeal ones would have led to decreased fitness, so the bacterial genes had to be kept, and the archaeal ones were lost instead (Figure 53).

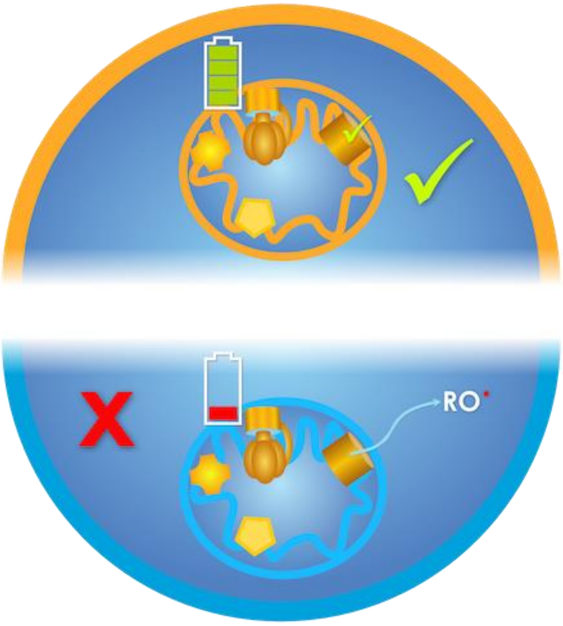


Figure 53. Chapter 4: Adaptation of mitochondrial bioenergetic proteins to the bacterial membrane underlies the prevalence of bacterial over archaeal phospholipids in the origin of eukaryotes

FECA would have had an archaeal plasma membrane. I argue that bacterial phospholipids were selected over their archaeal counterparts because changing the mitochondrial membranes to archaeal ones would have led to loss of energy and deleterious behaviour of membrane proteins, potentially including reactive-oxygen (RO*) species leakage.

I tested this hypothesis by demonstrating that horizontal gene transfers between the two prokaryotic domains are less likely if the gene encodes a membrane protein; that is, it is easier for species of one of the prokaryotic domains to pick up a protein from the other if the protein is not membrane-bound. I argued that this is because membrane proteins would have to sit on a foreign membrane, whereas cytosolic proteins sit on a more familiar aqueous medium. In addition, I provided computational-chemistry evidence that shows that membrane proteins from one domain behave differently in membranes with archaeal and bacterial glycerol-phosphate backbones, and specifically that the energy of a protein is lower (more optimal) in the phospholipid membranes of its own domain. These results shed light on one of the many unresolved questions in the origin of the eukaryotic cell.

I concluded the research with **Chapter 5**, where I discussed the roles of membrane proteins in adaptation in general. I reported that membrane proteins have on average fewer detectable orthologues than water-soluble proteins, and that this pattern remains across the tree of life although, interestingly, the effect is largest for prokaryotes, somewhat smaller for unicellular eukaryotes, and smaller still for multicellular eukaryotes. I analysed 64 species, and found the same pattern in all 64: membrane proteins always have considerably fewer orthologues (on average less than half in most species). This could be due to two main causes: either detection algorithms such as BLAST fail at finding the homologues, and/or the absences are true gene losses (Figure 54).

I confirmed previous reports in mammals, yeast and parasites, that membrane proteins evolve faster overall than water-soluble proteins, which supports the first scenario, and extended this to a general principle across all three domains of life. Given a certain threshold for homologue detection in BLAST, the faster-evolving membrane proteins will tend to cross it sooner than the more sequence-conserved water-soluble proteins. I further explored the noticeable observation that their aqueous sections evolve faster than the membrane-spanning ones and, more importantly, that amongst the aqueous sections the outside-facing ones evolve faster than their inside-facing counterparts. This observation is widespread for different types of proteins across the tree of life, and I argued that it suggests the operation of a major evolutionary force: the outside evolves faster because it is the outside that most frequently and strongly interacts with the environment. Adapting to new functions or environments implies that the outside experiences higher selective pressures, while the inside is subject to stronger homeostasis.

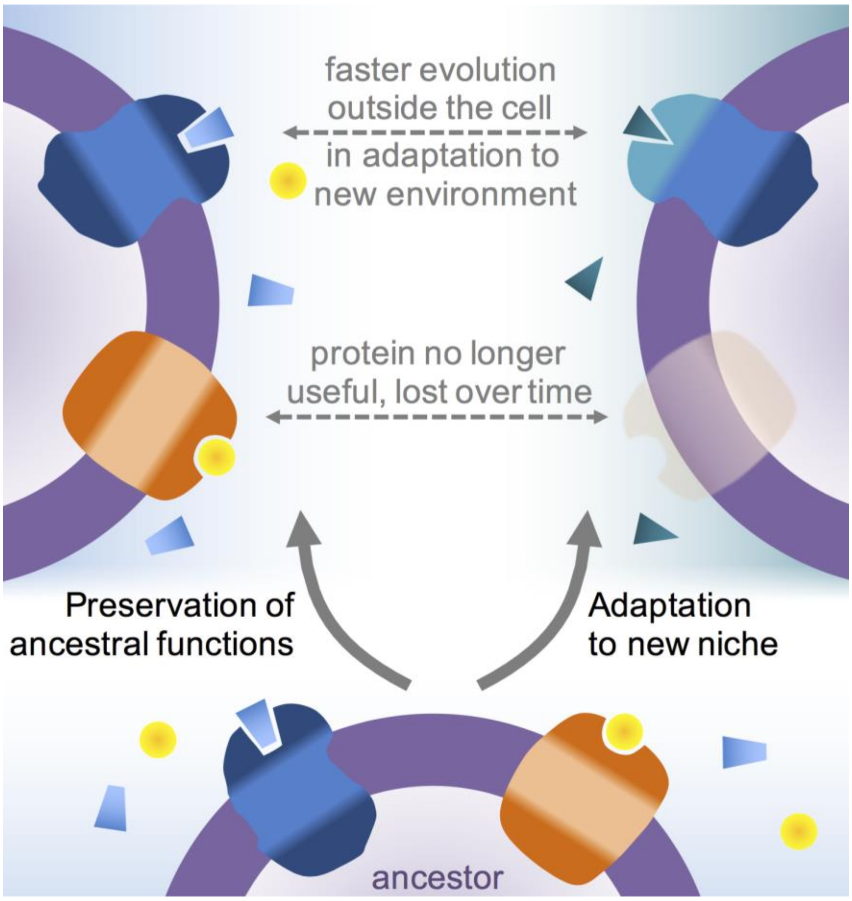

faster evolution outside the cell in adaptation to new environment
protein no longer useful, lost over time
Preservation of ancestral functions
Adaptation to new niche
ancestor

Figure 54. Chapter 5: Evolution is faster outside the cell because of adaptation to new niches

As emerging species colonise new environments and specialise in new functions, selective pressure is likely to be stronger on the outside. This leads to faster evolution outside the cell, and to loss of membrane proteins rendered useless in the new environment.

Regarding the second mechanism to explain decreased orthology in membrane proteins (true gene losses in the middle of Figure 54), it follows that once an incipiently diverging species colonises a new environment, adapts to a changing environment, or specifies in a new function, some of its membrane proteins will be rendered useless, either because an ancient external substrate is no longer found in the new environment, or because it is no longer relevant to the life history of the cell and its new functions. True gene losses should be prevalent in speciation and adaptation in general. To test this, I repeated the analysis of orthologue counts above, but focusing on closely related species. For this I assumed that if a predictably ancestral gene is not detected in some members of a group of closely related species, it is likely that the gene has truly been lost, assuming that the sequences cannot have diverged beyond recognition since speciation took place. I thus picked proteins shared by at least half of the members of the clade, and then compared the numbers of orthologues that each protein had, finding again that membrane proteins are shared by fewer species. That is, membrane proteins have been lost more often than water-soluble proteins amongst closely related species. This demonstrated that both mechanisms in Figure 54, i.e. divergence beyond recognition and true gene losses, are at play.

These findings ultimately imply that, from a phylogenetic point of view, the majority of phylogenetic differences between strains and species should be observed for membrane proteins, a prediction that can be tested straightforwardly.

An immediate prediction of the hypothesis is that the effect should be smaller for unicellular eukaryotes, many of whose membrane proteins sit on internal membranes subject to homeostasis, than for prokaryotes, and smaller still in multicellular eukaryotes, which benefit from the multiple layers of homeostasis provided by tissues, organs, and full bodies. As mentioned above, this is indeed the case. In conclusion, I provide an explanation for why evolution is faster outside the cell, and given that over half of drug targets are membrane proteins, this may help explain why it is often so difficult to advance a drug from testing on animal models into clinical trials on humans.

6.3 Open questions

6.3.1 The ancestral nature of the Na^+/H^+ antiporter (SPAP)

If the evolutionary path outlined in Chapter 2 and Figure 51 is correct, then the SPAP should be an ancestral protein, and traces of this might still be found in present-day organisms. I have indeed shown that a given SPAP sequence retrieves successful hits across the tree of life. This does not rule out horizontal gene transfer, though; i.e. it is possible that several of the hits I have recovered are a product of either recent or ancient horizontal acquisitions. There are many independent SPAP proteins in modern species. I predict that at least one of these should branch as deeply as the ribosome, tRNAs, and the ATPase, (although again, this assumes that differential losses and replacements of the ancestral version with a more efficient horizontal acquisition will not have managed to fully obliterate the ancient pattern). One potential caveat is that the SPAP is a simple gene, in contrast with the ribosome and ATPase, which are complex multi-subunit systems that should be more difficult to transfer horizontally; in addition, the usefulness of SPAP would predictably vary in environments with different salinity, such that there would be correspondingly different pressures to either acquire or lose it. Both factors should make the detection of an ancestral SPAP more difficult than for the ribosome or ATPase, yet recent findings seem to suggest that LUCA did have a SPAP (Sousa et al. 2016).

6.3.2 Are alternative archaeal phospholipid backbones ancestral?

No archaea have been observed with the bacterial G3P backbone, however not all archaea use G1P as their main phospholipid backbone either. Chemical analysis of marine sediments has produced lipids that do not have the three-carbon glycerol backbone of most living beings but instead have a four- or five-carbon analogue (Zhu et al. 2014). The presence of ether linkages in these lipids quite likely betrays their archaeal procedence, but the species are not yet known.

This use of alternative backbones could be due to either ecology or contingency. If contingency, i.e. if the archaeal organisms that use these backbones derived it independently from the ones that developed G1PDH, then these alternative backbone-bearing organisms should branch very deeply in the tree of the archaea.

6.3.3 Can simple isoprene amphiphiles originate abiotically, self-assemble into vesicles, and flip-flop?

Proton gradients are readily dissipated by flip-flop in fatty-acid vesicles (Kamp et al. 1995), which I argue was a necessity for a potential autotrophic origin of life in alkaline hydrothermal vents (Chapter 2). There is ample literature on fatty acid vesicles at the origin of life (Deamer and Nichols 1989; Monnard and Deamer 2002; e.g. Chen and Szostak 2004), but it is reasonable to consider whether isoprene amphiphiles could provide an alternative building block for the earliest protocells. Relatedly, how easy is it to synthesise isoprene amphiphiles abiotically? Their role at the origin of life was suggested over two decades ago (Ourisson and Nakatani 1994), and although abiotic synthetic mechanisms have remained more elusive than for fatty acids (Deamer et al. 2002), there has been some theoretical (Aylward 2008) and experimental (Désaubry et al. 2003) progress. The polar ends of the amphiphilic molecules synthesised are alcohols, though, which would not be readily protonated and therefore would not transfer H^+ by flip-flop. The possibility of simple isoprenoid amphiphile membranes is nevertheless worth investigating further.

6.3.4 Dual homochirality and the undetected importance of parallel evolution

I argued in Chapter 3 that archaea developed G1PDH from the duplication and neo-functionalisation of an ancestral and already available pro-*R* gene, while bacteria followed an independent and analogous route by neo-functionalising a duplicated pro-*S* enzyme gene. If so, and as I posit above, this may well be a fortunate event for those of us studying deep evolution. Instead, the early archaea and bacteria could have followed the same route independently, and both duplicated and neo-functionalised the same gene. If that had been the case, it would be almost impossible to assess whether LUCA had the gene for producing glycerol-phosphate backbones from dihydroxyacetone phosphate, and the deep split simply reflects the antiquity of both domains, or whether evolution occurred in parallel. In fact, most parsimoniously the logical yet erroneous conclusion would be that the gene was present in LUCA and both domains naturally inherited it. This was fortunately not the case for glycerol-phosphate backbones, but it suggests the possibility of daunting prospects when considering how many cases of parallel evolution like the hypothetical one suggested above may have actually happened and are currently ignored and perhaps even untraceable. One significant example is in DNA replication, which remains unclear whether it was ancestral or evolved independently in the two domains (Leipe et al. 1999).

6.3.5 Homochirality of sugar- and amino acid- synthesising enzymes

If homochirality is enforced on homologous genes by inheritance, and if it arises as an inevitable consequence of biochemical catalysis, as demonstrated by the dual homochirality of phospholipid backbones, then it should be easy to show that each of the enzymes that catalyse the reactions that produce the chiral alpha carbons in amino acids belong to families in which all the homologues share the same chiral preference. A similar result should be found for sugars (although in this case the analysis would be less trivial due to the multiplicity of chiral centres in most sugars, including ribose).

6.3.6 The origin of stereospecificity and homochirality at the origin of enzymes and cofactors

At the origin of enzymatic catalysis, when short peptides or ribozymes lacking genetic encoding started to catalyse the first proto-bio-chemical reactions, were they intrinsically stereospecific, however short? It is likely that this was the case, unless glycine content was high (glycine is the only achiral amino acid, and therefore cannot provide stereospecificity). Does this change for reactions catalysed by simple cofactors such as iron, nickel, magnesium, or other ions? It should be possible to test whether short homochiral oligomers associated to simple cofactors select for specific enantiomers, and whether the mirror images prefer the opposite enantiomeric substrate. If no enantioselectivity is observed, the lengths of the polymers could be increased to determine the at which pint one mirror image is favoured over the other.

6.3.7 The evolution of pumping

The results in Chapter 2 predict that active ion pumps must have arisen after the sodium/proton antiporter (SPAP), and that only then did selection favour the evolution of ion-tight membranes with glycerol-phosphate backbones. Given that SPAP on its own facilitated the spread and colonisation of regions with shallower and more intermittent gradients (Figure 15C, p. 70), pumping is expected to have arisen independently in more than one population. This would therefore predict differences in the mechanisms of pumping. That seems to be the case. The simplest and arguably most ancient chemiosmotic circuits are those of the methanogens (archaea) and acetogens (bacteria), which grow from H_2 and CO_2 alone via the Wood-Ljungdahl acetyl-CoA pathway (Fuchs 2011). These cells have a single membrane ion pump and lack respiratory chains with quinones and cytochromes (Fuchs 2011; Buckel and Thauer 2013). The acetyl-CoA pathway is the only exergonic carbon fixation pathway; it is short and linear, and contains numerous inorganic (iron-nickel-sulphur) clusters, all of which point to its ancient origins (Russell and Martin 2004). While the deepest branches among the archaea and bacteria are hard to constrain phylogenetically, some studies do indeed indicate that methanogens (Kelly et al. 2011; Nelson-Sathi et al. 2015; Raymann et al. 2015) and acetogens (Ciccarelli et al. 2006) are ancient and potentially ancestral to their respective domains. However, there is a deep split in the biochemistry of the acetyl-CoA pathway, specifically in the mechanism of electron bifurcation (Sojo et al. 2016), the process by which both groups generate electrochemical ion gradients (Buckel and Thauer 2013). This deep split in the mechanism of pumping is in fact entirely consistent with the prediction put forward here that active pumping should have arisen independently in archaea and bacteria. Specifically, the divergence in the acetyl-CoA pathway can be explained by the direction of ion flux through Ech (Figure 55). Ech drew on natural proton gradients to drive carbon reduction in LUCA (Figure 55A), and methanogens continued to do so, obliging them to evolve a separate ion pump (Mtr) and the biochemistry to sustain it (Figure 55B). In contrast, acetogens simply reversed the direction of ion flux through Ech, giving them a ready-made pump, but obliging them to evolve a different pathway for carbon fixation (Figure 55C).

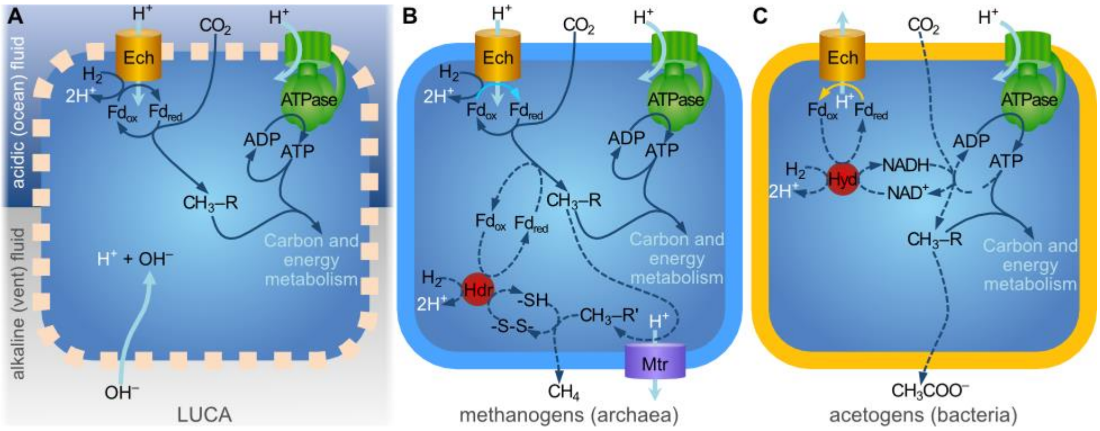


Figure 55. Evolution of pumping in archaea and bacteria

(A) Proposed carbon and energy metabolism powered by natural proton gradients in an ancestral protocell with leaky membranes. Ech: energy-converting hydrogenase; Fd_{ox}: ferredoxin; Fd_{red}: reduced ferredoxin. For simplicity, this figure only depicts the $–CH_3$ branch of a postulated ancestral acetyl-CoA pathway. R is one of a number of cofactors that differ between methanogens and acetogens. The direction of proton flow through Ech is critical and differs in (C). **(B)** Simplified carbon and energy metabolism of methanogens (archaea). Ech reduces ferredoxin using a proton gradient as in (A), but there is now a requirement to pump ions to regenerate membrane potential. This is achieved by electron bifurcation, using H_2 to simultaneously reduce ferredoxin and a heterodisulfide (-S–S-). Methanogenesis regenerates membrane potential via a new ion pump (Mtr), which may have evolved from a Na^+/H^+ antiporter. **(C)** Simplified carbon and energy metabolism of acetogens (bacteria). Ech reverses, oxidising ferredoxin to generate membrane potential. There is now a need to regenerate reduced ferredoxin, achieved via a distinct form of electron bifurcation that uses H_2 to simultaneously reduce ferredoxin and NAD^+. However, acetogens can no longer use ferredoxin to reduce CO_2, hence had to evolve a new pathway of carbon metabolism, using ATP and NADH in place of ferredoxin. New pathways of energy metabolism are depicted with dotted lines. Taken from Sojo et al. (Sojo et al. 2016).

These conclusions should be testable by increasing genomic data and improved phylogenetic methods. The prediction is that they will confirm the deep branching of methanogens and acetogens.

6.3.8 The early evolution of additional symporters and antiporters

The results in Chapter 2 bring forward the possibility that other types of transporters, besides the sodium-proton antiporter (SPAP), may have evolved early. Modern cells are characterised not only by low internal sodium contents, but also by high internal potassium (Mulkidjanian et al. 2012). Therefore, it could be fruitful to study the potential early evolution of transporters such as spontaneous K^+ channels, or H^+/Cl^- symporters.

6.3.9 Faster evolution outside the cell in diversification of membrane lipids

Trivially, different lipids behave differently. Following the arguments in Chapter 5, in which I show that membrane proteins evolve faster and are lost more often than water-soluble proteins because of adaptation to new niches, it should be productive to analyse whether the differences in membrane lipids within domains are also caused at least in part by the same mechanism. Multicellular eukaryotes, in which different lipids become more prevalent in membranes in different tissues, should provide good subjects for testing.

6.3.10 The adaptive role of horizontal gene transfers of membrane proteins

In relation to the results in Chapter 5, membrane proteins should be more likely to be acquired horizontally than water soluble proteins, since they should have a greater effect on adaptation to new environments. However, this is in direct contradiction to the predictions of Chapter 4, specifically that it should be more difficult to acquire foreign membrane proteins than foreign water-soluble proteins, simply because of the physical interactions between the membrane protein and the surrounding phospholipids. A thorough study of horizontal gene transfers of membrane-bound versus water-soluble proteins within and across domains should prove fruitful.

6.3.11 The origin of life at alkaline hydrothermal vents

The origin of life remains one of the most fundamental unresolved problems in biology. There are many open questions, including the prebiotic synthesis of life's first molecules from the simplest building blocks, inorganic hydrogen (H_2) and carbon dioxide (CO_2), assuming an autotrophic origin of life in alkaline hydrothermal vents, as reviewed in Sojo et al. (Sojo et al. 2016). It is also important to show whether amphiphiles can form viable membranes under these conditions, contrary to some predictions in the literature (Deamer and Georgiou 2015). Microfluidics should prove a useful tool for simulating the conditions in alkaline hydrothermal vents and testing these hypotheses.

6.4 Conclusions

Life is characterised by disequilibria and homeostasis across membranes. Membranes and their associated proteins play crucial roles in most cellular processes, and their importance is correspondingly central to evolution.

The last universal common ancestor (LUCA) should have had membranes made from organic amphiphiles such as fatty acids, but these membranes would have been leaky to H^+. LUCA's proton–leaky membranes allowed it to exploit natural ion gradients such as those provided by alkaline hydrothermal vents. Modern membranes evolved later, after the origin of pumping and, predictably, also after the evolution of a sodium/proton antiporter (SPAP). This means that LUCA's membranes could not have had a glycerol-phosphate backbone, which makes phospholipids much less permeable than simpler amphiphiles such as single fatty acids. This explains why the archaea and bacteria have different stereochemistries in their glycerol-phosphate backbones, and why the corresponding synthesising enzymes are unrelated: they evolved independently, and later, after the divergence of the ancestors of the two domains.

This dichotomy in the glycerol phospholipid backbones points to a general underlying cause for the strong one-handedness, or *homochirality*, of terrestrial life, most notably in D-sugars and L-amino acids. But there is no great mystery in biochemical chirality: the very process of enzymatic catalysis implies that, at the atomic level, enantiomers (mirror images of otherwise the same molecule) are entirely different molecules, and an enzyme that becomes specific for a catalysis that involves one isomer need have no particular affinity for the other. The evolution of novel functions in biology is often dictated by contingency: gene duplications facilitate the adaptation of gene products into new roles, but these new roles are inevitably constrained by the nature of the ancestral sequence. In terms of enzymatic catalysis, this means that stereochemical preferences will be preserved. Ultimately, homochirality arises inevitably because enzymes can only become efficient by picking a single orientation, and this has to be one of two in the case of the R/S symmetry of chiral carbon atoms or their pro-chiral precursors. More broadly, in the evolution of a biochemical pathway, a chiral choice at the origin of an early catalyst is inexorably imposed on the enzymes that evolve later.

Long after the divergence of the two prokaryotic domains, a member of the bacteria established an endosymbiotic association inside a member of the archaea, giving rise to the eukaryotic cell. This meant that there were both archaeal and bacterial membranes in the first eukaryotic common ancestor (FECA), yet the fact that all modern eukaryotes have only bacterial membranes suggests that, by the time the last eukaryotic common ancestor (LECA) had evolved, the archaeal analogues had been largely lost. Since many duplicated functions were retained, it seems unlikely that genome-size pressures forced this loss, so it is possible that the reduction was due to deleterious interactions in a hybrid membrane. I suggest that bacterial lipids had to be kept because of their increasingly important association to bioenergetic proteins in the mitochondrial membrane.

Membrane proteins are dramatically less conserved than water-soluble proteins, across all three domains in the tree of life. Given that trans-membrane proteins link the interior of the cell to their surrounding environment, they interact more frequently with the exterior, and are therefore more closely involved in environmental adaptation than are cytosolic proteins. Over time, this means that they evolve faster, particularly on the outside. More drastically, some membrane proteins will be of no use in a new environment, and may be lost due to genome reduction. This effect should be stronger in prokaryotes, weaker in unicellular eukaryotes (which have organellar membranes), and weakest in multicellular eukaryotes (with multiple levels of homeostasis); this is indeed the case.

It is my hope that this thesis will serve to highlight the importance of membranes and particularly the disequilibria across them in the origin and evolution of life.

REFERENCES

Akanni WA, Siu-Ting K, Creevey CJ, McInerney JO, Wilkinson M, Foster PG, Pisani D. 2015. Horizontal gene flow from Eubacteria to Archaebacteria and what it means for our understanding of eukaryogenesis. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140337.

Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. 2007. Molecular Biology of the Cell. 5th ed. New York, NY, USA: Garland Science

Allen JF. 1993. Control of gene expression by redox potential and the requirementfor chloroplast and mitochondrial genomes. *J. Theor. Biol.* 165:609–631.

Allen JF. 2003. The function of genomes in bioenergetic organelles. *Philos. Trans. R. Soc. B Biol. Sci.* 358:19–38.

Allen JF. 2010. Redox homeostasis in the emergence of life. On the constant internal environment of nascent living cells. *J. Cosmol.* 10:3362–3373.

Allen JF. 2015. Why chloroplasts and mitochondria retain their own genomes and genetic systems: Colocation for redox regulation of gene expression. *Proc. Natl. Acad. Sci. U.S.A.* 112:201500012.

Allers T, Mevarech M. 2005. Archaeal genetics - the third way. *Nat. Rev. Genet.* 6:58–73.

Altenhoff AM, Škunca N, Glover N, Train C-M, Sueki A, Piližota I, Gori K, Tomiczek B, Müller S, Redestig H, et al. 2015. The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Res.* 43:D240–D249.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

Amend JP, McCollom TM. 2009. Energetics of biomolecule synthesis on early Earth. In: Zaikowski L, Friedrich JM, Seidel SR, editors. Chemical Evolution II: From the Origins of Life to Modern Society. Vol. 1025. ACS Symposium Series. American Chemical Society. p. 63–94.

Aravind L, Tatusov RL, Wolf YI, Walker DR, Koonin E V. 1998. Evidence for massive gene exchange between archaeal and bacterial hyperthermophiles. *Trends Genet.* 14:442–444.

Araya-Secchi R, Garate JA, Holmes DS, Perez-Acle T. 2011. Molecular dynamics study of the archaeal aquaporin AqpM. *BMC Genomics* 12:S8.

Arbing MA, Chan S, Shin A, Phan T, Ahn CJ, Rohlin L, Gunsalus RP. 2012. Structure of the surface layer of the methanogenic archaean Methanosarcina acetivorans. *Proc. Natl. Acad. Sci. U.S.A.* 109:11812–11817.

Arndt NT, Nisbet EG. 2012. Processes on the young Earth and the habitats of early life. *Annu. Rev. Earth Planet. Sci.* 40:521–549.

Arnold LJ, You K, Allison WS, Kaplan NO. 1976. Determination of the hydride transfer stereospecificity of nicotinamide adenine dinucleotide linked oxidoreductases by proton magnetic resonance. *Biochemistry* 15:4844–4849.

Aylward N. 2008. The synthesis of terpenes in prebiotic molecular evolution on Earth. *Proc. 1st WSEAS Int. Conf. Biomed. Electron. Biomed. informatics*:202–207.

Baaske P, Weinert FM, Duhr S, Lemke KH, Russell MJ, Braun D. 2007. Extreme accumulation of nucleotides in simulated hydrothermal pore systems. *Proc. Natl. Acad. Sci. U.S.A.* 104:9346–9351.

Baradaran R, Berrisford JM, Minhas GS, Sazanov LA. 2013. Crystal structure of the entire respiratory complex I. *Nature* 494:443–448.

Baranova E, Fronzes R, Garcia-Pino A, Van Gerven N, Papapostolou D, Péhau-Arnaudet G, Pardon E, Steyaert J, Howorka S, Remaut H. 2012. SbsB structure and lattice reconstruction unveil Ca2+ triggered S-layer assembly. *Nature* 487:119–122.

Bassilana M, Damiano E, Leblanc G. 1984. Kinetic properties of Na(+) -H(+) antiport in *Escherichia coli* membrane vesicles: Effects of imposed electrical potential, proton gradient, and internal pH. *Biochemistry* 23:5288–5294.

Baymann F, Lebrun E, Brugna M, Schoepp-Cothenet B, Giudici-Orticoni M-T, Nitschke W. 2003. The redox protein construction kit: pre-last universal common ancestor evolution of energy-conserving enzymes. *Phil. Trans. R. Soc. B* 358:267–274.

Benner S. 1982. The stereoselectivity of alcohol dehydrogenases: A stereochemical imperative? *Experientia* 38:633–637.

Berendsen HJC, van der Spoel D, van Drunen R. 1995. GROMACS: A message-passing parallel molecular dynamics implementation. *Comput. Phys. Commun.* 91:43–56.

Berg IA, Kockelkorn D, Ramos-Vera WH, Say RF, Zarzycki J, Hügler M, Alber BE, Fuchs G. 2010. Autotrophic carbon fixation in archaea. *Nat. Rev. Microbiol.* 8:447–460.

Berg OG, Kurland CG. 2000. Why mitochondrial genes are most often found in nuclei. *Mol. Biol. Evol.* 17:951–961.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

Bernsel A, Viklund H, Falk J, Lindahl E, von Heijne G, Elofsson A. 2008. Prediction of membrane-protein topology from first principles. *Proc. Natl. Acad. Sci. U.S.A.* 105:7177–7181.

Bernsel A, Viklund H, Hennerdal A, Elofsson A. 2009. TOPCONS: consensus prediction of membrane protein topology. *Nucleic Acids Res.* 37:W465–W468.

Blackmond DG. 2004. Asymmetric autocatalysis and its implications for the origin of homochirality. *Proc. Natl. Acad. Sci. U.S.A.* 101:5732–5736.

Bonner WA, Kavasmaneck PR, Martin FS, Flores JJ. 1975. Asymmetric adsorption by quartz: A model for the prebiotic origin of optical activity. *Orig. Life* 6:367–376.

Boothby TC, Tenlen JR, Smith FW, Wang JR, Patanella KA, Osborne E, Tintori SC, Li Q, Jones CD, Yandell M, et al. 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U.S.A.*

112:15976–15981.

Brack A, Spach GG, Synthesis PI, Studies C. 1979. β-structures of polypeptides with L- and D-residues. Part I. *J. Mol. Evol.* 13:35–46.

Bräsen C, Esser D, Rauch B, Siebers B. 2014. Carbohydrate Metabolism in Archaea: Current Insights into Unusual Enzymes and Pathways and Their Regulation. *Microbiol. Mol. Biol. Rev.* 78:89–175.

Breslow R. 2011. The origin of homochirality in amino acids and sugars on prebiotic earth. *Tetrahedron Lett.* 52:4228–4232.

Brocks JJ. 1999. Archean Molecular Fossils and the Early Rise of Eukaryotes. *Science* 285:1033–1036.

Brooks DJ, Fresco JR. 2002. Increased frequency of cysteine, tyrosine, and phenylalanine residues since the last universal ancestor. *Mol. Cell. Proteomics* 1:125–131.

Buckel W, Thauer RK. 2013. Energy conservation via electron bifurcating ferredoxin reduction and proton/Na+ translocating ferredoxin oxidation. *Biochim. Biophys. Acta* 1827:94–113.

Budin I, Bruckner RJ, Szostak JW. 2009. Formation of protocell-like vesicles in a thermal diffusion column. *J. Am. Chem. Soc.* 131:9628–9629.

Chen IA, Szostak JW. 2004. A kinetic study of the growth of fatty acid vesicles. *Biophys. J.* 87:988–998.

Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* 311:1283–1287.

Cline DB. 2005. On the physical origin of the homochirality of life. *Eur. Rev.* 13:49–59.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423.

Cohen KM, Finney SC, Gibbard PL, Fan J-X. 2013. The ICS International Chronostratigraphic Chart. *Episodes* 36:199–204.

Colquhoun D. 2014. An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. Open Sci.* 1:140216.

Conant GC, Wolfe KH. 2008. Turning a hobby into a job: How duplicated genes find new functions. *Nat. Rev. Genet.* 9:938–950.

Corrigan JJ. 1969. D-Amino Acids in Animals. *Science* 164:142–149.

Cramer CJ. 2004. Essentials of Computational Chemistry, 2nd Edition. 2nd ed. Chichester, West Sussex, UK: Wiley-Blackwell

Dagan T, Martin W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc. Natl. Acad. Sci. U.S.A.* 104:870–875.

Dagan T, Roettger M, Bryant D, Martin W. 2010. Genome networks root the tree of life between prokaryotic domains. *Genome Biol. Evol.* 2:379–392.

Dalrymple GB. 2001. The age of the Earth in the twentieth century: a problem (mostly) solved. *Geol. Soc. London, Spec. Publ.* 190:205–221.

Damer B, Deamer D, van Kranendonk M, Walter M. 2016. An origin of life through three coupled phases in cycling hydrothermal pools with distribution and adaptive radiation to marine stromatolites. In: Proceedings of the 2016 Gordon Research Conference on the Origins of Life.

Damer B, Deamer D. 2015. Coupled phases and combinatorial selection in fluctuating hydrothermal pools: a scenario to guide experimental approaches to the origin of cellular life. *Life* 5:872–887.

Darwin C, Wallace A. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. *J. Proc. Linn. Soc. London. Zool.* 3:45–62.

Darwin C. 1838. Notebook B: [Transmutation of species (1837-1838)].

Darwin C. 1845. Letter 820, to C.H. Smith (26 Jan 1845). *Darwin Corresp. Proj.* Available from: http://www.darwinproject.ac.uk/letter/entry-820

Darwin C. 1859. On the origin of species by means of natural selection. London: John Murray

Darwin C. 1863. Correspondence to J.D. Hooker, in “Life and Letters of Charles Darwin - Volume 2.” null

Deamer D, Bramhall J. 1986. Permeability of lipid bilayers to water and ionic solutes. *Chem. Phys. Lipids.* 40:167–188.

Deamer D, Dworkin J. 2005. Chemistry and physics of primitive membranes. *Top. Curr. Chem.* 259:1–27.

Deamer D, Dworkin JP, Sandford SA, Bernstein MP, Allamandola LJ. 2002. The first cell membranes. *Astrobiology* 2:371–381.

Deamer D, Nichols J. 1989. Proton Flux Mechanisms in Model and Biological Membranes. *J. Memb. Biol.* 103:91–103.

Deamer D, Weber AL. 2010. Bioenergetics and life’s origins. *Cold Spring Harb. Perspect. Biol.* 2:a004929.

Deamer D. 2008. Origins of life: How leaky were primitive cells? *Nature* 454:37–38.

Deamer DW, Georgiou CD. 2015. Hydrothermal conditions and the origin of cellular life. *Astrobiology* 15:1091–1095.

Deamer DW, Nichols JW. 1983. Proton-hydroxide permeability of liposomes. *Proc. Natl. Acad. Sci. U.S.A.* 80:165–168.

Denayer T, Stöhr T, Van Roy M. 2014. Animal models in translational medicine: Validation and prediction. *New Horizons Transl. Med.* 2:5–11.

Désaubry L, Nakatani Y, Ourisson G. 2003. Toward higher polyprenols under “prebiotic” conditions. *Tetrahedron Lett.* 44:6959–6961.

Dirac PAM. 1929. Quantum mechanics of many-electron systems. *Proc. R. Soc.* ...

123:714–733.

Domański J, Stansfeld PJ, Sansom MSP, Beckstein O. 2010. Lipidbook: a public repository for force-field parameters used in membrane simulations. *J. Membr. Biol.* 236:255–258.

Doolittle W. 1999. Phylogenetic classification and the universal tree. *Science* 284:2124–2128.

Doolittle WF, Boucher Y, Nesbø CL, Douady CJ, Andersson JO, Roger a J. 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Phil. Trans. R. Soc. B* 358:39–57.

Doolittle WF, Mariscal C. 2015. Eukaryotes first: how could that be? *Phil. Trans. R. Soc. B* 370:20140322.

Doolittle WF. 1999. Lateral genomics. *Trends Biochem. Sci.* 24:M5–M8.

Doolittle WF. 2000. Uprooting the tree of life. *Sci. Am.* 282:90–95.

Doolittle WF. 2014. How natural a kind is “eukaryote?” *Cold Spring Harb. Perspect. Biol.* 6:1–12.

Ducluzeau A-L, Schoepp-Cothenet B, Baymann F, Russell MJ, Nitschke W. 2014. Free energy conversion in the LUCA: Quo vadis? *Biochim. Biophys. Acta* 1837:982–988.

de Duve C. 1995. Vital dust: life as a cosmic imperative. New York: Basic Books

Edgell DR, Doolittle WF. 1997. Archaea and the origin(s) of DNA replication proteins. *Cell* 89:995–998.

Eme L, Sharpe SC, Brown MW, Roger AJ. 2014. On the age of eukaryotes: evaluating evidence from fossils and molecular clocks. *Cold Spring Harb. Perspect. Biol.* 6:1–16.

Engel M, Macko S. 1997. Isotopic evidence for extraterrestrial non- racemic amino acids in the Murchison meteorite. *Nature* 389:265–268.

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G, Henze K, Kretschmann E, Richly E, Leister D, et al. 2004. A genome phylogeny for mitochondria among α-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Mol. Biol. Evol.* 21:1643–1660.

Etzold C, Deckers-Hebestreit G, Altendorf K. 1997. Turnover number of *Escherichia coli* F_OF_1 ATP synthase for ATP synthesis in membrane vesicles. *Eur. J. Biochem.* 243:336–343.

Ferguson SJ. 2010. ATP synthase: from sequence to ring size to the P/O ratio. *Proc. Natl. Acad. Sci. U.S.A.* 107:16755–16756.

Fischer E. 1894. Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte der Dtsch. Chem. Gesellschaft* 27:2985–2993.

Fisher HF, Conn EE, Vennesland B, Westheimer FH. 1953. The enzymatic transfer of hydrogen I. The reaction catalyzed by alcohol dehydrogenase. *J. Biol. Chem.* 202:687–697.

Forterre P, Bergerat A, López-García P. 1996. The unique DNA topology and DNA topoisomerases of hyperthermophilic archaea. *FEMS Microbiol. Rev.* 18:237–248.

Franzosa EA, Xue R, Xia Y. 2013. Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. *Genome Biol. Evol.* 5:734–744.

Fuchs G, Stupperich E, Jaenchen R. 1980. Autotrophic CO2 fixation in Chlorobium limicola. Evidence against the operation of the Calvin cycle in growing cells. *Arch. Microbiol.* 128:56–63.

Fuchs G. 2011. Alternative pathways of carbon dioxide fixation: insights into the early evolution of life? *Annu. Rev. Microbiol.* 65:631–658.

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T. 1989. Evolution of the vacuolar H+-ATPase: implications for the origin of eukaryotes. *Proc. Natl. Acad. Sci. U.S.A.* 86:6661–6665.

Goldman D. 1943. Potential, impedance, and rectification in membranes. *J. Gen. Physiol.* 27:37–60.

Gray MW, Burger C, Structural CV. 1999. Mitochondrial evolution. *Science* 283:1476–1481.

Griffith F. 1927. The significance of pneumococcal types. *J. Hyg. (Lond).* XXVII:113–159.

Haeckel E. 1874. The Gastraea-theory, the phylogenetic classification of the animal kingdom and the homology of the germ-lamellæ. *Q. J. Microsc. Sci.* 2:142–165.

Haggerty LS, Jachiet PA, Hanage WP, Fitzpatrick DA, Lopez P, O'Connell MJ, Pisani D, Wilkinson M, Bapteste E, McInerney JO. 2014. A pluralistic account of homology: adapting the models to the data. *Mol. Biol. Evol.* 31:501–516.

Hammes GG, Hilborn DA. 1971. Steady state kinetics of soluble and membrane-bound mitochondrial ATPase. *Biochim. Biophys. Acta* 233:580–590.

Hanczyc M, Fujikawa S, Szostak J. 2003. Experimental models of primitive cellular compartments: encapsulation, growth, and division. *Science* 302:618–622.

Hanford M, Peeples T. 2002. Archaeal tetraether lipids. *Appl. Biochem. Biotechnol.* 97:45–62.

Hannaert V, Brinkmann H, Nowitzki U, Lee J a, Albert M a, Sensen CW, Gaasterland T, Müller M, Michels P, Martin W. 2000. Enolase from Trypanosoma brucei, from the amitochondriate protist Mastigamoeba balamuthi, and from the chloroplast and cytosol of Euglena gracilis: pieces in the evolutionary puzzle of the eukaryotic glycolytic pathway. *Mol. Biol. Evol.* 17:989–1000.

Hanson K, Rose I. 1975. Interpretations of Enzyme Reaction Stereospecificity. *Acc. Chem. Res.* 8:1–10.

Hanson K. 1972. Enzyme symmetry and enzyme stereospecificity. *Annu. Rev. Plant Physiol.* 23:335–366.

Harold FM. 1986. The vital force: a study of bioenergetics. New York, NY, USA: W.H. Freeman

Hartman H, Fedorov A. 2002. The origin of the eukaryotic cell: a genomic investigation. *Proc. Natl. Acad. Sci. U.S.A.* 99:1420–1425.

Hedderich R. 2004. Energy-converting [NiFe] hydrogenases from archaea and extremophiles: ancestors of complex I. *J. Bioenerg. Biomembr.* 36:65–75.

Hedin LE, Illergård K, Elofsson A. 2011. An introduction to membrane proteins. *J. Proteome Res.* 10:3324–3331.

von Heijne G. 1986a. Why mitochondria need a genome. *FEBS Lett.* 198:1–4.

von Heijne G. 1986b. The distribution of positively charged residues in bacterial inner membrane proteins correlates with the trans-membrane topology. *EMBO J.* 5:3021–3027.

von Heijne G. 1992. Membrane protein structure prediction. Hydrophobicity analysis and the positive-inside rule. *J. Mol. Biol.* 225:487–494.

Hellmer J, Pätzold R, Zeilinger C. 2002. Identification of a pH regulated Na+/H+ antiporter of *Methanococcus jannaschii*. *FEBS Lett.* 527:245–249.

Henze K, Badr A, Wettern M, Cerff R, Martin W. 1995. A nuclear gene of eubacterial origin in *Euglena gracilis* reflects cryptic endosymbioses during protist evolution. *Proc. Natl. Acad. Sci. U.S.A.* 92:9122–9126.

Herbert S. 2005. Charles Darwin, Geologist. Cornell University Press

Herschy B, Whicher A, Camprubi E, Watson C, Dartnell L, Ward J, Evans JRG, Lane N. 2014. An origin-of-life reactor to simulate alkaline hydrothermal vents. *J. Mol. Evol.* 79:213–227.

Hess B, Kutzner C, van der Spoel D, Lindahl E. 2008. GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation. *J. Chem. Theory Comput.* 4:435–447.

Hessa T, Meindl-Beinker NM, Bernsel A, Kim H, Sato Y, Lerch-Bader M, Nilsson I, White SH, von Heijne G. 2007. Molecular code for transmembrane-helix recognition by the Sec61 translocon. *Nature* 450:1026–1030.

Hilario E, Gogarten JP. 1993. Horizontal transfer of ATPase genes — the tree of life becomes a net of life. *Biosystems* 31:111–119.

Hinchliffe A. 2008. Molecular Modelling for Beginners. Wiley-Blackwell

Hodgkin A, Katz B. 1949. The effect of sodium ions on the electrical activity of the giant axon of the squid. *J. Physiol.* 108:37–77.

Holmes AM, Solari R, Holgate ST. 2011. Animal models of asthma: value, limitations and opportunities for alternative approaches. *Drug Discov. Today* 16:659–670.

Huerta-Cepas J, Dopazo J, Gabaldón T. 2010. ETE: a python Environment for Tree Exploration. *BMC Bioinformatics* 11:24.

Humphrey W. 1996. VMD: Visual molecular dynamics. *J. Mol. Graph.* 14:33–38.

Hunte C, Screpanti E, Venturi M, Rimon A, Padan E, Michel H. 2005. Structure of a Na+/H+ antiporter and insights into mechanism of action and regulation by pH. *Nature* 435:1197–1202.

Iborra FJ, Kimura H, Cook PR. 2004. The functional organization of mitochondrial genomes in human cells. *BMC Biol.* 2:1–14.

Jacob F, Ryter A, Cuzin F. 1966. On the association between DNA and membrane in bacteria. *Proc. R. Soc. B* 164:267–278.

Jensen JH. 2010. Molecular Modeling Basics. Boca Raton, FL, USA: CRC Press

Jones JE. 1924. On the Determination of Molecular Fields. II. From the Equation of State of a Gas. *Proc. R. Soc. A Math. Phys. Eng. Sci.* 106:463–477.

Jorissen A, Cerf C. 2002. Asymmetric photoreactions as the origin of biomolecular homochirality: a critical review. *Orig. Life Evol. Biosph.* 32:129–142.

Joyce GF. 1994. Foreword. In: Deamer DW, Fleischaker GR, editors. Origins of life: the central concepts. Boston, MA, USA: Jones and Bartlett Publishers. p. 431.

Julenius K, Pedersen AG. 2006. Protein evolution is faster outside the cell. *Mol. Biol. Evol.* 23:2039–2048.

Kagawa Y, Ohta T, Abe Y, Endo H, Yohda M, Kato N, Endo I, Hamamoto T, Ichida M, Hoaki T, et al. 1995. Gene of Heat Shock Protein of Sulfur-Dependent Archaeal Hyperthermophile Desulfurococcus. *Biochem. Biophys. Res. Commun.* 214:730–736.

Käll L, Krogh A, Sonnhammer ELL. 2005. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 21 Suppl 1:i251–i257.

Kamp F, Zakim D, Zhang F, Noy N, Hamilton JA. 1995. Fatty acid flip-flop in phospholipid bilayers is extremely fast. *Biochemistry* 34:11928–11937.

Kandasamy SK, Larson RG. 2006. Molecular dynamics simulations of model trans-membrane peptides in lipid bilayers: a systematic investigation of hydrophobic mismatch. *Biophys. J.* 90:2326–2343.

Karagounis G, Coumoulos G. 1938. A new method of resolving a racemic compound. *Nature* 141:917–918.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

Katz LA. 2015. Recent events dominate interdomain lateral gene transfers between prokaryotes and eukaryotes and, with the exception of endosymbiotic gene transfers, few ancient transfer events persist. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140324.

Kawai F. 1999. Sphingomonads involved in the biodegradation of xenobiotic polymers. *J. Ind. Microbiol. Biotechnol.* 23:400–407.

Kawasaki T, Suzuki K, Shimizu M, Ishikawa K, Soai K. 2006. Spontaneous absolute asymmetric synthesis in the presence of achiral silica gel in conjunction with asymmetric autocatalysis. *Chirality* 18:479–482.

Kelley DS, Karson JA, Blackman DK, Früh-Green GL, Butterfield DA, Lilley MD, Olson EJ, Schrenk MO, Roe KK, Lebon GT, et al. 2001. An off-axis hydrothermal vent field near the Mid-Atlantic Ridge at 30 degrees N. *Nature* 412:145–149.

Kelly S, Wickstead B, Gull K. 2011. Archaeal phylogenomics provides evidence in support of a methanogenic origin of the Archaea and a thaumarchaeal origin for the eukaryotes. *Proc. R. Soc. B* 278:1009–1018.

Kelman LM, Kelman Z. 2004. Multiple origins of replication in archaea. *Trends Microbiol.* 12:399–401.

Koga Y, Kyuragi T, Nishihara M, Sone N. 1998. Did archaeal and bacterial cells arise independently from noncellular precursors? A hypothesis stating that the advent of membrane phospholipid with enantiomeric glycerophosphate backbones caused the separation of the two lines of descent. *J. Mol. Evol.* 46:54–63.

Koga Y, Sone N, Noguchi S, Morii H. 2014. Transfer of Pro- R Hydrogen from NADH to Dihydroxyacetonephosphate by sn -Glycerol-1-phosphate Dehydrogenase from the Archaeon Methanothermobacter thermautotrophicus. *Biosci. Biotechnol. Biochem.* 67:1605–1608.

Kondepudi DK, Nelson GW. 1985. Weak neutral currents and the origin of biomolecular chirality. *Nature* 314:438–441.

Koonin E V, Makarova KS, Aravind L. 2001. Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55:709–742.

Koonin E V, Martin W. 2005. On the origin of genomes and cells within inorganic compartments. *Trends Genet.* 21:647–654.

Koonin E V, Novozhilov AS. 2009. Origin and evolution of the genetic code: the universal enigma. *IUBMB Life* 61:99–111.

Koonin E V. 2009. On the origin of cells and viruses: primordial virus world scenario. *Ann. N. Y. Acad. Sci.* 1178:47–64.

Koonin E V. 2015. The turbulent network dynamics of microbial evolution and the statistical tree of life. *J. Mol. Evol.* 80:244–250.

Koumandou VL, Wickstead B, Ginger ML, van der Giezen M, Dacks JB, Field MC. 2013. Molecular paleontology and complexity in the last eukaryotic common ancestor. *Crit. Rev. Biochem. Mol. Biol.* 48:373–396.

Koutsovoulos G, Kumar S, Laetsch DR, Stevens L, Daub J, Conlon C, Maroon H, Thomas F, Aboobaker A, Blaxter M. 2015. The genome of the tardigrade Hypsibius dujardini. *bioRxiv*:033464.

Krebs HA. 1935. Metabolism of amino-acids: Deamination of amino-acids. *Biochem. J.* 29:1620–1644.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580.

Ku C, Nelson-sathi S, Roettger M, Sousa FL, Lockhart PJ, Bryant D, Hazkani-covo E, Mcinerney JO, Landan G, Martin WF. 2015. Endosymbiotic origin and differential loss of eukaryotic genes. *Nature* 524:427–432.

Lake JA. 2015. Eukaryotic Origins. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140321.

Lane N, Allen JF, Martin W. 2010. How did LUCA make a living? Chemiosmosis in the origin of life. *BioEssays* 32:271–280.

Lane N, Martin W. 2010. The energetics of genome complexity. *Nature* 467:929–934.

Lane N, Martin WF. 2012. The origin of membrane bioenergetics. *Cell* 151:1406–1416.

Lane N. 2010. Life Ascending: The Ten Great Inventions of Evolution. London, UK: Profile Books

Lane N. 2014. Bioenergetic constraints on the evolution of complex life.Keeling PJ, Koonin E V, editors. *Cold Spring Harb. Perspect. Biol.* 6:a015982.

Lane N. 2015. The Vital Question: Energy, Evolution, and the Origins of Complex Life. New York: WW Norton & Co.

Lanyi JK. 1998. Understanding structure and function in the light-driven proton pump bacteriorhodopsin. *J. Struct. Biol.* 124:164–178.

Lee A, Rana BK, Schiffer HH, Schork NJ, Brann MR, Insel P a., Weiner DM. 2003. Distribution analysis of nonsynonymous polymorphisms within the G-protein-coupled receptor gene family. *Genomics* 81:245–248.

Leipe DD, Aravind L, Koonin E V. 1999. Did DNA replication evolve twice independently? *Nucleic Acids Res.* 27:3389–3401.

Lemkul J a, Bevan DR. 2011. Characterization of interactions between PilA from Pseudomonas aeruginosa strain K and a model membrane. *J. Phys. Chem. B* 115:8004–8008.

Li YD, Xie ZY, Du YL, Zhou Z, Mao XM, Lv LX, Li YQ. 2009. The rapid evolution of signal peptides is mainly caused by relaxed selection on non-synonymous and synonymous sites. *Gene* 436:8–11.

Liao BY, Weng MP, Zhang J. 2010. Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biol. Evol.* 2:39–43.

Liebgott P-P, Leroux F, Burlat B, Dementin S, Baffert C, Lautier T, Fourmond V, Ceccaldi P, Cavazza C, Meynial-Salles I, et al. 2010. Relating diffusion along the substrate tunnel and oxygen sensitivity in hydrogenase. *Nat. Chem. Biol.* 6:63–70.

Lindahl E, Hess B, van der Spoel D. 2001. GROMACS 3.0: a package for molecular simulation and trajectory analysis. *J. Mol. Model.* 7:306–317.

Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J. 2000. Molecular Cell Biology. 4th ed. New York, NY, USA: W. H. Freeman

Lombard J, López-García P, Moreira D. 2012a. The early evolution of lipid membranes and the three domains of life. *Nat. Rev. Microbiol.* 10:507–515.

Lombard J, López-García P, Moreira D. 2012b. An ACP-Independent Fatty Acid Synthesis Pathway in Archaea: Implications for the Origin of Phospholipids. *Mol. Biol. Evol.* 29:3261–3265.

Londei P. 2006. Translational mechanisms and protein synthesis. In: Garrett RA, Klenk H-P, editors. Archaea: Evolution, Physiology, and Molecular Biology. Malden, MA, USA: Blackwell Publishing Ltd. p. 219–228.

López-García P, Moreira D. 2015. Open questions on the origin of eukaryotes. *Trends Ecol. Evol.*:1–12.

Lupas AN, Ponting CP, Russell RB. 2001. On the evolution of protein folds: are similar motifs in different protein folds the result of convergence, insertion, or relics of an ancient peptide world? *J. Struct. Biol.* 134:191–203.

Madigan MT, Martinko JM, Stahl DA, Clark DP. 2011. Brock Biology of Microorganisms. Pearson Education

Mansy SS, Schrum JP, Krishnamurthy M, Tobé S, Treco D a, Szostak JW. 2008. Template-directed synthesis of a genetic polymer in a model protocell. *Nature* 454:122–125.

Margulis L. 2004. Serial endosymbiotic theory (SET) and composite individuality. *Microbiol. Today* 31:172–174.

Marreiros BC, Batista AP, Duarte AMS, Pereira MM. 2013. A missing link between complex I and group 4 membrane-bound [NiFe] hydrogenases. *Biochim. Biophys. Acta* 1827:198–209.

Martin W, Baross J, Kelley D, Russell MJ. 2008. Hydrothermal vents and the origin of life. *Nat. Rev. Microbiol.* 6:805–814.

Martin W, Kowallik K. 1999. Annotated English translation of Mereschkowsky's 1905 paper "Über Natur und Ursprung der Chromatophoren im Pflanzenreiche." *Eur. J. Phycol.* 34:287–295.

Martin W, Müller M. 1998. The hydrogen hypothesis for the first eukaryote. *Nature* 392:37–41.

Martin W, Roettger M. 2012. Modern endosymbiotic theory : Getting lateral gene transfer into the equation. *J. Endocytobiosis Cell Res.*:1–5.

Martin W, Russell MJ. 2003. On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. *Phil. Trans. R. Soc. B* 358:59–83.

Martin W, Russell MJ. 2007. On the origin of biochemistry at an alkaline hydrothermal vent. *Phil. Trans. R. Soc. B* 362:1887–1925.

Martin W, Schnarrenberger C. 1997. The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. *Curr. Genet.* 32:1–18.

Martin WF, Garg S, Zimorski V. 2015. Endosymbiotic theories for eukaryote origin. *Philos. Trans. R. Soc. B Biol. Sci.* 370:20140330.

Mason SF. 1984. Origins of biomolecular handedness. *Nature* 311:19–23.

Maynard Smith J, Szathmáry E. 1997. The major transitions in evolution. Oxford, U.K.: Oxford University Press

Maynard Smith J, Szathmáry E. 1999. The origins of life: From the birth of life to the origin of language. Oxford, U.K.: Oxford University Press

Maynard Smith J. 1993. The role of sex in bacterial evolution. *J. Hered.* 84:326–327.

McCammon J, Gelin B, Karplus M. 1977. Dynamics of folded proteins. *Nature* 267:585–590.

McInerney J, Pisani D, O'Connell MJ. 2015. The ring of life hypothesis for eukaryote origins is supported by multiple kinds of data. *Philos. Trans. R. Soc. B Biol. Sci.*

370:20140323.

McInerney JO, Martin WF, Koonin E V., Allen JF, Galperin MY, Lane N, Archibald JM, Embley TM. 2011. Planctomycetes and eukaryotes: A case of analogy not homology. *BioEssays* 33:810–817.

van Meer G, Voelker DR, Feigenson GW. 2008. Membrane lipids: where they are and how they behave. *Nat. Rev. Mol. Cell Biol.* 9:112–124.

Mereschkowsky C. 1905. Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biologisches Centralblatt 25/18: 38–604, ed. *J. Rosenthal*.

Mereschkowsky C. 1910. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biol. Cent.* 30:353–367.

Meuer J, Kuettner HC, Zhang JK, Hedderich R, Metcalf WW. 2002. Genetic analysis of the archaeon Methanosarcina barkeri Fusaro reveals a central role for Ech hydrogenase and ferredoxin in methanogenesis and carbon fixation. *Proc. Natl. Acad. Sci. U.S.A.* 99:5632–5637.

Michener CD, Corliss JO, Cowan RS, Raven PH, Sabrowsky CW, Squires DF, Wharton GW. 1970. Systematics in support of biological research. Washington DC, USA: Division of Biology and Agriculture, National Research Council

Miller SL. 1953. A production of amino acids under possible primitive Earth conditions. *Science* 117:528–529.

Mira A, Ochman H, Moran N a. 2001. Deletional bias and the evolution of bacterial genomes. *Trends Genet.* 17:589–596.

Mitchell P. 1957. The origin of life and the formation and organising functions of natural membranes. In: Oparin AI, Pasynskii AG, Braunshtein AE, Pavlovskaya TE, editors. Proceedings of the First International Symposium on the Origin of Life on the Earth. Moscow: Academy of Sciences (USSR). p. 229–234.

Mitchell P. 1961. Coupling of phosphorylation to electron and hydrogen transfer by a chemi-osmotic type of mechanism. *Nature* 191:144–148.

Mitchell P. 1966. Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. *Biol. Rev.* 41:444–501.

Monnard PA, Deamer DW. 2002. Membrane self-assembly processes: Steps toward the first cellular life. *Anat. Rec.* 268:196–207.

Moran N a. 2002. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108:583–586.

Moran U, Phillips R, Milo R. 2010. SnapShot: key numbers in biology. *Cell* 141:1262–1262.e1.

Mott ML, Berger JM. 2007. DNA replication initiation: mechanisms and regulation in bacteria. *Nat. Rev. Microbiol.* 5:343–354.

Mulkidjanian AY, Bychkov AY, Dibrova D V., Galperin MY, Koonin E V. 2012. Origin of first cells at terrestrial, anoxic geothermal fields. *Proc. Natl. Acad. Sci. U.S.A.* 109:E821–E830.

Mulkidjanian AY, Galperin MY, Koonin E V. 2009. Co-evolution of primordial membranes and membrane proteins. *Trends Biochem. Sci.* 34:206–215.

Mulkidjanian AY, Makarova KS, Galperin MY, Koonin E V. 2007. Inventing the dynamo machine: the evolution of the F-type and V-type ATPases. *Nat. Rev. Microbiol.* 5:892–899.

Nagata Y, Futamura A, Miyauchi K, Takagi M. 1999. Two different types of dehalogenases, LinA and LinB, involved in γ- hexachlorocyclohexane degradation in Sphingomonas paucimobilis UT26 are localized in the periplasmic space without molecular processing. *J. Bacteriol.* 181:5409–5413.

Nagle JF, Zhang R, Tristram-Nagle S, Sun W, Petrache HI, Suter RM. 1996. X-ray structure determination of fully hydrated L alpha phase dipalmitoylphosphatidylcholine bilayers. *Biophys. J.* 70:1419–1431.

Nelson-Sathi S, Dagan T, Landan G, Janssen A, Steel M, McInerney JO, Deppenmeier U, Martin WF. 2012. Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc. Natl. Acad. Sci. U.S.A.*:6–11.

Nelson-Sathi S, Sousa FL, Roettger M, Lozada-Chávez N, Thiergart T, Janssen A, Bryant D, Landan G, Schönheit P, Siebers B, et al. 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517:77–80.

Nelson DL, Cox MM. 2013. Lehninger Principles of Biochemistry. 6th ed. New York, NY, USA: W.H. Freeman & Company

Nicholls DG, Ferguson SJ. 2013. Bioenergetics. Fourth Ed. London: Academic Press

Nichols J, Deamer D. 1980. Net proton-hydroxyl permeability of large unilamellar liposomes measured by an acid-base titration technique. *Proc. Natl. Acad. Sci. U.S.A.* 77:2038–2042.

Nozaki Y, Tanford C. 1981. Proton and hydroxide ion permeability of phospholipid vesicles. *Proc. Natl. Acad. Sci. U.S.A.* 78:4324–4328.

O’Donnell M, Langston L, Stillman B. 2013. Principles and concepts of DNA replication in bacteria, archaea, and eukarya. *Cold Spring Harb. Perspect. Biol.* 5:a010108.

Oberai A, Ihm Y, Kim S, Bowie JU. 2006. A limited universe of membrane protein families and folds. *Protein Sci.* 15:1723–1734.

Oberai A, Joh NH, Pettit FK, Bowie JU. 2009. Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc. Natl. Acad. Sci. U.S.A.* 106:17747–17750.

Ohno S, Wolf U, Atkin NB. 1968. Evolution from fish to mammals by gene duplication. *Hereditas* 59:169–187.

Ourisson G, Nakatani Y. 1994. The terpenoid theory of the origin of cellular life: the evolution of terpenoids to cholesterol. *Chem. Biol.* 1:11–23.

Overington J, Al-Lazikani B, Hopkins A. 2006. How many drug targets are there? *Nat. Rev. Drug Discov.* 5:993–996.

Oxford Dictionaries. 2015a. Life. *Oxford Univ. Press*. Available from: http://www.oxforddictionaries.com/definition/english/life

Oxford Dictionaries. 2015b. Chiral. *Oxford Univ. Press*.

Oxford Dictionaries. 2016. Phylogeny. *Oxford Univ. Press*. Available from: http://www.oxforddictionaries.com/definition/english/phylogeny

Le Page M. 2014. Meet your maker: Homing in on the ancestor of all life. *New Sci.* 223:30–33.

Pál C, Papp B, Lercher MJ. 2005. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat. Genet.* 37:1372–1375.

Pascal R, Boiteau L. 2011. Energy flows, metabolism and translation. *Phil. Trans. R. Soc. B* 366:2949–2958.

Pasteur L. 1848. Sur les relations qui peuvent exister entre la forme cristalline, la composition chimique et le sens de la polarisation rotatoire. *Ann. Chim. Phys.* 24:442–459.

Penny D, Poole A. 1999. The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.* 9:672–677.

Peretó J, López-García P, Moreira D. 2004. Ancestral lipid biosynthesis and early membrane evolution. *Trends Biochem. Sci.* 29:469–477.

Philippe H, Douady CJ. 2003. Horizontal gene transfer and phylogenetics. *Curr. Opin. Microbiol.* 6:498–505.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F. 2005. Heterotachy and long-branch attraction in phylogenetics. *BMC Evol. Biol.* 5:50.

Phillips G, Chikwana VM, Maxwell A, El-Yacoubi B, Swairjo M a, Iwata-Reuyl D, de Crécy-Lagard V. 2010. Discovery and characterization of an amidinotransferase involved in the modification of archaeal tRNA. *J. Biol. Chem.* 285:12706–12713.

Pinti D. 2005. The origin and evolution of the oceans. *Lect. Astrobiol.* I:83–112.

Pisani D, Cotton J a., McInerney JO. 2007. Supertrees disentangle the chimerical origin of eukaryotic genomes. *Mol. Biol. Evol.* 24:1752–1760.

Pizzarello S, Cronin J. 2000. Non-racemic amino acids in the Murray and Murchison meteorites. *Geochim. Cosmochim. Acta* 64:329–338.

Plotkin JB, Dushoff J, Fraser HB. 2004. Detecting selection using a single genome sequence of M. tuberculosis and P. falciparum. *Nature* 428:942–945.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* 5:e9490.

Proskurowski G, Lilley MD, Kelley DS, Olson EJ. 2006. Low temperature volatile production at the Lost City Hydrothermal Field, evidence from a hydrogen stable isotope geothermometer. *Chem. Geol.* 229:331–343.

Puigbò P, Lobkovsky AE, Kristensen DM, Wolf YI, Koonin E V. 2014. Genomes in turmoil: Quantification of genome dynamics in prokaryote supergenomes. *BMC Biol.* 12:66.

R Core Team. 2014. R: A Language and Environment for Statistical Computing.

Rachel R, Wyschkony I, Riehl S, Huber H. 2002. The ultrastructure of Ignicoccus:

evidence for a novel outer membrane and for intracellular vesicle budding in an archaeon. *Archaea* 1:9–18.

Ravenhall M, Škunca N, Lassalle F, Dessimoz C. 2015. Inferring horizontal gene transfer. *PLoS Comput. Biol.* 11:e1004095.

Rawls KS, Martin JH, Maupin-Furlow JA. 2011. Activity and transcriptional regulation of bacterial protein-like glycerol-3-phosphate dehydrogenase of the haloarchaea in Haloferax volcanii. *J. Bacteriol.* 193:4469–4476.

Raymann K, Brochier-Armanet C, Gribaldo S. 2015. The two-domain tree of life is linked to a new root for the Archaea. *Proc. Natl. Acad. Sci. U.S.A.* 112:201420858.

Razin S, Tully JG. 1970. Cholesterol requirement of mycoplasmas. *J. Bacteriol.* 102:306–310.

Rehling P, Brandner K, Pfanner N. 2004. Mitochondrial import and the twin-pore translocase. *Nat. Rev. Mol. Cell Biol.* 5:519–530.

Reynolds SM, Käll L, Riffle ME, Bilmes JA, Noble WS. 2008. Transmembrane topology and signal peptide prediction using dynamic bayesian networks. *PLoS Comput. Biol.* 4:e1000213.

Rich PR. 2003. The molecular machinery of Keilin's respiratory chain. *Biochem. Soc. Trans.* 31:1095–1105.

Rivera MC, Lake JA. 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature* 431:152–155.

Rizzo PJ. 2003. Those amazing dinoflagellate chromosomes. *Cell Res.* 13:215–217.

Rohlin L, Leon DR, Kim U, Loo J a, Ogorzalek Loo RR, Gunsalus RP. 2012. Identification of the major expressed S-layer and cell surface-layer-related proteins in the model methanogenic archaea: Methanosarcina barkeri Fusaro and Methanosarcina acetivorans C2A. *Archaea* 2012:873589.

Rose PW, Prlić A, Bi C, Bluhm WF, Christie CH, Dutta S, Green RK, Goodsell DS, Westbrook JD, Woo J, et al. 2015. The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Res.* 43:D345–D356.

Russell MJ, Daniel RM, Hall AJ, Sherringham JA. 1994. A hydrothermally precipitated catalytic iron sulphide membrane as a first step toward life. *J. Mol. Evol.* 39:231–243.

Russell MJ, Daniel RM, Hall AJ. 1993. On the emergence of life via catalytic iron-sulphide membranes. *Terra Nova* 5:343–347.

Russell MJ, Hall AJ, Turner D. 1989. In vitro growth of iron sulphide chimneys: possible culture chambers for origin-of-life experiments. *Terra Nova* 1:238–241.

Russell MJ, Martin W. 2004. The rocky roots of the acetyl-CoA pathway. *Trends Biochem. Sci.* 29:358–363.

Sagan L. 1967. On the origin of mitosing cells. *J. Theor. Biol.* 14:255–274.

Sanders CR, Mittendorf KF. 2011. Tolerance to changes in membrane lipid composition as a selected trait of membrane proteins. *Biochemistry* 50:7858–

7867.

Sazanov L, Hinchliffe P. 2006. Structure of the hydrophilic domain of respiratory complex I from *Thermus thermophilus*. *Science* 311:1430–1436.

Schlegel K, Leone V, Faraldo-Gómez JD, Müller V. 2012. Promiscuous archaeal ATP synthase concurrently coupled to Na^+ and H^+ translocation. *Proc. Natl. Acad. Sci. U.S.A.* 109:947–952.

Schlegel S, Klepsch M, Gialama D, Wickström D, Slotboom DJ, de Gier J-W. 2010. Revolutionizing membrane protein overexpression in bacteria. *Microb. Biotechnol.* 3:403–411.

Schmidt TH, Kandt C. 2012. LAMBADA and InflateGRO2: efficient membrane alignment and insertion of membrane proteins for molecular dynamics simulations. *J. Chem. Inf. Model.* 52:2657–2669.

Schrödinger LLC. 2010. The PyMOL Molecular Graphics System, Version 1.3.

Schrödinger E. 1944. What is life? Cambridge, UK: Cambridge University Press

Sczepanski JT, Joyce GF. 2014. A cross-chiral RNA polymerase ribozyme. *Nature* 515:440–442.

Senior AE, Wise JG. 1983. The proton-ATPase of bacteria and mitochondria. *J. Memb. Biol.* 73:105–124.

Shimada H, Yamagishi A. 2011. Stability of heterochiral hybrid membrane made of bacterial sn-G3P lipids and archaeal sn-G1P lipids. *Biochemistry* 50:4114–4120.

Siebers B, Schönheit P. 2005. Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. *Curr. Opin. Microbiol.* 8:695–705.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7:539.

Singer SJ, Nicolson GL. 1972. The fluid mosaic model of the structure of cell membranes. *Science* 175:720–731.

Skulachev VP. 1988. Membrane bioenergetics. Berlin, Germany: Springer-Verlag

Sojo V, Herschy B, Whicher A, Camprubi E, Lane N. 2016. The origin of life in alkaline hydrothermal vents. *Astrobiology* 16:181–197.

Sojo V, Pomiankowski A, Lane N. 2014. A bioenergetic basis for membrane divergence in archaea and bacteria. *PLoS Biol.* 12:e1001926.

Sojo V. 2015. On the biogenic origins of homochirality. *Orig. Life Evol. Biosph.* 45:219–224.

Soucy SM, Huang J, Gogarten JP. 2015. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* 16:472–482.

Sousa FL, Nelson-Sathi S, Martin WF. 2016. One step beyond a ribosome: the ancient anaerobic core. *Biochim. Biophys. Acta*:(in press).

Sousa FL, Thiergart T, Landan G, Nelson-Sathi S, Pereira IAC, Allen JF, Lane N, Martin WF. 2013. Early bioenergetic evolution. *Phil. Trans. R. Soc. B* 368:1–30.

Spang A, Saw JH, Jørgensen SL, Zaremba-Niedzwiedzka K, Martijn J, Lind AE, van Eijk R, Schleper C, Guy L, Ettema TJG. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521:173–179.

van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC. 2005. GROMACS: fast, flexible, and free. *J. Comput. Chem.* 26:1701–1718.

Stanier RY, van Niel CB. 1962. The concept of a bacterium. *Arch. Mikrobiol.* 42:17–35.

Stock D, Leslie A, Walker J. 1999. Molecular architecture of the rotary motor in ATP synthase. *Science* 286:1700–1705.

Storbeck S, Rolfes S, Raux-Deery E, Warren MJ, Jahn D, Layer G. 2010. A novel pathway for the biosynthesis of heme in Archaea: genome-based bioinformatic predictions and experimental evidence. *Archaea* 2010:175050.

Streif S, Staudinger WF, Marwan W, Oesterhelt D. 2008. Flagellar rotation in the archaeon Halobacterium salinarum depends on ATP. *J. Mol. Biol.* 384:1–8.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* 34:W609–W612.

Szostak JW, Bartel DP, Luisi PL. 2001. Synthesizing life. *Nature* 409:387–390.

Taglicht D, Padan E, Schuldiner S. 1991. Overproduction and purification of a functional Na+/H+ antiporter coded by nhaA (ant) from Escherichia coli. *J. Biol. Chem.* 266:11289–11294.

Taglicht D, Padan E, Schuldiner S. 1993. Proton-Sodium Stoichiometry of NhaA, an Electrogenic Antiporter from *Escherichia coli*. *J. Biol. Chem.*:5382–5387.

Tatsuta T, Scharwey M, Langer T. 2014. Mitochondrial lipid trafficking. *Trends Cell Biol.* 24:44–52.

Tatum EL, Lederberg J. 1947. Gene Recombination in the Bacterium Escherichia coli. *J. Bacteriol.* 53:673–684.

Timmis JN, Ayliffe MA, Huang CY, Martin W. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat. Rev. Genet.* 5:123–135.

Tourasse NJ, Li WH. 2000. Selective constraints, amino acid composition, and the rate of protein evolution. *Mol. Biol. Evol.* 17:656–664.

Tovar J, León-Avila G, Sánchez LB, Sutak R, Tachezy J, van der Giezen M, Hernández M, Müller M, Lucocq JM. 2003. Mitochondrial remnant organelles of Giardia function in iron-sulphur protein maturation. *Nature* 426:172–176.

Tsirigos KD, Peters C, Shu N, Käll L, Elofsson A. 2015. The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.* 43:W401–W407.

Tusnády GE, Dosztányi Z, Simon I. 2004. Transmembrane proteins in the Protein Data Bank: identification and classification. *Bioinformatics* 20:2964–2972.

Uchiyama I, Higuchi T, Kawai M. 2010. MBGD update 2010: toward a comprehensive resource for exploring microbial genome diversity. *Nucleic Acids Res.* 38:D361–D365.

Uchiyama I. 2003. MBGD: microbial genome database for comparative analysis. *Nucleic Acids Res.* 31:58–62.

Uchiyama I. 2007. MBGD: a platform for microbial comparative genomics based on the automated construction of orthologous groups. *Nucleic Acids Res.* 35:D343–D346.

Urata H, Sasaki R, Morita H, Kusumoto M, Ogawa Y, Mitsuda K, Akagi M. 2005. Kinetic analysis of hydrolytic reaction of homo- and heterochiral adenylyl(3'-5')adenosine isomers: breaking homochirality reduces hydrolytic stability of RNA. *Chem. Commun. (Camb).*:2578–2580.

Valentine D. 2007. Adaptations to energy stress dictate the ecology and evolution of the Archaea. *Nat. Rev. Microbiol.* 5:1070–1077.

Verhees CH, Kengen SWM, Tuininga JE, Schut GJ, Adams MWW, De Vos WM, Van Der Oost J. 2003. The unique features of glycolytic pathways in Archaea. *Biochem. J.* 375:231–246.

Viklund H, Bernsel A, Skwark M, Elofsson A. 2008. SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24:2928–2929.

Viklund H, Elofsson A. 2008. OCTOPUS: improving topology prediction by two-track ANN-based preference scores and an extended topological grammar. *Bioinformatics* 24:1662–1668.

Vinogradov AD. 1998. Catalytic properties of the mitochondrial NADH-ubiquinone oxidoreductase (complex I) and the pseudo-reversible active/inactive enzyme transition. *Biochim. Biophys. Acta* 1364:169–185.

Volkman SK, Hartl DL, Wirth DF, Nielsen KM, Choi M, Batalov S, Zhou Y, Plouffe D, Le Roch KG, Abagyan R, et al. 2002. Excess polymorphisms in genes for membrane proteins in Plasmodium falciparum. *Science* 298:216–218.

van de Vossenberg JLCM, Ubbink-Kok T, Elferink MGL, Driessen AJM, Konings WN. 1995. Ion permeability of the cytoplasmic membrane limits the maximum growth temperature of bacteria and archaea. *Mol. Microbiol.* 18:925–932.

Wächtershäuser G. 2003. From pre-cells to Eukarya – a tale of two lipids. *Mol. Microbiol.* 47:13–22.

Wasserstein RL, Lazar NA. 2016. The ASA's statement on p-values: context, process, and purpose. *Am. Stat.* (in press):doi:10.1080/00031305.2016.1154108.

Werner F, Grohmann D. 2011. Evolution of multisubunit RNA polymerases in the three domains of life. *Nat. Rev. Microbiol.* 9:85–98.

Williams T a., Embley TM. 2014. Archaeal "dark matter" and the origin of eukaryotes. *Genome Biol. Evol.* 6:474–481.

Williams TA, Foster PG, Cox CJ, Embley TM. 2013. An archaeal origin of eukaryotes supports only two primary domains of life. *Nature* 504:231–236.

Williams TA, Foster PG, Nye TMW, Cox CJ, Embley TM. 2012. A congruent phylogenomic signal places eukaryotes within the Archaea. *Proc. R. Soc. B* 279:4870–4879.

Woese C. 1998. The universal ancestor. *Proc. Natl. Acad. Sci. U.S.A.* 95:6854–6859.

Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* 87:4576–4579.

Woese CR. 1987. Bacterial evolution. *Microbiology* 51:221–271.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–1591.

Yoshida M, Muneyuki E, Hisabori T. 2001. ATP synthase-a marvellous rotary engine of the cell. *Nat. Rev. Mol. Cell Biol.* 2:669–677.

You K, Jr LA, Allison W, Kaplan N. 1978. Enzyme stereospecificities for nicotinamide nucleotides. *Trends Biochem. Sci.* 3:265–268.

Young D. 2001. Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems. Wiley-Blackwell

Zhang J, Kumar S. 1997. Detection of convergent and parallel evolution at the amino acid sequence level. *Mol. Biol. Evol.* 14:527–536.

Zhu C, Meador TB, Dummann W, Hinrichs KU. 2014. Identification of unusual butanetriol dialkyl glycerol tetraether and pentanetriol dialkyl glycerol tetraether lipids in marine sediments. *Rapid Commun. Mass Spectrom.* 28:332–338.

Zillig W. 1991. Comparative biochemistry of Archaea and Bacteria. *Curr. Opin. Genet. &amp; Dev.* 1:544–551.

Zuckerkandl E, Pauling L. 1965. Molecules as documents of evolutionary history. *J. Theor. Biol.* 8:357–366.

This sketch is ***most*** *imperfect; but in so short a space I cannot make it better. Your imagination must fill up very wide blanks.*

Charles Darwin (1858)